

Probabilistic and Causal Inference

The Works of Judea Pearl

Hector Geffner
Rina Dechter
Joseph Y. Halpern
(Editors)



ASSOCIATION FOR COMPUTING MACHINERY

**Probabilistic
and Causal Inference:
The Works of Judea Pearl**

ACM Books

Editors in Chief

Sanjiva Prasad, *Indian Institute of Technology (IIT) Delhi, India*

Marta Kwiatkowska, *University of Oxford, UK*

Charu Aggarwal, *IBM Corporation, USA*

ACM Books is a new series of high-quality books for the computer science community, published by ACM in collaboration with Morgan & Claypool Publishers. ACM Books publications are widely distributed in both print and digital formats through booksellers and to libraries (and library consortia) and individual ACM members via the ACM Digital Library platform.

Event Mining for Explanatory Modeling

Laleh Jalali, *University of California, Irvine (UCI), Hitachi America Ltd.*

Ramesh Jain, *University of California, Irvine (UCI)*

2021

Intelligent Computing for Interactive System Design: Statistics, Digital Signal Processing, and Machine Learning in Practice

Editors: Parisa Eslambolchilar, *Cardiff University, Wales, UK*

Andreas Komninos, *University of Patras, Greece*

Mark Dunlop, *Strathclyde University, Scotland, UK*

2021

Semantic Web for the Working Ontologist: Effective Modeling for Linked Data, RDFS, and OWL, Third Edition

Dean Allemang, *Working Ontologist LLC*

Jim Hendler, *Rensselaer Polytechnic Institute*

Fabien Gandon, *INRIA*

2020

Code Nation: Personal Computing and the Learn to Program Movement in America

Michael J. Halvorson, *Pacific Lutheran University*

2020

Computing and the National Science Foundation, 1950–2016: Building a Foundation for Modern Computing

Peter A. Freeman, *Georgia Institute of Technology*

W. Richards Adrion, *University of Massachusetts Amherst*

William Aspray, *University of Colorado Boulder*

2019

Providing Sound Foundations for Cryptography: On the work of Shafi Goldwasser and Silvio Micali

Oded Goldreich, *Weizmann Institute of Science*

2019

Concurrency: The Works of Leslie Lamport

Dahlia Malkhi, *VMware Research* and *Calibra*

2019

The Essentials of Modern Software Engineering: Free the Practices from the Method Prisons!

Ivar Jacobson, *Ivar Jacobson International*

Harold "Bud" Lawson, *Lawson Konsult AB (deceased)*

Pan-Wei Ng, *DBS Singapore*

Paul E. McMahon, *PEM Systems*

Michael Goedicke, *Universität Duisburg–Essen*

2019

Data Cleaning

Ihab F. Ilyas, *University of Waterloo*

Xu Chu, *Georgia Institute of Technology*

2019

Conversational UX Design: A Practitioner's Guide to the Natural Conversation Framework

Robert J. Moore, *IBM Research–Almaden*

Raphael Arar, *IBM Research–Almaden*

2019

Heterogeneous Computing: Hardware and Software Perspectives

Mohamed Zahran, *New York University*

2019

Hardness of Approximation Between P and NP

Aviad Rubinfeld, *Stanford University*

2019

The Handbook of Multimodal-Multisensor Interfaces, Volume 3: Language Processing, Software, Commercialization, and Emerging Directions

Editors: Sharon Oviatt, *Monash University*

Björn Schuller, *Imperial College London and University of Augsburg*

Philip R. Cohen, *Monash University*

Daniel Sonntag, *German Research Center for Artificial Intelligence (DFKI)*

Gerasimos Potamianos, *University of Thessaly*

Antonio Krüger, *Saarland University and German Research Center for Artificial Intelligence (DFKI)*

2019

Making Databases Work: The Pragmatic Wisdom of Michael Stonebraker

Editor: Michael L. Brodie, *Massachusetts Institute of Technology*

2018

The Handbook of Multimodal-Multisensor Interfaces, Volume 2: Signal Processing, Architectures, and Detection of Emotion and Cognition

Editors: Sharon Oviatt, *Monash University*

Björn Schuller, *University of Augsburg and Imperial College London*

Philip R. Cohen, *Monash University*

Daniel Sonntag, *German Research Center for Artificial Intelligence (DFKI)*

Gerasimos Potamianos, *University of Thessaly*

Antonio Krüger, *Saarland University and German Research Center for Artificial Intelligence (DFKI)*

2018

Declarative Logic Programming: Theory, Systems, and Applications

Editors: Michael Kifer, *Stony Brook University*

Yanhong Annie Liu, *Stony Brook University*

2018

The Sparse Fourier Transform: Theory and Practice

Haitham Hassanieh, *University of Illinois at Urbana-Champaign*

2018

The Continuing Arms Race: Code-Reuse Attacks and Defenses

Editors: Per Larsen, *Immunant, Inc.*

Ahmad-Reza Sadeghi, *Technische Universität Darmstadt*

2018

Frontiers of Multimedia Research

Editor: Shih-Fu Chang, *Columbia University*

2018

Shared-Memory Parallelism Can Be Simple, Fast, and Scalable

Julian Shun, *University of California, Berkeley*

2017

Computational Prediction of Protein Complexes from Protein Interaction Networks

Sriganesh Srihari, *The University of Queensland Institute for Molecular Bioscience*

Chern Han Yong, *Duke-National University of Singapore Medical School*

Limsoon Wong, *National University of Singapore*

2017

The Handbook of Multimodal-Multisensor Interfaces, Volume 1: Foundations, User Modeling, and Common Modality Combinations

Editors: Sharon Oviatt, *Incaa Designs*

Björn Schuller, *University of Passau and Imperial College London*

Philip R. Cohen, *Voicebox Technologies*

Daniel Sonntag, *German Research Center for Artificial Intelligence (DFKI)*
Gerasimos Potamianos, *University of Thessaly*
Antonio Krüger, *Saarland University and German Research Center for Artificial Intelligence (DFKI)*
2017

Communities of Computing: Computer Science and Society in the ACM
Thomas J. Misa, Editor, *University of Minnesota*
2017

Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining
ChengXiang Zhai, *University of Illinois at Urbana–Champaign*
Sean Massung, *University of Illinois at Urbana–Champaign*
2016

An Architecture for Fast and General Data Processing on Large Clusters
Matei Zaharia, *Stanford University*
2016

Reactive Internet Programming: State Chart XML in Action
Franck Barbier, *University of Pau, France*
2016

Verified Functional Programming in Agda
Aaron Stump, *The University of Iowa*
2016

The VR Book: Human-Centered Design for Virtual Reality
Jason Jerald, *NextGen Interactions*
2016

Ada's Legacy: Cultures of Computing from the Victorian to the Digital Age
Robin Hammerman, *Stevens Institute of Technology*
Andrew L. Russell, *Stevens Institute of Technology*
2016

Edmund Berkeley and the Social Responsibility of Computer Professionals
Bernadette Longo, *New Jersey Institute of Technology*
2015

Candidate Multilinear Maps
Sanjam Garg, *University of California, Berkeley*
2015

Smarter Than Their Machines: Oral Histories of Pioneers in Interactive Computing
John Cullinane, *Northeastern University; Mossavar-Rahmani Center for Business and Government, John F. Kennedy School of Government, Harvard University*
2015

A Framework for Scientific Discovery through Video Games

Seth Cooper, *University of Washington*

2014

Trust Extension as a Mechanism for Secure Code Execution on Commodity Computers

Bryan Jeffrey Parno, *Microsoft Research*

2014

Embracing Interference in Wireless Systems

Shyamnath Gollakota, *University of Washington*

2014



Photo credit: UCLA Samueli School of Engineering

Probabilistic and Causal Inference: The Works of Judea Pearl

Hector Geffner, editor

ICREA and Universitat Pompeu Fabra

Rina Dechter, editor

University of California, Irvine

Joseph Y. Halpern, editor

Cornell University

ACM Books #36



Copyright © 2022 by Association for Computing Machinery

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means—electronic, mechanical, photocopy, recording, or any other except for brief quotations in printed reviews—without the prior permission of the publisher.

Designations used by companies to distinguish their products are often claimed as trademarks or registered trademarks. In all instances in which the Association of Computing Machinery is aware of a claim, the product names appear in initial capital or all capital letters. Readers, however, should contact the appropriate companies for more complete information regarding trademarks and registration.

Probabilistic and Causal Inference: The Works of Judea Pearl
Hector Geffner, Rina Dechter, Joseph Y. Halpern, Editors

books.acm.org
<http://books.acm.org>

ISBN: 978-1-4503-9586-1 hardcover
ISBN: 978-1-4503-9587-8 paperback
ISBN: 978-1-4503-9588-5 EPUB
ISBN: 978-1-4503-9589-2 eBook

Series ISSN: 2374-6769 print 2374-6777 electronic

DOIs:

10.1145/3501714 Book	10.1145/3501714.3501738 Chapter 22
10.1145/3501714.3501715 Preface	10.1145/3501714.3501739 Chapter 23
10.1145/3501714.3501716 Credits	10.1145/3501714.3501740 Chapter 24
10.1145/3501714.3501717 Chapter 1	10.1145/3501714.3501741 Chapter 25
10.1145/3501714.3501718 Chapter 2	10.1145/3501714.3501742 Chapter 26
10.1145/3501714.3501719 Chapter 3	10.1145/3501714.3501743 Chapter 27
10.1145/3501714.3501720 Chapter 4	10.1145/3501714.3501744 Chapter 28
10.1145/3501714.3501721 Chapter 5	10.1145/3501714.3501745 Chapter 29
10.1145/3501714.3501722 Chapter 6	10.1145/3501714.3501746 Chapter 30
10.1145/3501714.3501723 Chapter 7	10.1145/3501714.3501747 Chapter 31
10.1145/3501714.3501724 Chapter 8	10.1145/3501714.3501748 Chapter 32
10.1145/3501714.3501725 Chapter 9	10.1145/3501714.3501749 Chapter 33
10.1145/3501714.3501726 Chapter 10	10.1145/3501714.3501750 Chapter 34
10.1145/3501714.3501727 Chapter 11	10.1145/3501714.3501751 Chapter 35
10.1145/3501714.3501728 Chapter 12	10.1145/3501714.3501752 Chapter 36
10.1145/3501714.3501729 Chapter 13	10.1145/3501714.3501753 Chapter 37
10.1145/3501714.3501730 Chapter 14	10.1145/3501714.3501754 Chapter 38
10.1145/3501714.3501731 Chapter 15	10.1145/3501714.3501755 Chapter 39
10.1145/3501714.3501732 Chapter 16	10.1145/3501714.3501756 Chapter 40
10.1145/3501714.3501733 Chapter 17	10.1145/3501714.3501757 Chapter 41
10.1145/3501714.3501734 Chapter 18	10.1145/3501714.3501758 Chapter 42
10.1145/3501714.3501735 Chapter 19	10.1145/3501714.3501759 Chapter 43
10.1145/3501714.3501736 Chapter 20	10.1145/3501714.3501760 Bios/Index
10.1145/3501714.3501737 Chapter 21	

A publication in the ACM Books series, #36

Editors in Chief: Sanjiva Prasad, *Indian Institute of Technology (IIT) Delhi, India*

Marta Kwiatkowska, *University of Oxford, UK*

Charu Aggarwal, *IBM Corporation, USA*

Area Editor: M. Tamer Ozsü, *University of Waterloo*

This book was typeset in Arnhem Pro 10/14 and Flama using pdf \TeX .

First Edition

10 9 8 7 6 5 4 3 2 1

Contents

Preface xxv

Credits xxvii

PART I INTRODUCTION 1

Chapter 1 Biography of Judea Pearl by Stuart J. Russell 3

References 9

Chapter 2 Turing Award Lecture 11

References 27

Chapter 3 Interview by Martin Ford 29

References 42

Chapter 4 An Interview with Ron Wassertein on How *The Book of Why* Transforms Statistics 43

Chapter 5 Selected Annotated Bibliography by Judea Pearl 49

Search and Heuristics 49

Bayesian Networks 50

Causality 51

Causal, Casual, and Curious 53

PART II HEURISTICS 57

Chapter 6 Introduction by Judea Pearl 59

References 60

Chapter 7 Asymptotic Properties of Minimax Trees and Game-Searching

Procedures 61

Judea Pearl

Abstract 61

- 7.1 The Probability of Winning a Standard h -level Game Tree with Random WIN Positions 62
 - 7.2 Game Trees with an Arbitrary Distribution of Terminal Values 65
 - 7.3 The Mean Complexity of Solving (h, d, P_0) -game 69
 - 7.4 Solving, Testing, and Evaluating Game Trees 75
 - 7.5 Test and, if Necessary, Evaluate—The SCOUT Algorithm 78
 - 7.6 Analysis of SCOUT's Expected Performance 79
 - 7.7 On the Branching Factor of the ALPHA-BETA $(\alpha-\beta)$ procedure 85
- References 88

Chapter 8 The Solution for the Branching Factor of the Alpha-Beta Pruning Algorithm and its Optimality 91

Judea Pearl

- 8.1 Introduction 92
 - 8.2 Analysis 94
 - 8.3 Conclusions 101
- References 102

Chapter 9 On the Discovery and Generation of Certain Heuristics 103

Judea Pearl

Abstract 103

- 9.1 Introduction: Typical Uses of Heuristics 103
 - 9.2 Mechanical Generation of Admissible Heuristics 114
 - 9.3 Can a Program Tell an Easy Problem When It Sees One? 117
 - 9.4 Conclusions 119
- References 121

PART III PROBABILITIES 123

Chapter 10 Introduction by Judea Pearl 125

References 126

Chapter 11 Reverend Bayes on Inference Engines: A Distributed Hierarchical Approach 129

Judea Pearl

Abstract 129

- 11.1 Introduction 129
- 11.2 Definitions and Nomenclature 131
- 11.3 Structural Assumptions 131
- 11.4 Combining Top and Bottom Evidences 132
- 11.5 Propagation of Information Through the Network 134
- 11.6 A Token Game Illustration 135
- 11.7 Properties of the Updating Scheme 136
- 11.8 A Summary of Proofs 136
- 11.9 Conclusions 137
- References 137

Chapter 12 Fusion, Propagation, and Structuring in Belief Networks 139

Judea Pearl

Abstract 139

- 12.1 Introduction 140
- 12.2 Fusion and Propagation 148
- 12.3 Structuring Causal Trees 169
- 12.A Appendix A. Derivation of the Updating Rules for Singly Connected Networks 181
- 12.B Appendix B. Conditions for Star-decomposability 183
- Acknowledgments 185
- References 186

Chapter 13 GRAPHOIDS: Graph-Based Logic for Reasoning about Relevance Relations Or When Would x Tell You More about y If You Already Know z ? 189

Judea Pearl and Azaria Paz

Abstract 189

- 13.1 Introduction 190
- 13.2 Probabilistic Dependencies and their Graphical Representation 192
- 13.3 GRAPHOIDS 195
- 13.4 Special Graphoids and Open Problems 196
- 13.5 Conclusions 198
- References 199

Chapter 14 System Z: A Natural Ordering of Defaults with Tractable Applications to Nonmonotonic Reasoning 201

Judea Pearl

Abstract 201

- 14.1 Description 201
- 14.2 Consequence Relations 204
- 14.3 Illustrations 206
- 14.4 The Maximum Entropy Approach 208
- 14.5 Conditional Entailment 210
- 14.6 Conclusions 211
- Acknowledgments 211
- 14.I Appendix I: Uniqueness of The Minimal Ranking Function 211
- 14.II Appendix II: Rational Monotony of Admissible Rankings 213
- References 214

PART IV CAUSALITY 1988–2001 215

Chapter 15 Introduction by Judea Pearl 217

References 219

Chapter 16 Equivalence and Synthesis of Causal Models 221

TS Verma and Judea Pearl

Abstract 221

- 16.1 Introduction 222
- 16.2 Patterns of Causal Models 224
- 16.3 Embedded Causal Models 227
- 16.4 Applications to the Synthesis of Causal Models 231
- IC-Algorithm (Inductive Causation) 232
- Acknowledgments 234
- References 234

Chapter 17 Probabilistic Evaluation of Counterfactual Queries 237

Alexander Balke and Judea Pearl

Abstract 237

- 17.1 Introduction 237
- 17.2 Notation 240
- 17.3 Party Example 241
- 17.4 Probabilistic vs. Functional Specification 242
- 17.5 Evaluating Counterfactual Queries 245
- 17.6 Party Again 248
- 17.7 Special Case: Linear-Normal Models 250
- 17.8 Conclusion 252
- Acknowledgments 253

References 253

Chapter 18 Causal Diagrams for Empirical Research (With Discussions) 255

Judea Pearl

Summary 255

Some key words 255

- 18.1 Introduction 255
- 18.2 Graphical Models and the Manipulative Account of Causation 258
- 18.3 Controlling Confounding Bias 262
- 18.4 A Calculus of Intervention 265
- 18.5 Graphical Tests of Identifiability 269
- 18.6 Discussion 275
 - Acknowledgments 277
- 18.A Appendix 278
 - References 279
- 18.I Discussion of ‘Causal Diagrams for Empirical Research’ by J. Pearl 282
- 18.II Discussion of ‘Causal Diagrams for Empirical Research’ by J. Pearl 283
- 18.III Discussion of ‘Causal Diagrams for Empirical Research’ by J. Pearl 285
- 18.IV Discussion of ‘Causal Diagrams for Empirical Research’ by J. Pearl 287
- 18.V Discussion of ‘Causal Diagrams for Empirical Research’ by J. Pearl 290
- 18.VI Discussion of ‘Causal Diagrams for Empirical Research’ by J. Pearl 292
- 18.VII Discussion of ‘Causal Diagrams for Empirical Research’ by J. Pearl 296
- 18.VIII Discussion of ‘Causal Diagrams for Empirical Research’ by J. Pearl 299
- 18.IX Discussion of ‘Causal Diagrams for Empirical Research’ by J. Pearl 300
- 18.X Rejoinder to Discussions of ‘Causal Diagrams for Empirical Research’ 303
 - Additional References 313

Chapter 19 Probabilities of Causation: Three Counterfactual Interpretations and Their Identification 317

Judea Pearl

Abstract 317

- 19.1 Introduction 318
- 19.2 Structural Model Semantics (A Review) 321
- 19.3 Necessary and Sufficient Causes: Conditions of Identification 331
- 19.4 Examples and Applications 342
- 19.5 Identification in Non-Monotonic Models 351
- 19.6 From Necessity and Sufficiency to “Actual Cause” 354
- 19.7 Conclusion 364
- 19.A Appendix: The Empirical Content of Counterfactuals 365
 - References 368

Chapter 20 Direct and Indirect Effects 373

Judea Pearl

Abstract 373

20.1 Introduction 373

20.2 Conceptual Analysis 375

20.3 Formal Analysis 380

20.4 Conclusions 390

Acknowledgments 390

References 391

PART V CAUSALITY 2002–2020 393

Chapter 21 Introduction by Judea Pearl 395

References 396

Chapter 22 Comment: Understanding Simpson’s Paradox 399

Judea Pearl

22.1 The History 399

22.2 A Paradox Resolved 402

22.3 Armistead’s Critique 408

22.4 Conclusions 409

References 410

Chapter 23 Graphical Models for Recovering Probabilistic and Causal Queries from Missing Data 413

Karthika Mohan and Judea Pearl

Abstract 413

23.1 Introduction 413

23.2 Missingness Graph and Recoverability 414

23.3 Recovering Probabilistic Queries by Sequential Factorization 416

23.4 Recoverability in the Absence of an Admissible Sequence 418

23.5 Non-recoverability Criteria for Joint and Conditional Distributions 419

23.6 Recovering Causal Queries 420

23.7 Attrition 422

23.8 Related Work 423

23.9 Conclusion 424

Acknowledgments 424

References 424

23.A Appendix 426

Chapter 24 Recovering from Selection Bias in Causal and Statistical Inference 433*Elias Bareinboim, Jin Tian and Judea Pearl*Abstract **433**24.1 Introduction **433**24.2 Recoverability without External Data **437**24.3 Recoverability with External Data **440**24.4 Recoverability of Causal Effects **444**24.5 Conclusions **447**Acknowledgments **447**References **447****Chapter 25 External Validity: From *Do*-Calculus to Transportability Across Populations 451***Judea Pearl and Elias Bareinboim*Abstract **451**Key words and phrases **451**25.1 Introduction: Threats vs. Assumptions **452**25.2 Preliminaries: The Logical Foundations of Causal Inference **454**25.3 Inference Across Populations: Motivating Examples **461**25.4 Formalizing Transportability **465**25.5 Transportability of Causal Effects—A Graphical Criterion **471**25.6 Conclusions **475**25.A Appendix **477**Acknowledgments **478**References **478****Chapter 26 Detecting Latent Heterogeneity 483***Judea Pearl*Abstract **483**Keywords **483**26.1 Introduction **483**26.2 Covariate-Induced Heterogeneity **485**26.3 Latent Heterogeneity between the Treated and Untreated **488**26.4 Three Ways of Detecting Heterogeneity **490**26.5 Example: Heterogeneity in Recruitment **495**26.6 Conclusions **497**Acknowledgments **498**Declaration of Conflicting Interests **498**Funding **498**

- References 498
- Author Biography 501
- 26.A Appendix A (An Extreme Case of Latent Heterogeneity) 501
- 26.B Appendix B (Assessing Heterogeneity in Structural Equation Models) 503

PART VI CONTRIBUTED ARTICLES 507

Chapter 27 On Pearl's Hierarchy and the Foundations of Causal Inference 509

Elias Bareinboim, Juan D. Correa, Duligur Ibeling and Thomas Icard

- Abstract 509
- 27.1 Introduction 510
- 27.2 Structural Causal Models and the Causal Hierarchy 514
- 27.3 Pearl Hierarchy—A Logical Perspective 524
- 27.4 Pearl Hierarchy—A Graphical Perspective 533
- 27.5 Conclusions 551
- Acknowledgments 552
- References 552

Chapter 28 The Tale Wags the DAG 557

Philip Dawid

- Abstract 557
- 28.1 Introduction 557
- 28.2 The Ladder of Causation 558
- 28.3 Ground Level: Syntax 559
- 28.4 Rung 1: Seeing 560
- 28.5 Rung 2: Doing 564
- 28.6 Rung 3: Imagining 569
- 28.7 Conclusion 571
- References 572

Chapter 29 Instrumental Variables with Treatment-induced Selection: Exact Bias Results 575

Felix Elwert and Elan Segarra

- 29.1 Introduction 575
- 29.2 Causal Graphs 577
- 29.3 Instrumental Variables 579
- 29.4 Selection Bias in IV: Qualitative Analysis 580
- 29.5 Selection Bias in IV: Quantitative Analysis 581
- 29.6 Conclusion 588
- 29.A Appendix 589
- References 591

Chapter 30 Causal Models and Cognitive Development 593*Alison Gopnik*References **601****Chapter 31 The Causal Foundations of Applied Probability and Statistics 605***Sander Greenland*Abstract **605**

- 31.1 Introduction: Scientific Inference *is* a Branch of Causality Theory **606**
- 31.2 Causality is Central Even for Purely Descriptive Goals **608**
- 31.3 The Strength of Probabilistic Independence Demands Physical Independence **609**
- 31.4 The Superconducting Supercollider of Selection **610**
- 31.5 Data and Algorithms are Causes of Reported Results **611**
- 31.6 Getting Causality into Statistics by Putting Statistics into Causal Terms from the Start **612**
- 31.7 Causation in 20th-century Statistics **613**
- 31.8 Causal Analysis versus Traditional Statistical Analysis **614**
- 31.9 Relating Causality to Traditional Statistical Philosophies and “Objective” Statistics **616**
- 31.10 Discussion **618**
- 31.11 Conclusion **619**
- 31.A Appendix **619**
- Acknowledgments **620**
- References **620**

Chapter 32 Pearl on Actual Causation 625*Christopher Hitchcock*Abstract **625**

- 32.1 Introduction **625**
- 32.2 Actual Causation **625**
- 32.3 Causal Models and But-for Causation **626**
- 32.4 Pre-emption and Lewis **631**
- 32.5 Intransitivity and Overdetermination **634**
- 32.6 Pearl’s Definitions of Actual Causation **637**
- 32.7 Pearl’s Achievement **642**
- References **643**

Chapter 33 Causal Diagram and Social Science Research 647*Kosuke Imai*

- 33.1 Graphical Causal Models and Social Science Research **647**

- 33.2 Two Applications of Graphical Causal Models 648
- 33.3 The Future of Causal Research in the Social Sciences 652
- References 652

Chapter 34 Causal Graphs for Missing Data: A Gentle Introduction 655

Karthika Mohan

- 34.1 Introduction 655
- 34.2 Missingness Graphs 656
- 34.3 Recoverability 658
- 34.4 Testability 664
- References 666

Chapter 35 A Note of Appreciation 667

Azaria Paz

- 35.1 A Magic Square 668
- 35.2 A Magic Shield of David 668

Chapter 36 Causal Models for Dynamical Systems 671

Jonas Peters, Stefan Bauer and Niklas Pfister

- Abstract 671
- 36.1 Introduction 671
- 36.2 Chemical Reaction Networks and ODEs 675
- 36.3 Causal Kinetic Models 677
- 36.4 Challenges in Causal Inference for ODE-based Systems 681
- 36.5 From Invariance to Causality and Generalizability 682
- 36.6 Conclusions 683
- Acknowledgments 684
- References 684

Chapter 37 Probabilistic Programming Languages: Independent Choices and Deterministic Systems 691

David Poole and Frank Wood

- 37.1 Probabilistic Models and Deterministic Systems 693
- 37.2 Possible Worlds Semantics 694
- 37.3 Inference 700
- 37.4 Learning 703
- 37.5 Other Issues 704
- 37.6 Causal Models 705
- 37.7 Some Pivotal References 705

- 37.8 Conclusion 706
- References 707

Chapter 38 An Interventionist Approach to Mediation Analysis 713

James M. Robins, Thomas S. Richardson and Ilya Shpitser

- 38.1 Introduction 713
- 38.2 Approaches to Mediation Based on Counterfactuals Defined in Terms of the Mediator: The CDE and PDE 715
- 38.3 Interventionist Theory of Mediation 726
- 38.4 Path-Specific Counterfactuals 747
- 38.5 Conclusion 754
- Acknowledgments 754
- 38.A Appendix 754
- References 761

Chapter 39 Causality for Machine Learning 765

Bernhard Schölkopf

Abstract 765

- 39.1 Introduction 765
- 39.2 The Mechanization of Information Processing 766
- 39.3 From Statistical to Causal Models 769
- 39.4 Levels of Causal Modeling 773
- 39.5 Independent Causal Mechanisms 774
- 39.6 Cause–Effect Discovery 780
- 39.7 Half-sibling Regression and Exoplanet Detection 782
- 39.8 Invariance, Robustness, and Semi-supervised Learning 783
- 39.9 Causal Representation Learning 790
- 39.10 Personal Notes and Conclusion 793
- Acknowledgments 795
- References 795

Chapter 40 Why Did They Do That? 805

Ross Shachter and David Heckerman

Abstract 805

- 40.1 Introduction 805
- 40.2 Some Examples 806
- 40.3 Back to the Garden of Eden 807
- 40.4 Decision Theory and Decision Analysis 808
- 40.5 Back Again in the Garden of Eden 810

- 40.6 Conclusion: God's Decision 811
- References 812

Chapter 41 Multivariate Counterfactual Systems and Causal Graphical Models 813

Ilya Shpitser, Thomas S. Richardson and James M. Robins

- 41.1 Introduction 813
- 41.2 Graphs, Non-parametric Structural Equation Models, and the *g-do* Operator 820
- 41.3 The Potential Outcomes Calculus and Identification 833
- 41.4 Identification in Hidden Variable Causal Models 835
- 41.5 Conclusion 844
- Acknowledgments 845
- 41.A Appendix 845
- References 848

Chapter 42 Causal Bayes Nets as Psychological Theory 853

Steven A. Sloman

- Abstract 853
- 42.1 The Human Conception of Causality 854
- 42.2 Core Properties 856
- 42.3 The Broader Perspective: The Community of Knowledge 859
- 42.4 Collective Causal Models 861
- 42.5 Conclusion 863
- Acknowledgments 864
- References 864

Chapter 43 Causation: Objective or Subjective? 867

Wolfgang Spohn

- Abstract 867
- 43.1 Causation: A Bunch of Attitudes 867
- 43.2 The Model Relativity of Causation 871
- 43.3 Laws 874
- 43.4 Probability 878
- Acknowledgments 886
- References 886

Editors' Biographies 889

Index 893

Preface

In March 2010, we, the editors, (Rina, Hector, and Joe) organized a meeting in honor of Judea Pearl at the University of California, Los Angeles (UCLA), and edited a book made up of contributions from a number of authors about Judea's work (*Heuristics, Probability and Causality: A Tribute to Judea Pearl*, College Publications, 2010). Judea was happily surprised by the meeting and by the book, but after opening it he couldn't avoid being himself and only half jokingly said: "How come there are no papers by me?" Thus, this book is different from the previous one, also edited in Judea's honor, as it focuses on Judea's works, while also containing a number of contributions and commentaries by Judea's colleagues. The book was commissioned by the Association for Computing Machinery (ACM) as part of the ACM-Morgan Claypool series dedicated to the ACM Turing Award winners.

Judea Pearl won the ACM Turing Award prize in 2011 for "fundamental contributions to artificial intelligence through the development of a calculus for probabilistic and causal reasoning." Judea is the creator of Bayesian networks, a mathematical formalism for defining complex probability models, as well as the main algorithms that are used for inference in these models. The work not only revolutionized the field of artificial intelligence (AI) but also became an important tool for many other branches of engineering and the natural sciences. Judea later created a mathematical framework for causal and counterfactual inference that also is having a significant impact in the social sciences.

Judea started his research work in AI in the mid-1970s. AI has changed a great deal since then; arguably no one has played a larger role in that change than Judea. Judea Pearl's work made probability the prevailing language of modern AI and, perhaps more significantly, it placed the elaboration of crisp and meaningful models, and of effective computational mechanisms, at the center of AI research. His work is conveyed in hundreds of scientific publications and three landmark books *Heuristics: Intelligent Search Strategies for Computer Problem Solving* (Addison-Wesley Longman Publishing Co, Boston 1984), *Probabilistic Reasoning*

in *Intelligent Systems* (Morgan Kaufmann, California, 1988), and *Causality* (Cambridge University Press, New York 2000). His “burning questions” were (and still are): “How does the human mind ‘do it?’” and “How can a ‘stupid robot’ do it?” He set out to answer these questions with an unusual combination of intuition, passion, creativity, intellectual honesty, and technical skill.

For three decades now, Judea’s work has been focused on causality and counterfactuals, notions that are central to human reasoning, machine learning, and AI, and which have attracted the attention of philosophers for centuries. Central to the Causal Revolution, advocated and facilitated by him with the help of students and colleagues, is the language of causal diagrams. Judea finds the plain language of probability theory (or data) suitable only for associational reasoning, rung one of his three-level hierarchy (seeing), which also includes two other levels: interventions (doing) and counterfactuals (imagining). A technical introduction to these ideas can be found in his book *Causal Inference in Statistics: A Primer*, with Madeilyn Glymour and Nicholas P. Jewell (Wiley, London 2016). His most recent book, *The Book of Why: The New Science of Cause and Effect*, with Dana Mackenzie (Basic Books, New York 2018), is a more general and delightful introduction written for the general public.

This volume is organized into six parts, starting with an introduction with a biography, interviews, and transcript of Judea Pearl’s Turing Award Lecture (delivered at the Association for the Advancement of Artificial Intelligence [AAAI] 2012), followed by selected seminal works by Judea and, in some cases, co-authors, organized into three themes: heuristic search, probabilities, and causality, the latter divided in two periods, 1998–2001 and 2001–2020. Judea Pearl himself was kind enough to write the introductions to the latter four parts. This is followed by articles and commentaries by distinguished colleagues from areas like machine learning and AI, computer science and engineering, statistics and the natural sciences, cognitive science, social sciences, and philosophy. The wide variety of areas shows the reach of Judea’s ideas and impact.

Two of the editors of this volume, Rina and Hector, are former students of Judea, and the third, Joe, is a close colleague and collaborator. The three of us would like to thank all the authors who contributed to the volume, both for their articles and their support. We also thank Kaoru Mulvihill, Judea’s assistant for almost 30 years, whose help with the book has been invaluable. It has been a privilege to edit this second book in Judea’s honor. We are indeed Judea’s fans: we love Judea and we admire him as an advisor, as a scientist, and as a great human being. It has been a unique privilege to know Judea and to learn from him.

Hector Geffner, Rina Dechter, and Joe Halpern

Credits

- Chapter 1 S. Russell. *Judea Pearl*. ACM Turing Awards page, https://amturing.acm.org/award_winners/pearl_2658896.cfm.
- Chapter 2 J. Pearl. *Turing Award Lecture*. ACM Turing Awards page, https://amturing.acm.org/vp/pearl_2658896.cfm.
- Chapter 3 M. Ford. 2018. *Architects of Intelligence: The Truth about AI from the People Building it*. Martin Ford, Packt Publishing, Mumbai.
- Chapter 4 R. Wasserstein. 2018. *Interview*. *Amstat News*, 8 September.
- Chapter 7 J. Pearl. 1980. Asymptotic properties of minimax trees and game-searching procedures. *Artif. Intell.* 14, 2, 113–138. DOI: [https://doi.org/10.1016/0004-3702\(80\)90037-5](https://doi.org/10.1016/0004-3702(80)90037-5).
- Chapter 8 J. Pearl. 1982. The solution for the branching factor of the alpha–beta pruning algorithm and its optimality. *Commun. ACM* 25, 8. DOI: <https://doi.org/10.1145/358589.358616>.
- Chapter 9 J. Pearl. 1983. On the discovery and generation of certain heuristics. *AI Mag.* 4, 1, 23. DOI: <https://doi.org/10.1609/aimag.v4i1.385>.
- Chapter 11 J. Pearl. 1982. Reverend Bayes on inference engines: A distributed hierarchical approach. *Proceeding of the 2nd AAAI Conference on Artificial Intelligence*, 133–136.
- Chapter 12 J. Pearl. 1986. Fusion, propagation, and structuring in belief networks. *Artif. Intell.* 29, 3, 241–288. DOI: [https://doi.org/10.1016/0004-3702\(86\)90072-X](https://doi.org/10.1016/0004-3702(86)90072-X).
- Chapter 13 J. Pearl and A. Paz. 1985. *GRAPHOIDS: A Graph-based Logic for Reasoning about Relevance Relationships*, Benedict du Boulay, David C. Hogg, Luc Steels: *Advances in Artificial Intelligence II*, Seventh European Conference on Artificial Intelligence, ECAI 1986, Brighton, UK, July 20-25, 1986, Proceedings. North-Holland 1987, ISBN 0-444-70279-2.

- Chapter 14 J. Pearl. 1990. System Z: A natural ordering of defaults with tractable applications to non-monotonic reasoning. In *Proceedings of the 3rd Conference on Theoretical Aspects of Reasoning about Knowledge (TARK)*. Morgan Kaufmann Publishers Inc., 121–135.
- Chapter 16 T. S. Verma and J. Pearl. 1991. Equivalence and synthesis of causal models. *Proceedings of the 7th Conference on Uncertainty in Artificial Intelligence*, Morgan Kaufmann Publishers Inc., 255–270.
- Chapter 17 A. Balke and J. Pearl. 1994. Probabilistic evaluation of counterfactual queries. *Proceedings of the 12th AAAI Conference on Artificial Intelligence*, 230–237.
- Chapter 18 J. Pearl. 1995. Causal diagrams for empirical research (with discussion). *Biometrika* 82, 4, 669–688. DOI: <https://doi.org/10.1093/biomet/82.4.669>.
- Chapter 19 J. Pearl. 1999. Probabilities of causation: Three counterfactual interpretations and their identification. *Synthese* 121, 93–149. DOI: <https://doi.org/10.1023/A:1005233831499>.
- Chapter 20 J. Pearl. 2001. Direct and indirect effects. *Proceedings of the 17th Conference on Uncertainty in Artificial Intelligence*, 411–420.
- Chapter 22 J. Pearl. 2014. Comment: Understanding Simpson’s paradox. *Am. Stat.* 68, 1, 8–13. DOI: <https://doi.org/10.1080/00031305.2014.876829>.
- Chapter 23 K. Mohan and J. Pearl. 2014. Graphical models for recovering probabilistic and causal queries from missing data. *Proceedings of the 27th International Conference on Neural Information Processing Systems* 1, 1520–1528.
- Chapter 24 E. Bareinboim, J. Tian, and J. Pearl. 2014. Recovering from election bias in causal and statistical inference. *Proceedings of the 28th AAAI Conference on Artificial Intelligence*.
- Chapter 25 J. Pearl and E. Bareinboim. 2014. External validity: From do-calculus to transportability across populations. *Stat. Sci.* 29, 4, 579–595. DOI: <https://doi.org/10.1214/14-STS486>.
- Chapter 26 J. Pearl. 2017. Detecting latent heterogeneity. *Sociol. Meth. Res.* 46, 370–389. DOI: <https://doi.org/10.1177/0049124115600597>.
- Chapter 28 P. Dawid. “The glass is falling hour by hour ...” from “Bagpipe music” by Louis MacNeice.

I
PART

INTRODUCTION

1

Biography of Judea Pearl by Stuart J. Russell



Judea Pearl created the representational and computational foundation for the processing of information under uncertainty.

He is credited with the invention of *Bayesian networks*, a mathematical formalism for defining complex probability models, as well as the principal algorithms used for inference in these models. This work not only revolutionized the field of artificial intelligence but also became an important tool for many other branches of engineering

and the natural sciences. He later created a mathematical framework for causal inference that has had significant impact in the statistical, health, and social sciences.

Judea Pearl was born on September 4, 1936, in Tel Aviv, which was at that time administered under the British Mandate for Palestine. He grew up in Bnei Brak, a Biblical town his grandfather came to re-establish in 1924. In 1956, after serving in the Israeli army and joining a kibbutz, Judea decided to study engineering. He attended the Technion, where he met his wife, Ruth, and received a BS degree in electrical engineering in 1960. Recalling the Technion faculty members in a 2012 interview in the *Technion Magazine*, he emphasized the thrill of discovery:

Originally published in ACM Turing Awards page, https://amturing.acm.org/award_winners/pearl_2658896.cfm.

Professor Franz Olendorf always spoke as if he was personally present in Cavendish laboratory, where the electron was discovered, Professor Abraham Ginzburg made us feel the winds blowing in our face as we travelled along those line integrals in the complex plane. And Professor Amiram Ron gave us the feeling that there is still something we can add to Maxwell's theory of electromagnetic waves.

Judea then went to the United States for graduate study, receiving an MS in electronics from Newark College of Engineering in 1961, an MS in physics from Rutgers University in 1965, and a PhD in electrical engineering from the Polytechnic Institute of Brooklyn in the same year. The title of his PhD thesis was *Vortex Theory of Superconductive Memories*; the term “Pearl vortex” has become popular among physicists to describe the type of superconducting current he studied. He worked at RCA Research Laboratories in Princeton, New Jersey, on superconductive parametric amplifiers and storage devices, and at Electronic Memories, Inc., in Hawthorne, California, on advanced memory systems. Despite the apparent focus on physical devices, Pearl reports being motivated even then by potential applications to intelligent systems.

When industrial research on magnetic and superconducting memories was curtailed by the advent of large-scale semiconductor memories, Pearl decided to move into academia to pursue his long-term interest in perception and reasoning. In 1969, he joined the faculty of the University of California, Los Angeles, initially in Engineering Systems, and in 1970, he received tenure in the newly formed Computer Science Department. In 1976, he was promoted to full professor. In 1978, he founded the Cognitive Systems Laboratory—a title that emphasized his desire to automate human cognition. The laboratory's research facility was Pearl's office, on the door of which hung a permanent sign reading, “Don't knock. Experiments in progress.”

Pearl's reputation in computer science was established initially not in probabilistic reasoning—a highly controversial topic at that time—but in combinatorial search. A series of journal papers beginning in 1980 culminated in the publication of the book *Heuristics: Intelligent Search Strategies for Computer Problem Solving* [Pearl 1984]. This work included many new results on traditional search algorithms, such as A^* , and on game-playing algorithms, raising artificial intelligence (AI) research to a new level of rigor and depth. It also set out new ideas on how admissible heuristics might be derived automatically from relaxed problem definitions, an approach that has led to dramatic advances in planning systems. Despite the book's formal style, it drew its inspiration from, as Pearl said, “the ever-amazing observation of how much people can accomplish with that simplistic,

unreliable information source known as *intuition*.” Ira Pohl wrote in 2011 that “The impact of Pearl’s monograph was transformative ... [The book] was a tour de force summarizing the work of three decades.”

Soon after arriving at UCLA, Pearl began teaching courses on probability and decision theory, which was a rarity in computer science departments at that time. Probabilistic methods had been tried in the 1960s and found wanting; a system for estimating the probability of a disease given n possible symptoms was thought to require a set of probability parameters whose size is exponential in n . The 1970s, on the other hand, saw the rise of *knowledge-based systems*, based primarily on logical rules or on rules augmented with “certainty factors.”

Pearl believed that sound probabilistic analysis of a problem would give intuitively correct results, even in those cases where rule-based systems behaved incorrectly. One such case had to do with the ability to reason both *causally* (from cause to effect) and *diagnostically* (from effect to cause). “If you used diagnostic rules, you could not do prediction, and if you used predictive rules you could not reason diagnostically, and if you used both, you ran into positive-feedback instabilities, something we never encountered in probability theory.” Another case concerned the “explaining-away” phenomenon, whereby the degree of belief in any cause of a given effect is increased when the effect is observed, but then decreases when some other cause is found to be responsible for the observed effect. Rule-based systems could not exhibit the explaining-away phenomenon, whereas it happens automatically in probabilistic analysis.

In addition to these basic qualitative questions, Pearl was inspired by David Rumelhart’s 1976 paper on reading comprehension. As he wrote later in his 1988 book,

In this paper, Rumelhart presented compelling evidence that text comprehension must be a distributed process that combines both top-down and bottom-up inferences. Strangely, this dual mode of inference, so characteristic of Bayesian analysis, did not match the capabilities of either the “certainty factors” calculus or the inference networks of PROSPECTOR¹—the two major contenders for uncertainty management in the 1970s. I thus began to explore the possibility of achieving distributed computation in a “pure” Bayesian framework.

Pearl realized that the concept of *conditional independence* would be the key to constructing complex probability models with polynomially many parameters and

1. An expert system that finds ore deposits from geological information; created in the 1970s by Richard Duda, Peter Hart, and others at Stanford Research Institute (SRI).

to organizing distributed probability computations. The paper “Reverend Bayes on inference engines: A distributed hierarchical approach” [Pearl 1982] introduced probability models defined by directed acyclic graphs and derived an exact, distributed, asynchronous, linear-time inference algorithm for trees—an algorithm we now call *belief propagation*, the basis of turbo codes. There followed a period of remarkable creative output for Pearl, with more than 50 papers covering exact inference for general graphs, approximate inference algorithms using Markov chain Monte Carlo, conditional independence properties, learning algorithms, and more, leading up to the publication of *Probabilistic Reasoning in Intelligent Systems* [Pearl 1988]. This monumental work combined Pearl’s philosophy, his theories of human cognition, and all his technical material into a persuasive whole that sparked a revolution in the field of artificial intelligence. Within just a few years, leading researchers from both the logical and the neural-network camps within AI had adopted a probabilistic—often simply called the *modern*—approach to AI.

Pearl’s Bayesian networks provided a syntax and a calculus for multivariate probability models, in much the same way that George Boole provided a syntax and a calculus for logical models. Theoretical and algorithmic questions associated with Bayesian networks form a significant part of the modern research agenda for machine learning and statistics. Their use has also permeated other areas, such as natural language processing, computer vision, robotics, computational biology, and cognitive science. As of 2012, some 50,000 publications have appeared with Bayesian networks as a primary focus.

Even while developing the theory and technology of Bayesian probability networks, Pearl suspected that a different approach was needed to address the issue of *causality*, which had been one of his concerns for many years. In his 2000 book *Causality* [Pearl 2000], he described his early interest as follows:

I got my first hint of the dark world of causality during my junior year of high school. My science teacher, Dr. Feuchtwanger, introduced us to the study of logic by discussing the 19th century finding that more people died from smallpox inoculations than from smallpox itself. Some people used this information to argue that inoculation was harmful when, in fact, the data proved the opposite, that inoculation was saving lives by eradicating smallpox.

“And here is where logic comes in,” concluded Dr. Feuchtwanger, “To protect us from cause–effect fallacies of this sort.” We were all enchanted by the marvels of logic, even though Dr. Feuchtwanger never actually showed us how logic protects us from such fallacies.

It doesn't, I realized years later as an artificial intelligence researcher. Neither logic nor any branch of mathematics had developed adequate tools for managing problems, such as the smallpox inoculations, involving cause-effect relationships.

A Bayesian network such as *Smoking* \rightarrow *Cancer* fails to capture causal information; indeed, it is mathematically equivalent to the network *Cancer* \rightarrow *Smoking*. The key characteristic of a *causal network* is the way in which it captures the potential effect of exogenous intervention. In a causal network $X \rightarrow Y$, *intervening* to set the value of Y should leave one's prior belief in X unchanged and simply break the link from X to Y ; thus, *Smoking* \rightarrow *Cancer* as a causal network captures our beliefs about how the world works (inducing cancer in a subject does not change one's belief in whether the subject is a smoker), whereas *Cancer* \rightarrow *Smoking* does not (inducing a subject to smoke does change one's belief that the subject will develop cancer). This simple asymmetry prompted Pearl to develop a new calculus, called the *do-calculus*, which led to a complete mathematical framework for formulating causal models and for analyzing data to determine causal relationships. This work has overturned the long-held belief in statistics that causality can be determined only from controlled random trials—which are impossible in areas such as the biological and social sciences. Referring to this work, Phil Dawid (Professor of Statistics at Cambridge) remarks that Pearl is “the most original and influential thinker in statistics today.” Chris Winship (Professor of Sociology at Harvard) writes that, “Social science will be forever in his debt.”

In 2010, a Symposium was held at UCLA in Pearl's honor, and a Festschrift was published containing papers in all the areas covered by his research [Dechter et al. 2010]. The volume also contains reminiscences from former students and other researchers in the field. Ed Purcell, Pearl's first PhD student, wrote, “In class I was immediately impressed and enchanted by Judea's knowledge, intelligence, brilliance, warmth and humor. His teaching style was engaging, interactive, informative and fun.” Hector Geffner, a PhD student in the late 1980s, wrote, “He was humble, fun, unassuming, respectful, intelligent, enthusiastic, full of life, very easy to get along with, and driven by a pure and uncorrupted *passion for understanding*.” Nils Nilsson, former Professor and Chair of the Computer Science Department at Stanford and an AI pioneer, described Pearl as “a towering figure in our field.”

Pearl's outside interests include music (several early conferences were entertained by his impromptu piano renditions and very realistic trumpet imitations), philosophy, and early books—particularly the great works of science throughout history, of which he possesses several first editions. Judea and Ruth Pearl had three children, Tamara, Daniel, and Michelle. Since Daniel's kidnapping and murder in

Pakistan in 2002, Professor Pearl has devoted a significant portion of his time and energy to the Daniel Pearl Foundation, which he and his wife founded to promote Daniel's values of "uncompromised objectivity and integrity; insightful and unconventional perspective; tolerance and respect for people of all cultures; unshaken belief in the effectiveness of education and communication; and the love of music, humor, and friendship."

Pearl has donated a major portion of the Turing Prize money to support the projects of the Daniel Pearl Foundation and another portion to promote the introduction of causal inference in statistics education.

BIRTH:

September 4, 1936, Tel Aviv, Israel.

EDUCATION:

BS, Electrical Engineering (Technion, 1960); MS, Electronics (Newark College of Engineering, 1961); MS, Physics (Rutgers University, 1965); PhD, Electrical Engineering (Polytechnic Institute of Brooklyn, 1965).

EXPERIENCE:

Research Engineer, New York University Medical School (1960–1961); Instructor, Newark College of Engineering (1961); Member of Technical Staff, RCA Research Laboratories, Princeton, New Jersey (1961–1965); Director, Advanced Memory Devices, Electronic Memories, Inc., Hawthorne, California (1966–1969); Assistant Professor of Engineering Systems, UCLA (1969–1970); Associate Professor of Computer Science, UCLA (1970–1976); Director, Cognitive Systems Laboratory, UCLA (from 1978); Professor of Computer Science, UCLA (from 1976—Emeritus since 1994); Professor of Statistics, UCLA (from 1996—Emeritus since 1994); President, Daniel Pearl Foundation (from 2002); International Advisory Board, NGO Monitor (from 2011); Chancellor's Professor of Computer Science Department, UCLA (since 2014).

HONORS AND AWARDS:

RCA Laboratories Achievement Award (1963); NATO Senior Fellowship in Science (1974); Pattern Recognition Society Award for an Outstanding Contribution (1978); Fellow, IEEE (1988); Fellow, American Association of Artificial Intelligence (1990); Named "The Most Published Scientist in the Artificial Intelligence Journal" (1991); Member, National Academy of Engineering (1995); UCLA Faculty Research Lecturer of the Year (1996); IJCAI Research Excellence Award (1999); AAAI Classic Paper Award (2000); Lakatos Award, London School of Economics and Political Science

(2001); Corresponding Member, Spanish Academy of Engineering (2002); Pekeris Memorial Lecture (2003); ACM Allen Newell Award (2003); Purpose Prize (2006); Honorary Doctorate, University of Toronto (2007); Honorary Doctorate, Chapman University (2008); Benjamin Franklin Medal in Computers and Cognitive Science (2008); Festschrift and Symposium in honor of Judea Pearl (2010); Rumelhart Prize Symposium in honor of Judea Pearl (2011); David E. Rumelhart Prize (2011); IEEE Intelligent Systems' AI Hall of Fame (2011); ACM Turing Award (2011); Technion Harvey Prize (2012); Alumni Award NYU-Polytechnic (2013); Member, National Academy of Sciences (2014); Honorary Doctorate Degree, Texas A&M (2014); Honorary Doctorate, Carnegie Mellon University (2015); John C. Cassel Memorial Lecture (2015); CMU Dickson Prize (2015); AIJ Classic Paper Award (2015); ACM Fellow (2015); SER Sells Award (2016); Doctor Philosophiae Honoris Causa, Hebrew University of Jerusalem (2018); Honorary Doctorate, Yale University (2018); UCLA Edward A. Dickson Award (2018); AMS Ulf Grenander Prize (2018); Fellow, American Statistical Association (2019); Distinguished Honorary Fellow Royal Statistical Society (2020).

References

- R. Dechter, H. Geffner, and J. Y. Halpern. 2010. *Heuristics, Probability and Causality: A Tribute to Judea Pearl*, College Publications.
- J. Pearl. 1982. Reverend Bayes on inference engines: A distributed hierarchical approach. In *Proceedings of the 2nd AAAI Conferences on Artificial Intelligence*. 1982, 133–136.
- J. Pearl. 1984. *Heuristics: Intelligent Search Strategies for Computer Problem Solving*. Addison-Wesley, Boston, MA.
- J. Pearl. 1988. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, Burlington, MA.
- J. Pearl. 2000. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, Cambridge. DOI: <https://doi.org/10.1017/S0266466603004109>.



Turing Award Lecture

Transcript of Judea Pearl's Turing Award Lecture, *The Mechanization of Causal Inference: A 'Mini' Turing Test and Beyond*, presented at the 26th Association for the Advancement of Artificial Intelligence (AAAI) Conference, held in Toronto, Canada, in July 2012. The transcript has been lightly edited for clarity. The introduction is by Professor Kelly Gotlieb.

Kelly Gotlieb. It is my great pleasure to welcome you to the ACM Alan Turing Lecture. This annual presentation is delivered by the winner of the ACM Alan Turing Award, which is named for the great British mathematician and computer scientist Alan M. Turing, the originator of the Turing Test, and whose 100th birthday we've been celebrating.

The Turing Award is often referred to as the “Nobel Prize of Computing,” and is the most prestigious prize a computer scientist can receive; it carries a \$250,000 prize generously provided by Intel and Google.

This year's recipient of the ACM Turing Award, and our lecturer this morning, is Judea Pearl, Professor of Computer Science and Statistics at the University of California in Los Angeles. He received this honor in recognition of his fundamental contribution to artificial intelligence as a result of the development of a calculus for probabilistic and causal reasoning.

So, you can see it is quite fitting that he addresses this audience at this conference, seeing he is one of the true pioneers in advancing both the science and the art of artificial intelligence. And I do not give the term “art” loosely because if you know any of Professor Pearl's works or books, you'll know that he is as much a philosopher as a scientist.

The subject for his talk this morning is “The Mechanization of Causal Inference: A 'Mini' Turing Test and Beyond.” It is my privilege to introduce Judea Pearl.

Judea Pearl. Thank you, Kelly, for a wonderful introduction. I'm very glad to be here. I did request to deliver the Turing lecture at AAAI because you, AAAI students and researchers, were with me at an early stage of this game, and deserve to hear a progress report about what happened in this adventure since we last played in the sandbox and built those castles together.

Also, I think it is important that I pay tribute to AAAI for nurturing my work when it was not exactly fashionable. I want to thank all of you for being partners in the development of the things I'm going to talk about: colleagues, co-authors, co-principal investigators, students, and reviewers. I do not know if I should thank my reviewers as well [LAUGHTER].

Three of my most important works were published in the proceedings of AAAI, so I would like to start with those.¹ The first, presented at AAAI 1982 in Pittsburgh, was my first paper on belief propagation in trees. The second was presented at AAAI 1994 and it was a paper with Adnan Darwiche on the *do calculus*. I'm sure that it wouldn't have been published in any other conference proceedings, in Statistics or any other field. The third was presented in the same conference, AAAI 1994, and it was the paper with Alexander Balke on "Probabilistic Evaluation of Counterfactual Queries." I chose those three papers because their titles are closely related to the names of the three-layer hierarchy of causal reasoning that we have today. They established a very solid kind of hierarchy that is rarely mixed, in the sense that you can syntactically tell if a sentence is probabilistic, causal, or counterfactual.

But this is not a lecture about my work; it is a lecture about Turing. So, let me start with Turing and his Turing Test in the article in *Mind Magazine* in 1950: a test that I think is an engine behind much of the work that is done in AI.

Turing's answer to the question of, "Can computers think?" was very simple. "Yes, if it acts like it thinks," where "acting" means that it provides reasonable answers to non-trivial questions about a story, a topic, or a situation. Many of us are working on mini-Turing Tests in various fields. I will consider questions that involve causal inference.

Here is how Turing described a hypothetical conversation with the machine. First was the question about poetry. And the answer, of course, is evasive, although with some human element to it: "I never could write poetry."

The second question is about arithmetic: "Can you add that and that," and the answer is also human. You pause for 30 seconds, and then you give the answer. This is also a very simple domain.

1. The three AAAI papers that Judea Pearl is referring to are: "Reverend Bayes on inference engines: A distributed hierarchical approach" [Pearl 1982]; "Symbolic causal networks for reasoning about action and plans" [Darwiche and Pearl 1994]; and "Probabilistic evaluation of counterfactual queries" [Balke and Pearl 1994].

And then Turing said, “Let’s look at chess. Do you play chess?” “Yes.” “I have a King on my K1, and no other pieces; you have only King at K6 and Rook at R1. It is your move. What do you play?” And of course the machine answers after a pause, “Checkmate.”

So, these were the questions exemplified in Turing’s first paper: questions about various domains like arithmetic, poetry, and chess, all of which admit reasonable answers.

But then Turing talks about a “child machine,” which is essentially machine learning. “Why don’t we start with a child machine?” It should be easier, he said, because the child does not need as much background as we expect adults to have. “Our hope is that there is so little mechanism in the child-brain that something like it can be easily programmed.” I think that Turing underestimated the role that vision and motor action play even in high level intelligence. We know, for example, that metaphors taken from the child world play a tremendous role in the child’s ability to handle mathematics.

Turing then made some statements about the connection between machine learning and evolution, and said: “The survival of the fittest is a slow method for learning. The experimenter [the programmer], by exercise of intelligence, should be able to speed it up. How? By creating artificial mutations where they are needed. If he can trace the cause of some weakness, he can then probably think of a kind of mutation which will improve it.” Turing’s idea was that the programmer would be able to trace shortcomings of the program to where they matter, and fix them. There was a great vision here because it leads to the question: Why shouldn’t a machine, having a blue print of itself be able to pinpoint the root causes of a weakness, and change priority among competing computational resources?

I will explain to you why I chose causal reasoning to be a domain that deserves to be called a “Mini Turing Test.” For this, imagine that you have Turing’s experimental setting with an interrogator asking a machine questions. The questions, however, are limited mainly to three types or modalities: *What is?*, *What if?*, and *Why?*

The story, that I used many times in my 1988 book *Probabilistic Reasoning* [Pearl 1988] and the 2000 book *Causality* [Pearl 2000], is as follows: You get out of your house and you see the pavement. The pavement may be wet or dry, it may also be slippery or not, it may have rained or not, the season may be dry or wet, and the sprinkler may have been on or off. These are five binary variables that can be used to generate many simple stories connected to your everyday experience. The task is to tell a story to the machine and the machine has to answer questions corresponding to the three modalities.

One simple question: if the season is dry and the pavement is slippery, did it rain? You expect an answer like: “It is unlikely. It is more likely that the sprinkler was on, with a very slight possibility that the pavement is not even wet.” There could indeed be other reasons for why the pavement is slippery. This is the kind of answer that you expect on the basis of observations alone.

Then comes a second question: “What if you see that the sprinkler is off?” A plausible answer is: “It is more likely then that it rained.” This is reasonable; it is an example of what is called “explaining away.”

Now a question about actions: “Do you mean that if we actually turn the sprinkler on, then rain will be less likely?” And you want the machine to say, “No, there is a difference between seeing and doing; the likelihood of rain would remain the same but the pavement will surely get wet.”

Finally, a question of counterfactual nature: “Suppose that you see that the sprinkler is on and the pavement is wet. Would the pavement be wet if the sprinkler were off?” I’ll explain why I’m so hung up on counterfactuals, but first I would like you to answer the question instead of the machine. What I expect the machine to say is, “The pavement would be dry then, because the season is likely dry.” Namely, you take the observation here, that the sprinkler is on, and you infer, “Oh, it must be a dry season.” Then, if the sprinkler were off, the past remains the same but the future changes, so the justification should be: “Because the season is likely dry and the pavement is wet.”

This is the kind of question/answer session that we expect for a toy problem. We all remember, however, Searle’s argument of the Chinese room that says that answering questions does not mean that a machine thinks or even understands the questions. To prove his point, Searle imagines that the machine takes the questions in Chinese and answers them using a rule book, where every sentence in Chinese has the answer printed there in Chinese or in English. He concludes that the machine can’t be said to understand Chinese just because it looks up the answers in the book.

What Searle overlooks is the fact that there are not enough molecules in the universe to make up such a book, because of the huge number of questions that may be asked. “So what?” you may ask. “Just because you have combinatorial difficulty, you conclude that the machine thinks?” [AUDIENCE LAUGHS] The answer is “Yes,” because when you have such a combinatorial problem to overcome, the only way to solve it is by taking advantage of the relevant constraints in the domain. And understanding and taking advantage of the relevant principles and constraints is what we mean by understanding.

Even for the sprinkler example, if, for the sake of argument, we consider ten binary variables and count the number of entries in the table that we would need

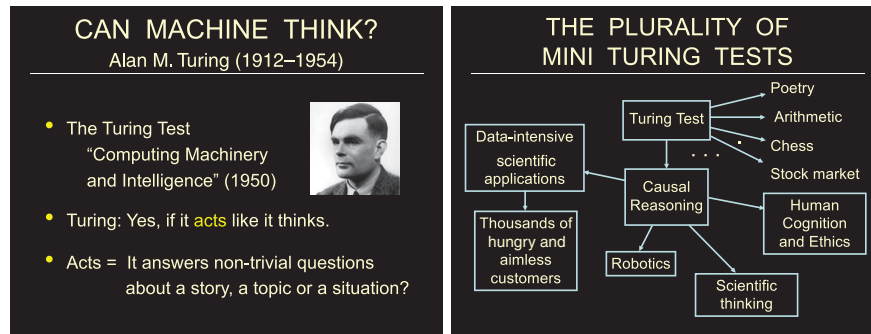


Figure 2.1 Turing Test and a plurality of mini-Turing Tests.

to use Searle’s Chinese-book method, it turns out that we would need on the order of 1,000 entries just for the probability. We need to multiply this number by another 1,000 to get the probabilities for all actions, and by an additional 1,000 to account for the counterfactual queries. So, we would need a billion-long table just to answer questions about the simple pavement story.

Yet even children can answer these questions quite intelligibly, and the question is, “How?” I’ll argue that there are important principles and constraints that enable the child to answer questions about observations, interventions, and counterfactuals, but before getting there, I’d like to explain why I think that the causal conversation is important (Figure 2.1, “The plurality of mini Turing tests”).

Causal reasoning is important because it is pervasive in human cognition and human ethics, and it is deeply entrenched in the cognitive development of children. In addition, causality is a building block of scientific thinking and crucial in robotics. Finally, and that’s the reason I have spent more than 25 years of my life on causality, there are many data-intensive applications that can benefit from any new insight in causal reasoning. There are thousands of hungry and aimless customers, not hungry for money since they are well endowed—all the pharmaceutical companies are part of this enterprise—but they are hungry for ideas, because causal reasoning has not been properly formalized in those fields. Thus, any insight that we get by trying to make a robot understand cause-and-effect could translate into methods that could save millions of lives and dollars in those fields.

Let me start with human cognition and ethics (Figure 2.2). I like to start with Adam and Eve—where else do you start? And you can see immediately that when God asked Adam, “Hey, did you eat from that tree?” Adam does not answer “yes” or “no.” He says instead, “She handed me the fruit and I ate.” You see: facts are for the gods; excuses are for men. [LAUGHTER] And Eve, of course, is no less expert in causal explanations, and says, “Don’t blame me. The serpent deceived

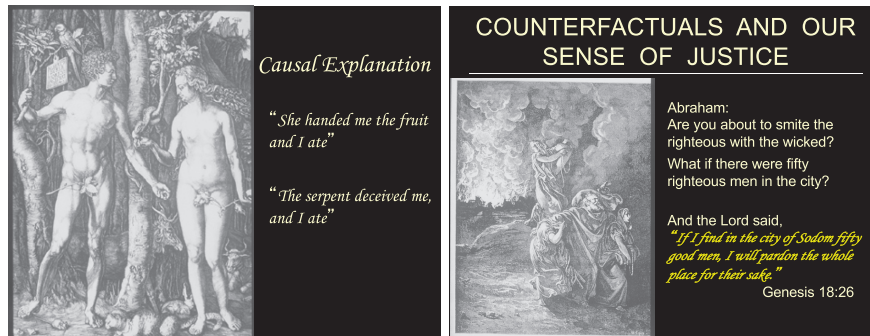


Figure 2.2 Causes, counterfactuals, and our sense of justice.

me and I ate.” Thus causal reasoning plays a key role in our sense of justice, and in the need to pass the buck to somebody else. [LAUGHTER]

You also remember when God told Abraham that he is about to destroy the cities of Sodom and Gomorrah, and Abraham said, “Are you about to smite the righteous with the wicked? You can’t do that. What if there were 50 righteous men in the city?” Here, you have the first counterfactual in the Bible. [AUDIENCE LAUGHS] “What if there were 50?” And look what God says: “If I find in the city of Sodom 50 good men, I will pardon the whole place for their sake.” Do you think that Abraham gave up at that point? No. He got down and said, “What about 45?” [AUDIENCE LAUGHS] “Are you going to make a big fuss for five people?” And God says, “No, I ain’t gonna destroy it,” and then he goes down to 40, and then 30, and 20, and 10, and you know what happened. The rest is history, and the question, of course, is what kind of game this is. Did Abraham doubt the ability of God to count or to distinguish the righteous from the wicked? No. Abraham was the first scientist: he tried to find a general rule. “Where is the threshold?” “What is the general rule for collective punishment?” [AUDIENCE LAUGHS] In that sense, he was the first scientist, because what is science all about? It is about the general rules; not about specific events.

So, here I go to science to prove to you that counterfactuals are indeed the basis for science. We all used to do problems in physics, for example, using Hooke’s law, which tells you that the length of the string Y is equal to a constant, say 2, times the weight X it supports. So if X is one kilogram, we have two equations: $Y = 2X$ and $X = 1$ (Figure 2.3). You may think that finding the length of the string Y is just arithmetic: you solve the two equations with the two unknowns, and obtain the values $Y = 2$ and $X = 1$. The question is: are the equations $Y = 2X$ and $X = 1$, and the equations $Y = 2$ and $X = 1$ equivalent? They are of course algebraically equivalent, as they have the same solution, but I will argue that they are not equivalent,

WHAT KIND OF QUESTIONS SHOULD THE ROBOT ANSWER?

- **Observational Questions:**
"What if we see A" (What is?)
- **Action Questions:**
"What if we do A?" (What if?)
- **Counterfactuals Questions:**
"What if we did things differently?" (Why?)
- **Options:**
"With what probability?"

THE CAUSAL HIERARCHY

WHY PHYSICS IS COUNTERFACTUAL

Scientific equations (e.g., Hooke's law) are non-algebraic
e.g., Length (Y) equals a constant (2) times the weight (X)

Correct notation:
(or)

$Y \leftarrow 2X$	$X = 1$	$X = 3$ $X = 3$
$X = 3$ $X = 1$	$Y = 2$	$Y = X + 1$
<u>Process information</u>	<u>The solution</u>	<u>Alternative</u>

Had X been 3, Y would be 6.
If we raise X to 3, Y would be 6.
Must "wipe out" X = 1.

Figure 2.3 The causal hierarchy and why physics is counterfactual.

because the equations on the left can answer questions that the ones on the right cannot.

For illustrating this difference, consider actually a system of equations $X = Y/2$ and $Y = X + 1$ which has the same solution $Y = 2$ and $X = 1$, along with the following question: "If we raise the weight X to 3, what would be the length Y ?" In the first system of equations $Y = 2X$ and $X = 1$, which captures Hooke's law and the unit body weight, the counterfactual question "if X had been 3" has the answer $Y = 6$, which can be obtained by wiping out the equation $X = 1$ and replacing it by $X = 3$. The new system of two equations, modified by the new information, gives us the answer $Y = 6$.

The system of equations $X = Y/2$ and $Y = X + 1$, on the other hand, has the same solutions as the equations $Y = 2X$ and $X = 1$, but if we apply the same method for answering the counterfactual query, and replace the equation $X = Y/2$ by $X = 3$, we obtain the answer $Y = 4$, which is wrong.

Every child in high school, when he or she solves physics problems, engages in counterfactual reasoning of this sort. The child knows which equations to write, which equations to wipe out, and which ones to keep. They keep the one that conveys the generic rule and wipe out the ones that are merely boundary conditions and subject to the antecedent of the counterfactual. If this is the case, the equality sign that we saw before in the equation $Y = 2X$ for expressing Hooke's law does not really represent an algebraic equality but something closer to an assignment statement in a programming language.

You can imagine that Nature, before determining the length of the spring, looks around for all variables that might possibly affect the length. She looks at the weight and says, "Ah, that is the one," then consults the weight on the spring, and finally determines the value of the length. So, this is the conception of Nature in physics: Nature looks at some variables, goes through some process, and then

assigns values to other variables. If that is so, then modeling Nature requires a different kind of algebra because the process involves wiping out equations. That is the meaning of arrows in the structure of causal graphs; it is a description of the strategy used by Nature.

The role of counterfactuals and causation in human reasoning has not escaped philosophers. Already at the time of the Greeks in 430 BC, Democritus said, “I would rather discover one causal relationship than be king of Persia.” King of Persia at that time was not exactly a dangerous occupation like it is today. [LAUGHTER] And Hume, of course, looked at that and said, “What is this idea of causation? I’ve got to solve it.” And he came out with the conception that causation is not a gift of the gods, but something that we learn from experience. Here is a famous paragraph: “We remember to have seen that species of object called ‘flame,’ and to have felt the species of sensations we call ‘heat.’ Without any further ceremony, we call the one ‘cause’ and the other ‘effect.’” So, it is a matter of determining regularity in nature that makes us come up with the label “cause.” There are obvious difficulties to that conception, of course, but the fact that generations of philosophers have stumbled on the difficulty of explaining what “cause” is, brings us to ask: “What gives us the audacity, here in AI, to think that we can add another iota to this long debate?”

The answer is simply that we do not have the luxury to philosophize. We need to build robots that understand what went wrong in the laboratory or the kitchen, and if they do not learn it by themselves, we need to teach them, so that they can act properly and answer queries about cause/effect relationships. And this is not a trivial thing to do because now the puzzles that philosophers have faced translate into engineering problems. The question of, “How do we acquire causal information from the environment?” is translated into, “How do we people conclude that the sprinkler caused the pavement to get wet?” And the question of “How do we people conclude that the sprinkler caused the pavement to get wet?” translates into, “How should a robot use causal information received from its creator-programmer to understand or to answer queries properly?”

The use of causal information may look trivial but it is not, because if you just follow the rules you get unexpected results. If the input is “If the grass is wet, then it rained” and “If you break this bottle, the grass will get wet,” you do not want an output such as “If we break the bottle, then it rained.” So, just rule-chaining is not going to do the work for us; we need something more.

And what is that something more? Before we get there, let me provide an outline of what I’m going to talk about (Figure 2.4). I’m going to talk about the three-level hierarchy first. The question “What if I see” is about probability and beliefs. The question “What if I do?” is about actions and interventions. Finally, the question

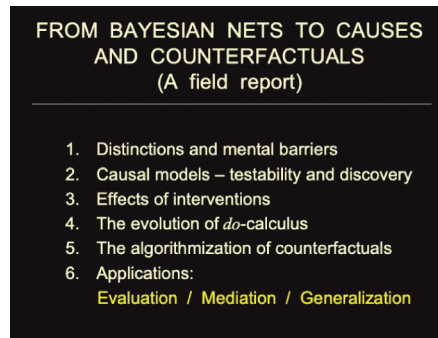


Figure 2.4 Roadmap: From Bayesian nets to causality and counterfactuals.

“What if I did things differently?” is about counterfactuals. You can decorate these questions with probabilities; namely, how likely are the answers, but that’s not essential.

The following is a field report of the journey that we took from the old days of Bayesian networks to causality and counterfactuals. We have to understand the distinctions and mental barriers that stood in our way. We have to talk also about what makes a model causal as opposed to something else, how a causal model can be tested, and how causal models and data are connected. If a model has testable implications, then you can hope to discover or learn the model from data. A model that does not have any testable implication cannot be discovered from data. Then I’ll talk about three themes: the effects of interventions, the evolution of the *do*-calculus, and the algorithmization of counterfactuals. I’ll also talk about applications: evaluation of plans and policies, mediation (i.e., distinguishing between direct and indirect causes), and generalization.

I start with the basic statistical problem and the paradigm that rules statistical thinking and most of machine learning. The idea is that someplace behind the scenes there is a Santa Claus called the “joint probability distribution” that occasionally, when he or she is gracious enough, spits out data. Our job is to infer Santa Claus’s properties: some aspect $Q(P)$ of the joint distribution function P from the data; for example, we might want to estimate the mean, come up with a classifier, or decide whether a customer who bought Product A will also buy Product B. This kind of question is neat and well-formulated because it can be neatly encapsulated in the language of probability theory. We even have a short sentence to express this question: “Find the conditional probability of B given A,” with conditional probability coming all the way from Reverend Bayes 250 years ago. The function P can be a very complex distribution defined on many variables, some continuous and some binary, and so on. Although this is not a simple computational problem,

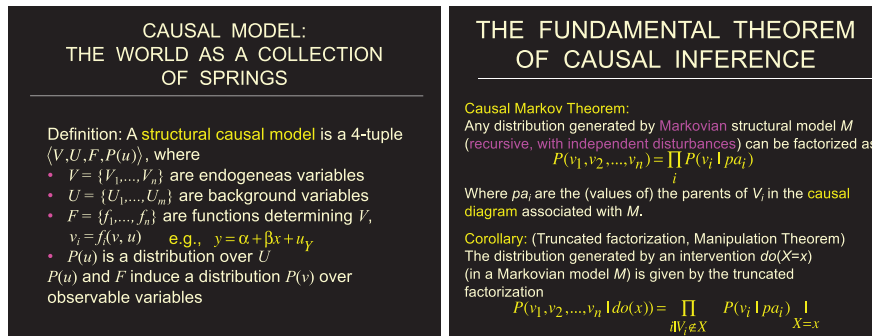


Figure 2.5 Structured causal models and truncated factorization.

the paradigm is clear enough. Causal reasoning, however, deals with a different paradigm.

You ask a question, for instance, “Infer whether customers who bought Product A would buy Product B if we double the price.” So, here we get up in the morning, whimsically greedy, and wonder what would happen if we raised the price. And we ask the question, “What will the probability of B given A be after we do something that perhaps has not been done before, like doubling the price of the product.” This is not even an aspect of the probability distribution P ; observing that the price has doubled (and what has happened as a consequence) is very different from doubling the price and seeing the consequence.

The counterfactual “had we doubled the price” is thus not an aspect or property of the Santa Claus. So, what is it? It is a property of a data-generating model that is behind the joint probability. As before, the joint probability spits out data, we get the samples, and we need to infer some property, but of what? Not of P , but of the data-generating model. This is the invariant strategy of Nature that I talked about before, sometimes called a “mechanism,” “recipe,” “law,” or “protocol”—all are counterfactual notions—by which Nature assigns values to variables.

This simple idea is torture for a statistician because it takes a leap of imagination to think of Nature rather than experiments or measurements. It is a traumatic experience for people outside artificial intelligence; I would like you to be aware of that if you ever talk to an outsider. [AUDIENCE LAUGHS]

Once we go there, let’s generalize it. Let’s imagine that the whole world is just a collection of springs. So, the model is fueled by a collection of functions that assign values to variables. Every variable is assigned a value that is a function of the other variables in the system (Figure 2.5). Some of the variables are exogenous; you do not care about their causes, but only about their effects. The rest are endogenous. And

our job is to encode this on a machine so that the machine can provide reasonable and plausible answers to our reasonable questions.

The equations that we had for the spring example are typical: after Nature spends some time, maybe a billionth of a second, looking at X , multiplying it by constant, adding to it some noise, and deciding that Y deserves the value y (great work, mother Nature!), our job is to decipher the strategy of Nature. If this sounds too ambitious, at the very least we should be able to answer counterfactual queries if we have enough data.

Let us illustrate this by considering a familiar digital circuit diagram. The circuit is an oracle for counterfactuals because if you look at the circuit you can answer a counterfactual question like “What if I were to replace this OR gate with an AND gate?” or “What if I were to connect this node Y to a power supply of 5 volts?” Even though the circuit designer never anticipated such crazy questions and events, the engineer glancing at the circuit has the ability to contemplate the answers and compute them correctly.

Where does this ability come from? It comes from some fundamental properties of the collection of functions and equations in the causal model. The fundamental one, from which everything else eventually derives, is that, if you happen to be lucky and your equations are recursive (no cycles there), and the disturbances happen to be independent of each other, then regardless of the functions that you have there and regardless of the distributions of disturbances, you can say something about the probability distribution of what you observe. So, the structure of that collection of springs determines something very basic in your distribution function, which has the form of a product and represents conditional independencies (Figure 2.5).

And from that comes the next corollary, which is the ability to answer questions about interventions. Once you have this product form, if somebody asks you, “And what if I take an action?,” the answer comes from the truncated factorized product (Figure 2.5). This is the same factorized product as before, but we delete from the product those variables that are forced to a constant (by the intervention) because those variables no longer listen to their parents.

Here is our sprinkler example again (Figure 2.6). Before you act, you have the diamond structure shown in the figure, which corresponds to the set of equations shown. But once you take an action like turning the sprinkler on, you must remove the causal influence of the variable *Season* on the variable *Sprinkler*, as Mr. Sprinkler no longer listens to its parent, and instead becomes enslaved to your muscles, which set the variable to a value.

This formalism for actions did not germinate in AI, but originated with an economist, Haavelmo. In 1943, he considered the problem of modeling

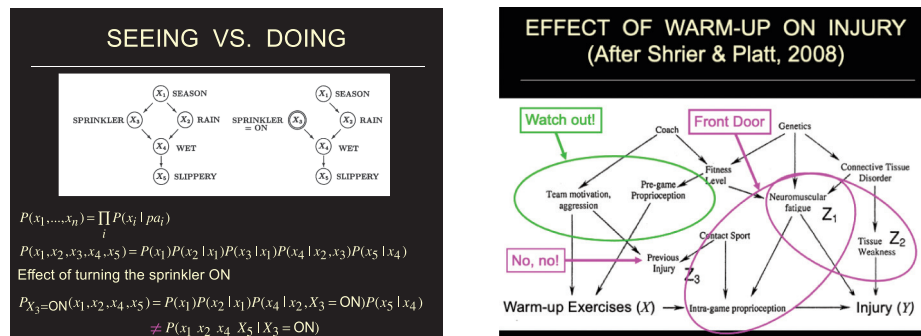


Figure 2.6 Structured causal models of two of the examples in the text.

government interventions in the economy, like fixing a price or imposing taxes, and he had the idea to model the effects of the actions by introducing changes in the equations. If the government does something like keeping a price constant, a term is added to the corresponding equation to balance the other terms, so that the price remains constant. Later on, this manipulation was replaced by Strotz and Wold, who “wiped out” the relevant equation and replaced it with a constant assignment. Then Spirtes, Glymour, and Scheines transformed this manipulation into a graphical surgery procedure, where you wipe out the arrows going into the manipulated variable, resulting in the truncated factorization. I took this all very seriously and said, “We have a new calculus that deserves algebraic support,” translated it into the *do*-calculus, and then applied it to counterfactuals. That has been the evolution of these ideas. Now we also have the unification with the Neyman–Rubin account in statistics, which also handles causality with counterfactuals.

How are counterfactuals handled, and what is the general model for counterfactuals? This is all very simple (Figure 2.7). You mutilate your model to take care of the antecedent of the counterfactual, and you solve the equation in the mutilated model. There’s nothing else to it; it’s embarrassingly simple. In this Definition, I simply say symbolically what I said verbally: you are in possession of a calculus because you have a semantics for joint counterfactuals. For any set of variables X, Y, Z, \dots , you can find the joint probability of Y taking a value y had X been x , and simultaneously, Z taking value z had W been w , and so on. The semantics determines the probability of any such sentence.

Specifically, the sentences can involve actions with the “*do*” operator and attributions, like “What is the likelihood that a patient would be alive today had he not taken the drug, given that in fact he is dead and he took the drug?” This is a sentence in the language, and the semantics is there. If you have the model,

COUNTERFACTUALS ARE EMBARRASINGLY SIMPLE

Definition:
The sentence: “*Y* would be *y* (in situation *u*), had *X* been *x*,” denoted $Y_x(u) = y$, means:
The solution for *Y* in a mutilated model M_x , (i.e., the equations for *X* replaced by $X = x$) with input $U = u$, is equal to *y*.

The Fundamental Equation of Counterfactuals:

$$\boxed{Y_x(u) = Y_{M_x}(u)}$$

COUNTERFACTUALS ARE EMBARRASINGLY SIMPLE

Definition:
The sentence: “*Y* would be *y* (in situation *u*), had *X* been *x*,” denoted $Y_x(u) = y$, means:
The solution for *Y* in a mutilated model M_x , (i.e., the equations for *X* replaced by $X = x$) with input $U = u$, is equal to *y*.

- **Joint probabilities of counterfactuals:**

$$P(Y_x = y, Z_w = z) = \sum_{u: Y_x(u)=y, Z_w(u)=z} P(u)$$

In particular:

$$P(y | do(x)) \triangleq P(Y_x = y) = \sum_{u: Y_x(u)=y} P(u)$$

$$P(Y_{x'} = y' | x, y) = \sum_{u: Y_{x'}(u)=y'} P(u | x, y)$$

Figure 2.7 Counterfactuals are simple.

you can compute the answer. Everybody knows how to solve equations, right? The semantics is extremely simple.

And Joe Halpern and David Galles came up with a complete axiomatization of that. Why do we need an axiomatization? So that if anybody says, “You can do counterfactuals differently,” you can compare the axioms and evaluate if they are equivalent or not. The workhorse is a composition axiom that tells you that if you do something that would have occurred anyhow, you have not done a thing. This sentence says, essentially, that our world is closer to our world than any other possible world, if you go to the possible worlds interpretation of it.

I’ll give you now an example of what you can do with it. You have a collection of equations and you think that Nature works like that. The first questions that you have to ask yourself are “Is this model testable?” or “Does the model have any testable implications?” As I said before, if it does not have testable implications, you cannot learn or verify the model. And the idea for the verification is very simple. Everything that we did with Bayesian nets translates now into Causal Bayesian nets, and the criterion of *d*-separation gives you a finite set of testable implications. Just look at the missing arrows: every one carries the promise of a test. If the test fails, the model is wrong.

What else can these models do for you? They can handle interventions; they are, indeed, an oracle for interventions. So, if you have questions like “What is the average causal effect of *X* on *Y*, given that you can measure variables *W* and *Z*” or “Can you do this without manipulation, just by observation?”, you can produce answers like “Yes, if you can measure variables like age or ethnicity.” Namely, you are guaranteed that you can answer the query without bias by simple adjustment (regression). Of course, these results are built on the assumptions encoded in the causal graph. Each missing link in the graph is an assumption of a causal nature, not of a statistical nature.

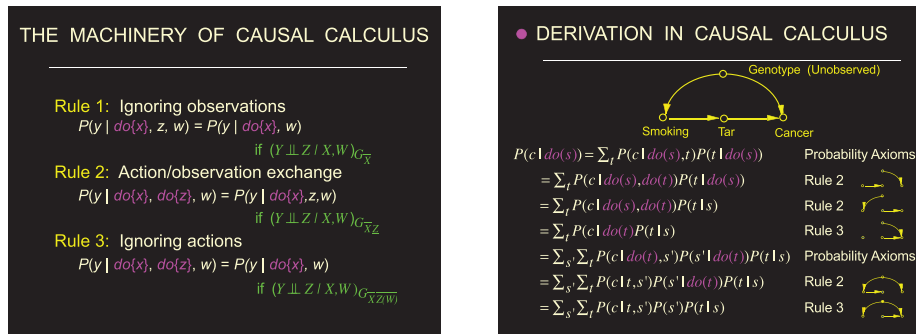


Figure 2.8 Causal calculus in action.

Here is another example, one which is highly applicable. You are in the sports medicine business, and you wonder whether warm-up is a cause of injury or prevents injuries in the game (Figure 2.6). It’s an extremely important question for our society, for our culture, right? You can take measurements of previous injuries, team aggressiveness, and so on. Which one would you measure? Each one takes a lot of dollars to measure. The answer is given to you automatically: “Thou shalt measure this, and you’re okay; thou shalt not measure that because you would get bias; thou shalt measure that—fine, and here there is another alternative.” Indeed, you can pick the measurements according to their cost and their reliability. There are three rules that drive this answering mechanism (Figure 2.8). The rules take the graph into account, are applied repeatedly, and produce the answer.

Another example: Does smoking cause cancer? The query given to you contains a causal symbol (Figure 2.8): the purple expression $do(s)$ stands for doing the action of smoking. We do not have the data for the effect of this action: we cannot conduct randomized experiments on smokers. So we have to answer the query analytically. We apply the rules one after the other until we get rid of all the purple expressions. Once we do this, it means that you can answer the query from data obtained by hands-off, passive observations. And you can answer the question quantitatively: this is the extent to which smoking causes cancer.

What else can this calculus do for you? Find equivalent models, identify counterfactual queries, mediation, which is about the distinction between direct and indirect effects, explanation, which is about finding the causes of observed effects, and transportability, which is about generalizing what you learn in one domain into another domain in which you cannot conduct any experiments.

Counterfactuals are very interesting because philosophers have gone through a great deal of pain to understand why we are able to agree on their truth value. Here is a typical example: “If Oswald didn’t kill Kennedy, someone else did” and

“If Oswald hadn’t killed Kennedy, someone else would have.” If I give you this pair of sentences, you’ll tell me “Yes” on the first one and “No” on the second. How are we able to agree on this? This was a puzzle for philosophers.

Hume tried to explain causes in terms of counterfactuals, and David Lewis tried to explain causes in those terms too. The puzzle that I faced was different. Why don’t we try to define counterfactuals in terms of causes, rather than the other way around? Are counterfactuals less problematic? Apparently so, because we do form consensus on counterfactuals. And these two pillars of philosophy tried indeed to define causes in terms of counterfactuals. To me it means that we do count on a counterfactual engine in our mind that is swift and reliable, and we form consensus because we share the architecture of this engine. So, this is an AI problem, not a philosophy problem.

Indeed, what Lewis came up with in his possible-worlds semantics for counterfactuals does not solve the consensus puzzle as it relies on assessing, for example, how close is a world in which we are all dead after Nixon presses the button, relative to a world in which Nixon presses the button but somebody disconnected the wires. That is a typical question in philosophy—assessment of how similar worlds are. In our structural world, you do not rely on similarity among worlds; you rely on equations which are common equations of physics, and mutilating those equations.

I will not have time to talk about the counterfactual triumph, which is the ability to distinguish between direct and indirect effects. It is an important distinction because we send people to prison if they are directly responsible for murder, and fine them if they are only indirectly responsible. So, it is a key notion in law, in ethics, and in understanding how the world works. However, it requires the ability to answer questions about different kinds of interventions—interventions where you enable and disable certain mechanisms, rather than fixing variables, as I mentioned before.

Direct and indirect effects is a booming field now in statistical epidemiology, called “mediation analysis.” And the impetus for that was counterfactuals. We were able to express the idea of indirect effects by counterfactuals, as you see here. What is the definition of “indirect effect?” It is the expected change in output when we keep the input constant but change the mediator. “What would you have gotten had the input changed?” is a nested counterfactual that is not about fixing the value of variables. It is now the accepted definition when you have indirect effects. That’s why I consider this account a triumph.

I’ll now talk about the next triumph: transportability. And I say it’s a triumph because here the *do*-calculus appeared out of the blue. We didn’t expect it to reveal its potency in an area like that, which has very little to do with interventions.

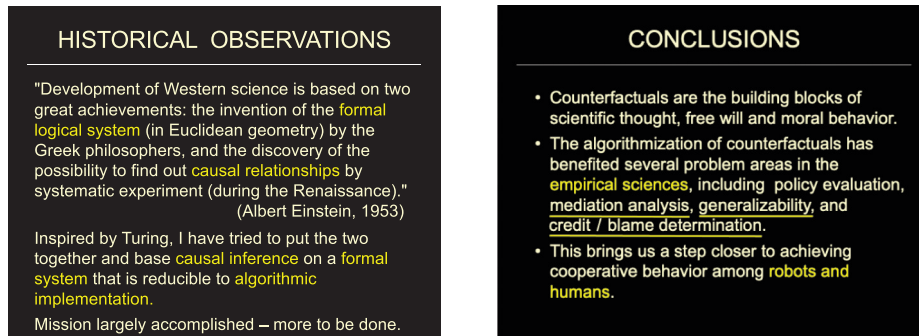


Figure 2.9 Logic and experiment for a science of cause and effect.

Imagine that we want to transfer relationships that we learn from experiments in one environment to a different environment in which no experiments can be conducted. So, we can think about training a robot in the cockpit and moving him or her to another environment where only observations are allowed, but no interventions.

How much of the causal knowledge that the robot acquired in the cockpit is transferable? We typically want a crisp logical answer, yes or no, regarding whether a certain relationship is or is not transferable given what we know about the two environments. And this has surprisingly a complete answer; that is, an answer that cannot be improved. When the method says that the information cannot be transferred, we also get an explanation for why, in terms of the assumptions about the disparities and commonalities between the two environments.

I think I'm close to the end of the talk. I have five seconds. [LAUGHTER]

I didn't talk about our new game, which is meta-analysis, in which big data comes to play. Imagine that you have data coming from 1,000 hospitals in the United States or worldwide, each one conducted under different conditions with different populations. You want to use all this data to come up with an answer to a query in another environment, where no measurements are allowed. All you know is the structure. Can you do it or not? We look for a crisp, yes or no answer. And if you can, how? So, I go through the "how" over many slides here, which I'll have to skip. Believe me, there is a method here, and there is a lot of work to be done in terms of decomposing the relationships into sub-relationships for picking up from every study the commonalities, and for putting them together to come up with an unbiased estimate.

It is time to move to the conclusions (Figure 2.9). Counterfactuals are the building blocks of scientific thought, free will, and moral behavior. The algorithmization of counterfactuals has benefited several problems in the empirical sciences, and

brings us a step closer to achieving cooperative behavior between humans and robots.

Historically—I have to play the sage at this point—Einstein noticed that there have been two major advances in Western science. One is the development of logic by the Greeks. The other is the recognition by Galileo that you can find cause-effect relationships from experiments. I’m following these paths, trying to combine the two: the logic of the Greeks with the experiments of Galileo, to come up with logically sound theories of causes and counterfactuals. Our mission is largely accomplished, but more remains to be done. Thank you. [APPLAUSE]

References

- A. Balke and J. Pearl. 1994. Probabilistic Evaluation of Counterfactual Queries. In *Proceedings of the AAAI-94*, Seattle, WA, Volume I, 230–237.
- A. Darwiche and J. Pearl. 1994. Symbolic Causal Networks for Reasoning about Actions and Plans. In *Proceedings of the AAAI-94*, Seattle, WA, Volume I, 238–244.
- J. Pearl. 1982. Reverend Bayes on inference engines: A distributed hierarchical approach. In *Proceedings of the 2nd AAAI Conferences on Artificial Intelligence*. 1982, 133–136.
- J. Pearl. 1988. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, Burlington, MA.
- J. Pearl. 2000. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, Cambridge.



Interview by Martin Ford

“The current machine learning concentration on deep learning and its non-transparent structures is a hang-up. They need to liberate themselves from this data-centric philosophy.”

Judea Pearl is known internationally for his contributions to artificial intelligence, human reasoning, and philosophy of science. He is particularly well known in the AI field for his work on probabilistic (or Bayesian) techniques and causality. He is the author of more than 450 scientific papers and three landmark books: Heuristics (1984) [1], Probabilistic Reasoning (1988) [3], and Causality (2000; 2009) [4]. His 2018 book, The Book of Why [6], makes his work on causation accessible to a general audience. In 2011, Judea received the Turing Award, which is the highest honor in the field of computer science and is often compared to the Nobel Prize.

MARTIN FORD: You’ve had a long and decorated career. What path led you to get started in computer science and artificial intelligence?

JUDEA PEARL: I was born in Israel in 1936, in a town named Bnei Brak. I attribute a lot of my curiosity to my childhood and to my upbringing, both as part of Israeli society and as a lucky member of a generation that received a unique and inspiring education. My high-school and college teachers were top-notch scientists who had come from Germany in the 1930s, and they couldn’t find a job in academia, so they taught in high schools. They knew they would never get back to academia, and they saw in us the embodiment of their academic and scientific dreams. My generation were beneficiaries of this educational experiment—growing up under the mentorship of great scientists who happened to be high-school teachers. I never excelled in school, I was not the best, or even second best, I was always third or fourth, but I always got very involved in each area taught. And we were taught in a chronological way, focusing on the inventor or scientist behind the invention or theorem.

Originally published in M. Ford, *Architects of Intelligence: The truth about AI from the people building it*. Birmingham, UK: Packt Publishing, 357-372, 2018.

Republished with permission.

Because of this, we got the idea that science is not just a collection of facts, but a continuous human struggle with the uncertainties of nature. This added to my curiosity.

I didn't commit myself to science until I was in the army. I was a member of a Kibbutz and was about to spend my life there, but smart people told me that I would be happier if I utilized my mathematical skills. As such, they advised me to go and study electronics in Technion, the Israel Institute of Technology, which I did in 1956. I did not favor any particular specialization in college; but I enjoyed circuit synthesis and electromagnetic theory. I finished my undergraduate degree and got married in 1960. I came to the US with the idea of doing graduate work, getting my PhD, and going back.

MARTIN FORD: You mean you planned to go back to Israel?

JUDEA PEARL: Yes, my plan was to get a degree and come back to Israel. I first registered at the Brooklyn Polytechnic Institute (now part of NYU), which was one of the top schools in microwave communication at the time. However, I couldn't afford the tuition, I ended up employed at the David Sarnoff Research Center at the RCA laboratory in Princeton, New Jersey. There, I was a member of the computer memory group under Dr. Jan Rajchman, which was a hardware-oriented group. We, as well as everybody else in the country, were looking for different physical mechanisms that could serve as computer memory. This was because magnetic core memories became too slow, too bulky, and you had to string them manually.

People understood that the days of core memory were numbered, and everybody—IBM, Bell Labs, and RCA Laboratories—was looking for various phenomena that could serve as a mechanism to store digital information. Superconductivity was appealing at that time because of the speed and the ease of preparing the memory, even though it required cooling to liquid helium temperature. I was investigating circulating currents in superconductors, again for use in memory, and I discovered a few interesting phenomena there. There's even a Pearl vortex named after me, which is a turbulent current that spins around in superconducting films, and gives rise to a very interesting phenomenon that defies Faraday's law. It was an exciting time, both on the technological side and on the inspirational, scientific side.

Everyone was also inspired by the potential capabilities of computers in 1961 and 1962. No one had any doubt that eventually, computers would emulate most human intellectual tasks. Everyone was looking for tricks to accomplish those tasks, even the hardware people. We were constantly looking for ways of making associative memories, dealing with perception, object recognition, the encoding of

visual scenes; all the tasks that we knew were important for general AI. The management at RCA also encouraged us to come up with inventions. I remember our boss Dr. Rajchman visiting us once a week and asking if we had any new patent disclosures.

Of course, all work on superconductivity stopped with the advent of semiconductors, which, at the time, we didn't believe would take off. We didn't believe that miniaturization technology would succeed as it did. We also didn't believe they could overcome the vulnerability problem where the memory would be wiped if the battery ran out. Obviously, they did, and semiconductor technology wiped out all its competitors. At that point, I was working for a company called Electronic Memories, and the rise of semiconductors left me without a job. That was how I came to academia, where I pursued my old dreams of doing pattern recognition and image encoding.

MARTIN FORD: Did you go directly to UCLA from Electronic Memories?

JUDEA PEARL: I tried to go to the University of Southern California, but they wouldn't hire me because I was too sure of myself. I wanted to teach software, even though I'd never programmed before, and the Dean threw me out of his office. I ended up at UCLA because they gave me a chance of doing the things that I wanted to do, and I slowly migrated into AI from pattern recognition, image encoding, and decision theory. The early days of AI were dominated by chess and other game-playing programs, and that enticed me in the beginning, because I saw there a metaphor for capturing human intuition. That was and remained my life dream, to capture human intuition on a machine.

In games, the intuition comes about in the way you evaluate the strength of a move. There was a big gap between what machines can do and what experts can do, and the challenge was to capture experts' evaluation in the machine. I ended up doing some analytical work and came up with a nice explanation of what heuristics is all about, and an automatic way of discovering heuristics, it is still in use today. I believe I was the first to show that alpha-beta search is optimal, as well other mathematical results about what makes one heuristic better than another. All of that work was compiled in my book, *Heuristics*, which came out in 1984 [1]. Then expert systems came to the scene, and people were excited about capturing different kinds of heuristics—not the heuristic of a chess master, but the intuition of highly-paid professionals, like a physician or a mineral explorer. The idea was to emulate professional performance on a computer system, either to replace or to assist the professional. I looked at expert systems as another challenge of capturing intuition.

MARTIN FORD: Just to clarify, expert systems are mostly based on rules, correct? If this is true, then do that, etc.

JUDEA PEARL: Correct, it was based on rules, and the goal was to capture the mode of operation of an expert, what makes an expert decide one way or the other while engaging in professional work.

What I did, was to replace it with a different paradigm. For example, instead of modeling a physician—the expert—we modeled the disease. You don't have to ask the expert what they do. Instead, you ask, what kind of symptoms you expect to see if you have malaria or if you have the flu; and what do you know about the disease? On the basis of this information, we built a diagnosis system that could examine a collection of symptoms and come out with the suspected disease. It also works for mineral exploration, for troubleshooting, or for any other expertise.

MARTIN FORD: Was this based on your work on heuristics, or are you referring now to Bayesian networks?

JUDEA PEARL: No, I left heuristics the moment my book was published in 1984, and I started working on Bayesian networks and uncertainty management. There were many proposals at the time for managing uncertainties, but they didn't gel with the dictates of probability theory and decision theory, and I wanted to do it correctly and efficiently.

MARTIN FORD: Could you talk about your work on Bayesian networks? I know they are used in a lot of important applications today.

JUDEA PEARL: First, we need to understand the environment at the time. There was a tension between the scruffies and the neaties. The scruffies just wanted to build a system that works, not caring about guarantees or whether their methods comply with any theory or not. The neaties wanted to understand why it worked and make sure that they have performance guarantees of some kind.

MARTIN FORD: Just to clarify, these were nicknames for two groups of people with different attitudes.

JUDEA PEARL: Yes. We see the same tension today in the machine learning community, where some people like to get machines to do important jobs, regardless of whether they're doing it optimally or whether the system can explain itself—as long as the job is being done. The neaties would like to have explainability and transparency, systems that can explain themselves and systems that have performance guarantees.

Well, at that time, the scruffies were in command, and they still are today, because they have a good conduit to funders and to industry. Industry, however,

is short-sighted and requires short-term success, which creates an imbalance in research emphasis. It was the same in the Bayesian network days; the scruffies were in command. I was among the few loners who advocated doing things correctly by the rules of probability theory. The problem was that probability theory, if you adhere to it in the traditional way, would require exponential time and exponential memory, and we couldn't afford these two resources.

I was looking for a way of doing it efficiently, and I was inspired by the work of David Rumelhart, a cognitive psychologist who examined how children read text so quickly and reliably. His proposal was to have a multi-layered system going from the pixel level to the semantic level, then the sentence level and the grammatical level, and they all shake hands and pass messages to each other. One level doesn't know what the other's doing; it's simply passing messages. Eventually, these messages converge on the correct answer when you read a word like "the car" and distinguish it from "the cat," depending on the context in the narrative.

I tried to simulate his architecture in probability theory, and I couldn't do it very well until I discovered that if you have a tree as a structure connecting the modules, then you do have this convergence property. You can propagate messages asynchronously, and eventually, the system relaxes to the correct answer. Then we went to a polytree, which is a fancier version of a tree, and eventually, in 1985, I published a paper about general Bayesian networks [2].

This architecture really caught us by surprise because it was very easy to program. A programmer didn't have to use a supervisor to oversee all the elements, all they had to do was to program what one variable does when it wakes up and decides to update its information. That variable then sends messages to its neighbors. The neighbors send messages to their neighbors, and so on. The system eventually relaxes to the correct answer.

The ease of programming was the feature that made Bayesian networks acceptable. It was also made acceptable by the idea that you can program the disease and not the physician—the domain, and not the professional that deals with the domain—that made the system transparent. The users of the system understood why the system provided one result or another, and they understood how to modify the system when things changed in the environment. You had the advantage of modularity, which you get when you model the way things work in nature.

It's something that we didn't realize at the time, mainly because we didn't realize the importance of modularity. When we did, I realized that it is causality that gives us this modularity, and when we lose causality, we lose modularity, and we enter into no-man's land. That means that we lose transparency, we lose reconfigurability, and other nice features that we like. By the time that I published my book

on Bayesian networks in 1988, though, I already felt like an apostate because I knew already that the next step would be to model causality, and my love was already on a different endeavor.

MARTIN FORD: We always hear people saying that “correlation is not causation,” and so you can never get causation from the data. Bayesian networks do not offer a way to understand causation, right?

JUDEA PEARL: No, Bayesian networks could work in either mode. It depends on what you think about when you construct them.

MARTIN FORD: The Bayesian idea is that you update probabilities based on new evidence so that your estimate should get more accurate over time. That’s the basic concept that you’ve built into these networks, and you figured out a very efficient way to do that for a large number of probabilities. It’s clear that this has become a really important idea in computer science and AI because it’s used all over the place.

JUDEA PEARL: Using Bayes’ rule is an old idea; doing it efficiently was the hard part. That’s one of the things that I thought was necessary for machine learning. You can get evidence and use the Bayesian rule to update the system to improve its performance and improve the parameters. That’s all part of the Bayesian scheme of updating knowledge using evidence, it is probabilistic, not causal knowledge, so it has limitations.

MARTIN FORD: But it’s used quite frequently, for example, in voice recognition systems and all the devices that we’re familiar with. Google uses it extensively for all kinds of things.

JUDEA PEARL: People tell me that every cellphone has a Bayesian network doing error correction to minimize transmission noise. Every cellphone has a Bayesian network and belief propagation, that’s the name we gave to the message passing scheme. People also tell me that Siri has a Bayesian network in it, although Apple is too secretive about it, so I haven’t been able to verify it.

Although Bayesian updating is one of the major components in machine learning today, there has been a shift from Bayesian networks to deep learning, which is less transparent. You allow the system itself to adjust the parameters without knowing the function that connects input and output. It’s less transparent than Bayesian networks, which had the feature of modularity, and which we didn’t realize was so important. When you model the disease, you actually model the cause and effect relationship of the disease, not the expert, and you get modularity. Once we realize that, the question begs itself: What is this ingredient that you and I call

“cause and effect relationships”? Where does it reside, and how do you handle it? That was the next step for me.

MARTIN FORD: Let’s talk about causation. You published a very famous book on Bayesian networks, and it was really that paper that led to Bayesian techniques becoming so popular in computer science. But before that book was even published, you were already starting to think about moving on to focus on causation?

JUDEA PEARL: Causation was part of the intuition that gave rise to Bayesian networks, even though the formal definition of Bayesian networks is purely probabilistic. You do diagnostics, you make predictions, and you don’t deal with interventions. If you don’t need interventions, you don’t need causality—theoretically. You can do everything that a Bayesian network does with purely probabilistic terminology. However, in practice, people noticed that if you structure the network in the causal direction, things are much easier. The question was why.

Now we understand that we were craving for features of causality that we didn’t even know come from causality. These were: modularity, reconfigurability, transferability, and more. By the time I looked into causality, I had realized that the mantra “correlation does not imply causation” is much more profound than we thought. You need to have causal assumptions before you can get causal conclusions, which you cannot get from data alone. Worse yet, even if you are willing to make causal assumptions, you cannot express them.

There was no language in science in which you can express a simple sentence like “mud does not cause rain,” or “the rooster does not cause the sun to rise.” You couldn’t express it in mathematics, which means that even if you wanted to take it for granted that the rooster does not cause the sun to rise, you couldn’t write it down, you couldn’t combine it with data, and you couldn’t combine it with other sentences of this kind.

In short, even if you agree to enrich the data with causal assumptions, you couldn’t write down the assumptions. It required a whole new language. This realization was really a shock and a challenge for me because I grew up on statistics, and I believed that scientific wisdom lies in statistics. Statistics allows you to do induction, deduction, abduction, and model updating. And here I find the language of statistics crippled in hopeless helplessness. As a computer scientist, I was not scared because computer scientists invent languages to fit their needs. But what is the language that should be invented, and how do we marry this language with the language of data?

Statistics speaks a different language—the language of averages, of hypothesis testing, summarizing data and visualizing it from different perspectives. All of this is the language of data, and here comes another language, the language of

cause and effect. How do we marry the two so that they can interact? How do I take assumptions about cause and effect, combine them with the data that I have, and then get conclusions that tell me how nature works? That was my challenge as a computer scientist and as a part-time philosopher. This is essentially the role of a philosopher, to capture human intuition and formalize it in a way that it can be programmed on a computer. Even though philosophers don't think about the computer, if you look closely at what they are doing, they are trying to formalize things as much as they can with the language available to them. The goal is to make it more explicable and more meaningful. At this point, computer scientists can program a machine to perform cognitive functions that puzzle philosophers.

MARTIN FORD: Did you invent the technical language or the diagrams used for describing causation?

JUDEA PEARL: No, I didn't invent that. The basic idea was conceived in 1920 by a geneticist named Sewall Wright, who was the first to write down a causal diagram with arrows and nodes, like a one-way city map. He fought all his life to justify the fact that you can get things out of this diagram that statisticians could not get from regression, association, or from correlation. His methods were primitive, but they proved the point; he could indeed get things that the statisticians could not get.

What I did was to take Sewall Wright's diagrams seriously and invested into them all my computer science background, reformalized them, and exploited them to their utmost. I came up with a causal diagram as a means of encoding scientific knowledge and as a means of guiding machines in the task of figuring out cause-effect relationships in various sciences, from medicine, to education, to climate warming. These were all areas where scientists worry about what causes what, how nature transmits the information from cause to effect, what are the mechanisms involved, how do you control it, and how do you answer practical questions which involve cause-effect relationships.

This has been my life's challenge for the past 30 years. I published a book on that in 2000, with the second edition in 2009, called *Causality* [4]. I co-authored a gentler introduction in 2016 [5]. And this year, I co-authored *The Book of Why* [6], which is a general audience book explaining the challenge in down-to-earth terms, so that people can understand causality even without knowing equations. Equations of course help to condense things and to focus on things, but you don't have to be a rocket scientist to read *The Book of Why*. You just have to follow the conceptual development of the basic ideas. In that book, I look at history from a causal lens perspective; I asked what conceptual breakthroughs made a difference in the way we think about causation, rather than what experiments discovered one drug or another.

MARTIN FORD: I've been reading *The Book of Why* and I'm enjoying it. I think one of the main outcomes of your work is that causal models are now very important in the social and natural sciences. In fact, I just saw an article the other day, written by a quantum physicist who used causal models to prove something in quantum mechanics. So clearly your work has had a big impact in those areas.

JUDEA PEARL: I read that article. In fact, I put it on my next-to-read list because I couldn't quite understand the phenomena that they were so excited about.

MARTIN FORD: One of the main points I took away from *The Book of Why* is that, while natural and social scientists have really begun to use the tools of causation, you feel that the field of AI is lagging behind. You think AI researchers will have to start focusing on causation in order for the field to progress.

JUDEA PEARL: Correct. Causal modeling is not at the forefront of the current work in machine learning. Machine learning today is dominated by statisticians and the belief that you can learn everything from data. This data-centric philosophy is limited.

I call it curve fitting. It might sound derogatory, but I don't mean it in a derogatory way. I mean it in a descriptive sense that what people are doing in deep learning and neural networks is fitting very sophisticated functions to a bunch of points. These functions are very sophisticated, they have thousands of hills and valleys, they're intricate, and you cannot predict them in advance. But they're still just a matter of fitting functions to a cloud of points.

This philosophy has clear theoretical limitations, and I'm not talking about opinion, I'm talking about theoretical limitations. You cannot do counterfactuals, and you cannot think about actions that you've never seen before. I describe it in terms of three cognitive levels: seeing, intervening, and imagining. Imagining is the top level, and that level requires counterfactual reasoning: how would the world look like had I done things differently? For example, what would the world look like had Oswald not killed Kennedy, or had Hillary won the election? We think about those things and can communicate with those kinds of imaginary scenarios, and we are quite comfortable to engage in this "let's pretend" game.

The reason why we need this capability is to build new models of the world. Imagining a world that does not exist gives us the ability to come up with new theories, new inventions, and also to repair our old actions so as to assume responsibility, regret, and free will. All of this comes as part of our ability to generate worlds that do not exist but could exist, and still generate them widely, not wildly. We have rules for generating plausible counterfactuals that are not whimsical. They have their own inner structure, and once we understand this logic, we can build

machines that imagine things, that assume responsibility for their actions, and understand ethics and compassion.

I'm not a futurist and I try not to talk about things that I don't understand, but I did some thinking, and I believe I understand how important counterfactuals are in all these cognitive tasks that people dream of, which eventually will be implemented on a computer. I have a few basic sketches of how we can program free will, ethics, morality, and responsibility into machines, but these are in the realm of sketches. The basic thing is that we know today what it takes to interpret counterfactuals and understand cause and effect.

These are the mini-steps toward general AI, but there's a lot we can learn from these steps, and that's what I'm trying to get the machine learning community to understand. I want them to understand that deep learning is a mini-step toward general AI. We need to learn what we can from the way theoretical barriers were circumvented in causal reasoning, so that we can circumvent them in general AI.

MARTIN FORD: So, you're saying that deep learning is limited to analyzing data and that causation can never be derived from data alone. Since people are able to do causal reasoning, the human mind must have some built-in machinery that allows us to create causal models. It's not just about learning from data.

JUDEA PEARL: To create is one thing, but even if somebody creates it for us, our parents, our peers, our culture, we need to have the machinery to utilize it.

MARTIN FORD: Right. It sounds like a causal diagram, or a causal model is really just a hypothesis. Two people might have different causal models, and somewhere in our brain is some kind of machinery that allows us to continuously create these causal models internally, and that's what allows us to reason based on data.

JUDEA PEARL: We need to create them, to modify them, and to perturb them when the need arises. We used to believe that malaria is caused by bad air, now we don't. Now we believe it's caused by a mosquito called Anopheles. It makes a difference because if it is bad air, I will carry a breathing mask the next time I go to the swamp; and if it's an Anopheles mosquito, I'll carry a mosquito net. These competing theories make a big difference in how we act in the world. The way that we get from one hypothesis to another was by trial and error; I call it playful manipulation.

This is how a child learns causal structure, by playful manipulation, and this is how a scientist learns causal structure—playful manipulation. But we have to have the abilities and the template to store what we learn from this playful manipulation so we can use it, test it, and change it. Without the ability to store it in a parsimonious encoding, in some template in our mind, we cannot utilize it, nor can we

change it or play around with it. That is the first thing that we have to learn; we have to program computers to accommodate and manage that template.

MARTIN FORD: So, you think that some sort of built-in template or structure should be built into an AI system so it can create causal models? DeepMind uses reinforcement learning, which is based on practice or trial and error. Perhaps that would be a way of discovering causal relationships?

JUDEA PEARL: It comes into it, but reinforcement learning has limitations, too. You can only learn actions that have been seen before. You cannot extrapolate to actions that you haven't seen, like raising taxes, increasing the minimum wage, or banning cigarettes. Cigarettes have never been banned before, yet we have machinery that allows us to stipulate, extrapolate, and imagine what could be the consequences of banning cigarettes.

MARTIN FORD: So, you believe that the capability to think causally is critical to achieving what you'd call strong AI or AGI, artificial general intelligence?

JUDEA PEARL: I have no doubt that it is essential. Whether it is sufficient, I'm not sure. However, causal reasoning doesn't solve every problem of general AI. It doesn't solve the object recognition problem, and it doesn't solve the language understanding problem. We basically solved the cause-effect puzzle, and we can learn a lot from these solutions so that we can help the other tasks circumvent their obstacles.

MARTIN FORD: Do you think that strong AI or AGI is feasible? Is that something you think will happen someday?

JUDEA PEARL: I have no doubt that it is feasible. But what does it mean for me to say no doubt? It means that I am strongly convinced it can be done because I haven't seen any theoretical impediment to strong AI.

MARTIN FORD: You said that way back around 1961, when you were at RCA, people were already thinking about this. What do you think of how things have progressed? Are you disappointed? What's your assessment of progress in artificial intelligence?

JUDEA PEARL: Things are progressing just fine. There were a few slowdowns, and there were a few hang-ups. The current machine learning concentration on deep learning and its non-transparent structures is such a hang-up. They need to liberate themselves from this data-centric philosophy. In general, the field has been progressing immensely, because of technology and because of the people that the field attracts. The smartest people in science.

MARTIN FORD: Most of the recent progress has been in deep learning. You seem somewhat critical of that. You've pointed out that it's like curve fitting and it's not transparent, but actually more of a black-box that just generates answers.

JUDEA PEARL: It's curve fitting, correct, it's harvesting low-hanging fruits.

MARTIN FORD: It's still doing amazing things.

JUDEA PEARL: It's doing amazing things because we didn't realize there are so many low-hanging fruits.

MARTIN FORD: Looking to the future, do you think that neural networks are going to be very important?

JUDEA PEARL: Neural networks and reinforcement learning will all be essential components when properly utilized in causal modeling.

MARTIN FORD: So, you think it might be a hybrid system that incorporates not just neural networks, but other ideas from other areas of AI?

JUDEA PEARL: Absolutely. Even today, people are building hybrid systems when you have sparse data. There's a limit, however, to how much you can extrapolate or interpolate sparse data if you want to get cause-effect relationships. Even if you have infinite data, you can't tell the difference between A causes B and B causes A.

MARTIN FORD: If someday we have strong AI, do you think that a machine could be conscious, and have some kind of inner experience like a human being?

JUDEA PEARL: Of course, every machine has an inner experience. A machine has to have a blueprint of some of its software; it could not have a total mapping of its software. That would violate Turing's halting problem.

It's feasible, however, to have a rough blueprint of some of its important connections and important modules. The machine would have to have some encoding of its abilities, of its beliefs, and of its goals and desires. That is doable. In some sense, a machine already has an inner self, and more so in the future. Having a blueprint of your environment, how you act on and react to the environment, and answering counterfactual questions amount to having an inner self. Thinking: What if I had done things differently? What if I wasn't in love? All this involves manipulating your inner self.

MARTIN FORD: Do you think machines could have emotional experiences, that a future system might feel happy, or might suffer in some way?

JUDEA PEARL: That reminds me of *The Emotion Machine*, a book by Marvin Minsky. He talks about how easy it is to program emotion. You have chemicals

floating in your body, and they have a purpose, of course. The chemical machine interferes with, and occasionally overrides the reasoning machine when urgencies develop. So, emotions are just a chemical priority-setting machine.

MARTIN FORD: I want to finish by asking you about some of the things that we should worry about as artificial intelligence progresses. Are there things we should be concerned about?

JUDEA PEARL: We have to worry about artificial intelligence. We have to understand what we build, and we have to understand that we are breeding a new species of intelligent animals.

At first, they are going to be domesticated, like our chickens and our dogs, but eventually, they will assume their own agency, and we have to be very cautious about this. I don't know how to be cautious without suppressing science and scientific curiosity. It's a difficult question, so I wouldn't want to enter into a debate about how we regulate AI research. But we should absolutely be cautious about the possibility that we are creating a new species of super-animals, or in the best case, a species of useful, but exploitable, human beings that do not demand legal rights or minimum wage.

JUDEA PEARL was born in Tel Aviv and is a graduate of the Technion-Israel Institute of Technology. He came to the United States for postgraduate work in 1960, and the following year he received a master's degree in electrical engineering from Newark College of Engineering, now New Jersey Institute of Technology. In 1965, he simultaneously received a master's degree in physics from Rutgers University and a PhD from the Brooklyn Polytechnic Institute, now Polytechnic Institute of New York University. Until 1969, he held research positions at RCA David Sarnoff Research Laboratories in Princeton, New Jersey and Electronic Memories, Inc. Hawthorne, California.

Judea joined the faculty of UCLA in 1969, where he is currently a professor of computer science and statistics and director of the Cognitive Systems Laboratory. He is known internationally for his contributions to artificial intelligence, human reasoning, and philosophy of science. He is the author of more than 450 scientific papers and three landmark books: Heuristics (1984), Probabilistic Reasoning (1988), and Causality (2000; 2009).

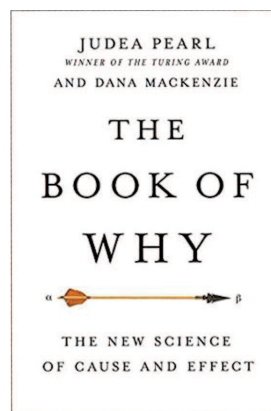
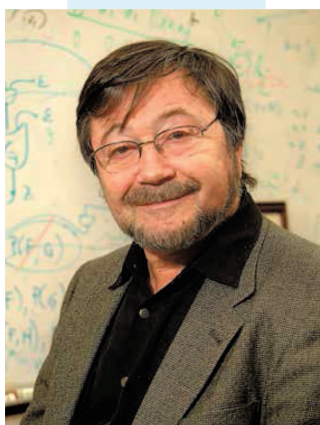
A member of the National Academy of Sciences, the National Academy of Engineering and a founding Fellow of the American Association for Artificial Intelligence, Judea is the recipient of numerous scientific prizes, including three awarded in 2011: the Association for Computing Machinery A.M. Turing Award for his fundamental contributions to artificial intelligence through the development of a calculus for probabilistic and causal reasoning, the David E. Rumelhart Prize for Contributions to the Theoretical Foundations of Human Cognition, and the Harvey Prize in Science and Technology from

Technion—Israel Institute of Technology. Other honors include the 2001 London School of Economics Lakatos Award in Philosophy of Science for the best book in the philosophy of science, the 2003 ACM Allen Newell Award for “seminal contributions that extend to philosophy, psychology, medicine, statistics, econometrics, epidemiology and social science,” and the 2008 Benjamin Franklin Medal for Computer and Cognitive Science from the Franklin Institute.

References

1. J. Pearl. 1984. *Heuristics: Intelligent Search Strategies for Computer Problem Solving*. Addison-Wesley.
2. J. Pearl. 1985. Bayesian networks: a model of self-activated memory for evidential reasoning. In *Proceedings, Cognitive Science Society*, 329–334.
3. J. Pearl. 1988. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, San Mateo, CA.
4. J. Pearl. 2000. *Causality: Models, Reasoning, and Inference* (Second Edition, 2009). Cambridge University Press.
5. J. Pearl, M. Glymour, and N. P. Jewell. 2016. *Causal Inference in Statistics: A Primer*. Wiley, New York, NY.
6. J. Pearl and D. Mackenzie. 2018. *The Book of Why*. Basic Books, New York, NY.

An Interview with Ron Wassertein on How *The Book of Why* Transforms Statistics



Judea Pearl's *The Book of Why* is a nontechnical book for the general public that discusses recent advances in causal inference.

*Judea Pearl, a longtime ASA member, was interviewed in November of 2012 (see <https://bit.ly/2LdNidA>) after receiving the Turing Award from the Association of Computing Machinery. He has recently published a book, *The Book of Why: The New Science of Cause and Effect* (with Dana Mackenzie), that aims to familiarize the general, nontechnical public with recent advances in causal inference. ASA Executive Director Ron Wasserstein interviews him again here to find out what message he thinks his new book sends to Amstat News readers.*

Originally published in *Interview. Amstat News*, 8 September, 2018. <https://magazine.amstat.org/blog/2018/08/01/judeapearl-interview/>.

Republished with permission.

***The Book of Why* is making a splash in statistics, as well as in machine learning and other data-intensive sciences. I would like to start with a question that you have probably heard many times: What brought you to write the book?**

I have official and unofficial answers to this question.

The official answers: First, I have found it both timely and exciting to lay before the public the amazing story of a science that has changed the way we understand scientific claims and yet has remained below the radar to the general public. As we enter the era of big data and machine learning, it is important to share with the public our current understanding of how this new science is likely to affect our lives in the 21st century.

Second, as a part-time philosopher, I have found it intriguing to narrate the history of statistics as viewed from the special lens of its orphaned sister: causation. The story of this “forbidden love” was never told before and, believe me, it is full of mystery, intrigue, personalities, dogmatic orthodoxy, and heroic champions of truth and conviction.

Finally, my unofficial reason is to incite a rebellious spirit among rank-and-file statisticians, so the excitement that currently fuels causality research in academia percolates down to education and to practice. In other words, I am impatient with the slow pace at which the tools of causal inference are becoming an organic part of statistical thinking.

You expressed a similar impatience in our interview six years ago. And you have initiated the ASA Causality in Statistical Education Award to close the growing gap between research and education. Hasn't this initiative met your expectations?

It has. But, with age, my impatience grew stronger and less forgiving. Of course, the availability of instructional material made it easier for instructors to introduce aspects of causal inference in graduate courses, but it was not sufficient to change the curriculum of undergraduate classes. Nor was it sufficient to reshape the minds of practicing statisticians or high-profile academics who are too busy to sort out what all the causal inference “hype” is about.

What *The Book of Why* is doing can be described as “the democratization of causal inference.” It awakens the untrained students to the realization that “it’s easy and who needs the ‘experts’ and all their quibbles?” As a result, the book is accomplishing what I have failed to achieve in the past 30 years through hard labor and scholarly discussion with the leading statisticians of our time—a mass uprising of common sense.

I have read that some statisticians find your claims to be “hard to swallow,” especially your characterization of causal inference as “The Causal Revolution” and your depiction of statisticians as antagonistic to causal thinking. Can you comment on these sentiments?

These are not only sentiments but natural complaints voiced by practicing statisticians who are genuinely surprised by how the history of statistics is viewed from the causal lens.

Take for instance the mantra “correlation does not imply causation,” which every statistics student has learned to chant, demonstrate, and internalize. *The Book of Why* dissects this mantra to far-reaching conclusions that seem indeed “hard to swallow,” even to seasoned statisticians.

First, it can be strengthened to assert that no causal conclusion can ever be obtained without some causal assumptions (or experiments) to support the conclusion. This is hard to swallow because it sounds circular, and because if you look at the statistical literature from 1832 to 1974, you will find many ideas about what is needed to substantiate causal conclusions (e.g., Yule, Fisher, Neyman, Hill, Cox, Cochran), but not one causal assumption—at least not formally.

This raises an interesting question: Why couldn’t these giants of statistics come up with a simple principle, telling us what assumptions are needed for establishing a given conclusion, and let us judge—for any given situation—whether it is plausible to make those assumptions? And here comes the second surprise that is even harder for people to swallow: Even if they knew the needed assumptions, statisticians could not have articulated them mathematically—they simply did not have the language to do so.

Readers refuse to accept this linguistic deficiency until I ask them to write down a mathematical expression for the sentence, “The rooster crow does not cause the sun to rise.” Failing this elementary exercise drives people to realize a totally new notational system is needed; the beautiful and powerful language of probability theory and its many extensions cannot make up for this deficiency.

The needed notation first came into being in 1920, when the geneticist Sewall Wright put down on paper a new mathematical object: A causal diagram. Thus, statistics was separated from causality, not by antagonism or disdain, but by a language barrier—the toughest barrier for humans to acknowledge and to cross. Now that the barrier is behind us, it is only natural we should call the crossing a “Causal Revolution.”

These are interesting theoretical points, but I wonder if they are likely to have significant impacts on the practice of statistics or on statistical education.

The most significant practical impact of the Causal Revolution would probably be a continuous erosion of the supremacy of randomized clinical trials (RCT) in the development and evaluation of drugs, therapeutical procedures, and social and educational policies. Last year, for example, the editors of one of the two leading medical journals in America stated that authors should not talk about causation unless they have conducted a randomized clinical trial.

Miguel Hernan of Harvard and several other specialists in public health vigorously protested this restriction, and Hernan wrote, “The biggest disservice of statistics to science has been to make ‘causal’ into a dirty word, the C-word that researchers have learned to avoid.”

Indeed, considering the practical difficulties of conducting an ideal RCT and its inherent sensitivity to sample selection bias, observational studies have a definite advantage: They interrogate the target populations at their natural habitats, not in artificial environments choreographed by experimental protocols.

The development of a new toolkit that allows scientists to estimate causal effects from observational studies now opens a wide variety of applications—from medicine to social science to ecology—free from problems of ethics, costs, and external validity that plague randomized clinical trials.

True, observational studies are necessarily sensitive to modeling assumptions that must be defended on scientific grounds. However, the transparency with which those conceptual assumptions are displayed, coupled with the ability of testing them against data, now make observational studies serious contenders to RCTs.

I would like to go back to education and ask what you believe would induce a typical statistics instructor to introduce aspects of causal inference in a standard statistics class.

Curious students who read *The Book of Why* will make it impossible for statistics instructors to skip such aspects.

Take for instance Simpson’s paradox, a phenomenon discussed in every statistics class, usually for the purpose of demonstrating that “correlation is not causation.” The discussion usually ends with a song of praise to statistical tables for showing us that the reversal can indeed occur in the data, hence the paradox does not exist. Done. Some instructors go a bit further and praise the table for protecting us from naïve beliefs in miracle drugs that are good for men, good for women, and bad for the population.

Now imagine an inquisitive student raising his/her hand and asking the very obvious question: So, what do we do if we find Simpson’s reversal in the data? Shall we believe the aggregated data or the disaggregated data? I do not believe

any instructor would in good faith be able to evade this question, suspecting the student knows the answer; it takes a few lines to describe. In other words, instructors would not be able to skip the causal implications of Simpson's paradox, as their professors did to them.

The same applies to Lord's paradox, spurious correlations, instrumental variables, confounders, and other causal concepts that were used to embarrass statistics instructors in the past.

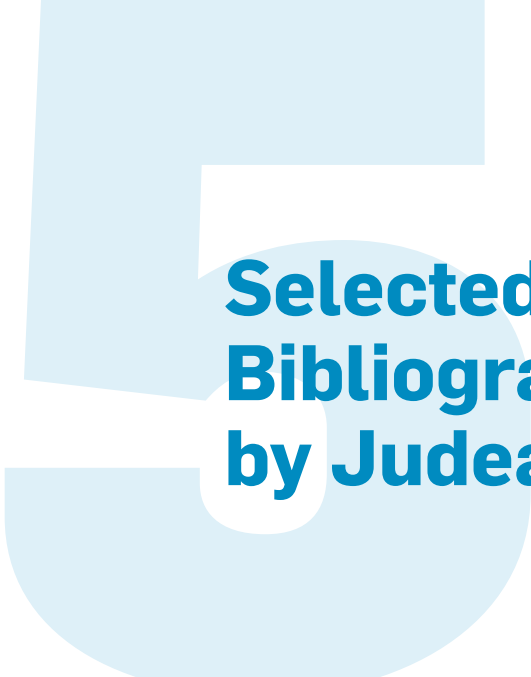
The graphical approach you advocate in the book is but one of several approaches currently used in causal inference. Would a reader versed in potential outcome analysis feel comfortable with your methodology?

Not only comfortable, but enlightened and liberated. Researchers entrenched in potential outcome analysis will discover, to their amazement, that the following three notorious weaknesses of potential outcomes can easily be overcome:

- Assumptions of “conditional ignorability,” which currently underlie every potential outcome study, can be made not because they facilitate available statistical routines, but when they are truly believed to hold in the world. They are, in fact, vividly displayed in our model of the world (i.e., the causal diagram), where they can be scrutinized for plausibility, completeness, and consistency.
- When assumptions of “conditional ignorability” do not hold, it is not the end of the world; the analysis can continue, and causal questions answered using other types of assumptions the model may license.
- Modeling assumptions need not remain opaque or data-blind; they can be tested for compatibility with the available data, and the model tells us how.

Making these three bullets available to researchers from the potential outcome camp will break through a wall of cultural isolation and enable them to communicate with the rest of the research community in a common, unified language.

To summarize, the democratization of causal inference is bringing about a globalization of common sense and a breakdown of cultural barriers. I am gratified to see *The Book of Why* contributing to this process. ■



Selected Annotated Bibliography by Judea Pearl

Search and Heuristics

1. J. Pearl. 1980. Asymptotic properties of minimax trees and game-searching procedures. *Artif. Intell.* 14, (September 1980), 113–138. DOI: [https://doi.org/10.1016/0004-3702\(80\)90037-5](https://doi.org/10.1016/0004-3702(80)90037-5). *One of the first papers to discover and analyze “phase transition” in combinatorial problem; introduced new mathematical techniques into the AI literature.*
2. J. Pearl. 1983. Knowledge versus search: A quantitative analysis using A*. *Artif. Intell.* 20, 1–13. DOI: [https://doi.org/10.1016/0004-3702\(83\)90013-9](https://doi.org/10.1016/0004-3702(83)90013-9). *Established the relationships between heuristic accuracy and search algorithm complexity.*
3. J. Pearl. 1983. On the nature of pathology in game searching. *Artif. Intell.* 20, 427–453. DOI: [https://doi.org/10.1016/0004-3702\(83\)90004-8](https://doi.org/10.1016/0004-3702(83)90004-8). *Proved that, under the standard model of game trees, deeper search does not necessarily improve play, and showed that this paradox is resolved by correct probabilistic updating of beliefs.*
4. R. Karp and J. Pearl. 1983. Searching for an optimal path in a tree with random costs. *Artif. Intell.* 21, 99–116. DOI: [https://doi.org/10.1016/S0004-3702\(83\)80006-X](https://doi.org/10.1016/S0004-3702(83)80006-X). *Identified a phase transition property for a very simple path-finding problem, with significant complexity implications.*
5. J. Pearl. 1983. On the discovery and generation of certain heuristics. *AI Magazine*, Winter/Spring, 23–33. DOI: <https://doi.org/10.1609/aimag.v4i1.385>. *The first paper on the systematic generation of admissible heuristics (lower bounds on*

optimal solution costs) by relaxing formally represented problem definitions; this idea led to dramatic advances in automated planning systems.

6. J. Pearl. 1984. *Heuristics: Intelligent Search Strategies for Computer Problem Solving*. Addison-Wesley. *Synthesized essentially everything known up to that point about intelligent methods for search and game playing, much of it Pearl's own work; also the first textbook to treat AI topics formally at a technically advanced level.*
7. R. Dechter and J. Pearl. 1985. Generalized best-first search strategies and the optimality of A^* . *J ACM* 32, 505–536. DOI: <https://doi.org/10.1145/3828.3830>. *Proved that A^* is the most efficient member of a very broad class of problem-solving algorithms.*

Bayesian Networks

8. J. Pearl. 1982. Reverend Bayes on inference engines: A distributed hierarchical approach. In *Proceedings, AAAI-82. The paper that began the probabilistic revolution in AI by showing how several desirable properties of reasoning systems can be obtained through sound probabilistic inference. It introduced tree-structured networks as concise representations of complex probability models, identified conditional independence relationships as the key organizing principle for uncertain knowledge, and described an efficient message-passing algorithm, which later became the engine for most practical applications of Bayesian Networks.*
9. J. Kim and J. Pearl. 1983. A computational model for combined causal and diagnostic reasoning in inference systems. In *Proceedings, IJCAI-83, 190–193. Generalized the tree-structured network to poly-trees, including colliders, and allows reasoning by “explaining away” competing causes.*
10. J. Pearl. 1985. Learning hidden causes from empirical data. In *Proceedings, IJCAI-85. Pioneered the field of “causal discovery” by developing algorithms that learn the structures of causal models from raw data.*
11. J. Pearl. 1986. On the logic of probabilistic dependencies. In *Proceedings, AAAI-86. One of several papers establishing the connection between graphical models and conditional independence relationships (d-separation), later labeled, “The 2nd Law of Causal Inference.”*
12. J. Pearl. 1986. Fusion, propagation and structuring in belief networks. *Artif. Intell.* 29, 241–288. DOI: [https://doi.org/10.1016/0004-3702\(86\)90072-X](https://doi.org/10.1016/0004-3702(86)90072-X). *The key technical paper on representation and inference in general Bayesian networks; by 1991 this had become the most cited paper in the Artificial Intelligence journal.*

13. J. Pearl and A. Paz. 1987. Graphoids: A graph-based logic for reasoning about relevance relations. In B. du Boulay et al. (Eds.), *Advances in Artificial Intelligence II*, North-Holland. *Establishes an axiomatic characterization of the properties that enable probabilities and other relational systems to be represented by graphs.*
14. J. Pearl. 1987. Evidential reasoning using stochastic simulation of causal models. *Artif. Intell.* 32, 245–257. DOI: [https://doi.org/10.1016/0004-3702\(87\)90012-9](https://doi.org/10.1016/0004-3702(87)90012-9). *Derived a general approximation algorithm for Bayesian network inference using Markov chain Monte Carlo (MCMC). This was the first significant use of MCMC in mainstream AI.*
15. J. Pearl. 1988. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann. *Explained the philosophical, cognitive, and technical basis for a probabilistic view of knowledge, reasoning, and decision-making. One of the most cited works in computer science, this book initiated the modern era in AI and established probabilistic inference as the standard of handling uncertainty in computer systems.*

Causality

16. J. Pearl and T.S. Verma. 1991. A theory of inferred causation. In *Proceedings, KR-91*. *Introduces minimal-model semantics as a basis for causal discovery and shows that causal directionality can be inferred from patterns of correlations without resorting to temporal information.*
17. J. Pearl. 1993. Graphical models, causality, and intervention. *Stat. Sci.* 8, 266–269. DOI: <https://www.jstor.org/stable/2245965>. *Introduces the back-door criterion for covariate selection, the first to guarantee bias-free estimation of causal effects.*
18. A. Balke and J. Pearl. 1994. Probabilistic Evaluation of Counterfactual Queries. In *Proceedings, National Conference on Artificial Intelligence*, 230–237. *Introduces the structural semantics of counterfactuals, later deemed “The First Law of Causal Inference.”*
19. J. Pearl. 1995. Causal diagrams for empirical research. *Biometrika.* 82, 4, 669–688. DOI: <https://doi.org/10.1093/biomet/82.4.669>. *Introduces the theory of causal diagrams and its associated do-calculus; the first (and still the only) mathematical method to enable a systematic removal of confounding bias in observations.*
20. J. Pearl. 1996. *The Art and Science of Cause and Effect*. UCLA Cognitive Systems Laboratory, Technical Report R-248. *Transcript of lecture given Thursday,*

October 29, 1996, as part of the UCLA 81st Faculty Research Lecture Series. Used later as epilogue to the book Causality (2000). Provides a panoramic view of the historical development of causal thoughts from antiquity to modern days.

21. J. Pearl. 2000. *Causality: Models, Reasoning, and Inference*. Cambridge University Press (Second Edition, 2009). *Building on theoretical results from 1987 to 2000, it lays out a complete framework for causal discovery, interventional analysis, and counterfactual reasoning, bringing mathematical rigor and conceptual clarity to an area previously considered off-limits. Winner of the 2001 Lakatos Prize for the most significant new work in the philosophy of science.*
22. J. Pearl. 2000. The logic of counterfactuals in causal inference (Discussion of “Causal inference without counterfactuals” by A.P. Dawid). *J. Am. Stat. Assoc.* 95, 428–435. *Demonstrates how counterfactual reasoning underlines scientific thought and argues against its exclusion from statistical analysis.*
23. J. Tian and J. Pearl. 2000. Probabilities of causation: Bounds and identification. *Ann. Math. Artif. Intell.* 28, 287–313. DOI: <https://doi.org/10.1023/A:1018912507879>. *Derives tight bounds on the probability that one observed event was the cause of another, in the legal sense of “but for,” thus providing a principled way of substantiating liability and responsibility from empirical data.*
24. J. Pearl. 2004. Robustness of causal claims. In *Proceedings, UAI-04*. *Offers a formal definition of robustness and develops a method for “sensitivity analysis,” i.e., assessing the degree to which causal claims are robust to model misspecification.*
25. J. Pearl. 2001. Direct and indirect effects. In *Proceedings, UAI-01*. *Establishes the theoretical basis of modern mediation analysis. Derives the “Mediation Formula” and provides graphical conditions for the identification of direct and indirect effect.*
26. J. Tian and J. Pearl. 2002. A general identification condition for causal effects. In *Proceedings, AAI-02*. *Uses the do-calculus to derive a general graphical condition for identifying causal effects from a combination of data and assumptions.*
27. J. Halpern and J. Pearl. 2005. Causes and explanations: A structural-model approach—Parts I and II. *Br J. Phil. Sci.* 56, 889–911. DOI: <https://www.jstor.org/stable/3541871>. *Establishes counterfactual conditions for one event to be perceived as the “actual cause” of another and for one event to provide an “explanation” of another.*
28. J. Pearl. 2009. Causal inference in statistics: An overview. *Stat. Surv.* 3, 96–146. DOI: <https://doi.org/10.1214/09-SS057>. *Describes a unified methodology for causal inference based on a symbiosis between graphs and counterfactual logic.*

29. J. Pearl. 2011. The algorithmization of counterfactuals. *Ann. Math. Artif. Intell.* 61, 29–39. DOI: <https://doi.org/10.1007/s10472-011-9247-9>. *Describes a computational model that explains how humans generate, evaluate, and distinguish counterfactual statements so swiftly and consistently.*
30. J. Pearl and E. Bareinboim. 2011. Transportability of causal and statistical relations: A formal approach. In *Proceedings, AAAI*. *Reduces the classical problem of external validity to mathematical transformations in the do-calculus, and establishes conditions under which experimental results can be generalized to new environments in which only passive observation can be conducted.*

Causal, Casual, and Curious

The entries below represent adventurous ideas and semi-heretical thoughts that emerged when, in 2013, I was given the opportunity to edit a fun section of the *Journal of Causal Inference* called “Causal, Casual, and Curious.” All the articles are available in the Internet.

31. J. Pearl. 2013. Linear models: A useful “microscope” for causal analysis. *J. Causal Inference.* 1, 1 (May 2013), 155–170. DOI: <https://doi.org/10.1515/jci-2013-0003>. *Demonstrates how causal phenomena of a non-trivial character can be understood, exemplified, and analyzed using linear structural equations, including Simpson’s paradox, case-control bias, selection bias, missing data, collider bias, reverse regression, bias amplification, near instruments, and measurement errors.*
32. J. Pearl. 2013. The curse of free-will and the paradox of inevitable regret. *J. Causal Inference.* 1, 2 (December 2013), 255–257. DOI: <https://doi.org/10.1515/jci-2013-0027>. *Challenges and clarifies the principles by which population data can be harnessed to guide personal decision-making, by examining situations in which an agent knows he/she will regret whatever action is taken.*
33. J. Pearl. 2014. Is scientific knowledge useful for policy analysis? A peculiar theorem says: No. *J. Causal Inference.* 2, 1 (March 2014), 109–112. DOI: <https://doi.org/10.1515/jci-2014-0017>. *Presents and resolves a paradox according to which the more we know about a problem domain the harder it is to predict the effects of policies.*
34. J. Pearl. 2014. Graphoids over counterfactuals. *J. Causal Inference.* 2, 2 (September 2014), 243–248. DOI: <https://doi.org/10.1515/jci-2014-0028>. *Augmenting the graphoid axioms with three additional rules enables us to handle*

independencies among observed as well as counterfactual variables, derive testable implications of ignorability assumptions, and test their identification.

35. J. Pearl. 2015. Conditioning on post-treatment variables. *J. Causal Inference*. 3, 1 (March 2015), 131–137. DOI: <https://doi.org/10.1515/jci-2015-0005>. *Includes Appendix (appended to published version). Compares ways of extracting information from post-treatment variables and clarifies the relationships between do-calculus conditioning and counterfactual conditioning.*
36. J. Pearl. 2015. Generalizing experimental findings. *J. Causal Inference*. 3, 2 (September 2015), 259–266. DOI: <https://doi.org/10.1515/jci-2015-0025>. *Compares ways in which researchers have attempted to generalize experimental finding across domains, and demonstrates that ignorability-based methods need to be enriched with structural assumptions in order to capture the full spectrum of conditions that permit generalizations.*
37. J. Pearl. 2016. The sure-thing principle. *J. Causal Inference*. 4, 1 (March 2016), 81–86. DOI: <https://doi.org/10.1515/jci-2016-0005>. *Traces the history of Jim Savage’s Sure Thing Principle, discusses its nuances, and evaluates its significance in the light of modern understanding of causal reasoning.*
38. J. Pearl. 2016. Lord’s paradox revisited—(Oh Lord! Kumbaya!). *J. Causal Inference*. 4, 2 (September 2016). DOI: <https://doi.org/10.1515/jci-2016-0021>. *Traces back Lord’s paradox from its original formulation in 1967, resolves it using modern tools of causal analysis, explains why it has resisted prior attempts at resolution, and addresses the general methodological issue of whether adjustments for preexisting conditions is justified in group comparison applications.*
39. J. Pearl. 2017. A linear “microscope” for interventions and counterfactuals. *J. Causal Inference*. 5, 1 (March 2017), 1–15. DOI: <https://doi.org/10.1515/jci-2017-0003>. *Using linear structural equations, the paper derives conditions for identifying total and direct effects, including the method of identifying counterfactual expressions, robustness to model misspecification, and generalization across populations.*
40. J. Pearl. 2017. Physical and metaphysical counterfactuals. *J. Causal Inference*. 5, 2 (September 2017). DOI: <https://doi.org/10.1515/jci-2017-0018>. *This paper leverages “imaging,” a process of “mass-shifting” among possible worlds, to define disjunctive counterfactuals, such as “had the color been either blue or purple.” It shows that every imaging operation can be given an interpretation in terms of a stochastic policy in which agents choose actions with certain probabilities.*
41. J. Pearl. 2018. What is gained from past learning. *J. Causal Inference*. 6, 1 (March 2018). DOI: <https://doi.org/10.1515/jci-2018-0005>. *Consider ways of*

leveraging previously learned information to novel situations so as to minimize the need for retraining, and shows that theoretical limitations exist on the amount of information that can be transported from previous learning. Robustness to changing environments depends on a delicate balance between the relations to be learned and the causal structure of the underlying model.

42. J. Pearl. 2018. Does obesity shorten life? Or is it the soda? On non-manipulable causes. *J. Causal Inference*. 6, 2 (September 2018). DOI: <https://doi.org/10.1515/jci-2018-2001>. *Non-manipulable factors, such as gender or race, have posed conceptual and practical challenges to causal analysts. The paper addresses this challenge in the context of public debates over the health cost of obesity, and offers a new perspective, based on the theory of structural causal models.*
43. J. Pearl. 2019. On the interpretation of do(x). *J. Causal Inference*. 7, 1 (March 2019), 1–6. *Provides empirically testable interpretation of the do(x) operator when applied to non-manipulable variables such as race, obesity, or cholesterol level, and ends with the conclusion that researchers need not distinguish manipulable from non-manipulable variables in their analyses.*
44. J. Pearl. 2019. Sufficient causes: On oxygen, matches, and fires. *J. Causal Inference*. 7, 2 (September 2019). *Demonstrates how counterfactuals can be used to compute the probability that one event was/is a sufficient cause of another, and how counterfactuals emerge organically from basic scientific knowledge.*



PART

HEURISTICS



Introduction by Judea Pearl

In the 1970s, heuristic search was considered the holy grail of artificial intelligence (AI), perhaps because of its universal applicability, ranging from game playing and planning, to natural language processing. Indeed, when I started teaching AI in 1971, heuristic search was the main topic of the course that, for me, was more than just a challenge of saving time and memory; it was a laboratory for combining two modes of human thought: knowledge and reasoning. The static evaluation function assigned to game positions represented knowledge, and the look-ahead procedure, followed by minimax, represented reasoning.

Typical questions that occupied researchers were: How is heuristic knowledge acquired, stored, and used by people?; how can it be represented and utilized by machines?; and what makes one heuristic succeed where others fail?

One of the main challenges to researchers at that time was the difficulty of predicting the performance of heuristic search methods. The relationship between the quality of the heuristic and the number of searches it saved was unclear. The alpha-beta pruning algorithm was by far the most efficient game searching algorithm then known, but its average performance remained an enigma, and so was its optimality.

The first two papers selected for this volume address these questions. The first uncovers an amazing convergence property of deep game trees when WIN-LOSS status is assigned randomly to terminal nodes [[Pearl 1980](#), Chapter 7]. The second uses this property to establish the average performance of alpha-beta and its optimality [[Pearl 1982](#), Chapter 8].

The third paper concerns the mechanical discovery of heuristics [[Pearl 1983](#), Chapter 9]. It follows the paradigm that heuristics are generated by solving simpli-

fied versions of the problems at hand. I am pleased to know that this method is still used in automated planning and other search-intensive applications.

References

- J. Pearl. 1980. Asymptotic properties of minimax tree and game-searching procedures. *Artificial Intelligence* 14, 2, 113–138. DOI: [https://doi.org/10.1016/0004-3702\(80\)90037-5](https://doi.org/10.1016/0004-3702(80)90037-5).
- J. Pearl. 1982. The solution for the branching factor of the alpha-beta pruning algorithm and its optimality. *Communications of the ACM* 25, 8, 559–564. DOI: <https://doi.org/10.1145/358589.358616>.
- J. Pearl. 1983. On the discovery and generation of certain heuristics. *AI Magazine* 4, 1, 23–33.

Asymptotic Properties of Minimax Trees and Game-Searching Procedures*

Judea Pearl

Abstract

The model most frequently used for evaluating the behavior of game-searching methods consists of a uniform tree of height h and a branching degree d , where the terminal positions are assigned random, independent and identically distributed values. This paper highlights some curious properties of such trees when h is very large and examines their implications on the complexity of various game-searching methods.

If the terminal positions are assigned a WIN-LOSS status with the probabilities P_0 and $1 - P_0$, respectively, then the root node is almost a sure WIN or a sure LOSS, depending on whether P_0 is higher or lower than some fixed-point probability $P^(d)$. When the terminal positions are assigned continuous real values, the minimax value of the root node converges rapidly to a unique predetermined value v^* , which is the $(1 - P^*)$ -fractile of the terminal distribution.*

Exploiting these properties we show that a game with WIN-LOSS terminals can be solved by examining, on the average, $O[(d)^{h/2}]$ terminal positions if $P_0 \neq P^$ and*

*This work was supported in part by the National Science Foundation Grants MCS 78-07468 and MCS 78-18924.

Recommended by Nils Nilsson.

Originally published in *Artificial Intelligence* **14** (1980), 113–138

Copyright 1980 by North-Holland Publishing Company. Republished with permission of Elsevier.

Original DOI: [10.1016/0004-3702\(80\)90037-5](https://doi.org/10.1016/0004-3702(80)90037-5)

$O[(P^*/(1 - P^*))^h]$ positions if $P_0 = P^*$, the former performance being optimal for all search algorithms. We further show that a game with continuous terminal values can be evaluated by examining an average of $O[(P^*/(1 - P^*))^h]$ positions, and that this is a lower bound for all directional algorithms. Games with discrete terminal values can, in almost all cases, be evaluated by examining an average of $O[(d)^{h/2}]$ terminal positions. This performance is optimal and is also achieved by the ALPHA-BETA procedure.

7.1 The Probability of Winning a Standard h -level Game Tree with Random WIN Positions

We consider a class of two-person perfect information games represented by the tree of Figure 7.1. Two players, called MAX and MIN, take alternate turns. In each turn a player may choose one out of d legal moves. The game lasts exactly n move-cycles or $h = 2n$ moves, at which point a terminal position is reached. Each terminal position constitutes either a WIN or a LOSS for the first player. We assume that the assignment of labels to the d^h terminal positions is done at random, prior to the beginning of the game, and that each terminal position may receive a WIN with probability P_0 (and a LOSS with probability $1 - P_0$) independently of other assignments. We shall refer to such a tree as a (h, d, P_0) -tree.

The first quantity we wish to compute is P_n , the probability that MAX can force a WIN given that it is his turn to move and that exactly n move-cycles are left in the game. Similarly, we denote by Q_n the probability that MAX can force a WIN given that it is MIN's turn to move and there are a total of $2n - 1$ individual moves

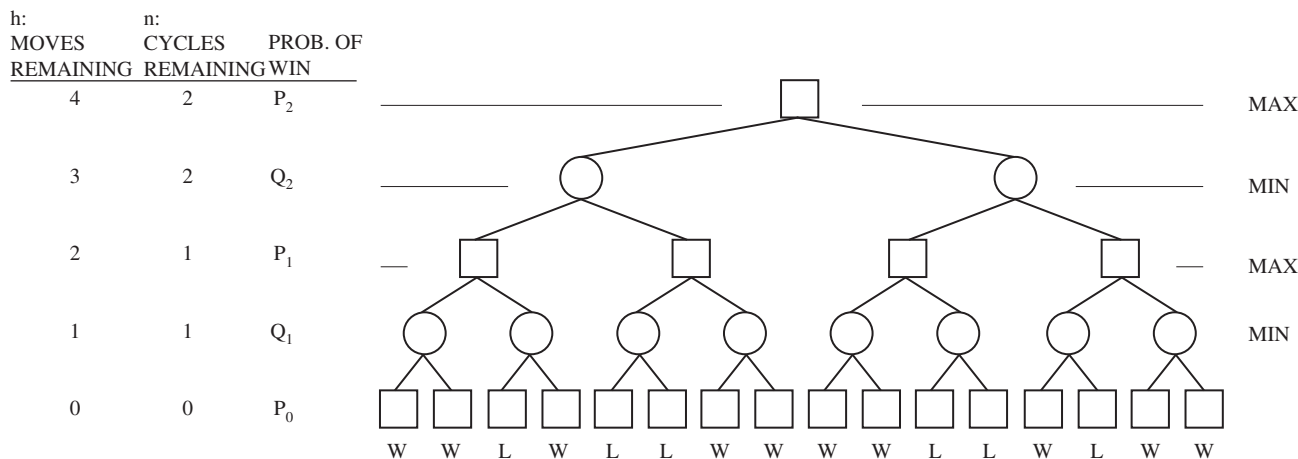


Figure 7.1 A uniform binary game tree with two move-cycles. $h = 4, n = 2, d = 2$.

left in the game. Clearly, P_n and Q_n are calculated prior to examining the terminal positions. Once the WIN-LOSS assignment is known, each node of the tree can be unequivocally labeled either a WIN or a LOSS.

For a MAX node (h even) to be a WIN, at least one of its d successors must be a WIN; therefore:

$$1 - P_n = (1 - Q_n)^d. \quad (7.1)$$

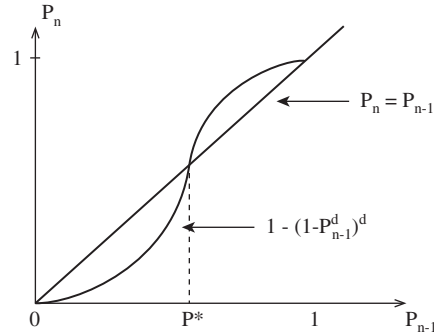
Also, for a MIN position (h odd) to be a WIN, all of its d successors must be a WIN; thus:

$$Q_n = P_{n-1}^d. \quad (7.2)$$

Combining 7.1 and 7.2 we obtain the recursive relationship:

$$P_n = 1 - (1 - P_{n-1}^d)^d. \quad (7.3)$$

The asymptotic behavior of P_n for large n can be inferred from the diagram below:



The curve $P_n = 1 - (1 - P_{n-1}^d)^d$ intersects the line $P_n = P_{n-1}$ in three points: two stable points $P_{n-1} = 0$ and $P_{n-1} = 1$, and one unstable point at $P_{n-1} = P^*$. P^* is the unique solution of the equation: $(1 - x^d)^d - (1 - x) = 0$ in the range $0 < x < 1$ or more conveniently, the positive root of the equation $x^d + x - 1 = 0$. It can be easily ascertained that every root of the latter equation is also a root of the former.

The significance of the probability P^* lies in the fact that if the terminal positions are assigned a WIN with probability $P_0 = P^*$, then, prior to examining any of these positions, MAX is assured a probability P^* of winning the game from any of his moves, regardless of the height of the tree.

Most significantly, if P_0 is slightly different than P^* , we have:

$$\lim_{n \rightarrow \infty} P_n(P_0) = \begin{cases} 1 & \text{if } P_0 > P^*, \\ 0 & \text{if } P_0 < P^*. \end{cases} \tag{7.4}$$

This means that when $P_0 > P^*$, MAX is almost assured a WIN if n is large enough, whereas he faces an almost sure LOSS in the case where $P_0 < P^*$. To illustrate this phenomenon, consider a binary game ($d = 2$) with five move-cycles ($n = 5$). P^* is the solution to $x^2 + x - 1 = 0$, or $P^* = 1/2(\sqrt{5} - 1) = 0.6180339$. If all we know about the terminal positions is that 61.80% of them are WIN's, then we also know that the first player to move has a 61.8% chance of being able to force a WIN. However, if only 50% of the terminal positions are winning, his chances to force a WIN drop to 1.95%, whereas when $P_0 = 70\%$, his chances increase to 98.5%. These numbers become much more dramatic in higher trees, as shown in Figure 7.2.

It is simple to show that the slope at the transition region is increasing exponentially with n :

$$\left. \frac{d}{dP_0}(P_n) \right|_{P_0=P^*} = \left[\frac{d(1-P^*)}{P^*} \right]^{2n} \quad \text{and} \quad \frac{d(1-P^*)}{P^*} > 1 \quad \text{for } d > 1. \tag{7.5}$$

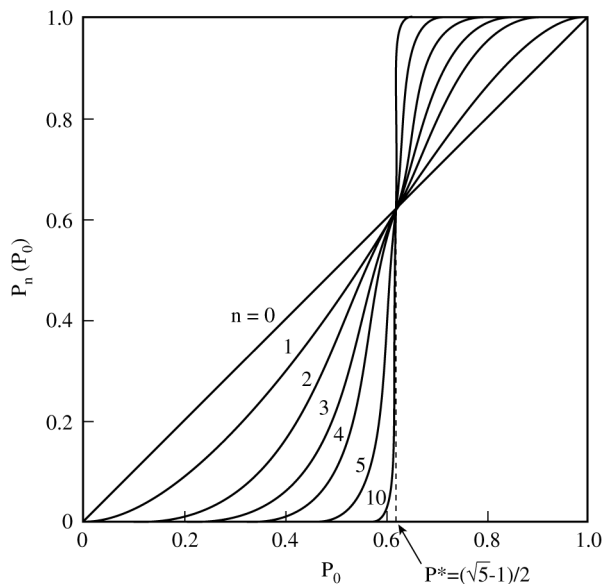


Figure 7.2 The probability of winning a n -cycle game (P_n) versus the probability of winning a terminal position (P_0), for a binary ($d = 2$) tree.

Also, a more detailed analysis shows that for sufficiently large n , P_n converges toward its asymptotic values at a super-exponential rate, i.e., for every $0 < \delta < 1$ we can find two integers n_1 and n_2 , such that:

$$\begin{aligned} P_n &\leq (\delta)^{d^{(n-n_1)}} && \text{for all } n > n_1 \text{ and } P_0 < P^*, \\ 1 - P_n &\leq (\delta)^{d^{(n-n_2)}} && \text{for all } n > n_2 \text{ and } P_0 > P^* \end{aligned} \quad (7.6)$$

where n_1 and n_2 are functions of δ and $P^* - P_0$. It is, thus, safe to conclude that for sufficiently large n , the function $P_n(P_0)$ resembles a step function with an extremely narrow transition region around P^* .

7.2 Game Trees with an Arbitrary Distribution of Terminal Values

Consider a uniform tree (constant d) where the terminal nodes are assigned numerical values, $V_0(S_1), V_0(S_2), \dots, V_0(S_{d^h})$, and assume the latter to be independent identically distributed random variables, drawn from a common distribution function $F_{V_0}(v) = P(V_0 \leq v)$. We shall refer to a tree drawn from such an ensemble as a (h, d, F) -tree and calculate the distribution of the minimax value of the root node.

Denoting the minimax values of nodes at the n th cycles by $V_n(S)$ for MAX nodes and by $U_n(S)$ for MIN nodes, we have:

$$\begin{aligned} V_n(S) &= \max[U_n(S_1), U_n(S_2), \dots, U_n(S_d)], \\ U_n(S) &= \min[V_{n-1}(S_1), V_{n-1}(S_2), \dots, V_{n-1}(S_d)] \end{aligned} \quad (7.7)$$

where S_1, S_2, \dots, S_d denote the d successors of S . The distribution of $V_n(S)$ is obtained by writing:

$$F_{V_n}(v) \triangleq P[V_n(S) \leq v] = \prod_{i=1}^d P[U_n(S_i) \leq v] = [F_{U_n}(v)]^d, \quad (7.8)$$

$$1 - F_{U_n}(v) \triangleq P[U_n(S) > v] = \prod_{i=1}^d P[V_{n-1}(S_i) > v] = [1 - F_{V_{n-1}}(v)]^d \quad (7.9)$$

yielding the recursive relation:

$$F_{V_n}(v) = \left\{ 1 - [1 - F_{V_{n-1}}(v)]^d \right\}^d. \quad (7.10)$$

Note that (7.8), (7.9), and (7.10) are identical to (7.1), (7.2), and (7.3), respectively, if one identifies $1 - F_{V_n}(v)$ with P_n and $1 - F_{U_n}(v)$ with Q_n . This is not surprising since for any fixed v , the propositions ' $V(S_i) > v$ ' propagate by the same logic as the propositions ' S_i is a WIN'; MAX nodes function as OR gates and MIN nodes perform an AND logic.

From the fact that P_n converges to a step-function as $n \rightarrow \infty$ (see (7.4)), we must conclude that $F_{V_n}(v)$, likewise, satisfies:

$$\lim_{n \rightarrow \infty} F_{V_n}(v) = \begin{cases} 0 & F_{V_0}(v) < 1 - P^*, \\ 1 - P^* & F_{V_0}(v) = 1 - P^*, \\ 1 & F_{V_0}(v) > 1 - P^*. \end{cases} \quad (7.11)$$

Assume, for the moment, that the terminal values V_0 are continuous random variables and that the distribution $F_{V_0}(v)$ is strictly increasing in the range $0 < F_{V_0} < 1$. In this case $F_{V_0}(v)$ has a unique inverse and the condition $F_{V_0}(v) = 1 - P^*$ is satisfied by one unique value of v which we call v^* :

$$v^* = F_{V_0}^{-1}(1 - P^*). \quad (7.12)$$

(7.11) then implies that when the game tree is sufficiently tall, the cumulative distribution of the root-node value approaches a step function in v , and that the transition occurs at a unique value v^* which is the $(1 - P^*)$ -fractile of the terminal distribution $F_{V_0}(\cdot)$. That implies that the density of $V_n(S)$, $f_{V_n}(v)$, becomes highly concentrated around v^* or, in other words, that the root-node value is almost certain to fall in the very close neighborhood of v^* . It appears that the repeated application of alternating MIN-MAX operations on the terminal values has the effect of filtering out their uncertainties until the result emerges at the high levels of the tree as an almost certain, predetermined, quantity.

This is a rather remarkable phenomenon which deserves to be decorated by a theorem.

Theorem 7.1 *The root value of a (h, d, F) -tree with continuous strictly increasing terminal distribution F converges, as $h \rightarrow \infty$ (in probability) to the $(1 - P^*)$ -fractile of F , where P^* is the solution of $x^d + x - 1 = 0$.*

If the terminal values are discrete: $v_1 < v_2 < \dots < v_M$, then the root value converges to a definite limit iff $1 - P^ \neq F_{V_0}(v_i)$ for all i , in which case the limit is the smallest v_i satisfying:*

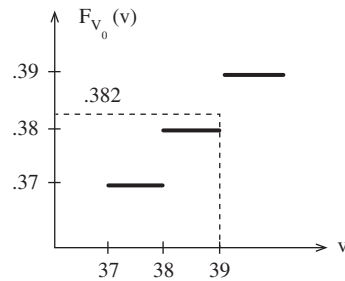
$$F_{V_0}(v_{i-1}) < 1 - P^* < F_{V_0}(v_i).$$

The second part of Theorem 7.1 becomes evident by writing:

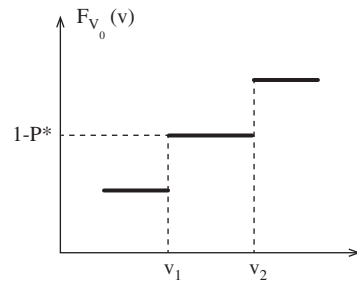
$$P[V_n(S) = v_i] = F_{V_n}(v_i) - F_{V_n}(v_{i-1}).$$

If $1 - P^*$ can be ‘sandwiched’ between two successive levels of F in such a way that $F_{V_0}(v_{i-1}) < 1 - P^* < F_{V_0}(v_i)$, then according to (7.11) $F_{V_n}(v_i) \rightarrow 1$, $F_{V_n}(v_{i-1}) \rightarrow 0$, and consequently $P[V_n(S) = v_i] \rightarrow 1$.

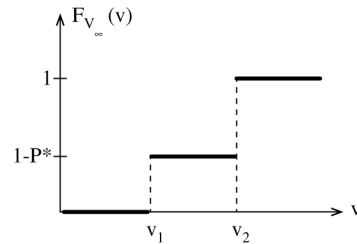
The remarkable feature of this phenomenon is that Theorem 7.1 holds for any arbitrary distribution of the terminal values. Thus, for example, the root value of a binary tree ($d = 2$) with terminal values uniformly distributed between 0 and 1 would converge to the value $1 - 1/2(\sqrt{5} - 1) = 0.382 \dots$. If the terminal values are integers, uniformly distributed between 1 and 100, then $F_{V_0}(38) < 1 - P^* = 0.382 < F_{V_0}(39)$.



Therefore, the root value will converge to the integer 39. Exceptions to the theorem would occur only in rare pathological cases where $1 - P^*$ coincides exactly with one of the plateaus of $F_{V_0}(v)$, as shown in the diagram below.



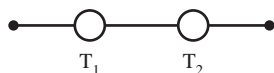
In such a case, the asymptotic distribution of the root node would go from 0 to 1 in two steps, one at v_1 and the other at v_2 , as is shown below:



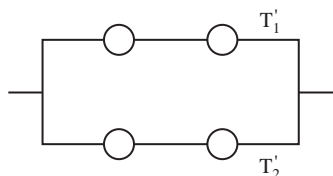
This implies that $V_n(S)$ does not converge to a single limit but may assume two possible values; in a fraction P^* of the instances it will be assigned the value v_2 and

in the remaining instances the value v_1 . In fact, Section 7.1 dealt with such a case where, if $P_0 = P^*$, the status of the root-node remains undetermined between WIN ($V_n = 1$) and LOSS ($V_n = 0$).

For the reader's amusement, another manifestation of Theorem 7.1 will be mentioned, unrelated to game trees. Consider a large collection of unreliable electrical components (e.g., light bulbs) whose times to failure are identically distributed random variables. Connect two of them in series:



The failure time of this series connection is given by $\min[T_1, T_2]$. Now connect two such circuits in parallel:



The failure time of the parallel circuit is equal to that of the longest surviving branch, i.e., $\max(T_1', T_2')$. Continue the process, alternately connecting duplicate circuits in series and in parallel, for n cycles. What can be said about the limiting distribution of the failure time, T_n , of the entire circuit? Clearly, T_n is equal to the minimax value of the root-node in an n -cycle binary tree with terminal values determined by the failure times of the individual components. According to Theorem 7.1, T_n converges to a definite value given by the $(1 - P^*)$ -fractile of the terminal distribution. Thus, assuming that n is sufficiently large, the entire circuit should fail at a predictable, precise time, which is quite remarkable considering the fact that the circuit is assembled from a host of independent, unreliable, and unpredictable components.

At this point a natural question to ask is how fast the density distribution of the root value contracts to its final value v^* . The answer is that the width of the density function decreases exponentially with n . The range of values W_ϵ (around v^*) which contains all but 2ϵ of the total area under the density function can be shown (for $d = 2$) to be proportional to $(\log 1/2\epsilon)^{0.584} 2^{-n}$.

This finding raises some interesting questions regarding the advisability of searching deep uniform game trees. If the final values of these trees can be predetermined with virtual certainty, why spend the exponentially growing effort demanded by an exact evaluation? Instead of insisting on selecting the best first

move, we might as well select just any move at random. The expected loss of opportunity induced by such selection is guaranteed not to exceed some predetermined limit which diminishes exponentially with the height of the remaining tree. It makes more sense to reserve one's computational powers for the end-game where the shallowness of the trees is accompanied by more widely varying node values.

These arguments touch on the more general question of how the willingness to act somewhat suboptimally can be converted into computational savings, a question which we hope to study more fully in future studies. At this point, it suffices to state that the uniform tree model with independent and identically distributed terminal values was not devised as a practical game playing tool but rather as a test bed for comparing the efficiencies of various exact-search methods. We shall pursue this plan in the remaining part of this report.

7.3 The Mean Complexity of Solving (h, d, P_0) -game

Solving a game tree means deciding whether the root-node is a WIN or a LOSS. An absolute lower bound on the number of terminal node examinations needed for establishing the status of the root-node is given by the following argument. If the root node is a WIN, then there exists a subtree (called a solution tree) consisting of one branch emanating from each MAX node and all branches emanating from each MIN node, terminating at a set of WIN terminal positions. Similarly, if the game is a LOSS, such a solution tree exists for the opponent, terminating at all LOSS nodes. In either case, the number of terminal positions in a solution tree is d^n (representing a branching factor d in each move-cycle) or $d^{h/2}$ where h is the number of individual moves. The number of terminal node examinations required to solve the game must exceed $d^{h/2}$ since, regardless of how the solution tree was discovered, one must still ascertain that all its $d^{h/2}$ terminal nodes are WIN in order to defend the proposition 'root is a WIN'. Thus, $d^{h/2}$ represents the number of terminal nodes inspected by a non-deterministic algorithm which solves the (h, d, P_0) -game and is, therefore, a lower bound for all deterministic algorithms.

It is not hard to show that any algorithm solving the (h, d, P_0) -game would, in the worst case, inspect all d^h terminal positions. This can be done by cleverly arranging the terminal assignments in such a way that a decision could not be reached until the last node is inspected. Since the difference between $d^{h/2}$ and d^h may be quite substantial, it is interesting to evaluate the expected number of terminal examinations where the expectation is taken with respect to all possible WIN-LOSS assignments to the terminal nodes.

Definition Let A be a deterministic algorithm which solves the (h, d, P_0) -game and let $I_A(h, d, P_0)$ denote the expected number of terminal positions examined by A .

The quantity:

$$r_A(d, P_0) = \lim_{h \rightarrow \infty} [I_A(h, d, P_0)]^{1/h}$$

is called the *branching factor* corresponding to the algorithm A .

Definition Let C be a class of algorithms capable of solving a general (h, d, P_0) -tree. An algorithm A is said to be *asymptotically optimal over C* if for some P_0 and all d :

$$r_A(d, P_0) \leq r_B(d, P_0) \quad \forall B \in C.$$

Definition An algorithm A is said to be *directional* if for some linear arrangement of the terminal nodes it never selects for examination a node situated to the left of a previously examined node.

Simply stated, an algorithm is directional if it always examines nodes from left to right, regardless of the content of the nodes examined.

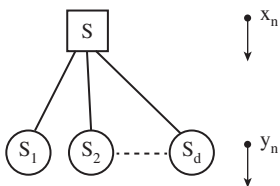
We now compute the branching factor of a simple directional algorithm, called SOLVE, given by the following informal description¹:

Algorithm SOLVE(S): To solve S , start solving its successors from left to right.

If S is MAX, return a WIN as soon as one successor is found to be a WIN; return a LOSS if all successors of S are found to be a LOSS.

If S is MIN, return a LOSS as soon as one successor is found to be a LOSS; return a WIN if all successors of S are found to be a WIN.

To compute $I_{\text{SOLVE}}(h, d, P_0)$ we consider the n th cycle preceding the terminal positions. Let x_n stand for the expected number of terminal nodes inspected in solving the root S of an n -cycle tree, and y_n the expected number of inspections used for solving any of the successors of S .



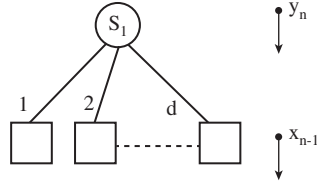
The probability of issuing a WIN after solving the k th successor is $(1 - Q_n)^{k-1} Q_n$. Such an event requires an average of $(k - 1)y_n^- + y_n^+$ terminal inspections, where y_n^- and y_n^+ stand for the mean number of inspections required for establishing a LOSS

1. A more formal definition is given by the flow-chart of Figure 7.5, with the few simple modifications discussed at the head of Section 7.4.

or a WIN, respectively. Also, the event of issuing a LOSS for S carries a probability $(1 - Q_n)^d$ and a mean expenditure of dy_n^- inspections. Therefore:

$$\begin{aligned}
 x_n &= \sum_{k=1}^d Q_n(1 - Q_n)^{k-1}[(k - 1)y_n^- + y_n^+] + d(1 - Q_n)^d y_n^- \\
 &= y_n^+ Q_n \frac{1 - (1 - Q_n)^d}{Q_n} + y_n^- (1 - Q_n) \frac{1 - (1 - Q_n)^d}{Q_n} \\
 &= [y_n^+ Q_n + y_n^- (1 - Q_n)] \frac{[1 - (1 - Q_n)^d]}{Q_n} \quad (\text{using (7.1) and (7.2)}) \\
 &= y_n \frac{P_n}{P_{n-1}^d}. \tag{7.13}
 \end{aligned}$$

Now examine the solution of any successor of S , say S_1 .



The event of issuing a LOSS after solving its k th successor has a probability $P_{n-1}^{k-1}(1 - P_{n-1})$ and carries a mean expenditure of $(k - 1)x_{n-1}^+ + x_{n-1}^-$ inspections. The event of exiting with a WIN involves solving all d successors and, therefore, occurs with probability P_{n-1}^d and costs an average of dx_{n-1}^+ inspections. Thus:

$$\begin{aligned}
 y_n &= \sum_{k=1}^d P_{n-1}^{k-1}(1 - P_{n-1})[(k - 1)x_{n-1}^- + x_{n-1}^-] + dP_{n-1}^d x_{n-1}^+ \\
 &= [x_{n-1}^+ P_{n-1} + x_{n-1}^- (1 - P_{n-1})] \frac{(1 - P_{n-1}^d)}{1 - P_{n-1}} \\
 &= x_{n-1} \frac{1 - P_{n-1}^d}{1 - P_{n-1}}. \tag{7.14}
 \end{aligned}$$

Combining (7.13) and (7.14) we obtain:

$$\frac{x_n}{x_{n-1}} = \frac{1 - Q_n}{Q_n} \cdot \frac{P_n}{1 - P_{n-1}} = \frac{P_n \cdot (1 - P_{n-1}^d)}{P_{n-1}^d \cdot (1 - P_{n-1})}. \tag{7.15}$$

Since x_n is equivalent to $I_{\text{SOLVE}}(2n, d, P_0)$ and $x_0 = 1$, we can state:

Theorem 7.2 The expected number of terminal position in a (h, d, P_0) -tree examined by the SOLVE algorithm is given by:

$$I_{\text{SOLVE}}(h, d, P_0) = \prod_{i=1}^{h/2} \frac{P_i(1 - P_{i-1}^d)}{P_{i-1}^d(1 - P_{i-1})} \tag{7.16}$$

where $P_i, i = 1, \dots, \frac{1}{2}h$, is related to P_0 by (7.3).

Theorem 7.2 permits an easy calculation of $I_{\text{SOLVE}}(h, d, P_0)$ for wide ranges of d and h , as shown in Figure 7.3. In the special case where $P_0 = P^*$ all terms in the

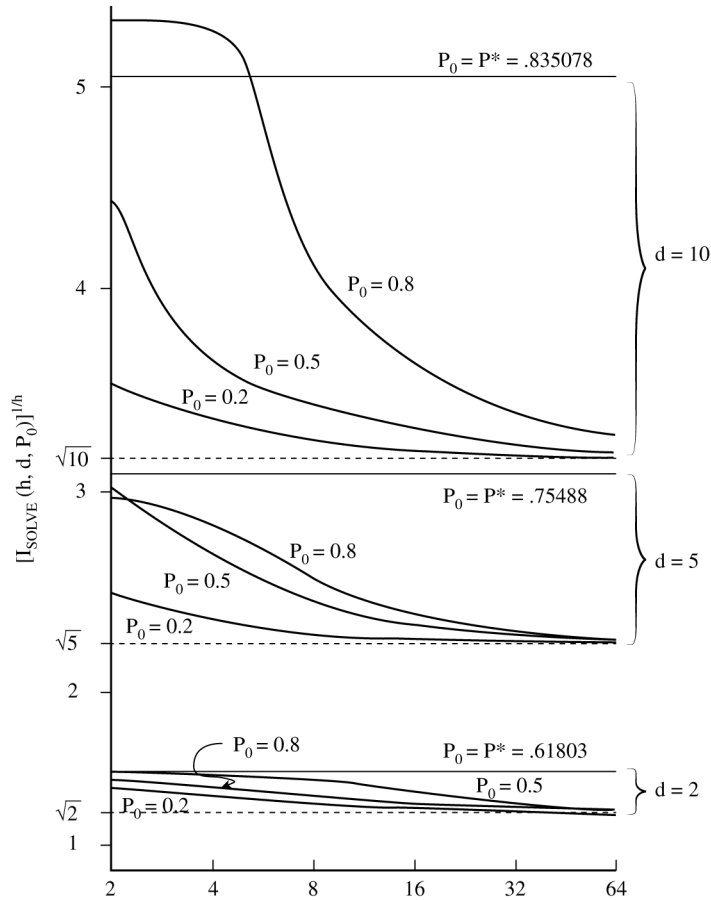


Figure 7.3 The expected number of terminal nodes examined by SOLVE (normalized by $(\cdot)^{1/h}$ to represent an effective branching factor).

product of (7.16) are equal and, using $P^{*d} = 1 - P^*$, (7.16) reduces to:

$$I_{\text{SOLVE}}(h, d, P^*) = \left(\frac{P^*}{1 - P^*} \right)^h. \quad (7.17)$$

Note that

$$\lim_{P_{n-1} \rightarrow 0} \frac{x_n}{x_{n-1}} = \lim_{P_{n-1} \rightarrow 1} \frac{x_n}{x_{n-1}} = d$$

which implies:

$$\lim_{n \rightarrow \infty} \frac{x_n}{x_{n-1}} = \begin{cases} d & P_0 \neq P^*, \\ \left(\frac{P^*}{1 - P^*} \right)^2 & P_0 = P^*. \end{cases} \quad (7.18)$$

This limit, combined with the very rapid convergence of P_n (see (7.6)), leads directly to the asymptotic branching factor of SOLVE:

Corollary 7.1 *The branching factor of the SOLVE algorithm is given by:*

$$r_{\text{SOLVE}}(d, P_0) = \begin{cases} d^{1/2} & P_0 \neq P^*, \\ \frac{P^*}{1 - P^*} & P_0 = P^* \end{cases} \quad (7.19)$$

where P^* is the positive solution of $x^d + x - 1 = 0$.

Recalling that $d^{1/2}$ is an absolute lower bound for the branching factor of any tree solving algorithm, we conclude:

Corollary 7.2 *SOLVE is asymptotically optimal for $P_0 \neq P^*$.*

For finite values of h or for $P_0 = P^*$ we have no guarantee that SOLVE is optimal. Non-directional algorithms, such as that proposed by Stockman [7] may outperform SOLVE. However, Corollary 7.2 states that for very deep trees the savings could not be substantial in all cases where $P_0 \neq P^*$.

Any directional algorithm which is governed by a successor-ordering scheme identical to that of SOLVE must examine all the nodes examined by SOLVE. This is so because if some left-to-right algorithm B skips a node visited by SOLVE, a WIN-LOSS assignment can be found which would render the conclusion of SOLVE contrary to that of B . Thus B could not be a general algorithm for solving all (h, d, P_0) -trees. Now, since $I_{\text{SOLVE}}(h, d, P_0)$ is independent on the particular choice of ordering scheme, we may conclude that SOLVE is optimal over the class of directional game-solving algorithms. This leads to:

Corollary 7.3 *The optimal branching factor of any directional algorithm which solves a general (h, d, P_0) -tree is given by (7.19).*

The case $P_0 = P^*$ deserves a special attention. Although it is not very likely to occur in practical WIN-LOSS games, it plays an important role in the analysis of the $\alpha - \beta$ procedure. We conclude this section by examining the behavior of $r_{\text{SOLVE}}(d, P^*) = P^*/(1 - P^*)$ for large values of d . Writing:

$$q(d) = 1 - P^*(d) \quad (7.20)$$

the defining equation for $q(d)$ becomes:

$$q(d) = [1 - q(d)]^d \quad (7.21)$$

which can be satisfied only when:

$$\lim_{d \rightarrow \infty} q(d) = 0. \quad (7.22)$$

Taking log on both sides of (7.21), gives:

$$\log q(d) = d \log[1 - q(d)] = -d[q(d) + O(q^2)] \quad (7.23)$$

or:

$$q(d) = (1/d) \log 1/q(d) \quad (7.24)$$

By repeated iteration, the solution of (7.24) can be written:

$$q(d) = 1/d[\log d - \log \log d + \log \log \log d - \dots] \quad (7.25)$$

from which we see that for large d :

$$q(d) = \frac{\log d}{d} + O\left(\frac{\log \log d}{d}\right). \quad (7.26)$$

This result was also shown by Baudet [1] using a slightly different method. Substituting (7.26) in (7.19), the asymptotic behavior of $r_{\text{SOLVE}}(d, P^*)$ becomes:

$$r_{\text{SOLVE}}(d, P^*) = \frac{d}{\log d} \left[1 + O\left(\frac{\log \log d}{\log d}\right) \right]. \quad (7.27)$$

the log-log graph of Figure 7.4 depicts $r_{\text{SOLVE}}(d, P^*)$ for the range $2 \leq d \leq 10,000$. It is shown to be in remarkable agreement with the formula $(0.925)d^{0.74741}$, while

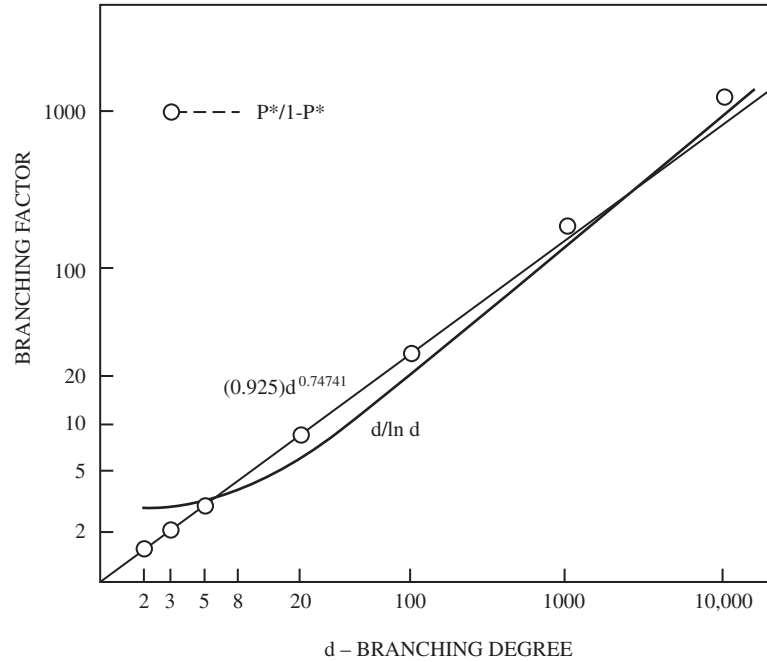


Figure 7.4 Worst case branching-factor for the SOLVE algorithm $\left[r_{\text{SOLVE}}(d, P^*) = \frac{P^*(d)}{1-P^*(d)} \right]$.

the asymptotic expression $d/\log d$ becomes a better approximation only for $d > 5000$.

7.4 Solving, Testing, and Evaluating Game Trees

When the terminal positions are assigned real values, the root-node must be *evaluated* rather than *solved*. The SOLVE algorithm discussed in Section 7.3 is insufficient to fully evaluate a (h, d, F_{V_0}) -game tree because it produces a binary WIN-LOSS outcome rather than the (real) minimax value $V(S)$ of the rootnode. It can, however, be used to test the proposition ' $V(S) > v$ ', where v is any fixed reference value chosen for the test. We simply interpret any terminal node t for which $V_0(t) > v$ as a WIN position (otherwise it is a LOSS), and apply SOLVE directly. If it issues a WIN, the proposition ' $V(S) > v$ ' is proven, otherwise we deduce ' $V(S) \leq v$ '. This procedure, which we call $\text{TEST}(S, v, >)$, is described in algorithmic details in Figure 7.5. An almost identical algorithm, $\text{TEST}(S, v, \geq)$, could be used to test whether $V(S) \geq v$ by simply permitting equality in all the comparison tests of Figure 7.5.

From the structural identity of SOLVE and TEST, it is clear that the expected number of nodes inspected by TEST, $I_{\text{TEST}}(h, d, F_{V_0}, v)$ is equal to that inspected by SOLVE if the terminal WIN labels are assigned with probability

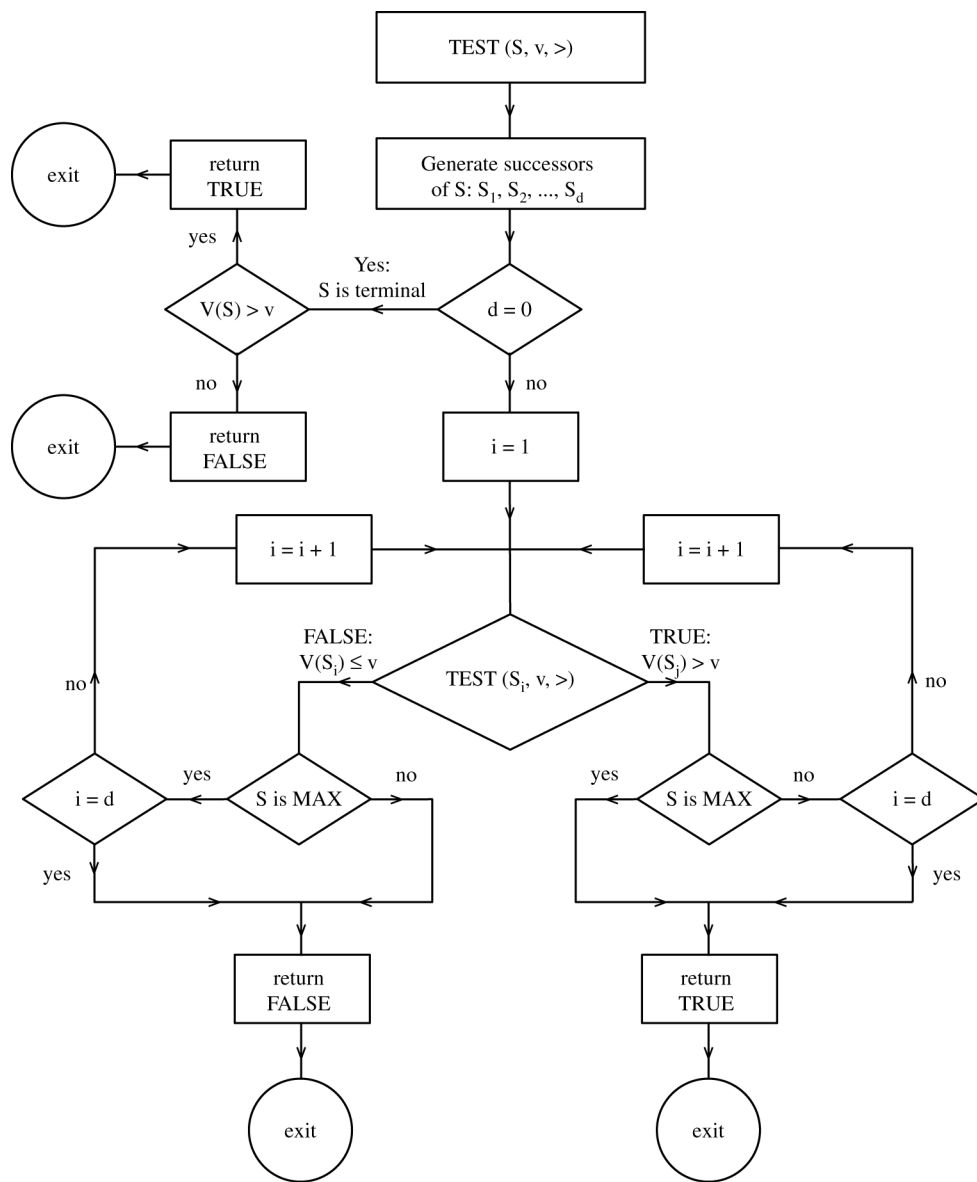


Figure 7.5 A flow-chart of the $\text{TEST}(S, v, >)$ procedure which tests whether the minimax value of position S exceeds a reference v .

$P_0 = P[V_0(t) > v] = 1 - F_{V_0}(v)$. Therefore:

$$I_{\text{TEST}}(h, d, F_{V_0}, v) = I_{\text{SOLVE}}(h, d, 1 - F_{V_0}(v)). \tag{7.28}$$

(7.28), combined with (7.19), yields:

Theorem 7.3 *The expected number of terminal positions examined by the TEST algorithm in testing the proposition ‘ $V(S) > v$ ’ for the root of a (h, d, F_{V_0}) -tree, has a branching-factor:*

$$r_{\text{TEST}}(d, F_{V_0}, v) = \begin{cases} d^{1/2} & \text{if } v \neq v^*, \\ \frac{P^*}{1 - P^*} & \text{if } v = v^* \end{cases} \quad (7.29)$$

where v^* satisfies $F_{V_0}(v^*) = 1 - P^*$

From the fact that TEST is directional and SOLVE is optimal we can also conclude:

Corollary 7.4 *The optimal branching factor of any directional algorithm which tests whether the root node of a (h, d, F_{V_0}) -tree exceeds a specified reference v is given by $r_{\text{TEST}}(d, F_{V_0}, v)$ in (7.29).*

Note that when the terminal values are continuous (and $F_{V_0}(v)$ strictly increasing) Theorem 7.1 states that $V(S)$ converges to v^* for very large h . Thus, although testing the proposition ‘ $V(S) > v$ ’ is easier for $v \neq v^*$, the outcomes of such tests are almost trivial. The most informative test is that which verifies whether $V(S) > v^*$, and such a test, according to (7.29) is indeed the hardest.

When the terminal positions are assigned discrete values then unless $1 - P^*$ coincides with one of the plateaus of F_{V_0} , the equation $F_{V_0}(v^*) = 1 - P^*$ would not have a solution, and the limiting root value would converge to the smallest v' satisfying $F_{V_0}(v') > 1 - P^*$. Thus all inequality propositions could be tested with a branching factor $d^{1/2}$.

Consider now the minimum number of terminal node examinations required to *evaluate* a game tree. At the best possible case, even if someone hands us for free the true value of S , any evaluation algorithm should be able to defend the proposition ‘ $V(S) = v$ ’ i.e., to defend the pair of propositions ‘ $V(S) \geq v$ ’ and ‘ $V(S) \leq v$ ’. Since the solution tree required for the verification of an inequality proposition contains $d^{h/2}$ terminal positions and since the sets of terminal positions participating in the defense of each of these inequalities are mutually exclusive, save for the one position satisfying $V_0(t) = V(S)$, we have:

Corollary 7.5 *Any procedure which evaluates a (h, d, F_{V_0}) -tree must examine at least $2d^{h/2} - 1$ terminal nodes.*

We assumed, of course, that the probability of two or more terminal nodes satisfying $V_0(t) = V(S)$ is zero, and that h is even. This result (in a slightly different form) was also proven by Knuth and Moore [3]. Earlier, Slagle and Dixon [6] proved that the $\alpha - \beta$ procedure achieves this optimistic bound if the successor positions are perfectly ordered.

Let us consider now the more interesting question of estimating $I(h, d, F_{V_0})$, the *expected* number of terminal examinations required for evaluating (h, d, F_{V_0}) -game trees. Let $I_D(h, d)$ be the minimal value of $I(h, d, F_{V_0})$ achieved by any directional algorithm under the worst-case F_{V_0} .

$$I_D(h, d) \triangleq \min_{\substack{A \\ \text{directional}}} \max_F I_A(h, d, F). \quad (7.30)$$

Every algorithm which evaluates a game tree must examine at least as many nodes as that required for testing whether the root value is greater than some reference v . This is so because an evaluation procedure produces a more informative outcome than any inequality test, and moreover, one can always use the value $V(S)$ to deduce all inequality propositions regarding $V(S)$. This fact combined with the optimality of TEST over the class of directional algorithms (see Corollary 7.4) leads to:

$$I_D(h, d) \geq \left(\frac{P^*}{1 - P^*} \right)^h. \quad (7.31)$$

The right-hand side of (7.31) is obtained when the terminal positions are assigned continuous values and TEST is given the task of verifying ' $V(S) > v^*$ '. This leads directly to:

Theorem 7.4 *The expected number of terminal positions examined by any directional algorithm which evaluates a (h, d) -game tree with continuous terminal values must have a branching factor greater or equal to $P^*/(1 - P^*)$.*

The quantity $P^*/(1 - P^*)$ was shown by Baudet [1] to be a lower bound for the branching factor of the $\alpha - \beta$ procedure. Theorem 7.4 extends the bound to all directional game-evaluating algorithms.

In the next section we will present a straightforward evaluation algorithm called SCOUT which actually achieves the branching factor $P^*/(1 - P^*)$, thus establishing the asymptotic optimality of SCOUT over the class of directional algorithms, including the $\alpha - \beta$ procedure.

7.5 Test and, if Necessary, Evaluate—The SCOUT Algorithm

SCOUT evaluates a MAX position S by first evaluating its left most successor S_1 , then 'scouting' the remaining successors, from left to right, to determine if any meets the condition $V(S_k) > V(S_1)$. If the inequality is found to hold for S_k , this node is then evaluated exactly and its value $V(S_k)$ is used for subsequent 'scoutings' tests. Otherwise S_k is exempted from evaluation and S_{k+1} selected for a test. When all successors have been either evaluated or tested and found unworthy of evaluation, the last value obtained is issued as $V(S)$. An identical procedure is used

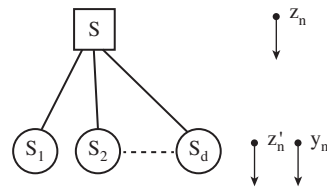
for evaluating a MIN position S , save for the fact that the event $V(S_k) \geq V(S_1)$ now constitutes grounds for exempting S_k from evaluation. The flow-chart of Figure 7.6 describes SCOUT in algorithmic details, calling on the TEST algorithm of Figure 7.5 to perform the inequality checks.

At first glance it appears that SCOUT is very wasteful; any node S_k which is found to fail a test criterion is submitted back for evaluation. The terminal nodes inspected during such a test may (and in fact will) be revisited during the evaluation phase. An exact mathematical analysis, however, reveals that the amount of waste is not substantial and that SCOUT, in spite of some duplicated effort, still achieves the optimal branching factor $P^*/(1 - P^*)$.

Two factors work in favor of SCOUT: (1) Most tests would result in exempting the tested node (and all its descendents) from any further evaluation, and (2) testing for inequality using the TEST(S, v) procedure is relatively speedy. In the worst possible case TEST only consumes an average of $(P^*/(1 - P^*))^h$ inspections which according to (7.31) is below the average consumption of the best directional evaluation procedure. The superiority of TEST stems from the fact that it induces many cutoffs not necessarily permitted by EVAL or any other evaluation scheme. As soon as *one* successor of a MAX node meets the criterion $V(S_k) > v$, all other successors can be ignored. EVAL, by contrast, would necessitate a further examination of the remaining successors to determine if any would possess a value higher than $V(S_k)$.

7.6 Analysis of SCOUT's Expected Performance

Let S be a MAX node rooting an n -cycle tree ($h = 2n$) with a uniform branching degree d . Let z_n denote the expected number of terminal examinations undertaken by SCOUT. These examinations consist of those performed during the EVAL(S_k) phases ($k = 1, \dots, d$) plus those performed during the TEST($S_k, v, >$) phases ($k = 2, \dots, d$). Since the subtrees emanating from the successors of S all have identically distributed terminal values, the number of positions examined in each EVAL(S_k), phase have identical expectations, called z'_n . Let v_k be the test criterion during the TEST($S_k, v, >$) phase, and let $y_n^+(k)$ and $y_n^-(k)$ have the same interpretations as in



Section 7.3. The event that S_k is found to satisfy the criterion $V(S_k) > v_k$ would consume a mean expenditure of $y_n^+(k) + z'_n$ inspections while a successor found to refute

this test would consume, on the average, only $y_n^-(k)$ inspections. Thus, if q_k stands for the probability that successor S_k would require an evaluation, we have:

$$\begin{aligned} z_n &= z'_n + \sum_{k=2}^d q_k(z'_n + y_n^+(k)) + \sum_{k=2}^d (1 - q_k)y_n^-(k) \\ &= z'_n[1 + \sum_{k=2}^d q_k] + \sum_{k=2}^d y_n(k). \end{aligned} \quad (7.32)$$

Since S_k would require an evaluation iff $V(S_k) > \max[V(S_1), V(S_2), \dots, V(S_{k-1})]$ and all the node-values at any given level are independent, identically distributed and continuous random variables, we have:

$$q_k = \frac{1}{k}, \quad k = 2, \dots, d. \quad (7.33)$$

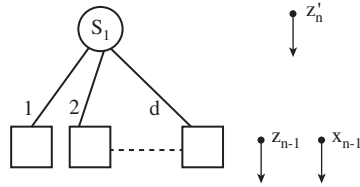
Moreover, since we are interested in a worst case analysis, each $y_n(k)$ can be replaced by its highest possible value. This occurs when the probability that any given terminal position t satisfies $V(t) > v_k$ is equal to the fixed point probability P^* . From (7.13) and (7.14) $y_n(k)$, in such a case, would be given by $(P^*/(1 - P^*))^{2n-1}$ and we can write:

$$z_n = z'_n \zeta(d) + \left(\frac{P^*}{1 - P^*} \right)^{2n-1} (d - 1) \quad (7.34)$$

where

$$\zeta(d) = \sum_{k=1}^d \frac{1}{k}. \quad (7.35)$$

Note that this approximation is not too pessimistic in light of the fact that for large n the values of all nodes converge rapidly toward the limiting value v^* and, therefore, most tests would employ a threshold level v_k from the neighborhood of v^*



To compute the solution of (7.34) we now examine the expected number of inspections z'_n employed while evaluating any of the successors of S , say S_1 . Since S_1 is a MIN position a successor would be submitted for evaluation iff its value is

proven to be *below* the threshold level propagating from the left. Each evaluation would require an average of z_{n-1} inspections and each test would consume at most an average of $[x_{n-1} \approx (P^*/(1-P^*))^{2n-2}]$ terminal inspections. Consequently, using an argument similar to the one above, we obtain:

$$z'_n = z_{n-1}\zeta(d) + \left(\frac{P^*}{1-P^*}\right)^{2n-2} (d-1) \quad (7.36)$$

which, combining (7.34) and (7.36), yields:

$$z_n = z_{n-1}\zeta^2(d) + (d-1) \left(\frac{P^*}{1-P^*}\right)^{2n-2} \left[\zeta(d) + \left(\frac{P^*}{1-P^*}\right)\right]. \quad (7.37)$$

(7.37) is a linear difference equation of the form:

$$z_n = \alpha z_{n-1} + K\beta^n \quad (7.38)$$

with

$$\begin{aligned} z_0 &= 1, \\ K &= (d-1) \left(\frac{1-P^*}{P^*}\right)^2 \left[\zeta(d) + \left(\frac{P^*}{1-P^*}\right)\right], \\ \beta &= \left(\frac{P^*}{1-P^*}\right)^2, \\ \alpha &= \zeta^2(d). \end{aligned} \quad (7.39)$$

Its solution is given by:

$$z_n = \alpha^n + K\beta \frac{\beta^n - \alpha^n}{\beta - \alpha}. \quad (7.40)$$

Clearly, it is the relative size of α and β which governs the asymptotic behavior of z_n for large values of n . However, since for all d we have:

$$\frac{P^*(d)}{1-P^*(d)} > \zeta(d)$$

(e.g., for $d \rightarrow \infty$, $P^*/(1-P^*) = O(d/\log d)$ while $\zeta(d) = O(\log d)$) β would become the dominant factor, and we can write:

$$z_n \sim \frac{K}{\beta - \alpha} \beta^{n+1} \quad (7.41)$$

or equivalently (with $h = 2n$):

$$I_{\text{SCOUT}}(h, d, F) \sim \frac{(d-1)}{\frac{P^*}{1-P^*} - \zeta(d)} \left(\frac{P^*}{1-P^*} \right)^h. \quad (7.42)$$

Theorem 7.5 *The expected number of terminal examinations performed by SCOUT in the evaluation of (h, d) -game trees with continuous terminal values has a branching factor:*

$$r_{\text{SCOUT}} = \frac{P^*}{1-P^*}. \quad (7.43)$$

So far, our analysis was based on the assumption that the terminal nodes may be assigned continuous values. We will now demonstrate that I_{SCOUT} is substantially reduced if the terminal nodes are assigned only discrete values.

Let's ignore the rare case where $1 - P^*$ coincides exactly with one of the plateaus of F . When coincidence does not occur, we showed in Section 7.2 that the values of all nodes at sufficiently high levels converge to the same limit, given by the lowest terminal value v' satisfying $F_{V_0}(v') > 1 - P^*$. This convergence has two effects on the complexity of SCOUT as analyzed in (7.32): first, q_k is no longer equal to $1/k$ but rather, converges to zero at high n for all $k > 1$. The reason for this is that in order for $V(S_k)$ to be greater than $V(S_1)$ (which is most probably equal to v') it must exceed $V(S_1)$ by a finite positive quantity and, at a very high h , finite differences between any two nodes are extremely rare. Second, the threshold levels v_k against which the $\text{TEST}(S_k, v, >)$ procedures are performed are no longer close to v^* but differ from it by finite amounts. Under such conditions the proposition ' $V(S_k) > v_k$ ' can be tested more efficiently since $r_{\text{TEST}} = d^{1/2}$ (see (7.29)).

Applying these considerations to the analysis of z_n in (7.32) gives:

$$r_{\text{SCOUT}} \sim d^{1/2} \quad (7.44)$$

and we obtain:

Theorem 7.6 *The expected number of terminal positions examined by the SCOUT procedures in evaluating a (h, d, F_{V_0}) -game with discrete terminal values has a branching factor:*

$$r_{\text{SCOUT}} = d^{1/2} \quad (7.45)$$

with exceptions only when one of the discrete values, v^ , satisfies $F_{V_0}(v^*) = 1 - P^*$.*

Corollary 7.6 *For games with discrete terminal values satisfying the conditions of Theorem 7.6, the SCOUT procedure is asymptotically optimal over all evaluation algorithms.*

Of course, the transition from $r_{\text{SOLVE}} = P^*/(1 - P^*)$ in the continuous case to $r_{\text{SOLVE}} = d^{1/2}$ in the discrete case does not occur abruptly. When the quantization levels are very close to each other it takes many more levels before SCOUT begins to acquire the lower branching factor of $d^{1/2}$. In fact, using the discussion of Section 7.1, it is possible to compute at what height SCOUT begins to act more efficiently. For example, if the terminal values are integers, uniformly distributed from 1 to M , we know that at very high levels of the tree the values of all nodes will converge to I^* , where I^* is the lowest integer satisfying $F_{V_0}(I^*) > 1 - P^*$. The probability that a node n cycles away from the bottom would acquire this value is:

$$P[V_n(S) = I^*] = F_{V_n}(I^*) - F_{V_n}(I^* - 1). \quad (7.46)$$

If $F_{V_n}(I^*)$ and $F_{V_n}(I^* - 1)$ are very close to each other, $P[V_n(S) = I^*]$ will be governed by the linear regions of the curves in Figure 7.2. Therefore, we can write:

$$P[V_n(S) = I^*] \approx \frac{dF_{V_n}(v)}{dF_{V_0}(v)} \Big|_{F_{V_0}=1-P^*} [F_{V_0}(I^*) - F_{V_0}(I^* - 1)]$$

and, using (7.5) and (7.10):

$$P[V_n(S) = I^*] \cong \left[\frac{d(1 - P^*)}{P^*} \right]^{2n} 1/M. \quad (7.47)$$

In order for the TEST procedures nested in SCOUT to achieve a branching factor of $d^{1/2}$ the parameter P_{n-1} appearing in (7.15) must be sufficiently close to zero. But this is achieved when $P[V_n(S) = I^*]$ approaches unity, i.e., when h satisfies:

$$h \geq \frac{\log M}{\log \frac{d(1-P^*)}{P^*}} \triangleq h_0(M, d). \quad (7.48)$$

Thus, above the critical level $h_0(M, d)$ it becomes fairly sure that every node has a minimax value I^* and, consequently, the SCOUT procedure would have to expand only $d^{1/2}$ nodes per level. Note that the critical height increases logarithmically with the number of quantization levels M .

Several improvements could be applied to the SCOUT algorithm to render it more efficient. For example, when a TEST procedure issues a non-exempt verdict, it could also return a new reference value and some information regarding how the decision was obtained in order to minimize the number of nodes to be inspected by EVAL. The main reasons for introducing SCOUT have been its conceptual and analytic simplicity and the fact that it possesses the lowest branching

factor of any algorithm known to date. However, the potential of SCOUT as a practical game-searching procedure should not be dismissed altogether. Recent simulation studies using the game of Kalah show² that the efficiency of SCOUT, even in its unpolished version, compares favorably with that of the α - β procedure.

7.7 On the Branching Factor of the ALPHA-BETA (α - β) Procedure

The reader is assumed to be familiar with the basic features of the ALPHA-BETA (α - β) pruning method. Descriptions of the method can be found in the textbooks by Nilsson [4, Section 4] and Slagle [5, pp. 16–24]. A historical survey of the development of the concept is given by Knuth and Moore [3, Section 5].

The fact that the number of terminal nodes examined by α - β may vary from $(d^{\lfloor h/2 \rfloor} + d^{\lceil h/2 \rceil} - 1)$ to d^h was shown by Slagle and Dixon [6] and elaborated by Knuth and Moore [3].

The analysis of expected performance using uniform trees with random terminal values has begun with Fuller *et al.* [2] who obtained formulas by which the average number of terminal examinations can be computed. Unfortunately, the formulas are very complicated and would not facilitate an asymptotic analysis. Simulation studies conducted by Fuller *et al.* led to the estimate:

$$r_{\alpha-\beta} \approx d^{0.72}.$$

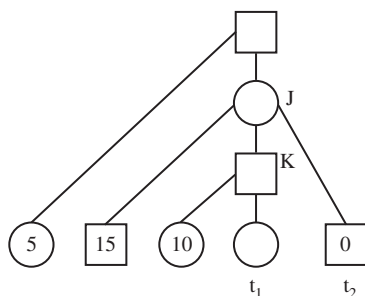
Knuth and Moore [3] have analyzed a less powerful but simpler version of the α - β procedure by ignoring deep cutoffs. They have shown that the branching factor of this simplified model is $O(d/(\log d))$ and speculated that the inclusion of deep cutoffs would not alter this behavior substantially. However, the gap between the upper and lower bounds for the branching factor remained appreciable, even for the simplified model.

A more recent paper by Baudet [1] contains several improvements. Starting by considering possible equalities between terminal values, Baudet derived a general formula for $I_{\alpha-\beta}$ (deep cutoffs included) from which the branching factor can be estimated. In particular, Baudet shows that for bivalued terminal positions $r_{\alpha-\beta}$ could be as high as $P^*/(1 - P^*)$ (a special case of (7.19)); and for the continuous case that $r_{\alpha-\beta}$ lower bounded by $r_{\alpha-\beta} \geq P^*/(1 - P^*)$ (a special case of (7.31)). A tighter upper bound for $r_{\alpha-\beta}$ was then computed which significantly narrowed the gap left by Knuth and Moore to less than 20% in the range $2 \leq d \leq 32$.

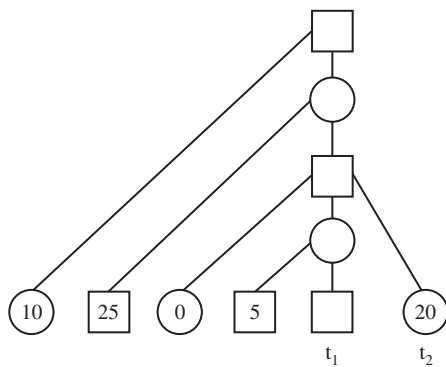
In view of the fact that the SCOUT algorithm was found to achieve the lower bound $P^*/(1 - P^*)$, we were first led to believe that α - β , which appears to be much more economical than SCOUT, also achieves this bound and that the uncertainty

2. Peter Homeiert, personal communication.

concerning the actual branching factor of $\alpha\text{-}\beta$ has finally been eliminated. However, after several futile attempts to prove the superiority of $\alpha\text{-}\beta$ over SCOUT, we found counter-examples demonstrating that SCOUT's extra caution in testing prior to evaluation may sometimes pay off, causing it to skip nodes which would be visited by $\alpha\text{-}\beta$. In the diagram below, the node marked t_1 would be examined by the $\alpha\text{-}\beta$ procedure but ignored by SCOUT.



When J is submitted to the test $\text{TEST}(J, 5, >)$, the zero value assigned to node t_2 causes the test to fail, whereas during the $\text{TEST}(K, 5, >)$ phase, t_1 is skipped by virtue of its elder sibling having the value 10. $\alpha\text{-}\beta$, on the other hand, has no way of finding out the low value of t_2 before t_1 is examined.



The converse situation can, of course, also be demonstrated. The diagram above shows how a node (t_1) which is visited by SCOUT is cut off by $\alpha\text{-}\beta$. However, the asymptotic performance of SCOUT is at least as good as that of $\alpha\text{-}\beta$ by virtue of Theorem 7.4 and the fact that $\alpha\text{-}\beta$ is directional.

Our inability to demonstrate the asymptotic equivalence of SCOUT and $\alpha\text{-}\beta$ on a node by node basis leaves the branching factor of the $\alpha\text{-}\beta$ procedure enigmatic and renders its asymptotic optimality unsettled. We wish to conjecture, though, that $\alpha\text{-}\beta$ probably does reach the branching factor $P^*/(1 - P^*)$ and that it is, therefore, asymptotically optimal over all directional algorithms. It would simply be

too amusing to find a wasteful procedure such as SCOUT outperforming the α - β procedure.³

However, the uncertainty regarding the branching factor of the α - β procedure only pertains to continuous valued trees. We shall next demonstrate that when the terminal positions are assigned discrete values, the α - β procedure attains the absolute minimal branching factor of $d^{1/2}$.

The fact that at high levels almost all nodes attain the same minimax value, v^* , makes it increasingly probable that the α - β cutoff conditions⁴ are met successfully at all nodes where they are applicable and this, in turn, gives rise to a branching factor of $d^{1/2}$. For, consider the top m cycles of a $(n + m)$ -cycle tree, if all cutoff conditions are met at this portion of the tree, only $2d^m - 1$ n th level nodes need be expanded. Therefore, denoting by x_n the expected number of terminal positions examined by α - β evaluating any n th level node, and by $P_{\alpha-\beta}(n, m)$ the probability that all cutoff conditions are satisfied whenever applicable, we can write:

$$\frac{x_{n+m}}{x_n} \leq (2d^m - 1)P_{\alpha-\beta}(n, m) + d^{2m}[1 - P_{\alpha-\beta}(n, m)]. \quad (7.49)$$

On the other hand, the event of meeting all cutoff conditions is subsumed by the event that all the $2d^m - 1$ nodes expanded attain the limit value v^* , and consequently:

$$P_{\alpha-\beta}(n, m) \geq P[V_{n-1}(S) = v^*]^{(2d^m-1)}.$$

Now, letting $m = n^2$ and recalling (7.6) that $P[V_{n-1}(S) = v^*]$ approaches unity at a super exponential rate:

$$1 - P[V_{n-1}(s) = v^*] \leq (\delta)^{d^{n-n_0}} \quad \text{for } n > n_0$$

where δ is a fraction strictly smaller than 1, and n_0 a function of $\delta, F_{V_0}(v_i)$ and $F_{V_0}(v_{i-1})$ (see Theorem 7.1), we obtain:

$$\lim_{n \rightarrow \infty} d^{2m}[1 - P_{\alpha-\beta}(n, m)] = 0$$

and from (7.49):

$$\frac{x_{n-n^2}}{x_n} = O(2d^{n^2}).$$

3. This conjecture has recently been confirmed (see Pearl, J., "The Solution for the Branching Factor of the Alpha-Beta Pruning Algorithm," UCLA-ENG-CSL-8019, School of Engineering and Applied Science. University of California, Los Angeles. May 1980).

4. [2] contains an elaborate description of the α - β cutoff conditions, using a notation similar to ours.

The effective branching factor for the entire $(n + n^2)$ -cycle tree, even assuming that every node expanded at the n th cycle requires the examination of all d^{2n} terminal nodes under it, becomes:

$$r_{\alpha-\beta} = \lim_{n \rightarrow \infty} [2d^{n^2} \cdot d^{2n}]^{1/2(n+n^2)} = d^{1/2}.$$

We summarize this result by stating:

Theorem 7.7 *The expected number of terminal positions examined by the ALPHA-BETA procedure in evaluating a (h, d, F) -game with discrete terminal values has a branching factor $r_{\alpha-\beta} = d^{1/2}$ with exceptions only when one of the discrete values, v^* , satisfies $F(v^*) = 1 - P^*$.*

Corollary 7.7 *For games with discrete terminal values satisfying the conditions of Theorem 7.7, the α - β procedure is asymptotically optimal over all evaluation algorithms.*

Paralleling our discussion of the SCOUT algorithm, α - β too does not acquire the more efficient branching factor of $d^{1/2}$ by an abrupt transition from the continuous to the discrete case. If the terminal values are drawn from M equally likely integers, (7.48) provides an estimate for the height $h_0(M, d)$ at which the search would become more efficient. Note, however, that it is not the total number of quantization levels v_1, v_2, \dots, v_M which affects the search efficiency but rather the distances of $F_{V_0}(v_i)$ and $F_{V_0}(v_{i-1})$ from $1 - P^*$. Thus, coarser quantizations in the neighborhood of v^* have a more significant role in speeding up the α - β procedure.

Recently, Stockman [7] has introduced a non-directional algorithm which examines fewer nodes than α - β . The magnitude of this improvement has not been evaluated yet, but the superiority of Stockman's algorithm could be one of the following two types. It may either possess a reduced branching factor, or it may exhibit a marginal improvement at low h 's which disappears on taller trees. If the superiority is of the former type, it must be singular to the continuous case because in the discrete case, Corollary 7.7 states that α - β is asymptotically optimal over all algorithms, directional as well as non-directional.

It would still be interesting, though, to find out if any non-directional algorithm can solve a (h, d, P^*) -game with branching factor lower than $P^*/(1 - P^*)$. If such an algorithm exists it could be incorporated into SCOUT (replacing TEST) and thus enabling it to evaluate continuous valued game trees with a branching factor lower than $P^*/(1 - P^*)$.

References

1. Baudet, G.M., On the branching factor of the alpha-beta Pruning algorithm, *Artificial Intelligence* **10** (1978) 173-199.

2. Fuller, S.H., Gaschnig, J.G. and Gillogly, J.J., An analysis of the alpha-beta Pruning algorithm. Department of Computer Science Report, Carnegie-Mellon University (July 1973).
3. Knuth, D.E. and Moore, R.N., An analysis of alpha-beta Pruning, *Artificial Intelligence* 6 (1975) 293–326.
4. Nilsson, N.J., *Problem-Solving Methods in Artificial Intelligence* (McGraw-Hill, New York, 1971).
5. Slagle, J.R., *Artificial Intelligence: The Heuristic Programming Approach* (McGraw-Hill, New York, 1971).
6. Slagle, J.R. and Dixon, J.K., Experiments with some programs that search game trees, *J. ACM* 2 (1969) 189–207.
7. Stockman, G., A minimax algorithm better than alpha-beta?, *Artificial Intelligence* 12 (1979) 179–196.

Received 2 January 1980; revised version received 12 March 1980

The Solution for the Branching Factor of the Alpha–Beta Pruning Algorithm and its Optimality

Judea Pearl

This paper analyzes $N_{n,d}$, the average number of terminal nodes examined by the α - β pruning algorithm in a uniform game tree of degree n and depth d for which the terminal values are drawn at random from a continuous distribution. It is shown that increasing the search depth by one extra step would increase $N_{n,d}$ by

This work was supported in part by the National Science Foundation Grants MCS 78-07468 and MCS 78-18924.

An early version of this paper was presented at the *8th International Conference on Automata Languages and Algorithms*, Acre, Israel, July 3–17, 1981.

M. Douglas McIlroy, former editor of *Programming Techniques and Data Structures*, of which Ellis Horowitz is the current editor.

Author's Present Address: Judea Pearl, Cognitive Systems Laboratory, Departments of Computer Science and Engineering Systems, University of California, Los Angeles, Los Angeles, CA 90024.

Permission to copy without fee all or part of this material is granted provided that the copies are not made or distributed for direct commercial advantage, the ACM copyright notice and the title of the publication and its date appear, and notice is given that copying is by permission of the Association for Computing Machinery. To copy otherwise, or to republish, requires a fee and/or specific permission.

Programming Techniques and Data Structures

Originally published in *Communications of the ACM* August 1982 Volume 25 Number 8

© 1982 ACM 0001-0782/82/0800-0559 \$00.75. All rights reserved.

Original DOI: [10.1145/358589.358616](https://doi.org/10.1145/358589.358616)

a factor (called the *branching factor*) $\mathcal{R}_{\alpha-\beta}(n) = \xi_n/1 - \xi_n \approx n^{3/4}$ where ξ_n is the positive root of $x^n + x - 1 = 0$. This implies that for a given search time allotment, the α - β pruning allows the search depth to be increased by a factor $\approx 4/3$ over that of an exhaustive minimax search. Moreover, since the quantity $(\xi_n/1 - \xi_n)^d$ has been identified as an absolute lower bound for the average complexity of all game searching algorithms, the equality $\mathcal{R}_{\alpha-\beta}(n) = \xi_n/1 - \xi_n$ now renders α - β *asymptotically optimal*.

CR Categories and Subject Descriptors: I.2.8 [Artificial Intelligence]: Problem Solving, Control Methods and Search—*graph and tree search strategies, heuristic methods*; F.2.2 [Analysis of Algorithms and Problem Complexity]: Nonnumerical Algorithms and Problems—*complexity of proof procedures, sorting and searching*; G.2.2 [Discrete Mathematics]: Graph Theory—*graph algorithms, trees*

General Term: Algorithms

Additional Key Words and Phrases: alpha-beta search, game searching, games, minimax algorithms, branch and bound search, average case analysis

8.1 Introduction

8.1.1 Informal Description of the α - β Procedure

The α - β pruning algorithm is the most commonly used procedure in game playing applications, where it serves to speed up game searching without loss of information. The algorithm determines the minimax value of the root of a game tree by traversing the tree in a predetermined order, for example, from left to right, skipping all those nodes that can no longer influence the minimax value of the root.

The method is demonstrated in Figure 8.1 which shows a binary game tree of depth $d = 4$ with nodes at maximizing levels (called MAX nodes) and at minimizing levels (called MIN nodes) represented by squares and circles, respectively. The numbers inside the terminal squares represent evaluations of the game positions at the frontier of the search tree, while those at higher levels are the minimax values computed by the α - β procedure. The heavy branches represent the search tree actually generated by the α - β procedure as it traverses the game tree from left to right. Nodes not on that search tree are skipped (or “cutoff”) by α - β , as they cannot provide useful information.

The rationale for node skipping can be explained by examining the nodes labeled *A*, *B*, and *C*, in Figure 8.1. The purpose of exploring node *B* has been to find out if the value of *A* can be reduced below 10, which is the value established

$N_{n,d}$, the average number of terminal nodes examined during the search, has become a standard yardstick for the complexity of the search method. Additionally, the significant parameter for very deep trees is the *branching factor*

$$\mathcal{R}_{\alpha-\beta} = \lim_{d \rightarrow \infty} (N_{n,d})^{1/d}$$

which measures the effective number of branches actually explored by α - β from a typical node of the search tree.

Slagle and Dixon [8] showed that the number of terminal nodes examined by α - β must be at least $n^{\lfloor d/2 \rfloor} + n^{\lceil d/2 \rceil} - 1$ but may, in the worst case, reach the entire set of n^d terminal nodes. The analysis of expected performance using uniform trees with random terminal values began with Fuller, Gaschnig, and Gillogly [2] who obtained formulas by which the average number of terminal examinations $N_{n,d}$ can be computed. Unfortunately, the formula would not facilitate asymptotic analysis; simulation studies led to the estimate $\mathcal{R}_{\alpha-\beta} \approx (n)^{0.72}$.

Knuth and Moore [3] analyzed a less powerful but simpler version of the α - β procedure by ignoring deep cutoffs. They showed that the branching factor of this simplified model is $O(n/\log n)$ and speculated that the inclusion of deep cutoffs would not alter this behavior substantially. A more recent study by Baudet [1] confirmed this conjecture by deriving an integral formula for $N_{n,d}$ (deep cutoffs included), from which the branching factor can be estimated. In particular, Baudet shows that $\mathcal{R}_{\alpha-\beta}$ is bounded by $\xi_n/1 - \xi_n \leq \mathcal{R}_{\alpha-\beta} \leq M_n^{1/2}$, where ξ_n is the positive root of $x^n + x - 1 = 0$ and M_n is the maximal value of the polynomial $P(x) = (1 - x^n/1 - x)[1 - (1 - x^n)^n/x^n]$ in the range $0 \leq x \leq 1$. Pearl [5] has shown both that $\xi_n/1 - \xi_n$ lower bounds the branching factor of every directional game searching algorithm and that an algorithm exists (called SCOUT) that actually achieves this bound. Tarsi [10] has very recently shown that $\xi_n/1 - \xi_n$ also lower bounds the branching factor of nondirectional algorithms. Thus, the enigma of whether α - β is optimal remains contingent upon determining the exact magnitude of $\mathcal{R}_{\alpha-\beta}$ within the range delineated by Baudet.

This paper now shows that the branching factor of α - β indeed coincides with the lower bound $\xi_n/1 - \xi_n$, thus establishing the asymptotic optimality of α - β over the class of all game searching algorithms.

8.2 Analysis

8.2.1 An Integral Formula for $N_{n,d}$

Our starting point will be an examination of the conditions under which an arbitrary node J is generated by the α - β algorithm. If all terminal values to the left of

J are given, one can perform a simple test to determine whether or not J will be generated. For a MAX node J , form the path leading from the root to J , and define the following quantities:

$A(J)$ = the highest minimax value among all left-siblings of odd ancestors of J

$B(J)$ = the lowest minimax value among all left-siblings of even ancestors of J

J will be generated by α - β if and only if

$$A(J) < B(J).$$

The same criterion holds when J is a MIN node, except that $A(J)$ is computed over even ancestors and $B(J)$ over odd ancestors of J . A special definition is required to include so-called *critical* nodes for which the corresponding sets of left-siblings are empty [7].

The reader can easily verify that in Figure 8.1 all nodes generated satisfy the criterion above while all those satisfying $A(J) \geq B(J)$ can provide no information beyond that which has already been gathered by the search and will be cut off. For example, for the rightmost leaf node we have:

$$A(J) = \max[\min(3, 22), 10] = 10$$

$$B(J) = \min\{20, \max[\min(5, 6), \min(4, 7)]\} = 5$$

and since $A(J) > B(J)$, it is not generated by the α - β search.

The criterion above was first derived by Fuller et al. [2] and is a useful tool for computing $N_{n,d}$, the average number of terminal nodes examined by α - β . One need only compute the probability $P[A(J) < B(J)]$ for every node J , then sum these probabilities over all terminal nodes.

$$N_{n,d} = \sum_{J \text{ terminal}} P[A(J) < B(J)]$$

This procedure may seem like a major undertaking. Fortunately, when the terminal values are drawn independently from a common distribution function $f_0(x) = P[V_0 \leq x]$, very simple propagation rules govern the distributions of the minimax values at higher levels of the tree. For example, if V_k stands for the minimax value of a MIN node at level k of the tree, then its distribution f_k is related to that of its direct descendants by

$$f_k(x) = 1 - [1 - f_{k-1}(x)]^n$$

and to that of its grandsons by

$$f_k(x) = 1 - \{1 - [f_{k-2}(x)]^n\}^n$$

From these recursions one can compute the distributions $F_{A(J)}(x)$ and $F_{B(J)}(x)$ of the random variables $A(J)$ and $B(J)$ for any terminal node J . Moreover, since $A(J)$ and $B(J)$ are independent and continuous (for noncritical nodes) we have

$$P[A(J) < B(J)] = \int_{x=-\infty}^{\infty} F_{A(J)}(x)F'_{B(J)}(x) dx$$

and $N_{n,d}$ becomes

$$N_{n,d} = \int_{x=-\infty}^{\infty} \left[\sum_{J \text{ terminal}} F_{A(J)}(x)F'_{B(J)}(x) \right] dx + n^{\lfloor d/2 \rfloor} + n^{\lceil d/2 \rceil} - 1$$

where the terms added to the integral represent the number of critical nodes, all of which are examined. The summation inside the integral can be performed using the recursion relations above (see Roizen [7]) and lead to the following theorem

Theorem 8.1 Let $f_0(x) = x$, and, for $i = 1, 2, \dots$, define

$$f_i(x) = 1 - \{1 - [f_{i-1}(x)]^n\}^n$$

$$r_i(x) = \frac{1 - [f_{i-1}(x)]^n}{1 - f_{i-1}(x)}$$

$$s_i(x) = \frac{f_i(x)}{[f_{i-1}(x)]^n}$$

$$R_i(x) = r_1(x) \times \dots \times r_{\lfloor i/2 \rfloor}(x)$$

$$S_i(x) = s_1(x) \times \dots \times s_{\lfloor i/2 \rfloor}(x)$$

The average number $N_{n,d}$ of terminal nodes examined by the α - β pruning algorithm in a uniform game tree of degree n and depth d for which the bottom values are drawn from a continuous distribution is given by

$$N_{n,d} = n^{\lfloor d/2 \rfloor} + \int_0^1 R'_d(t)S_d(t) dt \quad (8.1) \quad \blacksquare$$

An identical expression for $N_{n,d}$ was first derived by Baudet ([1], Theorem 4.2) starting with discrete terminal values and progressively refining their quantization levels.

8.2.2 Evaluation of $\mathcal{R}_{\alpha-\beta}$

The difficulty in estimating the integral in Equation (8.1) stems from the recursive nature of $f_i(x)$ which tends to obscure the behavior of the integrand. We circumvent this difficulty by substituting for $f_0(x)$ another function $\phi(x)$ which makes the regularity associated with each successive iteration more transparent.

The value of the integral in Equation (8.1) does not depend on the exact nature of $f_0(x)$ as long as it is monotone from some interval $[a, b]$ onto the range $[0, 1]$. This is evident by noting that by substituting $f_0(x) = \phi(x)$ the integral becomes

$$\int_{x=a}^b \frac{dR_d[\phi(x)]}{dx} S_d[\phi(x)] dx = \int_{\phi=0}^1 \frac{dR_d(\phi)}{d\phi} S_d(\phi) d\phi$$

which is identical to that in Equation (8.1). This invariance reflects the fact that the search procedure depends only on the relative order of the d^n terminal values, not on their magnitudes, and since any continuous distribution of the terminal values generates all ranking permutations with equal probabilities, $N_{n,d}$ will not be affected by the *shape* of that distribution. Consequently, $f_0(x)$ which represents the terminal values' distribution, may assume an arbitrary form, subject to the usual constraints imposed on continuous distributions.

A convenient choice for the distribution $f_0(x)$ would be a characteristic function $\phi(x)$ that would render the distributions of the minimax value of every node in the tree identical in shape. Such a characteristic distribution indeed exists [6] and satisfies the functional equation

$$\phi(x) = g[\phi(ax)] \quad (8.2)$$

where

$$g(\phi) = 1 - (1 - \phi^n)^n \quad (8.3)$$

and a is a real-valued parameter to be determined by the requirement that Equation (8.2) possess a nontrivial solution for $\phi(x)$. This choice of $\phi(x)$ renders the functions $\{f_i(x)\}$ in Theorem 8.1 identical in shape, save for a scale factor. Accordingly, we can write

$$f_i(x) = \phi(x/a^i) \quad (8.4)$$

$$r_i(x) = r(x/a^{i-1}) \quad (8.5)$$

$$s_i(x) = s(x/a^{i-1}) \quad (8.6)$$

where

$$r(x) = \frac{1 - [\phi(x)]^n}{1 - \phi(x)} \quad (8.7)$$

and

$$s(x) = \frac{1 - \{1 - [\phi(x)]^n\}^n}{[\phi(x)]^n} \quad (8.8)$$

Equation (8.2), known as the Poincaré Equation [4], has a nontrivial solution $\phi(x)$ with the following properties [6]:

(i) $\phi(0) = \xi_n$ where ξ_n is the root of

$$x^n + x - 1 = 0 \quad (8.9)$$

(ii)
$$a = \frac{1}{g'(\xi_n)} = \left[\frac{\xi_n}{n(1 - \xi_n)} \right]^2 < 1 \quad (8.10)$$

(iii) $\phi'(0)$ can be chosen arbitrarily, for example, $\phi'(0) = 1$

(iv)
$$\begin{aligned} x(\phi) &= \lim_{k \rightarrow \infty} a^k [g^{-k}(\phi) - \xi_n] \\ \phi(x) &\underset{x \rightarrow \infty}{\approx} 1 - (n)^{-n/n-1} \exp[-(x)^{-\ln(n)/\ln(a)}] \\ \phi(x) &\underset{x \rightarrow -\infty}{\approx} (n)^{-1/n-1} \exp[-(x)^{-\ln(n)/\ln(a)}] \end{aligned}$$

However, only properties (8.9) and (8.10) will play a role in our analysis. Most significantly, parameter a , which is an implicit function of n , remains lower than 1 for all n .

Substituting Equations (8.4), (8.5), and (8.6) into Equation (8.1) and considering, without loss of generality, the case where d is an even integer, $d = 2h$, we obtain

$$N_{n,d} = n^h + \int_{x=-\infty}^{\infty} \pi_h(x) \left(\sum_{i=1}^h \frac{r'_i(x)}{r_i(x)} \right) dx \quad (8.11)$$

where

$$\pi_h(x) = \prod_{i=0}^{h-1} p(x/a^i), \quad (8.12)$$

$$p(x) = r(x)s(x) = P[\phi(x)], \quad (8.13)$$

and

$$P(\phi) = \frac{1 - \phi^n}{1 - \phi} \frac{1 - (1 - \phi^n)^n}{\phi^n} \quad (8.14)$$

Using Equations (8.5) and (8.7), it can be easily shown that $r'_i(x)/r_i(x)$ satisfies

$$\frac{r'_i(x)}{r_i(x)} \leq n \phi'(x/a^{i-1})l/a^{i-1} \quad (8.15)$$

and consequently, Equation (8.11) becomes

$$N_{n,d} \leq n^h + n \int_{-\infty}^{\infty} \pi_h(x) \cdot \left[\sum_{i=1}^h \phi'(x/a^{i-1})l/a^{i-1} \right] dx \quad (8.16)$$

We now wish to bound the term $\pi_h(x)$ from above.

An examination of $p(x) = P[\phi(x)]$ [Equations (8.13) and (8.14)] reveals that $p(x)$ is unimodal in x , $p(0) = [\xi_n/1 - \xi_n]^2$, and that $p(x)$ lies above the asymptotes $p(-\infty) = p(+\infty) = n$. Moreover, the maximum of $P(\phi)$ occurs below $\phi = \xi_n$ and, consequently, $p(x)$ attains its maximum M_n below $x = 0$.

At this point, were we to use the bound $\pi_h(x) \leq M_n^h$ in (8.16), it would result in $N_{n,d} < n^h + nhM_n^h$ and lead to Baudet's bound $\mathcal{R}_{\alpha-\beta} \leq M_n^{1/2}$. Instead, a tighter bound can be established by exploiting the unique relationships between the factors of $\pi_h(x)$.

Lemma 8.1 Let $x_0 < 0$ be the unique negative solution of $p(x_0) = p(0)$. $\pi_h(x)$ attains its maximal value in the range $a^{h-1}x_0 \leq x \leq 0$.

Proof. Since $p(x)$ is unimodal we have $p(x) < p(0)$ and $p'(x) > 0$ for all $x < x_0$. Consequently, for all $x < x_0$, any decrease in the magnitude of $|x|$ would result in increasing $p(x)$, that is, $p(cx) > p(x)$ for all $0 \leq c < 1$. Now consider $\pi_h(ax)$.

$$\begin{aligned} \pi_h(ax) &= p(x/a^{h-2})p(x/a^{h-3}) \cdots p(x)p(ax) \\ &= \pi_h(x)p(ax)/p(x/a^{h-1}); \end{aligned}$$

for all x' satisfying $x'/a^{h-1} < x_0$ we must have $p(ax') > p(x'/a^{h-1})$ (using $c = a^h < 1$) and $\pi_h(ax') > \pi_h(x')$, implying that $\pi_h(x')$ could not be maximal. Consequently, for $\pi_h(x')$ to be maximal, x' must be in the range $x_0a^{h-1} \leq x' \leq 0$. ■

Lemma 8.2 $\pi_h(x)$ can be bounded by

$$\pi_h(x) \leq A(n)[p(0)]^h \quad (8.17)$$

where $A(n)$ is a constant multiplier independent on h .

Proof. Since $p(x)$ is continuous, there exists a positive constant α such that $p(x) \leq p(0) - \alpha x$ for all $x \leq 0$. Consequently, using Lemma 8.1, we can write

$$\begin{aligned} \max_x \pi_h(x) &= \max_{a^{h-1}x_0 \leq x \leq 0} \pi_h(x) \\ &\leq \max_{a^{h-1}x_0 \leq x \leq 0} \prod_{i=0}^{h-1} (p(0) - \alpha x/a^i) \\ &\leq [p(0)]^h \max_{a^{h-1}x_0 \leq x \leq 0} \exp\left(\sum_{i=0}^{h-1} -\frac{\alpha x}{a^i p(0)}\right) \\ &= [p(0)]^h \exp\left[\frac{-\alpha x_0}{p(0)} a^{h-1} \sum_{i=0}^{h-1} 1/a^i\right] \\ &\leq [p(0)]^h \exp\left[\frac{-\alpha x_0}{p(0)(1-a)}\right] \end{aligned}$$

Selecting $A(n) = \exp[-\alpha x_0/p(0)(1-a)]$ proves the Lemma. ■

Theorem 8.2 The branching factor of the α - β procedure for a uniform tree of degree n is given by

$$\mathcal{R}_{\alpha-\beta} = \frac{\xi_n}{1 - \xi_n} \quad (8.18)$$

where ξ_n is the positive root of the equation $x^n + x - 1 = 0$.

Proof. Substituting (8.17) in (8.16) yields

$$\begin{aligned} N_{n,d} &\leq n^h + n A(n)[p(0)]^h \\ &\quad \cdot \int_{-\infty}^{\infty} \sum_{i=0}^{h-1} (1/a^i) \phi'(x/a^i) dx \\ &\simeq n^h + n A(n)[p(0)]^h h \end{aligned}$$

Finally, using $p(0) = (\xi_n/1 - \xi_n)^2 > n$, we obtain

$$\mathcal{R}_{\alpha-\beta} = \lim_{h \rightarrow \infty} (N_{n,d})^{1/2h} \leq \xi_n/1 - \xi_n \quad (8.19)$$

This, together with Baudet's lower bound $\mathcal{R}_{\alpha-\beta} \geq \xi_n/1-\xi_n$, completes the proof of Theorem 8.2. ■

8.3 Conclusions

The asymptotic behavior of $\mathcal{R}_{\alpha-\beta}$ is $O(n/\log n)$, as predicted by Knuth's analysis [3]. However, for moderate values of n ($n \leq 1000$), $\xi_n/1-\xi_n$ is fitted much better by the formula $(0.925)n^{0.747}$ (see Figure 4 of [5]), which vindicates the simulation results of Fuller et al. [2]. This approximation offers a more meaningful appreciation of the pruning power of the $\alpha-\beta$ algorithm. Roughly speaking, a fraction of only $(0.925)n^{0.747}/n \approx n^{-1/4}$ of the legal moves will be explored by $\alpha-\beta$. Alternatively, for a given search time allotment, the $\alpha-\beta$ pruning allows the search depth to be increased by a factor $\log n/\log \mathcal{R}_{\alpha-\beta} \approx 4/3$ over that of an exhaustive minimax search.

The establishment of the precise value of $\mathcal{R}_{\alpha-\beta}$ for continuous-valued trees, together with a previous result that $\mathcal{R}_{\alpha-\beta} = n^{1/2}$ for almost all discrete-valued trees [5], completes the characterization of the asymptotic behavior of $\alpha-\beta$ and settles the question of its optimality. The fact that $\alpha-\beta$ is asymptotically optimal (that is, achieves the lowest possible branching factor) over the class of directional algorithms follows directly from Equation (8.18) and a previous result [5] that $\xi_n/1-\xi_n$ lower bounds the branching factor of any directional algorithm. However, the possible existence of some nondirectional algorithm outperforming $\alpha-\beta$ and exhibiting a branching factor lower than $\xi_n/1-\xi_n$ has remained unsettled until very recently. Indeed, Stockman [9] introduced a nondirectional algorithm called SSS* which consistently examines fewer nodes than $\alpha-\beta$. Hopes were then raised that the superiority of Stockman's algorithms reflected an improved branching factor over that of $\alpha-\beta$.

These possibilities have all been eliminated by a more recent result by Tarsi [10]. Considering a standard bi-valued game tree in which the terminal nodes are assigned the values 1 and 0 with the probabilities ξ_n and $1-\xi_n$, respectively, Tarsi's result states that *any* algorithm which solves such a game tree must, on the average, examine at least $(\xi_n/1-\xi_n)^d$ terminal positions. At the same time the task of solving any bi-valued game tree is equivalent to the task of verifying an inequality proposition regarding the minimax value of a continuous-valued game tree [5] of identical structure, and, consequently, the former cannot be more complex than the latter. Thus, the quantity $(\xi_n/1-\xi_n)^d$ should also lower bound the expected number of nodes examined by any algorithm searching a continuous-valued game tree. This, together with Equation (8.18), establishes the asymptotic optimality of $\alpha-\beta$.

Received 5/80; revised 7/81; accepted 11/81

References

1. Baudet, G.M. On the branching factor of the alpha-beta pruning algorithm. *Artificial Intelligence* 10, 2 (April 1978), 173–199.
2. Fuller, S.H., Gaschnig, J.G., and Gillogly, J.J. An analysis of the alpha-beta pruning algorithm. Department of Computer Science Report, Carnegie-Mellon University, (1973).
3. Knuth, D.E., and Moore, R.N. An analysis of alpha-beta pruning. *Artificial Intelligence* 6 (1975), 293–326.
4. Kuczma, M. *Functional Equations in a Single Variable*. Polish Scientific Publishers, Warszawa, (1968), p. 141.
5. Pearl, J. Asymptotic properties of minimax trees and game-searching procedures. *Artificial Intelligence* 14, 2 (Sept. 1980), 113–138.
6. Pearl, J. A space-efficient on-line method of computing quantile estimates *J. of Algorithms* 2, 2 (June 1981) 24–28.
7. Roizen, I. On the average number of terminal nodes examined by alpha-beta. UCLA-ENG-CSL-8108, Cognitive Systems Laboratory, University of California, Los Angeles, (1981).
8. Slagle, J.R., and Dixon, J.K. Experiments with some programs that search game trees. *JACM* 16, 2 (April 1969) 189–207.
9. Stockman, G. A minimax algorithm better than alpha-beta? *Artificial Intelligence* 12, 2 (Aug. 1979), 179–196.
10. Tarsi, M. Optimal searching of some game trees. UCLA-ENG- CSL-8108, Cognitive Systems Laboratory, University of California, Los Angeles, (1981). (To appear in *JACM*.)

On the Discovery and Generation of Certain Heuristics

Judea Pearl

Abstract

This paper explores the paradigm that heuristics are discovered by consulting simplified models of the problem domain. After describing the features of typical heuristics on some popular problems, we demonstrate that these heuristics can be obtained by the process of deleting constraints from the original problem and solving the relaxed problem which ensues. We then outline a scheme for generating such heuristics mechanically, which involves systematic refinement and deletion of constraints from the original problem specification until a semi-decomposable model is identified. The solution to the latter constitutes a heuristic for the former.

9.1 Introduction: Typical Uses of Heuristics

Heuristics are methods and criteria for judging the relative merits of alternative courses of planning or action. There is hardly any intellectual activity which does not rely on heuristics of some kind. The decision to begin reading this paper, for example, reflects a tacit use of heuristics which has lured the reader to invest

Prepared for *UCLA-Computer Science Department Quarterly*, Spring 1982, this work was supported in part by the National Science Foundation Grant MCS 81 14209

Technical Report R-39

Originally published in *AI Magazine* Volume 4 Number 1 (1983)

© AAIL. Republished with permission from AAIL.

Original DOI: [10.1609/aimag.v4i1.385](https://doi.org/10.1609/aimag.v4i1.385)

time and effort in anticipation of certain benefits. Although such anticipations may occasionally be disappointed, on the whole they are essential to planning our everyday activities.

Complex combinatorial problems require the use of heuristics if a reasonably “good” solution is to be produced within practical time constraints. We shall demonstrate this point using three simple problems (readers familiar with the properties of A^* may skip to section [9.1.3]. Where do these heuristics come from?):

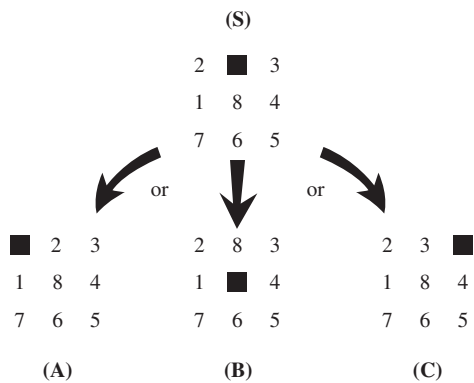


Figure 9.1 A goal tree in the 8-puzzle

The 8-puzzle. This simple puzzle is a one-person game, the objective of which is to rearrange a given configuration of eight uniquely numbered tiles on a 3×3 board into another given configuration by iteratively sliding one of the tiles into empty location, as in Figure 9.1 above:

Assume that in this example the objective is to reach the goal state:

1 2 3
8 ■ 4
7 6 5

Which of the three alternatives, (A), (B), or (C) appears most promising? The answer can, of course, be obtained by searching the graph associated with the puzzle and finding which of the three states leads to the shortest path to the goal. The notorious combinatorial explosion, however, makes this method utterly impractical when the distance to the goal is large and/or when larger boards (e.g., 4×4) are involved. The very search for a solution path requires the use of judgments to decide at any point which search avenue is the most promising.

To assist a computer in solving path-finding problems of this type, the programmer is usually required to provide a rule for computing an *estimate* of the proximity between two given configurations. The most popular rules for the 8-puzzle are: h_1 = the number of tiles by which the two configurations differ, and h_2 = the sum of the distances of the mismatched tiles from their proper destinations. (The black position is not counted.) The appropriate distance measure, in this case, is the sum of the coordinate differences, also known as the Manhattan or city-block distance. For instance, in the example above we can compute:

$$\begin{aligned} h_1(A) &= 2 & h_1(B) &= 3 & h_1(C) &= 4 \\ h_2(A) &= 2 & h_2(B) &= 4 & h_2(C) &= 4 \end{aligned}$$

These heuristic functions are intuitively appealing and readily computable, and may be used to prune the space of possibilities in such a way that only configurations lying close to the solution path will actually be explored. An algorithm which exhibits such pruning will be discussed later.

Finding the shortest path in a road map. Given a road map such as the one shown in Figure 9.2, it is desired to find the shortest path between city A and city B. If the intercity distances are presented in the form of a distance-matrix $d(i, j)$, there is no way for the search program to judge a priori that city C, unlike city D, lies way out of the natural direction from A to B and, consequently, that city D is better candidate from which to pursue the search. At the same time, the preference of D over C is obvious to anyone who glances at the map. What extra information does the map provide which is not made explicit in the distance table?

One possible answer is that the human observer exploits vision machinery to estimate the Euclidean distances in the map and, since the air distance from D to B is shorter than that between C and B, city D appears as a more promising candidate from which to launch the search. That same information can also be used by the machine if each city is assigned a heuristic function $h(*)$ equal to the air distance between that city and the goal B. A tentative choice between pursuing the search from city C or city D should depend, then, on the magnitude of the cost estimate $d(A, C) + h(C)$ relative to the estimate $d(A, D) + h(D)$.

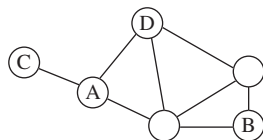


Figure 9.2 A graph expressing the shortest path problem of going from A to B.

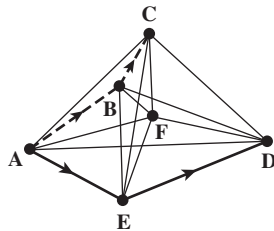


Figure 9.3 Search for cheapest path in graph.

9.1.1 The Traveling Salesman Problem (TSP)

Here we must find the cheapest tour, that is, the cheapest path which visits every node once and only once, and returns to the initial node, in a complete graph of N nodes with each edge assigned a non-negative cost.

It is well known that the TSP is NP-hard and that all known algorithms require an exponential time in the worst case. However, the use of good bounding functions often enables us (using the branch-and-bound algorithm) to find the optimal tour in much less time. What is a bounding function? Consider the graph below where the two marked paths ABC and AED represent two partial tours currently being considered by the search procedure. Which of the two, if properly completed to form a circuit, is more likely to be part of the optimal solution? Clearly, the overall solution cost is given by the cost of completing the tour added to the cost of the initial subtour, and so the answer lies in how cheaply we can complete the tour through the remaining nodes. However, since the computational effort required to find the optimal completion cost is almost as hard as that of finding the entire optimal tour, we must settle for an *estimate* of the completion cost. Given such estimates, the decision of which subtour to extend first would depend on which one, by combining the cost of the explored part with the estimate of its completion, offers a lower *overall* cost estimate. It can be shown that if at every stage of the search we select for exploration that partial tour with the lowest estimated cost, and if the estimates of the completion costs are consistently optimistic (underestimates), then the first tour to be completed by the search is also the optimal one.

What easily computable function would yield an optimistic, yet not too unrealistic, estimate of the subtour completion cost? People, when first asked to “invent” such a function, usually provide easily computable, but too simplistic answers. For example, the cheapest edge or two-edge path connecting the end of the initial subtour, bypassing all unvisited cities or going through one other city, respectively. These functions, while being optimistic, grossly underestimate the completion cost. Upon deeper thought, more realistic estimates are formed, and the two which have received the greatest attention in the literature are: (1) The cheapest

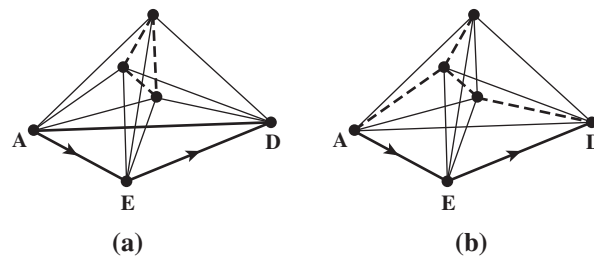
2nd degree graph going through the remaining nodes (Lawler and Wood, 1966), and (2) the cost of the minimum spanning tree (MST) through all remaining nodes (Held and Karp, 1971). The first is obtained by solving the so-called optimal assignment problem using $O(N^3)$ steps, while the second requires $O(N^2)$ steps.

That these two functions provide optimistic estimates of the completion cost is apparent when we consider that completing the tour requires the choice of a path going through all unvisited cities. A path is both a special case of a graph with degree 2 and also a special case of a spanning tree. Hence, the set of all completion paths is included in the sets of objects over which the optimization takes place, and so, the solution found must have a lower cost than that obtained by optimizing over the set of path only.

Figure 9.4a below shows the shape of a graph that may be found by solving the assignment problem instead of completing subtour AED. The completion part is not a single path but a collection of loops. Figure 9.4b shows a minimum-spanning-tree completion of subtour AED. In this case the object found is a tree containing a node with a degree higher than 2.

9.1.2 Some Properties of Heuristics

The three examples in the preceding section are typical of a problem-solving method called “state-approach” (Nilsson, 1971), where the search for a solution to a posed problem is formulated as the search for a path in a state-space graph. A path to a given node in such a graph represents a code for a subset of potential solutions: the arcs represent transformations of those codes, which correspond to finer partitions of the parent subsets. In the 8-puzzle the transformations corresponded to the actual legal moves of the game. In the road map and the TSP the transformations consisted of concatenating a partially explored path with one more edge. Tasks such as theorem proving, robot planning, and speech recognition can naturally be represented as path-finding problems using the state-space



Figures 9.4a and 9.4b Graphs showing use of assignment problem heuristic for finding a cheapest tour in Traveling Salesman Problem.

approach. Even constraint-satisfaction problems, such as the 8-queens problem, which at first glance bear no mention of graphs or paths, can be formulated in state-space if we regard the computations which allow us to scan the space of possible objects systematically as arcs in a graph.

An algorithm known as A^* (Hart et al., 1968) has become popular in the Artificial Intelligence literature as an efficient way of using heuristics to solve path-finding problems. Unlike conventional shortest-path algorithms, the state-space graph is not available explicitly but rather is generated incrementally during the search itself using the transformation rules. A^* uses heuristic information to search the state-space graph in a directed fashion, making explicit only that portion of the graph which is absolutely necessary for finding an optimal solution.

We shall say that a node n is *expanded* when all possible transformation rules are applied to it and the resultant nodes, called successors of n , are generated. Any node which is expanded is called CLOSED, and any node generated but not yet expanded is called OPEN. At each step of the search, A^* selects for expansion that OPEN node which has the lowest cost estimate $f(n)$. The estimate $f(n)$ consists of two components:

$$f(n) = g(n) + h(n),$$

where $g(n)$ is the minimal cost so far encountered from the root to n , and $h(n)$ is an estimate of the cost required to complete the path from n to a goal state. A^* halts when it attempts to expand a node which satisfies the goal condition.

The most significant theoretical result regarding the behavior of A^* is its *admissibility* property: if for every node n , $h(n)$ does not exceed the actual optimal completion cost $h^*(n)$, then A^* terminates with the minimal cost path to a goal. An estimate $h(n)$ satisfying the inequality $h(n) \leq h^*(n)$ for every node in the graph is called an *admissibility heuristic*.

The second important property of A^* is called *consistency*. A heuristic function $h(\bullet)$ is said to be consistent if it satisfies the triangle inequality:

$$h(n) \geq c(n, n') + h(n') \leq h(n) \leq h(n')$$

and:

$$h(n_g) = 0$$

where n' is any successor of n , $c(n, n')$ is the cost of the edge from n to n' and n_g is any node satisfying the goal conditions. The importance of consistency lies in guaranteeing that A^* will never reopen a CLOSED node. The reason is that A^* , when it selects a node for expansion, has already traced the optimal path to that node

and so it need not test whether shorter paths may be found in the future. It can be shown that consistency implies admissibility but not vice versa.

The *power* of the heuristic estimate h is measured by the amount of pruning induced by h and depends, of course, on the accuracy of the estimate. If $h(\bullet)$ estimates the completion cost precisely, then A^* will only explore nodes lying along an optimal path. Otherwise A^* will expand any open node satisfying the inequality:

$$g(n) + h(n) < h^*(s)$$

where h^* is the cost of the optimal path from the initial node. Clearly, the higher the value of h the fewer nodes will be expanded by A^* , as long as h remains admissible. In the 8-puzzle example, for instance, since h_2 is generally larger and never lower than h_1 , it is a more powerful heuristic and will give rise to a more efficient search.

9.1.3 Where do these Heuristics Come from?

We have seen a few examples of heuristic functions which were devised by clever individuals to assist in the solution of combinatorial problems. We now focus our attention on the mental process by which these heuristics are “discovered,” with a view toward emulating the process mechanically.

The word “discovery” carries with it an aura of mystery, since it is normally attached to mental processes which leave no memory trace of their intermediate steps. It is an appropriate term for the process of generating heuristics, since tracing back the intermediate steps evoked in this process is usually a difficult task. For example, although we can argue convincingly that h_2 , the sum of the distances in the 8-puzzle, is an optimistic estimate of the number of moves required to achieve the goal, it is hard to articulate the mechanism by which this function was discovered or to invent additional heuristics of similar merit.

Articulating the rationale for one’s conviction in certain properties of human-devised heuristics may, however, provide clues as to the nature of the discovery process itself. Examining again the h_2 heuristic for the 8-puzzle, note how surprisingly easy it is to convince people that h_2 is admissible. After all, the formal definition of admissibility contains a universal qualifier ($\leq n$) which, at least in principle, requires that the inequality $h(n) \neq h^*(n)$ be verified for every node in the graph. Such exhaustive verification is, of course, not only impractical but also inconceivable. Evidently the mental process we employ in verifying such propositions is similar to that of symbolic proofs in mathematics, where the truth of universally quantified statements (say, that there are infinitely many prime numbers) is established by a sequence of inference rules applied to other statements (axioms) without exhaustive enumeration.

An even more surprising aspect of our conviction in the truth of $h_2(n) \leq h^*(n)$ is the fact that $h^*(n)$, by its very nature, is an unknown quantity for almost every node in the graph; not knowing $h^*(n)$ was the very reason for seeking its estimate $h(n)$. How can we, then, become so absolutely convinced in the validity of the assertion $h_2(n) \leq h^*(n)$? Clearly, the verification of this assertion performed in a code where $h^*(n)$ does not possess an explicit representation.

In the case of the road map problem our conviction the admissibility of the air distance heuristic is explainable. Here, based on our deeply entrenched knowledge of the properties of Euclidean spaces, we may argue that a straight line between any two points is shorter than any alternative connection between these points and, hence, that the air distance to the goal constitutes an admissible heuristic for the problem. In the 8-puzzle and the Traveling Salesman Problem, however, such a universal assertion cannot be drawn directly from our culture or experience, and must be defended, therefore, by more elaborate arguments based on more fundamental principles.

If we try to articulate the rationale for our confidence in the admissibility of h_2 for the 8-puzzle, we may encounter arguments such as the following:

1. Consider any solution to the goal, not necessarily an optimal one. In order to satisfy the goal conditions, each tile must trace some trajectory from its original location to its destination, and the overall cost (number of steps) of the solution is the sum of the costs of the individual trajectories. Every trajectory must consist of at least as many steps as that given by the Manhattan-distance between the tile's origin and its destination, and, hence, the sum of the distances cannot exceed the overall cost of the solutions.
2. If I were able to move each tile independently of the others, I would pick up tile #1 and move it, in steps, along the shortest path to its destination, do the same with tile #2, and so on until all tiles reach their goal locations. On the whole, I will have to spend at least as many steps as that given by the sum of the distances. The fact that in the actual game tiles tend to interfere with each other can only make things worse. Hence, ...

The first argument is analytical. It selects one property which must be satisfied by every solution, such as in providing a homeward-trajectory for every tile, and asks for the minimum cost required for maintaining just that property. The second argument is operational. It describes a procedure for solving a similar, auxiliary problem where the rules of the game have been relaxed. Instead of the conventional 8-puzzle whose tiles are kept confined in a 2-dimensional plane, we now imagine a relaxed puzzle whose tiles are permitted to climb on top of each other. Instead

of seeking heuristic *function* $h(n)$ to approximate $h^*(n)$, we can actually compute the solution using the relaxed version of the puzzle, count the number of steps required, and use this count as an estimate of $h^*(n)$.

The last scheme leads to the general paradigm expounded in this paper: *Heuristics are discovered by consulting simplified models of the problem domain*. We shall later explicate what is meant by a *simplified* model and how to go about finding such a one. The preceding example, however, specifically demonstrates the use of one important class of simplified models: that generated by *removing constraints* which forbid or penalize certain moves in the original problem. We shall call models obtained by such constraint-deletion processes *relaxed models*.

Let us examine first how the constraint-deletion scheme may work in the Traveling Salesman Problem. We are required to find an estimate for the cheapest completion path which starts at city D, ends at City A, and goes through every city in the set S of the unvisited cities. A path is a connected graph of degree 2, except for the end-points, which are of degree 1, a definition which we can express as a conjunction of three conditions: (1) being a graph, (2) being connected, (3) being of degree 2. If we delete the requirement that the completion graph be connected, we get the assignment heuristic of Figure 9.4a. Similarly, if we delete the constraint that the graph be of degree 2, we get the minimum-spanning-tree (MST heuristic depicted in Figure 9.4b). An even richer set of heuristics evolves by relaxing the condition that the task completed by a graph.

An alternative way of leading toward the MST heuristic is to imagine a salesman employed under the following cost arrangement: he has to pay from his own pocket for any trip in which he visits a city for the first time, but can get a free ride back to any city which he visited before. It is not hard to see that under such relaxed cost conditions the salesman would benefit from visiting some cities more than once and that the optimal tour strategy is to pay only for those trips which are part of the minimum-spanning-tree. If instead of this cost arrangement the salesman gets one free ride *from* any city visited twice, the optimal tour may be made up of smaller loops, and the assignment problem ensues. Thus, by adding free rides to the original cost structure we create relaxed problems whose solutions can be taken as heuristics for the original problem.

It is interesting to note that heuristics generated by optimizations over relaxed models are guaranteed to be *consistent*. This is easily verified by inspecting Figure 9.5, where $h(n)$ and $h(n')$ are the heuristics assigned to nodes n and n' , respectively. These heuristics stand for the minimum cost of completing the solution from the corresponding nodes in some relaxed model common to both nodes. $h(n)$, representing an optimal solution (geodesic), must satisfy $h(n) \leq c'(n, n') + h(n')$, where $c'(n, n')$ is the relaxed cost of the edge (n, n') , or else $c'(n, n') + h(n')$, instead

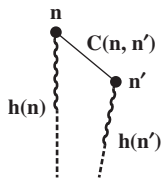


Figure 9.5 For consistent heuristic h , $h(n) \leq c(n, n') + h(n')$.

of $h(n)$, would constitute the optimal cost from n . The relaxed edge-cost $c'(n, n')$ cannot, by definition, exceed the original cost $c(n, n')$, thus:

$$h(n) \leq c(n, n') + h(n')$$

which, together with the technical condition $h(n_g) = 0$, completes the requirements for consistency.

This feature has both computational and psychological implications. Computationally, it guarantees that a search algorithm, A^* , guided by any heuristic evolving from a relaxed model, would be spared the effort of reopening CLOSED nodes or even testing whether a newly generated node has been expanded before. The pointers assigned to any node expanded by such algorithms are already directed along the optimal path to that node.

Psychologically, if we assume that people discover heuristics by consulting relaxed models, we can explain now why most man-made heuristics are both admissible and consistent. The reader may wish to test this point by attempting to generate heuristics for the 8-puzzle or the TSP that are admissible but not consistent. The difficulties encountered in such attempts should strengthen the reader's conviction that the most natural process for generating heuristics is by relaxation, where consistency surfaces as an automatic bonus.

Before proceeding to outline how systematic relaxations can be used to generate heuristics mechanically, it is important to note that not every relaxed model is automatically simpler than the original. Assume, for example, that in the 8-puzzle, in addition to the conventional moves, we also allow a checker-like jumping of tiles across the main diagonals. The puzzle thus created is obviously more relaxed than the original, but it is not at all clear that the complexity of searching for an optimal solution in the relaxed puzzle is lower than that associated with the original problem. True, adding shortcuts makes the search graph of the relaxed model somewhat shallower, yet it also becomes bushier: A^* must now examine the extra moves available at every decision junction, even if they lead nowhere.

Moreover, relaxation is not the only scheme which may simplify problems. In fact, simplified models can easily be obtained by a process opposite to that

of relaxation, such as loading the original model with additional constraints. Of course, the solutions to over-constrained models would no longer be admissible. However, in problems where one settles for finding any path to the goal, not necessarily the optimal, simplified over-constrained models may be very helpful. The most popular way of constraining models is to assume that a certain portion of the solution is given a-priori. This assumption cuts down on the number of remaining variables which of course, results in a speedy search for the completion portion. For example, if one arbitrarily selects an initial subtour through M cities in the TSP, an overconstrained problem ensues which is simpler than the original because it involves only $N - M$ cities. The cost associated with the solution to such a problem constitutes an upper bound to h^* and can be used to cut down the storage requirement of A^* . Every node is OPEN whose admissible evaluation $f(n) = g(n) + h(n)$ exceeds that upper bound can be permanently removed from memory without endangering the optimality of the resulting solution.

A third important class of simplified models can be obtained by probabilistic considerations. In certain cases we may possess sufficient knowledge about the problem domain to permit an estimate of the *most probable* cost of the completion path. Consider, for example, the problem of finding the cheapest cost path in a tree where all arc costs are known to be drawn independently from a common distribution function, with mean μ . If N stands for the number of arcs remaining between a node n and the goal, then for large N the cost of any path to the goal is known to be highly peaked about μN . Therefore, if we are sure that only one path leads from n to the goal, we can take the value μN as an estimate of h^* and use $f - g + \mu N$ as a node-rating function in A^* . If several paths lead from n to the goal set and if the structure of these paths is fairly regular, probability calculus can be invoked to estimate the most likely cost of the cheapest one among them, and that estimate can be used as h in A^* . Alternatively, probabilistic models often predict that reasonable solution paths are likely to exhibit certain distinctive properties in their behavior, e.g. that the cost along the path will increase gradually at a predetermined rate. Hence, an irrevocable search strategy can be employed which prunes away any path found to behave at variance with such expectations (Karp and Pearl, 1983).

By their very nature, probability-based models cannot guarantee the optimality of the solution. Although, in most cases, they produce accurate estimates of the completion cost, occasionally these turn out to be grossly overestimated, which may cause A^* to terminate prematurely with a suboptimal solution. On the other hand, if one does not insist on finding an exact optimal solution all the time but settles instead for finding a “good” solution most of the time, probability-based heuristics can, in many cases, reduce the complexity of combinatorial problems from exponential to polynomial.

Another class of simplified models used in heuristic reasoning is analogical or metaphorical models. Here the auxiliary model draws its power not from a structural simplicity inherent in the *problem*, but rather from matching the machinery and expertise accumulated by the *problem solver*. For example, the game of tic-tac-toe appears simpler to us than its isomorphic number-scrabble game (Newell and Simon, 1972) because the former evokes an expertise gathered by our visual machinery which has not yet been acquired by our arithmetic reasoner. Likewise, the use of visual imagery in solving complex mathematical or programming problems takes advantage of the special purpose machinery which evolution has bestowed upon us for processing visual information and manipulating physical objects. The use of analogical models by computers would only be beneficial when we learn how to build an efficient data-driven expert system for at least one problem domain of sufficient richness, e.g., physical objects.

9.2 Mechanical Generation of Admissible Heuristics

We return now to the relaxation scheme and to show how the deletion of constraints can be systematic to the point that natural heuristics, such as those demonstrated in the Introduction Section, can be generated by mechanical means. The constraint-relaxation scheme is particularly suitable for this purpose because many problem domains are conveniently formalizable by the explicit representation of the constraints which govern the applicability and impact of the various transformations in that domain. Take, for instance, the 8-puzzle problem. It is utterly impractical to specify the set of legal moves by an exhaustive list of pairs, describing the states before and after the application of each move. A much more natural representation of the puzzle would specify the available moves by two sets of conditions, one which must hold true before a given move is applicable and one which must prevail after the move is applied. In the robot-planning program STRIPS (Fikes and Nilsson, 1971), for example, actions are represented by three lists: (1) a *precondition-list*, a conjunction of predicates which must hold true before the action can be applied; (2) an *add-list*, a list of predicates which are to be added to the description of the world-state as a result of applying the action; and (3) a *delete-list*, a list of predicates that are no longer true once the action is applied and should, therefore, be deleted from the state description. We shall use this representation in formalizing the relaxation scheme for the 8-puzzle.

We start with a set of three primitive predicates:

- ON(x, y) : tile x is on cell y
- CLEAR(y) : cell y is clear of tiles
- ADJ(y, z) : cell y is adjacent to cell z ,

where the variable x is understood to stand for the tiles X_1, X_2, \dots, X_8 and where the variables y and z range over the set of cells C_1, C_2, \dots, C_9 . Although the predicate CLEAR can be defined in terms of ON:

$$\text{CLEAR}(y) \Leftrightarrow \forall(x) \sim \text{ON}(x, y),$$

it is convenient to carry CLEAR explicitly as if it were an independent predicate. Using these primitives, each state will be described by a list of 9 predicates, such as:

$$\text{ON}(X_1, C_1), \text{ON}(X_2, C_2), \dots, \text{ON}(X_8, C_8), \text{CLEAR}(C_9),$$

together with the board configuration:

$$\text{ADJ}(C_1, C_2), \text{ADJ}(C_1, C_4), \dots,$$

The move corresponding to transferring tile x from location y to location z will be described by the three lists:

MOVE (x, y, z) :

precondition list : ON (x, y) , CLEAR (z) , ADJ (y, z)

add list : ON (x, y) , CLEAR (y)

delete list : ON (x, y) , CLEAR (z)

The problem is defined as finding a sequence of applicable instantiations for the basic operator MOVE (x, y, z) that will transform the initial state into a state satisfying the goal criteria.

Let us now examine the effect of relaxing the problem by deleting the two conditions, CLEAR (z) and ADJ (y, z) , from the precondition list. The resultant puzzle permits each tile to be taken from its current position and be placed on any desired cell with one move. The problem can be readily solved using a straightforward control scheme: at any state find any tile which is not located on the required cell. Let this tile be X_1 , its current location Y_1 , and its required location Z_1 . Apply the operator MOVE (X_1, Y_1, Z_1) , and repeat the procedure on the prevailing state until all tiles are properly located.

Clearly, the number of moves required to solve any such problem is exactly the number of tiles which are misplaced in the initial state. If one submits this relaxed problem to a mechanical problem-solver and counts the number of moves required, the heuristic $h_1(\bullet)$ ensues.

Imagine now that instead of deleting the two conditions, only CLEAR(z) is deleted while ADJ(y, z) remains. The resultant model permits each tile to be moved into an adjacent location regardless of its being occupied by another tile. This obviously leads to the $h_2(\bullet)$ heuristic: the sum of the Manhattan-distances.

The next deletion follows naturally; let us retain the condition CLEAR(z) and delete ADJ(y, z). The resultant model permits transferring any tile to the empty spot, even when the two cells are not adjacent. The problem of reconfiguring the initial state with such operators is equivalent to that of sorting a list of elements by swapping the locations of two elements at a time, where every swap must exchange one marked element (the blank) with some other element. The optimal solution to this swap-sort problem can be obtained using the following “greedy” algorithm:

If the current empty cell y is to be covered by tile x , move x into y . Otherwise (if y is to remain empty in the goal state), move into y any arbitrary misplaced tile. Repeat.

The resulting cost of this model, h_3 , is mentioned neither in the Introductory Section 1 nor in textbooks on heuristic search. It is not the kind of heuristic that is likely to be discovered by the novice, and it was first introduced by Gaschnig (1979) eleven years after A^* was exemplified using h_1 and h_2 . Although h_3 turns out to be only slightly better than h_1 , its late discovery, coupled with the fact that it evolves so naturally from the constraint-deletion scheme, illustrates that the method of systematic deletions is capable of generating non-trivial heuristics.

The reader may presume that the space of deletions is now exhausted; deleting ON(x, y) leads again to h_1 , while retaining all three conditions brings us back to the original problem. Fortunately, the space of deletions can be further refined by enriching the set of elementary predicates. There is no reason, for instance, why the relation ADJ(y, z) need be taken as elementary – one may wish to express this relation as a conjunction of two other relations:

$$\text{ADJ}(y, z) \Leftrightarrow \text{NEIGHBOR}(y, z) \wedge \text{SAME-LINE}(y, z)$$

Deleting any one of these new relations will result in a new model, closer to the original than that created by deleting the entire ADJ predicate.

Thus, if we equip our program with a large set of predicates or with facilities to generate additional predicates, the space of deletions can be refined progressively, each refinement creating new problems closer and closer to the original. The resulting problems, however, may not lend themselves to easy solutions and may turn out to be even harder than the original problem. Therefore, the search for a model in the space of deletions cannot proceed blindly but must be directed toward finding a model which is both easy to solve and not too far from the original.

This begs the following question: Can the program tell an easy problem from a hard one without actually trying to solve them? This will be discussed in the next section.

9.3 Can a Program Tell an Easy Problem When It Sees One?

We now come to the key issue in our heuristic-generation scheme. We have seen how problem models can be relaxed with various degrees of refinements. We have also seen that relaxation without simplification is a futile excursion. Ideally, then, we would like to have a program that evaluates the degree of simplification provided by any candidate relaxation and uses this evaluation to direct the search in modelspace toward a model which is both simple and close to the original. This, however, may be asking for too much. The most we may be able to obtain is a program that recognizes a simple model when such a one happens to be generated by some relaxation. Of course, we do not expect to be able to prove mechanically propositions such as “This class of problems cannot be solved in polynomial-time”. Instead, we should be able to recognize a subclass of easy problems, those possessing salient feature advertising their simplicity.

Most of the examples discussed above possess such features. They can be solved by “greedy,” hill-climbing methods without backtracking, and the feature that makes them amenable to such methods is their *decomposability*. Take, for instance, the most relaxed model for the 8-puzzle, where each tile can be lifted and placed on any cell with no restrictions. We know that this problem is simple, without actually solving it, by virtue of the fact that all the goal conditions, $ON(X_1, C_1), ON(X_2, C_2), \dots$, can be satisfied *independently* of each other. Each element in this list of subgoals can be satisfied by one operator without undoing the effect of previous operators and without affecting the applicability of future operators.

Similar conditions prevail in the model corresponding to h_2 , except where a sequence of operators is required to satisfy each of the goal conditions. The sequences, however, are again independent in both applicability and effects, a fact which is discernible mechanically from the formal specification of the model, since the subgoals themselves define the desired partition of the operators. Any operator whose add-list contains the predicate $ON(X_i, y)$ will be directed only toward satisfying the subgoal $ON(X_i, C_i)$ and can be proven to be non-interfering with any operator containing the predicate $ON(X_j, y)$ in its add-list.

Most automatic problem-solvers are driven by mechanisms which attempt to break down a given problem into its constituent subproblems as dictated by the goal description. For example, the General-Problem-Solver (Ernst and Newell, 1969) is controlled by “differences,” a set of features which make the goal different from the current state. The programmer has to specify, though, along what dimensions

these differences are measured, which difference are easier to remove, what operators have the potential of reducing each of the differences, and under what conditions each reduction operator is applicable. In STRIPS, most of these decisions are made mechanically on the basis of the 3-list description of operators. Actions are brought up for consideration by virtue of their add-list containing predicates which can bridge the gap between the desired goal and the current state. If the current state does not possess the conditions necessary for enacting a useful difference-reducing transformation, a new subgoal must be created to satisfy the missing conditions. Thus the complexity of this “end-means” strategy increases sharply when subgoals begin to interact with each other.

The simplicity of decomposable problems, on the other hand, stems from the fact that each of the subgoals can be satisfied independently of each other; thus the overall goal can be achieved in a time equal to the number of conjuncts in the goal description multiplied by the time required for satisfying a single conjunct in isolation. It is a version of the celebrated ‘divide-and-conquer’ principle, where the division is dictated by the primitive conjuncts defining the goal conditions. For example, in an $N \times N$ -puzzle we have N^2 conjuncts defining the goal state configuration, and the solution of each subproblem, namely finding the shortest path for one tile using a relaxed model with single-cell moves, can be obtained in $O(N^2)$ steps even by an uninformed, breadthfirst, algorithm. Thus the optimal solution for the overall relaxed $N \times N$ -puzzle can be obtained in $O(N^4)$ steps, which is substantially better than the exponential complexity normally encountered in the non-relaxed version of the $N \times N$ -puzzle.

The relaxed $N \times N$ -puzzle is an example of a complete independence between the subgoals, where an operator leading toward a given subgoal is neither hindered nor assisted by any operator leading toward another subgoal. Such complete independence is a rare case in practice and can only be achieved after deleting a large fraction of the applicability constraints. It turns out, however, that simplicity can also be achieved in much weaker forms of independence which we shall call *semi-decomposable* models.

Take, for example, the minimum-spanning-tree problem. If the goal is defined as a conjunction of $N-1$ conditions:

$$\text{CONNECTED (city } i, \text{ city } 1) \quad i = 2, 3, \dots, N$$

and each elementary operator consists of adding an edge between a connected and an unconnected city, we have a semi-decomposable structure. Even though no operator may undo the labor of previous operators or hinder the applicability of future operators, some degree of coupling remains since each operator *enables* a

different set of applicable operators. This form of coupling, which was not present in the relaxed 8-puzzle, may make the cost of the solution depend on the order in which the operators are applied. Fortunately, the MST problem possesses another feature, *commutativity*, which renders the greedy algorithm “cheapest-subgoal-first” optimal. *Commutativity* implies that the internal order at which a given set of operators is applied does not alter the set of operators applicable in the future. This property, too, should be discernible from the formal specification of the domain model and, once verified, would identify the “greedy” strategy which yields an optimal solution.

Another type of semi-decomposable problem is exemplified by the swap-sort model of the 8-puzzle where the subgoals interact with respect to both applicability and effects. Moving a given tile into the empty cell clearly disqualifies the applicability of all operators which move other tiles into that particular cell. Additionally, if at a certain stage the predicate specifying the correct position of the empty cell is already satisfied, it would be impossible to satisfy additional subgoals without first falsifying this predicate, hopefully on a temporary basis only.

In spite of these couplings, the feature which renders this puzzle simple, admitting a greedy algorithm, is the existence of a partial order on the subgoals and their associated operators such that the operators designated for any subgoal g may influence only subgoals of lower order than g , leaving all other subgoals unaffected. In our simple example, establishing the correct position of the blank is a subgoal of a higher order than all the other subgoals and should, therefore, be attempted last. To find such a partial order of subgoals from the problem specification is similar to finding a triangular connection matrix in GPS; programs for computing this task have been reported in the literature ([Ernst and Goldstein, 1982](#)).

9.4 Conclusions

This paper outlined a natural scheme for devising heuristics for combinatorial problems. First, the problem domain is formulated in terms of the 3-list operators which transform the states in the domain and the conditions which define the goal states. Next, the preconditions which limit the applicability of the operators are refined and partially deleted, and a relaxed model submitted to an evaluator which determines whether it is semi-decomposable. The space of all such deletions constitutes a meta-search-space of relaxed models from which the most restrictive semi-decomposable element is to be selected. The search starts either at the original model from which precondition conjuncts are deleted one at a time, or at the trivial model containing no preconditions to which restrictions are added sequentially until decomposability is destroyed. The model selected, together with

the “greedy” strategy which exploits its decomposability, constitutes the heuristic for the original domain.

Although the effort invested in searching for an appropriate relaxed model may seem heavy, the payoffs expected are rather rewarding. Once a simplified model is found, it can be used to generate heuristics for *all* instances of the original problem domain. For example, if our scheme is applied successfully to the TSP model, it may discover a $O(N^2)$ heuristic superior to (more constrained than) the MST. Such a heuristic will be applicable to every instance of the TSP problem and, when incorporated into A^* , will yield optimal TSP solutions in shorter times than the MST heuristic. We currently do not know if such heuristics exist. Equally challenging are routing problems encountered in communication networks and VLSI designs which, unlike the TSP, have not been the focus of a long theoretical research, but where the problem of devising effective heuristics remains, nonetheless, a practical necessity.

Future progress in this area hinges on developing techniques for recognizing simple, decomposable problems when such are present, and on manipulating the space of deletions systematically in order that such problem be, in fact, synthesized.

9.4.1 Bibliographical and Historical Remarks

The ideas expressed in the preceding sections have been developed independently by several people, including: J Gaschnig, M Somalvico, M. Valtorta, D. Kibler and myself.

The notion of viewing heuristics as information provided by simplified models was first communicated to me by Stan Rosenschein in 1979. Aside from its popular use in operations research (Lawler and Wood, 1966), (Held and Karp, 1971), the auxiliary problem approach was formally introduced to AI-type problems by Gaschnig (1979), Guida and Somalvico (1979) and Banerji (1980). Gaschnig described the spaces of auxiliary problems as “subgraphs” and “supergraphs” obtained by deleting or adding edges to the original problem graph. Guida and Somalvico use propositional representation of constraints similar to that of the “Mechanical Generation of Admissible Heuristics” section and propose the use of relaxed models for generating admissible heuristics. A slightly different formulation is also given in Kibler (1982).

Sacerdoti’s planning system ABSTRIPS (Sacerdoti, 1974) also uses constraints relaxation for creating simplified problem spaces. ABSTRIPS first synthesizes a global abstract plan, then searches for a detailed mode of its implementation. The abstract planning phase differs from the fully detailed one in that the operators invoked by the former lack some of the preconditions spelled out in the

latter. Although the program does not make explicit use of a numerical evaluation function, the search schedule is determined by progress achieved in the abstract planning phase and so, in effect, it can be thought of as being guided by the advice of a relaxed model.

Valtorta (1981) presents a proof of the consistency of relaxation-based heuristics, and an analysis of the overall complexity of searching both the original and the auxiliary problems. He shows that if the auxiliary problems are solved by the blind search, then the overall complexity will be worse than simply executing a breadth-first search on the original problem. This result emphasizes the importance of searching for decomposable structures where optimal solutions can be found by “greedy” algorithms without resorting to breadth-first search. The use of systematic deletions in search for decomposable problems was proposed by Pearl (1982) and is currently pursued at UCLA. Other approaches to automatic generation of heuristics are reported by Banerji (1980) and Ernst and Goldstein (1982).

References

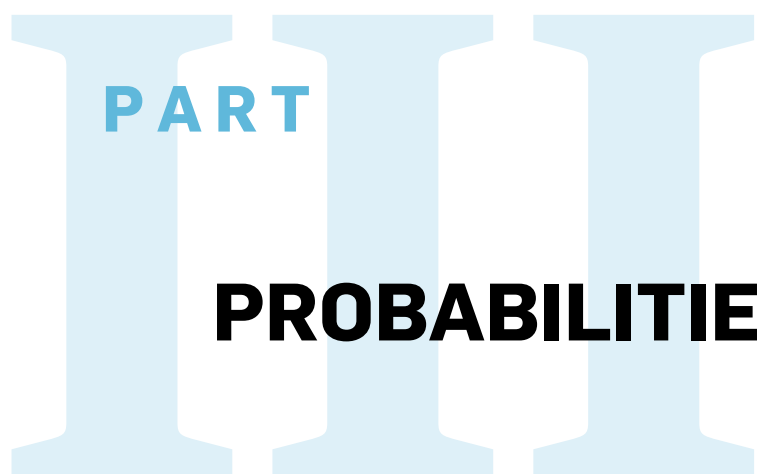
- Banerji, R. B. 1980. *Artificial Intelligence A Theoretical Approach*, Amsterdam: North Holland.
- Ernst, G. W. and Goldstein, M. M. 1982. Mechanical Discovery of Classes of Problem-Solving Strategies. *JACM*, 29(1):1–23.
- Ernst, G. W. and Newell, A. 1969. *GPS: A Case Study in Generality and Problem Solving*, New York: Academic Press.
- Fikes, R. E. and Nilsson, N. J. 1971. “STRIPS: A New Approach to the Application of Theorem Proving to Problem Solving.” *Artificial Intelligence*, 2:184–208.
- Gaschnig, J. 1979. A problem similarity approach to devising heuristics: First results. *Proc. IJCAI*, 6, 301–307.
- Guida, G. and Somalvico, M. 1979. A method for computing heuristics in problem solving. *Information Sciences*, 19:251–259.
- Hart, P., Nilsson, N., and Raphael, B. 1968. A formal basis for heuristic determination of minimum cost paths. *IEEE Trans. System Sci. Cybernetics*, 4:100–107.
- Held, M. and Karp, R. 1971. The traveling salesman problem and minimum spanning trees: Part II. *Mathematical Programming*, 1:6–25.
- Karp, R. M. and Pearl, J. 1983. Searching for the cheapest path in a tree with random costs. UCLA-ENG-CSL-8275 (To appear in *Artificial Intelligence*.) Since published: “Searching for an Optimal Path in a Tree with Random Costs,” *Artificial Intelligence*, 21(1–2):99–116, March 1983.
- Kibler, D. 1982. Natural generation of admissible heuristics. University of California at Irvine, Dept of Computer Science.
- Lawler, E. L. and Wood, D. E. 1966. Branch-and-bound methods: A survey. *Operations Research*, 14:699–719.
- Newell, A. and Simon, H. 1972. *Human Problem Solving*. New York:Prentice-Hall.

Nilsson, N. J. 1971. *Problem Solving Methods in Artificial Intelligence*. New York: McGraw-Hall.

Pearl, J. 1982. On the discovery and generation of certain heuristics. *UCLA Computer Science Quarterly*, Vol 2. Also in *AI Magazine*, Winter/Spring, 23–33, 1983.

Sacerdoti, E. D. 1974. Planning in a hierarchy of abstraction spaces *Artificial Intelligence*, 5:115–135.

Valtorta, M. 1981. A result on the computational complexity of heuristic estimates for the A* algorithm. Ph.D. thesis, Department of Computer Science, Duke University.



PART
PROBABILITIES

100

Introduction by Judea Pearl

“Reverend Bayes on inference engines” [Pearl 1982, Chapter 11] was my first paper on belief propagation using a Bayesian network. The title was chosen because I needed every bit of reverence I could muster to argue, in 1982, for the restoration of probabilistic methods in artificial intelligence (AI) systems, primarily for using Bayesian conditioning as a normative way of updating knowledge in light of new evidence.

The origin of these ideas and my struggle to gain acceptance in mainstream AI are narrated in the paper “A personal journey into Bayesian networks” [Pearl 2018], which also summarizes the main contenders to probability theory in the 1970s, and what gave me the “nerve” to argue against them.

The advantage of message passing architecture was key to its final acceptance, but was not appreciated at the time. When I presented it at Stanford in 1982, the audience could not understand why I emphasized computational issues instead of the accuracy of probability judgments (from experts). Little did we imagine that Bayesian networks would eventually be constructed with thousands of variables and operate coherently and effectively in many complex applications.

The article “Fusion propagation and structuring in belief networks” [Pearl 1986, Chapter 12] (hereafter *Fusion*) was the culmination of a series of papers¹ in which I explored the possibility of representing and manipulating probabilistic knowledge in graphical forms, later called *belief networks* (also known as *Bayesian networks*).

1. “Reverend Bayes on inference engines: A distributed hierarchical approach” [Pearl 1982]; “A computational model for causal and diagnostic reasoning in inference system” [Kim and Pearl 1983]; “How to do with probabilities what people say you can’t” [Pearl 1985a]; and “Bayesian networks: A model of self-activated memory for evidential reasoning” [Pearl 1985b].

I coined the name *Bayesian networks* in 1985 only when I was satisfied that an algorithm exists (called loop-cut conditioning in *Fusion*) that correctly updates probabilities in a network of arbitrary topology. The name *Bayesian networks* was chosen to emphasize three of their most important characteristics. (1) Allowing and even inviting subjective knowledge; (2) Employing Bayes conditionalization as a means for updating beliefs; and (3) Listening attentively to the asymmetry between cause and effect. All three of these ingredients shine like a beacon in Bayes' original paper of 1763.

On the anecdotal side, many readers were intrigued by the lengthy review process for *Fusion* (received January 1982; revised version received February 1986): It indeed took four years to get the article accepted, but the reviewers were not at fault. The article simply got lost (literally!) twice, which was not entirely without virtue; each time the editor (Patrick Hayes) asked me to replace a lost copy, I would seize the opportunity and send an improved version.

Readers now say it was worth the wait. *Fusion* turned out to be my most cited paper and has recently won the "Classic Paper Award" from the *Artificial Intelligence Journal* (2015). It is arguably the article that introduced probabilistic reasoning and graphical methods to AI, as well as to computer science in general.

The theory of "graphoids," described in the third paper, [Pearl and Paz 1987, Chapter 13] was conceived in the summer of 1985, when Azaria Paz visited UCLA and he and I began to explore what graphs and probabilities have in common. It led to a theory of dependence, that is, a set of axioms capturing the relation: "X is independent of Y given that we know Z," which also capture graph separation: "Vertex X is separated from vertex Y if we remove the set of vertices Z." These axioms connect graph with probabilities and with other systems of information dependency and, as such, confer legitimacy on Bayesian networks and other graphical representations.

"System Z," the fourth article in this section, [Pearl 1990, Chapter 14] represents my brief excursion into non-monotonic reasoning, which occupied a large circle of AI researchers in the 1980s. How can we represent the dual character of "beliefs" that, like logical propositions, have definitive truth values and, like probabilities, undergo changes when new evidence arrive? Anecdotally, I named it "System-Z" because the first sound I made after seeing the result was "Gee! It's easy!"

References

- J. Kim and J. Pearl. 1983. A computational model for combined causal and diagnostic reasoning in inference systems. In *Proceedings, IJCAI-83*, 190–193.
- J. Pearl. 1982. Reverend Bayes on inference engines: A distributed hierarchical approach. In *Proceedings, AAAI-82*. Pittsburgh, PA, 133–136.

- J. Pearl. 1985a. How to do with probabilities what people say you can't. In *Proceedings of the 2nd IEEE Conference on Artificial Intelligence Applications*. Also in Charles L. Wesibin (Ed.), *AI Applications*. North-Holland, Amsterdam, 6–12, 1988.
- J. Pearl. 1985b. Bayesian networks: A model of self-activated memory for evidential reasoning. In *Proceedings of the 7th Conference of the Cognitive Science Society*, 329–334.
- J. Pearl. 1986. Fusion, propagation and structuring in belief networks. *Artificial Intelligence*, 29, 241–288. DOI: [https://doi.org/10.1016/0004-3702\(86\)90072-X](https://doi.org/10.1016/0004-3702(86)90072-X).
- J. Pearl and A. Paz. 1987. Graphoids: A graph-based logic for reasoning about relevance relations. In B. du Boulay et al. (Eds.), *Advances in Artificial Intelligence II*. North-Holland.
- J. Pearl. 1990. System Z: A natural ordering of defaults with tractable applications to nonmonotonic reasoning. In R. Parikh (Ed.), *Theoretical Aspects on Reasoning about Knowledge*. Morgan Kaufmann, San Mateo, CA, 121–135.
- J. Pearl. 2018. *A Personal Journey into Bayesian Network*. UCLA Computer Science Department, Technical Report R-476, May 2018.

Reverend Bayes on Inference Engines: A Distributed Hierarchical Approach

Judea Pearl

Abstract

This paper presents generalizations of Bayes likelihood-ratio updating rule which facilitate an asynchronous propagation of the impacts of new beliefs and/or new evidence in hierarchically organized inference structures with multi-hypotheses variables. The computational scheme proposed specifies a set of belief parameters, communication messages and updating rules which guarantee that the diffusion of updated beliefs is accomplished in a single pass and complies with the tenets of Bayes calculus.

11.1 Introduction

This paper addresses the issue of efficiently propagating the impact of new evidence and beliefs through a complex network of hierarchically organized inference

The paper “An Essay Towards Solving a Problem in the Doctrine of Chances by the late Rev. Mr. Bayes”, Phil. Trans. of Royal Soc., 1763, marks the beginning of the science of inductive reasoning. Supported in part by the National Science Foundation, Grant IST 80 19045.

Pearl, J., “Reverend Bayes on Inference Engines: a Distributed Hierarchical Approach,” Originally published in *Proceedings, AAAI National Conference on AI*, Pittsburgh, PA, 133-136, August 1982.

From: AAAI-82 Proceedings. Copyright © 1982, AAAI (www.aaai.org). Republished with permission from AAAI.

rules. Such networks find wide applications in expert-systems [1], [2], [3], speech recognition [4], situation assessment [5], the modelling of reading comprehension [6] and judicial reasoning [7].

Many AI researchers have accepted the myth that a respectable computational model of inexact reasoning must distort, modify or ignore at least some principles of probability calculus. Consequently, most AI systems currently employ ad-hoc belief propagation rules which may hinder both the inferential power of these systems and their acceptance by their intended users. The primary purpose of this paper is to examine what computational procedures are dictated by traditional probabilistic doctrines and whether modern requirements of local asynchronous processing render these doctrines obsolete.

We shall assume that beliefs are expressed in probabilistic terms and that the propagation of beliefs is governed by the traditional Bayes transformations on the relation $P(D|H)$, which stands for the judgmental probability of data D (e.g., a combination of symptoms) given the hypothesis H (e.g., the existence of a certain disease). The unique feature of hierarchical inference systems is that the relation $P(D|H)$ is computable as a cascade of local, more elementary probability relations involving intervening variables. Intervening variables, (e.g., organisms causing a disease) may or may not be directly observable. Their computational role, however, is to provide a conceptual summarization for loosely coupled subsets of observational data so that the computation of $P(H|D)$ can be performed by local processes, each employing a relatively small number of data sources.

The belief maintenance architecture proposed in this paper is based on a distributed asynchronous interaction between cooperating knowledge sources without central supervision similar to that used in the HEARSAY system [4]. We assume that each variable (i.e., a set of hypotheses) is represented by a separate processor which both maintains the parameters of belief for the host variable and manages the communication links to and from the set of neighboring, logically related variables. The communication lines are assumed to be open at all times, i.e., each processor may at any time interrogate its message-board for revisions made by its neighbors, update its own belief parameters and post new messages on its neighbors' boards. In this fashion the impact of new evidence may propagate up and down the network until equilibrium is reached.

The asynchronous nature of this model requires a solution to an instability problem. If a stronger belief in a given hypothesis means a greater expectation for the occurrence of a certain supporting evidence and if, in turn, a greater certainty in the occurrence of that evidence adds further credence to the hypothesis,

how can one avoid an infinite updating loop when the two processors begin to communicate with one another? Thus, a second objective of this paper is to present an appropriate set of belief parameters, communication messages and updating rules which guarantee that the diffusion of updated beliefs is accomplished in a single pass and complies with the tenets of Bayes calculus.

A third objective is to demonstrate that proper Bayes inference can be accomplished among multi-valued variables and that, contrary to the claims made by Pednault, Zucker and Muresan [8], this does not render conditional independence incompatible with the assumption of mutual exclusivity and exhaustivity.

11.2 Definitions and Nomenclature

A node in an inference net represents a variable name. Each variable represents a finite partition of the world given by the variable values or states. It may be a name for a collection of hypotheses (e.g., identity of organism: ORG_1, ORG_2, \dots) or for a collection of possible observations (e.g., patient's temperature: high, medium, low). Let a variable be labeled by a capital letter, e.g., A, B, C, \dots , and its various states subscripted, e.g., A_1, A_2, \dots

An inference net is a directed acyclical graph where each branch $\textcircled{A} \rightarrow \textcircled{B}$ represents a family of rules of the form: if A_i then B_j . The uncertainties in these rules are quantified by a conditional probability matrix, $\underline{M}(B|A)$, with entries: $M(B|A)_{ij} = P(B_j|A_i)$. The presence of a branch between A and B signifies the existence of a direct communication line between the two variables. The directionality of the arrow designates A as the set of hypotheses and B as the set of indicators or manifestations for these hypotheses. We shall say that B is a son of A and confine our attention to trees, where every node has only one multi-hypotheses father and where the leaf nodes represent observable variables.

In principle, the model can also be generalized to include some graphs (multiple parents), keeping in mind that the states of each variable in the tree may represent the power set of multi-parent groups in the corresponding graph.

11.3 Structural Assumptions

Consider the following segment of the tree (Figure 11.1): The likelihood of the various states of B would, in general, depend on the entire data observed so far, i.e., data from the tree rooted at B , the tree rooted at C and the tree above A . However, the fact that B can communicate directly only with its father (A) and its sons (F and E) means that the influence of the entire network above B on B

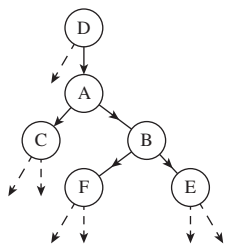


Figure 11.1 A Bayesian tree.

is completely summarized by the likelihood it induces on the states of A. More formally, let $D_d(B)$ stand for the data obtained from the tree rooted at B, and $D^u(B)$ for the data obtained from the network above B. The presence of only one link connecting $D^u(B)$ and (B) implies:

$$P(B_j | A_i, D^u(B)) = P(B_j | A_i) \quad (11.1)$$

This structural assumption of local communication immediately dictates what is normally called “Conditional Independence”; if C and B are siblings and A is their parent, then

$$P(B_j, C_k | A_i) = P(B_j | A_i) \cdot P(C_k | A_i) \quad (11.2)$$

because the data $C = C_k$ is part of $D^u(B)$ and hence (11.1) implies $P(B_j | C_k, A_i) = P(B_j | A_i)$, from which (11.2) follows.

Note the difference between the weak form of conditional independence in (11.2) and the over-restrictive form adapted by Pednault et al. [8], who also asserted independence with respect to the complements \bar{A}_i .

11.4 Combining Top and Bottom Evidences

Our structural assumption (11.1) also dictates how evidences above and below some variable B should be combined. Assume we wish to find the likelihood of the states of B induced by some data D, part of which, $D^u(B)$, comes from above B and part, $D_d(B)$, from below. Bayes theorem, together with (11.1), yields the product rule:

$$P(B_i | D^u(B), D_d(B)) = \alpha P[D_d(B) | B_i] \cdot P[B_i | D^u(B)], \quad (11.3)$$

where α is a normalization constant. This is a generalization of the celebrated Bayes formula for binary variables:

$$O(H | E) = \lambda(E)O(H) \quad (11.4)$$

where $\lambda(E) = P(E|H)/P(E|\bar{H})$ is known as the likelihood ratio, and $O(H) = P(H)/P(\bar{H})$ as the prior odds [2].

Equation (11.3) generalizes (11.4) in two ways. First, it permits the treatment of non-binary variables where the mental task of estimating $P(E|\bar{H})$ is often unnatural, and where conditional independence with respect to the negations of the hypotheses is normally violated (i.e., $P(E_1, E_2|\bar{H}) \neq P(E_1|\bar{H})P(E_2|\bar{H})$). Second, it identifies a surrogate to the prior probability term for any intermediate node in the tree, even after obtaining some evidential data. According to (11.3), the multiplicative role of the prior probability in Equation (11.4) is taken over by the conditional probability of a variable based only on the evidence gathered by the network above it, excluding the data collected from below. Thus, the product rule (11.3) can be applied to any node in the network, without requiring prior probability assessments.

The root is the only node which requires a prior probability estimation. Since it has no network above, $D^u(B)$ should be interpreted as the available background knowledge which remains unexplicated by the network below. This interpretation renders $P(B_i|D^u(B))$ identical to the classical notion of subjective prior probability. The probabilities of all other nodes in the tree are uniquely determined by the arc-matrices, the data observed and the prior probability of the root.

Equation (11.3) suggests that the probability distribution of every variable in the network can be computed if the node corresponding to that variable contains the parameters

$$\lambda(B_i) \triangleq P(D_d(B)|B_i) \quad (11.5)$$

and

$$q(B_i) \triangleq P(B_i|D^u(B)). \quad (11.6)$$

$q(B_i)$ represents the anticipatory support attributed to B_i by its ancestors and $\lambda(B_i)$ represents the evidential support received by B_i from its diagnostic descendants. The total strength of belief in B_i would be given by the product

$$P(B_i) = \alpha\lambda(B_i)q(B_i). \quad (11.7)$$

Whereas only two parameters, $\lambda(E)$ and $O(H)$, were sufficient for binary variables, an n-state variable needs to be characterized by two n-tuples:

$$\begin{aligned} \underline{\lambda}(B) &= \lambda(B_1), \lambda(B_2), \dots, \lambda(B_n) \\ \underline{q}(B) &= q(B_1), q(B_2), \dots, q(B_n). \end{aligned}$$

11.5 Propagation of Information Through the Network

Assuming that the vectors $\underline{\lambda}$ and \underline{q} are stored with each node of the network, our task is now to prescribe how the influence of new information spreads through the network. Traditional probability theory, together with some efficiency considerations [9], dictate the following propagation scheme which we first report without proofs.

1. Each processor (Figure 11.2) computes two message vectors: \underline{p} and \underline{r} . \underline{p} is sent to every son while \underline{r} is delivered to the father. The message \underline{p} is identical to the probability distribution of the sender and is computed from $\underline{\lambda}$ and \underline{q} using Equation (11.7). \underline{r} is computed from $\underline{\lambda}$ using the matrix multiplication:

$$\underline{r} = \underline{M} \cdot \underline{\lambda} \tag{11.8}$$

where \underline{M} is the matrix quantifying the link to the father. Thus, the dimensionality of \underline{r} is equal to the number of hypotheses managed by the father. Each component of \underline{r} represents the diagnostic contribution of the data below the host processor to the belief in one of the father's hypotheses.

2. When processor B is called to update its parameters, it simultaneously inspects the $\underline{p}(A)$ message communicated by the father A and the messages $\underline{r}_1, \underline{r}_2, \dots$, communicated by each of its sons and acknowledges receiving the latter. Using these inputs, it then updates $\underline{\lambda}$ and \underline{q} as follows:

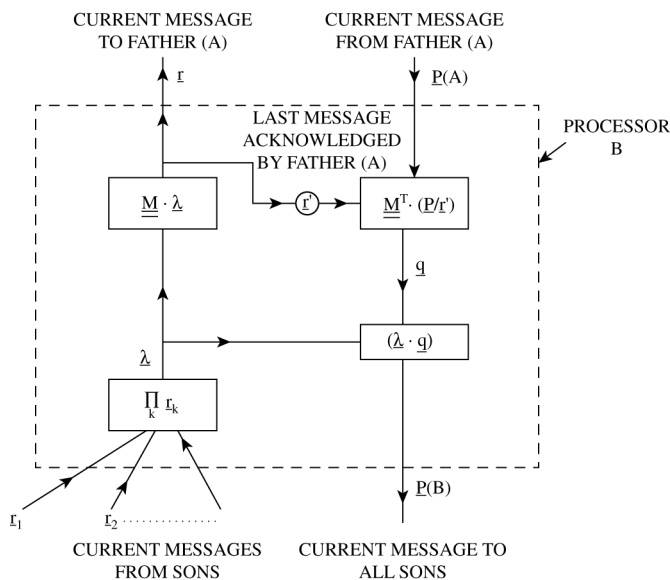


Figure 11.2 Node Processors.

3. Bottom-up propagation: $\underline{\lambda}$ is computed using a term-by-term multiplication of the vectors $\underline{r}_1, \underline{r}_2, \dots$:

$$\lambda(B_i) = (\underline{r}_1)_i \times (\underline{r}_2)_i \times \dots = \prod_k (\underline{r}_k)_i \quad (11.9)$$

4. Top-down propagation: \underline{q} is computed using:

$$q(B_i) = \beta \sum_j P(B_i | A_j) P(A_j) / (\underline{r}')_j \quad (11.10)$$

where β is a normalization constant and \underline{r}' is the last message from B to A acknowledged by the father A. (The division by \underline{r}' amounts to removing from $\underline{P}(A)$ the contribution due to $D_d(B)$ as dictated by the definition of \underline{q} in Equation (11.6)).

5. Using the updated values of $\underline{\lambda}$ and \underline{q} , the messages \underline{P} and \underline{r} are then recomputed as in step 1 and are posted on the message-boards dedicated for the sons and the father, respectively. This updating scheme is shown schematically in the diagram below, where multiplications and divisions of any two vectors stand for term-by-term operations.

The terminal nodes in the tree require special boundary conditions. Here we have to distinguish between the two cases:

1. Anticipatory node: an observable variable whose state is still unknown. For such variables, \underline{P} should be equal to \underline{q} and, therefore, we should set $\underline{\lambda} = (1, 1, \dots, 1)$ (also implying $\underline{r} = (1, 1, \dots, 1)$).
2. Data-node: an observable variable with a known state. Following Equation (11.5), if the j^{th} state of B was observed to be true, set $\underline{\lambda} = (0, 0 \dots 0, 1, 0 \dots)$ with 1 at the j^{th} position.

Similarly, the boundary conditions for the root node is obtained by substituting the prior probability instead of the message $\underline{P}(A)$ expected from the father.

11.6 A Token Game Illustration

Figure 11.3 shows six successive stages of belief propagation through a simple binary tree, assuming that updating is activated by changes in the belief parameters of neighboring processes. Initially (Figure 11.3a), the tree is in equilibrium and all terminal nodes are anticipatory. As soon as two data nodes are activated (Figure 11.3b), white tokens are placed on their links, directed towards their fathers. In the next phase, the fathers, activated by these tokens, absorb the latter and manufacture the appropriate number of tokens for their neighbors (Figure 11.3c), white tokens for their fathers and black ones for the children (the links through which

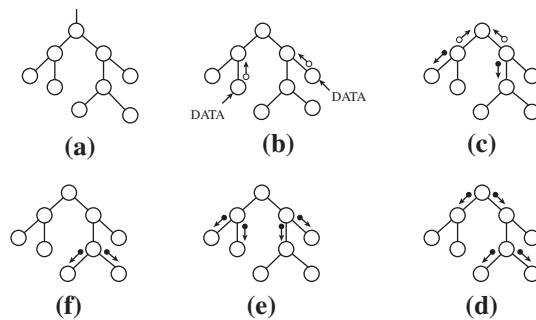


Figure 11.3 The message passing scheme in action.

the absorbed tokens have entered do not receive new tokens, thus reflecting the division of \underline{p} by \underline{r}). The root node now receives two white tokens, one from each of its descendants. That triggers the production of two black tokens for top-down delivery (Figure 11.3d). The process continues in this fashion until, after six cycles, all tokens are absorbed and the network reaches a new equilibrium.

11.7 Properties of the Updating Scheme

1. The local computations required by the proposed scheme are efficient in both storage and time. For an m -ary tree with n states per node, each processor should store $n^2 + mn + 2n$ real numbers, and perform $2n^2 + mn + 2n$ multiplications per update. These expressions are on the order of the number of rules which each variable invokes.
2. The local computations are entirely independent of the control mechanism which activates the updating sequence. They can be activated by either data-driven or goal driven (e.g., requests for evidence) control strategies, by a clock or at random.
3. New information diffuses through the network in a single pass. Infinite relaxations have been eliminated by maintaining a two-parameter system (\underline{q} and \underline{r}) to decouple top and bottom evidences. The time required for completing the diffusion (in parallel) is equal to the diameter of the network.

11.8 A Summary of Proofs

From the fact that $\underline{\lambda}$ is only influenced by changes propagating from the bottom and \underline{q} only by changes from the top, it is clear that the tree will reach equilibrium after a finite number of updating steps. It remains to show that, at equilibrium, the updated parameters $P(V_i)$, in every node V , correspond to the correct probabilities $P(V_i | D^u(V), D_d(V))$ or (see Equation (11.3)), that the equilibrium values of $\lambda(V_i)$

and $q(V_i)$ actually equal the probabilities $P(D_d(V) | V_i)$ and $P(V_i | D^u(V))$. This can be shown by induction bottom-up for $\underline{\lambda}$ and then top-down for \underline{q} .

Validity of $\underline{\lambda}$: $\underline{\lambda}$ is certainly valid for leaf nodes, as was explained above in setting the boundary conditions. Assuming that the $\underline{\lambda}$'s are valid at all children of node B, the validity of $\underline{\lambda}(B)$ computed through steps (11.8) and (11.9) follows directly from the conditional independence of the data beneath B's children (Equation (11.2)).

Validity of \underline{q} : if all the $\underline{\lambda}$'s are valid, then \underline{P} is valid for the root node. Assuming now that $\underline{P}(A)$ is valid, let us examine the validity of $\underline{q}(B)$, where B is any child of A. By definition (Equation (11.6)), $\underline{q}(B)$ should satisfy:

$$q(B_i) = P(B_i | D^u(B)) = \sum_j P(B_i | A_j) P(A_j | D^u(A), D_d(S))$$

where S denotes the set of B's siblings. The second factor in the summation differs from $P(A_j) = P(A_j | D^u(A), D_d(A))$ in that the latter has also incorporated B's message (r'_j) in the formation of $\lambda(A_j)$ (Equation (11.9)). When we divide $P(A_j)$ by (r'_j) , as prescribed in (11.10), the correct probability ensues.

11.9 Conclusions

The paper demonstrates that the centuries-old Bayes formula still retains its potency for serving as the basic belief revising rule in large, multi-hypotheses, inference systems. It is proposed, therefore, as a standard point of departure for more sophisticated models of belief maintenance and inexact reasoning.

References

- [1] Shortliffe, E.H., and Buchanan, B.G., "A Model of Inexact Reasoning in Medicine". *Math.Biosci.*, 23 (1975), 351-379.
- [2] Duda, R.O., Hart, P.E. and Nilsson, N.J., "Subjective Bayesian Methods for Rule-Based Inference Systems". Tech. Note 124, AI Center, SRI International, Menlo Park, CA; also *Proc. 1976 NCC (AFIPS Press)*.
- [3] Duda, R., Hart, P., Barrett, P., Gashnig, J., Konolige, K., Reboh, R. and Slocum J., "Development of the Prospector Consultation System for Mineral Exploration". AI Center, SRI International, Menlo Park, CA, Sept. 1976.
- [4] Lesser, V.R. and Erman, L.D., "A Retrospective View of HEARSAY II Architecture". *Proc. 5th Int. Joint Conf. AI, Cambridge, MA, 1977*, 790-800.
- [5] DDI Handbook for Decision Analysis, Decision and Design Inc., McLean, VA, 1973.
- [6] Rumelhart, D.E., "Toward an Interactive Model of Reading". *Center for Human Info. Proc. CHIP-56*, UC La Jolla, March 1976.

- [7] Schum, D. and Martin, A., "Empirical Studies of Cascaded Inference in Jurisprudence: Methodological Consideration". Rice Univ., Psychology Research Report, #80-01, May 1980.
- [8] Pednault, E.P.D., Zucker, S.W. and Muresan, L.V., "On the Independence Assumption Underlying Subjective Bayesian Updating". *Art. Intel.*, Vol. 16, No. 2, May 1981, 213-222.
- [9] Pearl, J., "Belief Propagation in Hierarchical Inference Structures". UCLA-ENG-CSL-8211, UC Los Angeles, January 1982.



Fusion, Propagation, and Structuring in Belief Networks

Judea Pearl

Abstract

Belief networks are directed acyclic graphs in which the nodes represent propositions (or variables), the arcs signify direct dependencies between the linked propositions, and the strengths of these dependencies are quantified by conditional probabilities. A network of this sort can be used to represent the generic knowledge of a domain expert, and it turns into a computational architecture if the links are used not merely for storing factual knowledge but also for directing and activating the data flow in the computations which manipulate this knowledge.

The first part of the paper deals with the task of fusing and propagating the impacts of new information through the networks in such a way that, when equilibrium is reached, each proposition will be assigned a measure of belief consistent with the axioms of probability theory. It is shown that if the network is singly connected (e.g. tree-structured), then probabilities can be updated by local propagation in an isomorphic network of parallel and autonomous processors and that the impact of new information can be imparted to all propositions in time proportional to the longest path in the network.

The second part of the paper deals with the problem of finding a tree-structured representation for a collection of probabilistically coupled propositions using auxiliary

This work was supported in part by the National Science Foundation, Grant #DSR 83-13875.

Recommended by Patrick Hayes

Originally published in *Artificial Intelligence* 29 (1986) 241–288

0004-3702/86/\$3.50 © 1986, Elsevier Science Publishers B.V. (North-Holland). Republished with permission of Elsevier.

Original DOI: [10.1016/0004-3702\(86\)90072-X](https://doi.org/10.1016/0004-3702(86)90072-X)

(dummy) variables, colloquially called “hidden causes.” It is shown that if such a tree-structured representation exists, then it is possible to uniquely uncover the topology of the tree by observing pairwise dependencies among the available propositions (i.e., the leaves of the tree). The entire tree structure, including the strengths of all internal relationships, can be reconstructed in time proportional to $n \log n$, where n is the number of leaves.

12.1 Introduction

This study was motivated by attempts to devise a computational model for humans’ inferential reasoning, namely, the mechanism by which people integrate data from multiple sources and generate a coherent interpretation of that data. Since the knowledge from which inferences are drawn is mostly judgmental—subjective, uncertain and incomplete—a natural place to start would be to cast the reasoning process in the framework of probability theory. However, the mathematician who approaches this task from the vantage point of probability theory may dismiss it as a rather prosaic exercise. For, if one assumes that human knowledge is represented by a joint probability distribution, $P(x_1, \dots, x_n)$, on a set of propositional variables, x_1, \dots, x_n , the task of drawing inferences from observations amounts to simply computing the probabilities of a small subset, H_1, \dots, H_k , of variables called hypotheses, conditioned upon a group of instantiated variables, e_1, \dots, e_m , called evidence. Indeed, computing $P(H_1, \dots, H_k | e_1, \dots, e_m)$ from a given joint distribution on all propositions is merely arithmetic tedium, void of theoretical or conceptual interest.

It is not hard to see that this textbook view of probability theory presents a rather distorted picture of human reasoning and misses its most interesting aspects. Consider, for example, the problem of encoding an arbitrary joint distribution, $P(x_1, \dots, x_n)$, on a computer. If we need to deal with n propositions, then to store $P(x_1, \dots, x_n)$ explicitly would require a table with 2^n entries—an unthinkably large number, by any standard. Moreover, even if we found some economical way of storing $P(x_1, \dots, x_n)$ (or rules for generating it), there would still remain the problem of manipulating it to compute the probabilities of propositions which people consider interesting. For example, computing the marginal probability $P(x_i)$ would require summing $P(x_1, \dots, x_n)$ over all 2^{n-1} combinations of the remaining $n-1$ variables. Similarly, computing the conditional probability $P(x_1 | x_j)$ from its textbook definition $P(x_1 | x_j) = P(x_1, x_j) / P(x_j)$ would involve dividing two marginal probabilities, each resulting from summation over an exponentially large number of variable combinations. Human performance, by contrast, exhibits a different complexity ordering: probabilistic judgments on a small number of propositions (especially

two-place conditional statements such as the likelihood that a patient suffering from a given disease will develop a certain type of complication) are issued swiftly and reliably, while judging the likelihood of a conjunction of many propositions entails a great degree of difficulty and hesitancy. This suggests that the elementary building blocks which make up human knowledge are not the entries of a joint-distribution table but, rather, the low-order marginal and conditional probabilities defined over small clusters of propositions.

Further light on the structure of probabilistic knowledge can be shed by observing how people handle the notion of independence. Whereas a person may show reluctance to giving a numerical estimate for a conditional probability $P(x_i|x_j)$, that person can usually state with ease whether x_i and x_j are dependent or independent, namely, whether or not knowing the truth of x_j will alter the belief in x_i . Likewise, people tend to judge the three-place relationships of conditional dependency (i.e., x_i influences x_j given x_k) with clarity, conviction, and consistency.

This suggests that the notions of dependence and conditional dependence are more basic to human reasoning than are the numerical values attached to probability judgments. (This is contrary to the picture painted in most textbooks on probability theory, where the latter is presumed to provide the criterion for testing the former.) Moreover, the nature of probabilistic dependency between propositions is similar in many respects to that of connectivity in graphs. For instance, we find it plausible to say that a proposition q affects proposition r *directly*, while s influences r *indirectly*, via q . Similarly, we find it natural to identify a set of direct justifications for q to sufficiently shield it (q) from all other influences and to describe them as the direct neighbors of q [5]. These graphical metaphors suggest that the fundamental structure of human knowledge can be represented by dependency graphs and that mental tracing of links in these graphs are the basic steps in querying and updating that knowledge.

12.1.1 Belief Networks

Assume that we decide to represent our perception of a certain problem domain by sketching a graph in which the nodes represent propositions and the links connect those propositions that we judge to be *directly* related. We now wish to quantify the links with weights that signify the strength and type of dependencies between the connected propositions. If these weights are to reflect summaries of actual experiences, we must first attend to two problems: *consistency* and *completeness*. Consistency guarantees that we do not overload the graph with an excessive number of parameters; overspecification may lead to contradictory conclusions, depending on which parameter is consulted first. Completeness

protects us from underspecifying the graph dependencies and guarantees that our conclusion-generating routine will not get deadlocked for lack of information.

One of the attractive features of the traditional joint-distribution representation of probabilities is the transparency by which one can synthesize consistent probability models or detect inconsistencies therein. In this representation, all we need to do to create a complete model, free of inconsistencies, is to assign nonnegative weights to the atomic compartments in the space (i.e., conjunctions of propositions), just making sure the sum of the weights equals one. By contrast, the synthesis process in the graph representation is more hazardous. For example, assume you have three propositional variables, x_1, x_2, x_3 , and you want to express their dependencies by specifying the three pairwise probabilities $P(x_1, x_2), P(x_2, x_3), P(x_3, x_1)$. It turns out that this will normally lead to inconsistencies; unless the parameters given satisfy some nonobvious relationship, there exists no probability model that will support all three inputs. By contrast, if we specify the probabilities on only two pairs, incompleteness results; many models exist which conform to the input specification, and we will not be able to provide answers to all probabilistic queries.

Fortunately, the consistency-completeness issue has a simple solution stemming from the chain-rule representation of joint distributions. Choosing an arbitrary order d on the variables x_1, \dots, x_n , we can write¹:

$$\begin{aligned} P(x_1, x_2, \dots, x_n) \\ = P(x_n | x_{n-1}, \dots, x_1) \cdots P(x_3 | x_2, x_1) P(x_2 | x_1) P(x_1). \end{aligned}$$

In this formula, each factor contains only one variable on the left side of the conditioning bar and, in this way, the formula can be used as a prescription for consistently quantifying the dependencies among the nodes of an arbitrary graph. Suppose we are given a directed acyclic graph G in which the arrows pointing at each node x_i emanate from a set S_i of parent nodes judged to be directly influencing x_i , and we wish to quantify the strengths of these influences in a complete and consistent way. If, by direct parents we mean a set of variables which, once we fix their values, would shield x_i from the influence of all other predecessors of x_i (i.e., $P(x_i | S_i) = P(x_i | x_1, \dots, x_{i-1})$), then the chain-rule formula states that a separate assessment of each child-parents relationship should suffice. We need only assess

1. Probabilistic formulae of this kind are shorthand notation for the statement that for any instantiation i of the variables x_1, x_2, \dots, x_n , the probability of the joint event $(x_1 = i_1) \& (x_2 = i_2) \& \dots \& (x_n = i_n)$ is equal to the product of the probabilities of the corresponding conditional events $(x_1 = i_1), (x_2 = i_2 \text{ if } x_1 = i_1), (x_3 = i_3 \text{ if } (x_2 = i_2 \& x_1 = i_1)), \dots$. For this expansion to be valid, we must require that $P(E) > 0$ for all conditioning events E .

the conditional probabilities, $P(x_i|S_i)$, by some functions, $F_i(x_i, S_i)$, and make sure these assessments satisfy

$$\sum_{x_i} F_i(x_i, S_i) = 1, \quad 0 \leq F_i(x_i, S_i) \leq 1,$$

where the summation ranges over all values of x_i . This specification is complete and consistent because the product form

$$P(x_1, \dots, x_n) = \prod_i F_i(x_i, S_i)$$

constitutes a joint probability distribution that supports the assessed quantities. In other words, if we compute the conditional probabilities $P(x_i|S_i)$ dictated by $P(x_1, \dots, x_n)$, the original assessments $F_i(x_i, S_i)$ will be recovered:

$$P(x_i|S_i) = \frac{P(x_i, S_i)}{P(S_i)} = \frac{\sum_{x_j \notin (x_i \cup S_i)} P(x_1, \dots, x_n)}{\sum_{x_j \notin S_i} P(x_1, \dots, x_n)} = F_i(x_i, S_i).$$

So, for example, the distribution corresponding to the graph of Figure 12.1 can be written by inspection:

$$\begin{aligned} P(x_1, x_2, x_3, x_4, x_5, x_6) \\ = P(x_6|x_5)P(x_5|x_2, x_3)P(x_4|x_1, x_2)P(x_3|x_1)P(x_2|x_1)P(x_1). \end{aligned}$$

This also leads to a simple method of constructing a dependency-graph representation for any given joint distribution $P(x_1, \dots, x_n)$. We start by imposing an arbitrary order d on the set of variables, x_1, \dots, x_n , then choose x_1 as a root of the graph and assign to it the marginal probability $P(x_1)$ dictated by $P(x_1, \dots, x_n)$. Next, we form

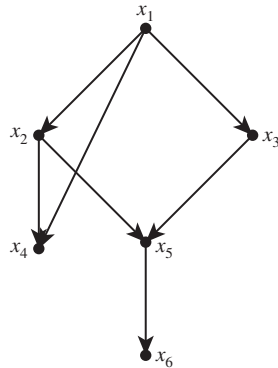


Figure 12.1 A typical Bayesian network representing the distribution $P(x_1, \dots, x_6) = P(x_6|x_5)P(x_5|x_2, x_3)P(x_4|x_1, x_2)P(x_3|x_1)P(x_2|x_1)P(x_1)$.

a node to represent x_2 ; if x_2 is dependent on x_1 , a link from x_1 to x_2 is established and quantified by $P(x_2|x_1)$. Otherwise, we leave x_1 and x_2 unconnected and assign the prior $P(x_2)$ to node x_2 . At the i th stage, we form the node x_i and establish a group of directed links to x_i from the smallest subset of nodes $S_i \subseteq \{x_1, \dots, x_{i-1}\}$ satisfying the condition

$$P(x_i|S_i) = P(x_i|x_{i-1}, \dots, x_1).$$

It can be shown that the set of subsets satisfying this condition is closed under intersection; therefore, the minimal subset S_i is unique. Thus, the distribution, $P(x_1, \dots, x_n)$, together with the order d uniquely identify a set of parent nodes for each variable x_i , and that constitutes a full specification of a directed acyclic graph which represents many of the independencies imbedded in $P(x_1, \dots, x_n)$.

In expert-systems applications where, instead of a numerical representation for $P(x_i, \dots, x_n)$, we have only intuitive understanding of the major constraints in the domain, the graph can still be configured by the same modular method as before, except that the parent set S_i must be selected judgmentally. The addition of any new node x_i to the network requires only that the expert identify a set S_i of variables which “directly influence” x_i , locally assess the strength of this relation and make no commitment regarding the effect of x_i on other variables, outside S_i . Even though each judgment is performed locally, their sum total is guaranteed to be consistent. This model-building process permits people to express qualitative relationships perceived to be essential, and the network preserves these qualities, despite sloppy assignments of numerical estimates. In Figure 12.1, for example, the fact that x_6 can tell us nothing new about x_3 once we know x_5 , will remain part of the model, no matter how carelessly the numbers are assigned.

Graphs constructed by this method will be called *belief networks*, *Bayesian networks*, or *influence networks* interchangeably, the former two to emphasize the judgmental origin and the probabilistic nature of the quantifiers, the latter to reflect the directionality of the links. When the nature of the interactions is perceived to be causal, then the term, *causal network*, may also be appropriate. In general, however, an influence network may also represent associative or inferential dependencies, in which case the directionality of the arrows mainly provides computational convenience [10]. An alternative graphical representation, using undirected graphs, is provided by the so-called Markov fields approach [12] and will not be discussed here. For comparison of properties and applications, see [15, 24, 32].

In the strictest sense, these networks are not graphs but hypergraphs because to describe the dependency of a given node on its k parents requires a function of

$k + 1$ arguments which, in general, could not be specified by k two-place functions on the individual links. This, however, does not diminish the advantages of the network representation because the essential interactions between the variables are still displayed by the connecting links. If the number of parents k is large, estimating $P(x_i|S_i)$ may be troublesome because, in principle, it requires a table of size 2^k . In practice, however, people conceptualize causal relationships by forming hierarchies of small clusters of variables (see Section 12.3.1) and, moreover, the interactions among the factors in each cluster are normally perceived to fall into one of a few prestored, prototypical structures, each requiring about k parameters. Common examples of such prototypical structures are: noisy OR gates (i.e., any one of the factors is likely to trigger the effect), noisy AND gates and various enabling mechanisms (i.e., factors identified as having no influence of their own except enabling other influences to become effective).

Note that the topology of a Bayes network can be extremely sensitive to the node ordering d ; a network with a tree structure in one ordering may turn into a complete graph if that ordering is reversed. For example, if x_1, \dots, x_n stands for the outcomes of n independent coins, and x_{n+1} represents the output of a detector triggered if any of the coins comes up head, then the influence network will be an inverted tree of n arrows pointing from each of the variables x_1, \dots, x_n toward x_{n+1} . On the other hand, if the detector's outcome is chosen to be the first variable, say x_0 , then the underlying influence network will be a complete graph.

This order sensitivity may at first seem paradoxical; d can be chosen arbitrarily, whereas people have fairly uniform conceptual structures, e.g., they agree on whether a pair of propositions are directly or indirectly related. The answer to this apparent paradox lies in the fact that the consensus about the structure of influence networks stems from the dominant role *causality* plays in the formation of these networks. In other words, the standard ordering imposed by the direction of causation indirectly induces identical topologies on the networks that people adopt for encoding experiential knowledge. It is tempting to speculate that, were it not for the social convention of adopting a standard ordering of events conforming to the flow of time and causation, human communication (as we now know it) would be impossible.

12.1.2 Conditional Independence and Graph Separability

To facilitate the verification of dependencies among the variables in a Bayes network, we need to establish a clear correspondence between the topology of the network and various types of independence. Normally, independence between variables connotes lack of connectivity between their corresponding nodes. Thus, it would be ideal to require that, should the removal of some subset S of nodes from

the network render nodes x_i and x_j disconnected, then such separation indicates genuine independence between x_i and x_j , conditioned on S :

$$P(x_i|x_j, S) = P(x_i|S).$$

This would provide a clear graphical representation for the notion that x_j does not affect x_i directly but, rather, its influence is mediated by the variables in S . Unfortunately, a network constructed to satisfy this correspondence for any arbitrary S would normally fail to display an important class of independencies [24]. For example, in such a network, two variables which are marginally independent will appear directly connected, merely because there exists some other variable that depends on both.

Bayes' networks, on the other hand, allow representation of this class of independencies, but only at the cost of a slightly more complex criterion of separability, one which takes into consideration the directionality of the arrows in the graph. Consider a triplet of variables, x_1, x_2, x_3 , where x_1 is connected to x_3 via x_2 . The two links, connecting the pairs (x_1, x_2) and (x_2, x_3) , can join at the midpoint, x_2 , in one of three possible ways:

- (1) tail-to-tail, $x_1 \leftarrow x_2 \rightarrow x_3$,
- (2) head-to-tail, $x_1 \rightarrow x_2 \rightarrow x_3$ or $x_1 \leftarrow x_2 \leftarrow x_3$,
- (3) head-to-head, $x_1 \rightarrow x_2 \leftarrow x_3$.

If we assume that x_1, x_2, x_3 are the only variables involved, it is clear from the method of constructing the network that, in cases (1) and (2), x_1 and x_3 are conditionally independent, given x_2 , while in case (3), x_1 and x_3 are marginally independent (i.e., $P(x_3|x_1) = P(x_3)$) but may become dependent, given the value of x_2 . Moreover, if x_2 in case (3) has descendants x_4, x_5, \dots , then x_1 and x_3 may also become dependent if any one of those descendant variables is instantiated. These considerations motivate the definition of a qualified version of path connectivity, applicable to paths with directed links and sensitive to all the variables for which values are known at a given time.

- Definition 12.1.1** (a) A subset of variables S_e is said to *separate* x_i from x_j if all paths between x_i and x_j are *separated* by S_e .
 (b) A path P is *separated* by a subset S_e of variables if at least one pair of successive links along P is *blocked* by S_e .

We next introduce a nonconventional criterion under which a pair of converging arrows is said to be *blocked* by S_e .

- Definition 12.1.2** (a) Two links meeting head-to-tail or tail-to-tail at node X are *blocked by* S_e if X is in S_e .
- (b) Two links meeting head-to-head at node X are *blocked by* S_e if neither X nor any of its descendants is in S_e .

This modified definition of separation provides a graphical criterion for testing conditional independence: if S_e separates x_i from x_j , then x_i is conditionally independent of x_j , given S_e . The procedure involved in testing this modified criterion is slightly more complicated than the conventional test for deciding whether S_e is a separating cutset and can be handled by visual inspection. In Figure 12.1, for example, one can easily verify that variables x_2 and x_3 are separated by $S_e = \{x_1\}$ or $S_e = \{x_1, x_4\}$ because the two paths between x_2 and x_3 are blocked by either one of these subsets. However, x_2 and x_3 are not separated by $S_e = \{x_1, x_6\}$ because x_6 , as a descendant of x_5 , “unblocks” the head-to-head connection at x_5 , thus opening a pathway between x_2 and x_3 .

Although the structure of Bayes’ networks, together with the directionality of its links, depends strongly on the node ordering used in the network construction, conditional independence is a property of the underlying distribution and is, therefore, order-invariant. Thus, if we succeed in finding an ordering d in which a given conditional independence relationship becomes graphically transparent, that relationship remains valid even though it may not induce a graph-separation pattern in networks corresponding to other orderings. This permits the use of Bayes’ networks for identifying by inspection a *screening neighborhood* for any given node, namely, a set S of variables that renders a given variable independent of every variable not in S . The separation criterion for Bayes’ networks guarantees that the union of the following three types of neighbors is sufficient for forming a screening neighborhood: direct parents, direct successors and all direct parents of the latter. Thus, in a Markov chain, the screening neighborhood of any nonterminal node consists of its two immediate neighbors while, in trees, the screening neighborhood consists of the (unique) father and the immediate successors. In Figure 12.1, however, the screening neighborhood of x_3 is $\{x_1, x_5, x_2\}$.

12.1.3 An Outline and Summary of Results

The first part of this paper (Section 12.2) deals with the task of fusing and propagating the impacts of new evidence and beliefs through Bayesian networks in such a way that, when equilibrium is reached, each proposition will be assigned a certainty measure consistent with the axioms of probability theory. We first argue (Section 12.2.1) that any viable model of human reasoning should be able to perform this task by a self-activated propagation mechanism, i.e., by an array of simple

and autonomous processors, communicating locally via the links provided by the belief network itself. In Section 12.2.2 we then show that these objectives can be fully realized in tree-structured networks, where each node has only one father. In Section 12.2.3 we extend the result to networks with multiple parents that are singly connected, i.e., there exists only one (undirected) path between any pair of nodes. In both cases, we identify belief parameters, communication messages and updating rules which guarantee that equilibrium is reached in time proportional to the longest path in the network and that, at equilibrium, each proposition will be accorded a belief measure consistent with probability theory. Several approaches toward achieving autonomous propagation in multiply connected networks are discussed in Section 12.2.4.

The second part of the paper (Section 12.3) expands on one of these approaches by examining the feasibility of preprocessing a belief network and turning it permanently into a tree by introducing dummy variables. In Section 12.3.1 we argue that such a technique mimics the way people develop causal models, that dummy variables correspond to the mental constructs known as “hidden causes” and that humans’ relentless search for causal models is motivated by their desire to achieve computational advantages similar to those offered by tree-structured belief networks. After defining (in Section 12.3.2) the notions of star-decomposability and tree-decomposability, Section 12.3.3 treats triplets of propositional variables and asks under what conditions one is justified in attributing the observed dependencies to one central cause represented by a fourth variable. We show that these conditions are readily testable and that, when the conditions are satisfied, the parameters specifying the relations between the visible variables and the central cause can be uniquely determined. In Section 12.3.4 we extend these results to the case of a tree with n leaves. We show that, if there exists a set of dummy variables which decompose a given Bayes network into a tree, then the uniqueness of the triplets’ decomposition enables us to configure that tree from pairwise dependencies among the variables. Moreover, the configuration procedure involves only $O(n \log n)$ steps. In Section 12.3.5 we evaluate the merits of this method and address the difficult issues of estimation and approximation.

12.2 Fusion and Propagation

12.2.1 Autonomous Propagation as a Computational Paradigm

Once a belief network is constructed, it can be used to represent the generic knowledge of a given domain and can be consulted to reason about the interpretation of specific input data. The interpretation process involves instantiating a set of

variables corresponding to the input data, calculating its impact on the probabilities of a set of variables designated as hypotheses and, finally, selecting the most likely combinations of these hypotheses. In general, this process can be carried out by an external interpreter which may have access to all parts of the network, may use its own computational facilities and may schedule its computational steps so as to take full advantage of the network topology with respect to the incoming data. However, the use of such an interpreter appears foreign to the reasoning process normally exhibited by humans [30]. Our limited short-term memory and narrow focus of attention, combined with our inability to shift rapidly between alternative lines of reasoning, suggests that our reasoning process is fairly local, progressing incrementally along pre-established pathways. Moreover, the speed and ease with which we perform some of the low-level interpretive functions, such as recognizing scenes, reading text and even understanding stories, strongly suggest that these processes involve a significant amount of parallelism, and that most of the processing is done *at the knowledge level* itself, not external to it.

A paradigm for modeling such phenomena would be to view an influence network not merely as a passive parsimonious code for storing factual knowledge but also as a computational architecture for reasoning about that knowledge. That means that the links in the network should be treated as the only pathways and activation centers that direct and propel the flow of data in the process of querying and updating beliefs. Accordingly, we assume that each node in the network is designated a separate processor, which both maintains the parameters of belief for the host variable and manages the communication links to and from the set of neighboring, conceptually related, variables. The communication lines are assumed to be open at all times, i.e., each processor may, at any time, interrogate the belief parameters associated with its neighbors and compare them to its own parameters. If the compared quantities satisfy some local constraints, no activity takes place. However, if any of these constraints are violated, the responsible node is activated to set its violating parameter straight. This, of course, will activate similar revisions at the neighboring nodes and will set up a multidirectional propagation process, until equilibrium is reached.

The main reason for this distributed message-passing paradigm is that it leads to a “transparent” revision process, in which the intermediate steps can be given an intuitively meaningful interpretation. Since a distributed process restricts each computational step to obtain inputs only from neighboring, semantically related variables, and since the activation of these steps proceeds along semantically familiar pathways, people find it easy to give meaningful interpretation to the individual steps, thus establishing confidence in the final result. Additionally, it is possible to generate qualitative justifications mechanically by tracing the sequence

of operations along the activated pathways and giving them causal or diagnostic interpretations using appropriate verbal expressions.

The ability to update beliefs by an autonomous propagation mechanism also has a profound effect on sequential implementations of evidential reasoning. Of course, when this architecture is simulated on sequential machines, the notion of autonomous processors working simultaneously in time is only a metaphor; however, it signifies the complete separation of the stored knowledge from the control mechanism—the proclaimed, yet rarely achieved, goal of rule-based architectures. This separation guarantees the ultimate flexibility for a sequential controller; the computations can be performed in any order, without the need to remember or verify which parts of the network have or have not already been updated. Thus, for example, belief updating may be activated by changes occurring in logically related propositions, by requests for evidence arriving from a central supervisor, by a pre-determined schedule or entirely at random. The communication and interaction among individual processors can be simulated using a blackboard architecture [17], where each proposition is designated specific areas of memory to access and modify. Additionally, the uniformity of this propagation scheme renders it natural for formulation in object-oriented languages: each node is an object of the same generic type, and the belief parameters are the messages by which interacting objects communicate.

In AI, constraint-propagation mechanisms have been found essential in several applications, e.g., vision [27, 35] and truth maintenance [20]. However, their use in evidential reasoning has been limited to non-Bayesian formalisms (e.g. [19, 30]). There have been several reasons for this.

First, the conditional probabilities characterizing the links in the network do not seem to impose definitive constraints on the probabilities that can be assigned to the nodes. The quantifier $P(A|B)$ only restricts the belief accorded to A in a very special set of circumstances, namely, when B is known to be true with absolute certainty and when no other evidential data is available. Under normal circumstances, all internal nodes in the network will be subject to some uncertainty and, more seriously, after the arrival of evidence e , the posterior beliefs in A and B are no longer related by $P(A|B)$ but by $P(A|B, e)$, which may be totally different. The result is that any arbitrary assignment of beliefs to propositions A and B can be consistent with the value of $P(A|B)$ initially assigned to the link connecting them; in other words, among these parameters, no violation of constraint can be detected locally.

Next, the difference between $P(A|B, e)$ and $P(A|B)$ suggests that the weights on the links should not remain fixed but should undergo constant adjustment as new evidence arrives. Not only would this entail enormous computational overhead,

but it would also obliterate the advantages normally associated with propagation through fixed networks of constraints.

Finally, the fact that evidential reasoning involves both top-down (predictive) and bottom-up (diagnostic) inferences has caused apprehensions that, once we allow the propagation process to run its course unsupervised, pathological cases of instability, deadlock, and circular reasoning will develop [19]. Indeed, if a stronger belief in a given hypothesis means greater expectation for the occurrence of its various manifestations and if, in turn, a greater certainty in the occurrence of these manifestations adds further credence to the hypothesis, how can one avoid infinite updating loops when the processors responsible for these propositions begin to communicate with one another? Such apprehensions are not unique to probabilistic reasoning but should be considered in any hierarchical model of cognition where mutual reinforcement takes place between lower and higher levels of processing, e.g., connectionist models of reading [29] and language production [4].

This paper demonstrates that coherent and stable probabilistic reasoning *can* be accomplished by local propagation mechanisms while keeping the weights on the links constant throughout the process. This is made possible by characterizing the belief in each proposition by a *list* of parameters, each representing the degree of support the host proposition obtains from one of its neighbors. In the next two subsections we show that maintaining such a breakdown record of the sources of belief facilitates local updating of beliefs and that the network relaxes to a stable equilibrium, consistent with the axioms of probability theory, in time proportional to the network diameter. This record of parameters is also postulated as the mechanism which permits people to retrace reasoned assumptions for the purposes of modifying the model and generating explanatory arguments.

12.2.2 Belief Propagation in Trees

We shall first consider tree-structured influence networks, i.e., one in which every node, except one called “root,” has only one incoming link. We allow each node to represent a multivalued variable which may represent a collection of mutually exclusive hypotheses (e.g., identity of organism: ORG_1, ORG_2, \dots) or a collection of possible observations (e.g. patient’s temperature: high, medium, low). Let a variable be labeled by a capital letter, e.g., A, B, C, \dots , and its possible values subscripted, e.g., A_1, A_2, \dots, A_n . Each directed link $A \rightarrow B$ is quantified by a fixed conditional probability matrix, $M(B|A)$, with entries: $M(B|A)_{ij} = P(B_j|A_i)$. Normally, the directionality of the arrow designates A as the set of causal hypotheses and B as the set of consequences or manifestations for these hypotheses.

Example 12.2.1 Assume that in a certain trial there are three suspects, one of whom has definitely committed a murder, and that the murder weapon, showing some fingerprints, was later found by the police. Let A stand for the identity of the last user of the weapon, namely, the killer. Let B stand for the identity of the last holder of the weapon, i.e., the person whose fingerprints were left on the weapon, and let C represent the possible readings that may be obtained in a fingerprint-testing laboratory.

The relations between these three variables would normally be conceptualized by the chain $A \rightarrow B \rightarrow C$; A generates expectations about B , and B generates expectations about C , but A has no influence on C once we know the value of B .

To represent the common-sense knowledge that, under normal circumstances, the killer is expected to be the last to hold the weapon, we may use the 3×3 conditional probability matrix:

$$P(B_j|A_i) = \begin{cases} 0.80, & \text{if } A_i = B_j, i, j = 1, 2, 3, \\ 0.10, & \text{if } A_i \neq B_j, i, j = 1, 2, 3. \end{cases}$$

To represent the reliability of the laboratory test, we use a matrix $P(C_k|B_j)$, satisfying

$$\sum_k P(C_k|B_j) = 1 \quad \text{for all } j.$$

Each entry in this matrix represents an if-then rule of the type:

If the fingerprint is of suspect B_j then expect reading of the type C_k , with certainty $P(C_k|B_j)$

Note that this rule convention is at variance with that used in many expert systems (e.g., MYCIN), where rules point from evidence to hypothesis (e.g., if symptom, then disease), thus denoting a flow of mental inference. By contrast, the arrows in Bayes' networks point from causes to effects or from conditions to consequence, thus denoting a flow of constraints in the physical world. The reason for this choice is that people often prefer to encode experiential knowledge in causal schemata [34] and, as a consequence, rules expressed in causal format are assessed more reliably.²

2. It appears that, by and large, frames used to index human memory are organized to evoke *expectations* rather than *explanations*. The reason could, perhaps, be attributed to the fact that expectation-evoking frames normally consist of more stable relationships. For example, $P(B_j|C_k)$ in Example 12.2.1 would vary drastically with the proportion of people who have type B_j fingerprints. $P(C_k|B_j)$, on the other hand, depends merely on the similarity between the type of fingerprint that suspect B_j has and the readings observed in the lab; it is perceived to be a stable local property of the laboratory procedure, independent of other information regarding suspect B_j .

Incoming information may be of two types: *specific evidence* and *virtual evidence*. Specific evidence corresponds to direct observations which validate, with certainty, the values of some variables in the network. Virtual evidence corresponds to judgments based on undisclosed observations which affect the belief in some variables in the network. Such evidence is modeled by dummy nodes, representing the undisclosed observations, connected by unquantified (dummy) links to the variables affected by the observations. These links will carry only one-way information, from the evidence to the variables affected by it, but not vice versa. For example, if it is impractical for the fingerprint laboratory to disclose all possible readings (in variable C) or if the laboratory chose to base its finding on human judgment, C will be represented by a dummy node, and the link $B \rightarrow C$ will specify the relative degree to which each suspect is believed to be the owner of the fingerprint pattern examined. For example, the laboratory examiner may issue a report in the form of a list,

$$P(C_{\text{observed}}|B) = (0.80, 0.60, 0.50),$$

stating that he/she is 80% sure that the fingerprint belongs to suspect B_1 , 60% sure that it belongs to B_2 and 50% sure that it belongs to B_3 . Note that these numbers need not sum up to unity, thus permitting each judgment to be formed independently of the other, separately matching each suspect's finger-prints to those found on the weapon.

All incoming evidence, both specific and virtual, will be denoted by D to conote *data*, and will be treated by instantiating the variables corresponding to the evidence. For the sake of clarity, we will distinguish between the fixed conditional probabilities that label the links, e.g., $P(A|B)$, and the dynamic values of the updated node probabilities. The latter will be denoted by $\text{BEL}(A_i)$, which reflects the overall belief accorded to proposition $A = A_i$ by all data so far received. Thus,

$$\text{BEL}(A_i) \triangleq P(A_i|D)$$

where D is the value combination of all instantiated variables.

Consider the fragment of a tree, as depicted in Figure 12.2. The belief in the various values of B depends on three distinct sets of data: i.e., data from the tree rooted at B , from the tree rooted at C and from the tree above A . However, since A separates B from all variables except B 's descendants (see Section 12.1.2), the influence of the latter two sources of information on B are completely summarized by their combined effect on A . More formally: let D_B^- stand for the data contained in the tree rooted at B and D_B^+ for the data contained in the rest of the network.

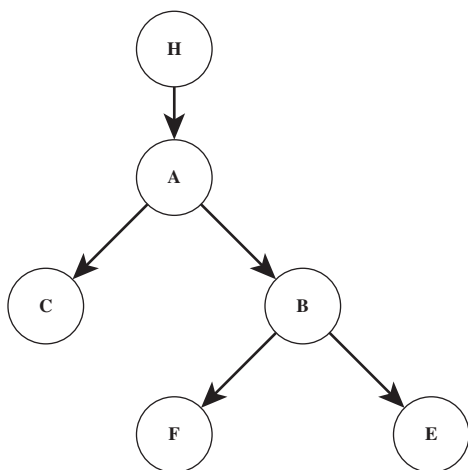


Figure 12.2 A segment of a tree illustrating data partitioning.

We have

$$P(B_j|A_i, D_B^+) = P(B_j|A_i) \tag{12.1}$$

which also leads to the usual “intersiblings” conditional independence:

$$P(B_j, C_k|A_i) = P(B_j|A_i) \cdot P(C_k|A_i), \tag{12.2}$$

since the proposition $C = C_k$ is part of D_B^+ .

12.2.2.1 Data Fusion

Assume we wish to find the belief induced on B by some data $D = D_B^- \cup D_B^+$. Bayes’ theorem, together with (12.1), yields the product rule

$$\text{BEL}(B_i) = P(B_i|D_B^+, D_B^-) = \alpha P[D_B^-|B_i] \cdot P[B_i|D_B^+], \tag{12.3}$$

where α is a normalizing constant. This is a generalization of the celebrated Bayes formula for binary variables

$$O(H|E) = \lambda(E)O(H), \tag{12.4}$$

where $\lambda(E) = P(E|H)/P(E|\bar{H})$ is known as the *likelihood ratio* and $O(H) = P(H)/P(\bar{H})$ as the *prior odds* [6].

As an example, let D_B^- represent the experience of examining the fingerprints left on the murder weapon, and let D_B^+ stand for all other testimonies heard in

the trial. $P(B_i|D_B^+)$ would then stand for our prior (before examining the fingerprints) belief that the i th suspect was the last to hold the weapon, and $P(D_B^-|B_i)$ would represent the report issued by the fingerprint laboratory. Taking, as before, $P(D_B^-|B) = (0.80, 0.60, 0.50)$, and assuming we have $P(B|D_B^+) = (0.60, 0.30, 0.10)$, our total belief in the assertions $B = B_i$ is given by

$$\begin{aligned} \text{BEL}(B) &= \alpha P(D_B^-|B)P(B|D_B^+) \\ &= \alpha(0.80, 0.60, 0.50)(0.60, 0.30, 0.10) \\ &= \alpha(0.48, 0.18, 0.05) \end{aligned}$$

and, to properly normalize $\text{BEL}(B)$, we set $\alpha = (0.48 + 0.18 + 0.05)^{-1}$ and obtain $\text{BEL}(B) = (0.676, 0.254, 0.07)$.

Equation (12.3) generalizes (12.4) in two ways. First, it permits the treatment of nonbinary variables where the mental task of estimating $P(E|\bar{H})$ is often unnatural and where conditional independence with respect to the negations of the hypotheses is normally violated (i.e., $P(E_1, E_2|\bar{H}) \neq P(E_1|\bar{H})P(E_2|\bar{H})$). Second, it identifies a surrogate to the prior probability term for every intermediate node in the tree, even *after* obtaining some evidential data.

In ordinary Bayesian updating of sequential data, it is often possible to recursively use the posterior odd as a new prior for computing the impact of the next item of evidence. However, this method works only when the items of evidence are mutually independent conditioned on the updated hypothesis, H , and will not be applicable to network updating because only variables which are separated from each other by H are guaranteed to be conditionally independent, given H . In general, therefore, it is not permissible to use the total posterior belief, updated by (12.3), as a new multiplicative prior for the calculation. Thus, the significance of (12.3) lies in showing that a product rule analogous to (12.4) can be applied to any node in the network without requiring a separate prior probability assessment. However, the multiplicative role of the prior probability has been taken over by that portion of belief contributed by evidence from the subtree *above* the updated variable, i.e., excluding the data collected from its descendants. The root is the only node which requires a prior probability estimation, and since it has no network above, D_{root}^+ should be interpreted as the background knowledge which remains unexplicated.

Equation (12.3) suggests that the probability distribution of every variable in the network can be computed if the node corresponding to that variable contains the parameters

$$\lambda(B_i) = P(D_B^-|B_i) \tag{12.5}$$

and

$$\pi(B_i) = P(B_i|D_B^+). \quad (12.6)$$

$\pi(B_i)$ represents the causal or *anticipatory* support attributed to B_i by the ancestors of B , and $\lambda(B_i)$ represents the diagnostic or *retrospective* support B_i receives from B 's descendants. The total strength of belief in B_i would be obtained by *fusing* these two supports via the product

$$\text{BEL}(B_i) = \alpha\lambda(B_i)\pi(B_i). \quad (12.7)$$

While two parameters, $\lambda(E)$ and $O(H)$, were sufficient for binary variables, an n -valued variable needs to be characterized by two n -tuples:

$$\lambda(B) = \lambda(B_1), \lambda(B_2), \dots, \lambda(B_n), \quad (12.8)$$

$$\pi(B) = \pi(B_1), \pi(B_2), \dots, \pi(B_n). \quad (12.9)$$

To see how information from several descendants fuse at node B , note that the data D_B^- in (12.5) can be partitioned into disjoint subsets, $D^{1-}, D^{2-}, \dots, D^{m-}$, one for each subtree emanating from (the m children of) B . Since B “separates” these subtrees, conditional independence holds:

$$\lambda(B_i) = P(D_B^-|B_i) = \prod_k P(D^{k-}|B_i), \quad (12.10)$$

so $\lambda(B_i)$ can be formed as a product of the terms $P(D^{k-}|B_i)$ if these are delivered to processor B as messages from its children. For instance if in our fingerprint example $P(D^{1-}|B) = (0.80, 0.60, 0.50)$ and $P(D^{2-}|B) = (0.30, 0.50, 0.90)$ represent two reports issued by two independent laboratories, then the overall diagnostic support $\lambda(B)$ attributable to the three possible states of B is

$$\lambda(B) = (0.80, 0.60, 0.50) \cdot (0.30, 0.50, 0.90) = (0.24, 0.30, 0.45).$$

This, combined with the previous causal support $\pi(B) = (0.60, 0.30, 0.10)$, yields an overall belief of

$$\begin{aligned} \text{BEL}(B) &= \alpha(0.24, 0.30, 0.45)(0.60, 0.30, 0.10) \\ &= (0.516, 0.322, 0.161). \end{aligned}$$

Thus, we see that, at each node of a Bayes tree, the fusion of all incoming data is purely multiplicative.

12.2.2.2 Propagation Mechanism

Assuming that the vectors λ and π are stored with each node of the network, our task is now to determine how the influence of new information will spread through the network, namely, how the parameters π and λ of a given node can be determined from the π 's and λ 's of its neighbors. This is done easily by conditioning (12.5) and (12.6) on all the values that the neighbors can assume. For example, suppose E is the k th son of B . To compute the k th multiplicand in the product of (12.10) from the value of $\lambda(E)$, we write

$$P(D^{k-}|B_i) = \sum_j P(D_E^-|B_i, E_j)P(E_j|B_i)$$

and obtain (using (12.1) and (12.5))

$$P(D^{k-}|B_i) = \sum_j \lambda(E_j)P(E_j|B_i).$$

Thus, $P(D^{k-}|B_i)$ is obtained by taking the λ -vector stored at the k th son of B and multiplying it by the fixed conditional-probability matrix that quantifies the link between B and E . Thus, the λ -vector of each node can be computed from the λ 's of its children by multiplying the latter by their respective link matrices and then multiplying the resultant vectors together, term-by-term, as shown in (12.10). Each multiplicand $P(D^{k-}|B)$ would be treated as a *message* sent by the k th son of B and, if the sending variable is named E , the message will be denoted by $\lambda_E(B)$,

$$\lambda_E(B_i) = \sum_j P(E_j|B_i)\lambda(E_j).$$

A similar analysis, applied to the vector π , shows that the π of any node can be computed from the π of its father and the λ 's of its siblings, again after multiplication by the corresponding link matrices. No direct communication with the siblings is necessary since the information required of them already resides at the father's site (for the purpose of calculating its λ , as in (12.10)) and can be sent down to the requesting son. This can be shown by conditioning $\pi(B)$ over the values of the parent A :

$$\begin{aligned} \pi(B_i) &= P(B_i|D^+(B)) \\ &= \sum_j P(B_i|A_j, D^+(B))P(A_j|D^+(B)) \\ &= \sum_j P(B_i|A_j), P(A_j|\text{all data excluding } D^-(B)) \\ &= \sum_j P(B_i|A_j) \left[\alpha\pi(A_j) \prod_m \lambda_m(A_j) \right] \end{aligned}$$

with m ranging over the siblings of B . The expression in the brackets contains parameters available to processor A , and it can be chosen, therefore, as the message $\pi_B(A)$ that A transmits to B .

Thus,

$$\pi(B_i) = \sum_j P(B_i|A_j)\pi_B(A_j), \quad (12.11)$$

where

$$\pi_B(A_j) = \alpha\pi(A_j) \prod_{m:\text{ sibling of } B} \lambda_m(A_j), \quad (12.12)$$

or, alternatively,

$$\pi_B(A_j) = \alpha' \frac{\text{BEL}(A_j)}{\lambda_B(A_j)}. \quad (12.13)$$

The division by $\lambda_B(A)$ amounts to removing from $\text{BEL}(A)$ the contribution of D_B^- as dictated by the definition of π in (12.6).

These results lead to the following propagation scheme:

Step 1. When processor B is activated to update its parameters, it simultaneously inspects the $\pi_B(A)$ message communicated by the father A and the messages $\lambda_1(B)$, $\lambda_2(B)$, ..., communicated by each of its sons. Using these inputs, it then updates its λ and π as follows:

Step 2. λ is computed using a term-by-term multiplication of the vectors $\lambda_1, \lambda_2, \dots$, (as in (12.10)):

$$\lambda(B_i) = \lambda_1(B_i) \times \lambda_2(B_i) \times \dots = \prod_k \lambda_k(B_i).$$

Step 3. π is computed using:

$$\pi(B_i) = \beta \sum_j P(B_i|A_j)\pi_B(A_j),$$

where β is a normalizing constant and $\pi_B(A)$ is the last message sent to B from the father A .

Step 4. Using the messages received, together with the updated values of λ and π , each processor then computes new π - and λ -messages to be posted on the message boards reserved for its sons and its father, respectively. These are computed as follows:

Step 5. Bottom-up propagation. The new message $\lambda_B(A)$ that B sends to its father (A) is computed by

$$\lambda_B(A_j) = \sum_i P(B_i|A_j)\lambda(B_i).$$

Step 6. Top-down propagation. The new message $\pi_E(B)$ that B sends to its k th child E is computed by

$$\pi_E(B_i) = \alpha\pi(B_i) \prod_{m \neq k} \lambda_m(B_i),$$

or, alternatively,

$$\pi_E(B_i) = \alpha' \frac{\text{BEL}(B_i)}{\lambda_E(B_i)}.$$

This updating scheme is shown schematically in Figure 12.3, where multiplications of any two vectors stand for term-by-term operations. There is no need,

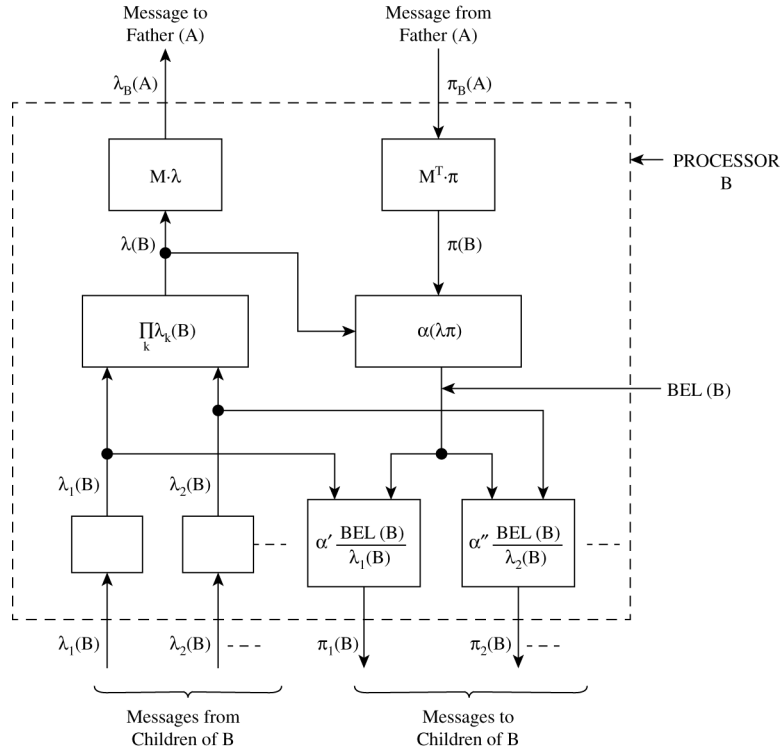


Figure 12.3 The internal structure of a single processor performing belief updating for variable B .

of course, to normalize the π -messages prior to transmission (only the $\text{BEL}(\cdot)$ expressions actually require normalization). This is done solely for the purpose of retaining the probabilistic meaning of these messages. Additional economy can be achieved by having each node B transmit a single message $\text{BEL}(B)$ to all its children and letting each child use (12.13) to uncover its appropriated π -message.

Terminal and data nodes in the tree require special treatments. Here we have to distinguish several cases:

- (1) *Anticipatory node*, a leaf node that has not been instantiated yet: For such variables, BEL should be equal to π and, therefore, we should set $\lambda = (1, 1, \dots, 1)$.
- (2) *Data node*, a variable with instantiated value: Following (12.5) and (12.6), if the j th state of B were observed to be true, we set $\lambda = \pi = (0, \dots, 0, 1, 0, \dots, 0)$ with 1 at the j th position.
- (3) *Dummy node*, a node B representing virtual or judgmental evidence bearing on A : We do not specify $\lambda(B)$ or $\pi(B)$ but, instead, post a $\lambda_B(A)$ message to A , where $\lambda_B(A_i) = K \cdot P(\text{observation}|A_i)$, and K is any convenient constant.
- (4) *Root node*: The boundary condition for the root node is established by setting $\pi(\text{root}) = \text{prior probability of the root variable}$.

Example 12.2.2 To illustrate these computations let us return to Example 12.2.1, and let us assume that based on all testimonies heard so far, our belief in the identity of the killer amounts to $\pi(A) = (0.8, 0.1, 0.1)$. Before obtaining any fingerprint information, B is an anticipatory node with $\lambda(B) = (1, 1, 1)$, which also yields $\lambda_B(A) = \lambda(A) = (1, 1, 1)$ and $\text{BEL}(A) = \pi(A)$. $\pi(B)$ can be calculated from (12.13) (using $\pi_B(A) = \pi(A)$ and $P(B_i|A_j) = 0.8$ if $i = j$), yielding

$$\pi(B) = \begin{bmatrix} 0.8 & 0.1 & 0.1 \\ 0.1 & 0.8 & 0.1 \\ 0.1 & 0.1 & 0.8 \end{bmatrix} \begin{bmatrix} 0.8 \\ 0.1 \\ 0.1 \end{bmatrix} = (0.66, 0.17, 0.17) = \text{BEL}(B).$$

Now assume that a laboratory report arrives, summarizing the test results (a virtual evidence C) by the message $\lambda_C(B) = \lambda(B) = (0.80, 0.60, 0.50)$. Node B updates its belief to read:

$$\begin{aligned} \text{BEL}(B) &= \alpha \lambda(B) \pi(B) = \alpha (0.80, 0.60, 0.50) (0.66, 0.17, 0.17) \\ &= (0.738, 0.142, 0.119) \end{aligned}$$

and computes a new message, $\lambda_B(A)$, for A :

$$\lambda_B(A) = M \cdot \lambda = \begin{bmatrix} 0.8 & 0.1 & 0.1 \\ 0.1 & 0.8 & 0.1 \\ 0.1 & 0.1 & 0.8 \end{bmatrix} \begin{bmatrix} 0.8 \\ 0.6 \\ 0.5 \end{bmatrix} = (0.75, 0.61, 0.54).$$

Upon receiving this message, node A sets $\lambda(A) = \lambda_B(A)$ and recomputes its belief to

$$\begin{aligned} \text{BEL}(A) &= \alpha \lambda(A) \pi(A) = \alpha(0.75, 0.61, 0.54)(0.8, 0.1, 0.1) \\ &= (0.84, 0.085, 0.076). \end{aligned}$$

Now assume that suspect A_1 produces a very strong alibi in his favor, suggesting that there are only 1 : 10 odds that he could have committed the crime. To fuse this information with all previous evidence, we link a new virtual-evidence node E directly to A and post the message $\lambda_E(A) = (0.10, 1.0, 1.0)$ on the link. $\lambda_E(A)$ combines with $\lambda_B(A)$ to yield

$$\begin{aligned} \lambda(A) &= \lambda_E(A) \lambda_B(A) = (0.075, 0.61, 0.54), \\ \text{BEL}(A) &= \alpha(A) \pi(A) \\ &= \alpha(0.075, 0.061, 0.54)(0.84, 0.85, 0.076) \\ &= (0.404, 0.333, 0.263) \end{aligned}$$

and generates the message $\pi_B(A) = \alpha \lambda_E(A) \pi(A) = \alpha(0.08, 0.1, 0.1)$ to B . Upon receiving $\pi_B(A)$, processor B updates its causal support $\pi(B)$ to read:

$$\pi(B) = \alpha' \begin{bmatrix} 0.8 & 0.1 & 0.1 \\ 0.1 & 0.8 & 0.1 \\ 0.1 & 0.1 & 0.8 \end{bmatrix} \begin{bmatrix} 0.08 \\ 0.10 \\ 0.10 \end{bmatrix} = (0.30, 0.35, 0.35)$$

and $\text{BEL}(B)$ becomes

$$\begin{aligned} \text{BEL}(B) &= \alpha \lambda(B) \pi(B) \\ &= \alpha(0.8, 0.6, 0.5)(0.334, 0.343, 0.317) \\ &= (0.423, 0.326, 0.251). \end{aligned}$$

The purpose of propagating beliefs top-down to sensory nodes such as B is two-fold—to guide data-acquisition strategies toward the most informative sensory nodes and to facilitate explanations which justify the system's inference steps.

Note that $BEL(A)$ cannot be taken as an updated prior of A for the purpose of calculating $BEL(B)$. In other words, it is wrong to update $BEL(B)$ via the textbook formula

$$BEL(B_i) = \sum_j P(B_i|A_j)BEL(A_j),$$

also known as Jeffrey's rule [11], because $BEL(A)$ itself was affected by information transmitted from B , and reflecting this information back to B would amount to counting the same evidence twice.

12.2.2.3 Illustrating the Flow of Belief

Figure 12.4 shows six successive stages of belief propagation through a simple binary tree, assuming that updating is triggered by changes in the belief parameters of neighboring processors. Initially (Figure 12.4(a)), the tree is in equilibrium, and all terminal nodes are anticipatory. As soon as two data nodes are activated (Figure 12.4(b)), white tokens are placed on their links, directed towards their fathers. In the next phase, the fathers, activated by these tokens, absorb them and manufacture the appropriate number of tokens for their neighbors (Figure 12.4(c)): white tokens for their fathers and black ones for the children. (The links through which the absorbed tokens have entered do not receive new tokens, thus reflecting the feature that a π -message is not affected by a λ -message crossing the same link.)

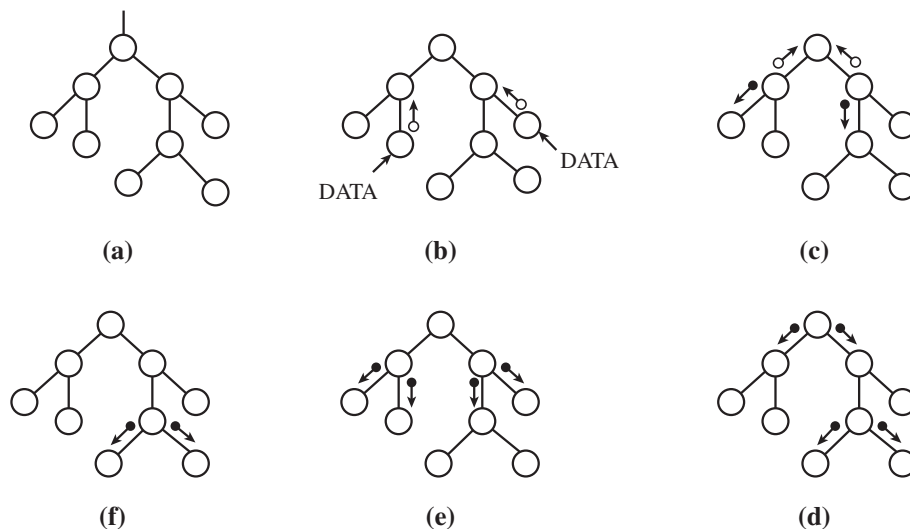


Figure 12.4 The impact of new data propagates through a tree by a message-passing process.

The root node now receives two white tokens, one from each of its descendants. That triggers the production of two black tokens for top-down delivery (Figure 12.4(d)). The process continues in this fashion until after six cycles, all tokens are absorbed, and the network reaches a new equilibrium.

As soon as a leaf node posts a token for its parent, it is ready to receive new data and, when this occurs, a new token is posted on the link, replacing the old one. In this fashion the inference network can also track a changing environment and provide coherent interpretation of signals emanating simultaneously from multiple sources.

12.2.2.4 Properties of the Updating Scheme

(1) The local computations required by the updating scheme are efficient in both storage and time. For an m -ary tree with n values per node, each processor should store $n^2 + mn + 2n$ real numbers and perform $2n^2 + mn + 2n$ multiplications per update.

(2) The local computations and the final belief distribution are entirely independent of the control mechanism that activates the individual operations. They can be activated by either data-driven or goal-driven (e.g., requests for evidence) control strategies, by a clock or at random.

(3) New information diffuses through the network in a single pass. Instabilities and indefinite relaxations have been eliminated by maintaining a two-parameter system (π and λ) to decouple causal support from diagnostic support. The time required for completing the diffusion (in parallel) is proportional to the diameter of the network.

12.2.3 Propagation in Singly Connected Networks

The tree structures treated in the preceding section require that exactly one variable be considered a cause of any other variable. This restriction simplifies computations, but its representational power is rather limited since it forces us to group together all causal factors sharing a common consequence into a single node. By contrast, when people associate a given observation with multiple potential causes, they weigh one causal factor against another as independent variables, each pointing to a specialized area of knowledge. As an illustration, consider the following situation:

Mr. Holmes received a phone call at work from his neighbor notifying him that she heard a burglar alarm sound from the direction of his home. As he is preparing to rush home, Mr. Holmes recalls that recently the alarm had

been triggered by an earthquake. Driving home, he hears a radio newscast reporting an earthquake 200 miles away. [14]

Mr. Holmes perceives two episodes which may be potential causes for the alarm sound, an attempted burglary and an earthquake. Even though burglaries can safely be assumed independent of earthquakes, the radio announcement still reduces the likelihood of a burglary, as it “explains away” the alarm sound. Moreover, the causal events are perceived as individual variables each pointing to a separate frame of knowledge.

This nonmonotonic interaction among multiple causes is a prevailing pattern of human reasoning. When a physician discovers evidence in favor of one disease, it reduces the likelihood of other diseases, although the patient might well be suffering from two or more disorders simultaneously. The same maxim also governs the interplay of other frame-like explanations (not necessarily causal). For example, it is essential for comprehending sentences such as “John could not walk straight, and I thought he got drunk again. However, seeing the blood on his shirt, I knew it was a different matter.”

This section extends the propagation scheme to graph structures which permit a node to have multiple parents and thus capture “sideways” interactions via common successors. However, the graphs are required to be *singly connected*, namely, one (undirected) path, at most, exists between any two nodes.

12.2.3.1 Fusion Equations

Consider a fragment of a singly connected network, depicted in Figure 12.5. The link $B \rightarrow A$ partitions the graph into two parts: an upper subgraph, G_{BA}^+ , and a lower subgraph G_{BA}^- . These two graphs contain two sets of *data*, which we shall call D_{BA}^+ and D_{BA}^- , respectively. Likewise, the links $C \rightarrow A$, $A \rightarrow X$, and $A \rightarrow Y$ define the subgraphs G_{CA}^+ , G_{AX}^- , and G_{AY}^- , which contain the data sets D_{CA}^+ , D_{AX}^- and D_{AY}^- , respectively. Since A is a common child of B and C , it does not separate G_{BA}^+ from G_{CA}^+ . However, it does separate the following three subgraphs: $G_{BA}^+ \cup G_{CA}^+$, G_{AX}^- and G_{AY}^- , and we can write

$$P(D_{AX}^-, D_{AY}^- | A_i, D_{BA}^+, D_{CA}^+) = P(D_{AX}^- | A_i) P(D_{AY}^- | A_i). \quad (12.14)$$

Thus, using Bayes' rule, the overall strength of belief in A_i can be written:

$$\begin{aligned} \text{BEL}(A_i) &= P(A_i | D_{BA}^+, D_{CA}^+, D_{AX}^-, D_{AY}^-) \\ &= \alpha P(A_i | D_{BA}^+, D_{CA}^+) P(D_{AX}^- | A_i) P(D_{AY}^- | A_i), \end{aligned} \quad (12.15)$$

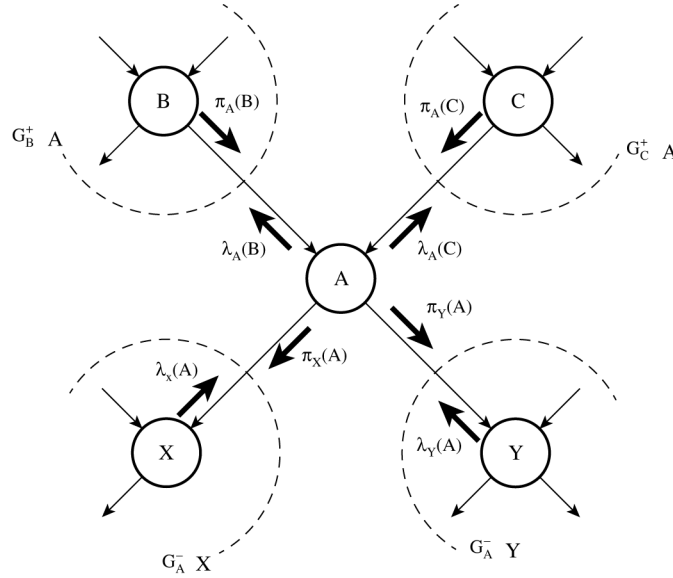


Figure 12.5 Fragment of a singly connected network with multiple parents, illustrating data partitioning and belief parameters.

where α is a normalizing constant. By further conditioning over the values of B and C (see Appendix 12.A), we get:

$$\text{BEL}(A_i) = \alpha P(D_{AX}^- | A_i) P(D_{AY}^- | A_i) \cdot \left[\sum_{jk} P(A_i | B_j, C_k) P(B_j | D_{BA}^+) P(C_k | D_{CA}^+) \right]. \quad (12.16)$$

Equation (12.16) shows that the probability distribution of each variable A in the network can be computed if three types of parameters are made available: (1) the current strength of the causal support, π , contributed by each incoming link to A :

$$\pi_A(B_j) = P(B_j | D_{BA}^+), \quad (12.17)$$

(2) the current strength of the diagnostic support, λ , contributed by each outgoing link from A :

$$\lambda_X(A_i) = P(D_{AX}^- | A_i), \quad (12.18)$$

and (3) the fixed conditional-probability matrix, $P(A|B, C)$, which relates the variable A to its immediate causes. Accordingly, we let each link carry two dynamic parameters, π and λ , and let each node store an encoding of $P(A|B, C)$.

With these parameters at hand, the fusion equation (12.16) becomes

$$\text{BEL}(A_i) = \alpha \lambda_X(A_i) \lambda_Y(A_i) \sum_{jk} P(A_i|B_j, C_k) \pi_A(B_j) \pi_A(C_k). \quad (12.19)$$

Alternatively, from two parameters, π and λ , residing on the same link, we can compute the belief distribution of the parent node by the product

$$\text{BEL}(B_j) = \alpha \pi_A(B_j) \lambda_A(B_j). \quad (12.20)$$

12.2.3.2 Propagation Equation

Assuming that the vectors π and λ are stored with each link, our task is now to prescribe how the influence of new information should spread through the network.

Updating λ

Starting from the definition of $\lambda_A(B_i) = P(D_{BA}^-|B_i)$, we partition the data D_{BA}^- into its components: A , D_{AX}^- , D_{AY}^- , and D_{CA}^+ , and summing over all values of A and C (see Appendix 12.A), we get:

$$\lambda_A(B_i) = \alpha \sum_j \left[\pi_A(C_j) \sum_k \lambda_X(A_k) \lambda_Y(A_k) P(A_k|B_i, C_j) \right]. \quad (12.21)$$

Equation (12.21) shows that only three parameters (in addition to the conditional probabilities $P(A|B, C)$) are needed for updating the diagnostic parameter vector $\lambda_A(B)$: $\pi_A(C)$, $\lambda_X(A)$, and $\lambda_Y(A)$. This is expected since D_{BA}^- is completely summarized by X , Y , and C .

Updating π

Similar manipulation on (12.17) (see Appendix 12.A) yields the following rule for updating the causal parameter $\pi_X(A)$:

$$\pi_X(A_i) = \alpha \lambda_Y(A_i) \left[\sum_{jk} P(A_i|B_j, C_k) \pi_A(B_j) \pi_A(C_k) \right]. \quad (12.22)$$

Thus, $\pi_X(A)$, like $\lambda_A(B)$, is also determined by three neighboring parameters: $\lambda_Y(A)$, $\pi_A(B)$, and $\pi_A(C)$.

Equations (12.21) and (12.22) demonstrate that a perturbation of the causal parameter π will not affect the diagnostic parameter λ on the same link, and vice versa. The two are orthogonal to each other since they depend on two disjoint sets of data. Therefore, any perturbation of beliefs due to new evidence propagates through the network and is absorbed at the boundary without reflection. A new state of equilibrium will be reached after a finite number of updates which, in the worst case, would be equal to the diameter of the network.

Equation (12.21) also reveals that if no data are observed below A (i.e., all λ 's pointing to A are unit vectors), then all λ 's emanating from A are unit vectors. This means that evidence gathered at a particular node does not influence its spouses until their common son gathers diagnostic support. This reflects the special connectivity conditions established in Section 12.1.2 and matches our intuition regarding multiple causes. In Mr. Holmes' case, for example, prior to the neighbor's telephone call, seismic data indicating an earthquake would not have influenced the likelihood of a burglary.

Although the treatment in this paper is restricted to discrete variables, (12.21) and (12.22) can be readily extended to handle continuous variables as well. The case of additive Gaussian variables is particularly attractive because all belief distributions and all the π - and λ -messages can be characterized by only two parameters each, the mean and the variance. Thus, the computations required are simpler, and matrix manipulations are avoided [23]. Distributed updating of noncausal, object-class hierarchies is described in [25].

12.2.4 Summary and Extensions for Multiply Connected Networks

The preceding two sections show that the architectural objectives of propagating beliefs coherently through an active network of primitive, identical, and autonomous processors can be fully realized in singly connected graphs. Instabilities due to bidirectional inferences are avoided by using multiple, source-identified belief parameters, and equilibrium is guaranteed to be reached in time proportional to the network diameter.

The primitive processors are simple and repetitive, and they require no working memory except that used in matrix multiplications. Thus, this architecture lends itself naturally to hardware implementation, capable of real-time interpretation of rapidly changing data. It also provides a reasonable model of neural nets involved in such cognitive tasks as visual recognition, reading comprehension [28] and associative retrieval [1], where unsupervised parallelism is an uncontested mechanism.

It is also interesting to note that the marginal conditional probabilities on the links of the network remain constant and retain their viability throughout the updating process. This is important because having to adjust the weights each time new data arrives would be computationally prohibitive. The stable viability of the marginal conditional probabilities may explain why people can assess the magnitude of these relationships better than those of any other probabilistic quantity. Apparently, these relationships have been chosen as the standard primitives for organizing and quantifying probabilistic knowledge in our long-term memory.

The efficacy of singly connected networks in supporting autonomous propagation raises the question of whether similar propagation mechanisms can operate

in less restrictive networks (like the one in Figure 12.1), where multiple parents of common children may possess common ancestors, thus forming loops in the underlying network. If we ignore the existence of loops and permit the nodes to continue communicating with each other as if the network were singly connected, messages may circulate indefinitely around these loops, and the process will not converge to the correct state of equilibrium.

A straightforward way of handling the network of Figure 12.1 would be to appoint a local interpreter for the loop x_1, x_2, x_3, x_5 that will account for the interactions between x_2 and x_3 . This amounts, basically, to collapsing nodes x_2 and x_3 into a single node representing the compound variable (x_2, x_3) . This method works well on small loops [32], but as soon as the number of variables exceeds 3 or 4, compounding requires handling huge matrices and masks the natural conceptual structure embedded in the original network.

A second method of propagation is based on “stochastic relaxation” [8] similar to that used by Boltzman machines [9]. Each processor examines the states of the variables within its screening neighborhood, computes a belief distribution for the values of its host variable, then randomly selects one of these values with probability given by the computed distribution. The value chosen will subsequently be interrogated by the neighbors upon computing their beliefs, and so on. This scheme is guaranteed convergence, but it usually requires very long relaxation times before reaching a steady state.

A third method called *conditioning* [22] is based on our ability to change the connectivity of a network and render it singly connected by instantiating a selected group of variables. In Figure 12.1, for example, instantiating x_1 to some value would block the pathway x_2, x_1, x_3 , and would render the rest of the network singly connected, so that the propagation techniques of the preceding section would be applicable. Thus, if we wish to propagate the impact of an observed datum, say at x_6 , to the entire network, we first assume $x_1 = 0$, propagate the impact of x_6 to the variables x_2, \dots, x_5 , repeat the propagation under the assumption $x_1 = 1$ and, finally, sum the two results weighted by the posterior probability $P(x_1|x_6)$. It can also be executed in parallel by letting each node receive, compute, and transmit several sets of parameters, one for each value of the conditioning variable(s). Conditioning provides a working solution in most practical cases, but it occasionally suffers from the inevitable combinatorial explosion—the number of messages may grow exponentially with the number of nodes required for breaking up all loops in the network.

The use of conditioning to facilitate propagation is not foreign to human reasoning. When we find it hard to estimate the likelihood of a given outcome, we often make hypothetical *assumptions* that render the estimation simpler and then

negate the assumptions to see if the results do not vary substantially. One of the most pervasive patterns of plausible reasoning is the maxim that, if two diametrically opposed assumptions impart two different degrees of confidence onto a proposition Q , then the unconditional degree of confidence merited by Q should be somewhere between the two. The terms “hypothetical” or “assumption-based” reasoning, “reasoning by cases,” and “envisioning” all refer to the same basic mechanism of selecting a key variable, binding it to some of its values, deriving the consequences of each binding separately, and integrating those consequences together.

Finally, a preprocessing approach, which is discussed more fully in Section 12.3, introduces auxiliary variables and permanently turns the network into a tree. To understand the basis of this method, consider, for example, the tree of Figure 12.2. The variables C, H, E, F are tightly coupled in the sense that no two of them can be separated by the others; therefore, if we were to construct a Bayesian network based on these variables *alone*, a complete graph would ensue. Yet, together with the intermediate variables A and B the interactions among the leaf variables are tree-structured, clearly demonstrating that some multiply connected networks can inherit all the advantages of tree representations by the introduction of a few dummy variables. In some respects, this method is similar to that of appointing external interpreters to handle nonseparable components of the graph, because the processors assigned to the dummy variables, like the external interpreters, serve no other function but that of mediation among the real variables. However, the dummy-variables scheme enjoys the added advantage of uniformity: the processors representing the dummy variables can be identical to those representing the real variables, in full compliance with our architectural objectives. Moreover, there are strong reasons to believe that the process of reorganizing data structures by adding fictitious variables mimics an important component of conceptual development in human beings—the evolution of causal models. These considerations are discussed in the section that follows.

12.3 Structuring Causal Trees

12.3.1 Causality, Conditional Independence, and Tree Architecture

Human beings exhibit an almost obsessive urge to conceptually mold empirical phenomena into structures of cause-and-effect relationships. This tendency is, in fact, so compulsive that it sometimes comes at the expense of precision and often requires the invention of hypothetical, unobservable entities such as “ego,” “elementary particles,” and “supreme beings” to make theories fit the mold of causal

schema. When we try to explain the actions of another person, for example, we invariably invoke abstract notions of mental states, social attitudes, beliefs, goals, plans, and intentions. Medical knowledge, likewise, is organized into causal hierarchies of invading organisms, physical disorders, complications, clinical states and, only finally, the visible symptoms.

We take the position that human obsession with causation, like many other psychological compulsions, is computationally motivated. Causal models are attractive only because they provide effective data structures for representing empirical knowledge—they can be queried and updated at high speed with minimal external supervision; so, it behooves us to take a closer look at the structure of causal models and determine what it is that makes them so effective. In other words, what are the computational assets of those fictitious variables called “causes” that make them worthy of such relentless human pursuit, and what renders causal explanations so pleasing and comforting, once they are found?

The paradigm expounded in this paper is that the main ingredient responsible for the pervasive role of causal models is their *centrally organized architecture*, i.e., an architecture in which dependencies among variables are mediated by one central mechanism.

If you ask n persons in the street what time it is, the answers will undoubtedly be very similar. Yet, instead of suggesting that, somehow, the answers evoked or the persons surveyed influence each other, we postulate the existence of a central cause, the standard time, and the commitment of each person to adhere to that standard. Thus, instead of dealing with a complex n -ary relation, the causal model in this example consists of a network of n binary relations, all connected star-like to one central node which serves to dispatch information to and from the connecting variables. Psychologically, this architecture is much more pleasing than one which entails intervariable communication. Since the activity of each variable is constrained by only one source of information (i.e., the central cause), no conflict in activity arises: any assignment of values consistent with the central constraints will also be globally consistent, and a change in any of the variables can communicate its impact to all other variables in only two steps.

Computationally speaking, such causes are merely names given to auxiliary variables which facilitate the efficient manipulation of the activities of the original variables in the system. They encode a summary of the interactions among the visible variables and, once calculated, permit us to treat the visible variables as if they were mutually independent.

The dual summarizing/decomposing role of a causal variable is analogous to that of an orchestra conductor: it achieves coordinated behavior through central communication and thereby relieves the players from having to communicate

directly with one another. In the physical sciences, a classical example of such coordination is exhibited by the construct of a *field* (e.g., gravitational, electric, or magnetic). Although there is a one-to-one mathematical correspondence between the electric field and the electric charges in terms of which it is defined, nearly every physicist takes the next step and ascribes physical reality to the electric field, imagining that in every point of space there is some real physical phenomenon taking place which determines both the magnitude and direction which tag the point. This psychological construct offers an advantage vital to understanding the development of electrical sciences: It decomposes the complex phenomena associated with interacting electric charges into two independent processes: (1) the creation of the field at a given point by the surrounding charges, and (2) the conversion of the field into a physical force once another charge passes near that point.

The advantages of centrally coordinated architectures are not unique to star-structured networks but are also present in tree structures since every internal node in the tree centrally coordinates the activities of its neighbors. In a management hierarchy, for example, where employees can communicate with each other only through their immediate superiors, the passage of information is swift, economical, conflict-free, and highly parallel. Likewise, we know that, if the interactions among a set of variables can be represented by a tree of binary constraints, then a globally consistent solution can be found in linear time, using backtrack-free search [3, 7]. These computational advantages of trees also retain their power when the relationships constraining the variables are probabilistic in nature.

In probabilistic formalisms, the topological concept of central coordination is embodied in the notion of *conditional independence*. In our preceding example, the answers to the question “What time is it?” would be viewed as random variables that are bound together by a *spurious correlation* [31, 33]; they become independent of each other once we know the state of the mechanism causing the correlation, i.e., the standard time. Thus, conditional independence captures both functions of our orchestra conductor: coordination and decomposition.

The most familiar connection between causality and conditional independence is reflected in the scientific notion of a *state*. It was devised to nullify the influence that the past exerts on the future by providing a sufficiently detailed description of the present. In probabilistic terms this came to be known as a Markov property; future events are conditionally independent of past events, given the current state of affairs. This is precisely the role played by the set of parents S_i in the construction of Bayesian networks (Section 12.1.1); they screen the variable x_i from the influence of all its other ancestors.

But conditional independence is not limited to separating the past from the future; it often applies to events occurring at the same time. Knowing the values

of the parent set S_i not only decouples x_i from its other ancestors but renders x_i independent of *all* other variables except its descendants. In fact, this sort of independence constitutes the most universal and distinctive characteristic featured by the notion of causality. In medical diagnosis, for example, a group of cooccurring symptoms often become independent of each other once we know the disease that caused them. When some of the symptoms directly influenced each other, the medical profession *invents* a name for that interaction (e.g., complication, clinical state, etc.) and treats it as a new auxiliary variable, which again assumes the decompositional role characteristic of causal agents; knowing the exact state of the auxiliary variable renders the interacting symptoms independent of each other. In other words, the auxiliary variables constitute a sufficient summary for determining the likely development of each individual symptom in the group; thus, additional knowledge regarding the states of the other symptoms becomes superfluous.

The continuous influx of such auxiliary concepts into our languages cast new light on the status of conditional independence in probabilistic modelling. Contrary to positions often found in the literature, conditional independence is not a “restrictive assumption” made for mathematical elegance; neither is it an occasional grace of nature for which we must passively wait. Rather, it is a mental construct that we actively create and a psychological necessity which our culture labors to satisfy.

The decompositional role of causal variables attains its ultimate realization in tree-structured networks, where every pair of nonadjacent variables becomes independent given a third variable on the path connecting the pair. Indeed, the speed, stability and autonomy of the updating scheme described in Section 12.2.2 draws its power from the high degree of decomposition provided by the tree structure. These computational advantages, we postulate, give rise to the satisfying sensation called “in-depth understanding,” which people experience when they discover causal models consistent with observations.

Given that tree dependence captures the main feature of causation and that it provides a convenient computational medium for performing interpretations and predictions, we now ask whether it is possible to reconfigure every belief network as a tree and, if so, how. First we assume that there exist dummy variables which decompose the network into a tree, and then ask whether the internal structure of such a tree can be determined from observations made solely on the leaves. If it can, then the structure found will constitute an operational definition for the hidden causes often found in causal models. Additionally, if we take the view that “learning” entails the acquisition of computationally effective representations of nature’s regularities, then procedures for configuring such trees may reflect an important component of human learning.

A related structuring task was treated by Chow and Liu [2], who also used tree-dependent random variables to approximate an arbitrary joint distribution. However, in Chow's trees all nodes denote observed variables; so, the conditional probability for any pair of variables is assumed to be given. By contrast, the internal nodes in our trees denote dummy variables, artificially concocted to make the representation tree-like. Since only the leaves are accessible to empirical observations, we know neither the conditional probabilities that link the internal nodes to the leaves nor the structure of the tree—these we would have to learn. A similar problem of configuring probabilistic models with hidden variables is mentioned by Hinton et al. [9] as one of the tasks that a Boltzman machine should be able to solve. However, it is not clear whether the relaxation techniques employed by the Boltzman machine can easily escape local minima and whether they can readily accept the constraint that the resulting structure be a tree. The method described in the following sections offers a solution to this problem, but it assumes some restrictive conditions: all variables are bivalued, a solution tree is assumed to exist, and the value of each interleaf correlation is precisely known.

12.3.2 Problem Definition and Nomenclature

Consider a set of n binary-valued random variables x_1, \dots, x_n with a given probability mass function $P(x_1, \dots, x_n)$. We address the problem of representing P as a marginal of an $(n+1)$ -variable distribution $P_s(x_1, \dots, x_n, w)$ that renders x_1, \dots, x_n conditionally independent given w , i.e.,

$$P_s(x_1, \dots, x_n, w) = \prod_{i=1}^n P_s(x_i|w)P_s(w), \quad (12.23)$$

$$P(x_1, \dots, x_n) = \alpha \prod_{i=1}^n P_s(x_i|w=1) + (1-\alpha) \prod_{i=1}^n P_s(x_i|w=0). \quad (12.24)$$

The functions $P_s(x_i|w)$, $w = 0, 1$, $i = 1, \dots, n$, can be viewed as 2×2 stochastic matrices relating each x_i to the central hidden variable w (see Figure 12.6(a)); hence, we name P_s a *star distribution* and call P *star-decomposable*. Each matrix contains two independent parameters, f_i and g_i , where

$$f_i = P_s(x_i = 1|w = 1), \quad g_i = P_s(x_i = 1|w = 0) \quad (12.25)$$

and the central variable w is characterized by its prior probability $P_s(w=1) = \alpha$ (see Figure 12.6(b)).

The advantages of having star-decomposable distributions are several. First, the product form of P_s in (12.23) makes it very easy to compute the probability of

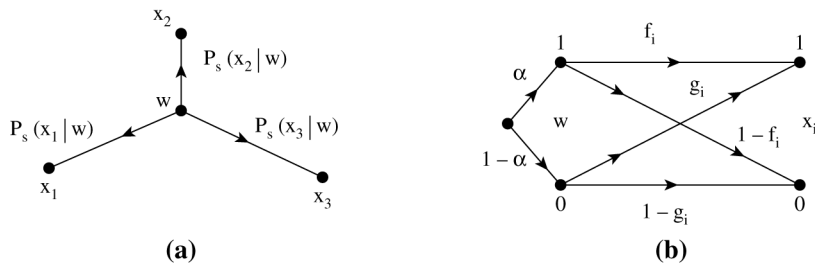


Figure 12.6 (a) Three random variables, x_1, x_2, x_3 connected to a central variable w by a star network. (b) Illustration of the three parameters, α, f_i, g_i , associated with each link.

any combination of variables. More importantly, the product form is also convenient for calculating the conditional probabilities, $P(x_i|x_j)$, describing the impact of an observation x_j on the probabilities of unobserved variables. The computation requires only two vector multiplications.

Unfortunately, when the number of variables exceeds 3, the conditions for star-decomposability become very stringent and are not likely to be met in practice. Indeed, a star-decomposable distribution for n variables has $2n + 1$ independent parameters, while the specification of a general distribution requires $2^n - 1$ parameters. Lazarfeld [16] considered star-decomposable distributions where the hidden variable w is permitted to range over λ values, $\lambda > 2$. Such an extension requires the solution of $\lambda n + \lambda - 1$ nonlinear equations to find the values of its $\lambda n + \lambda - 1$ independent parameters. In this paper, we pursue a different approach, allowing a larger number of binary hidden variables but insisting that they form a tree-like structure (see Figure 12.7), i.e., each triplet forms a star, but the central variables may differ from triplet to triplet. Trees often portray meaningful conceptual hierarchies and are, computationally, almost as convenient as stars.

We shall say that a distribution $P(x_1, x_2, \dots, x_n)$ is *tree-decomposable* if it is the marginal of a distribution

$$P_T(x_1, x_2, \dots, x_n, w_1, w_2, \dots, w_m), \quad m \leq n - 2$$

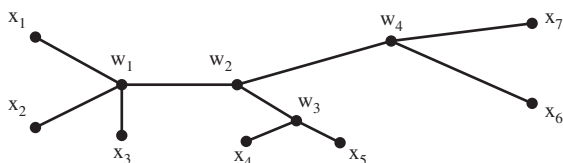


Figure 12.7 A tree containing four dummy variables and seven visible variables.

that supports a tree-structured network, such that w_1, w_2, \dots, w_m correspond to the internal nodes of a tree T and x_1, x_2, \dots, x_n to its leaves.

Note that if P_T supports a rooted tree T , then any two leaves are conditionally independent, given the value of any internal node on the path connecting them. These relationships between leaves and internal nodes are a property of the undirected tree, independent of the choice of root. Now, since a choice of a new root for T will create a tree T' which is also supported by P_T , we are permitted to treat T as an unrooted tree. Conversely, given an unrooted tree T and an assignment of variables to its nodes, the form of the corresponding distribution can be written by the following procedure: We first choose an arbitrary node as a root. This, in turn, defines a unique father $F(y_i)$ for each node $y_i \in \{x_1, \dots, x_n, w_1, \dots, w_m\}$ in T , except the chosen root, y_1 . The joint distribution is simply given by the product form:

$$P_T(x_1, \dots, x_n, w_1, \dots, w_m) = P(y_1) \prod_{i=2}^{m+n} P[y_i|F(y_i)]. \quad (12.26)$$

For example, if in Figure 12.7 we choose w_2 as the root, we obtain:

$$\begin{aligned} &P_T(x_1, \dots, x_7, w_1, \dots, w_4) \\ &= P(x_7|w_4)P(x_6|w_4)P(x_5|w_3)P(x_4|w_3) \\ &\quad \cdot P(x_3|w_1)P(x_2|w_1)P(x_1|w_1)P(w_1|w_2) \cdot P(w_3|w_2)P(w_4|w_2)P(w_2). \end{aligned}$$

Throughout this discussion we shall assume that each w has at least three neighbors; otherwise, it is superfluous. In other words, an internal node with two neighbors can simply be replaced by an equivalent direct link between the two. Similarly, we shall assume that all link matrices are nonsingular, conveying genuine dependencies between the linked variables; otherwise, the tree can be decomposed into disconnected components, i.e., a forest.

If we are given $P_T(x_1, \dots, x_n, w_1, \dots, w_m)$, then, clearly, we can obtain $P(x_1, \dots, x_n)$ by summing over w 's. We now ask whether the inverse transformation is possible, i.e., given a tree-decomposable distribution $P(x_1, \dots, x_n)$, can we recover its underlying extension $P_T(x_1, \dots, x_n, w_1, \dots, w_m)$? We shall show that: (1) the tree distribution P_T is unique, (2) it can be recovered from P using $n \log n$ computations, and (3) the structure of T is uniquely determined by the second-order probabilities of P . The construction method depends on the analysis of star-decomposability for triplets, which is presented next. (Impatient readers may skip this analysis and go directly to Theorem 12.3.1.)

12.3.3 Star-Decomposable Triplets

In order to test whether a given three-variable distribution $P(x_1, x_2, x_3)$ is star-decomposable, we first solve (12.24) and express the parameters α, f_i, g_i as a function of the parameters specifying P . This task was carried out by Lazarfeld [16] in terms of the seven joint-occurrence probabilities.

$$\begin{aligned} p_i &= P(x_i = 1), \\ p_{ij} &= P(x_i = 1, x_j = 1), \\ p_{ijk} &= P(x_i = 1, x_j = 1, x_k = 1), \end{aligned} \quad (12.27)$$

and led to the following solution:

Define the quantities,

$$[ij] = p_{ij} - p_i p_j, \quad (12.28)$$

$$S_i = \left[\frac{[ij][ik]}{[jk]} \right]^{1/2}, \quad (12.29)$$

$$\mu_i = \frac{(p_i p_{ijk} - p_{ij} p_{ik})}{[jk]}, \quad (12.30)$$

$$K = \frac{S_i}{p_i} - \frac{p_i}{s_i} + \frac{\mu_i}{S_i p_i}, \quad (12.31)$$

and let t be the solution of

$$t^2 + Kt - 1 = 0. \quad (12.32)$$

The parameters α, f_i, g_i are given by:

$$\alpha = t^2 / (1 + t^2), \quad (12.33)$$

$$f_i = p_i + S_i [(1 - \alpha) / \alpha]^{1/2}, \quad (12.34)$$

$$g_i = p_i - S_i [\alpha / (1 - \alpha)]^{1/2}. \quad (12.35)$$

Moreover, the differences $f_i - g_i$ are independent of p_{ijk} :

$$f_i - g_i = S_i = \left[\frac{[ij][ik]}{[jk]} \right]^{1/2}. \quad (12.36)$$

The conditions for star-decomposability are obtained by requiring that preceding solutions satisfy:

- (a) S_i should be real,
- (b) $0 \leq f_i \leq 1$,
- (c) $0 \leq g_i \leq 1$.

Using the variances

$$\sigma_i = [p_i(1 - p_i)]^{1/2} \quad (12.37)$$

and the correlation coefficients

$$\rho_{ij} = (p_{ij} - p_i p_j) / \sigma_i \sigma_j, \quad (12.38)$$

requirement (a) is equivalent to the condition that all three correlation coefficients are nonnegative. (If two of them are negative, we can rename two variables by their complements; the newly defined triplet will have all its pairs positively correlated.) We shall call triplets with this property *positively correlated*.

This, together with requirements (b) and (c), yields (see Appendix 12.B):

Theorem 12.3.1 *A necessary and sufficient condition for three dichotomous random variables to be star-decomposable is that they are positively correlated, and that the inequality,*

$$\frac{p_{ik} p_{ij}}{p_i} \leq p_{ijk} \leq \frac{p_{ik} p_{ij}}{p_i} + \sigma_j \sigma_k (\rho_{jk} - \rho_{ij} \rho_{ik}), \quad (12.39)$$

is satisfied for all $i \in \{1, 2, 3\}$. When this condition is satisfied, the parameters of the star-decomposed distribution can be determined uniquely, up to a complementation of the hidden variable w , i.e., $w \rightarrow (1 - w)$, $f_i \rightarrow g_i$, $\alpha \rightarrow (1 - \alpha)$.

Obviously, in order to satisfy (12.39), the term $(\rho_{jk} - \rho_{ij} \rho_{ik})$ must be nonnegative. This introduces a simple necessary condition for star-decomposability that may be used to quickly rule out many likely candidates.

Corollary 12.3.2 *A necessary condition for a distribution $P(x_1, x_2, x_3)$ to be star-decomposable is that all correlation coefficients obey the triangle inequality:*

$$\rho_{jk} \geq \rho_{jk} \rho_{ik}. \quad (12.40)$$

Inequality (12.40) is satisfied with equality if w coincides with x_i , i.e., when x_j and x_k are independent, given x_i . Thus, an intuitive interpretation of this corollary is that the correlation between any two variables must be stronger than that induced by their dependencies on the third variable; a mechanism accounting for direct dependencies must be present.

Having established the criterion for star-decomposability, we may address a related problem. Suppose P is not star-decomposable. Can it be approximated by a star-decomposable distribution \hat{P} that has the same second-order probabilities?

The preceding analysis contains the answer to this question. Note that the third-order statistics are represented only by the term p_{ijk} , and this term is confined by

(12.39) to a region whose boundaries are determined by second-order parameters. Thus, if we insist on keeping all second-order dependencies of P intact and are willing to choose p_{ijk} so as to yield a star-decomposable distribution, we can only do so if the region circumscribed by (12.39) is nonempty. This leads to the statement:

Theorem 12.3.3 *A necessary and sufficient condition for the second-order dependencies among the triplet x_1, x_2, x_3 to support a star-decomposable extension is that the six inequalities,*

$$\frac{p_{ij}p_{ik}}{p_i} \leq x \leq \frac{p_{ij}p_{ik}}{p_i} + \sigma_j\sigma_k(\rho_{jk} - \rho_{ij}\rho_{ik}), \quad i = 1, 2, 3, \quad (12.41)$$

possess a solution for x .

12.3.4 A Tree-Reconstruction Procedure

We are now ready to confront the central problem of this section—given a tree-decomposable distribution $P(x_1, \dots, x_n)$, can we uncover its underlying topology and the underlying tree-distribution $P_T(x_1, \dots, x_n, w_1, \dots, w_m)$?

The construction method is based on the observation that any three leaves in a tree have one, and only one, internal node that can be considered their *center*, i.e., it lies on all the paths connecting the leaves to each other. If one removes the center, the three leaves become disconnected from each other. This means that, if P is tree-decomposable, then the joint distribution of any triplet of variables x_i, x_j, x_k is star-decomposable, i.e., $P(x_i, x_j, x_k)$ uniquely determines the parameters α, f_i, g_i as in (12.33), (12.34), and (12.35), where α is the marginal probability of the central variable. Moreover, if we compute the star decompositions of two triplets of leaves, both having the same central node w , the two distributions should have the same value for $\alpha = P_T(w = 1)$. This provides us with a basic test for verifying whether two arbitrary triplets of leaves share a common center, and a successive application of this test is sufficient for determining the structure of the entire tree.

Consider a 4-tuple x_1, x_2, x_3, x_4 of leaves in T . These leaves are interconnected through one of the four possible topologies shown in Figure 12.8. The topologies differ in the identity of the triplets which share a common center. For example, in the topology of Figure 12.8(a) the pair $[(1, 2, 3), (1, 2, 4)]$ share a common center, and so does the pair $[(1, 3, 4), (2, 3, 4)]$. In Figure 12.8(b), on the other hand, the sharing

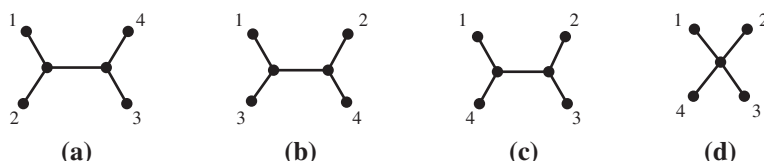


Figure 12.8 The four possible topologies by which four leaves can be related.

pairs are $[(1, 2, 4), (2, 4, 3)]$ and $[(1, 3, 4), (2, 1, 3)]$, and in Figure 12.8(d) all triplets share the same center. Thus, the basic test for center-sharing triplets enables us to decide the topology of any 4-tuple and, eventually, to configure the entire tree.

We start with any three variables x_1, x_2 , and x_3 , form their star decomposition, choose a fourth variable, x_4 , and ask to which leg of the star should x_4 be joined. We can answer this question easily by testing which pairs of triplets share centers, deciding on the appropriate topology and connecting x_4 accordingly. Similarly, if we already have a tree structure T_i , with i leaves, and we wish to know where to join the $(i + 1)$ th leaf, we can choose any triplet of leaves from T_i with central variable w and test to which leg of w should x_{i+1} be joined. This, in turn, identifies a subtree T'_i of T_i that should receive x_{i+1} and permits us to remove from further consideration the subtrees emanating from the unselected legs of w . Repeating this operation on the selected subtree T'_i will eventually reduce it to a single branch, to which x_{i+1} is joined.

It is possible to show [26] that, if we choose, in each state, a central variable that splits the available tree into subtrees of roughly equal size, the joining branch of x_{i+1} can be identified in, at most, $\log_{k/(k-1)}(i)$ tests, where k is the maximal degree of the T_i . This amounts to $O(n \log n)$ test for constructing an entire tree of n leaves.

So far, we have shown that the structure of the tree T can be uncovered uniquely. Next we show that the distribution P_T is, likewise, uniquely determined from P , i.e., that we can determine all the functions $P(x_i|w_j)$ and $P(w_j|w_k)$ in (12.26), for $i = 1, \dots, n$ and $j, k = 1, 2, \dots, m$. The functions $P(x_i|w_j)$ assigned to the peripheral branches of the tree are determined directly from the star decomposition of triplets involving adjacent leaves. In Figure 12.7, for example, the star decomposition of $P(x_1, x_2, x_5)$ yields $P(x_1|w_1)$ and $P(x_2|w_1)$. The conditional probabilities $P(w_j|w_k)$ assigned to interior branches are determined by solving matrix equations. For example, $P(x_1|w_2)$ can be obtained from the star decomposition of (x_1, x_5, x_7) , and it is related to $P(x_1|w_1)$ via

$$P(x_1|w_2) = \sum_{w_1} P(x_1|w_1)P(w_1|w_2).$$

This matrix equation has a solution for $P(w_1|w_2)$ because $P(x_1|w_1)$ must be nonsingular. It is only singular when $f_1 = g_1$, i.e., when x_1 is independent of w_1 and is therefore independent of all other variables. Hence, we can determine the parameters of the branches next to the periphery, use them to determine more interior branches, and so on, until all the interior conditional probabilities $P(w_i|w_j)$ are determined.

Next, we shall show that the tree structure can be recovered without resorting to third order probabilities; correlations among pairs of leaves suffice. This feature stems from the observation that, when two triplets of a 4-tuple are

star-decomposable with respect to the same central variable w (e.g. (1, 2, 3) and (1, 2, 4) in Figure 12.8(a)), then not only are the values of α the same, but the f - and g -parameters associated with the two common variables (e.g., 1 and 2 in Figure 12.8(a)) must also be the same. While the value of α depends on a third-order probability, the difference $f_i - g_i$ depends only on second-order terms via (12.36). Thus, requiring that $f_1 - g_1$ in Figure 12.8(a) obtain the same value in the star decomposition of (1, 2, 3) as in that of (1, 2, 4) leads to the equation:

$$[12][13]/[23] = [12][14]/[24] \tag{12.42}$$

which, using (12.28), yields

$$\rho_{13}\rho_{42} = \rho_{14}\rho_{32}. \tag{12.43}$$

An identical equality will be obtained for each $f_i - g_i$, $i = 1, 2, 3, 4$, relative to the topology of Figure 12.8(a). Similarly, the topology of Figure 12.8(b) dictates

$$\rho_{12}\rho_{43} = \rho_{14}\rho_{23} \tag{12.44}$$

and that of Figure 12.8(c) dictates:

$$\rho_{12}\rho_{34} = \rho_{13}\rho_{24}. \tag{12.45}$$

Thus, we see that each of these three topologies is characterized by its own distinct equality, while the topology of Figure 12.8(d) is distinguished by all three equalities holding simultaneously. This provides the necessary second-order criterion for deciding the topology of any 4-tuple tested: if the equality $\rho_{ij}\rho_{kl} = \rho_{ik}\rho_{jl}$ holds for some permutation of the indices, we decide on the topology



If it holds for two permutations with distinct topologies, the entire 4-tuple is star-decomposable. Note that the equality $\rho_{ij}\rho_{kl} = \rho_{ik}\rho_{jl}$ must hold for at least one permutation of the variables or else the 4-tuple would not be tree-decomposable.

12.3.5 Conclusions and Open Questions

This section provides an operational definition for entities called “hidden causes,” which are not directly observable but facilitate the acquisition of effective causal models from empirical data. Hidden causes are viewed as dummy variables which, if held constant, induce probabilistic independence among sets of visible variables. It is shown that if all variables are bivalued and if the activities of

the visible variables are governed by a tree-decomposable probability distribution, then the topology of the tree can be uncovered uniquely from the observed correlations between pairs of variables. Moreover, the structuring algorithm requires only $n \log n$ steps.

The method introduced in this paper has two major shortcomings: It requires precise knowledge of the correlation coefficients, and it works only when there exists an underlying model that is tree-structured. In practice, we often have only sample estimates of the correlation coefficients; therefore, it is unlikely that criteria based on equalities (as in (12.43)) will ever be satisfied exactly. It is possible, of course, to relax these criteria and make topological decisions by seeking proximities rather than equalities. For example, instead of searching for an equality $\rho_{ij}\rho_{kl} = \rho_{ik}\rho_{jl}$, we can decide the 4-tuple topology on the basis of the permutation of indices that minimizes the difference $\rho_{ij}\rho_{kl} - \rho_{ik}\rho_{jl}$. Experiments show, however, that the structure which evolves from such a method is very sensitive to inaccuracies in the estimates ρ_{ij} , because no mechanism is provided to retract erroneous decisions made in the early stages of the structuring process. Ideally, the topological membership of the $(i+1)$ th leaf should be decided not merely by its relations to a single triplet of leaves chosen to represent an internal node w but also by its relations to all previously structured triplets which share w as a center. This, of course, will substantially increase the complexity of the algorithm.

Similar difficulties plague the task of finding the best tree-structured *approximation* for a distribution which is not tree-decomposable. Even though we argued that natural data which lend themselves to causal modeling should be representable as tree-decomposable distributions, these distributions may contain internal nodes with more than two values. The task of determining the parameters associated with such nodes is much more complicated and, in addition, rarely yields unique solutions. Unique solutions, as shown in Section 12.3.4, are essential for building large structures from smaller ones. We leave open the question of explaining how approximate causal modeling, an activity which humans seem to perform with relative ease, can be embodied in computational procedures that are both sound and efficient.

12.A

12.A.1

Appendix A. Derivation of the Updating Rules for Singly Connected Networks

Updating BEL

Starting with

$$\text{BEL}(A_i) \triangleq P(A_i | D_{BA}^+, D_{CA}^+, D_{AX}^-, D_{AY}^-),$$

we apply Bayes' rule, and obtain

$$\text{BEL}(A_i) = \alpha P(D_{AX}^-, D_{AY}^- | A_i, D_{BA}^+, D_{CA}^+) P(A_i | D_{BA}^+, D_{CA}^+).$$

The conditional independence of (12.14) now yields (12.15):

$$\text{BEL}(A_i) = \alpha P(D_{AX}^-, A_i) P(D_{AY}^- | A_i) P(D_{AY}^- | A_i) P(A_i | D_{BA}^+, D_{CA}^+).$$

Conditioning and summing over the values of B and C , we get

$$\begin{aligned} \text{BEL}(A_i) &= \alpha P(D_{AX}^- | A_i) P(D_{AY}^- | A_i) \\ &\quad \cdot \sum_{B,C} P(A_i | D_{BA}^+, D_{CA}^+, B, C) P(B, C | D_{BA}^+, D_{CA}^+) \\ &= \alpha P(D_{AX}^- | A_i) P(D_{AY}^- | A_i) \\ &\quad \cdot \sum_{B,C} P(A_i | B, C) P(B | D_{BA}^+) P(C | D_{CA}^+) \end{aligned}$$

making use of the fact that B and C are independent, given data from nondescendants of A . This confirms (12.16):

$$\begin{aligned} \text{BEL}(A_i) &= \alpha P(D_{AX}^- | A_i) P(D_{AY}^- | A_i) \\ &\quad \cdot \left[\sum_{j,k} P(A_i | B_j, C_k) P(B_j | D_{BA}^+) P(C_k | D_{CA}^+) \right] \end{aligned}$$

and, using the λ - π notation

$$\lambda_X(A_i) = P(D_{AX}^- | A_i), \quad \pi_A(B_j) = P(B_j | D_{BA}^+),$$

we obtain (12.19)

$$\text{BEL}(A_i) = \alpha \lambda_X(A_i) \lambda_Y(A_i) \left[\sum_{j,k} P(A_i | B_j, C_k) \pi_A(B_j) \pi_A(C_k) \right].$$

12.A.2 Updating π

$$\begin{aligned} \pi_X(A_i) &= P(A_i | D_{AX}^+) = P(A_i | D - D_{AX}^-) \\ &= \text{BEL}(A_i | \lambda_X(A)) = (1, 1, \dots, 1) \\ &= \alpha \lambda_Y(A_i) \left[\sum_{j,k} P(A_i | B_j, C_k) \pi_A(B_j) \pi_A(C_k) \right], \end{aligned}$$

thus confirming (12.22).

12.A.3 Updating λ

$$\begin{aligned}
 \lambda_A(B_i) &= P(D_{AB}^-|B_i) = P(A, D_{AX}^-, D_{AY}^-, D_{CA}^+|B_i) \\
 &= \sum_{j,k} P(D_{AX}^-, D_{AY}^-, D_{CA}^+|B_i, C_j, A_k)P(C_j, A_k|B_i) \\
 &= \sum_{j,k} P(D_{AX}^-|A_k)P(D_{AY}^-|A_k)P(D_{CA}^+|C_j) \\
 &\quad \cdot P(A_k|B_i, C_j)P(C_j|B_i).
 \end{aligned}$$

But $P(C_j|B_i) = P(C_j)$ because B and C are marginally independent, and

$$P(D_{CA}^+|C_j)P(C_j) = \alpha P(C_j|D_{CA}^+)$$

by Bayes' rule. Therefore,

$$\begin{aligned}
 \lambda_A(B_i) &= \alpha \sum_{j,k} P(D_{AX}^-|A_k)P(D_{AY}^-|A_k)P(C_j|D_{CA}^+)P(A_k|C_j, B_i) \\
 &= \alpha \sum_{j,k} \lambda_X(A_k)\lambda_Y(A_k)\pi_A(C_j)P(A_k|B_i, C_j) \\
 &= \alpha \sum_j \left[\pi_A(C_j) \sum_k \lambda_X(A_k)\lambda_Y(A_k)P(A_k|B_i, C_j) \right],
 \end{aligned}$$

which confirms (12.21).

12.B Appendix B. Conditions for Star-decomposability

Let

$$\begin{aligned}
 p_i &= P(x_i = 1), \\
 p_{ij} &= P(x_i = 1, x_j = 1), \\
 p_{ijk} &= P(x_i = 1, x_j = 1, x_k = 1).
 \end{aligned} \tag{12.B.1}$$

The seven joint-occurrence probabilities, $p_1, p_2, p_3, p_{12}, p_{13}, p_{23}, p_{123}$, uniquely define the seven parameters necessary for specifying $P(x_1, x_2, x_3)$. For example:

$$\begin{aligned}
 P(x_1 = 1, x_2 = 1, x_3 = 0) &= p_{12} - p_{123}, \\
 P(x_1 = 1, x_2 = 0) &= p_1 - p_2, \text{ etc.}
 \end{aligned}$$

These probabilities will be used in the following analysis.

Assuming P is star-decomposable (equations (12.23) and (12.24)), we can express the joint-occurrence probabilities in terms of α, f_i, g_i and obtain seven equations for these seven parameters.

$$p_i = \alpha f_i + (1 - \alpha)g_i, \quad (12.B.2)$$

$$p_{ij} = \alpha f_i f_j + (1 - \alpha)g_i g_j, \quad (12.B.3)$$

$$p_{ijk} = \alpha f_i f_j f_k + (1 - \alpha)g_i g_j g_k. \quad (12.B.4)$$

These equations can be manipulated to yield product forms on the right-hand sides:

$$p_{ij} - p_i p_j = \alpha(1 - \alpha)(f_i - g_i)(f_j - g_j), \quad (12.B.5)$$

$$p_i p_{ijk} - p_{ij} p_{ik} = \alpha(1 - \alpha) f_i g_i (f_j - g_j)(f_k - g_k). \quad (12.B.6)$$

Equation (12.B.5) comprises three equations which can be solved for the differences $f_i - g_i, i = 1, 2, 3$, giving

$$f_i - g_i = S_i = \pm \left[\frac{[ij][ik]}{[jk]} \right]^{1/2}, \quad (12.B.7)$$

where the bracket $[ij]$ stands for the determinant

$$[ij] = p_{ij} - p_i p_j. \quad (12.B.8)$$

These, together with (12.B.2), determine f_i and g_i in terms of S_i and α (still unknown):

$$f_i = p_i + S_i [(1 - \alpha)/\alpha]^{1/2}, \quad (12.B.9)$$

$$g_i = p_i - S_i [\alpha/(1 - \alpha)]^{1/2}. \quad (12.B.10)$$

To determine α , we invoke (12.B.6) and obtain

$$[\alpha/(1 - \alpha)]^{1/2} = t \quad \text{or} \quad \alpha = t^2/(1 - t^2), \quad (12.B.11)$$

where t is a solution to

$$t^2 + Kt - 1 = 0, \quad (12.B.12)$$

and K is defined by:

$$K = \frac{S_i}{p_i} - \frac{p_i}{S_i} + \frac{\mu_i}{S_i p_i}, \quad (12.B.13)$$

$$\mu_i = [jk, i]/[jk] = (p_i p_{ijk} - p_{ij} p_{ik})/[jk]. \quad (12.B.14)$$

It can be easily verified that K (and, therefore, α) obtains the same value regardless of which index i provides the parameters in (12.B.13).

From (12.B.13) we see that the parameters S_i and μ_i of P govern the solutions of (12.B.12) which, in turn, determine whether P is star-decomposable via the resulting values of α, f_i, g_i . These conditions are obtained by requiring that:

- (a) S_i be real,
- (b) $0 \leq f_i \leq 1$,
- (c) $0 \leq g_i \leq 1$.

Requirement (a) implies that, of the three brackets in (12.B.7), either all three are nonnegative, or exactly two are negative. These brackets are directly related to the correlation coefficient via:

$$\rho_{ij} = [ij][p_i(1-p_i)]^{-1/2}[p_j(1-p_j)]^{-1/2} = [ij]/\sigma_i\sigma_j \quad (12.B.15)$$

and so, requirement (a) is equivalent to the condition that all three correlation coefficients are nonnegative. If two of them are negative, we can rename two variables by their complements; the newly defined triplet will have all its pairs positively correlated.

Now attend to requirement (b). Equation (12.B.9) shows that f_i can be negative only if S_i is negative, i.e., if S_i is identified with the negative square root in (12.B.7). However, the choice of negative S_i yields a solution (f'_i, g'_i, α') which is symmetrical to (f_i, g_i, α) stemming from a positive S_i , with $f'_i = g_i, g'_i = f_i, \alpha' = 1 - \alpha$. Thus, S_i and f_i can be assumed to be nonnegative, and it remains to examine the condition $f_i \leq 1$ or, equivalently, $t \geq S_i/(1-p_i)$ (see (12.B.9) and (12.B.11)). Imposing this condition in (12.B.12) translates to:

$$p_{ijk} \leq p_{ij}p_{ik}/p_i + \sigma_k\sigma_j[\rho_{jk} - \rho_{ij}\rho_{ik}]. \quad (12.B.16)$$

Similarly, inserting requirement (c), $g_i \geq 0$, in (12.B.12) yields the inequality:

$$p_{ik}p_{ij}/p_i \leq p_{ijk} \quad (12.B.17)$$

which, together with (12.B.16), lead to Theorem 12.3.1.

Acknowledgments

I thank many people for helping me prepare this manuscript. N. Dalkey has called my attention to the work of Lazarfeld [16] and has provided a continuous stream of valuable advice. Thomas Ferguson made helpful comments on Section 12.3. Jin

Kim is responsible for deriving the propagation equations of Section 12.2.3 [14] and for using the propagation scheme in his decision support system CONVINC [13]. Michael Tarsi has helped develop the tree-construction algorithm of Section 12.3.4 and has proved its optimality [26]. Eli Gafni has spent many hours discussing the relations among Markov fields, stochastic relaxation, and belief propagation. Ed Pednault, Nils Nilsson, Henry Kyburg, and Igor Roizen have thoroughly reviewed earlier versions of this paper and have suggested many improvements.

The results of Section 12.3 were presented at the International Joint Conference on Artificial Intelligence, University of California, Los Angeles, August 19–23, 1985.

References

1. Anderson, J.R., *The Architecture of Cognition* (Harvard University Press, Cambridge, MA, 1983).
2. Chow, C.K. and Liu, C.N., Approximating discrete probability distributions with dependence trees, *IEEE Trans. Inf. Theory* **14** (1968) 462–467.
3. Dechter, R. and Pearl, J., The anatomy of easy problems: A constraint-satisfaction formulation, in: *Proceedings Ninth International Joint Conference on Artificial Intelligence*, Los Angeles, CA, (1985) 1066–1072.
4. Dell, G.S., Positive feedback in hierarchical connectionist models: Applications to language production, *Cognitive Sci.* **9** (1) (1985) 3–24.
5. Doyle, J. A truth maintenance system, *Artificial Intelligence* **12** (1979) 231–272.
6. Duda, R.O., Hart, P.E. and Nilsson, N.J., Subjective Bayesian methods for rule-based inference systems, in: *Proceedings 1976 National Computer Conference (AFIPS Conference Proceedings)* **45** (1976) 1075–1082.
7. Freuder, E.C., A sufficient condition of backtrack-free search, *J. ACM* **29** (1) (1982) 24–32.
8. Geman, S. and Geman, D., Stochastic relaxations, Gibbs distributions, and the Bayesian restoration of images, *IEEE Trans. Pattern Anal. Machine Intelligence* **6** (6) (1984) 721–742.
9. Hinton, G.E., Sejnowski, T.J. and Ackley, D.H., Boltzman machines: Constraint satisfaction networks that learn, Tech. Rept. CMU-CS-84-119, Department of Computer Science, Carnegie-Mellon University, Pittsburgh, PA, 1984.
10. Howard, R.A. and Matheson, J.E., Influence diagrams, in: R.A. Howard and J.E. Matheson (Eds.), *The Principles and Applications of Decision Analysis* (Strategic Decisions Group, Menlo Park, CA, 1984).
11. Jeffrey, R., *The Logic of Decisions* (McGraw-Hill, New York, 1965).
12. Kemeny, J.G., Snell, J.L. and Knapp, A.W., *Denumerable Markov Chains* (Springer, Berlin, 2nd ed., 1976).

13. Kim, J., CONVINCENCE: A CONVERSATIONAL INFERENCE CONSOLIDATION ENGINE, Ph.D. Dissertation, University of California, Los Angeles, CA, 1983.
14. Kim, J. and Pearl, J., A computational model for combined causal and diagnostic reasoning in inference systems, in: *Proceedings Eighth International Joint Conference on Artificial Intelligence*. Karlsruhe, F.R.G. (1983) 190–193.
15. Lauritzen, S.L., *Lectures on Contingency Tables* (University of Aalborg Press, Aalborg, Denmark, 2nd ed., 1982).
16. Lazarsfeld, P.F., Latent structure analysis, in: S.A. Stouffer, L. Guttman, E.A. Suchman, P.F. Lazarsfeld, S.A. Star and J.A. Claussen (Eds.), *Measurement and Prediction* (Wiley, New York, 1966).
17. Lesser, V.R. and Erman, L.D., A retrospective view of HEARSAY II architecture, in: *Proceedings Fifth International Joint Conference on Artificial Intelligence*, Cambridge, MA (1977) 790–800.
18. Levy, H. and Low, D.W., A new algorithm for finding small cycle cutsets, Rept. G 320-2721, IBM Los Angeles Scientific Center, Los Angeles, CA, 1983.
19. Lowrance, J.D., Dependency-graph models of evidential support, COINS Tech. Rept. 82-26, University of Massachusetts at Amherst, MA, 1982.
20. McAllester, D., An outlook on truth maintenance, AIM-551, Artificial Intelligence Laboratory, MIT, Cambridge, MA, 1980.
21. Pearl, J., Reverend Bayes on inference engines: A distributed hierarchical approach, in: *Proceedings Second National Conference on Artificial Intelligence*, Pittsburgh, PA (1982) 133–136.
22. Pearl, J., A constraint-propagation approach to probabilistic reasoning, in: *Proceedings Workshop on Uncertainty and Probability in AI*, Los Angeles, CA (1985) 31–42; also in: L.N. Kanal and J.F. Lemmer (Eds.), *Uncertainty in Artificial Intelligence* (North-Holland, Amsterdam, 1986) 357–370.
23. Pearl, J., Distributed diagnosis in causal models with continuous variables, Tech. Rept. CSD-860051, Cognitive Systems Laboratory, Computer Science Department, University of California, Los Angeles, 1985.
24. Pearl, J. and Paz, A., Graphoids: A graph-based logic for reasoning about relevancy relations, Tech. Rep. CSD-850038, Cognitive Systems Laboratory, Computer Science Department, University of California, Los Angeles, 1985.
25. Pearl, J., On evidential reasoning in a hierarchy of hypotheses, *Artificial Intelligence* 28 (1986) 9–15.
26. Pearl, J. and Tarsi, M., Structuring causal trees, *J. Complexity* 2 (1) (1986) 60–77.
27. Rosenfeld, A., Hummel, A. and Zucker, S., Scene labelling by relaxation operations, *IEEE Trans. Syst. Man Cybern.* 6 (1976) 420–433.
28. Rumelhart, D.E., Toward an interactive model of reading, *Center for Human Information Proceedings CHIP-56*, University of California, San Diego, La Jolla, CA, 1976.

29. Rumelhart, D.E. and McClelland, J.L., An interactive activation model of context effects in letter perception: Part 2. The contextual enhancement effect and some tests and extensions of the model, *Psychol. Rev.* **89** (1982) 60–94.
30. Shastri, L. and Feldman, J.A., Semantic networks and neural nets, TR-131, Computer Science Department, The University of Rochester, Rochester, NY, 1984.
31. Simon, H.A., Spurious correlations: A causal interpretation, *J. Am. Stat. Assoc.* **49** (1954) 469–492.
32. Spiegelhalter, D.J., Probabilistic reasoning in predictive expert systems, in: L.N. Kanal and J.F. Lemmer (Eds.), *Uncertainty in Artificial Intelligence* (North-Holland, Amsterdam, 1986) 47–68.
33. Suppes, P., *A Probabilistic Theory of Causality* (North-Holland, Amsterdam, 1970).
34. Tverski, A. and Kahneman, D., Causal schemata in judgments under uncertainty, in: M. Fishbein (Ed.), *Progress in Social Psychology* (Erlbaum, Hillsdale, NJ., 1977).
35. Waltz, D.G., Generating semantic descriptions from drawings of scenes with shadows, AI TR-271, Artificial Intelligence Laboratory, Cambridge, MA, 1972.

Received January 1982; revised version received February 1986

GRAPHOIDS: Graph-Based Logic for Reasoning about Relevance Relations^{*} *Or*

When Would x Tell You More about y If You Already Know z ?

Judea Pearl and Azaria Paz[†]

Abstract

We consider 3-place relations $I(x, z, y)$ where x , y , and z are three non-intersecting sets of elements (e.g., propositions), and $I(x, z, y)$ stands for the statement: “Knowing z renders x irrelevant to y .” We give sufficient conditions on I for the existence

^{*}This work was supported in part by the National Science Foundation, Grants DCR 83-13875 & 85-01234.

[†]Technion – Israel Institute of Technology

Advances in Artificial Intelligence – II B. Du Boulay, D. Hogg and L. Steels (Editors)

Judea Pearl and Azaria Paz. 1986. Graphoids: graph-based logic for reasoning about relevance relations or when would x tell you more about y if you already know z ? In Proceedings of the 7th European Conference on Artificial Intelligence - Volume 2 (ECAI'86). North-Holland, 357–363. ISBN 9780444702791.

© Elsevier Science Publishers B.V. (North-Holland), 1987. Republished with permission from Elsevier.

of a (minimal) graph G such that $I(x, z, y)$ can be validated by testing whether z separates x from y in G . These conditions define a GRAPHOID. The theory of graphoids uncovers the axiomatic basis of information relevance (e.g., probabilistic dependencies) and ties it to vertex-separation conditions in graphs. The defining axioms can also be viewed as inference rules for deducing which propositions are relevant to each other, given a certain state of knowledge.

13.1 Introduction

Any system that reasons about knowledge and beliefs must make use of information about relevancies. If we have acquired a body of knowledge z and now wish to assess the truth of proposition x , it is important to know whether it would be worthwhile to consult another proposition y , which is not in z . In other words, before we consult y we need to know if its truth value can potentially generate new information relative to x , information not available from z . For example, in trying to predict whether I am going to be late for a meeting, it is normally a good idea to ask somebody on the street for the time. However, once I establish the precise time by listening to the radio, asking people for the time becomes superfluous and their responses would be irrelevant. Similarly, knowing the color of X 's car normally tells me nothing about the color of Y 's. However, if X were to tell me that he almost mistook Y 's car for his own, the two pieces of information become relevant to each other. What logic would facilitate this type of reasoning?

In probability theory, the notion of relevance is given precise quantitative underpinning using the device of conditional independence. A variable x is said to be independent of y given the information z if

$$P(x, y | z) = P(x | z) P(y | z).$$

However, it is rather unreasonable to expect people or machines to resort to numerical verification of equalities in order to extract relevance information. The ease and conviction with which people detect relevance relationships strongly suggest that such information is readily available from the organizational structure of human memory, not from numerical values assigned to its components. Accordingly, it would be interesting to explore how assertions about relevance can be tested in various models of memory and, in particular, whether such assertions can be derived by simple manipulations on graphs.

Graphs offer useful representations for a variety of phenomena. They give vivid visual display for the essential relations in the phenomenon and provide a convenient medium for people to communicate and reason about it. Graph-related concepts are so entrenched in our language that one wonders whether people can

in fact reason any other way, except by tracing links and arrows and paths in some mental representation of concepts and relations. Therefore, if we aspire to use non-numeric logic to mimic human reasoning about knowledge and beliefs, we should make sure that most derivational steps in that logic correspond to simple operations on some graphs.

When we deal with a phenomenon where the notion of neighborhood or connectedness is explicit (e.g., family relations, electronic circuits, communication networks, etc.) we have no problem configuring a graph which represents the main features of the phenomenon. However, in modelling conceptual relations such as causation, association and relevance, it is often hard to distinguish direct neighbors from indirect neighbors; so, the task of constructing a graph representation then becomes more delicate.

This paper studies the feasibility of devising graphoid representations for relational structures in which the notion of neighborhood is not specified in advance. Rather, what is given explicitly is the relation of “in betweenness.” In other words, we are given the means to test whether any given subset S of elements *intervenes* in a relation between elements x and y , but it remains up to us to decide how to connect the elements together in a graph that accounts for these interventions.

The notion of conditional independence in probability theory is a perfect example of such a relational structure. For a given probability distribution P and any three variables x, y, z , while it is fairly easy to verify whether knowing z renders x independent of y , P does not dictate which variables should be regarded as direct neighbors. Thus, many topologies might be used to display the dependencies embodied in P .

The theory of graphoids establishes a clear correspondence between probabilistic dependencies and graph representation. It tells us how to construct a unique edge-minimum graph G such that each time we observe a vertex x separated from y by a subset S of vertices, we can be guaranteed that variables x and y are independent given the values of the variables in S . Moreover, the set of neighbors assigned by G to each x coincides exactly with the boundary of x , i.e., the smallest set of variables needed to shield x from the influence of all other variables in the system. This construction is further extended by the theory of graphoids to cases where the notion of independence is not given probabilistically or numerically. We now ask what *logical* conditions should constrain the relationship: $I(x, z, y) =$ “knowing z renders x irrelevant to y ” so that we can validate it by testing whether z separates x from y in some graph G . We show that two main conditions (together with

symmetry and subset closure) are sufficient:

$$\text{weak closure for intersection: } I(x, z \cup w, y) \ \& \ I(x, z \cup y, w) \implies I(x, z, y \cup w) \quad (13.1)$$

$$\text{weak closure for union: } I(x, z, y \cup w) \implies I(x, z \cup w, y). \quad (13.2)$$

Loosely speaking, (13.1) states that if y does not affect x when w is held constant and if, simultaneously, w does not affect x when y is held constant, then neither w nor y can affect x . (13.2) states that learning an irrelevant fact (w) cannot help another irrelevant fact (y) become relevant. Condition (13.1) is sufficient to guarantee a unique construction of an edge-minimum graph G that validates $I(x, z, y)$ by vertex separation. Condition (13.2) guarantees that the neighborhoods defined by the edges of G coincide with the relevance boundaries defined by I . These two conditions are chosen as the defining axioms of graphoids, and are shown to account for the graphical properties of probabilistic dependencies.

This paper is organized as follows: In Section 13.2 we exemplify a graphoid system using probabilistic dependencies and their graphical representations. Section 13.3 introduces an axiomatic definition of graphoids, and states (without proofs) their graph-representation properties; the proofs can be found in [Pearl and Paz 1985]. Section 13.4 discusses a few extensions and outlines open problems.

13.2 Probabilistic Dependencies and their Graphical Representation

Let $U = \{\alpha, \beta, \dots\}$ be a finite set of discrete-valued random variables characterized by a joint probability function $P(\cdot)$, and let x, y , and z stand for any three subsets of variables in U . We say that x and y are conditionally independent given z if

$$P(x, y | z) = P(x | z) P(y | z) \quad \text{when } P(z) > 0. \quad (13.3)$$

Eq. (13.3) is a terse notation for the assertion that for any instantiation z_k of the variables in z and for any instantiation x_i and y_j of x and y , we have

$$P(x = x_i \text{ and } y = y_j | z = z_k) = P(x = x_i | z = z_k) P(y = y_j | z = z_k). \quad (13.4)$$

The requirement $P(z) > 0$ guarantees that all the conditional probabilities are well defined, and we shall henceforth assume that $P > 0$ for any instantiation of the variables in U . This rules out logical and functional dependencies among the variables, a case which would require special treatment.

We shall use $(x \perp z \perp y)_P$ or simply $(x \perp z \perp y)$ to denote the independence of x and y given z . Thus,

$$(x \perp z \perp y)_P \iff P(x, y | z) = P(x | z) P(y | z) \iff P(x | y, z) = P(x | z). \quad (13.5)$$

Note that $(x \perp z \perp y)$ implies the conditional independence of all pairs of variables $\alpha \varepsilon x$ and $\beta \varepsilon y$, but the converse is not necessarily true.

The relation $(x \perp z \perp y)$ satisfies the following logical independent properties:

$$\text{Symmetry:} \quad (x \perp z \perp y) \iff (y \perp z \perp x) \quad (13.6a)$$

$$\text{Closure for Subsets:} \quad (x \perp z \perp y, w) \implies P(x \perp z \perp y) \& (x \perp z \perp w) \quad (13.6b)$$

$$\text{Weak Closure for Intersection:} \quad (x \perp y, z \perp w) \& (x \perp y, w \perp z) \implies (x \perp y \perp z, w) \quad (13.6c)$$

$$\text{Weak Closure for Union:} \quad (x \perp y \perp z, w) \implies (x \perp y, z \perp w) \quad (13.6d)$$

$$\text{Contraction:} \quad (x \perp y, z \perp w) \& (x \perp y \perp z) \implies (x \perp y \perp z, w) \quad (13.6e)$$

While the properties in (13.5) characterize the numeric representation of P , those in (13.6) are purely logical, void of any association with numerical forms and can be viewed, therefore, as an axiomatic definition of conditional independence. A graphical interpretation for properties (13.6c) through (13.6e) can be obtained by envisioning the chain $x-y-z-w$ and associating the triplet $(x \perp z \perp y)$ with the statement “ z separates x from y ” or “ z intervenes between x and y .”

Ideally, dependent variables should be displayed as connected nodes in some graph G and independent variables as unconnected nodes. We would also like to require that if the removal of some subset S of nodes from the graph renders nodes x and y disconnected, written $\langle x|S|y \rangle_G$, then this separation should correspond to conditional independence between x and y given S , namely, $\langle x|S|y \rangle_G \implies (x \perp S \perp y)_P$ and conversely, $(x \perp S \perp y)_P \implies \langle x|S|y \rangle_G$.

This would provide a clear graphical representation for the notion that x does not affect y directly, that its influence is mediated by the variables in S . Unfortunately, we shall next see that these two requirements might be incompatible; there might exist no way to display all the dependencies and independencies embodied in P by vertex separation in a graph.

Definition An undirected graph G is a *dependency map* (D -map) of P if there is a one-to-one correspondence between the variables in P and the nodes of G , such that for all non-intersecting subsets, x, y, S of variables we have:

$$(x \perp S \perp y)_P \implies \langle x|S|y \rangle_G. \quad (13.7)$$

Similarly, G is an *Independency map* (I -map) of P if: $(x \perp S \perp y)_P \iff \langle x|S|y \rangle_G$ (13.8)

A D -map guarantees that vertices found to be connected are indeed dependent; however, it may occasionally display dependent variables as separated vertices. An I -map works the opposite way: it guarantees that vertices found to be separated always correspond to genuinely independent variables but does not guarantee that all those shown to be connected are in fact dependent. Empty graphs are trivial D -maps, while complete graphs are trivial I -maps.

Given an arbitrary graph G , the theory of *Markov Fields* [Lauritzen 1982] tells us how to construct a probabilistic model P for which G is both a D -map and an I -map. We now ask whether the converse construction is possible.

Lemma There are probability distributions for which no graph can be both a D -map and an I -map.

Proof. Graph separation always satisfies $\langle x | S_1 | y \rangle_G \implies \langle x | S_1 \cup S_2 | y \rangle_G$ for any two subsets S_1 and S_2 of vertices. Some P 's, however, may induce both $(x \perp S_1 \perp y)_P$ and NOT $(x | S_1 \cup S_2 \perp y)_P$. Such P 's cannot have a graph representation which is both an I -map and a D -map because D -mapness forces G to display S_1 as a cutset separating x and y , while I -mapness prevents $S_1 \cup S_2$ from separating x and y . No graph can satisfy these two requirements simultaneously. Q.E.D.

An example illustrating the conditions of the proof is an experiment with two coins and a bell that rings whenever the outcomes of the two coins are the same. If we ignore the bell, the coin outcomes are mutually independent, i.e., $S_1 = \emptyset$. However, if we notice the bell (S_2), then learning the outcome of one coin should change our opinion about the other coin.

Being unable to provide a graphical description for *all* independencies, we settle for the following compromise: we will consider only I -maps but will insist that the graphs in those maps capture as many of P 's independencies as possible, i.e., they should contain no superfluous edges.

Definition A graph G is a *minimal* I -map of P if no edge of G can be deleted without destroying its I -mapness.

Theorem 13.1 Every P has a (unique) minimal I -map G_0 (called the *MARKOV-NET* of P) constructed by connecting *only* pairs (α, β) for which

$$(\alpha \perp U - \alpha - \beta \perp \beta)_P \text{ is } FALSE \quad (13.9)$$

(i.e., deleting from the complete graph *all* edges (α, β) for which $(\alpha \perp U - \alpha - \beta \perp \beta)_P$).

Definition A *Markov boundary* $B_P(\alpha)$ of variable α is a minimal subset S that renders α independent of all other variables, i.e.,

$$(\alpha \perp S \perp U - S - \alpha)_P, \quad \alpha \notin S \quad (13.10)$$

and simultaneously, no proper subset S' of S satisfies $(\alpha \perp S' \mid U - S' - \alpha)_P$. If no S satisfies (13.10), define $B_P(\alpha) = U - \alpha$.

Theorem 13.2 Each variable α has a unique Markov boundary $B_P(\alpha)$ that coincides with the set of vertices $B_{G_0}(\alpha)$ adjacent to α in the Markov net G_0 .

The usefulness of Theorem 13.2 lies in the fact that in many cases it is the Markov boundaries $B_P(\alpha)$ that define the organizational structure of human memory. People find it natural to identify the immediate consequences and/or justifications of each action or event, and these relationships constitute the neighborhood semantics for inference nets used in expert systems [Duda et al. 1976]. The fact that $B_P(\alpha)$ coincides with $B_{G_0}(\alpha)$ guarantees that many independencies can be validated by tests for graph separation at the knowledge level itself [Pearl 1985].

13.3 GRAPHOIDS

Definition A *graphoid* is a set I of triplets (x, z, y) where x, z, y are three non-intersecting subsets of elements drawn from a finite collection $U = \{\alpha, \beta, \dots\}$, having the following four properties. (We shall write $I(x, y, z)$ to state that the triplet (x, y, z) belongs to graphoid I .)

$$\text{Symmetry} \quad I(x, z, y) \iff I(y, z, x) \quad (13.11a)$$

$$\text{Subset Closure} \quad I(x, z, y \cup w) \implies I(x, z, y) \ \& \ (x, z, w) \quad (13.11b)$$

$$\text{Intersection} \quad I(x, z \cup w, y) \ \& \ I(x, z \cup y, w) \implies I(x, z, y \cup w) \quad (13.11c)$$

$$\text{Union} \quad I(x, z, y \cup w) \iff I(x, z \cup w, y) \quad (13.11d)$$

For technical convenience we shall adopt the convention that I contains all triplets in which either x or y are empty, i.e., $I(x, z, \emptyset)$.

If U stands for the set of vertices in some graph G , and if we equate $I(x, z, y)$ with the statement: “ z separates between x and y ,” written $\langle x|z|y \rangle_G$, then the conditions in (13.11) are clearly satisfied. However, not all properties of graph separation are required for graphoids. For example, in graphs we always have $[\langle \alpha|z|\beta \rangle_G \ \& \ \langle \alpha|z|\gamma \rangle_G]$ iff $\langle \alpha|z|\beta \cup \gamma \rangle_G$ while property (13.11b) requires only the “if” part. Similarly, graph separation dictates $\langle x|z|y \rangle_G \implies \langle x|z \cup w|y \rangle_G, \forall w$, while (13.11d) severely restricts the conditions under which a separating set z can be enlarged by w .

Definition A graph G is said to be an I -map of I if there is a one-to-one correspondence between the elements in U and the vertices of G , such that, for all non-intersecting subsets x, y, S we have:

$$\langle x | S | y \rangle_G \implies I(x, S, y). \quad (13.12)$$

Theorem 13.3 Every graphoid I has a unique edge-minimum I -map G_0 . $G_0 = (U, E_0)$ is constructed by connecting *only* pairs (α, β) for which the triplet $(\alpha, U - \alpha - \beta, \beta)$ is not in I , i.e.,

$$(\alpha, \beta) \notin E_0 \quad \text{iff} \quad I(\alpha, U - \alpha - \beta, \beta). \quad (13.13)$$

Definition A *relevance sphere* $R_I(\alpha)$ of an element $\alpha \in U$ is any subset S of elements for which

$$I(\alpha, S, U - S - \alpha) \text{ and } \alpha \notin S. \quad (13.14)$$

Let $R_I^*(\alpha)$ stand for the set of all relevance spheres of α . A set is called a *relevance boundary* of α , denoted $B_I(\alpha)$, if it is in $R_I^*(\alpha)$ and if, in addition, none of its proper subsets is in $R_I^*(\alpha)$.

$B_I(\alpha)$ is to be interpreted as the smallest set that “shields” α from the influence of all other elements. Note that $R_I^*(\alpha)$ is non-empty because $I(x, z, \emptyset)$ guarantees that the set $S = U - \alpha$ satisfies (13.14).

Theorem 13.4 Every element $\alpha \in U$ in a graphoid I has a unique *relevance boundary* $B_I(\alpha)$. $B_I(\alpha)$ coincides with the set of vertices $B_{G_0}(\alpha)$ adjacent to α in the minimal graph G_0 .

Corollary 13.1 The set of relevance boundaries $B_I(\alpha)$ forms a *neighbor system*, i.e., a collection $B_I^* = \{B_I(\alpha) : \alpha \in U\}$ of subsets of U such that (i) $\alpha \notin B_I(\alpha)$, and (ii) $\alpha \in B_I(\beta)$ iff $\beta \in B_I(\alpha)$, $\alpha, \beta \in U$.

Corollary 13.2 The edge-minimum I -map G_0 can be constructed by connecting each α to all members of its relevance boundary $B_I(\alpha)$.

Thus we see that the major graphical properties of probabilistic dependencies are consequences of the intersection and union properties, (13.11c) and (13.11d), and will therefore be shared by all graphoids.

13.4 Special Graphoids and Open Problems

13.4.1 Graph-induced Graphoids

The most restricted type of graphoid is that which is isomorphic to some underlying graph, i.e., *all* triplets (x, z, y) in I reflect vertex-separation conditions in an actual graph.

Definition A graphoid I is said to be *graph-induced* if there exists a graph G such that

$$I(x, z, y) \iff \langle x | z | y \rangle_G. \quad (13.15)$$

Theorem 13.5 A necessary and sufficient condition for a graphoid I to be graph induced is that it satisfies the following five independent axioms:

$$I(x, z, y) \iff I(y, z, x) \quad (\text{symmetry}) \quad (13.16a)$$

$$I(x, z, y \cup w) \implies I(x, z, y) \ \& \ I(x, z, w) \quad (\text{subset closure}) \quad (13.16b)$$

$$I(x, z \cup w, y) \ \& \ I(x, z \cup y, w) \implies I(x, z, y \cup w) \quad (\text{intersection}) \quad (13.16c)$$

$$I(x, z, y) \implies I(x, z \cup w, z) \quad \forall w \subset U \quad (\text{strong union}) \quad (13.16d)$$

$$I(x, z, y) \implies I(x, z, \gamma) \ \text{or} \ I(\gamma, z, y) \quad \forall \gamma \notin x \cup z \cup y \quad (\text{transitivity}) \quad (13.16e)$$

Remarks (13.16c) and (13.16d) imply the converse of (13.16b). The union axiom (13.16d) is unconditional and therefore stronger than the one required for general graphoids (13.11d). It allows us to construct G_0 by simply deleting from a complete graph every edge (α, β) for which a triplet of the form (α, S, β) appears in I .

13.4.2 Probabilistic Graphoids

Definition A graphoid is called *probabilistic* if there exists a probability distribution P on the variables in U such that $I(x, z, y)$ iff x is independent of y given z , i.e.,

$$I(x, z, y) \iff (x \perp z \perp y)_P. \quad (13.17)$$

In other words, probabilistic graphoids capture the notion of conditional independence in Probability Theory (see Section 13.2).

Theorem 13.6 Every graph-induced graphoid is probabilistic.

Since every probabilistic-independence relation satisfies (13.6a)-(13.6e), a necessary condition for a graphoid to be probabilistic is that, in addition to (13.11), it also satisfies the contraction property (13.6e), i.e.,

$$I(x, y \cup z, w) \ \& \ I(x, y, z) \implies I(x, y, z \cup w). \quad (13.18)$$

(13.18) can be interpreted to state that if we judge w to be irrelevant (to x) after learning some irrelevant facts z , then w must have been irrelevant before learning z . Together with the union property (13.11d) it means that learning irrelevant facts should not alter the relevance status of other propositions in the system; whatever was relevant remains relevant and what was irrelevant remains irrelevant.

Conjecture The contraction property (13.18) is sufficient for a graphoid to be probabilistic.

Unlike the sufficiency condition for graph-induced graphoids, we found no way of constructing a distribution P that yields $I(x, z, y) \implies (x \perp z \perp y)_P$ for every I that satisfies (13.18).

13.4.3 Correlational Graphoids

Let U consist of n random variables u_1, u_2, \dots, u_n , and let z be a subset of U such that $|z| \leq n - 2$. The *partial correlation coefficient* of u_i and u_j with respect to z , denoted $\rho_{ij \cdot z}$, measures the correlation between u_i and u_j after subtracting from them the best linear estimates using the variables in z (Cramér 1946). In other words, $\rho_{ij \cdot z}$ measures the correlation that remains after removal of any part of the variation due to the influence of the variables in z .

Definition Let x, y, z be three nonintersecting subsets of U . A relation $I_c(x, y, z)$ is said to be *correlation-based* if for every $u_i \in x$ and $u_j \in y$ we have:

$$I_c(x, z, y) \iff \rho_{ij \cdot z} = 0. \quad (13.19)$$

In other words, x is considered irrelevant to y relative to z if every variable in x is uncorrelated with every variable in y , after removing the (linear) influence of the variables in z .

Theorem 13.7 Every correlation-based relation is a graphoid which, in addition to axioms (13.11), also satisfies the contraction property (13.18) and the converse of (13.11b), i.e.,

$$I(x, z, y) \text{ and } I(x, z, w) \implies I(x, z, y \cup w). \quad (13.20)$$

Conjecture Every graphoid satisfying (13.18) and (13.20) is isomorphic to some correlation-based relation.

13.5 Conclusions

We have shown that the essential qualities characterizing the probabilistic notion of conditional independence are captured by two logical axioms: weak closure for intersection (13.6c), and weak closure for union (13.6d). These two axioms enable us to construct an edge-minimum graph in which every cutset corresponds to a genuine independence condition, and these two axioms were chosen therefore as the logical basis for graphoid systems — a more general, nonprobabilistic formalism of relevance. Vertex separation in graphs, probabilistic independence and partial uncorrelatedness are special cases of graphoid systems where the two defining axioms are augmented with additional requirements.

The graphical properties associated with graphoid systems offer an effective inference mechanism for deducing, in any given state of knowledge, which propositional variables are relevant to each other. If we identify the relevance boundaries associated with each proposition in the system, and treat them as neighborhood relations defining a graph G_0 , then we can correctly deduce irrelevance relationships by testing whether the set of currently known propositions constitutes a cutset in G_0 .

References

- Cramér, H. (1946), *Mathematical Methods of Statistics*, Princeton University Press, Princeton, N.J.
- Duda, R. O., Hart, P. E., and Nilsson, N. J., (1976), Subjective Bayesian Methods for Rule-Based Inference Systems, *Proceedings, 1976 NCC (AFIPS)*, 45, 1075-1082.
- Lauritzen, S.L. (1982), *Lectures on Contingency Tables*, 2nd Ed., U. of Aalborg Press, Aalborg, Denmark.
- Pearl, J. (1985), Fusion, Propagation and Structuring in Bayesian Networks, Technical Report CSD-850022, June 1985.
- Pearl, J. & Paz, A., (1985), GRAPHOIDS: a Graph-Based Logic for Reasoning about Relevance Relations, *Technical Report CSD-850038*, December 1985.

System Z: A Natural Ordering of Defaults with Tractable Applications to Nonmonotonic Reasoning

Judea Pearl

Abstract

Recent progress towards unifying the probabilistic and model preference semantics for nonmonotonic reasoning has led to a remarkable observation: Any consistent system of default rules imposes an unambiguous and natural ordering on these rules which, to emphasize its simple and basic character, we term “Z-ordering.” This ordering can be used with various levels of refinement, to prioritize conflicting arguments, to rank the degree of abnormality of states of the world, and to define plausible consequence relationships. This paper defines the Z-ordering, briefly mentions its semantical origins, and illustrates two simple entailment relationships induced by the ordering. Two extensions are then described, maximum-entropy and conditional entailment, which trade in computational simplicity for semantic refinements.

14.1 Description

We begin with a set of rules $R = \{r: \alpha_r \rightarrow \beta_r\}$ where α_r and β_r are propositional formulas over a finite alphabet of literals, and \rightarrow denotes a new connective to be

This work was supported in part by National Science Foundation grant #IRI-86-10155 and Naval Research Laboratory grant #N00014-89-J-2007

Originally published in Theoretical Aspects of Reasoning About Knowledge, (TARK-III), R. Parikh (ed.), Morgan Kaufmann, 1990, pp. 121–135.

Original: https://ftp.cs.ucla.edu/pub/stat_ser/R131.pdf 7.9.2021

Republished with permission.

given default interpretations later on. A truth valuation of the literals in the language will be called a *model*. A model M is said to *verify* a rule $\alpha \rightarrow \beta$ if $M \models \alpha \wedge \beta$ (i.e., α and β are both true in M), and to *falsify* $\alpha \rightarrow \beta$ if $M \models \alpha \wedge \neg \beta$.

Given a set R of such rules, we first define the relation of *toleration*.

Definition 14.1 A set of rules $R' \subseteq R$ is said to *tolerate* an individual rule r , denoted $T(r | R')$, if the set of formulas $(\alpha_r \wedge \beta_r) \cup_{r' \in R'} (\alpha_{r'} \supset \beta_{r'})$ is satisfiable, i.e., if there exists a model that verifies r and does not falsify any of the rules in R' .

To facilitate the construction of the desired ordering, we now define the notion of *consistency*.

Definition 14.2 A set R of rules is said to be *consistent* if for every non-empty subset $R' \subseteq R$ there is at least one rule that is tolerated by all the others, i.e.,

$$\forall R' \subseteq R, \exists r' \in R', \text{ such that } T(r' | R' - r') \quad (14.1)$$

This definition, named *p-consistent* in [Adams 1975] and ϵ -consistent in [Pearl 1988], assures the existence of an *admissible* probability assignment when rules are given a probabilistic interpretation. In other words, if each rule $\alpha \rightarrow \beta$ is interpreted as a statement of high conditional probability, $P(\beta | \alpha) \geq 1 - \epsilon$, consistency assures that for every $\epsilon > 0$ there will be a probability assignment P (to models of the language) that satisfies all these statements simultaneously. An identical criterion of consistency also assures the existence of an *admissible* preference ranking on models, when each rule $\alpha \rightarrow \beta$ is given a model-preference interpretation, namely, β is true in all the most preferred models of α [Lehmann and Magidor 1988].

A slightly more elaborate definition of consistency applies to databases containing mixtures of defeasible and nondefeasible rules [Goldszmidt and Pearl 1989a]. Note that the condition of consistency is stronger than that of mere satisfiability. For example, the two rules $a \rightarrow b$ and $a \rightarrow \neg b$ are satisfiable (if a is false) but not consistent. Intuitively, consistency requires that in addition to satisfying the constraint associated with the rule $a \rightarrow b$, the truth of a should not be ruled out as an impossibility. This reflects the common understanding that a conditional sentence “if a then b ” is not fully satisfied by merely making a false; it requires that both a and b be true in at least one possible world, however unlikely.

The condition of consistency, Equation (14.1), leads to a natural ordering of the rules in R . Given a consistent R , we first identify every rule that is tolerated by all the other rules of R , assign to each such rule the label 0, and remove it from R . Next,

we attach a label 1 to every rule that is tolerated by all the remaining ones, and so on. Continuing in this way, we form an ordered partition of $R = (R_0, R_1, R_2, \dots, R_K)$, where

$$R_i = \{r: T(r | R - R_0 - R_1 - \dots - R_{i-1})\} \quad (14.2)$$

The label attached to each rule in the partition defines the Z -ranking or Z -ordering. The process of constructing this partition also amounts to testing the consistency of R , because it terminates with a full partition iff R is consistent [Goldszmidt and Pearl 1989a].

Theorem 14.1 The complexity of testing the consistency of a set of rules is $O[PS(n)N^2]$, where N is the number of rules, n the number of literals in R and $PS(n)$ the complexity of propositional satisfiability in the sublanguage characterizing the rules (e.g., $PS(n) = O(n)$ for Horn expressions).

Proof. Identifying R_0 takes $N \cdot PS(n)$ steps, identifying R_1 takes $(N - |R_0|)PS(n)$ steps, and so on. Thus, the total time it takes to complete the labeling is

$$\begin{aligned} PS(n)[N + (N - |R_0|) + (N - |R_0| - |R_1|) + \dots] &\leq PS(n)[N + (N - 1) + \dots] \\ &= PS(n) \frac{N^2}{2} \end{aligned} \quad (14.3)$$

In order to define the notions of entailment and consequence it is useful to translate the ranking among rules into preferences among models. The reason is that we wish to proclaim a formula g to be a plausible consequence of f , written $f \vdash g$, only if the constraints imposed by R would force the models of $f \wedge g$ to stand in some preference relation over those of $f \wedge \neg g$. For example, the traditional preferential criterion for g to be a rational consequence of f requires that all the most preferred models of f satisfy g , i.e., that all the most preferred models of f reside in $f \wedge g$ and none resides in $f \wedge \neg g$ [Shoham 1987]. We shall initially limit ourselves to such preference criteria that do not require substantial enumeration of models, i.e., that the preference between $f \wedge g$ and $f \wedge \neg g$ be readily tested using the partition defined in Equation (14.2). To that purpose, we propose the following ranking on models. Using $Z(r)$ to denote the label assigned to rule r ,

$$Z(r) = i \quad \text{iff} \quad r \in R_i, \quad (14.4)$$

we define the rank associated with a particular model M as the lowest integer n such that all rules having $Z(r) \geq n$ are satisfied by M ,

$$Z(M) = \min\{n: M \models (\alpha_r \supset \beta_r) \quad Z(r) \geq n\} \quad (14.5)$$

In other words, the rank of a model is equal to 1 plus the rank of the highest-ranked rule falsified by the model. The rank associated with a given formula f is now defined as the lowest Z of all models satisfying f ,

$$Z(f) = \min\{Z(M) : M \models f\} \quad (14.6)$$

Note that, once we establish the ranking of the rules, the complexity of determining the Z value of any given M is $O(N)$; we simply identify the highest Z rule that is falsified by M and add 1 to its Z. More significantly, determining the Z value of an arbitrary formula f requires at most N satisfiability tests; we search for the lowest i such that all rules having $Z(r) \geq i$ tolerate $f \rightarrow \text{true}$, i.e.,

$$Z(f) = \min\{i : T(f \rightarrow \text{true} \mid R_i, R_{i+1}, \dots)\} \quad (14.7)$$

Equation (14.5) defines a total order on models, with those receiving a lower Z interpreted as being more normal or more preferred. This ordering satisfies the constraints that for each rule $\alpha_r \rightarrow \beta_r$, β_r holds true in all the most-preferred models of α_r , namely, the usual preferential model interpretation of default rules. It can be shown (see Appendix 14.1) that the rankings defined by Equations (14.4) and (14.5) correspond to a special kind of a preferential structure; out of all rankings satisfying the rule constraints, the assignment defined in Equation (14.5) is the only one that is *minimal*, in the sense of assigning to each model the lowest possible ranking (or highest normality) permitted by the rules in R .

14.2 Consequence Relations

We are now ready to define two notions of nonmonotonic entailment. Given a knowledge base in the form of a consistent set R of rules, and some factual information f , we wish to define the conditions under which f can be said to entail a conclusion g , in the context of R .

Definition 14.3 0-entailment

g is said to be *0-entailed* by f in the context R , written $f \vdash_0 g$, if the augmented set of rules $R \cup f \rightarrow \neg g$ is inconsistent.

Theorem 14.2 0-entailment is semi-monotonic, i.e., if $R' \subseteq R$ then

$f \vdash_0 g$ under R whenever $f \vdash_0 g$ under R' .

The proof is immediate, from the fact that if $R' \cup f \rightarrow \neg g$ is inconsistent, then $R \cup f \rightarrow \neg g$ must be inconsistent as well. Semi-monotonicity reflects a strategy of extreme caution; no consequence will ever be issued if it is possible to add rules to R (consistently) in such a way as to render the conclusion no longer valid. Thus, 0-entailment generates the maximal set of “safe” conclusions that can be drawn

from R , and hence, was proposed in [Pearl 1989] as a *conservative core* that ought to be common to all non-monotonic formalisms.

0-entailment was named p -entailment by Adams [1975], ε -entailment by Pearl [1988] and r -entailment by Lehmann and Magidor [1988]. Probabilistically, 0-entailment guarantees that conclusions will receive arbitrarily high probabilities (i.e., $P(g|f) \rightarrow 1$) whenever the premises receive arbitrarily high probabilities (i.e., $P(\beta_r | \alpha_r) \rightarrow 1 \forall r \in R$). In the preferential model interpretation, 0-entailment guarantees that $\kappa(f \wedge g) < \kappa(f \wedge \neg g)$ holds in *all* admissible ranking functions κ , namely, in all ranking functions $\kappa(M)$ that satisfy the rule constraints

$$\kappa(\alpha_r \wedge \beta_r) < \kappa(\alpha_r \wedge \neg \beta_r) \forall r \in R \quad (14.8)$$

where, for every formula α ,

$$\kappa(\alpha) = \min\{\kappa(M) : M \models \alpha\}. \quad (14.9)$$

Due to its extremely conservative nature, 0-entailment does not properly handle irrelevant features, e.g., from $a \rightarrow c$ we cannot conclude $a \wedge b \rightarrow c$ even in cases where R makes no mention of b . To sanction such inferences we now define a more adventurous type of entailment.

Definition 14.4 1-entailment

A formula g is said to *1-entailed* by f , in the context R , (written $f \vdash_1 g$), if

$$Z(f \wedge g) < Z(f \wedge \neg g). \quad (14.10)$$

Namely, there exists an integer k such that the set of rules ranked higher or equal to k tolerates $f \rightarrow g$ but does not tolerate $f \rightarrow \neg g$. Note that, once we have the Z-rank of all rules, deciding 1-entailment for a given query requires at most $2(1 + \log |R|)$ satisfiability tests (using a binary-search strategy). 1-entailment can be given a clear motivation in preferential model semantics. Instead of insisting that $\kappa(f \wedge g) < \kappa(f \wedge \neg g)$ hold in *all* admissible ranking functions κ , as was done in 0-entailment, we only require that it holds in the unique admissible ranking that is minimal, namely, the Z-ranking (see Appendix 14.1).

Lehmann [1989] has extended 0-entailment in a slightly different way, introducing a consequence relation called *rational closure*. Rational closure is defined in terms of a relation called *more exceptional*, where a formula α is said to be more exceptional than β if

$$\alpha \vee \beta \vdash_0 \neg \alpha.$$

Based on this relation, Lehmann then used an inductive definition to assign a *degree* to each formula α in the language: $degree(\alpha) = i$ if $degree(\alpha)$ is not less than

i and every β that is less exceptional than α has $\text{degree}(\beta) < i$. Finally, a sentence $\alpha \rightarrow \beta$ was defined to be in the rational closure of R iff $\text{degree}(\alpha) < \text{degree}(\alpha \wedge \neg \beta)$.

Goldszmidt and Pearl [1989b] have recently shown that $\text{degree}(\alpha)$ is identical to $Z(\alpha)$ and, hence, rational closure is equivalent to 1-entailment. This endows the Z-ranking with an additional motivation in terms of exceptionality; $Z(\alpha) > Z(\beta)$ if α is more exceptional than β . Additionally, the computational procedure developed for 1-entailment renders membership in the rational closure decidable in at most $2(1 + \log |R|)$ satisfiability tests.

Lehmann [1989] has also shown that the rational closure can be obtained by syntactically closing the relation of 0-entailment under a rule suggested by Makinson called *rational monotony*. Rational monotony permits us to conclude $a \wedge b \vdash c$ from $a \vdash c$ as long as the consequence relation does not contain $a \vdash \neg b$. Rational monotony is induced by any admissible ranking function, not necessarily the minimal one defined by system-Z (see Appendix 14.II). Thus, 1-entailment can be thought of as an extension of 0-entailment to acquire properties that are sound in any individual (admissible) ranking function.

1-entailment, though more adventurous than 0-entailment, still does not go far enough, as is illustrated in the next section.

14.3 Illustrations

Consider the following collection of rules R :

r_1 : "Penguins are birds"	$p \rightarrow b$
r_2 : "Birds fly"	$b \rightarrow f$
r_3 : "Penguins do not fly"	$p \rightarrow \neg f$
r_4 : "Penguins live in the antarctic"	$p \rightarrow a$
r_5 : "Birds have wings"	$b \rightarrow w$
r_6 : "Animals that fly are mobile"	$f \rightarrow m$

It can be readily verified that r_6 , r_5 , and r_2 are each tolerated by all the other five rules in R . For example, the truth assignment ($p = 0, a = 0, f = 1, b = 1, w = 1, m = 1$) satisfies both

$$b \wedge w \wedge (p \supset b) \wedge (b \supset f) \wedge (p \supset \neg f) \wedge (p \supset a) \wedge (f \supset m)$$

and

$$b \wedge f \wedge (p \supset b) \wedge (b \supset w) \wedge (b \supset \neg f) \wedge (p \supset a) \wedge (f \supset m).$$

Thus, r_6 , r_5 and r_2 are each assigned a label 0 indicating that these rules pertain to the most normal state of affairs. No other rule can be labeled 0 because, once we

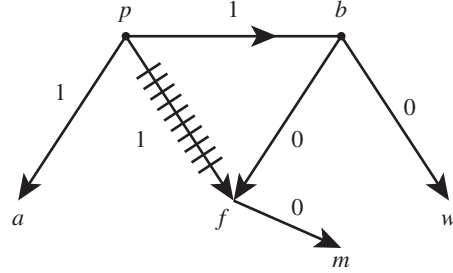


Figure 14.1 Collection of six rules for the example discussed in the text.

assign p the truth value 1, we must assign 1 to b and 0 to f , which is inconsistent with $b \supset f$. The remaining three rules can now be labeled 1, because each of the three is tolerated by the other two. A network describing the six rules and their Z-labels is shown in Figure 14.1.

The following are examples of plausible consequences one would expect to draw from R :

<i>0-entailed</i>	1-entailed	not-entailed
$b \wedge p \vdash \neg f$	$\neg b \vdash \neg p$	$p \vdash w$
$f \vdash \neg p$	$\neg f \vdash \neg b$	$p \wedge \neg a \vdash \neg f$
$b \vdash \neg p$	$b \vdash m$	$p \wedge \neg a \vdash w$
$p \wedge a \vdash b$	$\neg m \vdash \neg b$	
	$p \wedge \neg w \vdash b$	

For example, to test the validity of $b \wedge p \vdash_0 \neg f$ we add the rule $r_6: b \wedge p \rightarrow f$ to R , and realize that the augmented set becomes inconsistent; no rule in the set $\{b \wedge p \rightarrow f, p \rightarrow b, p \rightarrow \neg f\}$ can be tolerated by the other two.

1-entailment sanctions plausible inference patterns that are not 0-entailed, among them rule chaining, contraposition and the discounting of irrelevant features. For example, we cannot conclude by 0-entailment that birds are mobile, $b \vdash m$, because neither $b \rightarrow m$ nor $b \rightarrow \neg m$ would render R inconsistent. However, m is 1-entailed by b , because the rule $b \rightarrow m$ is tolerated by all rules in R while $b \rightarrow \neg m$ is tolerated by only those labeled 1. Thus,

$$Z(b \wedge m) < Z(b \wedge \neg m),$$

confirming Equation (14.10). Similarly, if c is an irrelevant feature (i.e., not appearing in R), we obtain $b \wedge c \vdash_1 f$ but not $b \wedge c \vdash_0 f$.

On the other hand, 1-entailment does not permit us to conclude that flying objects are birds ($f \vdash b$) or that penguins who do not live in the antarctic are still birds ($p \wedge \neg a \vdash b$). This is because negating these consequences will not change their Z-ratings — in testing $f \vdash_1 b$ we have $Z(f \wedge b) = Z(f \wedge \neg b) = 0$, while in testing $p \wedge \neg a \vdash_1 b$ we have $Z(p \wedge \neg a \wedge b) = Z(p \wedge \neg a \wedge \neg b) = 2$.

There are cases, however, where 1-entailment produces conclusions whose plausibility may be subject to dispute. For example,¹ if we add to Figure 14.1 the rule $c \rightarrow f$ we obtain $Z(c \rightarrow f) = 0$, which yields $c \vdash_1 \neg p$ and $c \wedge p \vdash_1 \neg f$. In other words, 1-entailment ranks the new class c to be as normal as birds, and penguins, by virtue of being exceptional kind of birds (relative to flying) are also treated as exceptional c 's. Were the database to contain no information relative to birds, penguins and c 's would be treated as equal status classes and the conclusion $p \wedge c \vdash \neg f$ would not be inferred. Thus, merely mentioning a property (f) by which a class (p) differs from its superclass (b) automatically brands that class (p) exceptional relative to any neutral class (c).

The main weakness of the system described so far is its inability to sanction property inheritance from classes to exceptional sub-classes. For example, neither of the two types of entailments can sanction the conclusion that penguins have wings ($p \rightarrow w$) by virtue of being birds (albeit exceptional birds). The reason is that the label 1 assigned to all rules emanating from p amounts to proclaiming penguins an exceptional type of birds in *all* respects, barred from inheriting *any* bird-like properties (e.g., laying eggs, having beaks, etc.). This is a drawback that cannot be remedied by methods based solely on the Z-ordering of defaults. The fact that $p \rightarrow w$ is tolerated by two extra rules ($p \rightarrow b$, and $b \rightarrow w$) on top of those tolerating $p \rightarrow \neg w$, remains undetected.

To sanction property inheritance, a more refined ordering is required which also takes into account the *number* of rules tolerating a formula, not merely their rank orders. One such refinement is provided by the maximum-entropy approach [Goldszmidt and Pearl 1989c] where each model is ranked by the sum of weights on the rules falsified by that model. Another refinement is provided by Geffner's conditional entailment [Geffner 1989], where the priority of rules induces a *partial* order on models. These two refinements will be summarized next.

14.4 The Maximum Entropy Approach

The maximum-entropy (ME) approach [Pearl 1988] is motivated by the convention that, unless mentioned explicitly, properties are presumed to be independent of one another; such presumptions are normally embedded in probability

1. This observation is due to Hector Geffner.

distributions that attain the maximum entropy subject to a set of constraints. Given a set R of rules and a family of probability distributions that are admissible relative to the constraints conveyed by R (i.e., $P(\beta_r \rightarrow \alpha_r) \geq 1 - \varepsilon \forall r \in R$), we can single out a distinguished distribution $P_{\varepsilon,R}^*$ having the greatest entropy $-\sum_M P(M) \log P(M)$, and define entailment relative to this distribution by

$$f \vdash_{ME} g \quad \text{iff} \quad P_{\varepsilon,R}^*(g | f) \xrightarrow{\varepsilon \rightarrow 0} 1. \quad (14.11)$$

An infinitesimal analysis of the ME approach also yields a ranking function κ on models, where $\kappa(M)$ now corresponds to the lowest exponent of ε in the expansion of $P_{\varepsilon,R}^*(M)$ into a power series in ε . Moreover, this ranking function can be encoded parsimoniously by assigning an integer weight w_r to each rule $r \in R$ and letting $\kappa(M)$ be the sum of the weights associated with the rules falsified by M . The weight w_r , in turn, reflects the “cost” we must add to each model M that falsifies rule r , so that the resulting ranking function would satisfy the constraint conveyed by R , namely,

$$\min\{\kappa(M): M \models \alpha_r \wedge \beta_r\} < \min\{\kappa(M): M \models \alpha_r \wedge \neg \beta_r\}, r \in R$$

These considerations lead to a set of $|R|$ non-linear equations for the weights w_r which, under certain conditions, can be solved by iterative methods. Once the rule weights are established, ME-entailment is determined by the criterion of Equation (14.11), translated to

$$f \vdash_{ME} g \quad \text{iff} \quad \min\{\kappa(M): M \models f \wedge g\} < \min\{\kappa(M): M \models f \wedge \neg g\}.$$

where

$$\kappa(M) = \sum_{r: M \models \alpha_r \wedge \neg \beta_r} w_r$$

We see that ME-entailment requires minimization over models, a task that may take exponential time. In practice, however, this minimization is accomplished quite effectively in databases of Horn expressions, yielding a reasonable set of inference patterns. For example, in the database of Figure 14.1, ME-entailment will sanction the desired consequences $p \vdash w$, $p \wedge \neg a \vdash \neg f$ and $p \wedge \neg a \vdash w$ and, moreover, it will avoid the undesirable pattern of concluding $c \wedge p \vdash \neg f$ from $R \cup \{c \rightarrow f\}$.

The weaknesses of the ME approach are two-fold. First, it does not properly handle causal relationships and, second, it is sensitive to the format in which the rules are expressed. This latter sensitivity is illustrated in the following example. From $R = \{\text{Swedes are blond, Swedes are well-mannered}\}$, ME will conclude that dark-haired Swedes are still well-mannered, while no such conclusion will be drawn

from $R = \{\text{Swedes are blond and well-mannered}\}$. This sensitivity might sometimes be useful for distinguishing fine nuances in natural discourse, concluding, for example, that mannerisms and hair color are two independent qualities. However, it stands at variance with one of the basic conventions of formal logic, which treats $a \rightarrow b \wedge c$ as a shorthand notation of $a \rightarrow b$ and $a \rightarrow c$ and, moreover, unlike 1-entailment it will conclude $c \wedge p \vdash_{ME} \neg f$ from $\Delta \cup \{c \rightarrow f\}$, where c is an irrelevant property.

The failure to respond to causal information (see Pearl [1988, pp. 463, 519] and Hunter [1989]) prevents the ME approach from properly handling tasks such as the Yale shooting problem [Hanks and McDermott 1986], where rules of causal character are given priority over other rules. This weakness may perhaps be overcome by introducing causal operators into the ME formulation, similar to the way causal operators are incorporated within other formalisms of nonmonotonic reasoning (e.g., Shoham [1986], Geffner [1989]).

14.5 Conditional Entailment

Geffner [1989] has overcome the weaknesses of 1-entailment by introducing two new refinements. First, rather than letting rule priorities dictate a ranking function on models, a partial order on models is induced instead. To determine the preference between two models, M and M' , we examine the highest priority rules that distinguish between the two, i.e., that are falsified by one and not by the other. If all such rules remain unfalsified in one of the two models, then this model is the preferred one. Formally, if $\Delta[M]$ and $\Delta[M']$ stand for the set of rules falsified by M and M' , respectively, then M is preferred to M' (written $M < M'$) iff $\Delta[M] \neq \Delta[M']$ and for every rule r in $\Delta[M] - \Delta[M']$ there exists a rule r' in $\Delta[M'] - \Delta[M]$ such that r' has a higher priority than r (written $r < r'$). Using this criterion, a model M will always be preferred to M' if it falsifies a proper subset of the rules falsified by M' . Lacking this feature in the Z-ordering has prevented 1-entailment from concluding $p \vdash w$ in the example of Section 14.3.

The second refinement introduced by Geffner is allowing the rule-priority relation, $<$, to become a partial order as well. This partial order is determined by the following interpretation of the rule $\alpha \rightarrow \beta$; if α is all that we know, then, regardless of other rules that R may contain, we are authorized to assert β . This means that $r: \alpha \rightarrow \beta$ should get a higher priority than any argument (a chain of rules) leading from α to $\neg \beta$ and, more generally, if a set of rules $R' \subset R$ does not tolerate r , then at least one rule in R' ought to have a lower priority than r . In Figure 14.1, for example, the rule $r_3: p \rightarrow \neg f$ is not tolerated by the set $\{r_1: p \rightarrow b, r_2: b \rightarrow f\}$, hence, we must have $r_1 < r_3$ or $r_2 < r_3$. Similarly, the rule $r_1: p \rightarrow b$ is not tolerated by $\{r_2, r_3\}$, hence, we also have $r_2 < r_1$ or $r_3 < r_1$. From the asymmetry and transitivity of $<$, these two

conditions yield $r_2 \prec r_3$ and $r_2 \prec r_1$. It is clear, then, that this priority on rules will induce the preference $M < M'$, whenever M validates $p \wedge b \wedge \neg f$ and M' validates $p \wedge b \wedge f$; the former falsifies r_2 , while the latter falsifies the higher priority rule r_3 . In general, we say that a proposition g is conditionally entailed by f (in the context of R) if g holds in all the preferred models of f induced by every priority ordering admissible with R .

Conditional entailment rectifies many of the shortcomings of 1-entailment as well as some weaknesses of ME-entailment. However, having been based on model minimization as well as on enumeration of subsets of rules, its computational complexity might be overbearing. A proof theory for conditional entailment can be found in Geffner [1989].

14.6 Conclusions

The central theme in this paper has been the realization that underlying any consistent system of default rules there is a natural ranking of these defaults and that this ranking can be used to induce preferences on models and plausible consequence relationships. We have seen that the Z-ranking emerges from both the probabilistic interpretation of defaults and their preferential model interpretation, and that two of its immediate entailment relations are decidable in $O(N^2)$ satisfiability tests. The major weakness of these entailment relationships has been the blockage of property inheritance across exceptional subclasses. Two refinements were described, maximum-entropy and conditional entailment, which properly overcome this weakness at the cost of a higher complexity. An open problem remains whether there exists a tractable approximation to the maximum entropy or the conditional entailment schemes which permits inheritance across exceptional subclasses and, at the same time, retains a proper handling of specificity-based priority.

Acknowledgments

I am indebted to Daniel Lehmann for sharing his thoughts on the relations between r -entailment, ε -entailment, rational closure, and maximum-entropy. Hector Geffner and Moises Goldszmidt have contributed many ideas, and are responsible for the developments described in Sections 14.4 and 14.5.

14.I

Definition

Appendix I: Uniqueness of The Minimal Ranking Function

A *ranking function* is an assignment of non-negative integers to the models of the language. A ranking function κ is said to be *admissible* relative to database R , if it satisfies

$$\min\{\kappa(M): M \models \alpha_r \wedge \beta_r\} < \min\{\kappa(M): M \models \alpha_r \wedge \neg \beta_r\} \quad (14I-1)$$

for every rule $r: \alpha_r \rightarrow \beta_r$ in R .

Let W stand for the set of models considered.

Definition A ranking function κ is said to be *minimal* if every other admissible ranking κ' satisfies $\kappa'(M) > \kappa(M')$ for at least one model $M' \in W$.

Clearly, every minimal ranking has the property of “local compactness,” namely, it is not possible to lower the rank of one model while keeping the ranks of all other models constant. Every such attempt will result in violating the constraint imposed by at least one rule in R . We will now show that local compactness is also a sufficient property for minimality, because there is in fact only one unique ranking that is locally compact.

Definition An admissible ranking function κ is said to be *compact* if, for every $M' \in W$, any ranking κ' satisfying

$$\begin{aligned}\kappa'(M) &= \kappa(M)M \neq M' \\ \kappa'(M) &< \kappa(M)M = M'\end{aligned}$$

is inadmissible.

Theorem (uniqueness):
Every consistent R has a unique compact ranking $Z(M)$ given by Equation (14.5).

Corollary Every consistent R has a unique minimal ranking given by the compact ranking $Z(M)$ of Equation (14.5).

Proof. We will prove that the ranking function Z given in Equation (14.5) is the unique compact ranking. First we show, by contradiction, that Z is indeed compact. Suppose it is possible to lower the rank $Z(M')$ of some model M' . Let $Z(M') = I$. From Equation (14.5) we know that M' falsifies some rule $r: \alpha \rightarrow \beta$ of rank $Z(r) = I - 1$, namely, $M' \models \alpha \wedge \neg \beta$, and there exists $\widehat{M} \models \alpha \wedge \beta$ having $Z(\widehat{M}) = I - 1$. Lowering the rank of M' below I , while keeping $Z(\widehat{M}) = I - 1$ would clearly violate the constraint imposed by the rule $\alpha \rightarrow \beta$ (see Equation (14I-1)). Thus, Z is compact.

We now prove that Z is unique. Suppose there exists some other compact ranking function κ that differs from Z on at least one model. We shall show that if there exists an M' such that $\kappa(M') < Z(M')$ then κ could not be admissible. while if there exists an M' such that $\kappa(M') > Z(M')$, then κ could not be compact. Assume $\kappa(M') < Z(M')$, let I be the lowest κ value for which such inequality holds, and let $Z(M') = J > I$. From Equation (14.5), M' falsifies some rule $\alpha \rightarrow \beta$ of rank $J - 1$, namely, $M' \models \alpha \wedge \neg \beta$ and every model M validating $\alpha \wedge \beta$ must obtain $Z(M) \geq J - 1$. By our assumption, $\kappa(M)$ must also assign to each such M a value not lower than

$J - 1 \geq I$. But this is incompatible with the constraint $\alpha \rightarrow \beta$ (see Equation (14I-1)). Thus, κ is inadmissible.

Now assume there is a non-empty set of models for which $\kappa(M) > Z(M)$, and let I be the lowest Z value in which $\kappa(M') > Z(M')$ holds for some model M' . We will show that κ could not be compact, because it should be possible to reduce $\kappa(M')$ to $Z(M')$ while keeping constant the κ of all other models. From $Z(M') = I$ we know that M' does not falsify any rule $\alpha' \rightarrow \beta'$ whose Z rank is higher than $I - 1$. Hence, we only need to watch whether the reduction of κ can violate rules r for which $Z(r) < I$. However, every such rule $r: \alpha \rightarrow \beta$ has a model $M \models \alpha \wedge \beta$ having $Z(M) < I$, and every such model was assumed to obtain a κ rank equal to that assigned by Z . Hence, none of these rules will be violated by lowering $\kappa(M')$ to $Z(M)$, QED.

14.II

Appendix II: Rational Monotony of Admissible Rankings

Theorem The consequence relation \vdash defined by the criterion

$$f \vdash g \quad \text{iff} \quad \kappa(f \wedge g) < \kappa(f \wedge \neg g)$$

is closed under rational monotony, for every admissible ranking function κ .

Proof. We need to show that for every three formulas a, b and c , if $a \vdash c$, then either $a \vdash \neg b$ or $a \wedge b \vdash c$. Assume $a \vdash c$ and $a \not\vdash \neg b$, namely,

- (i) $\kappa(a \wedge c) < \kappa(a \wedge \neg c)$
- (ii) $\kappa(a \wedge \neg b) \geq \kappa(a \wedge b)$,

we must prove

$$(iii) \quad \kappa(a \wedge b \wedge c) < \kappa(a \wedge b \wedge \neg c).$$

Rewriting (i) as

$$\kappa(a \wedge c) = \min\{\kappa(a \wedge c \wedge b), \kappa(a \wedge c \wedge \neg b)\} < \min\{\kappa(a \wedge b \neg c), \kappa(a \wedge \neg b \wedge \neg c)\} = \kappa(a \wedge \neg c)$$

we need to show only that the min on the left hand side is obtained at the second term, i.e., that

$$\min\{\kappa(a \wedge c \wedge b), \kappa(a \wedge c \wedge \neg b)\} = \kappa(a \wedge c \wedge \neg b).$$

But this is guaranteed by (ii), because the alternative possibility:

$$\kappa(a \wedge c \wedge b) < \kappa(a \wedge c \wedge \neg b)$$

together with (ii), would violate (i). QED

References

- [Adams 1975] Adams, E. 1975. *The logic of conditionals*. Dordrecht, The Netherlands: D. Reidel.
- [Geffner 1989] Geffner, H. 1989. Default reasoning: causal and conditional theories. UCLA Cognitive Systems Laboratory *Technical Report (R-137)*, December 1989. PhD. dissertation.
- [Goldszmidt and Pearl 1989a] Goldszmidt, M. and Pearl, J. 1989. On the consistency of defeasible databases. *Proc. 5th Workshop on Uncertainty in AI*, Windsor, Ontario, Canada, pp. 134-141.
- [Goldszmidt and Pearl 1989b] Goldszmidt, M. and Pearl, J. 1989. On the relation between rational closure and System-Z. UCLA Cognitive Systems Laboratory, *Technical Report (R-139)*, December 1989. Submitted. Since publishing, in *Proceedings of the Third International Workshop on Nonmonotonic Reasoning*, S. Lake Tahoe, CA, pp. 130-140.
- [Goldszmidt and Pearl 1989c] Goldszmidt, M. and Pearl, J. 1989. A maximum entropy approach to nonmonotonic reasoning. UCLA Cognitive Systems Laboratory, Technical Report R-132, in preparation. Since publishing, in *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 15 (no. 3) pp. 220-232.
- [Hanks and McDermott 1986] Hanks, S. and McDermott, D. V. 1986. Default reasoning, nonmonotonic logics, and the frame problem. *Proc., 5th Natl. Conf. on AI (AAAI-86)*, Philadelphia, pp. 328-33.
- [Hunter 1989] Hunter, D. 1989. Causality and maximum entropy updating. *Intl. Journal of Approximate Reasoning*. 3 (no. 1) pp. 87-114.
- [Lehmann 1989] Lehmann, D. 1989. What does a conditional knowledge base entail? *Proc. 1st Intl. Conf. on Principles of Knowledge Representation and Reasoning (KR'89)*, Toronto, May 1989, pp. 212-222, San Mateo: Morgan Kaufmann Publishers.
- [Lehmann and Magidor 1988] Lehmann, D. and Magidor, M. 1988. Rational logics and their models: a study in cumulative logics. Dept. of Computer Science, Hebrew University, Jerusalem, Israel, Technical Report #TR-88-16.
- [Pearl 1988] Pearl, J. 1988. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, San Mateo: Morgan Kaufmann Publishers.
- [Pearl 1989] Pearl, J. 1989. Probabilistic semantics for nonmonotonic reasoning: a survey. *Proc. 1st Intl. Conf. on Principles of Knowledge Representation and Reasoning (KR'89)*, Toronto, May 1989, pp. 505-516, San Mateo: Morgan Kaufmann Publishers.
- [Shoham 1986] Shoham, Y. 1986. Chronological ignorance: Time, nonmonotonicity, necessity, and causal theories. *Proc., 5th Natl. Conf. on AI (AAAI-86)*, Philadelphia, pp. 389-93.
- [Shoham 1987] Shoham, Y. 1987. Nonmonotonic logics: meaning and utility. *Proc. Intl. Joint Conf. on AI (IJCAI-87)*, Milan, pp. 388-393.

IV

PART

CAUSALITY 1988–2001

15

Introduction by Judea Pearl

When I started working on probabilistic reasoning, in the early 1980s, I thought that reasoning with uncertainty was the most important thing missing from artificial intelligence (AI). Moreover, I insisted that uncertainty be represented by probabilities, and this led to Bayesian networks, which I took to be the most plausible model of human cognition and decision-making. The message passing architecture that has evolved has given Bayesian networks a computational advantage over its rule-based rivals, primarily through its transparent semantics and programming simplicity. Given that we see certain facts, the network can swiftly compute the likelihood that certain other facts are true or false. Not surprisingly, Bayesian networks caught on in the AI community and even today are considered a leading paradigm in artificial intelligence for reasoning under uncertainty.

Although I am delighted with the ongoing success of Bayesian networks, they failed to bridge the gap between artificial and human intelligence. The missing ingredient was causality. True, causal ghosts were all over the place. The arrows invariably pointed from causes to effects, and practitioners often noted that diagnostic systems became unmanageable when the direction of the arrows was reversed. But for the most part we thought that this was a cultural habit, or an artifact of old thought patterns, not a central aspect of intelligent behavior.

At the time, I was so intoxicated with the power of probabilities that I considered causality a subservient concept, merely a convenience or a mental shorthand for expressing probabilistic dependencies and distinguishing relevant variables from irrelevant ones.

In my 1988 book *Probabilistic Reasoning in Intelligent Systems*, I wrote, “Causation is a language with which one can talk efficiently about certain structures of relevance relationships.” [Pearl 1988]. The words embarrass me today because “relevance” is so obviously an associational (rung-one) notion. Even by the time the

book was published, I knew in my heart that I was wrong. To my fellow computer scientists, the book became the bible of reasoning under uncertainty, but I was already feeling like an apostate.

Bayesian networks inhabit a world where all questions are reducible to probabilities, or degrees of association between variables. Using the ladder metaphor of the causal hierarchy, they could not ascend to the second or third rungs of the Ladder of Causation. Fortunately, they required only two slight twists to climb to the top. First, in 1991, the graph-surgery idea (that I learned from Peter Spirtes) empowered them to handle both observations and interventions. Another twist, in 1994, brought them to the third level and made them capable of handling counterfactuals. These two developments, described in the next section, made probabilities subservient to causality. While probabilities encode our beliefs about a static world, causality tells us whether and how probabilities change when the world changes, be it by intervention or by act of imagination. I am glad these early developments are given a stage in this volume.

The paper “Equivalence and synthesis of causal models” (Chapter 16) was a milestone in our transition from probabilistic to causal models [Verma and Pearl 1991]. Written barely a year after Thomas Verma proved the correctness of d -separation, this paper introduces several innovations that later became cornerstones of causal analysis. These include functional semantics for causal models, tests for Markov equivalence, inducing paths, Verma constraints, and the IC-algorithm, one of the first to discover causal structures from data. It was a true milestone.

The paper “Probabilistic evaluation of counterfactual queries” (Chapter 17) by Alex Balke and myself, was the first to demonstrate how structural models can be used to compute an arbitrary counterfactual expression [Balke and Pearl 1994]. Starting with the functional semantics of structural models, it introduced the 3-step procedure later dubbed “abduction, action, and prediction,” and invoked “response function variables” as nodes in the graph. These were later labeled “mapping variables” by Heckerman and Shachter [1995] and “principal stratification” by Frangakis and Rubin [2002].

“Causal diagrams for empirical research” (Chapter 18) was the culmination of a five-year effort to develop a comprehensive theory of identification of causal effects [Pearl 1995]. This includes: the back-door and front-door criteria, the *do*-calculus, surrogate experiments, and more. Semantically, the paper owes its inception to Peter Spirtes’ encoding of interventions as removal of arrows in the diagram. Professionally, I will always be indebted to Phil Dawid for his courage and leadership

in publishing this paper in *Biometrika*, thus introducing causal diagrams to empirical researchers. As expected, the discussions that followed were far from complementary. Paul Rosenbaum went as far as arguing that “no basis is given for believing that ... wiping out of equations predicts a certain physical reality.” Rubin and Imbens warned readers that “graphs lull researchers into a false sense of confidence.” Overall, this paper made causal diagrams the native language of causal inference.

In 1999, while writing chapter 9 of *Causality* [Pearl 2000], I stumbled upon a remarkable discovery. Although counterfactual descriptions of an individual behavior cannot in general be inferred from population data, they can nevertheless be bounded by combining data from both experimental and observational studies. Moreover, these bounds may sometimes collapse to point estimates. The paper “Probabilities of causation: Three counterfactual interpretations and their identification” (Chapter 19) describes these findings which have later inspired a fertile area of research, including personalized medicine, legal decisions, identifying causes of effects, and precision marketing [Tian and Pearl 2000; Pearl 2015; Li and Pearl 2019; Mueller et al. 2021].

“Direct and indirect effects” (Chapter 20) was my first paper on mediation analysis. It appeared a year after the publication of *Causality* [Pearl 2000], where I made the unfortunate suggestion that “the notion of indirect effect has no intrinsic operational meaning apart from providing a comparison between the direct and the total effects.” The full story of what made me change my mind is narrated in *The Book of Why* [Pearl and Mackenzie 2018] together with the enormous impact that “Direct and indirect effects” has had on mediation analysis. Recent concerns about “algorithmic fairness” and its societal implications have stimulated renewed interest in mediation analysis, especially the mediation formula derived in this paper.

References

- A. Balke and J. Pearl. 1994. Probabilistic evaluation of counterfactual queries. In *Proceedings of the Twelfth National Conference on Artificial Intelligence*. Volume I, MIT Press, Cambridge, MA, 230–237.
- C. Frangakis and D. Rubin. 2002. Principal stratification in causal inference. *Biometrics* 1, 21–29. DOI: <https://doi.org/10.1111/j.0006-341x.2002.00021.x>.
- D. Heckerman and R. Shachter. 1995. Decision-theoretic foundations for causal reasoning. *J. Artif. Intell. Res.* 3, 405–430. DOI: <https://doi.org/10.1613/jair.202>.
- A. Li and J. Pearl. 2019. Unit selection based on counterfactual logic. In *Proceedings, IJCAI-19*. 1793–1799.

- S. Mueller, A. Li, and J. Pearl. 2021. *Causes of effects: Learning individual responses from population data*. UCLA Cognitive Systems Laboratory, Technical Report R-505. <http://arxiv.org/abs/2104.13730>.
- J. Pearl. 1988. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann.
- J. Pearl. 1995. Causal diagrams for empirical research. *Biometrika* 82, 4.
- J. Pearl. 2000. *Causality: Models, Reasoning, and Inference*. Cambridge University Press.
- J. Pearl. 2015. Causes of effects and effects of causes. *J. Sociol. Methods Res.* 44, 1, 149–164. DOI: <https://doi.org/10.1177/0049124114562614>.
- J. Pearl and D. Mackenzie. 2018. *The Book of Why*. Basic Books, New York.
- J. Tian and J. Pearl. 2000. Probabilities of causation: Bounds and identification. *Ann. Math. Artif. Intell.* 28, 287–313. DOI: <https://doi.org/10.1023/A:1018912507879>.
- T. S. Verma and J. Pearl. 1991. Equivalence and synthesis of causal models. In *Proceedings of the 6th Conference in Artificial Intelligence*. Association for Uncertainty in AI, 220–227.

16

Equivalence and Synthesis of Causal Models

TS Verma[†] and Judea Pearl

Abstract

Scientists often use directed acyclic graphs (dags) to model the qualitative structure of causal theories, allowing the parameters to be estimated from observational data. Two causal models are equivalent if there is no experiment which could distinguish one from the other. A canonical representation for causal models is presented which yields an efficient graphical criterion for deciding equivalence, and provides a theoretical basis for extracting causal structures from empirical data. This representation is then extended to the more general case of an embedded causal model, that is, a dag in which only a subset of the variables are observable. The canonical representation presented here yields an efficient algorithm for determining when two embedded causal models reflect the same dependency information. This algorithm leads to a model theoretic definition of causation in terms of statistical dependencies.

This work was supported, in part, by NSF grant IRI-88-2144 and NRL grant N000-89-J-2007.

[†]University of California. Supported by an IBM graduate fellowship.

P.P. Bonissone, M. Henrion, L.N. Kanal and J.F. Lemmer (Editors)

Originally published in *Uncertainty in Artificial Intelligence 6*

© 1991 Elsevier Science Publishers B.V. All rights reserved. Republished with permission of Elsevier.

16.1 Introduction

The use of dags as a language for describing causal models has been popular in the behavioral sciences [Blalock 71], [Duncan 75] and [Wright 34], decision analysis [Howard and Matheson 81][Olmsted 84] and [Shachter 85] and evidential reasoning [Pearl 88], and has also received extensive theoretical studies [Geiger and Pearl 90], [Geiger et al 90], [Glymour et al 1987], [Pearl and Verma 87], [Shachter 85], [Smith 89], [Spirtes et al 90] and [Verma and Pearl 90]. One problem that has arisen in the course of these studies is that of non-uniqueness; it is quite common for two different causal models to be empirically indistinguishable, hence, equally predictive. This occurs when each of the two models can mimic the behavior of the other. Formally:

Definition 16.1 A causal theory is a pair $T = \langle D, \Theta_D \rangle$ consisting of a causal model D and a set of parameters Θ_D compatible with D . Θ_D assigns a function $x_i = f_i[\mathbf{pa}(x_i), \epsilon_i]$ and a probability measure g_i , to each $x_i \in U$, where $\mathbf{pa}(x_i)$ are the parents of x_i in D and each ϵ_i is a random disturbance distributed according to g_i , independently of the other ϵ 's and of any preceding variable $x_j : 0 < j < i$.

Definition 16.2 Two causal models D_1 and D_2 are **equivalent** if for every theory $T_1 = \langle D_1, \Theta_1 \rangle$ there is a theory $T_2 = \langle D_2, \Theta_2 \rangle$ such that T_1 and T_2 define the same probability distribution, and vice versa.

For example, consider the four causal models of Figure 16.1. The parameters required for the first model are $P(a)$, $P(b|a)$ and $P(c|b)$. The second requires estimations for $P(b)$, $P(a|b)$ and $P(c|b)$. It is easy to see that these two models are equivalent by the definition of conditional probability, i.e. $P(a)P(b|a) = P(ab) = P(b)P(b|a)$. Thus the values obtained for the first set of parameters completely determine the values of the second, and vice versa. Similarly, the third model is equivalent to the first two since its parameters, $P(c)$, $P(b|c)$ and $P(a|b)$ can be determined from either

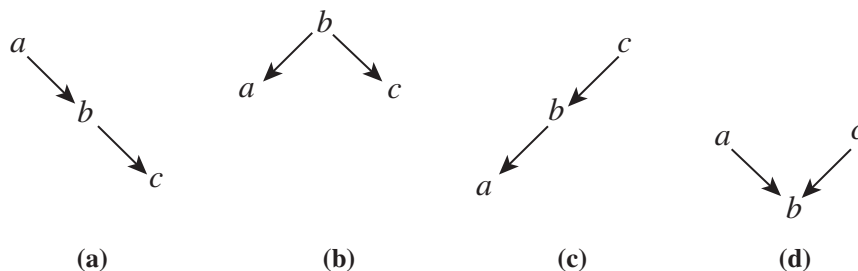


Figure 16.1 Three of the four models are equivalent.

of the first two sets. However, the fourth model is quite different; its parameters are $P(a)$, $P(c)$ and $P(b|ac)$ which cannot be determined from any of the previous sets.

The fact that the first three models are equivalent to each other but not the fourth is easily seen in terms of the independence information conveyed by the corresponding dags. The first three all represent the independence statement $I(a, b, c)$ which is read “ a is independent of c given b ” whereas the fourth represents the statement $I(a, \emptyset, c)$, which is read “ a is marginally independent of c ”. The statistical meaning of any causal model can be described completely and economically by its *stratified protocol*, which is a list of independence statements, each asserting that a variable is independent of its non-descendants, given its parents [Geiger and Pearl 90], [Pearl and Verma 87] and [Verma and Pearl 90]. Furthermore, any independence statement that logically follows from the stratified protocol can be graphically determined in linear time via the *d-separation* criterion [Geiger et al 89] and [Geiger et al 90]. Thus, the question of equivalence of causal models reduces to the question of equivalence of protocols: two dags are equivalent if and only if each dag’s protocol holds in the other [Pearl et al 89]. This solution is both intuitive and efficient. However, it has two drawbacks; it is difficult to process visually and it does not generalize to embedded causal models.

Embedded causal models are useful for modeling theories that cannot be modeled via simple dags. For example, if there are unobserved variables which cause spurious correlations between the observable variables it may be necessary to embed the observables in a larger dag containing “hidden” variables in order to build an accurate model. Even when there exists a simple causal model that fits a theory, it might be desirable to embed the model in a larger dag to satisfy some higher level constraints. For example, suppose that every causal model that fits a given set of data contains the link $a \rightarrow b$, but b is known to precede a . Under these circumstances, the simple causal models are inconsistent with our common notion of the temporal direction of causality; one way of avoiding this conflict is to hypothesize the existence of an unknown common cause, i.e. $a \leftarrow \alpha \rightarrow b$. See Figure 16.2 as well as Figures 16.4 and 16.5 for examples of the use of hidden variables, (denoted by greek letters).

Figure 16.2 illustrates a special problem that embedded causal models pose. Unlike simple causal models, the statistical meaning of an embedded causal model cannot be completely characterized by dependency information alone; two dependency equivalent causal models need not be equivalent in the general sense. For example both embedded causal models of Figure 16.2 represent the dependency statement $I(a, b, c)$, but the first model (a) imposes an additional constraint upon

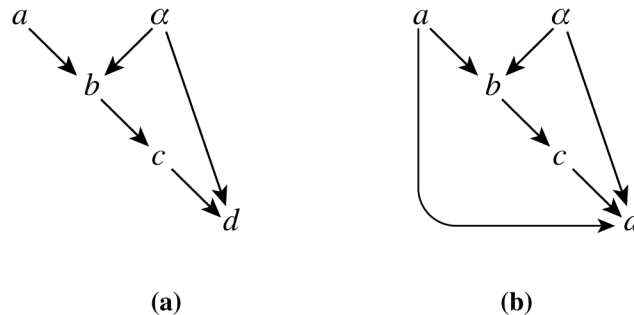


Figure 16.2 Two dependency equivalent embedded causal models which are not equivalent in general.

the set of distributions it can describe:

$$\sum_b P(b|a)P(d|abc) = f(c, d)$$

Fortunately, dependency equivalence is a tight enough necessary condition for equivalence that it permits many sound conclusions to be derived by graphical means.

This paper is organized as follows. Section 16.2 provides an efficient criterion for deciding the equivalence of two models, and a canonical representation called a *pattern* for describing the class of all models equivalent to a given dag. Section 16.3 extends this construction to the case of embedded causal models. Theorems will be stated without proofs, a full detail of which can be found in [Verma 91]. In Section 16.4, the Theorems of the previous two sections are applied to the problem of recovery of a causal model from statistical data.

16.2 Patterns of Causal Models

It is not difficult to observe that equivalent dags have common features. For example, two dags that represent equivalent causal models must have the same adjacency structure. Two nodes of a dag are adjacent, written \overline{ab} if either $a \rightarrow b$ or $a \leftarrow b$. That adjacency is invariant among equivalent dags follows from Lemma 16.1 which describes the principal relationship between adjacency and unseparability¹ (parts 1 and 2) as well as the relationships between separability and d-separation² given two particular special sets of nodes in the dags (parts 3 and 4). Let the ancestor set A_{ab} of a pair of variables a and b be defined as the union of the sets of

1. Two variables are unseparable just in case there is no set that d-separates them.
2. The predicate $I_D(\cdot)$ denotes d-separation in the dag D .

ancestors of a and b (less ab), and similarly, the parent set P_{ab} of the pair be defined as the union of the sets of parents of a and b (less ab).

Lemma 16.1 *Let a and b be two nodes of a dag D ; the following four conditions are equivalent:*

- (1) a and b are adjacent in D
- (2) a and b are unseparable in D
- (3) a and b are not d -separated by A_{ab} in D
- (4) a and b are not d -separated by P_{ab} in D

Proof. (Sketch) That (1) implies (2) follows from the fact that a link is a path which cannot be deactivated; and (2) trivially implies (3) since unseparability means the lack of d -separation in any context, including A_{ab} . Since every path activated by P_{ab} is also activated by A_{ab} , it follows that (3) implies (4). The final implication, that (4) implies (1) follows from the observation that if a and b are not d -separated given P_{ab} , then there must be active path between them. If this path contains a node, other than a or b , it would have to contain at least one head-to-head node since the path is active given P_{ab} . The head-to-head node nearest to a on the path would be a descendant of a , similarly the one nearest b would be a descendant of b , again because the path is active given P_{ab} . Any such of these head-to-head nodes would have to be in or be an ancestor of a node in P_{ab} for the path to be active, but the one nearest a could not be an ancestor of a , hence both it and a would be ancestors of b . Similarly, both b and the head to head node nearest it would have to be ancestors of a , but this would imply the existence of a directed loop, hence the path cannot contain any nodes other than a and b . Therefore the nodes are adjacent. ■

The major consequence of this lemma is that adjacency is a property determined solely by d -separation, hence remains invariant among equivalent dags.

A set of equivalent dags possesses another important invariant property, namely the directionality of the uncoupled head-to-head links (i.e. $a \rightarrow b \leftarrow c$ are *uncoupled* if a and c are not adjacent). There are other links whose directionality remains invariant, but these can easily be determined from the uncoupled head-to-head links. The following lemma summarizes this important class of links with invariant directionality.

Lemma 16.2 *In any dag D , if the nodes a, c, b form a chain \overline{acb} while a and b are d -separated by some set S but not Sc then $a \rightarrow c \leftarrow b$.*

Furthermore if $a \rightarrow c \leftarrow b$ then a and b are unseparable by any set containing c .

The proof of this lemma relies upon the inherent differences between a head-to-head junction and the other types of junctions (tail-to-tail and head-to-tail). The major ramification of Lemma 16.2 is that the directionality of a certain class of

links can be determined from d-separation alone. The implications this may have on the prospects of inferring causal relationships from independence statements are briefly discussed in Section 16.4 and in detail in [Verma 91].

Together, these lemmas form a necessary and sufficient condition for equivalence, previously stated in [Pearl et al 89]:

Theorem 16.1 *Two dags are equivalent if and only if they have the same links and same uncoupled head-to-head nodes.*

The proof of this theorem is based on the lemmas along with an inductive step showing that every active path in one dag has a corresponding active path in the other. The importance of Theorem 16.1 is that the equivalence of two causal models can be determined by a simple graphical criterion.

Since the two invariant properties of a dag identified in the lemmas are a sufficient condition for equivalence, they lead to a natural canonical representation of its equivalent class. Simply construct a *partially directed graph* by removing the arrowheads from any link of the dag that is not identified by Lemma 16.2. This partially-directed graph will be called the *rudimentary pattern* of the causal model. Since the rudimentary pattern can be defined solely in terms of d-separation, it follows that each equivalence class of causal models has a unique pattern; hence, two causal models are equivalent if and only if they have the same pattern. This is a useful view of the problem since the patterns can be constructed efficiently³.

Lemma 16.2 only identifies some of the invariant arrowheads of a causal model, but since identification of this class is sufficient for deciding equivalence, it follows that the remainder of the invariant arrowheads are completely determined by this class. It is not difficult to identify the remainder of the invariant arrowheads as some of the undirected links of a rudimentary pattern cannot be arbitrarily directed without either (1) creating a new uncoupled head-to-head node or (2) creating a directed loop. Since these undirected links are essentially constrained to a certain direction, it is desirable to define a *completed pattern* in which they are directed as constrained. The completed pattern reflects each and every invariant arrow head. Furthermore, both rudimentary patterns and completed patterns offer a compact summary of each and every dag in an equivalence class.

For example, in Figure 16.3, the rudimentary pattern (d) and the completed pattern (e) each summarizes the dags in the equivalence class $\{ (a), (b), (c) \}$. Any extension of either pattern into a full dag that does not create new uncoupled head-to-head nodes will be a dag in the equivalence class. There are three such extensions in the example of Figure 16.3.

3. Note that comparison of patterns is polynomial since the nodes are labeled.

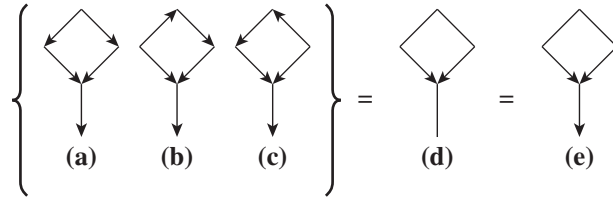


Figure 16.3 Equivalence class of models.

16.3 Embedded Causal Models

Partially-directed graphs offer an excellent tool for describing the equivalence classes of causal models; it would be desirable to find a similar structure for embedded causal models. Such a structure requires the ability to represent a direct non-causal correlation between two variables. In a simple dag, whenever two variables are unseparable, there must be a directed link between them, dictating that either the first causes the second or the second causes the first. There is no way to represent the existence of an unknown common cause, as illustrated in the following embedded causal model (Figure 16.4 (a)). Assume a, b, c and d are the observables and α is unobservable. There is no dag that can represent the dependencies between a, b, c and d using these variables only. However, the *hybrid graph* (Figure 16.4 (b)) which contains a *bi-directional* link does represent these dependencies. (Under a natural extension of d-separation [Verma 91].)

For hybrid graphs, the notation \overrightarrow{ab} denotes the existence of a link with at least an arrow head pointing at b , namely either $a \rightarrow b$ or $a \leftrightarrow b$, while \overline{ab} denotes the existence of a link without any constraints on its orientation. Thus, for example, when applied to a dag, \overline{ab} means $a \rightarrow b$ or $a \leftarrow b$; while in hybrid graphs \overline{ab} denotes the existence of any of the four possible types of links, (namely, $a - b, a \rightarrow b, a \leftarrow b$ and $a \leftrightarrow b$). Hybrid graphs can be used to represent *patterns* of embedded causal models according to the following definition.

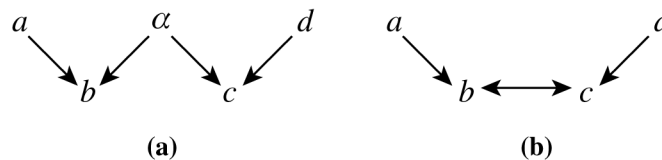


Figure 16.4 The representation of a hidden common cause.

Definition 16.3 Embedded Pattern

Given a dag D over the variables U_D , of which $U_O \subseteq U_D$ are observable, the rudimentary pattern P of D restricted to U_O is defined as the hybrid graph with fewest arrowheads that satisfies the following conditions:

- (1) $\overline{ab} \in P \Leftrightarrow \neg I_D(a, S, b) \forall S \subseteq U_O - ab$
- (2) \overrightarrow{ab} if $\exists c \in U_O$ such that: $\overline{abc} \in P, \overline{ac} \notin P$ and $\neg I_D(a, Sb, c) \forall S \subseteq U_O - abc$

Rudimentary embedded patterns can be extended into completed embedded patterns (or simply, embedded patterns) in much the same way that simple patterns are completed. The same constraints can be used for the completion, namely, no arrow head can be added to the pattern that would (1) create a new uncoupled head-to-head node or (2) create a strictly directed cycle. However, note that a strictly directed cycle contains only singly directed arrows.

While this defines a unique pattern for every embedded dag, it does so in terms of d-separation conditions over subsets all of U_O , which, in principle, might require an exponential number of tests. The next two lemmas show that patterns can be formed in polynomial time. Lemma 16.3 delineates the relationship between adjacency in the pattern and unseparability in the causal model (parts 1 and 2) and provides a practical criterion for determining separability in terms of a simple d-separation test (part 3) and a graphical test (part 4). The graphical test is defined in terms of an *inducing path*:

Definition 16.4 Inducing Path

An inducing path between the variables a and b of an embedded causal model is any path ρ satisfying the following two conditions:

- (1) Every observable node on ρ is head-to-head on ρ .
- (2) Every head-to-head node on ρ is in A_{ab} .

Lemma 16.3 Let P be the pattern of a dag D with respect to the observables $U_O \subset U_D$ and $a, b \in U_O$ be two observables; the following statements are equivalent:

- (1) a and b are adjacent in P
- (2) a and b are unseparable in D (over U_O)
- (3) a and b are not d-separated by $A_{ab} \cap U_O$ in D
- (4) a and b are connected by an inducing path in D

Proof. (Sketch) By definition, (1) is equivalent to (2) and (2) implies (3). To show that $\neg I_D(a, A_{ab} \cap U_O, b)$ implies the existence of an inducing path, consider that this dependency implies the existence of a path ρ , between a and b which is active given $A_{ab} \cap U_O$. Since $A_{ab} \cap U_O$ only contains ancestors of a and b it follows that every

head-to-head node on ρ must be in A_{ab} . Thus any observable node on ρ that is not head-to-head would be in $A_{ab} \cap U_O$ and would serve to deactivate the path, so every observable node on ρ must be head-to-head. Therefore ρ is an inducing path.

To show that the existence of an inducing path implies unseparability relative to U_O hence finish the proof, consider any two nodes a and b which are connected by an inducing path ρ . To show a and b are not d-separated in any context of U_O , consider any context S which deactivates ρ (if ρ is active for every context, then the two nodes are unseparable). Since the only observable nodes of ρ are head-to-head, only head-to-head nodes could serve to deactivate ρ . Each head-to-head node on ρ must be in A_{ab} and at least one must be inactive, given S (otherwise the path would be active given S). If all inactive head-to-head nodes are ancestors of a then consider the one closest to b , call it y . The portion of ρ between y and b is active, and the ancestry path from y to a can be added to form an active path between a and b given S . On the other hand, if any of the inactive head-to-head nodes is ancestor of b then pick the head-to-head ancestor of b which is closest to a on ρ and call it x . Every inactive head-to-head node between a and x must be an ancestor of a (if any exist), hence there must be an active path between a and x (either the portion of ρ between a and x , or the ancestry path from the head-to-head node between a and x which is closest to x concatenated with the portion of ρ from that node to x). Since x is an ancestor of b , the ancestry path from x to b can be concatenated to the path from a to x to form an active path between a and b given S . Thus a and b are unseparable. ■

Lemma 16.3 describes how links are induced in P by paths of D . The next lemma will describe how to determine the directionality of these links in terms of the inducing paths.

Lemma 16.4 *For any rudimentary pattern P , \overrightarrow{ab} if and only if there is a node c adjacent to b but not to a (in P) such that both edges \overline{ab} and \overline{bc} were induced by paths (of D) which ended pointing at b .*

Lemmas 16.3 and 16.4 provide a polynomial time algorithm for constructing the characteristic pattern of any embedded causal model. The final theorem completes the original task of deciding dependency equivalence.

Theorem 16.2 *Two embedded causal models are dependency equivalent if and only if they have identical completed patterns.*

Thus, Theorem 16.2 gives validity to the notion of a pattern as a characteristic representation of an embedded causal model. An interesting consequence of this theorem is given by the following corollary:

Corollary 16.1 *There are fewer than $5^{|U_O|^2}$ distinct embedded causal models containing $|U_O|$ variables; moreover, every embedded causal model is equivalent to a simple dag with fewer than $|U_O|^2$ variables.*

Part 1 follows from the fact that every embedded causal model is equivalent to its pattern, and every pattern contains fewer than $|U_O|$ edges (there are four types of edges). The second part stems from the fact that a bi-directional link $a \leftrightarrow b$ in a pattern can be represented by a single hidden common cause α of the observable variables, namely, $a \leftarrow \alpha \rightarrow b$.

Figure 16.5 contains three embedded causal models (a), (b) and (c) over the observable variables $\{a, b, c, d, e\}$ as well as their completed patterns (a'), (b') and (c') respectively. The patterns indicate that the first two causal models are equivalent to each other but not to the third; while d and e are marginally independent in

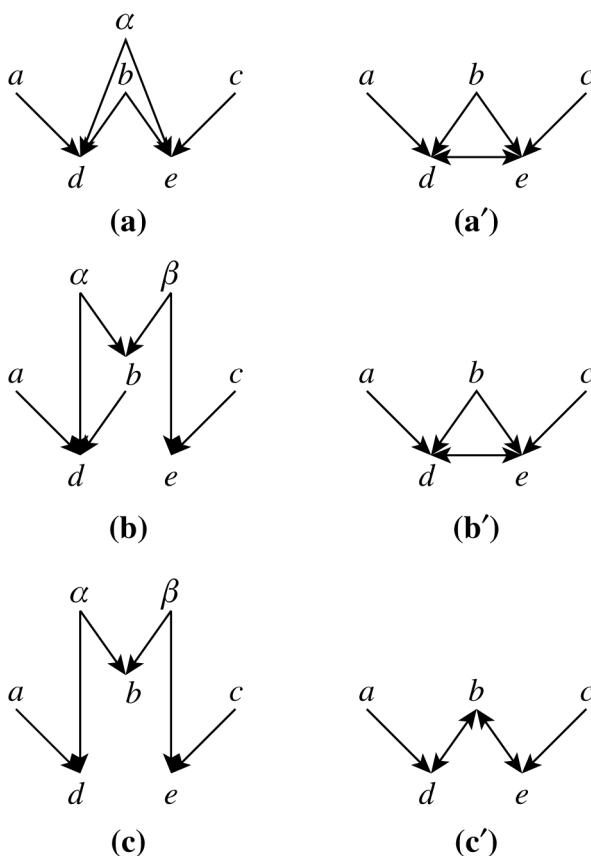


Figure 16.5 The patterns reveal which two models are dependency equivalent.

(c) they are dependent in both (a) and (b). Figure 16.5 (b) demonstrates that a hidden common cause is not equivalent to a bi-directional link since it is important to recognize the paths they may induce.

16.4 Applications to the Synthesis of Causal Models

The problem of deciding the equivalence of (embedded) causal models is fundamental to causal reasoning and theory building, as it allows us to determine which structural properties of the model (e.g. connectivity or directionality) can be substantiated by data and which serve merely for representational convenience. The canonical representations presented in this paper offer an efficient solution to this problem since they can be constructed (from the causal models) in polynomial time. They can also be used to solve the broader problem of model subsumption [Verma 91].

The construction of these canonical representations is based on (conditional) independence relationships, thus suggesting the possibility of extracting causal models directly from statistical information. Such application meets with the difficulty that, in general, probability distributions do not define unique graphical models. In other words, given that the data is generated by some causal theory $T = \langle D, \Theta \rangle$, it is always possible to contrive the parameters Θ to yield spurious independencies, not shown in D , that fit another theory $T' = \langle D', \Theta' \rangle$, with D' not equivalent to D . [Spirtes et al 90] show that, under some reasonable assumptions, the occurrence of such spurious independencies is a rare event of measure zero, and therefore argue that it is natural in causal modeling to assume that the underlying distribution is dag-isomorphic,⁴ albeit allowing for the inclusion of unobserved variables.

Under the assumption that the observed distribution is dag-isomorphic, Theorem 16.1 permits the recovery of the underlying structure uniquely, modulo the equivalence class defined by its pattern. One such recovery algorithm is proposed in [Spirtes et al 90] and several alternatives are discussed in the sequel.

Our basic algorithm has four parts; the first part is an application of Lemma 16.1 that identifies the links of the pattern. The second part of the algorithm is an application of Lemma 16.2 which adds directionality to some of the links, thus forming the rudimentary pattern. The third part of the algorithm consists of completing the rudimentary pattern into a full pattern; and the final part marks those links that are invariant over all dependency equivalent embedded causal models.

4. A probabilistic distribution is dag-isomorphic permitting all its dependencies and independencies to be displayed in some dag.

IC-Algorithm (Inductive Causation)Input: \hat{P} a sampled distribution.Output: $\text{core}(\hat{P})$ a marked hybrid acyclic graph.

1. For each pair of variables a and b , search for a set S_{ab} such that (a, S_{ab}, b) is in $I(\hat{P})$, namely a and b are independent in \hat{P} , conditioned on S_{ab} . If there is no such S_{ab} , place an undirected link between the variables.
2. For each pair of non-adjacent variables a and b with a common neighbor c , check if $c \in S_{ab}$.
If it is, then continue.
If it is not, then add arrowheads pointing at c , (i.e. $a \rightarrow c \leftarrow b$).
3. Form $\text{core}(\hat{P})$ by recursively adding arrowheads according to the following two rules:
If \overline{ab} and there is a strictly directed path from a to b then add an arrowhead at b .
If a and b are not adjacent but \overline{ac} and $c - b$, then direct the link $c \rightarrow b$.
4. If \overline{ab} then mark every uni-directed link $b \rightarrow c$ in which c is not adjacent to a .

The complexity of this algorithm is bounded by the first step, which by brute force would require an exponential search for the set S_{ab} . It can be greatly reduced by the generation of a Markov network. A Markov network is the undirected graph formed by linking every pair of variables a and b that are dependent given the rest of the variables (i.e. $\neg I(a, U - ab, b)$). The Markov network of a dag-isomorphic distribution has the property that the parents of any variable in the dag form a clique in the network. Since Lemma 16.1 states that any two variables a and b are separable if and only if they are separated by their parent set P_{ab} , the search for a separating set can be confined to the cliques that contain either a or b . Thus, the complexity is bounded, exponentially, by the size of the largest clique in the Markov network, and this coincides with the theoretical lower bound for recovery of a dag from independence information [Verma 91].

One drawback of the Markov network reduction is that it is not applicable to embedded causal models because it rests on part (4) of Lemma 16.1; no parallel lemma exists for embedded models. However, the basic algorithm stated above, by virtue of resting on Theorem 16.2 can be used to recover embedded causal model as well. The only difference is in the output; when the algorithm is applied to a dag-isomorphic distribution, every link is guaranteed to be assigned at most one arrowhead (a particular arrowhead may actually be assigned multiple times, but no

link will receive an arrowhead on both ends). However, when the distribution is isomorphic to an embedded dag it is possible for a link to be assigned an arrowhead on both ends, hence the recovery of a bi-directional link.

The invariant nature of the arrows in a pattern can form the basis for a general non-temporal definition of causation; one that determines the direction of causal influences from statistical data without resorting to chronological information, and one that applies to general distributions, including those that are not isomorphic to embedded dags. The essence of this definition can be articulated by taking as *models* of our theory the set \mathcal{P} of all patterns that are consistent with an observed distribution, namely, patterns that represent the minimal causal models of the distribution (see [Pearl and Verma 1991] for the definition of minimality).

Definition 16.5 Genuine and Potential Cause

c is a genuine cause of e if c causes e in every consistent model (i.e. every pattern of \mathcal{P} contains the directed arrow $c \rightarrow e$). c is a potential cause of e if c causes e in some consistent model (i.e. some pattern of \mathcal{P} contains $c \rightarrow e$) and e never causes c in any consistent model (i.e. no pattern of \mathcal{P} contains $c \rightarrow e$).

The IC-algorithm identifies every potential cause by assignment of a uni-directional link, and it marks those which are in fact genuine. The vertical arrow in Figure 16.3 (e) is an example of a genuine cause, since this arrow cannot be emulated by a hidden common cause of the two end points (in any consistent embedded model). The other arrows in Figure 16.3 (e) represent potential causes when viewed in the context of embedded models, because each can be represented by a common hidden cause in some equivalent causal model.

Since the number of patterns over $|U|$ variables is finite, Definition 16.5 is operational. However, the existence of an effective algorithm which can determine causation by means other than enumerating the patterns of \mathcal{P} is an open question. If the observed distribution is isomorphic to an embedded dag, then \mathcal{P} contains only one unique pattern; that which is generated by the recovery algorithm. This pattern contains all the information required for identifying the genuine and potential causes [Verma 91]. However, when applied to general distributions the arrows assigned in the generated pattern may or may not coincide with the model-theoretic definition of genuine and potential causes. A more detailed treatment of the model-theoretic definition of causation, including a set of sound causal relationships sanctioned by this definition can be found in [Pearl and Verma 91].

[Spirtes et al 90] have proposed an algorithm for identifying causal relationships which accepts many, but not all, of the genuine and potential causes in distributions that are isomorphic to embedded dags. The relationships identified

by [Spirtes et al 90] correspond to the singly directed arrows of the rudimentary pattern.

In practice, every recovery algorithm must face the problem of inferring independence relations from sampled data. The number of samples required to reliably test the assertion $I(a, S_{ab}, b)$ grows exponentially with the size of S_{ab} . A reasonable approximating algorithm for recovering a dag (or embedded dag) could be devised based upon the following redefinition of the independence relation:

Definition 16.6 Reliable Independence

$I(a, S, b)$ holds reliably whenever the set of hypotheses $\{P(a|S) = P(a|Sb)\}$ is confirmed for each instantiation of S for which a sufficient number of samples are available to reliably test the hypothesis.

This notion of reliable independence is captured by taking as a measure of dependency the (conditional) sample cross entropy [Pearl 88, page 392]:

$$\hat{H}(a, b|S) \stackrel{\text{def}}{=} \sum_{a,b,S} \hat{P}(a, b, S) \log \frac{\hat{P}(a, b|S)}{\hat{P}(a|S)\hat{P}(b|S)}$$

where \hat{P} stands for the sample frequency and the summation ranges over all instantiations of a , b and S . We see that terms involving small samples (i.e., low values of $\hat{P}(a, b, S)$) are automatically discounted relative to those of larger samples.

One issue that has not been addressed is that of deterministic nodes, such as those representing functional dependencies among variables. These nodes cannot be completely represented by the causal models considered in this paper, as they require a refinement of d -separation studied in [Geiger et al 89] and [Pearl et al 89]. The issues introduced by deterministic nodes are discussed in [Verma 91].

Acknowledgments

The problem of deciding equivalence of embedded causal models was posed by Clark Glymour and communicated to us by Dan Geiger. We wish to thank Peter Spirtes for pointing out a careless omission in [Pearl and Verma, 1991], which did not properly reflect the definition of an embedded pattern (Definition 16.3).

References

- [Blalock 71] H.M. Blalock, *Causal Models in The Social Sciences*. Macmillan, London, 1971.
- [Duncan 75] O.D. Duncan, *Introduction to Structural Equation Models*. Academic Press, New York, 1975.
- [Geiger et al 89] D. Geiger, T.S. Verma and J. Pearl, d -Separation: From Theorems to Algorithms, *Proceedings*, 5th Workshop on Uncertainty in AI, Windsor, Ontario, Canada, August 1989, pp. 118-124.

- [Geiger and Pearl 90] D. Geiger and J. Pearl, Logical and Algorithmic Properties of Independence and Their Application to Bayesian Networks, *Annals of Mathematics and AI*, vol. 2 no. 1-4 pp 165-178, 1990.
- [Geiger et al 90] D. Geiger and T.S. Verma and J. Pearl, Identifying Independence in Bayesian Networks, *Networks*, vol. 20 no. 5 pp 507-534, 1990.
- [Glymour et al 1987] C. Glymour, R. Scheines, P. Spirtes and K. Kelly. *Discovering Causal Structure*. Academic Press, New York, 1987.
- [Howard and Matheson 81] R.A. Howard and J.E. Matheson, Influence Diagrams, chapter 8, in *The Principles and Applications of Decision Analysis*, Vol. II, Strategic Decisions Group, Menlo Park, California, 1981.
- [Olmsted 84] S.M. Olmsted, On Representing and Solving Decision Problems, Ph.D. Thesis, Engineering-Economic Systems Dept., Stanford University, Stanford California, 1984.
- [Pearl et al 89] J. Pearl, D. Geiger and T.S. Verma, The Logic of Influence Diagrams, in R.M. Oliver and J.Q. Smith (Eds), *Influence Diagrams, Belief Networks and Decision Analysis*, John Wiley and Sons, Ltd., Sussex, England 1990. A shorter version, in *Kybernetika*, Vol. 25:2, 1989, pp. 33-44.
- [Pearl 88] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers, Inc, San Mateo, California, 1988.
- [Pearl and Verma 87] J. Pearl and T.S. Verma, The Logic of Representing Dependencies by Directed Graphs, *Proceedings*, AAAI Conference, Seattle, WA. July, 1987, pp. 374-379.
- [Pearl and Verma 91] J. Pearl and T.S. Verma, A Theory of Inferred Causation, in J. A. Allen, R. Fikes and E. Sandwall (editors) *Principles of Knowledge Representation and Reasoning: Proceedings of the Second International Conference*. San Mateo: Morgan Kaufmann. April 1991, pp. 441-452.
- [Shachter 85] R.D. Shachter, Evaluating Influence Diagrams, in A.P. Basu (Eds), *Reliability and Quality Control*, Elsevier, 1985, pp. 321-344.
- [Spirtes et al 90] P. Spirtes, C. Glymour and R. Scheines, Causality from Probability, in G. McKee, ed., *Evolving Knowledge in Natural and Artificial Intelligence*, Pitman, 1990.
- [Smith 89] J.Q. Smith, Influence Diagrams for Statistical Modeling, *The Annals of Statistics*, Vol. 17(2):654-672, 1989.
- [Verma and Pearl 90] T.S. Verma and J. Pearl, Causal Networks: Semantics and Expressiveness, in *Uncertainty in AI 4*, R. Shachter, T.S. Levitt and L.N. Kanal (eds), Elsevier Science Publishers, 1990, pp. 69-76.
- [Verma 91] T.S. Verma, Graphical Aspects of Causal Models. UCLA Computer Science Department, Cognitive Systems Laboratory, Technical Report R-191, September 1992.
- [Wright 34] S. Wright, The Method of Path Coefficients. *Ann. Math. Statistics* 5:161-215, 1934.

Probabilistic Evaluation of Counterfactual Queries

Alexander Balke* and Judea Pearl

Abstract

Evaluation of counterfactual queries (e.g., “If A were true, would C have been true?”) is important to fault diagnosis, planning, and determination of liability. We present a formalism that uses probabilistic causal networks to evaluate one’s belief that the counterfactual consequent, C , would have been true if the antecedent, A , were true. The antecedent of the query is interpreted as an external action that forces the proposition A to be true, which is consistent with Lewis’ *Miraculous Analysis*. This formalism offers a concrete embodiment of the “closest world” approach which (1) properly reflects common understanding of causal influences, (2) deals with the uncertainties inherent in the world, and (3) is amenable to machine representation.

17.1 Introduction

A counterfactual sentence has the form

If A were true, then C would have been true

where A , the counterfactual antecedent, specifies an event that is contrary to one’s real-world observations, and C , the counterfactual consequent, specifies a result that is expected to hold in the alternative world where the antecedent is true. A typical instance is “If Oswald were not to have shot Kennedy, then Kennedy would still

*University of California, Los Angeles

Originally published in *Proceedings of the Twelfth National Conference on Artificial Intelligence (AAAI-94)*, pp. 230-237, 1994.

Republished with permission from AAAI.

be alive” which presumes the factual knowledge of Oswald’s shooting Kennedy, contrary to the antecedent of the sentence.

The majority of the philosophers who have examined the semantics of counterfactual sentences (Goodman 1983; Harper, Stalnaker, & Pearce 1981; Nute 1980; Meyer & van der Hoek 1993) have resorted to some form of logic based on worlds that are “closest” to the real world yet consistent with the counterfactual’s antecedent. Ginsberg (1986), following a similar strategy, suggested that the logic of counterfactuals could be applied to problems in planning and diagnosis in Artificial Intelligence. The few other papers in AI that have focused on counterfactual sentences (e.g., (Jackson 1989; Pereira, Aparicio, & Alferes 1991; Boutilier 1992)) have mostly adhered to logics based on the “closest world” approach.

In the real world, we seldom have adequate information for verifying the truth of an indicative sentence, much less the truth of a counterfactual sentence. Except for the small set of relationships between variables which can be modeled by physical laws, most of the relationships in one’s knowledge base are non-deterministic. Therefore, it is more practical to ask not for the truth or falsity of a counterfactual, but for one’s degree of belief in the counterfactual consequent given the antecedent. To account for such uncertainties, (Lewis 1976) has generalized the notion of “closest world” using the device of “imaging”; namely, the closest worlds are assigned probability scores, and these scores are combined to compute the probability of the consequent.

The drawback of the “closest world” approach is that it leaves the precise specification of the closeness measure almost unconstrained. More specifically, it does not tell us how to encode distances in a way that would (1) conform to our perception of causal influences and (2) lend itself to economical machine representation. This paper can be viewed as a concrete explication of the closest world approach, one that satisfies the two requirements above.

The target of our investigation are counterfactual queries of the form:

If A were true, then what is the probability that C would have been true, given that we know B ?

The proposition B stands for the actual observations made in the real world (e.g., that Oswald did shoot Kennedy and that Kennedy is dead) which we make explicit to facilitate the analysis.

Counterfactuals are intertwined with notions of causality: We do not typically express counterfactual sentences without assuming a causal relationship between the counterfactual antecedent and the counterfactual consequent. For example, we can safely state “If the sprinkler were on, the grass would be wet”, but the contrapositive form of the same sentence in counterfactual form, “If the grass were

dry, then the sprinkler would not be on”, strikes us as strange, because we do not think the state of the grass has causal influence on the state of the sprinkler. Likewise, we do not state “All blocks on this table are green, hence, had this white block been on the table, it would have been green”. In fact, we could say that people’s use of counterfactual statements is aimed precisely at conveying generic causal information, uncontaminated by specific, transitory observations, about the real world. Observed facts often do reflect strange combinations of rare eventualities (e.g., all blocks being green) that have nothing to do with general traits of influence and behavior. The counterfactual sentence, however, emphasizes the law-like, necessary component of the relation considered. It is for this reason, we speculate, that we find such frequent use of counterfactuals in ordinary discourse.

The importance of equipping machines with the capability to answer counterfactual queries lies precisely in this causal reading. By making a counterfactual query, the user intends to extract the generic, necessary connection between the antecedent and consequent, regardless of the contingent factual information available at that moment.

Because of the tight connection between counterfactuals and causal influences, any algorithm for computing counterfactual queries must rely heavily on causal knowledge of the domain. This leads naturally to the use of probabilistic causal networks, since these networks combine causal and probabilistic knowledge and permit reasoning from causes to effects as well as, conversely, from effects to causes.

To emphasize the causal character of counterfactuals, we will adopt the interpretation in (Pearl 1993b), according to which a counterfactual sentence “If it were A , then B would have been” states that B would prevail if A were forced to be true by some unspecified action that is exogenous to the other relationships considered in the analysis. This action-based interpretation does not permit inferences from the counterfactual antecedent towards events that lie in its past. For example, the action-based interpretation would ratify the counterfactual

If Kennedy were alive today, then the country would have been in a better shape

but not the counterfactual,

If Kennedy were alive today, then Oswald would have been alive as well.

The former is admitted because the causal influence of Kennedy on the country is presumed to remain valid even if Kennedy became alive by an act of God. The second sentence is disallowed because Kennedy being alive is not perceived as

having causal influence on Oswald being alive. The information intended in the second sentence is better expressed in an indicative mood:

If Kennedy was alive today then he could not have been killed in Dallas, hence, Jack Ruby would not have had a reason to kill Oswald and Oswald would have been alive today.

Our interpretation of counterfactual antecedents, which is similar to Lewis' (1979) *Miraculous Analysis*, contrasts with interpretations that require that the counterfactual antecedent be consistent with the world in which the analysis occurs. The set of closest worlds delineated by the action-based interpretation contains all those which coincide with the factual world except on possible consequences of the action taken. The probabilities assigned to these worlds will be determined by the relative likelihood of those consequences as encoded by the causal network.

We will show that causal theories specified in functional form (as in (Pearl & Verma 1991; Druzdzel & Simon 1993; Poole 1993)) are sufficient for evaluating counterfactual queries, whereas the causal information embedded in Bayesian networks is not sufficient for the task. Every Bayes network can be represented by several functional specifications, each yielding different evaluations of a counterfactual. The problem is that, deciding what factual information deserves undoing (by the antecedent of the query) requires a model of temporal persistence, and, as noted in (Pearl 1993c), such a model is not part of static Bayesian networks. Functional specification, however, implicitly contains the temporal persistence information needed.

The next section introduces some useful notation for concisely expressing counterfactual sentences/queries. We then present an example demonstrating the plausibility of the external action interpretation adopted in this paper. We then demonstrate that Bayesian networks are insufficient for uniquely evaluating counterfactual queries whereas the functional model is sufficient. A counterfactual query algorithm is then presented, followed by a re-examination of the earlier example with a quantitative analysis using this algorithm. The final section contains concluding remarks.

17.2 Notation

Let the set of variables describing the world be designated by $X = \{X_1, X_2, \dots, X_n\}$. As part of the complete specification of a counterfactual query, there are real-world observations that make up the background context. These observed values will be represented in the standard form x_1, x_2, \dots, x_n . In addition, we must represent the value of the variables in the counterfactual world. To distinguish between x_i and the value of X_i in the counterfactual world, we will denote the latter with an asterisk;

thus, the value of X_i in the counterfactual world will be represented by x_i^* . We will also need a notation to distinguish between events that might be true in the counterfactual world and those referenced explicitly in the counterfactual antecedent. The latter are interpreted as being forced to the counterfactual value by an external action, which will be denoted by a hat (e.g., \hat{x}).

Thus, a typical counterfactual query will have the form “What is $P(c^* | \hat{a}^*, a, b)$?” to be read as “Given that we have observed $A = a$ and $B = b$ in the real world, if A were \hat{a}^* , then what is the probability that C would have been c^* ?”

17.3 Party Example

To illustrate the external-force interpretations of counterfactuals, consider the following interpersonal behaviors of Ann, Bob, and Carl:

- Ann sometimes goes to parties.
- Bob likes Ann very much but is not into the party scene. Hence, save for rare circumstances, Bob is at the party if and only if Ann is there.
- Carl tries to avoid contact with Ann since they broke up last month, but he really likes parties. Thus, save for rare occasions, Carl is at the party if and only if Ann is not at the party.
- Bob and Carl truly hate each other and almost always scuffle when they meet.

This situation may be represented by the diamond structure in Figure 17.1. The four variables $A, B, C,$ and S have the following domains:

$$\begin{aligned}
 a &\in \left\{ \begin{array}{l} a_0 \equiv \text{Ann is not at the party.} \\ a_1 \equiv \text{Ann is at the party.} \end{array} \right\} \\
 b &\in \left\{ \begin{array}{l} b_0 \equiv \text{Bob is not at the party.} \\ b_1 \equiv \text{Bob is at the party.} \end{array} \right\} \\
 c &\in \left\{ \begin{array}{l} c_0 \equiv \text{Carl is not at the party.} \\ c_1 \equiv \text{Carl is at the party.} \end{array} \right\} \\
 s &\in \left\{ \begin{array}{l} s_0 \equiv \text{No scuffle between Bob and Carl.} \\ s_1 \equiv \text{Scuffle between Bob and Carl.} \end{array} \right\}
 \end{aligned}$$

Now consider the following discussion between two friends (Laura and Scott) who did not go to the party but were called by Bob from his home ($b = b_0$):

- Laura: Ann must not be at the party, or Bob would be there instead of at home.
 Scott: That must mean that Carl is at the party!

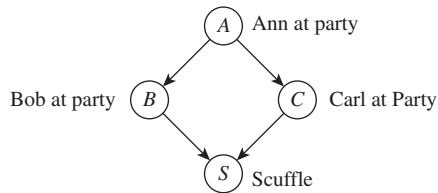


Figure 17.1 Causal structure reflecting the influence that Ann’s attendance has on Bob and Carl’s attendance, and the influence that Bob and Carl’s attendance has on their scuffling.

Laura: If Bob were at the party, then Bob and Carl would surely scuffle.

Scott: No. If Bob was there, then Carl would not be there, because Ann would have been at the party.

Laura: True. But if Bob were at the party even though Ann was not, then Bob and Carl would be scuffling.

Scott: I agree. It’s good that Ann would not have been there to see it.

In the fourth sentence, Scott tries to explain away Laura’s conclusion by claiming that Bob’s presence would be evidence that Ann was at the party which would imply that Carl was not at the party. Scott, though, analyzes Laura’s counterfactual statement as an indicative sentence by imagining that she had observed Bob’s presence at the party; this allows her to use the observation for abductive reasoning. But Laura’s subjunctive (counterfactual) statement should be interpreted as leaving everything in the past as it was (including conclusions obtained from abductive reasoning from real observations) while forcing variables to their counterfactual values. This is the gist of her last statement.

This example demonstrates the plausibility of interpreting the counterfactual statement in terms of an external force causing Bob to be at the party, regardless of all other prior circumstances. The only variables that we would expect to be impacted by the counterfactual assumption would be the descendants of the counterfactual variable; in other words, the counterfactual value of Bob’s attendance does not change the belief in Ann’s attendance from the belief prompted by the real-world observation.

17.4 Probabilistic vs. Functional Specification

In this section we will demonstrate that functionally modeled causal theories (Pearl & Verma 1991) are necessary for uniquely evaluating counterfactual queries, while the conditional probabilities used in the standard specification of Bayesian networks are insufficient for obtaining unique solutions.

Reconsider the party example limited to the two variables A and B , representing Ann and Bob's attendance, respectively. Assume that previous behavior shows $P(b_1 | a_1) = 0.9$ and $P(b_0 | a_0) = 0.9$. We observe that Bob and Ann are absent from the party and we wonder whether Bob would be there if Ann were there $P(b_1^* | \hat{a}_1^*, a_0, b_0)$. The answer depends on the mechanism that accounts for the 10% exception in Bob's behavior. If the reason Bob occasionally misses parties (when Ann goes) is that he is unable to attend (e.g., being sick or having to finish a paper for AAAI), then the answer to our query would be 90%. However, if the only reason for Bob's occasional absence (when Ann goes) is that he becomes angry with Ann (in which case he does exactly the opposite of what she does), then the answer to our query is 100%, because Ann and Bob's current absence from the party proves that Bob is not angry. Thus, we see that the information contained in the conditional probabilities on the observed variables is insufficient for answering counterfactual queries uniquely; some information about the mechanisms responsible for these probabilities is needed as well.

The functional specification, which provides this information, models the influence of A on B by a deterministic function

$$b = F_b(a, \epsilon_b)$$

where ϵ_b stands for all unknown factors that may influence B and the prior probability distribution $P(\epsilon_b)$ quantifies the likelihood of such factors. For example, whether Bob has been grounded by his parents and whether Bob is angry at Ann could make up two possible components of ϵ_b . Given a specific value for ϵ_b , B becomes a deterministic function of A ; hence, each value in ϵ_b 's domain specifies a *response function* that maps each value of A to some value in B 's domain. In general, the domain for ϵ_b could contain many components, but it can always be replaced by an equivalent variable that is minimal, by partitioning the domain into equivalence regions, each corresponding to a single response function (Pearl 1993a). Formally, these equivalence classes can be characterized as a function $r_b : \text{dom}(\epsilon_b) \rightarrow \mathbf{N}$, as follows:

$$r_b(\epsilon_b) = \begin{cases} 0 & \text{if } F_b(a_0, \epsilon_b) = 0 \text{ \& } F_b(a_1, \epsilon_b) = 0 \\ 1 & \text{if } F_b(a_0, \epsilon_b) = 0 \text{ \& } F_b(a_1, \epsilon_b) = 1 \\ 2 & \text{if } F_b(a_0, \epsilon_b) = 1 \text{ \& } F_b(a_1, \epsilon_b) = 0 \\ 3 & \text{if } F_b(a_0, \epsilon_b) = 1 \text{ \& } F_b(a_1, \epsilon_b) = 1 \end{cases}$$

Obviously, r_b can be regarded as a random variable that takes on as many values as there are functions between A and B . We will refer to this domain-minimal variable

as a *response-function variable*. r_b is closely related to the *potential response variables* in Rubin's model of counterfactuals (Rubin 1974), which was introduced to facilitate causal inference in statistical analysis (Balke & Pearl 1993).

For this example, the response-function variable for B has a four-valued domain $r_b \in \{0, 1, 2, 3\}$ with the following functional specification:

$$b = f_b(a, r_b) = h_{b,r_b}(a) \quad (17.1)$$

where

$$h_{b,0}(a) = b_0 \quad (17.2)$$

$$h_{b,1}(a) = \begin{cases} b_0 & \text{if } a = a_0 \\ b_1 & \text{if } a = a_1 \end{cases} \quad (17.3)$$

$$h_{b,2}(a) = \begin{cases} b_1 & \text{if } a = a_0 \\ b_0 & \text{if } a = a_1 \end{cases} \quad (17.4)$$

$$h_{b,3}(a) = b_1 \quad (17.5)$$

specify the mappings of the individual response functions. The prior probability on these response functions $P(r_b)$ in conjunction with $f_b(a, r_b)$ fully parameterizes the model.

Given $P(r_b)$, we can uniquely evaluate the counterfactual query “What is $P(b_1^* | \widehat{a}_1^*, a_0, b_0)$?” (i.e., “Given $A = a_0$ and $B = b_0$, if A were a_1 , then what is the probability that B would have been b_1 ?”). The action-based interpretation of counterfactual antecedents implies that the disturbance ϵ_b , and hence the response-function r_b , is unaffected by the actions that force the counterfactual values¹; therefore, what we learn about the response-function from the observed evidence is applicable to the evaluation of belief in the counterfactual consequent. If we observe (a_0, b_0) , then we are certain that $r_b \in \{0, 1\}$, an event having prior probability $P(r_b = 0) + P(r_b = 1)$. Hence, this evidence leads to an updated posterior probability for r_b (let $\vec{P}(r_b) = \langle P(r_b=0), P(r_b=1), P(r_b=2), P(r_b=3) \rangle$)

$$\begin{aligned} \vec{P}'(r_b) &= \vec{P}(r_b | a_0, b_0) = \\ &\left\langle \frac{P(r_b=0)}{P(r_b=0) + P(r_b=1)}, \frac{P(r_b=1)}{P(r_b=0) + P(r_b=1)}, 0, 0 \right\rangle. \end{aligned}$$

According to Equations (17.1)-(17.5), if A were forced to a_1 , then B would have been b_1 if and only if $r_b \in \{1, 3\}$, which has probability $P'(r_b=1) + P'(r_b=3) =$

1. An observation by D. Heckerman (personal communication).

$P'(r_b=1)$. This is exactly the solution to the counterfactual query,

$$P(b_1^* | \hat{a}_1^*, a_0, b_0) = P'(r_b=1) = \frac{P(r_b=1)}{P(r_b=0) + P(r_b=1)}.$$

This analysis is consistent with the *prior propensity account* of (Skyrms 1980).

What if we are provided only with the conditional probability ($P(b|a)$) instead of a functional model ($f_b(a, r_b)$ and $P(r_b)$)? These two specifications are related by:

$$\begin{aligned} P(b_1 | a_0) &= P(r_b=2) + P(r_b=3) \\ P(b_1 | a_1) &= P(r_b=1) + P(r_b=3). \end{aligned}$$

which show that $P(r_b)$ is not, in general, uniquely determined by the conditional distribution $P(b|a)$.

Hence, given a counterfactual query, a functional model always leads to a unique solution, while a Bayesian network seldom leads to a unique solution, depending on whether the conditional distributions of the Bayesian network sufficiently constrain the prior distributions of the response-function variables in the corresponding functional model.

In practice, specifying a functional model is not as daunting as one might think from the example above. In fact, it could be argued that the subjective judgments needed for specifying Bayesian networks (i.e., judgments about conditional probabilities) are generated mentally on the basis of a stored model of functional relationships. For example, in the noisy-OR mechanism, which is often used to model causal interactions, the conditional probabilities are derivatives of a functional model involving AND/OR gates, corrupted by independent binary disturbances. This model is used, in fact, to *simplify* the specification of conditional probabilities in Bayesian networks (Pearl 1988).

17.5 Evaluating Counterfactual Queries

From the last section, we see that the algorithm for evaluating counterfactual queries should consist of: (1) compute the posterior probabilities for the disturbance variables, given the observed evidence; (2) remove the observed evidence and enforce the value for the counterfactual antecedent; finally, (3) evaluate the probability of the counterfactual consequent, given the conditions set in the first two steps.

An important point to remember is that it is not enough to compute the posterior distribution of each disturbance variable (ϵ) separately and treat those variables as independent quantities. Although the disturbance variables are initially independent, the evidence observed tends to create dependencies among the parents of the observed variables, and these dependencies need to be represented

in the posterior distribution. An efficient way to maintain these dependencies is through the structure of the causal network itself.

Thus, we will represent the variables in the counterfactual world as distinct from the corresponding variables in the real world, by using a separate network for each world. Evidence can then be instantiated on the real-world network, and the solution to the counterfactual query can be determined as the probability of the counterfactual consequent, as computed in the counterfactual network where the counterfactual antecedent is enforced. But, the reader may ask, and this is key, how are the networks for the real and counterfactual worlds linked? Because any exogenous variable, ϵ_a , is not influenced by forcing the value of any endogenous variables in the model, the value of that disturbance will be identical in both the real and counterfactual worlds; therefore, a single variable can represent the disturbance in both worlds. ϵ_a thus becomes a common causal influence of the variables representing A in the real and counterfactual networks, respectively, which allows evidence in the real-world network to propagate to the counterfactual network.

Assume that we are given a *causal theory* $T = \langle D, \Theta_D \rangle$ as defined in (Pearl & Verma 1991). D is a directed acyclic graph (DAG) that specifies the structure of causal influences over a set of variables $X = \{X_1, X_2, \dots, X_n\}$. Θ_D specifies a functional mapping $x_i = f_i(\text{pa}(x_i), \epsilon_i)$ ($\text{pa}(x_i)$ represents the value of X_i 's parents) and a prior probability distribution $P(\epsilon_i)$ for each disturbance ϵ_i (we assume that ϵ_i 's domain is discrete; if not, we can always transform it to a discrete domain such as a response-function variable). A counterfactual query “What is $P(c^* | \hat{a}^*, \text{obs})$?” is then posed, where c^* specifies counterfactual values for a set of variables $C \subset X$, \hat{a}^* specifies forced values for the set of variables in the counterfactual antecedent, and obs specifies observed evidence. The solution can be evaluated by the following algorithm:

1. From the known causal theory T create a Bayesian network $\langle G, \mathcal{P} \rangle$ that explicitly models the disturbances as variables and distinguishes the real world variables from their counterparts in the counterfactual world. G is a DAG defined over the set of variables $V = X \cup X^* \cup \epsilon$, where $X = \{X_1, X_2, \dots, X_n\}$ is the original set of variables modeled by T , $X^* = \{X_1^*, X_2^*, \dots, X_n^*\}$ is their counterfactual world representation, and $\epsilon = \{\epsilon_1, \epsilon_2, \dots, \epsilon_n\}$ represents the set of disturbance variables that summarize the common external causal influences acting on the members of X and X^* . \mathcal{P} is the set of conditional probability distributions $P(V_i | \text{pa}(V_i))$ that parameterizes the causal structure G .

If $X_j \in \text{pa}(X_i)$ in D , then $X_j \in \text{pa}(X_i)$ and $X_j^* \in \text{pa}(X_i^*)$ in G ($\text{pa}(X_i)$ is the set of X_i 's parents). In addition, $\epsilon_i \in \text{pa}(X_i)$ and $\epsilon_i \in \text{pa}(X_i^*)$ in G . The conditional

probability distributions for the Bayesian network are generated from the causal theory:

$$P(x_i | \text{pa}_X(x_i), \epsilon_i) = \begin{cases} 1 & \text{if } x_i = f_i(\text{pa}_X(x_i), \epsilon_i) \\ 0 & \text{otherwise} \end{cases}$$

where $\text{pa}_X(x_i)$ is the set of values of the variables in $X \cap \text{pa}(x_i)$.

$$P(x_i^* | \text{pa}_{X^*}(x_i^*), \epsilon_i) = P(x_i | \text{pa}_X(x_i), \epsilon_i)$$

whenever $x_i = x_i^*$ and $\text{pa}_{X^*}(x_i^*) = \text{pa}_X(x_i)$. $P(\epsilon_i)$ is the same as specified by the functional causal theory T .

2. Observed evidence. The observed evidence *obs* is instantiated on the real world variables X corresponding to *obs*.
3. Counterfactual antecedent. For every forced value in the counterfactual antecedent specification $\hat{x}_i^* \in \hat{a}^*$, apply the action-based semantics of $\text{set}(X_i^* = \hat{x}_i^*)$ (see (Pearl 1993a; Spirtes, Glymour, & Scheines 1993)), which amounts to severing all the causal edges from $\text{pa}(X_i^*)$ to X_i^* for all $x_i^* \in \hat{a}^*$ and instantiating X_i^* to the value specified in \hat{a}^* .
4. Belief propagation. After instantiating the observations and actions in the network, evaluate the belief in c^* using the standard belief update methods for Bayesian networks (Pearl 1988). The result is the solution to the counterfactual query.

In the last section, we noted that the conditional distribution $P(x_k | \text{pa}(X_k))$ for each variable $X_k \in X$ constrains, but does not uniquely determine, the prior distribution $P(\epsilon_k)$ of each disturbance variable. Although the composition of the external causal influences are often not precisely known, a subjective distribution over response functions may be assessable. If a reasonable distribution can be selected for each relevant disturbance variable, the implementation of the above algorithm is straightforward and the solution is unique; otherwise, bounds on the solution can be obtained using convex optimization techniques. (Balke & Pearl 1993) demonstrates this optimization task in deriving bounds on causal effects from partially controlled experiments.

A network generated by the above algorithm may often be simplified. If a variable X_j^* in the counterfactual world is not a causal descendant of any of the variables mentioned in the counterfactual antecedent \hat{a}^* , then X_j and X_j^* will always have identical distributions, because the causal influences that functionally determine X_j and X_j^* are identical. X_j and X_j^* may therefore be treated as the same

variable. In this case, the conditional distribution $P(x_j | \text{pa}(x_j))$ is sufficient, and the disturbance variable ϵ_j and its prior distribution need not be specified.

17.6 Party Again

Let us revisit the party example. Assuming we have observed that Bob is not at the party ($b = b_0$), we want to know whether Bob and Carl would have scuffled if Bob were at the party (i.e., “What is $P(s_1^* | \hat{b}_1^*, b_0)$?”).

Suppose that we are supplied with the following causal theory for the model in Figure 17.1:

$$\begin{aligned} a &= f_a(r_a) &= h_{a,r_a}() \\ b &= f_b(a, r_b) &= h_{b,r_b}(a) \\ c &= f_c(a, r_c) &= h_{c,r_c}(a) \\ s &= f_s(b, c, r_s) &= h_{s,r_s}(b, c) \end{aligned}$$

where

$$\begin{aligned} P(r_a) &= \begin{cases} 0.40 & \text{if } r_a = 0 \\ 0.60 & \text{if } r_a = 1 \end{cases} \\ P(r_b) &= \begin{cases} 0.07 & \text{if } r_b = 0 \\ 0.90 & \text{if } r_b = 1 \\ 0.03 & \text{if } r_b = 2 \\ 0 & \text{if } r_b = 3 \end{cases} \\ P(r_c) &= \begin{cases} 0.05 & \text{if } r_c = 0 \\ 0 & \text{if } r_c = 1 \\ 0.85 & \text{if } r_c = 2 \\ 0.10 & \text{if } r_c = 3 \end{cases} \\ P(r_s) &= \begin{cases} 0.05 & \text{if } r_s = 0 \\ 0.90 & \text{if } r_s = 8 \\ 0.05 & \text{if } r_s = 9 \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

and

$$\begin{aligned} h_{a,0}() &= a_0 \\ h_{a,1}() &= a_1 \end{aligned}$$

$$\begin{aligned}
 h_{s,0}(b, c) &= s_0 \\
 h_{s,8}(b, c) &= \begin{cases} s_0 & \text{if } (b, c) \neq (b_1, c_1) \\ s_1 & \text{if } (b, c) = (b_1, c_1) \end{cases} \\
 h_{s,9}(b, c) &= \begin{cases} s_0 & \text{if } (b, c) \in \{(b_1, c_0), (b_0, c_1)\} \\ s_1 & \text{if } (b, c) \in \{(b_0, c_0), (b_1, c_1)\} \end{cases}
 \end{aligned}$$

The response functions for B and C (h_{b,r_b} and h_{c,r_c}) both take the same form as that given in Equation (17.5).

These numbers reflect the authors' understanding of the characters involved. For example, the choice for $P(r_b)$ represents our belief that Bob usually is at the party if and only if Ann is there ($r_b = 1$). However, we believe that Bob is sometimes ($\sim 7\%$ of the time) unable to go to the party (e.g., sick or grounded by his parents); this exception is represented by $r_b = 0$. In addition, Bob would sometimes ($\sim 3\%$ of the time) go to the party if and only if Ann is not there (e.g., Bob is in a spiteful mood); this exception is represented by $r_b = 2$. Finally, $P(r_s)$ represents our understanding that there is a slight chance (5%) that Bob and Carl would not scuffle regardless of attendance ($r_s = 0$), and the same chance ($P(r_s = 9) = 5\%$) that a scuffle would take place either outside or inside the party (but not if only one of them shows up).

Figure 17.2 shows the Bayesian network generated from step 1 of the algorithm. After instantiating the real world observations (b_0) and the actions (\hat{b}_1^*) specified by the counterfactual antecedent in accordance with steps 2 and 3, the network takes on the configuration shown in Figure 17.3.

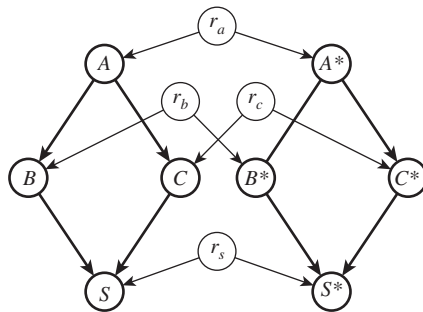


Figure 17.2 Bayesian model for evaluating counterfactual queries in the party example. The variables marked with * make up the counterfactual world, while those without *, the factual world. The r variables index the response functions.

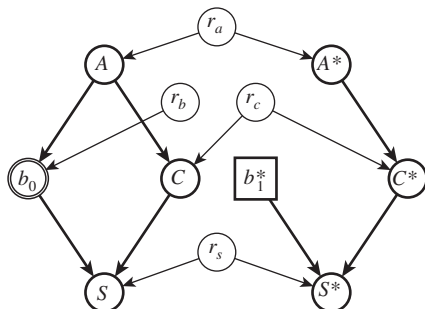


Figure 17.3 To evaluate the query $P(s_1^* | \hat{b}_1^*, b_0)$, the network of Figure 17.2 is instantiated with observation b_0 and action \hat{b}_1^* (links pointing to b_1^* are severed).

If we propagate the evidence through this Bayesian network, we will arrive at the solution

$$P(s_1^* | \hat{b}_1^*, b_0) = 0.79.$$

which is consistent with Laura’s assertion that Bob and Carl would have scuffled if Bob were at the party, given that Bob actually was not at the party. Compare this to the solution to the indicative query that Scott was thinking of:

$$P(s_1 | b_1) = 0.11.$$

that is, if we had observed that Bob was at the party, then Bob and Carl would probably not have scuffled. This emphasizes the difference between counterfactual and indicative queries and their solutions.

17.7 Special Case: Linear-Normal Models

Assume that knowledge is specified by the structural equation model

$$\vec{x} = B\vec{x} + \vec{\epsilon}$$

where B is a triangular matrix (corresponding to a causal model that is a DAG), and we are given the mean $\vec{\mu}_\epsilon$ and covariance $\Sigma_{\epsilon, \epsilon}$ of the disturbances $\vec{\epsilon}$ (assumed to be normal). The mean and covariance of the observable variables \vec{x} are then given by:

$$\vec{\mu}_x = S\vec{\mu}_\epsilon \tag{17.6}$$

$$\Sigma_{x,x} = S\Sigma_{\epsilon,\epsilon}S^t \tag{17.7}$$

where $S = (I - B)^{-1}$.

Under such a model, there are well-known formulas (Whittaker 1990, p. 163) for evaluating the conditional mean and covariance of \vec{x} under some observations \vec{o} :

$$\vec{\mu}_{x|o} = \vec{\mu}_x + \Sigma_{x,o}\Sigma_{o,o}^{-1}(\vec{o} - \vec{\mu}_o) \quad (17.8)$$

$$\Sigma_{x,x|o} = \Sigma_{x,x} - \Sigma_{x,o}\Sigma_{o,o}^{-1}\Sigma_{o,y} \quad (17.9)$$

where, for every pair of sub-vectors, \vec{z} and \vec{w} , of \vec{x} , $\Sigma_{z,w}$ is the sub-matrix of $\Sigma_{x,x}$ with entries corresponding to the components of \vec{z} and \vec{w} . Singularities of Σ terms are handled by appropriate means.

Similar formulas apply for the mean and covariance of \vec{x} under an action \vec{a} . B is replaced by the action-pruned matrix $\widehat{B} = [\widehat{b}_{ij}]$ defined by:

$$\widehat{b}_{ij} = \begin{cases} 0 & \text{if } X_i \in \vec{a} \\ b_{ij} & \text{otherwise} \end{cases} \quad (17.10)$$

The mean and covariance of \vec{x} under \widehat{B} is evaluated using Equations (17.6) and (17.7), where B is replaced by \widehat{B} :

$$\vec{\mu}_x = \widehat{S}\vec{\mu}_\epsilon \quad (17.11)$$

$$\widehat{\Sigma}_{x,x} = \widehat{S}\Sigma_{\epsilon,\epsilon}\widehat{S}^t \quad (17.12)$$

where $\widehat{S} = (I - \widehat{B})^{-1}$. We can then evaluate the distribution of \vec{x} under the action \vec{a} by conditioning on the value of the action \vec{a} according to Equations (17.8) and (17.9):

$$\vec{\mu}_{x|\vec{a}} \triangleq \vec{\mu}_{x|a} = \vec{\mu}_x + \widehat{\Sigma}_{x,a}\widehat{\Sigma}_{a,a}^{-1}(\vec{a} - \vec{\mu}_a) \quad (17.13)$$

$$\Sigma_{x,x|\vec{a}} \triangleq \widehat{\Sigma}_{x,x|a} = \widehat{\Sigma}_{x,x} - \widehat{\Sigma}_{x,a}\widehat{\Sigma}_{a,a}^{-1}\widehat{\Sigma}_{a,x} \quad (17.14)$$

To evaluate the counterfactual query $P(x^* | \widehat{a}^*o)$ we first update the prior distribution of the disturbances by the observations \vec{o} :

$$\begin{aligned} \vec{\mu}_\epsilon^o &\triangleq \vec{\mu}_{\epsilon|o} = \vec{\mu}_\epsilon + \Sigma_{\epsilon,\epsilon}S^t(S\Sigma_{\epsilon,\epsilon}S^t)^{-1}(\vec{o} - \vec{\mu}_o) \\ \Sigma_{\epsilon,\epsilon}^o &\triangleq \Sigma_{\epsilon,\epsilon|o} = \Sigma_{\epsilon,\epsilon} - \Sigma_{\epsilon,\epsilon}S^t(S\Sigma_{\epsilon,\epsilon}S^t)^{-1}S\Sigma_{\epsilon,\epsilon} \end{aligned}$$

We then evaluate the means $\vec{\mu}_{x^*|\widehat{a}^*o}$ and variances $\Sigma_{x^*,x^*|\widehat{a}^*o}$ of the variables in the counterfactual world (x^*) under the action \widehat{a}^* using Equations (17.13) and (17.14), with Σ^o and μ^o replacing Σ and μ .

$$\begin{aligned} \vec{\mu}_{x^*|\widehat{a}^*o} &\triangleq \vec{\mu}_{x^*|\widehat{a}}^o = \vec{\mu}_x^o + \widehat{\Sigma}_{x^*,a}^o(\widehat{\Sigma}_{a,a}^o)^{-1}(\vec{a} - \vec{\mu}_a^o) \\ \Sigma_{x^*,x^*|\widehat{a}^*o} &\triangleq \Sigma_{x^*,x^*|\widehat{a}}^o = \widehat{\Sigma}_{x^*,x^*}^o - \widehat{\Sigma}_{x^*,a}^o(\widehat{\Sigma}_{a,a}^o)^{-1}\widehat{\Sigma}_{a,x^*}^o \end{aligned}$$

where, from Equations (17.11) and (17.12), $\vec{\mu}_x^o = \widehat{S}\vec{\mu}_\epsilon^o$ and $\widehat{\Sigma}_{x^*,x^*}^o = \widehat{S}\Sigma_{\epsilon,\epsilon}^o\widehat{S}^t$.

It is clear that this procedure can be applied to non-triangular matrices, as long as S is non-singular. In fact, the response-function formulation opens the way to incorporate feedback loops within the Bayesian network framework.

17.8 Conclusion

The evaluation of counterfactual queries is applicable to many tasks. For example, determining liability of actions (e.g., “If you had not pushed the table, the glass would not have broken; therefore, you are liable”). In diagnostic tasks, counterfactual queries can be used to determine which tests to perform in order to increase the probability that faulty components are identified. In planning, counterfactuals can be used for goal regression or for determining which actions, if performed, could have avoided an observed, unexpected failure. Thus, counterfactual reasoning is an essential component in plan repairing, plan compilation and explanation-based learning.

In this paper we have presented formal notation, semantics, representation scheme, and inference algorithms that facilitate the probabilistic evaluation of counterfactual queries. World knowledge is represented in the language of modified causal networks, whose root nodes are unobserved, and correspond to possible functional mechanisms operating among families of observables. The prior probabilities of these root nodes are updated by the factual information transmitted with the query, and remain fixed thereafter. The antecedent of the query is interpreted as a proposition that is established by an external action, thus pruning the corresponding links from the network and facilitating standard Bayesian-network computation to determine the probability of the consequent.

At this time the algorithm has not been implemented but, given a subjective prior distribution over the response variables, there are no new computational tasks introduced by this formalism, and the inference process follows the standard techniques for computing beliefs in Bayesian networks (Pearl 1988). If prior distributions over the relevant response-function variables cannot be assessed, we have developed methods of using the standard conditional-probability specification of Bayesian networks to compute upper and lower bounds on counterfactual probabilities (Balke & Pearl 1994).

The semantics and methodology introduced in this paper can be adopted to nonprobabilistic formalisms as well, as long as they support two essential components: abduction (to abduce plausible functional mechanisms from the factual observations) and causal projection (to infer the consequences of the action-like antecedent). We should note, though, that the license to keep the response-function variables constant stems from a unique feature of counterfactual queries, where the factual observations are presumed to occur not

earlier than the counterfactual action. In general, when an observation takes place before an action, constancy of response functions would be justified if the environment remains relatively static between the observation and the action (e.g., if the disturbance terms ϵ_i represent unknown pre-action conditions). However, in a dynamic environment subject to stochastic shocks a full temporal analysis using temporally-indexed networks may be warranted or, alternatively, a canonical model of persistence should be invoked (Pearl 1993c).

Acknowledgments

The research was partially supported by Air Force grant #AFOSR, 90 0136, NSF grant #IRI-9200918, Northrop Micro grant #92-123, and Rockwell Micro grant #92-122. Alexander Balke was supported by the Fannie and John Hertz Foundation. This work benefitted from discussions with David Heckerman.

References

- Balke, A., and Pearl, J. 1993. Nonparametric bounds on causal effects from partial compliance data. Technical Report R-199, UCLA Cognitive Systems Lab.
- Balke, A., and Pearl, J. 1994. Bounds on probabilistically evaluated counterfactual queries. Technical report, UCLA Cognitive Systems Lab.
- Boutilier, C. 1992. A logic for revision and subjunctive queries. In *Proceedings Tenth National Conference on Artificial Intelligence*, 609–15. Menlo Park, CA: AAAI Press.
- Druzdzel, M. J., and Simon, H. A. 1993. Causality in bayesian belief networks. In *Proceedings of the 9th Annual Conference on Uncertainty in Artificial Intelligence (UAI-93)*, 3–11.
- Ginsberg, M. L. 1986. Counterfactuals. *Artificial Intelligence* 30:35–79.
- Goodman, N. 1983. *Fact, Fiction, and Forecast*. Cambridge, MA: Harvard University Press, 4th edition.
- Harper, W. L.; Stalnaker, R.; and Pearce, G., eds. 1981. *Ifs: Conditionals, Belief, Decision, Chance, and Time*. Boston, MA: D. Reidel.
- Jackson, P. 1989. On the semantics of counterfactuals. In *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence*, 1382–7 vol. 2. Palo Alto, CA: Morgan Kaufmann.
- Lewis, D. 1976. Probability of conditionals and conditional probabilities. *The Philosophical Review* 85:297–315.
- Lewis, D. 1979. Counterfactual dependence and time's arrow. *Noûs* 455–476.
- Meyer, J.-J., and van der Hoek, W. 1993. Counterfactual reasoning by (means of) defaults. *Annals of Mathematics and Artificial Intelligence* 9:345–360.
- Nute, D. 1980. *Topics in Conditional Logic*. Boston: D. Reidel.

- Pearl, J., and Verma, T. 1991. A theory of inferred causation. In *Principles of Knowledge Representation and Reasoning: Proceedings of the Second International Conference*, 441–452. San Mateo, CA: Morgan Kaufmann.
- Pearl, J. 1988. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Mateo, CA: Morgan Kaufman.
- Pearl, J. 1993a. Aspects of graphical models connected with causality. Technical Report R-195-LLL, UCLA Cognitive Systems Lab. Short version: *Statistical Science* 8(3):266–269.
- Pearl, J. 1993b. From Adams' conditionals to default expressions, causal conditionals, and counterfactuals. Technical Report R-193, UCLA Cognitive Systems Lab. To appear in *Festschrift for Ernest Adams*, Cambridge University Press, 1994.
- Pearl, J. 1993c. From conditional oughts to qualitative decision theory. In *Uncertainty in Artificial Intelligence: Proceedings of the Ninth Conference*, 12–20. Morgan Kaufmann.
- Pereira, L. M.; Aparicio, J. N.; and Alferes, J. J. 1991. Counterfactual reasoning based on revising assumptions. In *Logic Programming: Proceedings of the 1991 International Symposium*, 566–577. Cambridge, MA: MIT Press.
- Poole, D. 1993. Probabilistic Horn abduction and Bayesian networks. *Artificial Intelligence* 64(1):81–130.
- Rubin, D. B. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* 66(5):688–701.
- Skyrms, B. 1980. The prior propensity account of subjunctive conditionals. In Harper, W.; Stalnaker, R.; and Pearce, G., eds., *Ifs*. D. Reidel. 259–265.
- Spirtes, P.; Glymour, C.; and Scheines, R. 1993. *Causation, Prediction, and Search*. New York: Springer.
- Whittaker, J. 1990. *Graphical Models in Applied Multivariate Statistics*. New York: John Wiley & Sons.

Causal Diagrams for Empirical Research (With Discussions)

Judea Pearl

Summary

The primary aim of this paper is to show how graphical models can be used as a mathematical language for integrating statistical and subject-matter information. In particular, the paper develops a principled, nonparametric framework for causal inference, in which diagrams are queried to determine if the assumptions available are sufficient for identifying causal effects from nonexperimental data. If so the diagrams can be queried to produce mathematical expressions for causal effects in terms of observed distributions; otherwise, the diagrams can be queried to suggest additional observations or auxiliary experiments from which the desired inferences can be obtained.

Some key words

Causal inference; Graph model; Structural equations; Treatment effect.

18.1 Introduction

The tools introduced in this paper are aimed at helping researchers communicate qualitative assumptions about cause-effect relationships, elucidate the ramifications of such assumptions, and derive causal inferences from a combination of assumptions, experiments, and data.

Originally published in *Biometrika* (1995), 82, 4, pp. 669–710

Printed in Great Britain

Original DOI: [10.1093/biomet/82.4.669](https://doi.org/10.1093/biomet/82.4.669)

Republished with permission from *Biometrika*.

The basic philosophy of the proposed method can best be illustrated through the classical example due to Cochran (Wainer, 1989). Consider an experiment in which soil fumigants, X , are used to increase oat crop yields, Y , by controlling the eelworm population, Z , but may also have direct effects, both beneficial and adverse, on yields beside the control of eelworms. We wish to assess the total effect of the fumigants on yields when this study is complicated by several factors. First, controlled randomised experiments are infeasible: farmers insist on deciding for themselves which plots are to be fumigated. Secondly, farmers' choice of treatment depends on last year's eelworm population, Z_0 , an unknown quantity strongly correlated with this year's population. Thus we have a classical case of confounding bias, which interferes with the assessment of treatment effects, regardless of sample size. Fortunately, through laboratory analysis of soil samples, we can determine the eelworm populations before and after the treatment and, furthermore, because the fumigants are known to be active for a short period only, we can safely assume that they do not affect the growth of eelworms surviving the treatment. Instead, eelworm growth depends on the population of birds and other predators, which is correlated, in turn, with last year's eelworm population and hence with the treatment itself.

The method proposed in this paper permits the investigator to translate complex considerations of this sort into a formal language, thus facilitating the following tasks.

- (i) Explicate the assumptions underlying the model.
- (ii) Decide whether the assumptions are sufficient for obtaining consistent estimates of the target quantity: the total effect of the fumigants on yields.
- (iii) If the answer to (ii) is affirmative, provide a closed-form expression for the target quantity, in terms of distributions of observed quantities.
- (iv) If the answer to (ii) is negative, suggest a set of observations and experiments which, if performed, would render a consistent estimate feasible.

The first step in this analysis is to construct a causal diagram such as the one given in Figure 18.1, which represents the investigator's understanding of the major causal influences among measurable quantities in the domain. The quantities Z_1 , Z_2 and Z_3 denote, respectively, the eelworm population, both size and type, before treatment, after treatment, and at the end of the season. Quantity Z_0 represents last year's eelworm population; because it is an unknown quantity, it is represented by a hollow circle, as is B , the population of birds and other predators. Links in the diagram are of two kinds: those that connect unmeasured quantities are designated by dashed arrows, those connecting measured quantities by solid arrows.

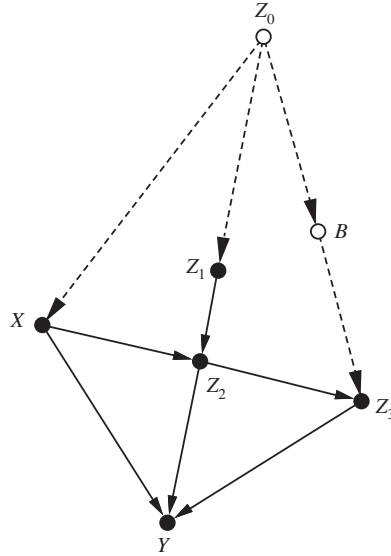


Figure 18.1 A causal diagram representing the effect of fumigants, X , on yields, Y .

The substantive assumptions embodied in the diagram are negative causal assertions, which are conveyed through the links missing from the diagram. For example, the missing arrow between Z_1 and Y signifies the investigator’s understanding that pre-treatment eelworms cannot affect oat plants directly; their entire influence on oat yields is mediated by post-treatment conditions, namely Z_2 and Z_3 . The purpose of the paper is not to validate or repudiate such domain-specific assumptions but, rather, to test whether a given set of assumptions is sufficient for quantifying causal effects from nonexperimental data, for example, estimating the total effect of fumigants on yields.

The proposed method allows an investigator to inspect the diagram of Figure 18.1 and conclude immediately the following.

- (a) The total effect of X on Y can be estimated consistently from the observed distribution of X, Z_1, Z_2, Z_3 and Y .
- (b) The total effect of X on Y , assuming discrete variables throughout, is given by the formula

$$\text{pr}(y | \check{x}) = \sum_{z_1} \sum_{z_2} \sum_{z_3} \text{pr}(y | z_2, z_3, x) \text{pr}(z_2 | z_1, x) \sum_{x'} \text{pr}(z_3 | z_1, z_2, x') \text{pr}(z_1, x'), \tag{18.1}$$

where the symbol \check{x} , read ‘ x check’, denotes that the treatment is set to level $X = x$ by external intervention.

- (c) Consistent estimation of the total effect of X on Y would not be feasible if Y were confounded with Z_3 ; however, confounding Z_2 and Y will not invalidate the formula for $\text{pr}(y|\check{x})$.

These conclusions can be obtained either by analysing the graphical properties of the diagram, or by performing a sequence of symbolic derivations, governed by the diagram, which gives rise to causal effect formulae such as (18.1).

The formal semantics of the causal diagrams used in this paper will be defined in § 18.2, following a review of directed acyclic graphs as a language for communicating conditional independence assumptions. Section 18.2.2 introduces a causal interpretation of directed graphs based on nonparametric structural equations and demonstrates their use in predicting the effect of interventions. Section 18.3 demonstrates the use of causal diagrams to control confounding bias in observational studies. We establish two graphical conditions ensuring that causal effects can be estimated consistently from nonexperimental data. The first condition, named the back-door criterion, is equivalent to the ignorability condition of Rosenbaum & Rubin (1983). The second condition, named the front-door criterion, involves covariates that are affected by the treatment, and thus introduces new opportunities for causal inference. In § 18.4, we introduce a symbolic calculus that permits the stepwise derivation of causal effect formulae of the type shown in (18.1). Using this calculus, § 18.5 characterises the class of graphs that permit the quantification of causal effects from nonexperimental data, or from surrogate experimental designs.

18.2 Graphical Models and the Manipulative Account of Causation

18.2.1 Graphs and Conditional Independence

The usefulness of directed acyclic graphs as economical schemes for representing conditional independence assumptions is well evidenced in the literature (Pearl, 1988; Whittaker, 1990). It stems from the existence of graphical methods for identifying the conditional independence relationships implied by recursive product decompositions

$$\text{pr}(x_1, \dots, x_n) = \prod_i \text{pr}(x_i | pa_i), \quad (18.2)$$

where pa_i stands for the realisation of some subset of the variables that precede X_i in the order (X_1, X_2, \dots, X_n) . If we construct a directed acyclic graph in which the variables corresponding to pa_i are represented as the parents of X_i , also called adjacent predecessors or direct influences of X_i , then the independencies implied by the decomposition (18.2) can be read off the graph using the following test.

Definition 18.1 *d*-separation

Let X , Y and Z be three disjoint subsets of nodes in a directed acyclic graph G , and let p be any path between a node in X and a node in Y , where by ‘path’ we mean any succession of arcs, regardless of their directions. Then Z is said to block p if there is a node w on p satisfying one of the following two conditions: (i) w has converging arrows along p , and neither w nor any of its descendants are in Z , or, (ii) w does not have converging arrows along p , and w is in Z . Further, Z is said to *d*-separate X from Y , in G , written $(X \perp\!\!\!\perp Y | Z)_G$, if and only if Z blocks every path from a node in X to a node in Y .

It can be shown that there is a one-to-one correspondence between the set of conditional independencies $X \perp\!\!\!\perp Y | Z$ (Dawid, 1979) implied by the recursive decomposition (18.2), and the set of triples (X, Z, Y) that satisfy the *d*-separation criterion in G (Geiger, Verma & Pearl, 1990).

An alternative test for *d*-separation has been given by Lauritzen et al. (1990). To test for $(X \perp\!\!\!\perp Y | Z)_G$, delete from G all nodes except those in $X \cup Y \cup Z$ and their ancestors, connect by an edge every pair of nodes that share a common child, and remove all arrows from the arcs. Then $(X \perp\!\!\!\perp Y | Z)_G$ holds if and only if Z is a cut-set of the resulting undirected graph, separating nodes of X from those of Y . Additional properties of directed acyclic graphs and their applications to evidential reasoning in expert systems are discussed by Pearl (1988), Lauritzen & Spiegelhalter (1988), Spiegelhalter et al. (1993) and Pearl (1993a).

18.2.2 Graphs as Models of Interventions

The use of directed acyclic graphs as carriers of independence assumptions has also been instrumental in predicting the effect of interventions when these graphs are given a causal interpretation (Spirtes, Glymour & Scheines, 1993, p. 78; Pearl, 1993b). Pearl (1993b), for example, treated interventions as variables in an augmented probability space, and their effects were obtained by ordinary conditioning.

In this paper we pursue a different, though equivalent, causal interpretation of directed graphs, based on nonparametric structural equations, which owes its roots to early works in econometrics (Frisch, 1938; Haavelmo, 1943; Simon, 1953). In this account, assertions about causal influences, such as those specified by the links in Figure 18.1, stand for autonomous physical mechanisms among the corresponding quantities, and these mechanisms are represented as functional relationships perturbed by random disturbances. In other words, each child-parent family in a directed graph G represents a deterministic function

$$X_i = f_i(p a_i, \varepsilon_i) \quad (i = 1, \dots, n), \quad (18.3)$$

where pa_i denote the parents of variable X_i in G , and ε_i ($1 \leq i \leq n$) are mutually independent, arbitrarily distributed random disturbances (Pearl & Verma, 1991). These disturbance terms represent exogenous factors that the investigator chooses not to include in the analysis. If any of these factors is judged to be influencing two or more variables, thus violating the independence assumption, then that factor must enter the analysis as an unmeasured, or latent, variable, to be represented in the graph by a hollow node, such as Z_0 or B in Figure 18.1. For example, the causal assumptions conveyed by the model in Figure 18.1 correspond to the following set of equations:

$$\begin{aligned} Z_0 &= f_0(\varepsilon_0), & Z_2 &= f_2(X, Z_1, \varepsilon_2), & B &= f_B(Z_0, \varepsilon_B), & Z_3 &= f_3(B, Z_2, \varepsilon_3), \\ Z_1 &= f_1(Z_0, \varepsilon_1), & Y &= f_Y(X, Z_2, Z_3, \varepsilon_Y), & X &= f_X(Z_0, \varepsilon_X). \end{aligned} \quad (18.4)$$

The equational model (18.3) is the nonparametric analogue of a structural equations model (Wright, 1921; Goldberger, 1972), with one exception: the functional form of the equations, as well as the distribution of the disturbance terms, will remain unspecified. The equality signs in such equations convey the asymmetrical counterfactual relation ‘is determined by’, forming a clear correspondence between causal diagrams and Rubin’s model of potential outcome (Rubin, 1974; Holland, 1988; Pratt & Schlaifer, 1988; Rubin, 1990). For example, the equation for Y states that, regardless of what we currently observe about Y , and regardless of any changes that might occur in other equations, if $(X, Z_2, Z_3, \varepsilon_Y)$ were to assume the values $(x, z_2, z_3, \varepsilon_Y)$, respectively, Y would take on the value dictated by the function f_Y . Thus, the corresponding potential response variable in Rubin’s model $Y_{(x)}$, the value that Y would take if X were x , becomes a deterministic function of Z_2, Z_3 and ε_Y , whose distribution is thus determined by those of Z_2, Z_3 and ε_Y . The relation between graphical and counterfactual models is further analysed by Pearl (1994a).

Characterising each child-parent relationship as a deterministic function, instead of by the usual conditional probability $\text{pr}(x_i | p_i)$, imposes equivalent independence constraints on the resulting distributions, and leads to the same recursive decomposition (18.2) that characterises directed acyclic graph models. This occurs because each ε_i is independent of all nondescendants of X_i . However, the functional characterisation $X_i = f_i(pa_i, \varepsilon_i)$ also provides a convenient language for specifying how the resulting distribution would change in response to external interventions. This is accomplished by encoding each intervention as an alteration to a selected subset of functions, while keeping the others intact. Once we know the identity of the mechanisms altered by the intervention, and the nature of the alteration, the overall effect can be predicted by modifying the corresponding equations in the model, and using the modified model to compute a new probability function.

The simplest type of external intervention is one in which a single variable, say X_i , is forced to take on some fixed value x_i . Such an intervention, which we call atomic, amounts to lifting X_i from the influence of the old functional mechanism $X_i = f_i(pa_i, \varepsilon_i)$ and placing it under the influence of a new mechanism that sets its value to x_i while keeping all other mechanisms unperturbed. Formally, this atomic intervention, which we denote by $\text{set}(X_i = x_i)$, or $\text{set}(x_i)$ for short, amounts to removing the equation $X_i = f_i(pa_i, \varepsilon_i)$ from the model, and substituting x_i for X_i in the remaining equations. The model thus created represents the system's behaviour under the intervention $\text{set}(X_i = x_i)$ and, when solved for the distribution of X_j , yields the causal effect of X_i on X_j , denoted by $\text{pr}(x_j | \check{x}_i)$. More generally, when an intervention forces a subset X of variables to fixed values x , a subset of equations is to be pruned from the model given in (18.3), one for each member of X , thus defining a new distribution over the remaining variables, which completely characterises the effect of the intervention. We thus introduce the following.

Definition 18.2 causal effect

Given two disjoint sets of variables, X and Y , the causal effect of X on Y , denoted $\text{pr}(y | \check{x})$, is a function from X to the space of probability distributions on Y . For each realisation x of X , $\text{pr}(y | \check{x})$ gives the probability of $Y = y$ induced on deleting from the model (18.3) all equations corresponding to variables in X and substituting x for X in the remainder.

An explicit translation of intervention into ‘wiping out’ equations from the model was first proposed by [Strotz & Wold \(1960\)](#), and used by [Fisher \(1970\)](#) and [Sobel \(1990\)](#). Graphical ramifications were explicated by [Spirtes et al. \(1993\)](#) and [Pearl \(1993b\)](#). A related mathematical model using event trees has been introduced by [Robins \(1986, pp. 1422–5\)](#).

Regardless of whether we represent interventions as a modification of an existing model as in Definition 18.2, or as a conditionalisation in an augmented model ([Pearl, 1993b](#)), the result is a well-defined transformation between the pre-intervention and the post-intervention distributions. In the case of an atomic intervention $\text{set}(X_i = x'_i)$, this transformation can be expressed in a simple algebraic formula that follows immediately from (18.3) and Definition 18.2:

$$\text{pr}(x_1, \dots, x_n | \check{x}'_i) = \begin{cases} \{\text{pr}(x_1, \dots, x_n)\} / \{\text{pr}(x_i | pa_i)\} & \text{if } x_i = x'_i, \\ 0 & \text{if } x_i \neq x'_i. \end{cases} \quad (18.5)$$

This formula reflects the removal of the terms $\text{pr}(x_i | pa_i)$ from the product in (18.2), since pa_i no longer influence X_i . Graphically, this is equivalent to removing the links between pa_i and X_i while keeping the rest of the network intact. Equation (18.5) can also be obtained from the G -computation formula of [Robins \(1986, p. 1423\)](#) and the Manipulation Theorem of [Spirtes et al. \(1993\)](#), who state

that this formula was ‘independently conjectured by Fienberg in a seminar in 1991’. Clearly, an intervention $\text{set}(x_i)$ can affect only the descendants of X_i in G . Additional properties of the transformation defined in (18.5) are given by Pearl (1993b).

The immediate implication of (18.5) is that, given a causal diagram in which all parents of manipulated variables are observable, one can infer post-intervention distributions from pre-intervention distributions; hence, under such assumptions we can estimate the effects of interventions from passive, i.e. nonexperimental observations. The aim of this paper, however, is to derive causal effects in situations such as Figure 18.1, where some members of pa_i may be unobservable, thus preventing estimation of $\text{pr}(x_i | pa_i)$. The next two sections provide simple graphical tests for deciding when $\text{pr}(x_j | \check{x}_i)$ is estimable in a given model.

18.3 Controlling Confounding Bias

18.3.1 The Back-Door Criterion

Assume we are given a causal diagram G together with nonexperimental data on a subset V_0 of observed variables in G , and we wish to estimate what effect the intervention $\text{set}(X_i = x_i)$ would have on some response variable X_j . In other words, we seek to estimate $\text{pr}(x_j | \check{x}_i)$ from a sample estimate of $\text{pr}(V_0)$.

The variables in $V_0 \setminus \{X_i, X_j\}$, are commonly known as concomitants (Cox, 1958, p. 48). In observational studies, concomitants are used to reduce confounding bias due to spurious correlations between treatment and response. The condition that renders a set Z of concomitants sufficient for identifying causal effect, also known as ignorability, has been given a variety of formulations, all requiring conditional independence judgments involving counterfactual variables (Rosenbaum & Rubin, 1983; Pratt & Schlaifer, 1988). Pearl (1993b) shows that such judgments are equivalent to a simple graphical test, named the ‘back-door criterion’, which can be applied directly to the causal diagram.

Definition 18.3 Back-door criterion

A set of variables Z satisfies the back-door criterion relative to an ordered pair of variables (X_i, X_j) in a directed acyclic graph G if: (i) no node in Z is a descendant of X_i , and (ii) Z blocks every path between X_i and X_j which contains an arrow into X_i . If X and Y are two disjoint sets of nodes in G , Z is said to satisfy the back-door criterion relative to (X, Y) if it satisfies it relative to any pair (X_i, X_j) such that $X_i \in X$ and $X_j \in Y$.

The name ‘back-door’ echoes condition (ii), which requires that only paths with arrows pointing at X_i be blocked; these paths can be viewed as entering X_i through the back-door. In Figure 18.2, for example, the sets $Z_1 = \{X_3, X_4\}$ and $Z_2 = \{X_4, X_5\}$

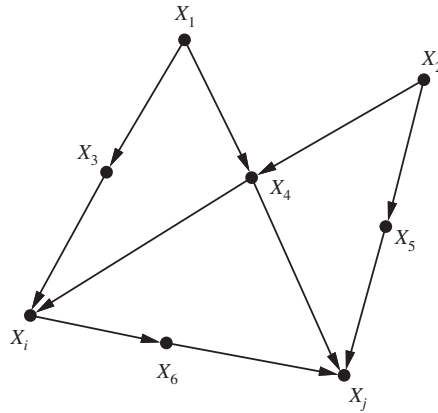


Figure 18.2 A diagram representing the back-door criterion; adjusting for variables $\{X_3, X_4\}$ or $\{X_4, X_5\}$ yields a consistent estimate of $\text{pr}(x_j | \tilde{x}_i)$.

meet the back-door criterion, but $Z_3 = \{X_4\}$ does not, because X_4 does not block the path $(X_i, X_3, X_1, X_4, X_2, X_5, X_j)$. An equivalent, though more complicated, graphical criterion is given in Theorem 7.1 of [Spirtes et al. \(1993\)](#). An alternative criterion using a single d -separation test will be established in § 18.4.4.

We summarise this finding in a theorem, after formally defining ‘identifiability’.

Definition 18.4 Identifiability

The causal effect of X on Y is said to be identifiable if the quantity $\text{pr}(y | \tilde{x})$ can be computed uniquely from any positive distribution of the observed variables that is compatible with G .

Identifiability means that $\text{pr}(y | \tilde{x})$ can be estimated consistently from an arbitrarily large sample randomly drawn from the joint distribution. To prove nonidentifiability, it is sufficient to present two sets of structural equations, both complying with (18.3), that induce identical distributions over observed variables but different causal effects.

Theorem 18.1 *If a set of variables Z satisfies the back-door criterion relative to (X, Y) , then the causal effect of X on Y is identifiable and is given by the formula*

$$\text{pr}(y | \tilde{x}) = \sum_z \text{pr}(y | x, z) \text{pr}(z). \quad (18.6)$$

Equation (18.6) represents the standard adjustment for concomitants Z when X is conditionally ignorable given Z ([Rosenbaum & Rubin, 1983](#)). Reducing ignorability conditions to the graphical criterion of Definition 18.3 replaces judgments about counterfactual dependencies with systematic procedures that can be applied

to causal diagrams of any size and shape. The graphical criterion also enables the analyst to search for an optimal set of concomitants, to minimise measurement cost or sampling variability.

18.3.2 The Front-Door Criteria

An alternative criterion, ‘the front-door criterion’, may be applied in cases where we cannot find observed covariates Z satisfying the back-door conditions. Consider the diagram in Figure 18.3. Although Z does not satisfy any of the back-door conditions, measurements of Z nevertheless enable consistent estimation of $\text{pr}(y | \check{x})$. This can be shown by reducing the expression for $\text{pr}(y | \check{x})$ to formulae computable from the observed distribution function $\text{pr}(x, y, z)$.

The joint distribution associated with Figure 18.3 can be decomposed into

$$\text{pr}(x, y, z, u) = \text{pr}(u) \text{pr}(x | u) \text{pr}(z | x) \text{pr}(y | z, u) \tag{18.7}$$

and, from (18.5), the causal effect of X on Y is given by

$$\text{pr}(y | \check{x}) = \sum_u \text{pr}(y | x, u) \text{pr}(u). \tag{18.8}$$

Using the conditional independence assumptions implied by the decomposition (18.7), we can eliminate u from (18.8) to obtain

$$\text{pr}(y | \check{x}) = \sum_z \text{pr}(z | x) \sum_{x'} \text{pr}(y | x', z) \text{pr}(x'). \tag{18.9}$$

We summarise this result by a theorem.

Theorem 18.2 *Suppose a set of variables Z satisfies the following conditions relative to an ordered pair of variables (X, Y) : (i) Z intercepts all directed paths from X to Y , (ii) there is no back-door path between X and Z , and (iii) every back-door path between Z and Y is blocked by X . Then the causal effect of X on Y is identifiable and is given by (18.9).*

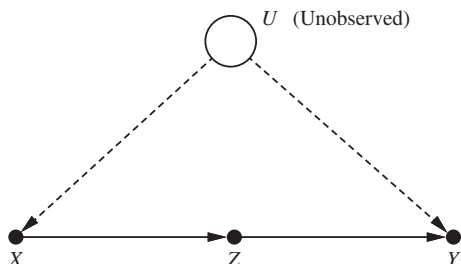


Figure 18.3 A diagram representing the front-door criterion.

The graphical criterion of Theorem 18.2 uncovers many new structures that permit the identification of causal effects from measurements of variables that are affected by treatments: see § 18.5.2. The relevance of such structures in practical situations can be seen, for instance, if we identify X with smoking, Y with lung cancer, Z with the amount of tar deposited in a subject's lungs, and U with an unobserved carcinogenic genotype that, according to some, also induces an inborn craving for nicotine. In this case, (18.9) would provide us with the means to quantify, from nonexperimental data, the causal effect of smoking on cancer, assuming, of course, that $\text{pr}(x, y, z)$ is available and that we believe that smoking does not have any direct effect on lung cancer except that mediated by tar deposits.

18.4 A Calculus of Intervention

18.4.1 General

This section establishes a set of inference rules by which probabilistic sentences involving interventions and observations can be transformed into other such sentences, thus providing a syntactic method of deriving or verifying claims about interventions. We shall assume that we are given the structure of a causal diagram G in which some of the nodes are observable while the others remain unobserved. Our main problem will be to facilitate the syntactic derivation of causal effect expressions of the form $\text{pr}(y | \check{x})$, where X and Y denote sets of observed variables. By derivation we mean step-wise reduction of the expression $\text{pr}(y | \check{x})$ to an equivalent expression involving standard probabilities of observed quantities. Whenever such reduction is feasible, the causal effect of X on Y is identifiable: see Definition 18.4.

18.4.2 Preliminary Notation

Let X , Y and Z be arbitrary disjoint sets of nodes in a directed acyclic graph G . We denote by $G_{\overline{X}}$ the graph obtained by deleting from G all arrows pointing to nodes in X . Likewise, we denote by $G_{\underline{X}}$ the graph obtained by deleting from G all arrows emerging from nodes in X . To represent the deletion of both incoming and outgoing arrows, we use the notation $G_{\overline{XZ}}$: see Figure 18.4 for illustration. Finally, $\text{pr}(y | \check{x}, z) := \text{pr}(y, z | \check{x}) / \text{pr}(z | \check{x})$ denotes the probability of $Y = y$ given that $Z = z$ is observed and X is held constant at x .

18.4.3 Inference Rules

The following theorem states the three basic inference rules of the proposed calculus. Proofs are provided in the 18.A.

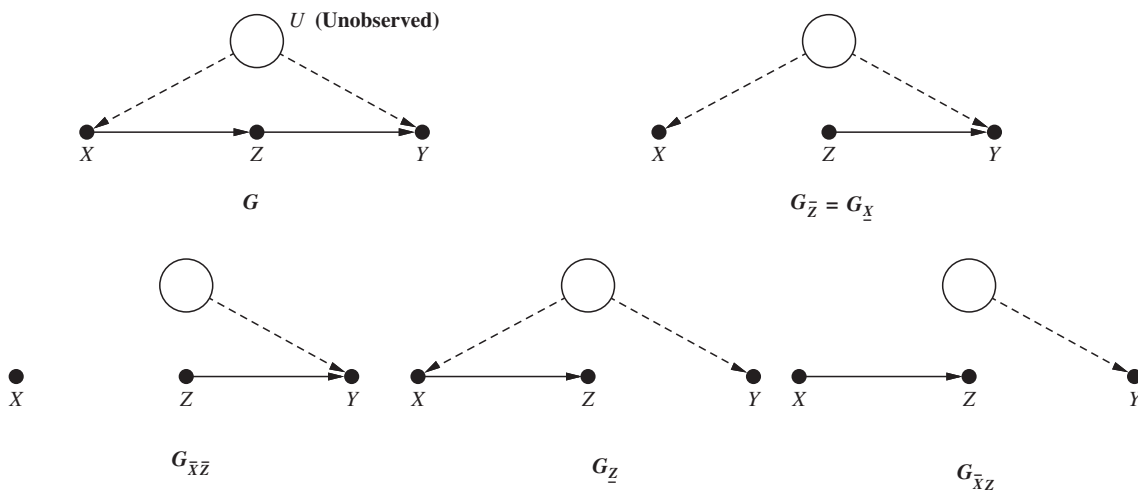


Figure 18.4 Subgraphs of G used in the derivation of causal effects.

Theorem 18.3 Let G be the directed graph associated with a causal model as defined in (18.3), and let $\text{pr}(\cdot)$ stand for the probability distribution induced by that model. For any disjoint subsets of variables X, Y, Z and W we have the following.

Rule 1 (insertion/deletion of observations):

$$\text{pr}(y | \check{x}, z, w) = \text{pr}(y | \check{x}, w) \text{ if } (Y \perp\!\!\!\perp Z | X, W)_{G_{\bar{X}}}. \tag{18.10}$$

Rule 2 (action/observation exchange):

$$\text{pr}(y | \check{x}, \check{z}, w) = \text{pr}(y | \check{x}, z, w) \text{ if } (Y \perp\!\!\!\perp Z | X, W)_{G_{\bar{X}Z}}. \tag{18.11}$$

Rule 3 (insertion/deletion of actions):

$$\text{pr}(y | \check{x}, \check{z}, w) = \text{pr}(y | \check{x}, w) \text{ if } (Y \perp\!\!\!\perp Z | X, W)_{G_{\bar{X}, \bar{Z}(w)}}, \tag{18.12}$$

where $Z(W)$ is the set of Z -nodes that are not ancestors of any W -node in $G_{\bar{X}}$.

Each of the inference rules above follows from the basic interpretation of the ‘ \check{x} ’ operator as a replacement of the causal mechanism that connects X to its pre-intervention parents by a new mechanism $X = x$ introduced by intervening force. The result is a submodel characterised by the subgraph $G_{\bar{X}}$, called the ‘manipulated graph’ by [Spirtes et al. \(1993\)](#), which supports all three rules.

Rule 1 reaffirms d -separation as a valid test for conditional independence in the distribution resulting from the intervention $\text{set}(X = x)$, hence the graph $G_{\bar{X}}$.

This rule follows from the fact that deleting equations from the system does not introduce any dependencies among the remaining disturbance terms: see (18.3).

Rule 2 provides a condition for an external intervention $\text{set}(Z = z)$ to have the same effect on Y as the passive observation $Z = z$. The condition amounts to $X \cup W$ blocking all back-door paths from Z to Y in $G_{\bar{X}}$, since $G_{\bar{X}Z}$ retains all, and only, such paths.

Rule 3 provides conditions for introducing or deleting an external intervention $\text{set}(Z = z)$ without affecting the probability of $Y = y$. The validity of this rule stems, again, from simulating the intervention $\text{set}(Z = z)$ by the deletion of all equations corresponding to the variables in Z .

Corollary 18.1 *A causal effect $q = \text{pr}(y_1, \dots, y_k | \check{x}_1, \dots, \check{x}_m)$ is identifiable in a model characterised by a graph G if there exists a finite sequence of transformations, each conforming to one of the inference rules in Theorem 18.3, which reduces q into a standard, i.e. check-free, probability expression involving observed quantities.*

Whether the three rules above are sufficient for deriving all identifiable causal effects remains an open question. However, the task of finding a sequence of transformations, if such exists, for reducing an arbitrary causal effect expression can be systematised and executed by efficient algorithms as described by Galles & Pearl (1995). As § 18.4.4 illustrates, symbolic derivations using the check notation are much more convenient than algebraic derivations that aim at eliminating latent variables from standard probability expressions, as in § 3.2.

18.4.4 Symbolic Derivation of Causal Effects: An Example

We now demonstrate how Rules 1–3 can be used to derive causal effect estimands in the structure of Figure 18.3 above. Figure 18.4 displays the subgraphs that will be needed for the derivations that follow.

Task 1: compute $\text{pr}(z | \check{x})$. This task can be accomplished in one step, since G satisfies the applicability condition for Rule 2, namely, $X \perp\!\!\!\perp Z$ in $G_{\bar{X}}$, because the path $X \leftarrow U \rightarrow Y \leftarrow Z$ is blocked by the converging arrows at Y , and we can write

$$\text{pr}(z | \check{x}) = \text{pr}(z | x). \quad (18.13)$$

Task 2: compute $\text{pr}(y | \check{z})$. Here we cannot apply Rule 2 to exchange \check{z} with z because $G_{\bar{Z}}$ contains a back-door path from Z to Y : $Z \leftarrow X \leftarrow U \rightarrow Y$. Naturally, we would like to block this path by measuring variables, such as X , that reside on that path. This involves conditioning and summing over all values of X :

$$\text{pr}(y | \check{z}) = \sum_x \text{pr}(y | x, \check{z}) \text{pr}(x | \check{z}). \quad (18.14)$$

We now have to deal with two expressions involving \check{z} , $\text{pr}(y|x, \check{z})$ and $\text{pr}(x|\check{z})$. The latter can be readily computed by applying Rule 3 for action deletion:

$$\text{pr}(x|\check{z}) = \text{pr}(x) \text{ if } (Z \perp\!\!\!\perp X)_{G_{\check{Z}}}, \quad (18.15)$$

since X and Z are d -separated in $G_{\check{Z}}$. Intuitively, manipulating Z should have no effect on X , because Z is a descendant of X in G . To reduce $\text{pr}(y|x, \check{z})$, we consult Rule 2:

$$\text{pr}(y|x, \check{z}) = \text{pr}(y|x, z) \text{ if } (Z \perp\!\!\!\perp Y|X)_{G_{\check{Z}}}, \quad (18.16)$$

noting that X d -separates Z from Y in $G_{\check{Z}}$. This allows us to write (18.14) as

$$\text{pr}(y|\check{z}) = \sum_x \text{pr}(y|x, z) \text{pr}(x) = E_x \text{pr}(y|x, z), \quad (18.17)$$

which is a special case of the back-door formula (18.6). The legitimising condition, $(Z \perp\!\!\!\perp Y|X)_{G_{\check{Z}}}$, offers yet another graphical test for the ignorability condition of Rosenbaum & Rubin (1983).

Task 3: compute $\text{pr}(y|\check{x})$. Writing

$$\text{pr}(y|\check{x}) = \sum_z \text{pr}(y|z, \check{x}) \text{pr}(z|\check{x}), \quad (18.18)$$

we see that the term $\text{pr}(z|\check{x})$ was reduced in (18.13) but that no rule can be applied to eliminate the ‘check’ symbol from the term $\text{pr}(y|z, \check{x})$. However, we can add a ‘check’ symbol to this term via Rule 2:

$$\text{pr}(y|z, \check{x}) = \text{pr}(y|z, \check{z}, \check{x}), \quad (18.19)$$

since the applicability condition $(Y \perp\!\!\!\perp Z|X)_{G_{\check{X}\check{Z}}}$, holds true. We can now delete the action \check{x} from $\text{pr}(y|z, \check{x})$ using Rule 3, $Y \perp\!\!\!\perp X|Z$ holds in $G_{\check{X}\check{Z}}$. Thus, we have

$$\text{pr}(y|z, \check{x}) = \text{pr}(y|z), \quad (18.20)$$

which was calculated in (18.17). Substituting (18.17), (18.20) and (18.13) back into (18.18) finally yields

$$\text{pr}(y|\check{x}) = \sum_z \text{pr}(z|x) \sum_{x'} \text{pr}(y|x', z) \text{pr}(x'), \quad (18.21)$$

which is identical to the front-door formula (18.9).

The reader may verify that all other causal effects, for example, $\text{pr}(y, z|\check{x})$ and $\text{pr}(x, z|\check{y})$, can likewise be derived through the rules of Theorem 18.3. Note that in all the derivations the graph G provides both the license for applying the inference rules and the guidance for choosing the right rule to apply.

18.4.5 Causal Inference by Surrogate Experiments

Suppose we wish to learn the causal effect of X on Y when $\text{pr}(y|\check{x})$ is not identifiable and, for practical reasons of cost or ethics, we cannot control X by randomised experiment. The question arises whether $\text{pr}(y|\check{x})$ can be identified by randomising a surrogate variable Z , which is easier to control than X . For example, if we are interested in assessing the effect of cholesterol levels X on heart disease, Y , a reasonable experiment to conduct would be to control subjects' diet, Z , rather than exercising direct control over cholesterol levels in subjects' blood.

Formally, this problem amounts to transforming $\text{pr}(y|\check{x})$ into expressions in which only members of Z carry the check symbol. Using Theorem 3 it can be shown that the following conditions are sufficient for admitting a surrogate variable Z : (i) X intercepts all directed paths from Z to Y , and (ii) $\text{pr}(y|\check{x})$ is identifiable in $G_{\check{Z}}$. Indeed, if condition (i) holds, we can write $\text{pr}(y|\check{x}) = \text{pr}(y|\check{x}, \check{z})$, because $(Y \perp\!\!\!\perp Z | X)_{G_{\check{Z}}}$. But $\text{pr}(y|\check{x}, \check{z})$ stands for the causal effect of X on Y in a model governed by $G_{\check{Z}}$ which, by condition (ii), is identifiable. Figures 18.7(e) and 18.7(h) below illustrate models in which both conditions hold. Translated to our cholesterol example, these conditions require that there be no direct effect of diet on heart conditions and no confounding effect between cholesterol levels and heart disease, unless we can measure an intermediate variable between the two.

18.5 Graphical Tests of Identifiability

18.5.1 General

Figure 18.5 shows simple diagrams in which $\text{pr}(y|\check{x})$ cannot be identified due to the presence of a bow pattern, i.e. a confounding arc, shown dashed, embracing a causal link between X and Y . A confounding arc represents the existence in the diagram of a back-door path that contains only unobserved variables and has no converging arrows. For example, the path X, Z_0, B, Z_3 in Figure 18.1 can be represented as a confounding arc between X and Z_3 . A bow-pattern represents an equation $Y = f_Y(X, U, \varepsilon_Y)$, where U is unobserved and dependent on X . Such an equation does not permit the identification of causal effects since any portion of the observed dependence between X and Y may always be attributed to spurious dependencies mediated by U .

The presence of a bow-pattern prevents the identification of $\text{pr}(y|\check{x})$ even when it is found in the context of a larger graph, as in Figure 18.5(b). This is in contrast to linear models, where the addition of an arc to a bow-pattern can render $\text{pr}(y|\check{x})$ identifiable. For example, if Y is related to X via a linear relation $Y = bX + U$, where U is an unobserved disturbance possibly correlated with X , then $b = \partial E(Y|\check{x})/\partial x$ is not identifiable. However, adding an arc $Z \rightarrow X$ to the structure, that is, finding a

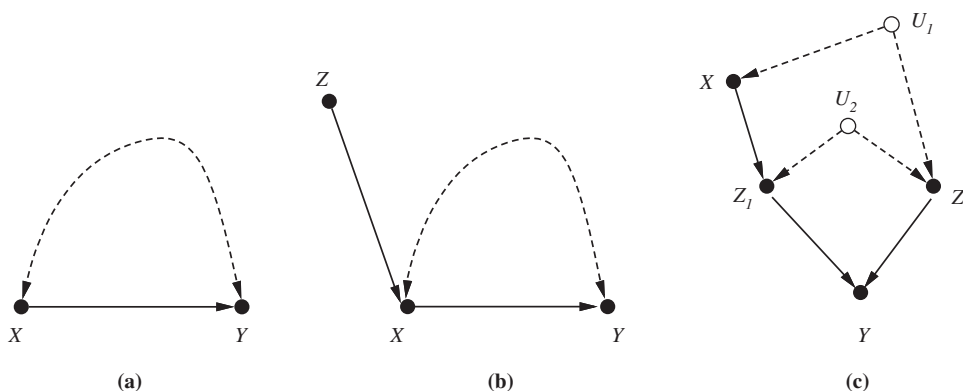


Figure 18.5 (a) A bow-pattern: a confounding arc embracing a causal link $X \rightarrow Y$, thus preventing the identification of $\text{pr}(y | \check{x})$ even in the presence of an instrumental variable Z , as in (b). (c) A bow-less graph still prohibiting the identification of $\text{pr}(y | \check{x})$.

variable Z that is correlated with X but not with U , would facilitate the computation of $E(Y | \check{x})$ via the instrumental-variable formula (Bowden & Turkington, 1984, p. 12; Angrist, Imbens & Rubin, 1995):

$$b := \frac{\partial}{\partial x} E(Y | \check{x}) = \frac{E(Y | z)}{E(X | z)} = \frac{R_{yz}}{R_{xz}}. \tag{18.22}$$

In nonparametric models, adding an instrumental variable Z to a bow-pattern, see Figure 18.5(b), does not permit the identification of $\text{pr}(y | \check{x})$. This is a familiar problem in the analysis of clinical trials in which treatment assignment, Z , is randomised, hence no link enters Z , but compliance is imperfect. The confounding arc between X and Y in Figure 18.5(b) represents unmeasurable factors which influence both subjects' choice of treatment, X , and response to treatment, Y . In such trials, it is not possible to obtain an unbiased estimate of the treatment effect $\text{pr}(y | \check{x})$ without making additional assumptions on the nature of the interactions between compliance and response (Imbens & Angrist, 1994), as is done, for example, in the approach to instrumental variables developed by Angrist et al. (1995). While the added arc $Z \rightarrow X$ permits us to calculate bounds on $\text{pr}(y | \check{x})$ (Robins, 1989, § 1g; Manski, 1990), and while the upper and lower bounds may even coincide for certain types of distributions $\text{pr}(x, y, z)$ (Balke & Pearl, 1994), there is no way of computing $\text{pr}(y | \check{x})$ for every positive distribution $\text{pr}(x, y, z)$, as required by Definition 18.4.

In general, the addition of arcs to a causal diagram can impede, but never assist, the identification of causal effects in nonparametric models. This is because such addition reduces the set of d -separation conditions carried by the diagram and, hence, if a causal effect derivation fails in the original diagram, it is bound to fail

in the augmented diagram as well. Conversely, any causal effect derivation that succeeds in the augmented diagram, by a sequence of symbolic transformations, as in Corollary 18.1, would succeed in the original diagram.

Our ability to compute $\text{pr}(y|\check{x})$ for pairs (x,y) of singleton variables does not ensure our ability to compute joint distributions, such as $\text{pr}(y_1, y_2|\check{x})$. Figure 18.5(c), for example, shows a causal diagram where both $\text{pr}(z_1|\check{x})$ and $\text{pr}(z_2|\check{x})$ are computable, but $\text{pr}(z_1, z_2|\check{x})$ is not. Consequently, we cannot compute $\text{pr}(y|\check{x})$. This diagram is the smallest graph that does not contain a bow-pattern and still presents an uncomputable causal effect.

18.5.2 Identifying Models

Figure 18.6 shows simple diagrams in which the causal effect of X on Y is identifiable. Such models are called identifying because their structures communicate a sufficient number of assumptions to permit the identification of the target quantity $\text{pr}(y|\check{x})$. Latent variables are not shown explicitly in these diagrams; rather, such variables are implicit in the confounding arcs, shown dashed. Every causal diagram with latent variables can be converted to an equivalent diagram involving measured variables interconnected by arrows and confounding arcs. This conversion corresponds to substituting out all latent variables from the structural equations of (18.3) and then constructing a new diagram by connecting any two variables X_i and X_j by (i) an arrow from X_j to X_i whenever X_j appears in the equation for X_i , and (ii) a confounding arc whenever the same ε term appears in both f_i and f_j . The result is a diagram in which all unmeasured variables are exogenous and mutually independent. Several features should be noted from examining the diagrams in Figure 18.6.

(i) Since the removal of any arc or arrow from a causal diagram can only assist the identifiability of causal effects, $\text{pr}(y|\check{x})$ will still be identified in any edge-subgraph of the diagrams shown in Figure 18.6. Likewise, the introduction of mediating observed variables onto any edge in a causal graph can assist, but never impede, the identifiability of any causal effect. Therefore, $\text{pr}(y|\check{x})$ will still be identified from any graph obtained by adding mediating nodes to the diagrams shown in Figure 18.6.

(ii) The diagrams in Figure 18.6 are maximal, in the sense that the introduction of any additional arc or arrow onto an existing pair of nodes would render $\text{pr}(y|\check{x})$ no longer identifiable.

(iii) Although most of the diagrams in Figure 18.6 contain bow-patterns, none of these patterns emanates from X as is the case in Figure 18.7(a) and (b) below. In general, a necessary condition for the identifiability of $\text{pr}(y|\check{x})$ is the absence of a confounding arc between X and any child of X that is an ancestor of Y .

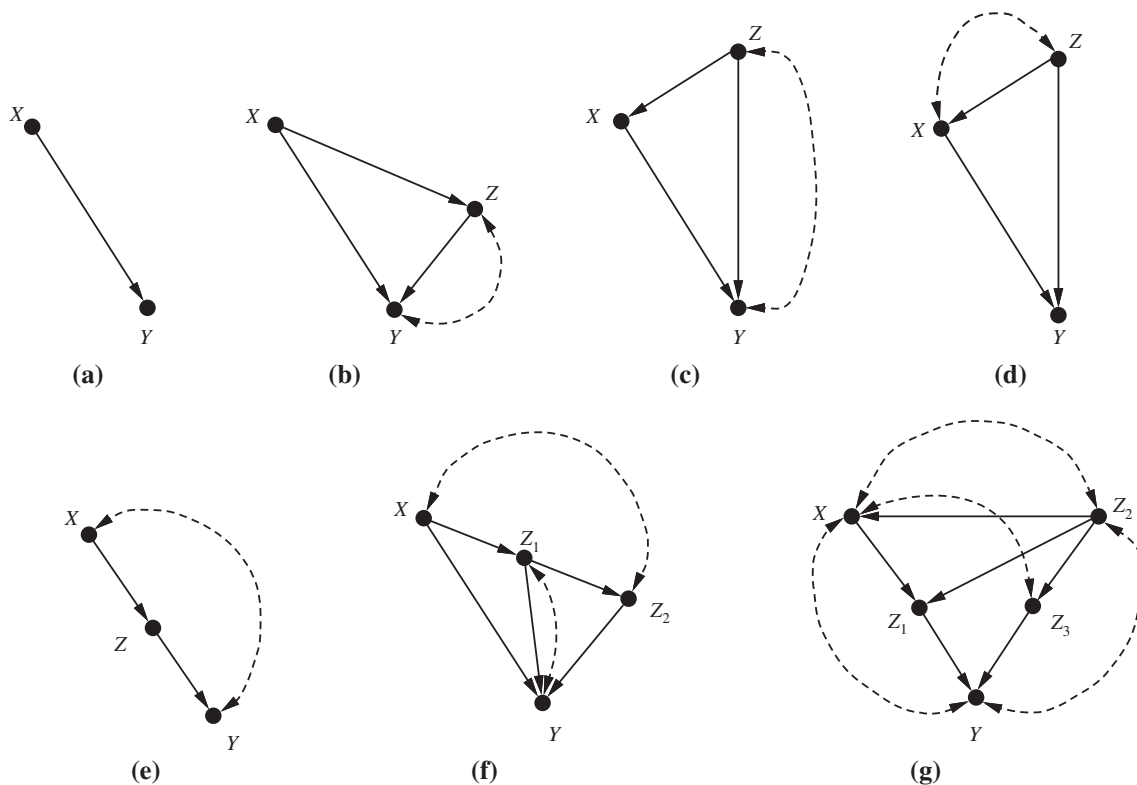


Figure 18.6 Typical models in which the effect of X on Y is identifiable. Dashed arcs represent confounding paths, and Z represents observed covariates.

(iv) Figures 18.6(a) and (b) contain no back-door paths between X and Y , and thus represent experimental designs in which there is no confounding bias between the treatment, X , and the response, Y ; that is, X is strongly ignorable relative to Y (Rosenbaum & Rubin, 1983); hence, $\text{pr}(y|\check{x}) = \text{pr}(y|x)$. Likewise, Figures 18.6(c) and (d) represent designs in which observed covariates, Z , block every back-door path between X and Y ; that is X is conditionally ignorable given Z (Rosenbaum & Rubin, 1983); hence, $\text{pr}(y|\check{x})$ is obtained by standard adjustment for Z , as in (18.6):

$$\text{pr}(y|\check{x}) = \sum_z \text{pr}(y|x, z) \text{pr}(z).$$

(v) For each of the diagrams in Figure 18.6, we can readily obtain a formula for $\text{pr}(y|\check{x})$, using symbolic derivations patterned after those in § 18.4.4. The derivation is often guided by the graph topology. For example, Figure 18.6(f) dictates the

following derivation. Writing

$$\text{pr}(y | \check{x}) = \sum_{z_1, z_2} \text{pr}(y | z_1, z_2, \check{x}) \text{pr}(z_1, z_2 | \check{x}),$$

we see that the subgraph containing $\{X, Z_1, Z_2\}$ is identical in structure to that of Figure 18.6(e), with Z_1, Z_2 replacing Z, Y , respectively. Thus, $\text{pr}(z_1, z_2 | \check{x})$ can be obtained from (18.14) and (18.21). Likewise, the term $\text{pr}(y | z_1, z_2, \check{x})$ can be reduced to $\text{pr}(y | z_1, z_2, x)$ by Rule 2, since $(Y \perp\!\!\!\perp X | Z_1, Z_2)_{G_{\check{x}}}$. Thus, we have

$$\text{pr}(y | \check{x}) = \sum_{z_1, z_2} \text{pr}(y | z_1, z_2, x) \text{pr}(z_1 | x) \sum_{x'} \text{pr}(z_2 | z_1, x') \text{pr}(x'). \quad (18.23)$$

Applying a similar derivation to Figure 18.6(g) yields

$$\text{pr}(y | \check{x}) = \sum_{z_1} \sum_{z_2} \sum_{x'} \text{pr}(y | z_1, z_2, x') \text{pr}(x') \text{pr}(z_1 | z_2, x) \text{pr}(z_2). \quad (18.24)$$

Note that the variable Z_3 does not appear in the expression above, which means that Z_3 need not be measured if all one wants to learn is the causal effect of X on Y .

(vi) In Figures 18.6(e), (f) and (g), the identifiability of $\text{pr}(y | \check{x})$ is rendered feasible through observed covariates, Z , that are affected by the treatment X , that is descendants of X . This stands contrary to the warning, repeated in most of the literature on statistical experimentation, to refrain from adjusting for concomitant observations that are affected by the treatment (Cox, 1958, p. 48; Rosenbaum, 1984; Pratt & Schlaifer, 1988; Wainer, 1989). It is commonly believed that, if a concomitant Z is affected by the treatment, then it must be excluded from the analysis of the total effect of the treatment (Pratt & Schlaifer, 1988). The reasons given for the exclusion is that the calculation of total effects amounts to integrating out Z , which is functionally equivalent to omitting Z to begin with. Figures 18.6(e), (f) and (g) show cases where one wants to learn the total effects of X and, still, the measurement of concomitants that are affected by X , for example Z or Z_1 , is necessary. However, the adjustment needed for such concomitants is nonstandard, involving two or more stages of the standard adjustment of (18.6): see (18.9), (18.23) and (18.24).

(vii) In Figures 18.6(b), (c) and (f), Y has a parent whose effect on Y is not identifiable, yet the effect of X on Y is identifiable. This demonstrates that local identifiability is not a necessary condition for global identifiability. In other words, to identify the effect of X on Y we need not insist on identifying each and every link along the paths from X to Y .

18.5.3 Nonidentifying Models

Figure 18.7 presents typical diagrams in which the total effect of X on Y , $\text{pr}(y | \check{x})$, is not identifiable. Noteworthy features of these diagrams are as follows.

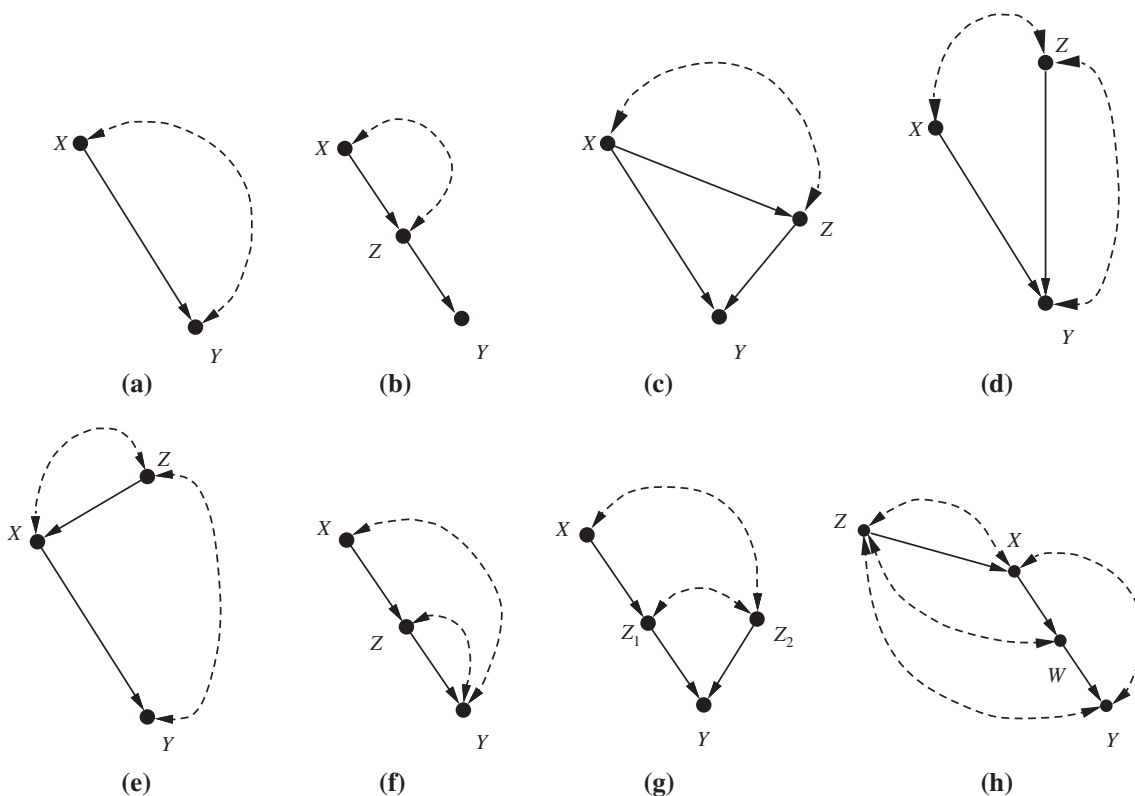


Figure 18.7 Typical models in which $\text{pr}(y | \tilde{x})$ is not identifiable.

(i) All graphs in Figure 18.7 contain unblockable back-door paths between X and Y , that is, paths ending with arrows pointing to X which cannot be blocked by observed nondescendants of X . The presence of such a path in a graph is, indeed, a necessary test for nonidentifiability. It is not a sufficient test, though, as is demonstrated by Figure 18.6(e), in which the back-door path (dashed) is unblockable, yet $\text{pr}(y | \tilde{x})$ is identifiable.

(ii) A sufficient condition for the nonidentifiability of $\text{pr}(y | \tilde{x})$ is the existence of a confounding path between X and any of its children on a path from X to Y , as shown in Figures 18.7(b) and (c). A stronger sufficient condition is that the graph contain any of the patterns shown in Figure 18.7 as an edge-subgraph.

(iii) Figure 18.7(g) demonstrates that local identifiability is not sufficient for global identifiability. For example, we can identify $\text{pr}(z_1 | \tilde{x})$, $\text{pr}(z_2 | \tilde{x})$, $\text{pr}(y | \tilde{z}_1)$ and $\text{pr}(y | \tilde{z}_2)$, but not $\text{pr}(y | \tilde{x})$. This is one of the main differences between nonparametric and linear models; in the latter, all causal effects can be determined from

the structural coefficients, each coefficient representing the causal effect of one variable on its immediate successor.

18.6 Discussion

The basic limitation of the methods proposed in this paper is that the results must rest on the causal assumptions shown in the graph, and that these cannot usually be tested in observational studies. In related papers (Pearl, 1994a, 1995) we show that some of the assumptions, most notably those associated with instrumental variables, see Figure 18.5(b), are subject to falsification tests. Additionally, considering that any causal inferences from observational studies must ultimately rely on some kind of causal assumptions, the methods described in this paper offer an effective language for making those assumptions precise and explicit, so they can be isolated for deliberation or experimentation and, once validated, integrated with statistical data.

A second limitation concerns an assumption inherent in identification analysis, namely, that the sample size is so large that sampling variability may be ignored. The mathematical derivation of causal-effect estimands should therefore be considered a first step toward supplementing estimates of these with confidence intervals and significance levels, as in traditional analysis of controlled experiments. Having nonparametric estimates for causal effects does not imply that one should refrain from using parametric forms in the estimation phase of the study. For example, if the assumptions of Gaussian, zero-mean disturbances and linearity are deemed reasonable, then the estimand in (18.9) can be replaced by $E(Y | \check{x}) = R_{xz} \beta_{zy \cdot x} x$, where $\beta_{zy \cdot x}$ is the standardised regression coefficient, and the estimation problem reduces to that of estimating coefficients. More sophisticated estimation techniques are given by Rubin (1978), Robins (1989, § 17), and Robins et al. (1992, pp. 331–3).

Several extensions of the methods proposed in this paper are possible. First, the analysis of atomic interventions can be generalised to complex policies in which a set X of treatment variables is made to respond in a specified way to some set Z of covariates, say through a functional relationship $X = g(Z)$ or through a stochastic relationship whereby X is set to x with probability $P^*(x | z)$. Pearl (1994b) shows that computing the effect of such policies is equivalent to computing the expression $\text{pr}(y | \check{x}, z)$.

A second extension concerns the use of the intervention calculus of Theorem 18.3 in nonrecursive models, that is, in causal diagrams involving directed cycles or feedback loops. The basic definition of causal effects in terms of ‘wiping out’ equations from the model (Definition 18.2) still carries over to nonrecursive systems (Strotz & Wold, 1960; Sobel, 1990), but then two issues must be addressed.

First, the analysis of identification must ensure the stability of the remaining submodels (Fisher, 1970). Secondly, the d -separation criterion for directed acyclic graphs must be extended to cover cyclic graphs as well. The validity of d -separation has been established for nonrecursive linear models and extended, using an augmented graph, to any arbitrary set of stable equations (Spirtes, 1995). However, the computation of causal effect estimands will be harder in cyclic networks, because symbolic reduction of $\text{pr}(y | \tilde{x})$ to check-free expressions may require the solution of nonlinear equations.

Finally, a few comments regarding the notation introduced in this paper. There have been three approaches to expressing causal assumptions in mathematical form. The most common approach in the statistical literature invokes Rubin's model (Rubin, 1974), in which probability functions are defined over an augmented space of observable and counterfactual variables. In this model, causal assumptions are expressed as independence constraints over the augmented probability function, as exemplified by Rosenbaum & Rubin's (1983) definitions of ignorability conditions. An alternative but related approach, still using the standard language of probability, is to define augmented probability functions over variables representing hypothetical interventions (Pearl, 1993b).

The language of structural models, which includes path diagrams (Wright, 1921) and structural equations (Goldberger, 1972) represents a drastic departure from these two approaches, because it invokes new primitives, such as arrows, disturbance terms, or plain causal statements, which have no parallels in the language of probability. This language has been very popular in the social sciences and econometrics, because it closely echoes statements made in ordinary scientific discourse and thus provides a natural way for scientists to communicate knowledge and experience, especially in situations involving many variables.

Statisticians, however, have generally found structural models suspect, because the empirical content of basic notions in these models appears to escape conventional methods of explication. For example, analysts have found it hard to conceive of experiments, however hypothetical, whose outcomes would be constrained by a given structural equation. Standard probability calculus cannot express the empirical content of the coefficient b in the structural equation $Y = bX + \varepsilon_Y$ even if one is prepared to assume that ε_Y , an unobserved quantity, is uncorrelated with X . Nor can any probabilistic meaning be attached to the analyst's excluding from this equation certain variables that are highly correlated with X or Y . As a consequence, the whole enterprise of structural equation modelling has become the object of serious controversy and misunderstanding among researchers (Freedman, 1987; Wermuth, 1992; Whittaker, 1990, p. 302; Cox & Wermuth, 1993).

To a large extent, this history of controversy stems not from faults in the structural modelling approach but rather from a basic limitation of standard probability theory: when viewed as a mathematical language, it is too weak to describe the precise experimental conditions that prevail in a given study. For example, standard probabilistic notation cannot distinguish between an experiment in which variable X is observed to take on value x and one in which variable X is set to value x by some external control. The need for this distinction was recognised by several researchers, most notably Pratt & Schlaifer (1988) and Cox (1992), but has not led to a more refined and manageable mathematical notation capable of reflecting this distinction.

The ‘check’ notation developed in this paper permits one to specify precisely what is being held constant and what is merely measured in a given study and, using this specification, the basic notions of structural models can be given clear empirical interpretation. For example, the meaning of b in the equation $Y = bX + \varepsilon_Y$ is simply $\partial E(Y | \check{x}) / \partial x$, namely, the rate of change, in x , of the expectation of Y in an experiment where X is held at x by external control. This interpretation holds regardless of whether ε_Y and X are correlated, for example, via another equation: $X = aY + \varepsilon_X$. Moreover, the notion of randomisation need not be invoked. Likewise, the analyst’s decision as to which variables should be included in a given equation is based on a hypothetical controlled experiment: a variable Z is excluded from the equation for Y if it has no influence on Y when all other variables, S_{YZ} , are held constant, that is, $\text{pr}(y | \check{z}, \check{s}_{YZ}) = \text{pr}(y | \check{s}_{YZ})$. In other words, variables that are excluded from the equation $Y = bX + \varepsilon_Y$ are not conditionally independent of Y given measurements of X , but rather conditionally independent of Y given settings of X . The operational meaning of the so-called ‘disturbance term’, ε_Y , is likewise demystified: ε_Y is defined as the difference $Y - E(Y | \check{s}_Y)$; two disturbance terms, ε_X and ε_Y , are correlated if $\text{pr}(y | \check{x}, \check{s}_{XY}) \neq \text{pr}(y | x, \check{s}_{XY})$; and so on.

The distinctions provided by the ‘check’ notation clarify the empirical basis of structural equations and should make causal models more acceptable to empirical researchers. Moreover, since most scientific knowledge is organised around the operation of ‘holding X fixed’, rather than ‘conditioning on X ’, the notation and calculus developed in this paper should provide an effective means for scientists to communicate subject-matter information, and to infer its logical consequences when combined with statistical data.

Acknowledgments

Much of this investigation was inspired by Spirtes et al. (1993), in which a graphical account of manipulations was first proposed. Phil Dawid, David Freedman, James

Robins and Donald Rubin have provided genuine encouragement and valuable advice. The investigation also benefitted from discussions with Joshua Angrist, Peter Bentler, David Cox, Arthur Dempster, David Galles, Arthur Goldberger, Sander Greenland, David Hendry, Paul Holland, Guido Imbens, Ed Learner, Rod McDonald, John Pratt, Paul Rosenbaum, Keunkwan Ryu, Glenn Shafer, Michael Sobel, David Tritchler and Nanny Wermuth. The research was partially supported by grants from Air Force Office of Scientific Research and National Science Foundation.

18.A Appendix

Proof of Theorem 18.3

(i) Rule 1 follows from the fact that deleting equations from the model in (18.8) results, again, in a recursive set of equations in which all ε terms are mutually independent. The d -separation condition is valid for any recursive model, hence it is valid for the submodel resulting from deleting the equations for X . Finally, since the graph characterising this submodel is given by $G_{\bar{X}}$, $(Y \perp\!\!\!\perp Z | X, W)_{G_{\bar{X}}}$ implies the conditional independence $\text{pr}(y | \check{x}, z, w) = \text{pr}(y | \check{x}, w)$ in the post-intervention distribution.

(ii) The graph $G_{\bar{X}Z}$ differs from $G_{\bar{X}}$ only in lacking the arrows emanating from Z , hence it retains all the back-door paths from Z to Y that can be found in $G_{\bar{X}}$. The condition $(Y \perp\!\!\!\perp Z | X, W)_{G_{\bar{X}Z}}$ ensures that all back-door paths from Z to Y in $G_{\bar{X}}$ are blocked by $\{X, W\}$. Under such conditions, setting $Z = z$ or conditioning on $Z = z$ has the same effect on Y . This can best be seen from the augmented diagram $G'_{\bar{X}}$, to which the intervention arcs $F_Z \rightarrow Z$ were added, where F_Z stands for the functions that determine Z in the structural equations (Pearl, 1993b). If all back-door paths from F_Z to Y are blocked, the remaining paths from F_Z to Y must go through the children of Z , hence these paths will be blocked by Z . The implication is that Y is independent of F_Z given Z , which means that the observation $Z = z$ cannot be distinguished from the intervention $F_Z = \text{set}(z)$.

(iii) The following argument was developed by D. Galles. Consider the augmented diagram $G'_{\bar{X}}$ to which the intervention arcs $F_Z \rightarrow Z$ are added. If $(F_Z \perp\!\!\!\perp Y | W, X)_{G'_{\bar{X}}}$, then $\text{pr}(y | \check{x}, \check{z}, w) = \text{pr}(y | \check{x}, w)$. If $(Y \perp\!\!\!\perp Z | X, W)_{G_{\bar{X}Z(w)}}$ and $(F_Z \not\perp\!\!\!\perp Y | W, X)_{G'_{\bar{X}}}$, there must be an unblocked path from a member $F_{Z'}$ of F_Z to Y that passes either through a head-to-tail junction at Z' , or a head-to-head junction at Z' . If there is such a path, let P be the shortest such path. We will show that P will violate some premise, or there exists a shorter path, either of which leads to a contradiction.

If the junction is head-to-tail, that means that $(Y \not\perp\!\!\!\perp Z' | W, X)_{G'_{\bar{X}}}$ but $(Y \perp\!\!\!\perp Z' | W, X)_{G'_{\bar{X}Z(w)}}$. So, there must be an unblocked path from Y to Z' that passes

through some member Z'' of $Z(W)$ in either a head-to-head or a tail-to-head junction. This is impossible. If the junction is head-to-head, then some descendant of Z'' must be in W for the path to be unblocked, but then Z'' would not be in $Z(W)$. If the junction is tail-to-head, there are two options: either the path from Z' to Z'' ends in an arrow pointing to Z'' , or in an arrow pointing away from Z'' . If it ends in an arrow pointing away from Z'' , then there must be a head-to-head junction along the path from Z' to Z'' . In that case, for the path to be unblocked, W must be a descendant of Z'' , but then Z'' would not be in $Z(W)$. If it ends in an arrow pointing to Z'' , then there must be an unblocked path from Z'' to Y in $G_{\bar{X}}$ that is blocked in $G_{\bar{X}Z(W)}$. If this is true, then there is an unblocked path from $F_{Z''}$ to Y that is shorter than P , the shortest path.

If the junction through Z' is head-to-head, then either Z' is in $Z(W)$, in which case that junction would be blocked, or there is an unblocked path from Z' to Y in $G_{\bar{X}Z(W)}$ that is blocked in $G_{\bar{X}}$. Above, we proved that this could not occur. So $(Y \perp\!\!\!\perp Z | X, W)_{G_{\bar{X}Z(W)}}$ implies $(F_Z \perp\!\!\!\perp Y | W, X)_{G'_{\bar{X}}}$ and thus $\text{pr}(y | \check{x}, \check{z}, w) = \text{pr}(y | \check{x}, w)$.

References

- Angrist, J. D., Imbens, G. W. & Rubin, D. B. (1995). Identification of causal effects using instrumental variables. *J. Am. Statist. Assoc.* **91**, 444–455, 1996.
- Balke, A. & Pearl, J. (1994). Counterfactual probabilities: Computational methods, bounds, and applications. In *Uncertainty in Artificial Intelligence*, Ed. R. Lopez de Mantaras and D. Poole, pp. 46–54. San Mateo, CA: Morgan Kaufmann.
- Bowden, R. J. & Turkington, D. A. (1984). *Instrumental Variables*. Cambridge, MA: Cambridge University Press.
- Cox, D. R. (1958). *The Planning of Experiments*. New York: John Wiley.
- Cox, D. R. (1992). Causality: Some statistical aspects. *J. R. Statist. Soc. A* **155**, 291–301.
- Cox, D. R. & Wermuth, N. (1993). Linear dependencies represented by chain graphs. *Statist. Sci.* **8**, 204–18.
- Dawid, A. P. (1979). Conditional independence in statistical theory (with Discussion). *J. R. Statist. Soc. B* **41**, 1–31.
- Fisher, F. M. (1970). A correspondence principle for simultaneous equation models. *Econometrica* **38**, 73–92.
- Freedman, D. (1987). As others see us: A case study in path analysis (with Discussion). *J. Educ. Statist.* **12**, 101–223.
- Frisch, R. (1938). Statistical versus theoretical relations in economic macrodynamics. *League of Nations Memorandum*. Reproduced (1948) in *Autonomy of Economic Relations*, Universitetets Socialøkonomiske Institutt, Oslo.
- Galles, D. & Pearl, J. (1995). Testing identifiability of causal effects. In *Uncertainty in Artificial Intelligence—11*, Ed. P. Besnard and S. Hanks, pp. 185–95. San Francisco, CA: Morgan Kaufmann.

- Geiger, D., Verma, T. S. & Pearl, J. (1990). Identifying independence in Bayesian networks. *Networks* **20**, 507–34.
- Goldberger, A. S. (1972). Structural equation models in the social sciences. *Econometrica* **40**, 979–1001.
- Haavelmo, T. (1943). The statistical implications of a system of simultaneous equations. *Econometrica* **11**, 1–12.
- Holland, P. W. (1988). Causal inference, path analysis, and recursive structural equations models. In *Sociological Methodology*, Ed. C. Clogg, pp. 449–84. Washington, D.C.: American Sociological Association.
- Imbens, G. W. & Angrist, J. D. (1994). Identification and estimation of local average treatment effects. *Econometrica* **62**, 467–76.
- Lauritzen, S. L., Dawid, A. P., Larsen, B. N. & Leimer, H. G. (1990). Independence properties of directed Markov fields. *Networks* **20**, 491–505.
- Lauritzen, S. L. & Spiegelhalter, D. J. (1988). Local computations with probabilities on graphical structures and their applications to expert systems (with Discussion). *J. R. Statist. Soc. B* **50**, 157–224.
- Manski, C. F. (1990). Nonparametric bounds on treatment effects. *Am. Econ. Rev., Papers Proc.* **80**, 319–23.
- Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems*. San Mateo, CA: Morgan Kaufmann.
- Pearl, J. (1993a). Belief networks revisited. *Artif. Intel.* **59**, 49–56.
- Pearl, J. (1993b). Comment: Graphical models, causality, and intervention. *Statist. Sci.* **8**, 266–9.
- Pearl, J. (1994a). From Bayesian networks to causal networks. In *Bayesian Networks and Probabilistic Reasoning*, Ed. A. Gammerman, pp. 1–31. London: Alfred Walter.
- Pearl, J. (1994b). A probabilistic calculus of actions. In *Uncertainty in Artificial Intelligence*, Ed. R. Lopez de Mantaras and D. Poole, pp. 452–62. San Mateo, CA: Morgan Kaufmann.
- Pearl, J. (1995). Causal inference from indirect experiments. *Artif. Intel. Med. J.*, **7**, 561–582, 1995.
- Pearl, J. & Verma, T. (1991). A theory of inferred causation. In *Principles of Knowledge Representation and Reasoning: Proceedings of the 2nd International Conference*, Ed. J. A. Allen, R. Fikes and E. Sandewall, pp. 441–52. San Mateo, CA: Morgan Kaufmann.
- Pratt, J. W. & Schlaifer, R. (1988). On the interpretation and observation of laws. *J. Economet.* **39**, 23–52.
- Robins, J. M. (1986). A new approach to causal inference in mortality studies with a sustained exposure period—applications to control of the healthy workers survivor effect. *Math. Model.* **7**, 1393–512.
- Robins, J. M. (1989). The analysis of randomized and non-randomized AIDS treatment trials using a new approach to causal inference in longitudinal studies. In *Health Service*

- Research Methodology: A Focus on AIDS*, Ed. L. Sechrest, H. Freeman and A. Mulley, pp. 113–59. Washington, D.C.: NCHSR, U.S. Public Health Service.
- Robins, J. M., Blevins, D., Ritter, G. & Wulfsohn, M. (1992). *G*-estimation of the effect of prophylaxis therapy for pneumocystis carinii pneumonia on the survival of AIDS patients. *Epidemiology* 3, 319–36.
- Rosenbaum, P. R. (1984). The consequences of adjustment for a concomitant variable that has been affected by the treatment. *J. R. Statist. Soc. A* 147, 656–66.
- Rosenbaum, P. & Rubin, D. (1983). The central role of propensity score in observational studies for causal effects. *Biometrika* 70, 41–55.
- Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* 66, 688–701.
- Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *Ann. Statist.* 7, 34–58.
- Rubin, D. B. (1990). Neyman (1923) and causal inference in experiments and observational studies. *Statist. Sci.* 5, 472–80.
- Simon, H. A. (1953). Causal ordering and identifiability. In *Studies in Econometric Method*, Ed. W. C. Hood and T. C. Hoopmans, Ch. 3. New York: John Wiley.
- Sobel, M. E. (1990). Effect analysis and causation in linear structural equation models. *Psychometrika* 55, 495–515.
- Spiegelhalter, D. J., Lauritzen, S. L., Dawid, A. P. & Cowell, R. G. (1993). Bayesian analysis in expert systems (with Discussion). *Statist. Sci.* 8, 219–47.
- Spirtes, P. (1995). Directed cyclic graphical representations of feedback models. In *Uncertainty in Artificial Intelligence 11*, Ed. P. Besnard and S. Hanks, pp. 491–98. San Mateo, CA: Morgan Kaufmann.
- Spirtes, P., Glymour, C. & Scheines, R. (1993). *Causation, Prediction, and Search*. New York: Springer-Verlag.
- Strotz, R. H. & Wold, H. O. A. (1960). Recursive versus nonrecursive systems: An attempt at synthesis. *Econometrica* 28, 417–27.
- Wainer, H. (1989). Eelworms, bullet holes, and Geraldine Ferraro: Some problems with statistical adjustment and some solutions. *J. Educ. Statist.* 14, 121–40.
- Wermuth, N. (1992). On block-recursive regression equations (with Discussion). *Brazilian J. Prob. Statist.* 6, 1–56.
- Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics*. Chichester: John Wiley.
- Wright, S. (1921). Correlation and causation. *J. Agric. Res.* 20, 557–85.

[Received May 1994. Revised February 1995]

18.1 Discussion of ‘Causal Diagrams for Empirical Research’ by J. Pearl

D. R. Cox* and Nanny Wermuth†

Judea Pearl has provided a general formulation for uncovering, under very explicit assumptions, what he calls the causal effect on y of ‘setting’ a variable x at a specified level, $\text{pr}(y|\tilde{x})$, as assessed in a system of dependencies that can be represented by a directed acyclic graph. His Theorem 18.3 then provides a powerful computational scheme.

The back-door criterion requires there to be no unobserved ‘common cause’ for x and y that is not blocked out by observed variables, that is at least one of the intermediate variables between x and y or the common cause is to be observed. It is precisely doubt about such assumptions that makes epidemiologists, for example, wisely in our view, so cautious in distinguishing risk factors from causal effects. The front-door criterion requires, first, that there be an observed variable z such that x affects y only via z . Moreover, an unobserved variable u affecting both x and y must have no direct effect on z . Situations where this could be assumed with any confidence seem likely to be exceptional.

We agree with Pearl that in interpreting a regression coefficient, or generalisation thereof, in terms of the effect on y of an intervention on x , it is crucial to specify what happens to other variables, observed and unobserved. Which are fixed, which vary essentially as in the data under analysis, which vary in some other way? If we ‘set’ diastolic blood pressure, presumably we must, at least for some purposes, also ‘set’ systolic blood pressure; and what about a host of biochemical variables whose causal interrelation with blood pressure is unclear? The difficulties here are related to those of interpreting structural equations with random terms, difficulties emphasised by Haavelmo many years ago; we cannot see that Pearl’s discussion resolves the matter.

The requirement in the standard discussion of experimental design that concomitant variables be measured before randomisation applies to their use for improving precision and detecting interaction. The use of covariates for detailed exploration of the relation between treatment effects, intermediate responses and final responses gets less attention than it deserves in the literature on design of experiments; see, however, the searching discussion in an agronomic context by

*Nuffield College

†Johannes Gutenberg-Universität Mainz

Fairfield Smith (1957). Graphical models and their consequences have much to offer here and we welcome Dr Pearl’s contribution on that account.

[Received May 1995]

18.II Discussion of ‘Causal Diagrams for Empirical Research’ by J. Pearl

A. P. Dawid*

The clarity which Pearl’s graphical account brings to the problems of describing and manipulating causal models is greatly to be welcomed. One point which deserves emphasis is the equivalence, for the purposes Pearl addresses, between the counterfactual functional representation (18.3), emphasised here, and the alternative formulation of Pearl (1993b), involving the incorporation into a ‘regular’ directed acyclic graph of additional nodes and links directly representing interventions. I must confess to a strong preference for the latter approach, which in any case is the natural framework for analysis, as is seen from the 18.A. In particular, although a counterfactual interpretation is possible, it is inessential: the important point is to represent clearly, by choice of the appropriate directed acyclic graph, the way in which an intervention set ($X = x$) disturbs the system, by specifying which conditional distributions are invariant under such an intervention. As (18.5) makes evident, the overall effect of intervention is then entirely determined by the conditional distributions describing the recursive structure, and in no way depends on the way in which these might be represented functionally as in (18.3). This is fortunate, since it is far easier to estimate conditional distributions than functional relationships.

There are contexts where distributions are not enough, and counterfactual relationships need to be assessed for valid inference. Perhaps the extension to non-recursive models mentioned in § 18.6 is one. More important is inquiry into the ‘causes of effects’, rather than the ‘effects of causes’ considered here. This arises in questions of legal liability: ‘Did Mr A’s exposure to radiation in his workplace cause his child’s leukaemia?’ Knowing that Mr A was exposed, and the child has developed leukaemia, the question requires us to assess, counterfactually, what would have happened to the child had Mr A not been exposed. For this, distributional models are insufficient: a functional or counterfactual model is essential.

*University College London

This raises the question as to how we can use scientific understanding and empirical data to construct the requisite causal model. By saying little about this specification problem, Pearl is in danger of being misunderstood to say that it is not important. To build either a distributional or a counterfactual causal model, we need to assess evidence on how interventions affect the system, and what remains unchanged. This will typically require a major scientific undertaking. Given this structure, distributional aspects can, in principle, be estimated from suitable empirical data, if only these are available, and we can then apply the manipulations described by Pearl to address problems of the ‘effects of causes’. But much more would be needed to address ‘causes of effects’, since counterfactual probabilities are, almost by definition, inaccessible to direct empirical study. Empirical data can be used to place bounds on these (Balke & Pearl, 1994), but these will usually only be useful when they essentially determine the functions in (18.3). And, for this, it will be necessary to conduct studies in which the variables ε_i are explicitly identified and observed. Thus the whole mechanism needs to be broken down into essentially deterministic sub-mechanisms, with randomness arising solely from incomplete observation. In most branches of science such a goal is quite unattainable.

I emphasise the distinction drawn above, between inference about ‘effects of causes’ and ‘causes of effects’, because it might be tempting to try to extend Pearl’s analysis, particularly in its formulation (18.3), to the latter problem. For both problems serious difficulties attend the initial model specification, but these are many orders of magnitude greater for ‘causes of effects’, and the inferences drawn will be very highly sensitive to the specification.

On a different point, I am intrigued by possible connexions between Pearl’s clear distinction between conditioning and intervening, and the prequential framework of Dawid (1984, 1991), especially as elaborated by Vovk (1993). Suppose A plays a series of games, involving coins, dice, roulette wheels, etc. At any point, the game chosen may depend on the observed history. We could model this dependence probabilistically, or leave it unspecified. Now suppose we are informed of the sequence of games actually played, and want to say something about their outcomes. In a fully probabilised model, we could condition on the games played, but this would involve unpleasant analysis, and be sensitive to assumptions. Alternatively, and seemingly very reasonably, we can use the ‘prequential model’, which treats the games as having been fixed in advance. This is obtained from a fully specified model, with its natural temporally defined causal model, by ‘setting’ the games, rather than conditioning on them.

[Received May 1995]

18.III Discussion of ‘Causal Diagrams for Empirical Research’ by J. Pearl

Stephen E. Fienberg*, Clark Glymour, and Peter Spirtes†

In recent years we have investigated the use of directed graphical models (Spirtes, 1995; Spirtes, Glymour & Scheines, 1993) in order to analyse predictions about interventions that follow from causal hypotheses. We therefore welcome Pearl’s development and exposition. Our goal here is to indicate some other virtues of the directed graph approach, and compare it to alternative formalisations.

Directed graph models have a dual role, explicitly representing substantive hypotheses about influence and implicitly representing hypotheses about conditional independence. We can connect the two dimensions, one causal and the other stochastic, by explicit mathematical axioms. For example, the causal Markov axiom requires that, in the graph, each variable be independent of its non-descendants conditional on its set of parents. The formalism allows one to hold causal hypotheses fixed while varying the axiomatic connexions to probabilistic constraints. In this way, one can prove the correctness of computable conditions for prediction, for the statistical equivalence of models, and for the possibility or impossibility of asymptotically correct model search, all under alternative axioms and under a variety of circumstances relevant to causal inference, including the presence of latent variables, sample selection bias, mixtures of causal structures, feedback, etc. Thus it is possible to derive Pearl’s Theorem 18.3, and other results in his paper, from the Markov condition alone, provided one treats a manipulation as conditionalisation on a ‘policy’ variable appropriately related to the variable manipulated. Further, two extensions of Theorem 18.3 follow fairly directly. First, if the sufficient conditions in Theorem 18.3 for the equalities of probabilities are violated, distributions satisfying the Markov condition exist for which the equalities do not hold. Secondly, if the Markov condition entails all conditional independencies holding in a distribution, an axiom sometimes called ‘faithfulness’, the conditions of Theorem 18.3 are also necessary for the equalities given there.

The graphical formalism captures many of the essential features common to statistical models that sometimes accompany causal or constitutive hypotheses, including linear and nonlinear regression, factor analysis, and both recursive and nonrecursive structural equation models. In many cases, these models are representable as graphical models with additional distribution assumptions. In some

*Carnegie Mellon University, Department of Statistics

†Carnegie Mellon University, Department of Philosophy

cases, the graphical formalism provides an alternative parametrisation of subsets of the distributions associated with a family of models, as, for example, for the graphical subset of distributions from the log-linear parametrisation of the multinomial family (Bishop, Fienberg & Holland, 1975; Whittaker, 1990). Directed graphs also offer an explicit representation of the connexion between causal hypotheses and independence and conditional independence hypotheses in experimental design, and, under various axioms, permit the mathematical investigation of relations between experimental and nonexperimental designs.

Rubin (1974), Rosenbaum & Rubin (1983), Holland (1988) and Pratt & Schlaifer (1988) have provided an important alternative treatment of the prediction of the results of interventions from partial causal knowledge. As Pearl notes, their approach, which involves conditional independence of measured and ‘counterfactual’ variables, gives results in agreement with the directed graphical approach under an assumption they refer to as ‘strong ignorability’. For example, a result given without proof by Pratt & Schlaifer provides a ‘sufficient and almost necessary’ condition for the equality of the probability of Y when X is manipulated, and the conditional probability of the counterfactual of Y on X . A direct analogue of their claim of sufficiency is provable from the Markov condition and necessity follows from the faithfulness condition, which is true with probability 1 for natural measures on linear and multinomial parameters. This offers a reasonable reconstruction of what they may have meant by ‘almost necessary’. The Rubin approach to prediction has some advantages over directed graph approaches, for example in the representation of circumstances in which features of units influence other units. The disadvantages of the framework stem from the necessity of formulating hypotheses explicitly in terms of the conditional independence of actual and counterfactual variables rather than in terms of variables directly influencing others. In our experience, even experts have difficulty reliably judging the conditional independence relations that do or do not follow from assumptions. For example, we have heard many statistically trained people deny, before doing the calculation, that the normality and independence of X , Y and e , coupled with the linear equation $Z = aX + bY + e$, entail that X and Y are dependent conditional on Z . For the same reason, the Rubin framework may make more difficult mathematical proofs of results about invariance, equivalence, search, etc.

There are at least two other alternative approaches to the graphical formalism: Robins’ (1986) G -computation algorithm for calculating the effects of interventions under causal hypotheses expressed as event trees, an extension of the Rubin approach; and Glenn Shafer’s (1996) more recent and somewhat different tree structure approach. Where both are applicable, they seem to give the same results

as do procedures Pearl describes for computing on directed graphs. An advantage of the directed graph formalism is the naturalness of the representation of influence. Questions regarding the relative power of these alternative approaches are as follows.

- (i) Is the graphical approach applicable to cases where the alternatives are not, particularly when there are structures in which it is not assumed that every variable either influences or is influenced by every other?
- (ii) Is the graphical approach faster in some instances, because the directed graphs can encode independencies in their structure while event trees cannot?
- (iii) Can the alternatives, like the graphical procedure, be extended to cases in which the distribution forced on the manipulated variable is continuous?

As far as we can tell, none of the approaches to date has been able to cope with causal language associated with explanatory variables in proportional hazards models, where the nonlinear structure does not lend itself naturally to conditional independence representations.

[Received April 1995]

18.IV Discussion of ‘Causal Diagrams for Empirical Research’ by J. Pearl

David Freedman*

Causal inference with nonexperimental data seems unjustifiable to many statisticians. For others, the trick can be done almost on a routine basis, with the help of regression and its allied techniques, like path analysis or simultaneous-equation models. However, typical regression studies are problematic, because inferences are conditional on unvalidated, even unarticulated, assumptions: for discussion and reviews of the literature, see [Freedman \(1991, 1995\)](#).

Deriving causation from association by regression depends on stochastic assumptions of the familiar kind, and on less familiar causal assumptions. Building on earlier work by [Holland \(1988\)](#) and [Robins \(1989\)](#) among others, Pearl develops a graphical language in which the causal assumptions are relatively easy to state. His formulation is both natural and interesting. It captures reasonably well

*University of California, Berkeley

one intuition behind regression analysis: causal inferences can be drawn from associational data if you are observing the results of a controlled experiment run by Nature, and the causal ordering of the variables is known. When these assumptions hold, there is identifiability theory that gives an intriguing description of permissible inferences.

Following Holland (1988), I state the causal assumptions along with statistical assumptions that, taken together, justify inference in conventional path models. There is an observational study with n subjects, $i = 1, \dots, n$. The data will be analysed by regression. There are three measured variables, X, Y, Z . The path diagram has arrows from X to Y ; then, from X and Y to Z . The diagram is interpreted as a set of assumptions about causal structure: the data result from coupling together two thought experiments, as specified below. Statistical analysis proceeds from the assumption that subjects are independent and identically distributed in certain respects. That is the basis for estimating regression functions, an issue Pearl does not address; customary tests of significance would follow too.

Random variables are represented in the usual way on a sample space Ω . With notation like Holland's, $Y_{i,x}(\omega)$ represents the Y -value for subject i at $\omega \in \Omega$, if you set the X -value to x . The thought experiments are governed by the following assumptions (18.IV.1) and (18.IV.2):

$$Y_{i,x}(\omega) = f(x) + \delta_i(\omega), \quad (18.IV.1)$$

$$Z_{i,x,y}(\omega) = g(x, y) + \varepsilon_i(\omega). \quad (18.IV.2)$$

The same f and g apply to all subjects. Additive disturbance terms help the regression functions f and g to be estimate, but more is required. Typically, linearity is assumed:

$$f(x) = a + bx, \quad g(x, y) = c + dx + ey. \quad (18.IV.3)$$

The δ 's are taken to be independent and identically distributed with mean 0 and finite variance, as are the ε 's; furthermore, the δ 's are taken as independent of the ε 's.

The experiments are coupled together to produce observables, as follows. Nature assigns X -values to the subjects at random, independently of the δ 's and ε 's. Finally, the data on subject i are modelled as

$$X_i(\omega), \quad Y_i(\omega) = f\{X_i(\omega)\} + \delta_i(\omega), \quad Z_i(\omega) = g\{X_i(\omega), Y_i(\omega)\} + \varepsilon_i(\omega).$$

Linearity of regression functions and normality of errors would be critical for small data sets; with more data, less is needed. Conditioning on the $\{X_i\}$ is a popular option.

The critical assumption is: if you intervene to set the value x for X on subject i in the first ‘experiment’, the Y -value responds according to (18.IV.1) above: the disturbance $\delta_i(\omega)$ is unaffected by intervention. If you set x and y as the values for X and Y on subject i in the second experiment, the Z -value responds according to (18.IV.2) above; again, $\varepsilon_i(\omega)$ is unaffected. In particular, the assignment by Nature of subjects to levels of X does not affect the δ ’s or ε ’s. Given this structure, the parameters a, b, c, d, e in (18.IV.1)–(18.IV.3) above can be estimated from nonexperimental data and used to predict the results of interventions: for instance, setting X to x and Y to y should make Z around $\hat{c} + \hat{d}x + \hat{e}y$.

Pearl says in § 18.6 that he gives a ‘clear empirical interpretation’ and ‘operational meaning’ to causal assumptions, and clarifies their ‘empirical basis’. There are two ways to read this:

- (i) assumptions that justify causal inference from regression have been stated quite sharply;
- (ii) feasible methods have been provided for validating these assumptions, at least in certain examples.

The first assertion seems right, indeed, that is one of the main contributions of the paper. The second reading, which is probably not the intended one, would be a considerable over-statement. Invariance of errors under hypothetical interventions is a tall order. How can we test that $Z_i(\omega)$ would have been $g(x, y) + \varepsilon_i(\omega)$ if only $X_i(\omega)$ had been set to x and $Y_i(\omega)$ to y ? What about the stochastic assumptions on δ and ε ? In the typical observational study, there is no manipulation of variables and precious little sampling. Validation of causal models remains an unresolved problem.

Pearl’s framework is more general than Equations (18.IV.1)–(18.IV.3) above, and the results are more subtle. Still, the causal laws, i.e. the analogues of the equations, are assumed rather than inferred from the data. One technical complication should be noted: in Pearl’s Equation (18.IV.3), distributions are identifiable but the ‘link functions’ f_i are not. The focus is qualitative rather than quantitative, so weaker invariance assumptions may suffice. More discussion of this point would be welcome.

Concomitants. Concomitant variables pose further difficulties (Dawid, 1979). Thus, in Equations (18.IV.1)–(18.IV.3) above, suppose X is a dummy variable for sex, Y is education and Z is income. Some would consider a counterfactual interpretation: How much would Harriet have made if she had been Harry and gone to college? Others would view X as not manipulable, even in principle. Setting a

subject's sex to M , even in a thought experiment, is then beside the point. [Robins \(1986, 1987a\)](#) offers another way to model concomitants.

Conclusions. Pearl has developed mathematical language in which causal assumptions can be discussed. The gain in clarity is appreciable. The next step must be validation: to make real progress, those assumptions have to be tested.

[Received April 1995]

18.V Discussion of 'Causal Diagrams for Empirical Research' by J. Pearl

[Guido W. Imbens and Donald B. Rubin*](#)

Judea Pearl presents a framework for deriving causal estimands using graphical models representing statements of conditional independence in models with independent and identically distributed random variables, and a 'set' notation with associated rules to reflect causal manipulations. This is an innovative contribution as it formalises the use of path-analysis diagrams for causal inference, traditionally popular in many fields including econometrics, e.g. [Goldberger \(1973\)](#). Because Pearl's technically precise framework separates issues of identification and functional form, often inextricably linked in the structural equations of literature, this paper should serve to make this extensive literature more accessible to statisticians and reduce existing confusion between statisticians and econometricians: see, e.g., the Discussion of [Wermuth \(1992\)](#).

Our discussion, however, focuses on this framework as an alternative to the practice in statistics, typically based on the potential outcomes framework for causal effects, or Rubin causal model ([Holland, 1986](#); [Rubin, 1974, 1978](#)), which itself is an extension of [Neyman's \(1923\)](#) formulation for randomised experiments as discussed by [Rubin \(1990\)](#). The success of these frameworks in defining causal estimands should be measured by their applicability and ease of formulating and assessing critical assumptions.

Much important subject-matter information is not conveniently represented by conditional independence in models with independent and identically distributed random variables. Suppose that, when a person's health status is 'good', there is no effect of a treatment on a final health outcome, but, when a person's health status is 'sick', there is an effect of this treatment, so there is dependence of final health status on treatment received conditional on initial health status.

*Harvard University

Although the available information is clearly relevant for the analysis, its incorporation, although immediate using potential outcomes, is not straightforward using graphical models.

Next, consider a two-treatment randomised experiment with imperfect compliance, so that the received treatment is, in general, not ignorable despite the ignorability of the assigned treatment. Assuming that any effect of the assigned treatment on the outcome works through the received treatment, one has the instrumental variables example in Figure 18.5(b), which excludes a direct effect of Z on Y , given X . In other work (‘Bayesian inference for causal effects in randomized experiments with noncompliance’, Working Paper 1976, Harvard Institute of Economic Research, Harvard University) we have discussed important distinctions between different versions of this exclusion restriction, which can be stated using potential outcomes but are blurred in graphical models. In that paper and related work (Imbens & Angrist, 1994; Angrist, Imbens & Rubin, 1995), we also stress the importance of the ‘monotonicity assumption’, requiring the absence of units taking the treatment if assigned to control and not taking it if assigned to treatment. This allows identification of the average effect of the treatment for the subpopulation of compliers without assuming a common, additive, treatment effect for all units. Yet the monotonicity assumption is difficult to represent in a graphical model without expanding it beyond the representation in Figure 18.5(b).

Complications also arise in Pearl’s framework when attempting to represent standard experimental designs (Cochran & Cox, 1957) having clustering of units in nests, split-plot randomisations, carryover treatments, etc.

A related reason for preferring the Rubin causal model is its explicit distinction between assignment mechanisms, which are often to some extent under the investigator’s control even in observational studies, and scientific models underlying the data, which are not. Consider the discussion of the equivalence of the Rosenbaum–Rubin condition of strong ignorability of the assignment mechanism and the back-door criterion. In general, the concept of ignorable assignment (Rubin, 1976, 1978) does not require the conditional independence used in Pearl’s analysis. For example, a sequential assignment mechanism with future treatment assignments dependent on observed outcomes of previous units is ignorable, but such an assignment mechanism apparently requires a very large graphical model with all units defined as separate nodes, thereby making Pearl’s results, which require ‘an arbitrary large sample randomly drawn from the joint distribution’, irrelevant.

Finally, consider the smoking–tar–cancer example in Figures 18.3, 18.4 and 18.6(e). Pearl claims that his analysis reveals that one can learn the effect of one’s smoking on one’s lung cancer from observational data, the only provision being

that ‘smoking does not have any direct effect on lung cancer except that mediated by tar deposits’, i.e. no direct arrow from X to Y . But this claim is misleading as there are many other provisions hidden by the lack of an arrow between X and Z . For example, suppose that smokers are more likely to live in cities, and therefore more likely to be exposed to tar through pollution, or that smokers are more likely to interact with smokers, and are therefore exposed to more second-hand smoke than nonsmokers, etc. In this example the role of ‘tar deposits’ as an outcome is confounded with its role as a cause whose assignment may partially depend on previous treatments and outcomes, as can occur in serial experiments (Herzberg & Cox, 1969).

Our overall view of Pearl’s framework is summarised by Hill’s concluding sentence (1971, p. 296). ‘Technical skills, like fire, can be an admirable servant and a dangerous master’. We feel that Pearl’s methods, although formidable tools for manipulating directed acyclical graphs, can easily lull the researcher into a false sense of confidence in the resulting causal conclusions. Consequently, until we see convincing applications of Pearl’s approach to substantive questions, we remain somewhat sceptical about its general applicability as a conceptual framework for causal inference in practice.

[Received April 1995]

18.VI Discussion of ‘Causal Diagrams for Empirical Research’ by J. Pearl

James M. Robins*

18.VI.A Introduction

Pearl has carried out two tasks. In the first, in § 18.2, using some results of Spirtes, Glymour & Scheines (1993), he showed that a nonparametric structural equations model depicted as a directed acyclic graph G implies that the causal effect of any variables $X \subseteq G$ on $Y \subseteq G$ is a functional of (i) the distribution function P_G of the variables in G , and (ii) the partial ordering of these variables induced by the directed graph. This functional is the g -computation algorithm functional, hereafter g -functional, of Robins (1986, p. 1423). In the second, in §§ 18.3–18.5, only a subset of the variables in G is observed. Given known conditional independence restrictions on P_G encoded as missing arrows on G , Pearl develops elegant graphical inference rules for determining identifiability of the g -functional from the law of the observed subset. Task 2 requires no reference to structural models or

*Harvard School of Public Health

to causality. A potential problem with Pearl’s formulation is that his structural model implies that all variables in G , including concomitants such as age or sex, are potentially manipulable. Below I describe a less restrictive model that avoids this problem but, when true, still implies that the g -functional equals the effect of the treatments X of interest on Y . This critique of Pearl’s structural model is unconnected with his graphical inference rules, which were his main focus and are remarkable and path-breaking, going far beyond my own and others’ results (Robins, 1986, § 8, Appendix F).

18.VI.B Task 1

18.VI.B.1 General

Robins (1986, 1987b) proposed a set of counterfactual causal models based on event trees, called causally interpreted structured tree graphs, hereafter causal graphs, that includes Pearl’s non-parametric structural equations model as a special case. These models extended Rubin’s (1978) ‘time-independent treatment’ model to studies with direct and indirect effects and time-varying treatments, concomitants, and outcomes. In this section, I describe some of these models.

18.VI.B.2 A Causal Model

Let $V_i = \{V_{1i}, \dots, V_{Mi}\}$ denote a set of temporally-ordered discrete random variables observed on the i th study subject, $i = 1, \dots, n$. Let $X_i := \{X_{1i}, \dots, X_{Ki}\} \subseteq V_i$ be temporally-ordered, potentially manipulable, treatment variables of interest. The effect of X_i on outcomes $Y_i \subseteq V_i \setminus X_i$ is defined to be $\text{pr}\{Y_i(x) = y\}$, where the counterfactual random variable $Y_i(x)$ denotes a subject’s Y value had all n subjects followed the generalised treatment regime $g = x := \{x_1, \dots, x_K\}$. Robins (1986) wrote $\text{pr}\{Y_i(x) = y\}$ as $\text{pr}(y | g = x)$. Pearl substitutes $\text{pr}(y | \check{x})$. We regard the

$$\{V_i, Y_i(x); x \in \text{support of } X_i\} \quad (i = 1, \dots, n)$$

as independent and identically distributed, and henceforth suppress the i subscript.

This formal set-up can accommodate a superpopulation model with deterministic outcomes and counterfactuals as did that of Rubin (1978). Suppose we regard the n study subjects as randomly sampled without replacement from a large superpopulation of N subjects, and our interest is in the causal effect of X on Y in the superpopulation. Then, even if for each superpopulation member, V and $Y(x)$ were deterministic nonrandom quantities, nonetheless, in the limit as $N \rightarrow \infty$ and $n/N \rightarrow 0$, we can model the data on the n study subjects as independent and identically distributed draws from the empirical distribution of the superpopulation.

We now show that $\text{pr}(y|g = x)$ is identified from the law of V if each component X_k of X is assigned at random given the past. Let L_k be the variables occurring between X_{k-1} and X_k , with L_1 being the variables preceding X_1 . Write $\bar{L}_k := (L_1, \dots, L_k)$, $L := \bar{L}_k$ and $\bar{X}_k := (X_1, \dots, X_k)$, and define $\bar{X}_0 = \bar{L}_0$ and \bar{V}_0 to be identically 0. In considering Task 1 I have proved the following (Robins, 1987b, Theorem AD.1 and its corollary).

Theorem *If, in Dawid's (1979) conditional independence notation, for all k ,*

$$Y(x) \perp\!\!\!\perp X_k \mid \bar{L}_k, \bar{X}_{k-1} = \bar{x}_{k-1}, \quad (18.VI.1)$$

$$X = x \Rightarrow Y(x) = Y, \quad (18.VI.2)$$

$$\text{pr}(X_k = x_k \mid \bar{X}_{k-1} = \bar{x}_{k-1}, \bar{L}_k) \neq 0, \quad (18.VI.3)$$

then

$$\text{pr}(y|g = x) = h(y|g = x), \quad (18.VI.4)$$

where

$$h(y|g = x) := \sum_{\bar{l}_k} \text{pr}(y \mid \bar{l}_k, \bar{x}_k) \prod_{k=1}^K \text{pr}(l_k \mid \bar{l}_{k-1}, \bar{x}_{k-1})$$

is the g -functional for x on y based on covariates L . If X is univariate,

$$h(y|g = x) = \sum_{l_1} \text{pr}(y \mid x, l_1) \text{pr}(l_1)$$

(Rosenbaum & Rubin, 1983).

Following Robins (1987b, p. 327), I shall refer to V as a $R(Y, g = x)$ causal graph whenever (18.VI.1) and (18.VI.2) above hold, where $R(Y, g = x)$ stands for 'randomised with respect to Y for treatment $g = x$ given covariates L '. Robins et al. (1992) called (18.VI.1) the assumption of no unmeasured confounders given L . Under the aforementioned superpopulation model, (18.VI.1) will hold in a true sequential randomised trial with X randomised and X_k -specific randomisation probabilities that depend only on the past $(\bar{L}_k, \bar{X}_{k-1})$. In observational studies, (18.VI.1) is untestable; investigators can at best hope to identify covariates L so that (18.VI.1) is approximately true. Equation (18.VI.2) is Rubin's (1978) stable unit treatment value assumption: it says Y and $Y(x)$ are equal for subjects with $X = x$, irrespective of other subjects' X values. Robins (1993) shows that

$$h(y|g = x) = E \left\{ I(X = x) I(Y = y) \left/ \prod_{k=1}^K \text{pr}(x_k \mid \bar{x}_{k-1}, \bar{L}_k) \right. \right\},$$

whose denominator clarifies the need for (18.VI.3). See also Rosenbaum & Rubin (1983).

18.VI.B.3 Relationship with Pearl’s Work

Suppose we represent our ordered variables $V = \{V_1, \dots, V_M\}$ by a directed acyclic graph G that has no missing arrows, so that $\bar{V}_{m-1} := \{V_1, \dots, V_{m-1}\}$ are V_m ’s parents. Then Pearl’s nonparametric structural equation model becomes

$$V_m = f_m(\bar{V}_{m-1}, \varepsilon_m), \quad (18.VI.5)$$

for $f_m(\cdot, \cdot)$ unrestricted ($m = 1, \dots, M$), and

$$\varepsilon_m \quad (1 \leq m \leq M) \quad (18.VI.6)$$

are jointly independent.

Pearl’s assumption of missing arrows on G is (i) more restrictive than (18.VI.5), and (ii) only relevant when faced with unobserved variables, as in Task 2. We now establish the equivalence between model (18.VI.5)–(18.VI.6) above and a particular causal graph, the finest fully randomised causal graph. For any $X \subset V$, $x \in \text{support } X$, let the counterfactual random variable $V_m(x)$ denote the value of V_m had X been manipulated to x .

Definitions 18.VI.1 Robins, 1986, pp. 1421–2

(a) We have that V is a finest causal graph if (i) all one-step ahead counterfactuals $V_m(\bar{v}_{m-1})$ exist, and (ii) V and the counterfactuals $V_m(x)$ for any $X \subset V$ are obtained by recursive substitution from the $V_m(\bar{v}_{m-1})$; for example

$$V_3 \equiv V_3\{V_1, V_2(V_1)\}, \quad V_3(v_1) = V_3\{v_1, V_2(v_1)\}.$$

(b) A finest causal graph V is a finest fully randomised causal graph if, for all m ,

$$\{V_{m+1}(\bar{V}_{m-1}, v_m), \dots, V_M(\bar{V}_{m-1}, v_m, \dots, v_{M-1})\} \perp\!\!\!\perp V_m \mid \bar{V}_{m-1}. \quad (18.VI.7)$$

For V to be a finest causal graph, all variables $V_m \in V$ must be manipulable. Equation (18.VI.7) above essentially says that each V_m was assigned at random given the past \bar{V}_{m-1} . In particular, (18.VI.7) would hold in a sequential randomised trial in which all variables in V , not just the treatments X of interest, are randomly assigned given the past.

Lemma 18.VI.1 (i) Equation (18.VI.5) above is equivalent to V ’s being a finest causal graph, and (ii) Equations (18.VI.5) and (18.VI.6) above are jointly equivalent to V ’s being a finest fully randomised causal graph.

Proof of Lemma. If (18.VI.5) holds, define $V_m(\bar{v}_{m-1})$ to be $f_m(\bar{v}_{m-1}, \varepsilon_m)$. Conversely, given $V_m(\bar{v}_{m-1})$, define $\varepsilon_m = \{V_m(\bar{v}_{m-1}): \bar{v}_{m-1} \in \text{support of } \bar{V}_{m-1}\}$ and set $f_m(\bar{v}_{m-1}, \varepsilon_m) = V_m(\bar{v}_{m-1})$. Part (ii) follows by some probability calculations. ■

The statement ‘ V a finest fully randomised causal graph’ implies that V is a $R(Y, g = x)$ causal graph, and thus, given (18.VI.3) above, that $\text{pr}(y | g = x) = h(y | g = x)$. The converse is false. For example, ‘ V a $R(Y, g = x)$ causal graph’ only requires that the treatments X of interest be manipulable.

18.VI.C Task 2

Given (18.VI.1)–(18.VI.3) above, to obtain $\text{pr}(y | g = x)$, we must compute $h(y | g = x)$. However, often data cannot be collected on a subset of the covariates $L \subseteq V$ believed sufficient to make (18.VI.1) above approximately true. Given a set of correct conditional independence restrictions on the law of V , encoded as missing arrows on a directed acyclic graph G over V , Pearl provides graphical inference rules for determining whether $h(y | g = x)$ is identified from the observed data. Pearl’s graphical inference rules are correct without reference to counterfactuals or causality when we define $\text{pr}(y | \check{x}, \check{z}, w)$ to be

$$h\{y, w | g = (x, z)\} / h\{w | g = (x, z)\}.$$

Unfortunately, since covariates are missing, an investigator must rely on often shaky subject matter beliefs to guide link-deletions. Pearl & Verma (1991) appear to argue, although I would not fully agree, that beliefs about causal associations are generally sharper and more accurate than those about noncausal associations. If so, it would be advantageous to have all potential links on G represent direct causal effects, which will be the case only if V is a finest fully randomised causal graph and would justify Pearl’s focus on nonparametric structural equation models.

[Received April 1995]

18.VII Discussion of ‘Causal Diagrams for Empirical Research’ by J. Pearl

Paul R. Rosenbaum*

18.VII.A Successful and Unsuccessful Causal Inference: Some Examples

Example 18.VII.1 Cameron & Pauling (1976) gave vitamin C to patients with advanced cancers and compared survival to untreated controls. They wrote: ‘Even though no formal

*University of Pennsylvania

process of randomisation was carried out ... we believe that [treated and control groups] come close to representing random subpopulations', expressing their belief in the following diagram.

$$(\text{Treatment}) \rightarrow (\text{Survival})$$

They concluded: '... there is strong evidence that treatment ... [with vitamin C] ... increases survival time'. [Moertel et al. \(1985\)](#) repeated this in a randomised trial, but found no evidence that vitamin C prolongs survival. Today, few believe vitamin C is effective against cancer. The studies have the same path diagram, but only the randomised trial gave the correct inference.

Example 18.VII.2 The Coronary Drug Project compared lipid-lowering drugs, including clofibrate, to placebo in a randomised trial ([May et al., 1981](#)). We focus on the comparison of placebo and clofibrate. A drug can work only if consumed, yielding the following diagram.

$$(\text{Assigned clofibrate or placebo}) \rightarrow (\text{Amount of clofibrate consumed}) \rightarrow (\text{Survival})$$

In the clofibrate group, the Project found 15% mortality at five years among good compliers who took their assigned clofibrate as opposed to 25% mortality among poor compliers who did not take their assigned clofibrate. [Theorem 18.2](#) suggests clofibrate prolongs survival. Alas, it does not. In the placebo group, the mortality rates among good compliers who took their placebo was 15% compared to 28% mortality among poor compliers who did not take their placebo. Total mortality was similar in the entire clofibrate and placebo groups. Again, the nonrandomised comparison of level of clofibrate gave the wrong inference while the randomised comparison of entire clofibrate and placebo groups gave the correct inference.

[Definition 18.2](#) is not a definition of causal effect, but rather an enormous web of assumptions. It asserts that a certain mathematical operation, namely this wiping out of equations and fixing of variables, predicts a certain physical reality, namely how changes in treatments, programmes and policies will change outcomes. No basis is given for believing that physical reality behaves this way. The examples above suggest it does not. See also [Box \(1966\)](#).

18.VII.B Warranted Inferences

We do not say an inference is justified because it depends upon assumptions. We distinguish warranted and unwarranted inferences. To say, as [Fisher \(1935\)](#) said,

that randomisation is the ‘reasoned basis for inference’ is to say it warrants a particular causal inference; a warrant is a reasoned basis. An assumption is not a basis for inference unless the assumption is warranted. Path diagrams allow one to make a large number of complex, interconnected assumptions, but this is not desirable, because it is much more difficult to ensure that the assumptions are warranted.

Inferences about treatment effects can sometimes be warranted by the following methods.

- (i) Care in research design, for instance random assignment of treatments, may provide a warrant.
- (ii) Insensitivity to substantial violations of assumptions may provide a warrant. For instance, the conclusion that heavy smoking causes lung cancer is highly insensitive to the assumption that smokers are comparable to nonsmokers (Cornfield et al., 1959; Rosenbaum, 1993, 1995).
- (iii) Confirmation of numerous, elaborate predictions of a simple causal theory may at times provide a warrant. Here is Fisher’s advice, as discussed by Cochran (1965, § 5):

About 20 years ago, when asked in a meeting what can be done in observational studies to clarify the step from association to causation, Sir Ronald Fisher replied: ‘Make your theories elaborate.’ The reply puzzled me at first, since by Occam’s razor, the advice usually given is to make theories as simple as is consistent with known data. What Sir Ronald meant, as subsequent discussion showed, was that when constructing a causal hypothesis one should envisage as many different consequences of its truth as possible, and plan observational studies to discover whether each of these is found to hold.

This advice is quite the opposite of finding the conditions that just barely identify a path model. Fisher is calling for a simple theory that makes extensive, elaborate predictions each of which can be contrasted with observable data to check the theory, that is an extremely overidentified model. See Rosenbaum (1984a, 1995) for related theory and practical examples.

[Received May 1995]

18.VIII Discussion of 'Causal Diagrams for Empirical Research' by J. Pearl

Glenn Shafer*

This is an innovative and useful paper. It establishes a framework in which both probability and causality have a place, and it uses this framework to unify and extend methods of causal inference developed in several branches of statistics.

Pearl's framework is the graphical model. He brings probability and causality together by giving this model two roles: (i) it expresses a joint probability distribution for a set of variables, and (ii) it tells how interventions can change these probabilities. I find this informative and attractive. When it fits a problem, it provides a clear understanding of causality. But how often does it fit? People tend to become uncomfortable as soon as we look at almost any extensive example. Even in Pearl's own examples, it is hard to agree that each causal connection is equivalent to an opportunity for intervention, or that the simplest interventions are those that remove a particular variable from the mechanism. If we try to do something about the birds, it will surely fall short of fixing their number at a desired level, and it may have other effects on the yield of the crop.

My inability to overcome objections of these kinds when I defend causal claims made for graphical models has led me to undertake a more fundamental analysis of causality in terms of probability trees. This analysis, which will be reported in a forthcoming book (Shafer, 1996), opens the way to generalising Pearl's ideas beyond their over-reliance on the idea of intervention.

A probability tree is causal if it is nature's tree i.e. if it shows how things happen step by step in nature. Causes are represented by steps in the tree. These steps determine the overall outcome, i.e. the path nature takes through the tree, and hence every variable. Some steps identify opportunities for intervention, but others simply represent how the world works. Variables are not causes, but they can be causally related. For example, two variables are independent in the probability-tree sense if they have no common causes: there is no step in the tree where both their probability distributions change. This implies that the variables are independent in the usual sense at every point in the tree. Similarly, two numerical variables are uncorrelated in the probability-tree sense if there is no step where both their expected values change, and this implies that they are uncorrelated in the usual sense at every point in the tree.

Pearl's graphical-model assumptions, explicit and implicit, correspond to the following statements about nature's probability tree: (i) if $i < j$, then X_i is settled

*Rutgers University

before X_j , and (ii) at any point in the tree where X_{i-1} is just settled, the probability of X_i eventually coming out equal to x_i is $p(x_i | pa_i)$, where pa_i is the value of X_i 's parents. These two conditions imply that Pearl's conditional independence relations, that each variable is independent of its nondescendants given its parents, hold at every point in the tree.

What is Pearl's $p(y | \check{x}_i)$ in probability-tree terms? It is an average of probabilities: we look at each point in the tree where X_{i-1} has just been settled, find the point following where X_i is settled to have the value x_i , and find the probability at that point that Y will come out equal to y . Then we average these probabilities of y , weighting each by the probability of the point where X_{i-1} was settled. This average tells us something about how steps in the direction of x_i , after X_{i-1} is settled, tend to promote y . It has causal meaning, for it describes how the world works, but it does not depend how well steps between X_{i-1} and x_i can be targeted by human intervention.

This idea of using averages to summarise the causal effect of steps in a probability tree does not depend on Pearl's graphical-model assumptions. What is needed, in general, is some way of describing a cut across nature's tree, in addition to the event or variable that identifies the following steps whose effect we want to average. In observational studies, the cut is often specified by choosing concomitants that are just settled there. In randomised studies, it can be specified by the conditions of the experiment, without explicit measurement.

Pearl's ideas can also be used in graphical models that make weaker assumptions about nature's tree. In particular, they can be used in path models, which represent only changes in expected values, not all changes in probabilities. These models are rather more flexible than Pearl's graphical models, contrary to the suggestion conveyed by Pearl's use of the term 'nonparametric'.

[Received April 1995]

18.IX Discussion of 'Causal Diagrams for Empirical Research' by J. Pearl

Michael E. Sobel*

18.IX.A Introduction

Pearl takes the view, widely held in the social and behavioural sciences, that structural equation models are useful for estimating effects that correspond to those obtained if a randomised or conditionally randomised experiment were

*University of Arizona

conducted. Typically, linear models are used, and parameters or functions of these interpreted as unit or average effects, direct or total. A few workers argue if endogenous variables are viewed as causes, these should be treated as exogenous, and a new hypothetical system, absent equations for these variables, considered. As the effects are defined in the new system, assumptions about the relationship between parameters of the old pre-intervention and new post-intervention systems are needed (Sobel, 1990).

Pearl neatly extends this argument. His Equation (18.IX.5) specifies the post-intervention probabilities $\text{pr}(x_1, \dots, x_n | \check{x}'_i)$ in terms of the model based pre-intervention probabilities. The effects of X_i on X_j are comparisons of $\text{pr}(x_j | \check{x}'_i)$ with $\text{pr}(x_j | \check{x}^*_i)$, where x'_i and x^*_i are distinct values of X_i . If (X_1, \dots, X_n) is observed, $\text{pr}(x_j | \check{x}'_i)$ is identified; Pearl considers the nontrivial case, giving sufficient conditions for identifiability.

Pearl suggests his results are equivalent to those in Rubin’s model for causal inference. For example, he claims that the back-door criterion is ‘equivalent to the ignorability condition of Rosenbaum & Rubin (1983)’; if so, it should follow that

$$\text{pr}(x_j | \check{x}'_i, w) = \text{pr}(x_{j_{x'_i}} | w),$$

the probability if all units in the subpopulation $W = w$ take value x'_i of the cause. Thus,

$$\text{pr}(x_j | \check{x}'_i) = \text{pr}(x_{j_{x'_i}}).$$

But strong ignorability, given covariates W , implies

$$\text{pr}(x_j | x'_i, w) = \text{pr}(x_{j_{x'_i}} | w);$$

the back-door criterion implies $\text{pr}(x_j | x'_i, w) = \text{pr}(x_j | \check{x}'_i, w)$. Neither condition implies the other; the two are only equivalent if

$$\text{pr}(x_{j_{x'_i}} | w) = \text{pr}(x_j | \check{x}'_i, w).$$

The assumption

$$\text{pr}(x_{j_{x'_i}} | w) = \text{pr}(x_j | \check{x}'_i, w),$$

and others like these, is the basis on which the paper rests, and for the equivalence claims made; such quantities need not be identical.

Suppose ignorability and the back-door criterion hold above. Then

$$\text{pr}(x_{j_{x'_i}} | w) = \text{pr}(x_j | \check{x}'_i, w);$$

equality is now a conclusion. If W is observed, ignorability implies

$$\text{pr}(x_{j_{x'_i}} | w) = \text{pr}(x_j | x'_i, w),$$

which can be computed directly. The problematic case in observational studies occurs when some covariates are unobserved. But supplementing ignorability with assumptions like those in this paper helps to identify effects in Rubin's model in such cases. To simplify, only atomic interventions will be considered and the outcome treated as a scalar quantity. Only the back-door criterion is considered, but Theorem 18.2, for example, could also be handled.

18.IX.B Ignorability and the Back-Door Criterion

For an atomic intervention, rule 2, which supports the back-door criterion, is

$$\text{pr}(x_j | \check{x}'_i, w) = \text{pr}(x_j | x'_i, w) \quad (18.IX.1)$$

if $(X_j \perp\!\!\!\perp X_i | W)_{G_{\underline{X}_i}}$. Assume, following Pearl's discussion of the back-door criterion, that W is observed. Ostensibly, (18.IX.1) looks similar to the assumption of strongly ignorable treatment assignment: $X_{j_{x'_i}} \perp\!\!\!\perp X_i | W$ for all values x_i of X_i , $0 < \text{pr}(X_i = x_i | W = w)$ for all (x_i, w) .

Lemma 18.IX.1 Equality in (18.IX.1) holds if $PA_i \not\subseteq W$,

$$X_j \perp\!\!\!\perp (PA_i \setminus W) | (X_i, W). \quad (18.IX.2)$$

Proof. This follows from

$$\text{pr}(x_j, w | \check{x}'_i) = \sum_{pa_i \setminus w} \frac{\text{pr}(x_j | w, x'_i, pa_i \setminus w) \text{pr}(w, x'_i, pa_i \setminus w)}{\text{pr}(x'_i | pa_i)}. \quad (18.IX.3)$$

■

Lemma 18.IX.2 If $PA_i \not\subseteq W$, the independence conditions in (18.IX.1) and (18.IX.2) are equivalent.

Proof. Suppose (18.IX.2) holds in G . In $G_{\underline{X}_i}$, any path p from X_i to X_j has the form $X_i \leftarrow M \cdots \rightarrow X_j$, where $M \in PA_i$. If $M \in PA_i \setminus W$, the subpath $M \cdots \rightarrow X_j$ in G is blocked by W , hence p is blocked by W in $G_{\underline{X}_i}$. Otherwise, p is a subpath of $l \rightarrow X_i \leftarrow M \cdots \rightarrow X_j$ in G , $l \in PA_i \setminus W$, which is blocked by W as it has arrows converging to X_i , implying p is blocked by W in $G_{\underline{X}_i}$. Conversely, in G any path p^* from

l to X_j has form (a) $l \cdots \rightarrow X_i \rightarrow \cdots \rightarrow X_j$, (b) $l \cdots \rightarrow X_i \leftarrow M \cdots \rightarrow X_j$, or (c) $l \cdots \rightarrow X_j$, where X_i does not appear. Here X_i blocks type (a) paths. Type (b) paths contain subpaths $X_i \leftarrow M \cdots \rightarrow X_j$; by hypothesis, W blocks these in G_{X_i} , implying W blocks p^* in G . Type (c) paths are subpaths of $X_i \leftarrow l \cdots \rightarrow X_j$; by hypothesis, W blocks these in G_{X_i} , implying W blocks p^* in G . ■

Theorem *If treatment assignment is strongly ignorable, given $(W, PA_i \setminus W)$, where $PA_i \setminus W \neq \emptyset$ is unobserved, (18.IX.2) above holds, and no node in W is a descendant of X_i , $\text{pr}(x_{j_{x'_i}})$ is identified, and $\text{pr}(x_{j_{x'_i}}) = \text{pr}(x_j | \check{x}'_i)$.*

Proof. We have

$$\text{pr}(x_j | x'_i, w, pa_i \setminus w) = \text{pr}(x_{j_{x'_i}} | w, pa_i \setminus w), \quad (18.IX.4)$$

$$\text{pr}(x_j | x'_i, w, pa_i \setminus w) = \text{pr}(x_j | x'_i, w), \quad (18.IX.5)$$

where (18.IX.4) follows from strong ignorability and (18.IX.5) from (18.IX.2); jointly, (18.IX.4) and (18.IX.5) imply

$$\text{pr}(x_{j_{x'_i}} | w, pa_i \setminus w) = \text{pr}(x_{j_{x'_i}} | w) = \text{pr}(x_j | x'_i, w). \quad (18.IX.6)$$

Since no node in W is a descendant of X_i ,

$$\text{pr}(x_{j_{x'_i}}) = \sum_w \text{pr}(x_j | x'_i, w) \text{pr}(w) = \text{pr}(x_j | \check{x}'_i). \quad \blacksquare$$

[Received May 1995]

18.X Rejoinder to Discussions of ‘Causal Diagrams for Empirical Research’

Judea Pearl

18.X.A General

The subject of causality seems inevitably to provoke controversy among scientists, perhaps because causality is so essential to our thought and yet so seldom discussed in the technical literature. I am pleased, therefore, to have the opportunity to respond to a range of concerns about the usefulness of the ideas developed in my paper.

18.X.B Graphs, Structural Equations and Counterfactuals

Underlying many of the discussions are queries about the assumptions, power and limitations of the three major notational schemes used in causal analysis: structural equations, graphs and the Neyman–Rubin–Holland model, henceforth called ‘counterfactual analysis’. Thus, it seems useful to begin by explicating the commonalities among the three representational schemes, as noted in the Discussions of Freedman, following Holland (1988), Robins and Sobel; I will start with a structural interpretation of counterfactual sentences and then provide a general translation from graphs back to counterfactuals.

The primitive object of analysis in the counterfactual framework is the unit-based response variable, denoted $Y(x, u)$ or $Y_x(u)$, read: ‘the value that Y would obtain in unit u , had X been x ’. This variable has a natural interpretation in structural equations model. Consider a set T of equations

$$X_i = f_i(PA_i, U_i) \quad (i = 1, \dots, n), \quad (18.X.1)$$

where the U_i are latent exogenous variables or disturbances and the PA_i are the observed explanatory variables. Equation (18.X.1) above is similar to (18.X.3) in my paper, except we no longer insist on the equations being recursive or on the U_i ’s being independent. Let U stand for the vectors (U_1, \dots, U_n) , let X and Y be two disjoint subsets of observed variables, and let T_x be the submodel created by replacing the equations corresponding to variables in X with $X = x$, as in Definition 18.2. The structural interpretation of $Y(x, u)$ is given by

$$Y(x, u) = Y_{T_x}(u), \quad (18.X.2)$$

namely, $Y(x, u)$ is the unique solution for Y under the realisation $U = u$ in the submodel T_x of T . While the term unit in the counterfactual literature normally stands for the identity of a specific individual in a population, a unit may also be thought of as the set of attributes that characterise that individual, the experimental conditions under study, the time of day, and so on, which are represented as components of the vector u in structural modelling. Equation (18.X.2) above forms a connection between the opaque English phrase ‘the value that Y would obtain in unit u , had X been x ’ and the physical processes that transfer changes in X into changes in Y . The formation of the submodel T_x represents a minimal change in model T needed for making x and u compatible; such a change could result either from external intervention or from a natural yet unanticipated eventuality.

Given this interpretation of $Y(x, u)$, it is instructive to contrast the methodologies of causal inference in the counterfactual and the structural frameworks. If U is treated as a random variable, then the value of the counterfactual $Y(x, u)$

becomes a random variable as well, denoted by $Y(x)$ or Y_x . The counterfactual analysis proceeds by imagining the observed distribution $\text{pr}(x_1, \dots, x_n)$ as the marginal distribution of an augmented probability function pr^* defined over both observed and counterfactual variables. Queries about causal effects, written $\text{pr}(y | \check{x})$ in the structural analysis, are phrased as queries about the marginal distribution of the counterfactual variable of interest, written $\text{pr}\{Y(x) = y\}$. The new entities $Y(x)$ are treated as ordinary random variables that are connected to the observed variables via the logical constraints (Robins, 1987b)

$$X = x \Rightarrow Y(x) = Y \quad (18.X.3)$$

and a set of conditional independence assumptions which the investigator must supply to endow the augmented probability, pr^* , with causal knowledge, paralleling the knowledge that a structural analyst would encode in equations or in graphs.

For example, to communicate the understanding that in a randomised clinical trial, see Figure 18.5(b), the way subjects react, Y , to treatments X is statistically independent of the treatment assignment Z , the analyst would write $Y(x) \perp\!\!\!\perp Z$. Likewise, to convey the understanding that the assignment process is randomised, hence independent of any variation in the treatment selection process, structurally written as $U_X \perp\!\!\!\perp U_Z$, the analyst would use the independence constraint $X(z) \perp\!\!\!\perp Z$.

A collection of constraints of this type might sometimes be sufficient to permit a unique solution to the query of interest, for example, $\text{pr}\{Y(x) = y\}$; in other cases, only bounds on the solution can be obtained. Section 18.6 explains why this approach is conceptually appealing to some statisticians, even though the process of eliciting judgments about counterfactual dependencies has so far not been systematised. When counterfactual variables are not viewed as by-products of a deeper, process-based model, it is hard to ascertain whether all relevant judgments have been articulated, whether the judgments articulated are redundant, or whether those judgments are self-consistent. The elicitation of such judgments can be systematised using the following translation from graphs.

Graphs provide qualitative information about the structure of both the equations in the model and the probability function $\text{pr}(u)$. Each parent-child family (PA_i, X_i) in a causal diagram G corresponds to an equation in the model (18.X.1) above. Additionally, the absence of dashed arcs between a node Y and a set of nodes Z_1, \dots, Z_k implies that the corresponding error variables, U_Y and $\{U_{Z_1}, \dots, U_{Z_k}\}$, are independent in $\text{pr}(u)$. These assumptions can be translated into the counterfactual notation using two simple rules; the first interprets the missing arrows in the graph, the second, the missing dashed arcs.

Rule 1: Exclusion restrictions. For every variable Y having parents PA_Y , and for every set of variables S disjoint of PA_Y , we have

$$Y(pa_Y) = Y(pa_Y, s). \quad (18.X.4)$$

Rule 2: Independence restrictions. If Z_1, \dots, Z_k is any set of nodes not connected to Y via dashed arcs, we have

$$Y(pa_Y) \perp\!\!\!\perp \{Z_1(pa_{Z_1}), \dots, Z_k(pa_{Z_k})\}. \quad (18.X.5)$$

For example, the graph in Figure 18.3, displaying the parent sets

$$PA_X = \{\emptyset\}, \quad PA_Z = \{X\}, \quad PA_Y = \{Z\}$$

encodes the following assumptions.

Assumption 18.X.1 Exclusion restrictions

We require

$$\begin{aligned} X(z) &= X(y) = X(z, y) = X(\emptyset) := X, \\ Z(y, x) &= Z(x), \quad Y(z) = Y(z, x). \end{aligned}$$

Assumption 18.X.2 Independence restrictions

We require

$$Z(x) \perp\!\!\!\perp \{X, Y(z)\}.$$

While it is not easy to see that these assumptions suffice for computing the causal effect $\text{pr}\{Y(x) = y\}$ using standard probability calculus together with axiom (18.X.3) above, the identifiability of $\text{pr}(y | \check{x})$ in the diagram of Figure 18.3 ensures this sufficiency.

In summary, the structural and counterfactual frameworks are complementary to each other. Structural analysts can interpret counterfactual sentences as constraints over the solution set of a given system of Equations (18.X.2) above and, conversely, counterfactual analysts can use the constraints over pr^* given by (18.X.4) and (18.X.5) above as a definition of the graphs, structural equations and the physical processes which they represent.

18.X.C The Equivalence of Counterfactual and Structural Analyses

Robins' discussion provides a concrete demonstration of the equivalence of the counterfactual and structural definitions of causal effects, $\text{pr}\{Y(x) = y\}$ and

$\text{pr}(y | \check{x})$, respectively. Whereas (18.X.2) above explicates counterfactual sentences in terms of operations on structural equations, Robins has done the converse by explicating the assumptions of a certain structural equations model in terms of counterfactual specifications. Specifically, starting with a complete directed acyclic graph with no confounding arcs, Robins translates the assumptions of error-independence in my (18.X.3) into the ignorability-type assumptions of his (18.X.1), and shows that causal effects can be expressed in the form of the g -functional in his (18.X.4), in full conformity with the post-intervention distribution in (18.X.5) in § 18.2.2. Note that in the structural equations framework the identifiability of causal effects in model (18.X.3) in my paper is almost definitional, because the post-intervention distribution (18.X.5) in § 18.2.2 follows immediately from the definition of an atomic intervention (Definition 18.2) and from the fact that deleting equations does not change the Markovian nature of (18.X.3), and hence the product form (18.X.2) applies. What is remarkable is that Robins has derived the same expression using counterfactual analysis, which, at least on the surface, is oblivious to meta-probabilistic notions such as equation deletion or error independence.

Robins’ approach to dealing with missing links and unmeasured variables is different from mine; it follows the algebraic reduction method illustrated in § 18.3.2. After writing the g -functional using both observed and unobserved variables as in my (18.7) and (18.8), Robins would attempt to use the independencies embodied in $\text{pr}(v)$ to eliminate the unobservables from the g -formula. Because the elimination only requires knowledge of the conditional independencies embodied in $\text{pr}(v)$, any dependency-equivalent graph of G can be used or, for that matter, any nongraphical encoding of those independencies, for example, $\text{pr}(v)$ itself. The price paid for this generality is complexity: many latent variables are being summed over unnecessarily, and I am not aware of any systematic way of eliminating the latent variables from the g -expression; see the transition from my (18.8) to (18.9).

The aim of §§ 18.4 and 18.5 of my paper is to demonstrate how the derivation can be systematised and simplified by abandoning this route and resorting instead to syntactic manipulation of formulae involving observed variables only. The derivation is guided by various subgraphs of G that depend critically on the causal directionality of the arrows, hence the conditional independencies carried by G will not suffice. It is quite possible that some of these manipulations could be translated to equivalent operations on probability distributions but, if we accept the paradigm that the bulk of scientific knowledge is organised in the form of qualitative causal models rather than probability distributions, I do not see tremendous benefit in such effort.

Sobel is correct in pointing out that the equivalence of the ignorability and the back-door conditions hinges upon the equality $\text{pr}\{Y(x) = y\} = \text{pr}(y | \check{x})$. Robins’

results and the translations of (18.X.4) and (18.X.5) above provide the basis for this equality. I am puzzled, though, by Rosenbaum's astonishment at the possibility that 'a certain mathematical operation, namely this wiping out of equations ..., predicts a certain physical reality'. While it may seem odd that post-Galilean scientists habitually expect reality to obey the predictions of mathematical operations, the perfect match between mathematical predictions based on Definition 18.2 and those obtained by other, less manageable approaches reaffirms the wisdom of this expectation; the scientific basis for deleting equations is given in the paragraph preceding Definition 18.2.

18.X.D Practical Versus Hypothetical Interventions

Freedman's concern that invariance of errors under interventions may be a 'tall order' is a valid concern when addressed to practical, not to hypothetical interventions. Given a structural equation $Y = f(X, U)$, the hypothetical atomic intervention $\text{set}(X = x)$ always leaves f and U invariant, by definition; see (18.X.2) above. The crucial point is that, in order to draw valid inferences about the effect of physically fixing X at x , we must assume that our means of fixing X possesses the local property of the operator $\text{set}(X = x)$, that is it affects only the mechanism controlling X , and leaves all other mechanisms, e.g. the function f , unaltered. If current technology is such that every known method of fixing X produces side effects, then those side effects should be specified and modelled as conjunctions of several atomic interventions. Naturally, causal theories can say nothing about interventions that might break down every mechanism in the system in a manner unknown to the modeller. Causal theories are about a class of interventions that affect a select set of mechanisms in a prescribed way.

Note that this locality assumption is tacitly embodied in every counterfactual utterance as well as in the counterfactual variable $Y(x)$ used in Rubin's model. When we say 'this patient would have survived had he taken the treatment', we exclude from consideration the eventuality that the patient takes the treatment but shoots himself. It is only by virtue of this locality assumption that we can predict the effect of practical interventions, e.g. how a patient would react to the legislated treatment, from counterfactual inferences about behaviour in a given experimental study.

Freedman's difficulty with unmanipulable concomitants such as age and sex is of a slightly different nature because, here, it seems that we lack the mental capacity to imagine even hypothetical interventions that would change these variables. Remarkably, however, people do not consider common expressions such as 'If you were younger' or 'Died from old age' to be as outrageous as manipulating one's age might suggest. Why? The answer, I believe, lies in the structural equations model

of (18.X.1) and (18.X.2) above. If age X is truly nonmanipulable, then the process determining X is considered exogenous to the system and X is modelled as a component of U , or a root node in the graph. As such, no manipulation is required for envisioning the event $X = x$; we can substitute $X = x$ in U without deleting any equations from the model and obtain $\text{pr}(y|\check{x}) = \text{pr}(y|x)$ for all x and y . Additionally, in employment discrimination cases, the focus of concern is not the effect of sex on salaries but rather the effect of the employer’s awareness of the plaintiff’s sex on salary. The latter effect is manipulable, both in principle and in practice.

Shafer’s uneasiness with the manipulative account of causation also stems from taking the notion of intervention, too literally, to mean human intervention. In the process of setting up the structural equations (18.X.1) above or their graphical abstraction the analyst is instructed to imagine hypothetical interventions as defined by the submodel T_x in Definition 18.2 and Equation (18.X.2) above, regardless of their feasibility. Such thought experiments, for example slowing down the moon’s velocity and observing the effect on the tides, are feasible to anyone who possesses a model of the processes that operate in a given domain.

The analysis in my paper invokes such hypothetical local manipulations, and I mean them to be as delicate and incisive as theory will permit; it does not insist on technologically feasible manipulations which, as Shafer and Freedman point out, might cause undesired side effects. Structural equations models, counterfactual sentences, and Shafer’s probability trees all invoke the same type of hypothetical scenarios, but I find an added clarity in imagining the desired scenario as triggered by some controlled wilful act, rather than by some uncontrolled natural phenomenon, e.g. the moon hitting a comet, which might have its own, undesired side effects, e.g. the comet creating its own effects on the tides.

I agree with Shafer that not every causal thought identifies opportunities for human intervention, but I would argue strongly that every causal thought is predicated upon some notion of a ‘change’. Therefore, a theory of how mechanisms are changed, assembled, replaced and broken down, be it by humans or by Nature, is essential for causal thinking.

18.X.E Intervention as Conditionalisation

I agree with Dawid that my earlier formulation (Pearl, 1993b), which incorporates explicit policy variables in the graph and treats intervention as conditionalisation on those variables, has several advantages over the functional representation emphasised here. Fienberg, Glymour & Spirtes articulate similar sentiments. Nonetheless, I am pursuing the functional representation, partly because it is a more natural framework for thinking about data-generating processes and partly

because it facilitates the identification of ‘causes of effects’, especially in nonrecursive systems.

Balke & Pearl (1994), for example, show that sharp informative bounds on ‘causes of effects’ can sometimes be obtained without identifying the functions f_i or the variables ε_i . Additionally, if we can assume the functional form of the equations, though not their parameters, then the standard econometric conditions of parameter identification are sufficient for consistently inferring ‘causes of effects’. Balke & Pearl (1995) demonstrate how linear, nonrecursive structural models can be used to estimate the probability that ‘event $X = x$ is the cause for effect E ’, by computing the counterfactual probability that, given effect E and observations O , ‘ E would not have been realised, had X not been x ’.

18.X.F Testing Versus using Assumptions

Freedman’s concern that ‘finding the mathematical consequences of assumptions matters, but connecting assumptions to reality matters too’ has also been voiced by other discussants, most notably Dawid and Rosenbaum. Testing hypotheses against data is indeed the basis of scientific inquiry, and my paper makes no attempt to minimise the importance of such tests. However, scientific progress also demands that we not re-test or re-validate all assumptions in every study but, rather, that we facilitate the transference of knowledge from one study to another, so that the conclusions of one study may be imposed as assumptions in the next. For example, the careful empirical work of Moertel et al. (1985), which, according to Rosenbaum’s discussion, refuted the hypothesis that vitamin C is effective against cancer, should not be wasted. Instead, their results should be imposed, e.g. as a missing causal link, in future studies involving vitamin C and cancer patients, so as to enable the derivation of new causal inferences. The transference of such knowledge requires a language in which the causal relationship ‘vitamin C does not affect survival’ receives symbolic representation. Such a language, to the best of my knowledge, so far has not become part of standard statistical practice. Moreover, a language for stating assumptions is not very helpful if it is not accompanied by the mathematical machinery for quickly drawing conclusions from those assumptions or reasoning backward and isolating assumptions that need be tested, justified, or reconsidered. Facilitating such reasoning comprises the main advantage of the graphical framework.

18.X.G Causation Versus Dependence

Cox & Wermuth welcome the development of graphical models but seem reluctant to use graphs for expressing substantive causal knowledge. For example, they refer to causal diagrams as ‘a system of dependencies that can be represented by a

directed acyclic graph’. I must note that my results do not generally hold in such a system of dependencies; they hold only in systems that represent causal processes of which statistical dependencies are but a surface phenomenon. Specifically, the missing links in these systems are defined by asymmetric exclusion restrictions, as in (18.X.4) above, not by conditional independencies. The difficulties that Smith (1957) encounters in defining admissible concomitants indeed epitomise the long-standing need for precise notational distinction between causal influences and statistical dependencies.

Another type of problem created by lack of such a distinction is exemplified by Cox & Wermuth’s ‘difficulties emphasised by Haavelmo many years ago’. These ‘difficulties’ are, see Discussions following Wermuth (1992) and Cox & Wermuth (1993): (i) the term ax in the structural equation $y = ax + \varepsilon$ normally does not stand for the conditional expectation $E(Y|x)$, and (ii) variables are excluded from the equation for reasons other than conditional independence. Haavelmo (1943), who emphasises these features in the context of nonrecursive equations, is very explicit about defining structural equations in terms of hypothetical experiments and, hence, does not view the difference between ax and $E(y|x)$ as a ‘difficulty’ of interpretation but rather as an important feature of a well-interpreted model, albeit one which requires a more elaborate estimation technique than least squares. Cox & Wermuth’s difficulty stems from the reality that certain concepts in science do require both a causal and a probabilistic vocabulary. The many researchers who embrace this richer vocabulary, e.g. Haavelmo, find no difficulty with the interpretation of structural equations. I therefore concur with Imbens & Rubin’s observation that the advent of causal diagrams should promote a greater understanding between statisticians and these researchers.

18.X.H Exemplifying Modelling Errors

Rosenbaum mistakenly perceives path analysis as a competitor to randomised experiments and, in attempting to prove the former inferior, he commits precisely those errors that most path analysts have learned to avoid. After reporting a randomised study (Moertel et al., 1985) that gave different results from those of a nonrandomised study (Cameron & Pauling, 1976), he concludes that ‘the studies have the same path diagram, but only the randomised trial gave the correct inference’. However, the two studies have different path diagrams. The diagram corresponding to the randomised trial is given in Figure 18.6(a), while the diagram corresponding to the nonrandomised trial is shown in Figure 18.7(a); the former is identifiable, the latter is not. Such modelling errors do not make the diagrams the same and do not invalidate the method.

In Rosenbaum's second example, with which he attempts to refute Theorem 18.2, he again introduces an incorrect diagram. The example involves a clinical trial in which compliance was imperfect, and the diagram corresponding to such trials is shown in Figure 18.5(b). Because a confounding back-door path exists between X and Y , the conditions of Theorem 18.2 are not satisfied, and the causal effect is not identifiable: see the discussion in the second paragraph of § 18.5, and a full analysis of noncompliance given by Pearl (1995). The chain diagram chosen by Rosenbaum implies a conditional independence relation that does not hold in the data reported. Thus, Rosenbaum's attempted refutation of Theorem 18.2 is based on a convenient, but incorrect, diagram.

18.X.1 The Myth of Dangerous Graphs

Imbens & Rubin perceive two dangers in using the graphical framework: (i) graphs hide assumptions; and (ii) graphs lull researchers into a false sense of confidence.

(i) Like all abstractions, graphs make certain features explicit while keeping details implicit, to be filled in by other means if the need arises. When an independence relationship does not obtain graphical representation, the information can be filled in from the numerical probabilities, or structural equations, that annotate the links of the graph. However, a graph never fails to display a dependency if the graph modeller perceives one; see (18.X.2) of my paper. Therefore, a graph analyst is protected from reaching invalid, unintended conclusions.

Imbens & Rubin's discussion of my smoking-tar-cancer example in Figures 18.3, 18.4 and 18.6(e) illustrate this point. Contrary to their statement, the provision that tar deposits not be confounded with smoking is not hidden in the graphical representation. Rather, it stands out as vividly as can be, in the form of a missing dashed arc between X and Z . I apologise that my terse summary gave the impression that a missing link between X and Y is the 'only provision' required. From the six provisions shown in the graph, I have elected to recall this particular one, but the vividness of the graph, condition (ii) of Theorem 18.2, Equation (18.13), and the entire analysis, see also (18.X.4) and (18.X.5) above, should convince Imbens and Rubin that such provisions have not been neglected. In fact, graphs provide a powerful deterrent against forgetting assumptions unmatched by any other formalism. Every pair of nodes in the graph waves its own warning flag in front of the modeller's eyes: 'Have you neglected an arrow or a dashed arc?' I consider these warnings to be a strength, not a weakness, of the graphical framework.

(ii) Imbens & Rubin's distrust of graphs would suggest, by analogy, that it is dangerous to teach differential calculus to physics students lest they become so enchanted by the convenience of the mathematics that they overlook the assumptions. Whilst we occasionally meet discontinuous functions that do not admit the

machinery of ordinary differential calculus, this does not make the calculus useless or harmful. Additionally, I do not think over-confidence is currently holding back progress in statistical causality. On the contrary, I believe that repeated warnings against confidence are mainly responsible for the neglect of causal analysis in statistical research, and that such warnings have already done more harm to statistics than graphs could ever do.

Finally, I would like to suggest that people will be careful with their assumptions if given a language that makes those assumptions and their implications transparent; moreover, when assumptions are transparent, they are likely to be widely discussed. No matter how powerful, a notational system that does not accommodate an explicit representation of familiar processes will only inhibit people from formulating and assessing assumptions. As a result, instead of being brought into the light, critical assumptions tend to remain implicit or informal, and important problems of causal inference go unexplored. Indeed, the theory of causal inference has so far had only minor impact on rank-and-file researchers, on the methods presented in statistics textbooks, and on public policy-making. I sincerely hope graphical methods can help change this situation, both by uncovering tangible new results and by transferring causal analysis from the academic to the laboratory.

[Received June 1995]

Additional References

- Balke, A. & Pearl, J. (1995). Counterfactuals and policy analysis in structural models. In *Uncertainty in Artificial Intelligence 11*, Ed. P. Besnard and S. Hanks, pp. 11–8. San Francisco, CA: Morgan Kaufmann.
- Bishop, Y. M. M., Fienberg, S. E. & Holland, P. W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. Cambridge, MA: MIT Press.
- Box, G. E. P. (1966). The use and abuse of regression. *Technometrics* **8**, 625–9.
- Cameron, E. & Pauling, L. (1976). Supplemental ascorbate in the supportive treatment of cancer: prolongation of survival times in terminal human cancer. *Proc. Nat. Acad. Sci. (USA)*, **73**, 3685–9.
- Cochran, W. G. (1965). The planning of observational studies of human populations (with Discussion). *J. R. Statist. Soc. A* **128**, 134–55.
- Cochran, W. G. & Cox, G. M. (1957). *Experimental Designs*, 2nd ed. New York: Wiley.
- Cornfield, J., Haenszel, W., Hammond, E., Lilienfeld, A., Shimkin, M. & Wynder, E. (1959). Smoking and lung cancer: Recent evidence and a discussion of some questions. *J. Nat. Cancer Inst.* **22**, 173–203.
- Dawid, A. P. (1984). Statistical theory. The prequential approach (with Discussion). *J. R. Statist. Soc. A* **147**, 278–92.

- Dawid, A. P. (1991). Fisherian inference in likelihood and prequential frames of reference (with Discussion). *J. R. Statist. Soc. B* **53**, 79–109.
- Fairfield Smith, H. (1957). Interpretation of adjusted treatment means and regressions in analysis of covariance. *Biometrics* **13**, 282–308.
- Fisher, R. A. (1935). *The Design of Experiments*. Edinburgh: Oliver and Boyd.
- Freedman, D. (1991). Statistical models and shoe leather (with Discussion). In *Sociological Methodology 1991*, Ed. P. Marsden, Ch. 10. Washington, D.C.: American Sociological Association.
- Freedman, D. (1995). Some issues in the foundation of statistics (with Discussion). *Foundat. Sci.* **1**, 19–83.
- Goldberger, A. S. (1973). Structural equation methods in the social sciences. *Econometrica* **40**, 979–1001.
- Herzberg, A. M. & Cox, D. R. (1969). Recent work on design of experiments: A bibliography and a review. *J. R. Statist. Soc. A* **132**, 29–67.
- Hill, A. B. (1971). *A Short Textbook of Medical Statistics*, 10th ed. Place: Lippincott.
- Holland, P. (1986). Statistical and causal inference. *J. Am. Statist. Assoc.* **81**, 945–70.
- May, G., DeMets, D., Friedman, L., Furberg, C. & Passamani, E. (1981). The randomized clinical trial: Bias in analysis. *Circulation* **64**, 669–73.
- Moertel, C., Fleming, T., Creagan, E., Rubin, J., O’Connell, M. & Ames, M. (1985). High-dose vitamin C versus placebo in the treatment of patients with advanced cancer who have had no prior chemotherapy: A randomized double-blind comparison. *New Engl. J. Med.* **312**, 137–41.
- Neyman, J. (1923). On the application of probability theory to agricultural experiments. Essay on Principles, Section 9. Transl. (1990) in *Statist. Sci.* **5**, 465–80.
- Robins, J. M. (1987a). A graphical approach to the identification and estimation of causal parameters in mortality studies with sustained exposure periods. *J. Chronic Dis.* **40**, Suppl. 2, 139S–161S.
- Robins, J. M. (1987b). Addendum to ‘A new approach to causal inference in mortality studies with sustained exposure periods—application to control of the healthy worker survivor effect’. *Comp. Math. Applic.* **14**, 923–45.
- Robins, J. M. (1993). Analytic methods for estimating HIV treatment and cofactor effects. In *Methodological Issues of AIDS Mental Health Research*, Ed. D. G. Ostrow and R. Kessler, pp. 213–90. New York: Plenum.
- Rosenbaum, P. R. (1984a). From association to causation in observational studies. *J. Am. Statist. Assoc.* **79**, 41–8.
- Rosenbaum, P. R. (1993). Hodges–Lehmann point estimates of treatment effect in observational studies. *J. Am. Statist. Assoc.* **88**, 1250–3.
- Rosenbaum, P. R. (1995). *Observational Studies*. New York: Springer-Verlag.
- Rubin, D. B. (1976). Inference and missing data (with Discussion). *Biometrika* **63**, 581–92.

- Shafer, G. (1996). *The Art of Causal Conjecture*. Cambridge, MA: MIT Press.
- Smith, H. F. (1957). Interpretation of adjusted treatment means and regressions in analysis of covariates. *Biometrics* **13**, 282–308.
- Vovk, V. G. (1993). A logic of probability, with application to the foundations of statistics (with Discussion). *J. R. Statist. Soc. B* **55**, 317–51.

Probabilities of Causation: Three Counterfactual Interpretations and Their Identification*

Judea Pearl

Abstract

According to common judicial standard, judgment in favor of plaintiff should be made if and only if it is “more probable than not” that the defendant’s action was the *cause* for the plaintiff’s damage (or death). This paper provides formal semantics, based on structural models of counterfactuals, for the probability that event x was a *necessary* or *sufficient* cause (or both) of another event y . The paper then explicates conditions under which the probability of necessary (or sufficient) causation can be learned from statistical data, and shows how data from both

*I am indebted to Sander Greenland for many suggestions and discussions concerning the treatment of causation in the epidemiological literature and potential applications of this analysis in practical epidemiological studies. Donald Michie and Jack Good are responsible for shifting my attention from PN to PS and PNS. Clark Glymour and Patricia Cheng have helped unravel some of the mysteries of causal power theory, and Michelle Pearl has provided useful pointers to the epidemiological literature. This investigation was supported in part by grants from NSF, AFOSR, ONR, and California MICRO program.

Originally published in *Synthese* **121**: 93–149, 1999.

© 2000 *Kluwer Academic Publishers*. Printed in the Netherlands. Republished with permission from Kluwer.

Original DOI: [10.1023/A:1005233831499](https://doi.org/10.1023/A:1005233831499)

experimental and nonexperimental studies can be combined to yield information that neither study alone can provide. Finally, we show that necessity and sufficiency are two independent aspects of causation, and that both should be invoked in the construction of causal explanations for specific scenarios.

19.1 Introduction

The standard counterfactual definition of causation¹ (i.e., that E would not have occurred if it were not for C), captures the notion of “necessary cause”. Competing notions such as “sufficient cause” and “necessary-and-sufficient cause” may be of interest in a number of applications,² and these, too, can be given concise counterfactual definitions. One advantage of casting aspects of causation in the language of counterfactuals is that the latter enjoys natural and formal semantics in terms of structural models (Galles and Pearl 1997, 1998; Halpern 1998; Pearl 2000), as well as effective procedures for computing probabilities of counterfactual expressions from a given causal theory (Balke and Pearl 1994a, 1995). These developments are reviewed in Section 19.2.

The purpose of this paper is to explore the counterfactual interpretation of necessary and sufficient causes, to illustrate the application of structural-model semantics (of counterfactuals) to the problem of identifying probabilities of causes, and to present, by way of examples, new ways of estimating probabilities of causes from statistical data. Additionally, the paper will argue that necessity and sufficiency are two distinct facets of causation that should be kept apart in any explication of “actual cause” and, using these two facets, we will show how certain problems associated with the standard counterfactual account of causation (Lewis 1986) can be resolved.

The results have applications in epidemiology, legal reasoning, artificial intelligence (AI), and psychology. Epidemiologists have long been concerned with estimating the probability that a certain case of disease is *attributable* to a particular exposure, which is normally interpreted counterfactually as “the probability that disease would not have occurred in the absence of exposure, given that disease

1. This definition dates back to Hume (1748, 115) and Mill (1843) and has been formalized and advocated in the philosophical work of Lewis (1986).

2. The distinction between necessary and sufficient causes goes back to Mill (1843), and has received semi-formal explications in the 1960s using the syntax of conditional probabilities (Good 1961) and logical implications (Mackie 1965). The basic limitations of the logical and probabilistic accounts are discussed in Kim (1971) and Pearl (1996, 1998) and stem primarily from lacking syntactic distinction between formulas that represent stable mechanisms and those that represent transitory logical or probabilistic relationships.

and exposure did in fact occur”. This counterfactual notion, which [Robins and Greenland \(1989\)](#) called the “probability of causation” measures how *necessary* the cause is for the production of the effect.³ It is used frequently in lawsuits, where legal responsibility is at the center of contention. We shall denote this notion by the symbol PN, an acronym for Probability of Necessity.

A parallel notion of causation, capturing how *sufficient* a cause is for the production of the effect, finds applications in policy analysis, AI, and psychology. A policy maker may well be interested in the dangers that a certain exposure may present to the healthy population ([Khoury et al. 1989](#)). Counterfactually, this notion can be expressed as the “probability that a healthy unexposed individual would have gotten the disease had he/she been exposed”, and will be denoted by PS (Probability of Sufficiency). A natural extension would be to inquire for the probability of necessary-and-sufficient causation, PNS, namely, how likely a given individual is to be affected both ways.

As the examples illustrate, PS assesses the presence of an active causal process capable of producing the effect, while PN emphasizes the absence of alternative processes, not involving the cause in question, still capable of sustaining the effect. In legal settings, where the occurrence of the cause (x) and the effect (y) are fairly well established, PN is the measure that draws most attention, and the plaintiff must prove that y would not have occurred *but for* x ([Robertson 1997](#)). Still, lack of sufficiency may weaken arguments based on PN ([Good 1993](#); [Michie 1997](#)).

It is known that PN is in general non-identifiable, namely, non-estimatable from frequency data involving exposures and disease cases ([Greenland and Robins 1988](#); [Robins and Greenland 1989](#)). The identification is hindered by two factors:

1. **Confounding:** exposed and unexposed subjects may differ in several relevant factors or, more generally, the cause and the effect may both be influenced by a third factor. In this case we say that the cause is not *exogenous* relative to the effect.
2. **Sensitivity to the generative process:** Even in the absence of confounding probabilities of certain counterfactual relationships cannot be identified from frequency information unless we specify the functional relationships

3. [Greenland and Robins \(1988\)](#) further distinguish between two ways of measuring probabilities of causation: the first (called “excess fraction”) concerns only *whether* the effect (e.g., disease) occurs by a particular time, while the second, (called “etiological fraction”) requires consideration of *when* the effect occurs. We will confine our discussion here to binary events occurring within a specified time period, hence, will not be concerned with the temporal aspects of etiological fractions.

that connect causes and effects. Functional specification is needed whenever the facts at hand (e.g., disease) might be affected by the counterfactual antecedent (e.g., exposure) (Balke and Pearl 1994b) (see example in Section 19.4.1).

Although PN is not identifiable in the general case, several formulas have nevertheless been proposed to estimate attributions of various kinds in terms of frequencies obtained in epidemiological studies (Breslow and Day 1980; Hennekens and Buring 1987; Cole 1997). Naturally, any such formula must be predicated upon certain implicit assumptions about the data-generating process. This paper explicates some of those assumptions and explores conditions under which they can be relaxed.⁴ It offers new formulas for PN and PS in cases where causes are confounded (with outcomes) but their effects can nevertheless be estimated (e.g., from clinical trials or from auxiliary measurements). We further provide a general condition for the identifiability of PN and PS when functional relationships are only partially known (Section 19.5).

Glymour (1998) has raised a number of issues concerning the identifiability of causal relationships when the functional relationships among the variables *are* known, but some variables are unobserved. These issues surfaced in connection with the psychological model introduced by Cheng according to which people assess the “causal power” between two events by estimating the probability of the effect in a hypothetical model in which certain elements are suppressed (Cheng 1997). In the examples provided, Cheng’s “causal power” coincides with PS and hence lends itself to counterfactual analysis. Accordingly we shall see that many of the issues raised by Glymour can be resolved and generalized using counterfactual analysis.

The distinction between *necessary*, and *sufficient* causes has important implications in AI, especially in systems that generate verbal explanations automatically. As can be seen from the epidemiological examples above, necessary causation is a concept tailored to a specific event under consideration, while sufficient causation is based on the general tendency of certain event *types* to produce other event types. Adequate explanations should respect both aspects. If we base explanations solely on generic tendencies (i.e., *sufficient* causation), we lose important specific information. For instance, aiming a gun at and shooting a person from 1000 meters away will not qualify as an explanation for that person’s death, due to the very low

4. A set of sufficient conditions for the identification of etiological fractions are given in Robins and Greenland (1989). These conditions, however, are too restrictive for the identification of PN, which is oblivious to the temporal aspects associated with etiological fractions.

tendency of typical shots fired from such long distances to hit their marks. The fact that the shot did hit its mark on that singular day, regardless of the reason, should carry decisive weight when we come to assess whether the shooter is the culprit for the consequence. If, on the other hand, we base explanations solely on singular-event considerations (i.e., *necessary* causation), then various background factors that are normally present in the world would awkwardly qualify as explanations. For example, the presence of oxygen in the room would qualify as an explanation for the fire that broke out, simply because the fire would not have occurred were it not for the oxygen. Clearly, some balance must be made between the necessary and the sufficient components of causal explanation, and the present paper illuminates this balance by formally explicating some of the basic relationships between the two components. Section 19.6 further discusses ways of incorporating singular-event information in the definition and evaluation of sufficient causation.

19.2 Structural Model Semantics (A Review)

This section presents a brief summary of the structural-equation semantics of counterfactuals as defined in Balke and Pearl (1995), Galles and Pearl (1997, 1998), and Halpern (1998). Related approaches have been proposed in Simon and Rescher (1966), Rubin (1974) and Robins (1986). For detailed exposition of the structural account and its applications see (Pearl 2000).

19.2.1 Definitions: Causal Models, Actions and Counterfactuals

A causal model is a mathematical object that assigns truth values to sentences involving causal and counterfactual relationships. Basic of our analysis are sentences involving actions or external interventions, such as, “*p* will be true if we do *q*” where *q* is any elementary proposition. Structural models are generalizations of the structural equations used in engineering, biology, economics and social science.⁵ World knowledge is represented as a collection of stable and autonomous relationships called “mechanisms”, each represented as an equation, and changes due to interventions or hypothetical novel eventualities are treated as local modifications of those equations.

Definition 19.1 Causal model

A *causal model* is a triple

$$M = \langle U, V, F \rangle$$

5. Similar models, called “neuron diagrams” (Lewis 1986, 200; Hall 1998) are used informally by philosophers to illustrate chains of causal processors.

where

- (i) U is a set of variables called *exogenous*, that are determined by factors outside the model.
- (ii) V is a set $\{V_1, V_2, \dots, V_n\}$ of variables, called *endogenous*, that are determined by variables in the model, namely, variables in $U \cup V$
- (iii) F is a set of functions $\{f_1, f_2, \dots, f_n\}$ where each f_i is a mapping from $U \times (V \setminus V_i)$ to V_i . In other words, each f_i tells us the value of V_i given the values of all other variables in $U \cup V$. Symbolically, the set of equations F can be represented by writing

$$v_i = f_i(pa_i, u_i) \quad i = 1, \dots, n$$

where pa_i is any realization of the unique minimal set of variables PA_i in V/V_i (connoting parents) that renders f_i nontrivial. Likewise, $U_i \subseteq U$ stands for the unique minimal set of variables in U that renders f_i nontrivial.

Every causal model M can be associated with a directed graph, $G(M)$, in which each node corresponds to a variable in V and the directed edges point from members of PA_i toward V_i . We call such a graph the *causal graph* associated with M . This graph merely identifies the endogenous variables PA_i that have direct influence on each V_i but it does not specify the functional form of f_i .

Definition 19.2 Submodel

Let M be a causal model, X be a set of variables in V , and x be a particular realization of X . A submodel M_x of M is the causal model

$$M_x = \langle U, V, F_x \rangle,$$

where

$$F_x = \{f_i : V_i \notin X\} \cup \{X = x\}. \quad (19.1)$$

In words, F_x is formed by deleting from F all functions f_i corresponding to members of set X and replacing them with the set of constant functions $X = x$.

Submodels are useful for representing the effect of local actions and hypothetical changes, including those dictated by counterfactual antecedents. If we interpret each function f_i in F as an independent physical mechanism and define the action $do(X = x)$ as the minimal change in M required to make $X = x$ hold true under any u , then M_x represents the model that results from such a minimal change, since it differs from M by only those mechanisms that directly determine the variables in X . The transformation from M to M_x modifies the algebraic content of F , which

is the reason for the name *modifiable structural equations* used in (Galles and Pearl 1998).⁶

Definition 19.3 Effect of action

Let M be a causal model, X be a set of variables in V , and x be a particular realization of X . The effect of action $do(X = x)$ on M is given by the submodel M_x .

Definition 19.4 Potential response

Let Y be a variable in V , and let X be a subset of V . The *potential response* of Y to action $do(x = X)$, denoted $Y_x(u)$, is the solution for Y of the set of equations F_x .⁷

We will confine our attention to actions in the form of “ $do(X = x)$ ”. Conditional actions, of the form “ $do(X = x)$ if $Z = z$ ” can be formalized using the replacement of equations by functions of Z , rather than by constants (Pearl 1994). We will not consider disjunctive actions, of the form “ $do(X = x \text{ or } X = x')$ ” since these complicate the probabilistic treatment of counterfactuals.

Definition 19.5 Counterfactual

Let Y be a variable in V , and let X a subset of V . The counterfactual sentence “The value that Y would have obtained, had X been x ” is interpreted as denoting the potential response $Y_x(u)$.⁸

This formulation generalizes naturally to probabilistic systems, as is seen below.

Definition 19.6 Probabilistic causal model

A *probabilistic causal model* is a pair

$$\langle M, P(u) \rangle,$$

where M is a causal model and $P(u)$ is a probability function defined over the domain of U .

6. Structural modifications date back to Marschak (1950) and Simon (1953). An explicit translation of interventions into “wiping out” equations from the model was first proposed by Strotz and Wold (1960) and later used in Fisher (1970), Sobel (1990), Spirtes et al. (1993), and Pearl (1995). A similar notion of sub-model is introduced in Fine (1985), though not specifically for representing actions and counterfactuals.

7. Galles and Pearl (1998) required that F_x has a unique solution, a requirement later relaxed by Halpern (1998). In this paper we are dealing with recursive systems (i.e., $G(M)$ is a cyclic) where uniqueness of solution is ensured.

8. The connection between counterfactuals and local actions (sometimes resembling “miracles”) is made in Lewis (1986) and is further elaborated in Balke and Pearl (1994a) and Heckerman and Shachter (1995).

$P(u)$, together with the fact that each endogenous variable is a function of U , defines a probability distribution over the endogenous variables. That is, for every set of variables $Y \subseteq V$, we have

$$P(y) \triangleq P(Y = y) = \sum_{\{u \mid Y(u)=y\}} P(u). \quad (19.2)$$

The probability of counterfactual statements is defined in the same manner, through the function $Y_x(u)$ induced by the submodel M_x :

$$P(Y_x = x) = \sum_{\{u \mid Y_x(u)=x\}} P(u). \quad (19.3)$$

Likewise a causal model defines a joint distribution on counterfactual statements, i.e., $P(Y_x = y, Z_w = z)$ is defined for any sets of variables Y, X, Z, W , not necessarily disjoint. In particular, $P(Y_x = y, X = x')$ and $P(Y_x = y, Y_{x'} = y')$ are well defined for $x, \neq x'$, and are given by

$$P(Y_x = y, X = x') = \sum_{\{u \mid Y_x(u)=y \& X(u)=x'\}} P(u), \quad (19.4)$$

and

$$P(Y_x = y, Y_{x'} = y') = \sum_{\{u \mid Y_x(u)=y \& Y_{x'}(u)=y'\}} P(u). \quad (19.5)$$

When x and x' are incompatible, Y_x and $Y_{x'}$ cannot be measured simultaneously, and it may seem meaningless to attribute probability to the joint statement “ Y would be y if $X = x$ and Y would be y' if $X = x'$ ”. Such concerns have been a source of recent objections to treating counterfactuals as jointly distributed random variables (Dawid 1997). The definition of Y_x and $Y_{x'}$ in terms of two distinct submodels, driven by a standard probability space over U , explains away these objections (see Appendix 19.A) and further illustrates that joint probabilities of counterfactuals can be encoded rather parsimoniously using $P(u)$ and F .

In particular, the probabilities of causation analyzed in this paper (see Equations (19.12)–(19.14)) require the evaluation of expressions of the form $P(Y_{x'} = y' \mid X = x, Y = y)$ with x and y incompatible with x' and y' respectively. Equation (19.4) allows the evaluation of this quantity as follows:

$$\begin{aligned} P(Y_{x'} = y' \mid X = x, Y = y) &= \frac{P(Y_{x'} = y', X = x, Y = y)}{P(X = x, Y = y)} \\ &= \sum_u P(Y_{x'}(u) = y') P(u \mid x, y). \end{aligned} \quad (19.6)$$

In other words, we first update $P(u)$ to obtain $P(u | x, y)$, then we use the updated distribution $P(u | x, y)$ to compute the expectation of the index function $Y_{x'}(u) = y'$.

19.2.2 Examples

Figure 19.1 describes the causal relationships among the season of the year (X_1), whether rain falls (X_2) during the season, whether the sprinkler is on (X_3) during the season, whether the pavement is wet (X_4), and whether the pavement is slippery (X_5). All variables in this graph except the root variable X_1 take a value of either “True” or “False” (encoded “1” and “0” for convenience). X_1 takes one of four values: “Spring”, “Summer”, “Fall”, or “Winter”. Here, the absence of a direct link between, for example, X_1 and X_5 , captures our understanding that the influence of the season on the slipperiness of the pavement is mediated by other conditions (e.g., the wetness of the pavement). The corresponding model consists of five functions, each representing an autonomous mechanism:

$$\begin{aligned}x_1 &= u_1, \\x_2 &= f_2(x_1, u_2), \\x_3 &= f_3(x_1, u_3), \\x_4 &= f_4(x_3, x_2, u_4), \\x_5 &= f_5(x_4, u_5).\end{aligned}\tag{19.7}$$

The exogenous variables U_1, \dots, U_5 , represent factors omitted from the analysis. For example, U_4 may stand for (unspecified) events that would cause the pavement to get wet ($x_4 = 1$) when the sprinkler is off ($x_2 = 0$) and it does not rain ($x_3 = 0$) (e.g., a leaking water pipe). These factors are not shown explicitly in Figure 19.1 to communicate, by convention, that the U 's are assumed independent of one another. When some of these factors are judged to be dependent, it is customary to encode such dependencies by augmenting the graph with double-headed arrows (Pearl 1995).

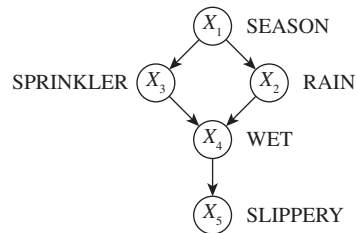


Figure 19.1 Causal graph illustrating causal relationships among five variables.

To represent the action “turning the sprinkler ON”, or $do(X_3 = \text{ON})$, we replace the equation $x_3 = f_3(x_1, u_3)$ in the model of Equation (19.7) with the equation $x_3 = 1$. The resulting submodel, $M_{X_3=\text{ON}}$, contains all the information needed for computing the effect of the action on the other variables. Note that the operation $do(X_3 = \text{ON})$ stands in marked contrast to that of *finding* the sprinkler ON; the latter involves making the substitution *without* removing the equation for X_3 , and therefore may potentially influence (the belief in) every variable in the network. In contrast, the only variables affected by the action $do(X_3 = \text{ON})$ are X_4 and X_5 , that is, the descendants of the manipulated variable X_3 . This mirrors the difference between *seeing* and *doing*: after observing that the sprinkler is ON, we wish to infer that the season is dry, that it probably did not rain, and so on; no such inferences should be drawn in evaluating the effects of the action “turning the sprinkler ON” that a person may consider taking.

This distinction obtains a vivid symbolic representation in cases where the U_i 's are assumed independent, because the joint distribution of the endogenous variables then admits the product decomposition

$$P(x_1, x_2, x_3, x_4, x_5) = P(x_1)P(x_2 | x_1)P(x_3 | x_1)P(x_4 | x_2, x_3)P(x_5 | x_4). \quad (19.8)$$

Similarly, the joint distribution associated with the submodel M_x representing the action $do(X_3 = \text{ON})$ is obtained from the product above by deleting the factor $P(x_3 | x_1)$ and substituting $x_3 = 1$.

$$P(x_1, x_2, x_4, x_5 | do(X_3 = \text{ON})) = P(x_1)P(x_2 | x_1)P(x_4 | x_2, x_3 = 1)P(x_5 | x_4). \quad (19.9)$$

The difference between the action $do(X_3 = \text{ON})$ and the observation $X_3 = \text{ON}$ is thus seen from the corresponding distributions. The former is represented by Equation (19.9), while the latter by *conditioning* Equation (19.8) on the observation, i.e.,

$$P(x_1, x_2, x_4, x_5 | X_3 = \text{ON}) = \frac{P(x_1)P(x_2 | x_1)P(x_3 = 1 | x_1)P(x_4 | x_2, x_3 = 1)P(x_5 | x_4)}{P(x_3 = 1)}.$$

Note that the conditional probabilities on the r.h.s. of Equation (19.9) are the same as those in Equation (19.8), and can therefore be estimated from pre-action observations, provided $G(M)$ is available. However, the pre-action distribution P together with the causal graph $G(M)$ is generally not sufficient for evaluating all counterfactuals sentences. For example, the probability that “the pavement would be slippery if the sprinkler were off, given that currently the pavement *is* slippery”, cannot be evaluated from the conditional probabilities $P(x_i | pa_i)$ alone;

the functional forms of the f_i 's (Equation (19.7)) are necessary for evaluating such queries (Balke and Pearl 1994b; Pearl 1996).

To illustrate the evaluation of counterfactuals, consider a deterministic version of the model given by Equation (19.7) assuming that the only uncertainty in the model lies in the identity of the season, summarized by a probability distribution $P(u_1)$ (or $P(x_1)$). We observe the ground slippery and the sprinkler on and we wish to assess the probability that the ground would be slippery had the sprinkler been off. Formally, the quantity desired is given by

$$P(X_{5_{x_3=0}} = 1 | X_5 = 1, X_3 = 1).$$

According to Equation (19.6), the expression above is evaluated by summing over all states of U that are compatible with the information at hand. In our example, the only state compatible with the evidence $X_5 = 1$ and $X_3 = 1$ is that which yields $X_1 = \text{Summer} \vee \text{Spring}$, and in this state $X_2 = \text{no-rain}$, hence $X_{5_{x_3=0}} = 0$. Thus, matching intuition, we obtain

$$P(X_{5_{x_3=0}} = 1 | X_5 = 1, X_3 = 1) = 0.$$

In general, the conditional probability of a counterfactual sentence “If it were A then B ”, given evidence e , can be computed in three steps:

1. **Abduction** – update $P(u)$ by the evidence e , to obtain $P(u | e)$.
2. **Action** – Modify M by the action $do(A)$, where A is the antecedent of the counterfactual, to obtain the submodel M_A .
3. **Deduction** – Use the updated probability $P(u | e)$ in conjunction with M_A to compute the probability of the counterfactual consequence B .

In temporal metaphors (Thomason and Gupta 1980), this 3-step procedure can be interpreted as follows: Step-1 explains the past (U) in light of the current evidence e , Step-2 bends the course of history (minimally) to comply with the hypothetical condition $X = x$ and, finally, Step-3 predicts the future (Y) based on our new understanding of the past and our new starting condition, $X = x$. Effective methods of computing probabilities of counterfactuals are presented in Balke and Pearl (1994a, 1995).

19.2.3 Relation to Lewis' Counterfactuals

The structural model of counterfactuals is closely related to Lewis's account (Lewis 1986),⁹ but differs from it in several important aspects. According to Lewis'

9. $Y_x(u) = y$ can be translated to “ $(X = x) > (Y = y)$ in world u ”.

account, one orders possible worlds by some measure of similarity, and the counterfactual $A > B$ is true in a world w just in case B is true in all the closest A -worlds to w . This semantics leaves two questions unsettled and problematic: 1. What choice of similarity measure would make counterfactual reasoning compatible with ordinary conception of cause and effect? 2. What mental representation of worlds ordering would render the computation of counterfactuals manageable and practical (in both man and machine)?¹⁰

Kit Fine's celebrated example (of Nixon pulling the trigger (Fine 1975)) demonstrates that similarity measures could not be arbitrary, but must respect our conception of causal laws.¹¹ Lewis (1979) has subsequently set up an intricate system of priorities among various dimensions of similarity: size of miracles (violations of laws), matching of facts, temporal precedence etc., to bring similarity closer to causal intuition. These difficulties do not enter the structural account. In contrast with Lewis' theory, counterfactuals are not based on an abstract notion of similarity among hypothetical worlds, but rests directly on the mechanisms (or "laws", to be fancy) that produce those worlds, and on the invariant properties of those mechanisms. Lewis' elusive "miracles" are replaced by principled mini-surgeries, $do(X = x)$, which represent the minimal change (to a model) necessary for establishing the antecedent $X = x$ (for all u). Thus, similarities and priorities, if they are ever needed, may be read into the $do(*)$ operator (see Goldszmidt and Pearl 1992), but do not govern the analysis.

The structural account answers the mental representational question by offering a parsimonious encoding of knowledge, from which causes, counterfactual and probabilities of counterfactuals can be derived by effective algorithms. This parsimony is acquired at the expense of generality; limiting the counterfactual antecedent to conjunction of elementary propositions prevents us from analyzing disjunctive hypotheticals such as "if Bizet and Verdi were compatriots".

19.2.4 Relation to Probabilistic Causality

The relation between the structural and probabilistic accounts of causality is best demonstrated when we make the Markov assumption (see Definition 19.15): 1. The equations $\{f_i\}$ are recursive (i.e., no feedback), and 2. The exogenous terms u_i are mutually independent. Under this assumption, which implies the

10. Since matching human intuition is the ultimate success criterion in most philosophical theories of causation, questions of cognitive compatibility must be considered an integral part of any such theory.

11. In this respect, Lewis' reduction of causes to counterfactuals is somewhat circular.

“screening-off” condition in the probabilistic accounts of causality, it can be shown (e.g., Pearl 1995) that the causal effect of a set X of decision variables on outcome variables Y is given by the formula:

$$P(Y = y | do(X = x)) = \sum_{pa_X} P(y | x, pa_X) P(pa_X), \quad (19.10)$$

where PA_X is the set of all parents of variables in X . Equation (19.10) calls for conditioning $P(y)$ on the event $X = x$ as well as on the parents of X , then averaging the result, weighted by the prior probabilities of those parents. This operation is known as “adjusting for PA_X ”.

Variations of this adjustment have been advanced by several philosophers as *definitions* of causality or of causal effects. Good (1961), for example, calls for conditioning on “the state of the universe just before” the occurrence of the cause. Suppes (1970) calls for conditioning on the entire past, up to the occurrence of the cause. Skyrms (1980, 133) calls for conditioning on “... maximally specific specifications of the factors outside of our influence at the time of the decision which are causally relevant to the outcome of our actions...”. The aim of conditioning in these proposals is, of course, to eliminate spurious correlations between the cause (in our case $X = x$) and the effect (in our case $Y = y$) and, clearly, the set PA_X of direct causes accomplishes this aim with great economy. However, the averaged conditionalization operation is not attached here as an add-on *adjustment*, aimed at irradiating spurious correlations. Rather, it emerges purely formally from the deeper principle of discarding the obsolete and preserving all the invariant information that the pre-action distribution can provide. Thus, while probabilistic causality first confounds causal effects $P(y | do(x))$ with epistemic conditionalization $P(y | x)$, then gets rid of spurious correlations through remedial steps of adjustment, the structural account defines causation directly in terms of Nature’s invariants (i.e., submodel M_x in Definition 19.3).

One tangible benefit of this conception is the ability to process commonplace causal statements in their natural deterministic habitat, without having to immerse them in nondeterministic decor. In other words, an event $X = x$ for which $P(x | pa_X) = 1$ (e.g., the output of a logic circuit), may still be a *cause* of some other event, $Y = y$. Consequently, probabilities of single-case causation are well defined, free of the difficulties that plague explications based on conditional probabilities. A second benefit lies in the generality of the structural equation model vis-à-vis probabilistic causality; interventions, causation and counterfactuals are well defined without invoking the Markov assumptions. Additionally, and most relevant to the topic of this paper, such ubiquitous notions as “probability of causation”

cannot easily be defined in the language of probabilistic causality (see discussion after Corollary 19.2, and Section 19.4.1).

Finally, we should note that the structural model, as it is presented in Section 19.2.1, is quasi-deterministic or Laplacian; chance arises only from unknown prior conditions as summarized in $P(u)$. Those who frown upon this classical approximation should be able to extend the results of this paper along more fashionable lines (see Appendix 19.A for an outline). However, considering that Laplace's illusion still governs human conception of cause and effect, I doubt that significant insight will be gained by such exercise.

19.2.5 Relation to Neyman–Rubin Model

Several concepts defined in Section 19.2.1 bear similarity to concepts in the potential-outcome model used by Neyman (1923) and Rubin (1974) in the statistical analysis of treatment effects. In that model, $Y_x(u)$ stands for the outcome of experimental unit u (e.g., an individual, or an agricultural lot) under experimental condition $X = x$, and is taken as a primitive, that is, as an undefined relationship, in terms of which one must express assumptions about background knowledge. In the structural model framework, the quantity $Y_x(u)$ is not a primitive, but is derived mathematically from a set of equations F that is modified by the operator $do(X = x)$. Assumptions about causal processes are expressed naturally in the form of such equations. The variable U represents any set of exogenous factors relevant to the analysis, not necessarily the identity of a specific individual in the population.

Using these semantics, it is possible to derive a complete axiomatic characterization of the constraints that govern the potential response function $Y_x(u)$ vis-à-vis those that govern directly observed variables, such as $X(u)$ and $Y(u)$ (Galles and Pearl 1998; Halpern 1998). These basic axioms include or imply relationships that were taken as given, and used extensively by statisticians who pursue the potential-outcome approach. Prominent among these we find the consistency condition (Robins 1987):

$$(X = x) \implies (Y_x = Y), \quad (19.11)$$

stating that if we intervene and set the experimental conditions $X = x$ equal to those prevailing before the intervention, we should not expect any change in the response variable Y . (For example, a subject who selects treatment $X = x$ by choice and responds with $Y = y$ would respond in exactly the same way to treatment $X = x$ under controlled experiment.) This condition is a proven theorem in structural-model semantics (Galles and Pearl 1998) and will be used in several of the

derivations of Section 19.3. Rules for translating the topology of a causal diagram into counterfactual sentences are given in (Pearl 2000, Chapter 7).

19.3 Necessary and Sufficient Causes: Conditions of Identification

19.3.1 Definitions, Notations, and Basic Relationships

Using the counterfactual notation and the structural model semantics introduced in Section 19.2.1, we give the following definitions for the three aspects of causation discussed in the introduction.

Definition 19.7 Probability of necessity (PN)

Let X and Y be two binary variables in a causal model M , let x and y stand for the propositions $X = \text{true}$ and $Y = \text{true}$, respectively, and x' and y' for their complements. The probability of necessity is defined as the expression

$$\begin{aligned} \text{PN} &\triangleq P(Y_{x'} = \text{false} \mid X = \text{true}, Y = \text{true}) \\ &\triangleq P(y'_{x'} \mid x, y). \end{aligned} \tag{19.12}$$

In other words, PN stands for the probability that event y would not have occurred in the absence of event x , ($y'_{x'}$), given that x and y did in fact occur.

Note a slight change in notation relative to that used in Section 19.2. Lower-case letters (e.g., x , y) denoted values of variables in Section 19.2, and now stand for propositions (or events). Note also the abbreviations y_x for $Y_x = \text{true}$ and y'_x for $Y_x = \text{false}$.¹² Readers accustomed to writing “ $A > B$ ” for the counterfactual “ B if it were A ” can translate Equation (19.12) to read $\text{PN} \triangleq P(x' > y' \mid x, y)$.

Definition 19.8 Probability of sufficiency (PS)

$$\text{PS} \triangleq P(y_x \mid y', x'), \tag{19.13}$$

PS measures the capacity of x to *produce* y and, since “production” implies a transition from the absence to the presence of x and y , we condition the probability $P(y_x)$ on situations where x and y are both absent. Thus, mirroring the necessity of x (as measured by PN), PS gives the probability that setting x would produce y in a situation where x and y are in fact absent.

12. These were proposed by Peyman Meshkat in class homework, and substantially simplify the derivations.

Definition 19.9 Probability of necessity and sufficiency (PNS)

$$\text{PNS} \triangleq P(y_x, y_{x'}). \quad (19.14)$$

PNS stands for the probability that y would respond to x both ways, and therefore measures both the sufficiency and necessity of x to produce y .

Associated with these three basic notions, there are other counterfactual quantities that have attracted either practical or conceptual interest. We will mention two such quantities, but will not dwell on their analyses, since these can be easily inferred from our treatment of PN, PS, and PNS.

Definition 19.10 Probability of disablement (PD)

$$\text{PD} \triangleq P(y_{x'} | y). \quad (19.15)$$

PD measures the probability that y would have been prevented if it were not for x ; it is therefore of interest to policy makers who wish to assess the social effectiveness of various prevention programs (Fleiss 1981, 75–76).

Definition 19.11 Probability of enablement (PE)

$$\text{PE} \triangleq P(y_x | y'),$$

PE is similar to PS, save for the fact that we do not condition on x' . It is applicable, for example, when we wish to assess the danger of an exposure on the entire population of healthy individuals, including those who were already exposed.

Although none of these quantities is sufficient for determining the others, they are not entirely independent, as shown in the following lemma.

Lemma 19.1 The probabilities of causation, PNS, PN and PS satisfy the following relationship:

$$\text{PNS} = P(x, y)\text{PN} + P(x', y')\text{PS}. \quad (19.16)$$

Proof of Lemma 19.1. Using the consistency conditions of Equation (19.11),

$$x \Rightarrow (y_x = y), \quad x' \Rightarrow (y_{x'} = y),$$

we can write

$$y_x \wedge y_{x'} = (y_x \wedge y_{x'}) \wedge (x \vee x') = (y \wedge x \wedge y_{x'}) \vee (y_x \wedge y' \wedge x').$$

Taking probabilities on both sides, and using the disjointness of x and x' , we obtain:

$$P(y_x, y_{x'}) = P(y_{x'}, x, y) + P(y_x, x', y') = P(y_{x'} | x, y)P(x, y) + P(y_x | x', y')P(x', y'),$$

which proves Lemma 19.1. ■

To put into focus the aspects of causation captured by PN and PS, it is helpful to characterize those changes in the causal model that would leave each of the two measures invariant. The next two lemmas show that PN is insensitive to the introduction of potential inhibitors of y , while PS is insensitive to the introduction of alternative causes of y .

Lemma 19.2 Let $\text{PN}(x, y)$ stand for the probability that x is a necessary cause of y , and $z = y \wedge q$ a consequence of y , potentially inhibited by q' . If $q \perp\!\!\!\perp \{X, Y_x, Y_{x'}\}$, then

$$\text{PN}(x, z) \triangleq P(z'_{x'} | x, z) = P(y'_{x'} | x, y) \triangleq \text{PN}(x, y).$$

Cascading the process $Y_x(u)$ with the link $z = y \wedge q$ amounts to inhibiting y with probability $P(q')$. Lemma 19.2 asserts that we can add such a link without affecting PN, as long as q is randomized. The reason is clear; conditioning on the event x and y implies that, in the scenario under consideration, the added link was not inhibited by q' .

Proof of Lemma 19.2.

$$\text{PN}(x, z) = P(z'_{x'} | x, z) = \frac{P(z'_{x'}, x, z)}{P(x, z)} = \frac{P(z'_{x'}, x, z | q)P(q) + P(z'_{x'}, x, z | q')P(q')}{P(z, x, q) + P(z, x, q')}. \quad (19.17)$$

Using $z = y \wedge q$, we have

$$q \Rightarrow (z = y), q \Rightarrow (z'_{x'} = y'_{x'}), \text{ and } q' \Rightarrow z',$$

therefore

$$\text{PN}(x, z) = \frac{P(y'_{x'}, x, y | q)P(q) + 0}{P(y, x, q) + 0} = \frac{P(y'_{x'}, x, y)}{P(y, x)} = P(y'_{x'} | xy) = \text{PN}(x, y). \quad \blacksquare$$

Lemma 19.3 Let $\text{PS}(x, y)$ stand for the probability that x is a sufficient cause of y , and let $z = y \vee r$ be a consequence of y , potentially triggered by r . If $r \perp\!\!\!\perp \{X, Y_x, Y_{x'}\}$, then

$$\text{PS}(x, z) = P(z_x | x', z') = P(y_x | x', y') = \text{PS}(x, y).$$

Lemma 19.3 asserts that we can add alternative independent causes (r), without affecting PS. The reason again is clear; conditioning on the event x' and y' implies that the added causes (r) were not active. The proof of Lemma 19.3 is similar to that of Lemma 19.2.

Definition 19.12 Identifiability

Let $Q(M)$ be any quantity defined on a causal model M . Q is *identifiable* in a class \mathbf{M} of models iff any two models M_1 and M_2 from \mathbf{M} that satisfy $P_{M_1}(v) = P_{M_2}(v)$ also satisfy $Q(M_1) = Q(M_2)$. In other words, Q is identifiable if it can be determined uniquely from the probability distribution $P(v)$ of the endogenous variables V .

The class \mathbf{M} that we will consider when discussing identifiability will be determined by assumptions that one is willing to make about the model under study. For example, if our assumptions consist of the structure of a causal graph G_0 , \mathbf{M} will consist of all models M for which $G(M) = G_0$. If, in addition to G_0 , we are also willing to make assumptions about the functional form of some mechanisms in M , \mathbf{M} will consist of all models M that incorporate those mechanisms, and so on.

Since all the causal measures defined above invoke conditionalization on y , and since y is presumed affected by x , the antecedent of the counterfactual y_x , we know that none of these quantities is identifiable from knowledge of the structure $G(M)$ and the data $P(v)$ alone, even under condition of no confounding. Moreover, none of these quantities determines the others in the general case. However, simple interrelationships and useful bounds can be derived for these quantities under the assumption of no-confounding, an assumption that we call *exogeneity*.

19.3.2 Bounds and Basic Relationships under Exogeneity**Definition 19.13 Exogeneity**

A variable X is said to be exogenous relative to Y in model M iff

$$P(y_x, y_{x'} | x) = P(y_x, y_{x'}), \quad (19.18)$$

namely, the way Y would potentially respond to conditions x or x' is independent of the actual value of X .

Equation (19.18) has been given a variety of (equivalent) definitions and interpretations. Epidemiologists refer to this condition as “no-confounding” (Robins and Greenland 1989), statisticians call it “as if randomized”, and Rosenbaum and Rubin (1983) call it “ignorability”. A graphical criterion ensuring exogeneity is the absence of a common ancestor of X and Y in $G(M)$. The classical econometric criterion for exogeneity (e.g., Dhrymes 1970, 169) states that X be independent of the error term in the equation for Y .¹³

13. This criterion has been the subject of relentless objections by modern econometricians (Engle et al. 1983; Hendry 1995; Imbens 1997), but see Aldrich (1993) and Galles and Pearl (1998) for a reconciliatory perspective on this controversy.

The importance of exogeneity lies in permitting the identification of $P(y_x)$, the *causal effect* of X on Y , since (using $x \Rightarrow (y_x = y)$)

$$P(y_x) = P(y_x | x) = P(y | x), \quad (19.19)$$

with similar reduction for $P(y_{x'})$.

Theorem 19.1 Under condition of exogeneity, PNS is bounded as follows:

$$\begin{aligned} & \max[0, P(y | x) + P(y' | x') - 1] \\ & \leq \text{PNS} \leq \min[P(y | x), P(y' | x')]. \end{aligned} \quad (19.20)$$

Both bounds are sharp in the sense that for every joint distribution $P(x, y)$ there exists a model $y = f(x, u)$, with u independent of x , that realizes any value of PNS permitted by the bounds.

Proof of Theorem 19.1. For any two events A and B we have tight bounds:

$$\max[0, P(A) + P(B) - 1] \leq P(A, B) \leq \min[P(A), P(B)]. \quad (19.21)$$

Equation (19.20) follows from (19.21) using $A = y_x$, $B = y'_{x'}$, $P(y_x) = P(y | x)$ and $P(y'_{x'}) = P(y' | x')$. ■

Clearly, if exogeneity cannot be ascertained, then PNS is bound by inequalities identical to those of Equation (19.20), with $P(y_x)$ and $P(y'_{x'})$ replacing $P(y | x)$ and $P(y' | x')$, respectively.

Theorem 19.2 Under condition of exogeneity, the probabilities PN, PS, and PNS are related to each other as follows:

$$\text{PN} = \frac{\text{PNS}}{P(y | x)}, \quad (19.22)$$

$$\text{PS} = \frac{\text{PNS}}{1 - P(y | x')}. \quad (19.23)$$

Thus, the bounds for PNS in Equation (19.20) provide corresponding bounds for PN and PS.

The resulting bounds for PN

$$\begin{aligned} & \frac{\max[0, P(y | x) + P(y' | x') - 1]}{P(y | x)} \\ & \leq \text{PN} \leq \frac{\min[P(y | x), P(y' | x')]}{P(y | x)}, \end{aligned} \quad (19.24)$$

have significant implications relative to both our ability to identify PN by experimental studies and the feasibility of defining PN in stochastic causal models. Replacing the conditional probabilities with causal effects (licensed by exogeneity), Equation (19.24) implies the following:

Corollary 19.2 Let $P(y_x)$ and $P(y'_{x'})$ be the causal effects established in an experimental study. For any point p in the range

$$\frac{\max[0, P(y_x) + P(y'_{x'}) - 1]}{P(y_x)} \leq p \leq \frac{\min[P(y_x), P(y'_{x'})]}{P(y_x)}, \quad (19.25)$$

we can find a causal model M that agrees with $P(y_x)$ and $P(y'_{x'})$ and for which $\text{PN} = p$.

This corollary implies that probabilities of causation cannot be defined uniquely in stochastic (non-Laplacian) models where, for each u , $Y_x(u)$ is specified in probability $P(Y_x(u) = y)$ instead of a single number.¹⁴ (See Example 1, Section 19.4.1.)

Proof of Theorem 19.2. Using $x \Rightarrow (y_x = y)$, we can write $x \wedge y_x = x \wedge y$, and obtain

$$\text{PN} = P(y'_{x'} | x, y) = P(y'_{x'}, x, y) / P(x, y), \quad (19.26)$$

$$= P(y'_{x'}, x, y_x) / P(x, y), \quad (19.27)$$

$$= P(y'_{x'}, y_x) P(x) / P(x, y), \quad (19.28)$$

$$= \frac{\text{PNS}}{P(y | x)}, \quad (19.29)$$

which establishes Equation (19.22). Equation (19.23) follows by identical steps. ■

For completion, we note the relationship between PNS and the probabilities of enablement and disablement:

$$\text{PD} = \frac{P(x)\text{PNS}}{P(y)}, \quad \text{PE} = \frac{P(x')\text{PNS}}{P(y')}. \quad (19.30)$$

19.3.3 Identifiability under Monotonicity and Exogeneity

Before attacking the general problem of identifying the counterfactual quantities in Equations (19.12)–(19.14) it is instructive to treat a special condition, called

14. [Robins and Greenland \(1989\)](#), who used a stochastic model of $Y_x(u)$, defined the probability of causation as

$$\text{PN}(u) = [P(y | x, u) - P(y | x', u)] / P(y | x, u),$$

instead of the counterfactual definition in Equation 19.12.

monotonicity, which is often assumed in practice, and which renders these quantities identifiable. The resulting probabilistic expressions will be recognized as familiar measures of causation that often appear in the literature.

Definition 19.14 Monotonicity

A variable Y is said to be monotonic relative to variable X in a causal model M iff the junction $Y_x(u)$ is monotonic in x for all u . Equivalently, Y is monotonic relative to X iff

$$y'_x \wedge y_{x'} = \text{false}. \quad (19.31)$$

Monotonicity expresses the assumption that a change from $X = \text{false}$ to $X = \text{true}$ cannot, under any circumstance make Y change from *true* to *false*.¹⁵ In epidemiology, this assumption is often expressed as “no prevention”, that is, no individual in the population can be helped by exposure to the risk factor. Angrist et al. (1996) used this assumption to identify treatment effects from studies involving non-compliance (see also Balke and Pearl (1997)). Glymour (1998) and Cheng (1997) resort to this assumption in using disjunctive or conjunctive relationships between causes and effects, excluding functions such as exclusive-or, or parity.

Theorem 19.3 Identifiability under exogeneity and monotonicity

If X is exogenous and Y is monotonic relative to X , then the probabilities PN, PS, and PNS are all identifiable, and are given by Equations (19.22)–(19.23) with

$$\text{PNS} = P(y|x) - P(y|x'). \quad (19.32)$$

The r.h.s. of Equation (19.32) is called “risk-difference” in epidemiology, and is also misnomered “attributable risk” (Hennekens and Buring 1987, 87).

From Equation (19.22) we see that the probability of necessity, PN, is identifiable and given by the *excess-risk-ratio*

$$\text{PN} = [P(y|x) - P(y|x')]/P(y|x), \quad (19.33)$$

often misnomered as the *attributable fraction* (Schlesselman 1982), *attributable-rate percent* (Hennekens and Buring 1987, 88), *attributed fraction for the exposed* (Kelsey et al. 1987, 38), or *attributable proportion* (Cole 1997). Taken literally, the ratio presented in (19.33) has nothing to do with attribution, since it is made up of statistical

15. Our analysis remains invariant to complementing x or y (or both), hence, the general condition of monotonicity should read: either $y'_x \wedge y_{x'} = \text{false}$ or $y'_{x'} \wedge y_x = \text{false}$. For simplicity, however, we will adhere to the definition in Equation (19.31).

terms and not of causal or counterfactual relationships. However, the assumptions of exogeneity and monotonicity together enable us to translate the notion of attribution embedded in the definition of PN (Equation (19.12)) into a ratio of purely statistical associations. This suggests that exogeneity and monotonicity were tacitly assumed by authors who proposed or derived Equation (19.33) as a measure for the “fraction of exposed cases that are attributable to the exposure”.

Robins and Greenland (1989) have analyzed the identification of PN under the assumption of stochastic monotonicity (i.e., $P(Y_x(u) = y) > P(Y_{x'}(u) = y)$) and have shown that this assumption is too weak to permit such identification; in fact, it yields the same bounds as in Equation (19.24). This indicates that stochastic monotonicity imposes no constraints whatsoever on the functional mechanisms that mediate between X and Y .

The expression for PS (Equation (19.23)), is likewise quite revealing

$$PS = [P(y|x) - P(y|x')]/[1 - P(y|x')], \quad (19.34)$$

as it coincides with what epidemiologists call the “relative difference” (Shep 1958), which is used to measure the *susceptibility* of a population to a risk factor x . Susceptibility is defined as the proportion of persons who possess “an underlying factor sufficient to make a person contract a disease following exposure” (Khoury et al. 1989). PS offers a formal counterfactual interpretation of susceptibility, which sharpens this definition and renders susceptibility amenable to systematic analysis. Khoury et al. (1989) have recognized that susceptibility in general is not identifiable, and have derived Equation (19.34) by making three assumptions: no confounding, monotonicity,¹⁶ and independence (i.e., assuming that susceptibility to exposure is independent of susceptibility to background not involving exposure). This last assumption is often criticized as untenable, and Theorem 19.3 assures us that independence is in fact unnecessary; Equation (19.34) attains its validity through exogeneity and monotonicity alone.

Equation (19.34) also coincides with what Cheng calls “causal power” (1997), namely, the effect of x on y after suppressing “all other causes of y ”. The counterfactual definition of PS, $P(y_x|x', y')$, suggests another interpretation of this quantity. It measures the probability that setting x would produce y in a situation where x and y are in fact absent. Conditioning on y' amounts to selecting (or hypothesizing) only those worlds in which “all other causes of y ” are indeed suppressed.

16. Monotonicity is not mentioned in (Khoury et al. 1989), but it must have been assumed implicitly to make their derivations valid.

It is important to note, however, that the simple relationships among the three notions of causation (Equations (19.22)–(19.23)) only hold under the assumption of exogeneity; the weaker relationship of Equation (19.16) prevails in the general, non-exogenous case. Additionally, all these notions of causation are defined in terms of the global relationships $Y_x(u)$ and $Y_{x'}(u)$ which is too crude to fully characterize the many nuances of causation; the detailed structure of the causal model leading from X to Y is often needed to explicate more refined notions, such as “actual cause” (see Section 19.6).

Proof of Theorem 19.3. Writing $y_{x'} \vee y'_x = \text{true}$, we have

$$y_x = y_x \wedge (y_{x'} \vee y'_x) = (y_x \wedge y_{x'}) \vee (y_x \wedge y'_x), \quad (19.35)$$

and

$$y_{x'} = y_{x'} \wedge (y_x \vee y'_x) = (y_{x'} \wedge y_x) \vee (y_{x'} \wedge y'_x) = y_{x'} \wedge y_x, \quad (19.36)$$

since monotonicity entails $y_{x'} \wedge y'_x = \text{false}$. Substituting Equation (19.36) into Equation (19.35) yields

$$y_x = y_{x'} \vee (y_x \wedge y'_x). \quad (19.37)$$

Taking the probability of Equation (19.37), and using the disjointness of $y_{x'}$ and y'_x , we obtain

$$P(y_x) = P(y_{x'}) + P(y_x, y'_x),$$

or

$$P(y_x, y'_x) = P(y_x) - P(y_{x'}). \quad (19.38)$$

Equation (19.38) together with the assumption of exogeneity (Equation (19.19)) establish Equation (19.32). ■

19.3.4 Identifiability under Monotonicity and Non-Exogeneity

The relations established in Theorems 19.1–19.3 were based on the assumption of exogeneity. In this section, we relax this assumption and consider cases where the effect of X on Y is confounded, i.e., $P(y_x) \neq P(y|x)$. In such cases $P(y_x)$ may still be estimated by auxiliary means (e.g., through adjustment of certain covariates, or through experimental studies) and the question is whether this added information can render the probability of causation identifiable. The answer is affirmative.

Theorem 19.4 If Y is monotonic relative to X , then PNS, PN, PS are identifiable whenever the causal effect $P(y_x)$ is identifiable and are given by

$$\text{PNS} = P(y_x, y_{x'}) = P(y_x) - P(y_{x'}), \quad (19.39)$$

$$\text{PN} = P(y_{x'} | x, y) = \frac{P(y) - P(y_{x'})}{P(x, y)}, \quad (19.40)$$

$$\text{PS} = P(y_x | x', y) = \frac{P(y_x) - P(y)}{P(x', y)}. \quad (19.41)$$

To appreciate the difference between Equations (19.40) and (19.33) we can expand $P(y)$ and write

$$\begin{aligned} \text{PN} &= \frac{P(y|x)P(x) + P(y|x')P(x') - P(y_{x'})}{P(y|x)P(x)} \\ &= \frac{P(y|x) - P(y|x')}{P(y|x)} + \frac{P(y|x') - P(y_{x'})}{P(x, y)}. \end{aligned} \quad (19.42)$$

The first term on the r.h.s. of Equation (19.42) is the familiar excess-risk-ratio as in Equation (19.33), and represents the value of PN under exogeneity. The second term represents the correction needed to account for X 's non-exogeneity, i.e., $P(y_{x'}) \neq P(y|x')$.

Equations (19.39)–(19.41) thus provide more refined measures of causation, which can be used in situations where the causal effect $P(y_x)$ can be identified through auxiliary means (see Example 4, Section 19.4.4). Note however that these measures are no longer governed by the simple relationships given in Equations (19.22)–(19.23). Instead, the governing relation is Equation (19.16).

Remarkably, since PS and PN must be non-negative, Equations (19.40)–(19.41) provide a simple necessary test for the assumption of monotonicity

$$P(y_x) \geq P(y) \geq P(y_{x'}), \quad (19.43)$$

which strengthen the standard inequalities

$$P(y_x) \geq P(x, y), P(y_{x'}) \geq P(x', y).$$

It can be shown that these inequalities are in fact sharp, that is, every combination of experimental and nonexperimental data that satisfy these inequalities can be generated from some causal model in which Y is monotonic in X . That the commonly made assumption of “no prevention” is not entirely exempt from empirical scrutiny should come as a relief to many epidemiologists. Alternatively, if the no-prevention assumption is theoretically unassailable, the inequalities of

Equation (19.43) can be used for testing the compatibility of the experimental and non-experimental data, namely, whether subjects used in clinical trials are representative of the target population, characterized by the joint distribution $P(x, y)$.

Proof of Theorem 19.4. Equation (19.39) was established in (19.38). To prove (19.41), we write

$$P(y_x | x', y') = \frac{P(y_x, x', y')}{P(x', y')} = \frac{P(y_x, x', y'_{x'})}{P(x', y')} \quad (19.44)$$

because $x' \wedge y' = x' \wedge y'_{x'}$ (by consistency). To calculate the numerator of Equation (19.44), we conjoin Equation (19.37) with x'

$$x' \wedge y_x = (x' \wedge y_{x'}) \vee (y_x \wedge y'_{x'} \wedge x'),$$

and take the probability on both sides, which gives (since $y_{x'}$ and $y'_{x'}$ are disjoint)

$$\begin{aligned} P(y_x, y'_{x'}, x') &= P(x', y_x) - P(x', y_{x'}) \\ &= P(x', y_x) - P(x', y) \\ &= P(y_x) - P(x, y_x) - P(x', y) \\ &= P(y_x) - P(x, y) - P(x', y) \\ &= P(y_x) - P(y). \end{aligned}$$

Substituting in Equation (19.44), we finally obtain

$$P(y_x | x', y') = \frac{P(y_x) - P(y)}{P(x', y)},$$

which establishes Equation (19.41). Equation (19.40) follows through identical steps. ■

One common class of models which permits the identification of $P(y_x)$ under conditions of non-exogeneity is called *Markovian*.

Definition 19.15 Markovian models

A causal model M is said to be Markovian if the graph $G(M)$ associated with M is acyclic, and if the exogenous factors u_i are mutually independent. A model is semi-Markovian iff $G(M)$ is acyclic and the exogenous variables are not necessarily independent. A causal model is said to be positive-Markovian if it is Markovian and $P(v) > 0$ for every v .

It is shown in Pearl (1993, 1995) that for every two variables, X and Y , in a positive-Markovian model M , the causal effect $P(y_x)$ is identifiable and is given by

$$P(y_x) = \sum_{pa_x} P(y|pa_x, x)P(pa_x), \quad (19.45)$$

where pa_x are (realizations of) the *parents* of X in the causal graph associate with M (see also Spirtes et al. (1993) and Robins (1986)). Thus, we can combine Equation (19.45) with Theorem 19.4 and obtain a concrete condition for the identification of the probability of causation.

Corollary 19.3 If in a positive-Markovian model M , the function $Y_x(u)$ is monotonic, then the probabilities of causation PNS, PS and PN are identifiable and are given by Equations (19.39)–(19.41), with $P(y_x)$ in Equation (19.45).

A broader identification condition can be obtained through the use of the back-door and front-door criteria (Pearl 1995), which are applicable to semi-Markovian models. These were further generalized in Galles and Pearl (1995)¹⁷ and lead to the following corollary:

Corollary 19.4 Let **GP** be the class of semi-Markovian models that satisfy the graphical criterion of Galles and Pearl (1995). If $Y_x(u)$ is monotonic, then the probabilities of causation PNS, PS and PN are identifiable in **GP** and are given by Equations (19.39)–(19.41), with $P(y_x)$ determined by the topology of $G(M)$ through the GP criterion.

19.4 Examples and Applications

19.4.1 Example 1: Betting against a Fair Coin

We must bet heads or tails on the outcome of a fair coin toss; we win a dollar if we guess correctly, lose if we don't. Suppose we bet heads and we win a dollar, without glancing at the outcome of the coin, was our bet a necessary cause (respectively, sufficient cause, or both) for winning?

Let x stand for “we bet on heads”, y for “we win a dollar”, and u for “the coin turned up heads”. The functional relationship between y , x and u is

$$y = (x \wedge u) \vee (x' \wedge u'), \quad (19.46)$$

17. Galles and Pearl (1995) provide an efficient method of deciding from the graph $G(M)$ whether $P(y_x)$ is identifiable and, if the answer is affirmative, deriving the expression for $P(y_x)$.

which is not monotonic but nevertheless permits us to compute the probabilities of causation from the basic definitions of Equations (19.12)–(19.14). To exemplify,

$$\text{PN} = P(y'_{x'} | x, y) = P(y'_{x'} | u) = 1,$$

because $x \wedge y \Rightarrow u$, and $Y_{x'}(u) = \text{false}$. In words, knowing the current bet (x) and current win (y) permits us to infer that the coin outcome must have been a head (u), from which we can further deduce that betting tails (x') instead of heads, would have resulted in a loss. Similarly,

$$\text{PS} = P(y_x | x', y') = P(y_x | u) = 1,$$

because $x' \wedge y' \Rightarrow u$, and

$$\begin{aligned} \text{PNS} &= P(y_x, y'_{x'}) \\ &= P(y_x, y'_{x'} | u)P(u) + P(y_x, y'_{x'} | u')P(u') \\ &= 1 \cdot \frac{1}{2} + 0 \cdot \frac{1}{2} = \frac{1}{2}. \end{aligned}$$

We see that betting heads has 50% chance of being a necessary-and-sufficient cause of winning. Still, once we win, we can be 100% sure that our bet was necessary for our win, and once we lose (say on betting tails) we can be 100% sure that betting heads would have been sufficient for producing a win. The empirical content of such counterfactuals is discussed in Appendix 19.A.

Note that these counterfactual quantities cannot be computed from the joint probability of X and Y without knowledge of the functional relationship in Equation (19.46) which tells us the (deterministic) policy by which a win or a loss is decided. This can be seen, for instance, from the conditional probabilities and causal effects associated with this example

$$P(y | x) = P(y | x') = P(y_x) = P(y_{x'}) = P(y) = \frac{1}{2},$$

because identical probabilities would be generated by a random payoff policy in which y is functionally independent of x , say by a bookie who watches the coin and ignores our bet. In such a random policy, the probabilities of causation PN, PS and PNS are all zero. Thus, according to our definition of identifiability (Definition 19.12), if two models agree on P and do not agree on a quantity Q , then Q is not identifiable. Indeed, the bounds delineated in Theorem 19.1 (Equation (19.20)) read $0 \leq \text{PNS} \leq \frac{1}{2}$, meaning that the three probabilities of causation cannot be determined from statistical data on X and Y alone, not even in a controlled experiment; knowledge of the functional mechanism is required, as in Equation (19.46).

It is interesting to note that whether the coin is tossed before or after the bet has no bearing on the probabilities of causation as defined above. This stands in contrast with some theories of probabilistic causality which attempt to avoid deterministic mechanisms by conditioning all probabilities on “the state of the world just before” the occurrence of the cause in question (x) (e.g., [Good 1961](#)). In the betting story above, the intention is to condition all probabilities on the state of the coin (u), but it is not fulfilled if the coin is tossed after the bet is placed. Attempts to enrich the conditioning set with events occurring after the cause in question have led back to deterministic relationships involving counterfactual variables (see [Cartwright 1989](#); [Eells 1991](#)).

One may argue, of course, that if the coin is tossed after the bet, then it is not at all clear what our winning would be had we bet differently; merely uttering our bet could conceivably affect the trajectory of the coin ([Dawid 1997](#)). This objection can be diffused by placing x and u in two remote locations and tossing the coin a split second after the bet is placed, but before any light ray could arrive from the betting room to the coin-tossing room. In such hypothetical situation the counterfactual statement: “our winning would be different had we bet differently” is rather compelling, even though the conditioning event (u) occurs after the cause in question (x). We conclude that temporal descriptions such as “the state of the world just before x ” cannot be used to properly identify the appropriate set of conditioning events (u) in a problem; a deterministic model of the mechanisms involved is needed for such identification.

19.4.2 Example 2: The Firing Squad

Consider a 2-man firing squad (see [Figure 19.2](#)) in which A and B are riflemen, C is the squad’s Captain who is waiting for the court order, U , and T is a condemned prisoner. Let u be the proposition that the court has ordered an execution, x the proposition stating that A pulled the trigger, and y that T is dead. Assume that $P(u) = \frac{1}{2}$, that A and B are perfectly accurate marksmen who are alert and law abiding, and that T is not likely to die from fright or other extraneous causes. We wish to compute the probability that x was a necessary (or sufficient, or both) cause for y (i.e., PN, PS, and PNS).

Definitions (19.7)–(19.9) permit us to compute these probabilities directly from the given causal model, since all functions and all probabilities are specified, with the truth value of each variable tracing that of U . Accordingly, we can write¹⁸

18. Recall that $P(Y_x(u') = \text{true})$ involves the submodel M_x , in which X is set to *true* independently of U . Thus, although under condition u' the captain has not given a signal, the potential outcome $Y_x(u')$ calls for hypothesizing rifleman- A pulling the trigger (x) despite a court order to stay the execution.

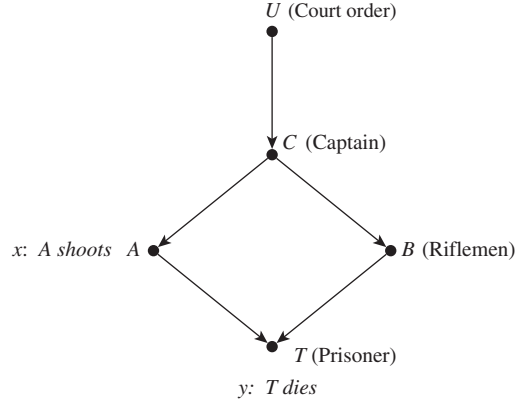


Figure 19.2 Causal relationships in the 2-man firing squad example.

$$\begin{aligned}
 P(y_x) &= P(Y_x(u) = true)P(u) + P(Y_x(u') = true)P(u') \\
 &= \frac{1}{2}(1 + 1) = 1.
 \end{aligned}
 \tag{19.47}$$

Similarly, we have

$$\begin{aligned}
 P(y_{x'}) &= P(Y_{x'}(u) = true)P(u) + P(Y_{x'}(u') = true)P(u') \\
 &= \frac{1}{2}(1 + 0) = \frac{1}{2}.
 \end{aligned}
 \tag{19.48}$$

To compute PNS, we need to evaluate the probability of the joint event $y_{x'} \wedge y_x$. Considering that these two events are jointly true only when $U = true$, we have

$$\begin{aligned}
 \text{PNS} &= P(y_x, y_{x'}) \\
 &= P(y_x, y_{x'} | u)P(u) + P(y_x, y_{x'} | u')P(u') \\
 &= \frac{1}{2}(1 + 0) = \frac{1}{2}.
 \end{aligned}
 \tag{19.49}$$

The calculation of PS and PN, likewise, are simplified by the fact that each of the conditioning events, $x \wedge y$ for PN and $x' \wedge y'$ for PS, is true in only one state of U . We thus have

$$\text{PN} = P(y'_{x'} | x, y) = P(y'_{x'} | u) = 0$$

reflecting the fact that, once the court orders an execution (u), T will die (y) from the shot of rifleman B , even if A refrains from shooting (x'). Indeed, upon learning of T 's death, we can categorically state that rifleman- A 's shot was *not* a necessary cause of the death.

Similarly,

$$PS = P(y_x | x', y') = P(y_x | u') = 1$$

matching our intuition that a shot fired by an expert marksman would be sufficient for causing the death of T , regardless of the court decision.

Note that Theorems 19.1 and 19.2 are not applicable to this example, because x is not exogenous; events x and y have a common cause (the Captain's signal) which renders $P(y | x') = 0 \neq P(y_{x'}) = \frac{1}{2}$. However, the monotonicity of Y (in x) permits us to compute PNS, PS and PN from the joint distribution $P(x, y)$ (using Equations (19.39)–(19.41)), instead of consulting the basic model. Indeed, writing

$$P(x, y) = P(x', y') = \frac{1}{2}, \quad (19.50)$$

$$P(x, y') = P(x', y) = 0, \quad (19.51)$$

we obtain

$$PN = \frac{P(y) - P(y_{x'})}{P(x, y)} = \frac{\frac{1}{2} - \frac{1}{2}}{\frac{1}{2}} = 0, \quad (19.52)$$

$$PS = \frac{P(y_x) - P(y)}{P(x', y')} = \frac{1 - \frac{1}{2}}{\frac{1}{2}} = 1, \quad (19.53)$$

as expected.

19.4.3 Example 3: The Effect of Radiation on Leukemia

Consider the following data (adapted from Finkelstein and Levin¹⁹ (1990)) comparing leukemia deaths in children in Southern Utah with high and low exposure to radiation from fallout from nuclear tests in Nevada. Given these data, we wish to estimate the probabilities that high exposure to radiation was a necessary (or sufficient or both) cause of death due to leukemia.

Assuming that exposure to nuclear radiation had no remedial effect on any individual in the study (i.e., monotonicity), the process can be modeled by a simple disjunctive mechanism represented by the equation

$$y = f(x, u, q) = (x \wedge q) \vee u, \quad (19.54)$$

19. The data in Finkelstein and Levin (1990) are given in person-year units. For the purpose of illustration we have converted the data to absolute numbers (of deaths and non-deaths) assuming a 10-year observation period.

where u represents “all other causes” of y , and q represents all “enabling” mechanisms that must be present for x to trigger y . Assuming q and u are both unobserved, the question we ask is under what conditions we can identify the probability of causation, PNS, PN, and PS, from the joint distribution of X and Y .

Since Equation (19.54) is monotonic in x , Theorem 19.3 states that all three quantities would be identifiable provided X is exogenous, namely, x should be independent of q and u . Under this assumption, Equations (19.32)–(19.34) further permit us to compute the probabilities of causation from frequency data. Taking fractions to represent probabilities, the data in Table 19.1 imply the following numerical results

$$\begin{aligned} \text{PNS} &= P(y|x) - P(y|x') \\ &= \frac{30}{30 + 69,130} - \frac{16}{16 + 59,010} = 0.0001625, \end{aligned} \tag{19.55}$$

$$\text{PN} = \frac{\text{PNS}}{P(y|x)} = \frac{\text{PNS}}{30/(30 + 69,130)} = 0.37535, \tag{19.56}$$

$$\text{PS} = \frac{\text{PNS}}{1 - P(y|x')} = \frac{\text{PNS}}{1 - 16/(16 + 59,010)} = 0.0001625. \tag{19.57}$$

Statistically, these figures mean: There is a 1.625 in ten thousand chance that a randomly chosen child would both die of leukemia if exposed and survive if not exposed. There is a 37.535% chance that a child who died from leukemia after exposure would have survived had he/she not been exposed. There is a 1.625 in ten-thousand chance that any unexposed surviving child would have died of leukemia had he/she been exposed.

Glymour (1998) analyzes this example with the aim of identifying the probability $P(q)$ (Cheng’s “causal power”) which coincides with PS (see Lemma 19.3). Glymour concludes that $P(q)$ is identifiable and is given by Equation (19.34), provided x , u , and q are mutually independent. Our analysis shows that Glymour’s result can

Table 19.1 Frequency data comparing leukemia deaths in children with high and low exposure to nuclear radiation

		Exposure	
		High	Low
		x	x'
Deaths	y	30	16
Survivals	y'	69,130	59,010

be generalized in several ways. First, since Y is monotonic in X , the validity of Equation (19.34) is assured even when q and u are dependent, because exogeneity merely requires independence between x and $\{u, q\}$ jointly. This is important in epidemiological settings, because an individual's susceptibility to nuclear radiation is likely to be associated with his/her susceptibility to other potential causes of leukemia (e.g., natural kinds of radiation).

Second, Theorem 19.2 assures us that the relationships between PN, PS and PNS (Equations (19.22)–(19.23)), which Glymour derives for independent q and u , should remain valid even when u and q are dependent.

Finally, Theorem 19.4 assures us that PN and PS are identifiable even when x is not independent of $\{u, q\}$, provided only that the mechanism of Equation (19.54) is embedded in a larger causal structure which permits the identification of $P(y_x)$. For example, assume that exposure to nuclear radiation (x) is suspect of being associated with terrain and altitude, which are also factors in determining exposure to cosmic radiation. A model reflecting such consideration is depicted in Figure 19.3, where W represents factors affecting both X and U . A natural way to correct for possible confounding bias in the causal effect of X on Y would be to adjust for W , that

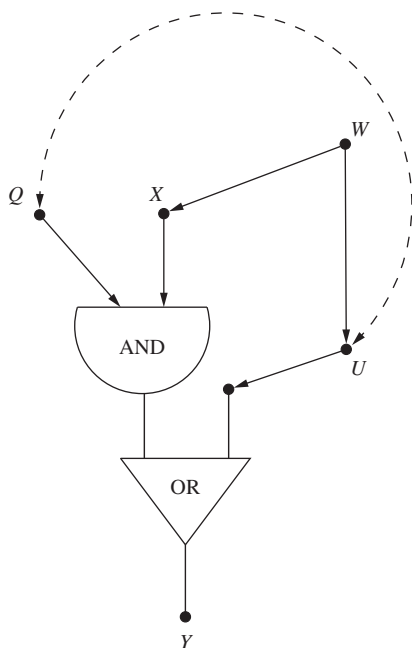


Figure 19.3 Causal relationships in the Radiation–Leukemia example. W represents confounding factors.

is, to calculate $P(y_x)$ using the adjustment formula

$$P(y_x) = \sum_w P(y|x, w)P(w), \quad (19.58)$$

(instead of $P(y|x)$) where the summation runs over levels of W . This adjustment formula, which follows from Equation (19.45), is correct regardless of the mechanisms mediating X and Y , provided only that W represents *all* common factors affecting X and Y (Pearl 1995). Theorem 19.4 instructs us to evaluate PN and PS by substituting (19.58) into Equations (19.40) and (19.41), respectively, and it assures us that the resulting expressions constitute consistent estimates of PN and PS. This consistency is guaranteed jointly by the assumption of monotonicity and by the (assumed) topology of the causal graph.

Note the monotonicity as defined in Equation (19.31) is a global property of all pathways between x and y . The causal model may include several nonmonotonic mechanisms along these pathways without affecting the validity of Equation (19.31). Arguments for the validity of monotonicity, however, must be based on substantive information, as it is not testable in general. For example, Robins and Greenland (1989) argue that exposure to nuclear radiation may conceivably be of benefit to some individuals, since such radiation is routinely used clinically in treating cancer patients.

19.4.4 Example 4: Legal Responsibility from Experimental and Nonexperimental Data

A lawsuit is filed against the manufacturer of drug x , charging that the drug is likely to have caused the death of Mr. A, who took the drug to relieve symptom S associated with disease D . The manufacturer claims that experimental data on patients with symptom S show conclusively that drug x may cause only negligible increase in death rates. The plaintiff argues, however, that the experimental study is of little relevance to this case, because it represents the effect of the drug on *all* patients, not on patients like Mr. A who actually died while using drug x . Moreover, argues the plaintiff, Mr. A is unique in that he used the drug on his own volition, unlike subjects in the experimental study who took the drug to comply with experimental protocols. To support this argument, the plaintiff furnishes non-experimental data indicating that most patients who chose drug x would have been alive if it were not for the drug. The manufacturer counter-argues by stating that: (1) counterfactual speculations regarding whether patients would or would not have died are purely metaphysical and should be avoided (Dawid 1997), and (2) non-experimental data should be dismissed a priori, on the ground that, such data may be highly biased; for example, incurable terminal patients might be more inclined to use drug x if

Table 19.2 Frequency data (hypothetical) obtained in experimental and non-experimental studies, comparing deaths among drug users (x) and non-users (x')

		Experimental				Non-Experimental	
		x	x'			x	x'
Deaths	y	16	14	Deaths	y	2	28
Survivals	y'	984	986	Survivals	y'	998	972

it provides them greater symptomatic relief. The court must now decide, based on both the experimental and non-experimental studies, what the probability is that drug x was in fact the cause of Mr. A's death.

The (hypothetical) data associated with the two studies are shown in Table 19.2. The experimental data provide the estimates

$$P(y_x) = 16/1000 = 0.016, \tag{19.59}$$

$$P(y_{x'}) = 14/1000 = 0.014. \tag{19.60}$$

The non-experimental data provide the estimates

$$P(y) = 30/2000 = 0.015, \tag{19.61}$$

$$P(y, x) = 2/2000 = 0.001. \tag{19.62}$$

Assuming that drug x can only cause, never prevent, death, Theorem 19.4 is applicable and Equation (19.40) gives

$$PN = \frac{P(y) - P(y_{x'})}{P(y, x)} = \frac{0.015 - 0.014}{0.001} = 1.00. \tag{19.63}$$

Thus, the plaintiff was correct; barring sampling errors, the data provide us with 100% assurance that drug x was in fact responsible for the death of Mr. A. Note that a straightforward use of the experimental excess-risk-ratio would yield a much lower (and incorrect) result:

$$\frac{P(y_x) - P(y_{x'})}{P(y_x)} = \frac{0.016 - 0.014}{0.016} = 0.125. \tag{19.64}$$

Evidently, what the experimental study does not reveal is that, given a choice, terminal patients stay away from drug x . Indeed, if there were any terminal patients who would choose (given the choice), then the control group (x') would have included some such patients (due to randomization) and then the proportion of deaths among the control group $P(y_{x'})$ should have been higher than $P(x', y)$, the population proportion of terminal patients avoiding x . However, the equality $P(y_{x'}) = P(y, x')$ tells us that no such patients were included in the control group, hence

(by randomization) no such patients exist in the population at large and, therefore, none of the patients who freely chose drug x was a terminal case; all were susceptible to x .

The numbers in Table 19.2 were obviously contrived to represent an extreme case, so as to facilitate a qualitative explanation of the validity of Equation (19.40). Nevertheless, it is instructive to note that a combination of experimental and non-experimental studies may unravel what experimental studies alone will not reveal and, in addition, that such combination may provide a test for the assumption of no-prevention, as outlined in Section 19.3.4 (Equation (19.43)).

19.5 Identification in Non-Monotonic Models

In this section we discuss the identification of probabilities of causation without making the monotonicity assumption. We will assume that we are given a causal model M in which all functional relationships are known, but since the exogenous variables U are not observed, their distributions are not known.

A straightforward way to identify any causal or counterfactual quantity (including PN, PS and PNS) would be to infer the probability distribution of the exogenous variables – that would amount to inferring the entire model, from which all quantities can be computed. Thus, our first step would be to study under what conditions the function $P(u)$ can be identified.

If M is Markovian, the problem can be analyzed by considering each parents-child family separately. Consider any arbitrary equation in M

$$\begin{aligned} y &= f(pa_Y, u_Y) \\ &= f(x_1, x_2, \dots, x_k, u_1, \dots, u_m), \end{aligned} \quad (19.65)$$

where $U_Y = \{U_1, \dots, U_m\}$ is the set of exogenous, possibly dependent variables that appear in the equation for Y . In general, the domain of U_Y can be arbitrary, discrete, or continuous, since these variables represent unobserved factors that were omitted from the model. However, since the observed variables are binary, there is only a finite number ($2^{(2^k)}$) of functions from PA_Y to Y and, for any point $U_Y = u$, only one of those functions is realized. This defines a partition of the domain of U_Y into a set S of equivalence classes, where each equivalence class $s \in S$ induces the same function $f^{(s)}$ from PA_Y to Y . Thus, as u varies over its domain, a set S of such functions is realized, and we can regard S as a new exogenous variable, whose values are the set $\{f^{(s)} : s \in S\}$ of functions from PA_Y to Y that are realizable in U_Y . The number of such functions will usually be smaller than $2^{(2^k)}$.²⁰

20. Balke and Pearl (1994a) called these S variables “response variables”, and Heckerman and Shachter (1995) called them “mapping variables”.

For example, consider the model described in Figure 19.3. As the exogenous variables (q, u) vary over their respective domains, the relation between X and Y spans three distinct functions

$$Y = \text{true}, Y = \text{false}, \quad \text{and} \quad Y = X.$$

The fourth possible function, $Y = \text{not-}X$, is never realized because $f_Y(\cdot)$ is monotonic. The cells (q, u) and (q', u) induce the same function between X and Y , hence they belong to the same equivalence class.

If we are given the distribution $P(u_Y)$, we can compute the distribution $P(s)$ and this will determine the conditional probabilities $P(y | pa_Y)$ by summing $P(s)$ over all those functions $f^{(s)}$ that map pa_Y into the value *true*,

$$P(y | pa_Y) = \sum_{s: f^{(s)}(pa_Y) = \text{true}} P(s). \quad (19.66)$$

To insure model identifiability it is sufficient that we can invert the process and determine $P(s)$ from $P(y | pa_Y)$. If we let the set of conditional probabilities $P(y | pa_Y)$ be represented by a vector \mathbf{p} (of dimensionality 2^k), and $P(s)$ by a vector \mathbf{q} , then the relation between \mathbf{q} and \mathbf{p} is linear and can be represented as a matrix multiplication (Balke and Pearl 1994b)

$$\mathbf{p} = \mathbf{R}\mathbf{q}, \quad (19.67)$$

where \mathbf{R} is a 0-1 matrix, with dimension $2^k \times |S|$. Thus, a sufficient condition for identification is simply that \mathbf{R} , together with the normalizing equation $\sum_j \mathbf{q}_j = 1$, be invertible.

In general, \mathbf{R} will not be invertible because the dimensionality of \mathbf{q} can be much larger than that of \mathbf{p} . However, in many cases, such as the Noisy-OR mechanism

$$Y = U_0 \bigvee_{i=1, \dots, k} (X_i \wedge U_i), \quad (19.68)$$

symmetry permits \mathbf{q} to be identified from $P(y | pa_Y)$ even when the exogenous variables U_0, U_1, \dots, U_k are not independent. This can be seen by noting that every point u for which $U_0 = \text{false}$ defines a unique function $f^{(s)}$ because, letting T be the set of indices i for which U_i is true, the relationship between PA_Y and Y becomes

$$Y = U_0 \bigvee_{i \in T} X_i \quad (19.69)$$

and, for $U_0 = \text{false}$, this equation defines a distinct function for each T . The number of induced functions is $2^k + 1$, which (subtracting 1 for normalization) is exactly

the number of distinct realizations of PA_Y . Moreover, it is easy to show that the matrix connecting \mathbf{p} and \mathbf{q} is invertible. We thus conclude that the probability of every counterfactual sentence can be identified in any Markovian model composed of Noisy-OR mechanisms, regardless of whether the exogenous variables in each family are mutually independent. The same holds of course for Noisy-AND mechanisms or any combination thereof, including negating mechanisms, provided that each family consists of one type of mechanism.

To generalize these results to mechanisms other than Noisy-OR and Noisy-AND, we note that although $f_Y(\cdot)$ in this example was monotonic (in each X_i), it was the redundancy of $f_Y(\cdot)$, not its monotonicity, that ensured identifiability. The following is an example of a monotonic function for which the \mathbf{R} matrix is not invertible

$$Y = (X_1 \wedge U_1) \vee (X_2 \wedge U_1) \vee (X_1 \wedge X_2 \wedge U_3).$$

It represents a Noisy-OR gate for $U_3 = \text{false}$, and becomes a Noisy-AND gate for $U_3 = \text{true}$, $U_1 = U_2 = \text{false}$. $U_1 = U_2 = \text{false}$. The number of equivalence-classes induced is six, which would require five independent equations to determine their probabilities; the data $P(y|pa_Y)$ provide only four such equations.

In contrast, the mechanism governed by the equation below, although non-monotonic, is invertible:

$$Y = \text{XOR}(X_1, \text{XOR}(U_2, \dots, \text{XOR}(U_{k-1}, \text{XOR}(X_k, U_k)))),$$

where XOR^* stands for Exclusive-OR. This equation induces only two functions from PA_Y to Y ;

$$Y = \begin{cases} \text{XOR}(X_1, \dots, X_k) & \text{if } \text{XOR}(U_1, \dots, U_k) = \text{false} \\ \neg \text{XOR}(X_1, \dots, X_k) & \text{if } \text{XOR}(U_1, \dots, U_k) = \text{true}. \end{cases}$$

A single conditional probability, say $P(y|x_1, \dots, x_k)$, would therefore suffice for computing the one parameter needed for identification: $P[\text{XOR}(U_1, \dots, U_k) = \text{true}]$.

We summarize these considerations with a theorem.

Definition 19.16 Local invertability

A model M is said to be locally invertible if for every variable $V_i \in V$ the set of $2^k + 1$ equations

$$P(y|pa_i) = \sum_{s: f_i^{(s)}(pa_i)=\text{true}} q_i(s), \quad (19.70)$$

$$\sum_s q_i(s) = 1 \quad (19.71)$$

has a unique solution for $q_i(s)$, where each $f_i^{(s)}(pa_i)$ corresponds to the function $f_i(pa_i, u_i)$ induced by u_i in equivalence-class s .

Theorem 19.5 Given a Markovian model $M = \langle U, V, \{f_i\} \rangle$ in which the functions $\{f_i\}$ are known and the exogenous variables U are unobserved, if M is locally invertible, then the probability of every counterfactual sentence is identifiable from the joint probability $P(v)$.

Proof. If Equation (19.70) has a unique solution for $q_i(s)$, we can replace U with S and obtain an equivalent model

$$M' = \langle S, V, \{f'_i\} \rangle \quad \text{where} \quad f'_i = f_i^{(s)}(pa_i).$$

M' together with $q_i(s)$ completely specifies a probabilistic model $\langle M', P(s) \rangle$ (due to the Markov property) from which probabilities of counterfactuals are derivable by definition. ■

Theorem 19.5 provides a sufficient condition for identifying probabilities of causation, but of course does not exhaust the spectrum of assumptions that are helpful in achieving identification. In many cases we might be justified in hypothesizing additional structure on the model, for example, that the U variables entering each family are themselves independent. In such cases, additional constraints are imposed on the probabilities $P(s)$ and Equation (19.70) may be solved even when the cardinality of S far exceeds the number of conditional probabilities $P(y|pa_Y)$.

19.6 From Necessity and Sufficiency to “Actual Cause”

19.6.1 The Role of Structural Information

In Section 19.3, we alluded to the fact that both PN and PS are global (i.e., input-output) features of a causal model, depending only on the function $Y_x(u)$, but not on the structure of the process mediating between the cause (x) and the effect (y). That such structure plays a role in causal explanation is seen in the following example.

Consider an electric circuit consisting of a light bulb and two switches, and assume that the light is turned on whenever either switch-1 or switch-2 is on. Assume further that, internally, when switch-1 is on it not only activates the light

but also disconnects switch-2 from the circuit, rendering it inoperative. From an input–output viewpoint, the light responds symmetrically to the two switches; either switch is sufficient to turn the light on. However, with both switches on, we would not hesitate to proclaim to switch-1 as the “actual cause” of the current flowing in the light bulb, knowing that, internally, switch-2 is totally disconnected in this particular state of affairs. There is nothing in PN and PS that could possibly account for this asymmetry; each is based on the response function $Y_x(u)$, and is therefore oblivious to the internal workings of the circuit.

This example is isomorphic to Suppes’ Desert Traveler, and belongs to a large class of counterexamples that were brought up against Lewis’ counterfactual account of causation. It illustrates how an event (e.g., switch-1 being on) can be considered a cause although the effect persists in its absence. Lewis’ (1986) answer to such counterexamples was to modify the counterfactual criterion and let x be a cause of y as long as there exists a counterfactual-dependence chain of intermediate variables between x to y , that is, the output of every link in the chain is counterfactually dependent on its input. Such a chain does not exist for switch-2, since it is disconnected when both switches are on.

Lewis’ chain criterion retains the connection between causation and counterfactuals, but it is rather ad-hoc; after all, why should the existence of a counterfactual-dependence chain be taken as a defining test for such crucial concepts as “actual cause”, by which we decide the guilt or innocence of defendants in a court of law? Another problem with Lewis’ chain is its failure to capture symmetric cases of overdetermination. For example, consider two switches connected symmetrically, such that each participates equally in energizing the light bulb. In this situation, our intuition regards each of the switches as a contributory actual cause of the light, though none passes the counterfactual test and none supports a counterfactual-dependence chain in the presence of the other.

An alternative way of using counterfactuals to define actual causes is proposed in (Pearl 1998). An event x is defined as the “actual cause” of event y (in a world u), if x passes the standard counterfactual test (i.e., $Y_x(u) = false$) in some mutilated model M' , minimally removed from M . In the symmetric two-switch example, we declare each switch to be an actual cause of the light because the light would be off if that switch were off, when we consider a slightly mutilated circuit, one in which the other switch is disconnected from the power source. The mutilated model M' , called a “causal beam”, is carefully constructed in (Pearl 1998) to ensure minimal deviation from the actual causal model M , considering the actual history of the world u .

The concept of causal sufficiency offers yet a third way of rescuing the counterfactual account of causation. Consider again the symmetric two-switch example

(or the firing squad example of Section 19.4.2). Both switches enjoy high PS value, because each would produce light from a state (u') of darkness, namely, a state in which the other switch is off. Likewise, the shot of each rifleman in Example 2 (Section 19.4.2) enjoys a PS value of unity (see Equation (19.53)), because each shot would cause the prisoner's death in the state u' in which the prisoner is alive, namely, the court orders no execution. Thus, if our intuition is driven by some strange mixture of sufficiency and necessity considerations, it seems plausible that we could formulate an adequate criterion for actual causation using the right mixture of PN and PS components.

Similar expectations are expressed in Hall (1998). In analyzing problems faced by the counterfactual approach, Hall makes the observation that there are two concepts of causation, only one of which is captured by the counterfactual account, and that failure to capture the second concept may well explain its clashes with intuition. Hall calls the first concept “dependence” and the second “production”. In the symmetrical two-switch example (an instance of “over-determination”), intuition considers each switch to be an equal “producer” of the light, while the counterfactual account tests for “dependence” only, and fails because the light does not “depend” on either switch alone.

The notions of dependence and production closely parallel those of necessity and sufficiency, respectively. Thus, our formulation of PS could well provide the formal basis for Hall's notion of production, and serve as a step toward the formalization of actual causation. For this program to succeed, several hurdles must be overcome, the most urgent being the problems of incorporating singular event information and structural information into PS. These will be discussed next.

19.6.2 Singular Sufficient Causes

So far we have explicated the necessity and sufficiency conceptions of causation in terms of their probabilities, but not as properties of a given specific scenario, dictated by a specific state of U . This stands in contrast with standard practice of first defining truth values of sentences in each specific world, then evaluating probabilities of sentences from probabilities of worlds. Lewis (1986) counterfactual account of causation, for example, assigns a truth value to the sentence “ x is a cause of y ” in each specific world (u), given by the conjunction $x \wedge y \wedge y'_x$. The question arises whether sentences about sufficient causation can likewise be given world-level truth values and, if they do, which worlds should provide those values, and how evidential information about those worlds should enter probability calculations.

Necessary causation can be formulated deterministically (at the world-level) in the standard counterfactual way:

Definition 19.17 Deterministic necessity

Event x is said to be a necessary cause of event y in a world u just in case the following hold in u :

1. $Y(u) = y$ and $X(u) = x$.
2. $Y_{x'}(u) \neq y$ for every $x' \neq x$.

Accordingly, if additional evidence e is available about our current world, it can easily be incorporated into the evaluation of PN as follows:

$$\text{PN}(x \rightarrow y | e) = P(y'_{x'} | x, y, e),$$

where $\text{PN}(x \rightarrow y | e)$ is the probability that x was a necessary cause of y , given evidence e .

Sufficient causation, on the other hand, requires a nonstandard deterministic (i.e., world-level) formulation.

Definition 19.18 Deterministic sufficiency

Event x is said to be a sufficient cause of event y in a world u just in case the following hold in u :

1. $Y(u) \neq y$ and $X(u) \neq x$.
2. $Y_x(u) = y$.

In words, x is a sufficient cause for y if x would produce y (counterfactually) in world u in which x and y are absent.

The nonstandard feature of this definition lies in requiring both the explanation (x) and the explanandum (y) to be false in any world u where the former pertains to cause the latter. Thus, it appears that nothing could possibly explain (by consideration of sufficiency) events that happened to materialize in the actual world. This feature reflects, of course our commitment to interpret sufficiency as the capacity to produce an effect and, as strange as it may sound, it is indeed impossible to talk about “ x producing y ” in a world (say ours) in which x and y are already true. The word “production” implies the establishment of new facts. Therefore to test production, we must step outside our world momentarily, imagine a new world with x and y absent, apply x , and see if y sets in.

This peculiar feature of sufficiency leads to difficulties in incorporating world-specific findings into the analysis. Consider a 1-man firing squad in which rifleman A has a hit rate of 99% and the prisoner has a small chance p of dying from fear. Our analysis of Section 19.3 indicates that PS equals 99%, independent of p . Now suppose we find that the bullet fired hit the prisoner’s leg, from which we conclude that the prisoner must have died from fear. Would this finding change our assessment of how sufficient A ’s shot was for causing T ’s death? There are grounds for

arguing that it should: although, in general, a shot from a rifleman like Mr. A would be 99% sufficient for the job, this particular shot was evidently of a different type, a peculiar type that scores zero on the accuracy and sufficiency scale.

However it is not at all trivial to formalize this argument using the logical machinery at our disposal. First, to properly incorporate the new piece of evidence, e : “The bullet was found in the prisoner’s leg” we need to know the structure of the causal process; the function $Y_x(u)$ in itself would be insufficient, for it does not tell us how the location of the bullet alters the chance of death. But even given the structure, say in the form of an intermediate variable denoting “Location of bullet”, we cannot simply add e to the conditioning part in the expression for PS, forming $P(y_x | x', y', e)$, as we did for PN. The location of the bullet was observed in the actual world, that is, after x was enacted and y verified, while the conditioning events, x' and y' pertain to a hypothetical world that existed prior to the action (x). Mixing the two without making this distinction leads to contradictions and misinterpretations. The expression $P(y_x | x', y', e)$ amounts to evaluating the probability that a living prisoner carrying a bullet in his leg would die if shot by Mr. A. This is certainly not the intended interpretation of PS and would not evaluate to zero as it should. As another example, if e stands for “bullet in the heart”, which conflicts with y' , we would be instructed into conditioning $P(y_x)$ on a contradictory event.

An attempt to place e in the consequent part of the counterfactual, forming $P(y_x, e | x', y')$, again does not accomplish our mission.²¹ It expresses the probability that, both, the shot would be sufficient to cause death and that a living prisoner would have a bullet in his leg; still far from the probability that a shot in the leg will suffice to cause death.

These difficulties stem from dealing with the dynamic process of “production” using a syntax that does not allow explicit reference to time. Fortunately, the difficulty can be resolved even in the confines of this syntax. Since the evidence e was obtained in a world created by the action x , and since events in such worlds are governed by the submodel M_x (see Section 19.2.1), the proper syntax for introducing such evidence would be to condition on the subscripted symbol e_x . This leads to:

Definition 19.19 Singular-event sufficiency

The probability that x was a sufficient cause of y given evidence e is defined as:²²

21. Related attempt to modify the consequent part is reported in Michie (1997), using an adaptation of Good’s measure of causal sufficiency, Q_{suf} .

22. Other expressions are also possible, for example, $P(y_x, e_x | x', y')$, which captures the capacity of x to produce both y and e . This expression suffers, however, from sensitivity to detail; elaborate descriptions of e would yield extremely low probabilities.

$$\text{PS}(x \rightarrow y | e) = P(y_x | e_x, x', y'). \quad (19.72)$$

To illustrate, assume Z stands for a 2-state variable “Location of bullet”, with z denoting “bullet in chest” and z' denoting “bullet not in chest”. Assuming further that the flow of causation is governed by the causal chain $X \rightarrow Z \rightarrow Y$, and that a bullet would cause death if and only if it ends up in the chest. It is not hard to show that Definition 19.19 yields

$$\begin{aligned} \text{PS}(x \rightarrow y | z) &= 1, \\ \text{PS}(x \rightarrow y | z') &= 0, \\ \text{PN}(x \rightarrow y | z) &= 1, \\ \text{PN}(x \rightarrow y | z') &= 1 - P(\text{death from fear}), \end{aligned} \quad (19.73)$$

as expected.

The next subsection illustrates the role of singular event information in a probabilistic analysis of Suppes’ desert traveller story.

19.6.3 Example: The Desert Traveler (after P. Suppes)

A desert traveller T has two enemies. Enemy-1 poisons T ’s canteen, and Enemy-2, unaware of Enemy-1’s action, shoots and empties the canteen.

A week later, T is found dead and the two enemies confess to action and intention. A jury must decide whose action was the cause of T ’s death.

Let u be the proposition that traveller’s first need of drink occurred after the shot was fired. Let x and p be the propositions “Enemy-2 shot”, and “Enemy-1 poisoned the water”, respectively, and let y denote “ T is dead”. In addition to these events we will make informal use of possible exceptions to the normal story, such as T surviving the ordeal or T suspecting that the water is poisoned.

The causal model underlying the story is depicted in Figure 19.4. The model is completely specified through the functions $f_i(pa_i, u)$ which are not shown explicitly in Figure 19.4, but are presumed to determine the value of each child variable from those of its parent variables in the graph, in accordance with our usual understanding of the story:

$$\begin{aligned} c &= p \wedge (u' \vee x'), \\ d &= x \wedge (u \vee p'), \\ y &= c \vee d. \end{aligned}$$

(We assume that T will not survive with empty canteen (x) even after drinking some unpoisoned water before the shot ($p' \wedge u'$).)

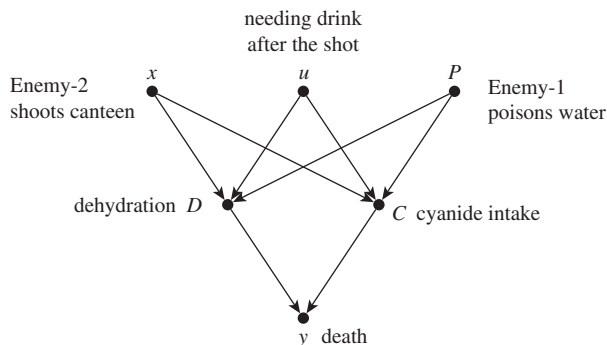


Figure 19.4 Causal relationships in the Desert-Traveler example.

19.6.3.1 Necessity and Sufficiency Ignoring Internal Structure

The global function $Y(x, p, u)$ is given by

$$y = x \vee p,$$

which is symmetric in x and p .

The calculations of $PS(x \rightarrow y) = PS(p \rightarrow y)$ and $PN(p \text{ g } y) = PN(x \text{ g } y)$, can proceed directly from their definitions, without resorting to structural information.

$$PS(x \rightarrow y) = P(y_x | x', y') = 1,$$

because (x', y') implies that no poison was added (p'), in which case $P(y_x)$ is 1, barring the unlikely event that T manages to survive with an empty canteen.

Similarly,

$$PN(x \rightarrow y) = P(y_{x'} | x, y) = 0.$$

If we wish to include the possibility of T surviving with either an empty canteen or a poisoned canteen, we have:

$$\begin{aligned}
 PS(x \rightarrow y) &= P(y_x | x', y') = 1 - P(\text{survival with empty canteen}) \\
 PS(p \rightarrow y) &= P(y_p | p', y') = 1 - P(\text{survival with poisoned water}). \tag{19.74}
 \end{aligned}$$

Note that $PN(x \rightarrow y)$ and $PN(p \rightarrow y)$ remain zero, unaffected by the possibility of survival, because T 's death (y) is taken as evidence that conditions necessary for survival did not in fact materialize (see Lemma 19.2).

19.6.3.2 Sufficiency and Necessity given Forensic Reports

Let c stand for: “Cyanide was found in T ’s body” and d for: T ’s body showed signs of dehydration”.

Incorporating the first evidence into the probability of sufficiency (Equation (19.72)), we have

$$\text{PS}(x \rightarrow y | c) = P(y_x | x', y', c_x).$$

The conditioning part instructs us to imagine a scenario in which Enemy-2 did not shoot, T did not die and cyanide would be found in T ’s body if Enemy-2 were to shoot. The one scenario which complies with these conditions is as follows: The water was poisoned, T drank the water before the time Enemy-2 was about to shoot (u'), (thus c_x is true despite x), and T was somehow rescued (y'). Under such a scenario, Enemy-2 shooting would not produce T ’s death, hence, $\text{PS}(x \rightarrow y | c) = 0$. This matches our intuition; upon learning that T ’s body contains cyanide, emptying the canteen is no longer death.

Now consider the evidence d : “dehydration”. To evaluate

$$\text{PS}(x \rightarrow y | d) = P(y_x | x', y', d_x)$$

we need first list all scenarios compatible with (x', y', d_x) , namely: no shot fired, T is alive and T would be dehydrated if Enemy-2 were to shoot. Two scenarios come to mind, one natural, the other bizarre.

Scenario 1: No shot fired, the water is poisoned, the poisoned water would be emptied if Enemy-2 were to shoot (u), and T would suffer dehydration. In this scenario x would produce death, unless T is rescued.

Scenario 2: No shot was fired, T would come to drink before the shot (if any) but would somehow suspect that the water is poisoned and refrain from drinking. This would cause dehydration by choice, and death unless rescued.

Summing over both scenarios, we obtain

$$\text{PS}(x \rightarrow y | d) = 1 - P(T \text{ survives in dehydration}).$$

To summarize, we now have

$$\text{PS}(x \rightarrow y) = 1 - P(\text{survival with empty canteen})$$

$$\text{PS}(x \rightarrow y | c) = 0$$

$$\text{PS}(x \rightarrow y | d) = 1 - P(T \text{ will be rescued after dehydration}). \quad (19.75)$$

Now consider the sufficiency of Enemy-1's action, in light of the two forensic reports. The conditioning part in

$$PS(p \rightarrow y | c) = P(y_p | p', y', c_p)$$

instructs us to imagine a scenario in which Enemy-1 did not poison the water, T did not die, but cyanide would be found in T 's body if Enemy-1 were to poison the water. This is the natural scenario to evolve if Enemy-2 did not shoot – T would die if the water were poisoned (y_p) unless rescued before the cyanide exerts its effect. Thus,

$$PS(p \rightarrow y | c) = 1 - P(\text{rescued after drinking cyanide}).$$

Finally, consider the evidence d : “dehydration”

$$PS(p \rightarrow y | d) = P(y_p | p', y', d_p).$$

We need first to list all scenarios compatible with (p', y', d_p) , namely: no poisoning occurred, T is alive and T would be dehydrated if enemy-1 were to poison the water. This is a bit hard to imagine, but not totally infeasible if we allow a special rescue operation: Enemy-2 shoots, the container is empty, T comes to drink after the shot is fired, dehydration occurs regardless of Enemy-1 action (d_p), but a rescue team revives T despite his state of dehydration.

In this scenario survival would occur even under p , therefore

$$PS(p \rightarrow y | d) = 0.$$

Summarizing:

$$PS(p \rightarrow y) = 1 - P(\text{survival with poisoned canteen})$$

$$PS(p \rightarrow y | c) = 1 - P(\text{rescue after drinking cyanide})$$

$$PS(p \rightarrow y | d) = 0. \tag{19.76}$$

19.6.3.3 Necessity Given Forensic Reports

The probabilities associated with necessary causation are usually easier to evaluate than their sufficiency counterparts, because the former call for scenarios that actually materialized in the story. To illustrate, let us evaluate the probability that Enemy-2 was a necessary cause of T 's death, given that cyanide was found in T 's body,

$$PN(x \rightarrow y | c) = P(y'_x | x, y, c).$$

The condition (x, y, c) can materialize only in state u' , where T drinks the poisoned water before the shot. Assuming this state, it is clear that T is doomed regardless of Enemy-2 action, and $y'_{x'}$ is false. Thus,

$$\text{PN}(x \rightarrow y | c) = 0.$$

Prospects of rescue, as we have mentioned before, do not alter this conclusion, because those are ruled out by the conditioning part.

A dehydration report would evoke the normal scenario, since

$$\text{PN}(x \rightarrow y | d) = P(y'_{x'} | x, y, d),$$

and condition (x, y, d) can materialize in state u : T reaches for drink after the shot is fired, finds the canteen empty, and suffers dehydration. In this state, $y'_{x'}$ is again false, because death would occur (from poison) even if Enemy-2 refrains from action (x'). Thus, as expected,

$$\text{PN}(x \rightarrow y | d) = 0.$$

For completeness, we evaluate the necessity ascribed to Enemy-1 action,

$$\begin{aligned} \text{PN}(p \rightarrow y | c) &= P(y'_{p'} | p, y, c) \\ &= P(T \text{ survives if not } p | u') = 0, \end{aligned} \tag{19.77}$$

because (p, y, c) implies that T drank the poisoned water before Enemy-2 fired and, in this state (u'), he would have died (from dehydration) even if Enemy-1 had not poisoned the water.

$$\begin{aligned} \text{PN}(p \rightarrow y | d) &= P(y'_{p'} | p, y, d) \\ &= P(T \text{ survives if not } p | u) = 0, \end{aligned} \tag{19.78}$$

because (p, y, d) implies that T reached for drink after Enemy-2 fired (u) and, in this state would have died (from dehydration) even if Enemy-1 had not poisoned the canteen.

Note that if we are not given any forensic report but assume, nevertheless, that such reports were available from the natural scenario in the story (i.e., u, x, p, d, y), then the probabilities of sufficiency would be (barring considerations of survival):

$$\begin{aligned} \text{PS}(x \rightarrow y | d) &= 1, \\ \text{PS}(p \rightarrow y | d) &= 0. \end{aligned} \tag{19.79}$$

These results coincide with those obtained from Lewis' analysis, using counterfactual-dependence chains. Whether this coincidence is universal, and whether it could serve as the basis for improving Lewis' account of causation remain a topic for future investigation.

19.7 Conclusion

This paper explicates and analyzes the necessary and sufficient components of causation. Using counterfactual interpretations that rest on structural-model semantics, the paper demonstrates how simple techniques of computing probabilities of counterfactuals can be used in computing probabilities of causes, deciding questions of identification, defining conditions under which probabilities of causes can be estimated from statistical data, and uncovering tests for assumptions that are routinely made (often unwittingly) by analysts and investigators.

On the practical side, the paper offers several useful tools to epidemiologists and health scientists. It formulates and calls attention to basic assumptions that must be ascertained before statistical measures such as excess-risk-ratio could represent causal quantities such as attributable-risk or probability of causes. It shows how data from both experimental and non-experimental studies can be combined to yield information that neither study alone can reveal. Finally, it provides tests for the commonly made assumption of "no prevention", and for the often asked question of whether a clinical study is representative of its target population.

On the conceptual side, we have seen that both the probability of necessity (PN) and probability of sufficiency (PS) play a role in our understanding of causation, and that both components have their logics and computational rules. Although the counterfactual concept of necessary cause (i.e., that an outcome would not have occurred "but for" the action) is predominant in legal settings (Robertson 1997) and in ordinary discourse, the sufficiency component of causation has a definite influence on causal thoughts.

The sufficiency component plays a major role in scientific and legal explanations, as can be seen from examples where the necessary component is dormant. Why do we consider striking a match to be a more adequate explanation (of a fire) than the presence of oxygen? Recasting the question in the language of PN and PS, we note that, since both explanations are necessary for the fire, each will command a PN of unity. (In fact PN is higher for the oxygen, if we allow for alternative ways of igniting a spark). Thus, it must be the sufficiency component alone that endows the match with greater explanatory power than the oxygen. If the probabilities associated with striking a match and the presence of oxygen are p_m and p_o , respectively, the PS measures associated with these explanations evaluate to $PS(\text{match}) = p_o$,

and $PS(\text{oxygen}) = p_m$, clearly favoring the match when $p_o \gg p_m$. Thus, a robot instructed to explain why a fire broke out has no choice but to consider both PN and PS in its deliberations.

Should PS enter legal considerations in criminal and tort law? I believe that it should (as does Good (1993)) because attention to sufficiency implies attention to the consequences of one's action. The person who lighted the match ought to have anticipated the presence of oxygen, whereas the person who supplied (or who could but failed to remove) the oxygen is not generally expected to have anticipated match-striking ceremonies.

However, what weight should the law assign to the necessary versus the sufficient component of causation? This question obviously lies beyond the scope of our investigation, and it is not at all clear who would be qualified to tackle the question or whether our legal system would be prepared to implement the recommendation. I am hopeful, however, that whoever undertakes to consider such questions will find the analysis in this paper to be of some use.

19.A Appendix: The Empirical Content of Counterfactuals

The word "counterfactual" is a misnomer, as it connotes a statement that stands contrary to facts or, at the very least, a statement that escapes empirical verification. Counterfactuals are in neither category; they are fundamental to scientific thought and carry as clear an empirical message as any scientific law.

Consider Ohm's law $V = IR$, the empirical content of this law can be encoded in two alternative forms.

1. **Predictive form:** If at time t_0 we measure current I_0 and voltage V_0 then, *ceteris paribus*, at any future times $t > t_0$, if the current flow will be $I(t)$ the voltage drop will be:

$$V(t) = \frac{V_0}{I_0} I(t).$$

2. **Counterfactual form:** If at time t_0 we measure current I_0 and voltage V_0 then, had the current flow at time t_0 been I' , instead of I_0 , the voltage drop would have been:

$$V' = \frac{V_0 I'}{I_0}.$$

On the surface, it seems that the predictive form makes meaningful and testable empirical claims while the counterfactual form merely speculates about events

that have not, and could not have occurred; as it is impossible to apply two different currents into the same resistor at the same time. However, if we interpret the counterfactual form to mean no more nor less than a conversational short hand of the predictive form, the empirical content of the former shines through clearly. Both enable us to make an infinite number of predictions from just one measurement (I_0, V_0), and both derive their validity from a scientific law (Ohm's law) which ascribes a time-invariant property (the ratio V/I) to any physical object.

I will adapt this predictive interpretation when I speak of counterfactuals, and I base this interpretation on the observation that counterfactuals, despite their a-temporal appearance, are invariably associated with some law-like, persistent relationships in the world. For example, the statement "had Germany not been punished so severely at the end of world-war I, Hitler would not have come to power" would sound bizarre to anyone who does not share our understanding that, as a general rule, "humiliation breeds discontent".

But if counterfactual statements are merely a round-about way of stating sets of predictions, why do we resort to such convoluted modes of expression instead of using the predictive mode directly? The answer, I believe, rests with the qualification "ceteris paribus" that accompanies the predictive claim, which is not entirely free of ambiguities. What should be held constant when we change the current in a resistor? The temperature? The laboratory equipments? The time of day? Certainly not the reading on the voltmeter? Such matters must be carefully specified when we pronounce predictive claims and take them seriously. Many of these specifications are implicit (hence superfluous) when we use counterfactual expressions, especially when we agree over the underlying causal model. For example, we do not need to specify under what temperature and pressure future predictions should hold true; these are implied by the statement "had the current flow at time t_0 been I' , instead of I_0 ". In other words, we are referring to precisely those conditions that prevailed in our laboratory at time t_0 . That statement also implies that we do not really mean for anyone to hold the reading on the voltmeter constant – only variables that, according to our causal model, are not affected by the counterfactual antecedent (I) are expected to remain constant for the predictions to hold true.

To summarize, I interpret a counterfactual statement to convey a set of predictions under a well defined set of conditions, those prevailing in the factual part of the statement. For these predictions to be valid, two components must remain invariants: the laws (or mechanisms) and the boundary conditions. Cast in the language of structural models, the laws correspond to the equations $\{f_i\}$ and the boundary conditions correspond to the state of the exogenous variables U . Thus, a

precondition for the validity of the predictive interpretation of a counterfactual statement is the assumption that U will remain the same at the time where our predictive claim is to be applied or tested.

This is best illustrated using the betting example of Section 19.4.1. The predictive interpretation of the counterfactual “Had I bet differently I would have lost a dollar” is the claim: “If my next bet is tails, I will lose a dollar”. For this claim to be valid, two invariants must be assumed: the payoff policy and the outcome of the coin. While the former is a plausible assumption in betting context, the latter would be realized in only rare circumstances. It is for this reason that the predictive utility of the statement “Had I bet differently I would have lost a dollar” is rather low, and some would even regard it as hind-sighted nonsense. (It is not hard however to imagine a lottery in which the payoff policy and the outcome of the random device remain constant for a short period of time, during which additional bets are accepted and processed. Most of those who play the stock market believe in strategies that allow an investor to quickly recover from a bad move.) At any rate, it is the persistence across time of U and $f(x, u)$ that endows counterfactual expressions with predictive power; take this persistence away, and the counterfactual loses its obvious economical utility.

I said “obvious” because there is an element of utility in counterfactuals that does not translate immediately to predictive payoff, and may explain, nevertheless, the ubiquity of counterfactuals in human discourse. I am thinking of explanatory value. Suppose, in the betting story, coins were tossed afresh for every bet. Is there no value whatsoever to the statement “Had I bet differently I would have lost a dollar?” I believe there is; it tells us that we are not dealing here with a whimsical bookie like the one who decides which way to spin our atoms and electrons, but one who at least glances at the bet, compares it to some standard, and decides a win or a Toss using a consistent policy. This information may not be very useful to us as players, but it may be useful to say state inspectors who come every so often to calibrate the gambling machines to ensure the State’s take of the profit. More significantly, it may be useful to us players, too, if we venture to cheat slightly, say by manipulating the trajectory of the coin, or by installing a tiny transmitter to tell us which way the coin landed. For such cheating to work, we should know the policy $y = f(x, u)$ and the statement “Had I bet differently I would have lost a dollar?” reveals important aspects of that policy.

Is it far fetched to argue for the merit of counterfactuals by hypothesizing unlikely situations where players cheat and rules are broken? I submit that such unlikely operations are the norm in gauging the explanatory value of sentences. In fact, it is the nature of any explanation, especially causal, that its utility be

amortized not over standard situations but, rather, over novel settings which require innovative manipulation of one's environment.

Recapping our discussion, we see that counterfactuals may earn predictive value under two conditions; (1) when the unobserved uncertainty-producing variables (U) remain constant (until our next prediction or action), (2) when the uncertainty-producing variables offer the potential of being observed sometime in the future (before our next prediction or action.) In both cases we also need to ensure that the outcome-producing mechanism $f(x, u)$ persists unaltered.

These conclusions raise interesting questions on the use of counterfactuals in microscopic phenomena, as none of these conditions holds for the type of uncertainty that we encounter in quantum theory. Heisenberg's dice is rolled afresh billions of times each second, and our measurement of u will never be fine enough to remove all uncertainty from the response equation $y = f(x, u)$. Thus, when we include quantum-level processes in our analysis we face a dilemma; either we disband all talk of counterfactuals (a strategy recommended by some researchers (Dawid 1997)) or we continue to use counterfactuals but limit their usage to situations where they assume empirical meaning. This amounts to keeping in the analysis only U 's that satisfy conditions (1) and (2) above. Instead of hypothesizing U 's that completely remove all uncertainties, we admit only those U 's that are either (1) persistent or (2) potentially observable.

Naturally, coarsening the granularity of the exogenous variables has its price tag; the mechanism equations $y = f(x, u)$ lose their deterministic character and should be made stochastic. Instead of constructing causal models from a set of deterministic equations $\{f_i\}$ we should consider models made up of stochastic functions $\{f_i^*\}$, where each f_i^* is a mapping from $V \cup U$ to some intrinsic probability distribution $P^*(v_i)$ over the states of V_i . This option lies beyond the scope of the present paper, but its basic character should follow from the three steps of abduction-action-deduction, abduction-action-deduction, outlined in Section 19.2.2.

References

- Aldrich, J.: 1993, 'Cowles' Exogeneity and Core Exogeneity', *Technical Report Discussion Paper 9308*, Department of Economics, University of Southampton, Southampton.
- Angrist, J. D., G. W. Imbens, and D. B. Rubin: 1996, 'Identification of Causal Effects Using Instrumental Variables (with Comments)', *Journal of the American Statistical Association* **91**, 444–472.
- Balke, A. and J. Pearl: 1994a, 'Counterfactual Probabilities: Computational Methods, Bounds and Applications', in R. Lopez de Mantaras and D. Poole (eds.), *Uncertainty in Artificial Intelligence 10*, Morgan Kaufmann, San Mateo, CA, pp. 46–54.

- Balke, A. and J. Pearl: 1994b, 'Probabilistic Evaluation of Counterfactual Queries', in *Proceedings of the Twelfth National Conference on Artificial Intelligence*, Volume I, MIT Press, Cambridge, MA, pp. 230–237.
- Balke, A. and J. Pearl: 1995, 'Counterfactuals and Policy Analysis in Structural Models', in P. Besnard and S. Hanks (eds.), *Uncertainty in Artificial Intelligence 11*, Morgan Kaufmann, San Francisco, CA, pp. 11–18.
- Balke, A. and J. Pearl: 1997, 'Nonparametric Bounds on Causal Effects from Partial Compliance Data', *Journal of the American Statistical Association* **92**, 1–6.
- Breslow, N. E. and N. E. Day: 1980, *Statistical Methods in Cancer Research Vol. 1; The Analysis of Case-Control Studies*, IARC, Lyon.
- Cartwright, N.: 1989, *Nature's Capacities and Their Measurement*, Clarendon, Oxford.
- Cheng, P. W.: 1997, 'From Covariation to Causation: A Causal Power Theory', *Psychological Review* **104**, 367–405.
- Cole, P.: 1997, 'Causality in Epidemiology Health Policy, and Law', *Journal of Marketing Research* **27**, 10279–10285.
- Dawid, A. P.: 1997, 'Causal Inference without Counterfactuals' (with discussion), *Journal of the American Statistical Association* **95**, 407–448.
- Dhrymes, P. J.: 1970, *Econometrics*, Springer-Verlag, New York.
- Eells, E.: 1991, *Probabilistic Causality*, Cambridge University Press, Cambridge.
- Engle, R. F., D.F. Hendry, and J.F. Richard: 1983, 'Exogeneity', *Econometrica* **51**, 277–304.
- Fine, K.: 1975, 'Review of Lewis' Counterfactuals', *Mind* **84**, 451–458.
- Fine, K.: 1985, *Reasoning with Arbitrary Objects*, B. Blackwell, New York.
- Finkelstein, M. O. and B. Levin: 1990, *Statistics for Lawyers*, Springer-Verlag, New York.
- Fisher, F. M.: 1970, 'A Correspondence Principle for Simultaneous Equations Models', *Econometrica* **38**, 73–92.
- Fleiss, J. L.: 1981, *Statistical Methods for Rates and Proportions* (second edition), John Wiley and Sons, New York.
- Galles, D. and J. Pearl: 1995, 'Testing Identifiability of Causal Effects', in P. Besnard and S. Hanks (eds.), *Uncertainty in Artificial Intelligence 11*, Morgan Kaufmann, San Francisco, CA, pp. 185–195. See also Pearl (2000), Chapter 4.
- Galles, D. and J. Pearl: 1997, 'Axioms of Causal Relevance', *Artificial Intelligence* **97**, 9–43.
- Galles, D. and J. Pearl: 1998, 'An Axiomatic Characterization of Causal Counterfactuals', *Foundations of Science* **3**, 151–182.
- Glymour, C.: 1998, 'Psychological and Normative Theories of Causal Power and Probabilities of Causes', in G. F. Cooper and S. Moral (eds.), *Uncertainty in Artificial Intelligence*, Morgan Kaufmann, San Francisco, CA, pp. 166–172.
- Goldszmidt, M. and J. Pearl: 1992, 'Rank-Based Systems: A Simple Approach to Belief Revision, Belief Update, and Reasoning about Evidence and Actions', in B. Nebel, C. Rich and W. Swartout (eds.), *Proceedings of the Third International Conference on*

- Knowledge Representation and Reasoning*, Morgan Kaufmann, San Francisco, CA, pp. 661–672.
- Good, I. J.: 1961, 'A Causal Calculus, I', *British Journal for the Philosophy of Science* **11**, 305–318.
- Good, I. J.: 1993, 'A Tentative Measure of Probabilistic Causation Relevant to the Philosophy of the Law', *J. Statist. Comput. and Simulation* **47**, 99–105.
- Greenland, S. and J. Robins: 1988, 'Conceptual Problems in the Definition and Interpretation of Attributable Fractions', *American Journal of Epidemiology* **128**, 1185–1197.
- Hall, N.: 1998, *Two Concepts of Causation*. Since published in John Collins, Ned Hall & Laurie Paul (eds.), *Causation and Counterfactuals*. MIT Press. pp. 225–276, 2004.
- Halpern, J. Y.: 1998, 'Axiomatizing Causal Reasoning', in G. F. Cooper and S. Moral (eds.), *Uncertainty in Artificial Intelligence*, Morgan Kaufmann, San Francisco, CA, pp. 202–210.
- Heckerman, D. and R. Shachter: 1995, 'Decision-Theoretic Foundations for Causal Reasoning', *Journal of Artificial Intelligence Research* **3**, 405–430.
- Hendry, D.: 1995, *Dynamic Econometrics*, Oxford University Press, New York.
- Hennekens, C. H. and J. E. Buring: 1987, *Epidemiology in Medicine*, Brown, Little, Boston.
- Hume, D.: 1748, *An Enquiry Concerning Human Understanding*, Open Court Press, LaSalle.
- Imbens, G. W.: 1997, 'Book Reviews', *Journal of Applied Econometrics* **12**.
- Kelsey J. L., A. S. Whittemore, A. S. Evans, and W. D. Thompson: 1987, *Methods in Observational Epidemiology*, Oxford University Press, New York.
- Khoury, M. J., W. D. Flanders, S. Greenland, and M. J. Adams: 1989, 'On the Measurement of Susceptibility in Epidemiology Studies', *American Journal of Epidemiology* **129**, 183–190.
- Kim, J.: 1971, 'Causes and Events: Mackie on Causation', *Journal of Philosophy* **68**, 426–471.
- Lewis, D.: 1979, 'Counterfactual Dependence and Time's Arrow', *Nous* **13**, 418–446.
- Lewis, D.: 1986, *Philosophical Papers*, Oxford University Press, New York.
- Mackie, J. L.: 1965, 'Causes and Conditions', *American Philosophical Quarterly* **2**, 261–264. Reprinted in E. Sosa and M. Tooley (eds.), *Causation*, Oxford University Press.
- Marschak, J.: 1950, 'Statistical Inference in Economics', in T. Koopmans (ed.), *Statistical Inference in Dynamic Economic Models*, John Wiley and Sons, New York, pp. 1–50.
- Michie, D.: 1997, 'Adapting Good's q Theory to the Causation of Individual Events'. Since published in *Machine Intelligence* **15**, 60–86, 1999.
- Mill, J. S.: 1843, *System of Logic*, Volume 1, John W Parker, London.
- Neyman, J.: 1923, 'On the Application of Probability Theory to Agricultural Experiments. Essays on Principles. Section 9', English Translation (1990), *Statistical Science* **5**(4), 465–480.
- Pearl, J.: 1993, 'Comment: Graphical Models, Causality and Interventions', *Statistical Science* **8**, 266–269.

- Pearl, J.: 1994, 'A Probabilistic Calculus of Actions', in R. Lopez de Mantaras and D. Poole (eds.), *Uncertainty in Artificial Intelligence 10*, Morgan Kaufmann, San Mateo, CA, pp. 454–462.
- Pearl, J.: 1995, 'Causal Diagrams for Experimental Research', *Biometrika* **82**, 669–710.
- Pearl, J.: 1996, 'Causation, Action, and Counterfactuals', in Y Shoham (ed.), *Theoretical Aspects of Rationality and Knowledge, Proceedings of the Sixth Conference*, Morgan Kaufmann, San Francisco, CA, pp. 51–73.
- Pearl, J.: 1998, 'On the Definition of Actual Cause', Technical Report R-259, Department of Computer Science, University of California, Los Angeles, CA. Also in Pearl (2000), Chapter 10.
- Pearl, J.: 2000, *Causality*, Cambridge University Press.
- Robertson, D. W.: 1997, 'The Common Sense of Cause in Fact', *Texas Law Review* **75**, 1765–1800.
- Robins, J. M. and S. Greenland, 1989, 'The Probability of Causation under a Stochastic Model for Individual Risk', *Biometrics* **45**, 1125–1138.
- Robins, J. M.: 1986, 'A New Approach to Causal Inference in Mortality Studies with a Sustained Exposure Period – Applications to Control of the Healthy Workers Survivor Effect', *Mathematical Modeling* **7**, 1393–1512.
- Robins, J. M.: 1987, 'A Graphical Approach to the Identification and Estimation of Causal Parameters in Mortality Studies with Sustained Exposure Period', *Journal of Chronic Diseases* **40**, 139–161S.
- Rosenbaum, P. and D. Rubin: 1983, 'The Central Role of Propensity Score in Observational Studies for Causal Effects', *Biometrika* **70**, 41–55.
- Rubin, D. B.: 1974, 'Estimating Causal Effects of Treatments in Randomized and Non-randomized Studies', *Journal of Educational Psychology* **66**, 688–701.
- Schlesselman, J. J.: 1982, *Case-Control Studies: Design Conduct Analysis*, Oxford University Press, New York.
- Shep, M. C.: 1958, 'Shall We Count the Living or the Dead?', *New England Journal of Medicine* **259**, 1210–1214.
- Simon, H. A. and N. Rescher: 1966, 'Cause and Counterfactual', *Philosophy and Science* **33**, 323–340.
- Simon, H. A.: 1953, 'Causal Ordering and Identifiability', in Wm. C. Hood and T. C. Koopmans (eds.), *Studies in Econometric Methods*, John Wiley and Sons, New York, pp. 49–74.
- Skyrms, B.: 1980, *Causal Necessity*, Yale University Press, New Haven, CT.
- Sobel, M. E.: 1990, 'Effect Analysis and Causation in Linear Structural Equation Models', *Psychometrika* **55**, 495–515.
- Spirtes, P., C. Glymour, and R. Scheines: 1993, *Causation, Prediction, and Search*, Springer-Verlag, New York.

Strotz, R. H. and H. O. A. Wold: 1960, 'Recursive versus Nonrecursive Systems: An Attempt at Synthesis', *Econometrica* **28**, 417–427.

Suppes, P.: 1970, *A Probabilistic Theory of Causality*, North-Holland Publishing Co., Amsterdam.

Thomason, R. and A. Gupta: 1980, 'A Theory of Conditionals in the Context of Branching Time', *Philosophical Review* **88**, 65–90.

20 Direct and Indirect Effects

Judea Pearl

Abstract

The direct effect of one event on another can be defined and measured by holding constant all intermediate variables between the two. Indirect effects present conceptual and practical difficulties (in nonlinear models), because they cannot be isolated by holding certain variables constant. This paper presents a new way of defining the effect transmitted through a restricted set of paths, without controlling variables on the remaining paths. This permits the assessment of a more natural type of direct and indirect effects, one that is applicable in both linear and nonlinear models and that has broader policy-related interpretations. The paper establishes conditions under which such assessments can be estimated consistently from experimental and nonexperimental data, and thus extends path-analytic techniques to nonlinear and nonparametric models.

20.1 Introduction

The distinction between total, direct, and indirect effects is deeply entrenched in causal conversations, and attains practical importance in many applications, including policy decisions, legal definitions and health care analysis. Structural equation modeling (SEM) (Goldberger 1972), which provides a methodology of defining and estimating such effects, has been restricted to linear analysis, and no comparable methodology has been devised to extend these capabilities to models

Originally published in *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, San Francisco, CA: Morgan Kaufmann, 411–420, 2001.

Copyright 2001 Morgan Kaufmann. Republished with permission of Elsevier.

involving nonlinear dependencies,¹ as those commonly used in AI applications (Hagenaars 1993, p. 17).

The causal relationship that is easiest to interpret, define and estimate is the *total effect*. Written as $P(Y_x = y)$, the total effect measures the probability that response variable Y would take on the value y when X is set to x by external intervention.² This probability function is what we normally assess in a controlled experiment in which X is randomized and in which the distribution of Y is estimated for each level x of X .

In many cases, however, this quantity does not adequately represent the target of investigation and attention is focused instead on the direct effect of X on Y . The term “direct effect” is meant to quantify an influence that is not mediated by other variables in the model or, more accurately, the sensitivity of Y to changes in X while all other factors in the analysis are held fixed. Naturally, holding those factors fixed would sever all causal paths from X to Y with the exception of the direct link $X \rightarrow Y$, which is not intercepted by any intermediaries.

Indirect effects cannot be defined in this manner, because it is impossible to hold a set of variables constant in such a way that the effect of X on Y measured under those conditions would circumvent the direct pathway, if such exists. Thus, the definition of indirect effects has remained incomplete, and, save for asserting inequality between direct and total effects, the very concept of “indirect effect” was deemed void of operational meaning (Pearl 2000, p. 165).

This paper shows that it is possible to give an operational meaning to both direct and indirect effects without fixing variables in the model, thus extending the applicability of these concepts to nonlinear and nonparametric models. The proposed generalization is based on a more subtle interpretation of “effects”, here called “descriptive” (see Section 20.2.2), which concerns the action of causal forces under natural, rather than experimental conditions, and provides answers to a broader class of policy-related questions. This interpretation yields the standard path-coefficients in linear models, but leads to different formal definitions and different estimation procedures of direct and indirect effects in nonlinear models.

Following a conceptual discussion of the descriptive and prescriptive interpretations (Section 20.2.2), Section 20.2.3 illustrates their distinct roles in decision-making contexts, while Section 20.2.4 discusses the descriptive basis and

1. A notable exception is the counterfactual analysis of Robins and Greenland (1992) which is applicable to nonlinear models, but does not incorporate path-analytic techniques.

2. The subscripted notation Y_x is borrowed from the potential-outcome framework of Rubin (1974). Pearl (2000) used, interchangeably, $P_x(y)$, $P(y|do(x))$, $P(y|\hat{x})$, and $P(y_x)$, and showed their equivalence to probabilities of subjunctive conditionals: $P((X = x) \square\rightarrow (Y = y))$ (Lewis 1973).

policy implications of indirect effects. Sections 20.3.2 and 20.3.3 provide, respectively, mathematical formulation of the prescriptive and descriptive interpretations of direct effects, while Section 20.3.4 establishes conditions under which the descriptive (or “natural”) interpretation can be estimated consistently from either experimental or nonexperimental data. Sections 20.3.5 and 20.3.6 extend the formulation and identification analysis to indirect effects. In Section 20.3.7, we generalize the notion of indirect effect to *path-specific effects*, that is, effects transmitted through any specified set of paths in the model.

20.2 Conceptual Analysis

20.2.1 Direct versus Total Effects

A classical example of the ubiquity of direct effects (Hesslow 1976) tells the story of a birth-control pill that is suspect of producing thrombosis in women and, at the same time, has a negative indirect effect on thrombosis by reducing the rate of pregnancies (pregnancy is known to encourage thrombosis). In this example, interest is focused on the direct effect of the pill because it represents a stable biological relationship that, unlike the total effect, is invariant to marital status and other factors that may affect women’s chances of getting pregnant or of sustaining pregnancy. This invariance makes the direct effect transportable across cultural and sociological boundaries and, hence, a more useful quantity in scientific explanation and policy analysis.

Another class of examples involves legal disputes over race or sex discrimination in hiring. Here, neither the effect of sex or race on applicants’ qualification nor the effect of qualification on hiring are targets of litigation. Rather, defendants must prove that sex and race do not *directly* influence hiring decisions, whatever indirect effects they might have on hiring by way of applicant qualification. This is made quite explicit in the following court ruling:

“The central question in any employment-discrimination case is whether the employer would have taken the same action had the employee been of a different race (age, sex, religion, national origin etc.) and everything else had been the same.” (Carson versus Bethlehem Steel Corp., 70 FEP Cases 921, 7th Cir. (1996), Quoted in Gastwirth 1997.)

Taking this criterion as a guideline, the direct effect of X on Y (in our case $X =$ gender $Y =$ hiring) can roughly be defined as the response of Y to change in X (say from $X = x^*$ to $X = x$) while keeping all other accessible variables at their

initial value, namely, the value they would have attained under $X = x^*$.³ This doubly-hypothetical criterion will be given precise mathematical formulation in Section 20.3, using the language and semantics of structural counterfactuals (Pearl 2000; Chapter 7).

As a third example, one that illustrates the policymaking ramifications of direct and total effects, consider a drug treatment that has a side effect – headache. Patients who suffer from headache tend to take aspirin which, in turn may have its own effect on the disease or, may strengthen (or weaken) the impact of the drug on the disease. To determine how beneficial the drug is to the population as a whole, under existing patterns of aspirin usage, the total effect of the drug is the target of analysis, and the difference $P(Y_x = y) - P(Y_{x^*} = y)$ may serve to assist the decision, with x and x^* being any two treatment levels. However, to decide whether aspirin should be encouraged or discouraged during the treatment, the *direct* effect of the drug on the disease, both with aspirin and without aspirin, should be the target of investigation. The appropriate expression for analysis would then be the difference $P(Y_{xz} = y) - P(Y_{x^*z} = y)$, where z stands for any specified level of aspirin intake.

In linear systems, direct effects are fully specified by the corresponding path coefficients, and are independent of the values at which we hold the intermediate variables (Z in our examples). In nonlinear systems, those values would, in general, modify the effect of X on Y and thus should be chosen carefully to represent the target policy under analysis. This lead to a basic distinction between two types of conceptualizations: *prescriptive* and *descriptive*.

20.2.2 Descriptive versus Prescriptive Interpretation

We will illustrate this distinction using the treatment-aspirin example described in the last section. In the prescriptive conceptualization, we ask whether a specific untreated patient would improve if treated, while holding the aspirin intake fixed at some predetermined level, say $Z = z$. In the descriptive conceptualization, we ask again whether the untreated patient would improve if treated, but now we hold the aspirin intake fixed at whatever level the patient currently consumes under no-treatment condition. The difference between these two conceptualizations lies in whether we wish to account for the natural relationship between the direct and the mediating cause (that is, between treatment and aspirin) or to modify that relationship to match policy objectives. We call the effect computed from the descriptive perspective the *natural* effect, and the one computed from the prescriptive perspective the *controlled* effect.

3. Robins and Greenland (1992) have adapted essentially the same criterion (phrased differently) for their interpretation of “direct effect” in epidemiology.

Consider a patient who takes aspirin if and only if treated, and for whom the treatment is effective only when aspirin is present. For such a person, the treatment is deemed to have no natural direct effect (on recovery), because, by keeping the aspirin at the current, pre-treatment level of zero, we ensure that the treatment effect would be nullified. The controlled direct effect, however, is nonzero for this person, because the efficacy of the treatment would surface when we fix the aspirin intake at non-zero level. Note that the descriptive formulation requires knowledge of the individual natural behavior—in our example, whether the untreated patient actually uses aspirin—while the prescriptive formulation requires no such knowledge.

This difference becomes a major stumbling block when it comes to estimating *average* direct effects in a population of individuals. At the population level, the prescriptive formulation is pragmatic; we wish to predict the difference in recovery rates between treated and untreated patients when a prescribed dose of aspirin is administered to all patients in the population—the actual consumption of aspirin under uncontrolled conditions need not concern us. In contrast, the descriptive formulation is attributional; we ask whether an observed improvement in recovery rates (again, between treated and untreated patients) is attributable to the treatment itself, as opposed to preferential use of aspirin among treated patients. To properly distinguish between these two contributions, we therefore need to measure the improvement in recovery rates while making each patient take the same level of aspirin that he/she took before treatment. However, as [Robins and Greenland \(1992\)](#) pointed out, such control over individual behavior would require testing the same group of patients twice (i.e., under treatment and no treatment conditions), and cannot be administered in experiments with two different groups, however randomized. (There is no way to determine what level of aspirin an untreated patient would take if treated, unless we actually treat that patient and, then, this patient could no longer be eligible for the untreated group.) Since repeatable tests on the same individuals are rarely feasible, the descriptive measure of the direct effect is not generally estimable from standard experimental studies. In Section 20.3.4 we will analyze what additional assumptions are required for consistently estimating this measure, the *average natural direct effect*, from either experimental or observational studies.

20.2.3 Policy Implications of the Descriptive Interpretation

Why would anyone be interested in assessing the average natural direct effect? Assume that the drug manufacturer is considering ways of eliminating the adverse side-effect of the drug, in our case, the headache. A natural question to ask is

whether the drug would still retain its effectiveness in the population of interest. The controlled direct effect would not give us the answer to this question, because it refers to a specific aspirin level, taken uniformly by all individuals. Our target population is one where aspirin intake varies from individual to individual, depending on other factors beside drug-induced headache, factors which may also cause the effectiveness of the drug to vary from individual to individual. Therefore, the parameter we need to assess is the average natural direct effect, as described in the Subsection 20.2.2.

This example demonstrates that the descriptive interpretation of direct effects is not purely “descriptive”; it carries a definite operational implications, and answers policy-related questions of practical significance. Moreover, note that the policy question considered in this example cannot be represented in the standard syntax of $do(x)$ operators—it does not involve fixing any of the variables in the model but, rather, modifying the causal paths in the model. Even if “headache” were a genuine variable in our model, the elimination of drug-induced headache is not equivalent to setting “headache” to zero, since a person might get headache for reason other than the drug. Instead, the policy option involves the de-activation of the causal path from “drug” to “headache”.

In general, the average natural direct effect would be of interest in evaluating policy options of a more refined variety, ones that involve, not merely fixing the levels of the variables in the model, but also determining how these levels would influence one another.

Typical examples of such options involve choosing the *manner* (e.g., instrument, or timing) in which a given decision is implemented, or choosing the agents that should be *informed*, about the decision. A firm often needs to assess, for example, whether it would be worthwhile to conceal a certain decision from a competitor. This amounts, again, to evaluating the natural direct effect of the decision in question, unmediated by the competitor’s reaction. Theoretically, such policy options could conceivably be represented as (values of) variables in a more refined model, for example one where the concept “the effect of treatment on headache” would be given a variable name, and where the manufacturer decision to eliminate side-effects would be represented by fixing this hypothetical variable to zero. The analysis of this paper shows that such unnatural modeling techniques can be avoided, and that important nonstandard policy questions can be handled by standard models, where variables stands for directly measurable quantities.

20.2.4 Descriptive Interpretation of Indirect Effects

The descriptive conception of direct effects can easily be transported to the formulation of indirect effects; oddly, the prescriptive formulation is not transportable.

Returning to our treatment-aspirin example, if we wish to assess the *natural* indirect effect of treatment on recovery for a specific patient, we withhold treatment and ask, instead, whether that patient would recover if given as much aspirin as he/she would have taken if he/she had been under treatment. In this way, we insure that whatever changes occur in the patient's condition are due to treatment-induced aspirin consumption and not to the treatment itself. Similarly, at the population level, the natural indirect effect of the treatment is interpreted as the improvement in recovery rates if we were to withhold treatment from all patients but, instead, let each patient take the same level of aspirin that he/she would have taken under treatment. As in the descriptive formulation of direct effects, this hypothetical quantity involves nested counterfactuals and will be identifiable only under special circumstances.

The prescriptive formulation has no parallel in indirect effects, for reasons discussed in the introduction section; there is no way of preventing the direct effect from operating by holding certain variables constant. We will see that, in linear systems, the descriptive and prescriptive formulations of direct effects lead, indeed, to the same expression in terms of path coefficients. The corresponding linear expression for indirect effects, computed as the difference between the total and direct effects, coincides with the descriptive formulation but finds no prescriptive interpretation.

The operational implications of indirect effects, like those of natural direct effect, concern nonstandard policy options. Although it is impossible, by controlling variables, to block a direct path (i.e., a single edge), if such exists, it is nevertheless possible to block such a path by more refined policy options, ones that deactivate the direct path through the manner in which an action is taken or through the mode by which a variable level is achieved. In the hiring discrimination example, if we make it illegal to question applicants about their gender, (and if no other indication of gender are available to the hiring agent), then any residual sex preferences (in hiring) would be attributable to the indirect effect of sex on hiring. A policy maker might well be interested in predicting the magnitude of such preferences from data obtained prior to implementing the no-questioning policy, and the average indirect effect would then provide the sought for prediction. A similar refinement applies in the firm-competitor example of the preceding subsection. A firm might wish to assess, for example, the economical impact of bluffing a competitor into believing that a certain decision has been taken by the firm, and this could be implemented by (secretly) instructing certain agents to ignore the decision. In both cases, our model may not be sufficiently detailed to represent such policy options in the form of variable fixing (e.g., the agents may not be represented as intermediate nodes between the decision and its effect) and the task

amounts then to evaluating the average natural indirect effects in a coarse-grain model, where a direct link exists between the decision and its outcome.

20.3 Formal Analysis

20.3.1 Notation

Throughout our analysis we will let X be the control variable (whose effect we seek to assess), and let Y be the response variable. We will let Z stand for the set of all intermediate variables between X and Y which, in the simplest case considered, would be a single variable as in Figure 20.1(a). Most of our results will still be valid if we let Z stand for any set of such variables, in particular, the set of Y 's parents excluding X .

We will use the counterfactual notation $Y_x(u)$ to denote the value that Y would attain in unit (or situation) $U = u$ under the control regime $do(X = x)$. See Pearl (2000, Chapter 7) for formal semantics of these counterfactual utterances. Many concepts associated with direct and indirect effect require comparison to a reference value of X , that is, a value relative to which we measure changes. We will designate this reference value by x^* .

20.3.2 Controlled Direct Effects (review)

Definition 20.1 Controlled unit-level direct-effect; qualitative

A variable X is said to have a controlled direct effect on variable Y in model M and situation $U = u$ if there exists a setting $Z = z$ of the other variables in the model and two values of X , x^* and x , such that

$$Y_{x^*z}(u) \neq Y_{xz}(u) \quad (20.1)$$

In words, the value of Y under $X = x^*$ differs from its value under $X = x$ when we keep all other variables Z fixed at z . If condition (20.1) is satisfied for some z , we say that the transition event $X = x$ has a controlled direct-effect on Y , keeping the reference point $X = x^*$ implicit.

Clearly, confining Z to the parents of Y (excluding X) leaves the definition unaltered.

Definition 20.2 Controlled unit-level direct-effect; quantitative

Given a causal model M with causal graph G , the controlled direct effect of $X = x$ on Y in unit $U = u$ and setting $Z = z$ is given by

$$CDE_z(x, x^*; Y, u) = Y_{xz}(u) - Y_{x^*z}(u) \quad (20.2)$$

where Z stands for all parents of Y (in G) excluding X .

Alternatively, the ratio $Y_{xz}(u)/Y_{x^*z}(u)$, the proportional difference $(Y_{xz}(u) - Y_{x^*z}(u))/Y_{x^*z}(u)$, or some other suitable relationship might be used to quantify the magnitude of the direct effect; the difference is by far the most common measure, and will be used throughout this paper.

Definition 20.3 Average controlled direct effect

Given a probabilistic causal model $\langle M, P(u) \rangle$, the controlled direct effect of event $X = x$ on Y is defined as:

$$CDE_z(x, x^*; Y) = E(Y_{xz} - Y_{x^*z}) \quad (20.3)$$

where the expectation is taken over u .

The distribution $P(Y_{xz} = y)$ can be estimated consistently from experimental studies in which both X and Z are randomized. In nonexperimental studies, the identification of this distribution requires that certain “no-confounding” assumptions hold true in the population tested. Graphical criteria encapsulating these assumptions are described in Pearl (2000, Sections 4.3 and 4.4).

20.3.3 Natural Direct Effects: Formulation

Definition 20.4 Unit-level natural direct effect; qualitative

An event $X = x$ is said to have a natural direct effect on variable Y in situation $U = u$ if the following inequality holds

$$Y_{x^*}(u) \neq Y_{x, Z_{x^*}(u)}(u) \quad (20.4)$$

In words, the value of Y under $X = x^*$ differs from its value under $X = x$ even when we keep Z at the same value ($Z_{x^*}(u)$) that Z attains under $X = x^*$.

We can easily extend this definition from events to variables by defining X as having a natural direct effect on Y (in model M and situation $U = u$) if there exist two values, x^* and x , that satisfy (20.4). Note that this definition no longer requires that we specify a value z for Z ; that value is determined naturally by the model, once we specify x , x^* , and u . Note also that condition (20.4) is a direct literal translation of the court criterion of sex discrimination in hiring (Section 20.2.1) with $X = x^*$ being a male, $X = x$ a female, $Y = 1$ a decision to hire, and Z the set of all other attributes of individual u .

If one is interested in the magnitude of the natural direct effect, one can take the difference

$$Y_{x, Z_{x^*}(u)}(u) - Y_{x^*}(u) \quad (20.5)$$

and designate it by the symbol $NDE(x, x^*; Y, u)$ (acronym for Natural Direct Effect). If we are further interested in assessing the average of this difference in a population of units, we have:

Definition 20.5 Average natural direct effect

The average natural direct effect of event $X = x$ on a response variable Y , denoted $NDE(x, x^*; Y)$, is defined as

$$NDE(x, x^*; Y) = E(Y_{x, Z_{x^*}}) - E(Y_{x^*}) \quad (20.6)$$

Applied to the sex discrimination example of Section 20.2.1 (with $x^* = \text{male}$, $x = \text{female}$, $y = \text{hiring}$, $z = \text{qualifications}$), Equation (20.6) measures the expected change in male hiring, $E(Y_{x^*})$, if employers were instructed to treat males' applications as though they were females'.

20.3.4 Natural Direct Effects: Identification

As noted in Section 20.2, we cannot generally evaluate the average natural direct-effect from empirical data. Formally, this means that Equation (20.6) is not reducible to expressions of the form

$$P(Y_x = y) \text{ or } P(Y_{xz} = y);$$

the former governs the causal effect of X on Y (obtained by randomizing X) and the latter governs the causal effect of X and Z on Y (obtained by randomizing both X and Z).

We now present conditions under which such reduction is nevertheless feasible.

Theorem 20.1 Experimental identification

If there exists a set W of covariates, nondescendants of X or Z , such that

$$Y_{xz} \perp\!\!\!\perp Z_{x^*} | W \quad \text{for all } z \text{ and } x \quad (20.7)$$

(read: Y_{xz} is conditionally independent of Z_{x^*} , given W), then the average natural direct-effect is experimentally identifiable, and it is given by

$$NDE(x, x^*; Y) = \sum_{w, z} [E(Y_{xz} | w) - E(Y_{x^*z} | w)] P(Z_{x^*} = z | w) P(w). \quad (20.8)$$

Proof. The first term in (20.6) can be written

$$E(Y_{x, Z_{x^*}}) = \sum_w \sum_z E(Y_{xz} | Z_{x^*} = z, W = w) P(Z_{x^*} = z | W = w) P(W = w). \quad (20.9)$$

Using (20.7), we obtain:

$$E(Y_{x,Z_{x^*}}) = \sum_w \sum_z E(Y_{xz} = y|W = w)P(Z_{x^*} = z|W = w)P(W = w). \quad (20.10)$$

Each factor in (20.10) is identifiable; $E(Y_{xz} = y|W = w)$, by randomizing X and Z for each value of W , and $P(Z_{x^*} = z|W = w)$ by randomizing X for each value of W . This proves the assertion in the theorem. Substituting (20.10) into (20.6) and using the law of composition $E(Y_{x^*}) = E(Y_{x^*Z_{x^*}})$ (Pearl 2000, p. 229) gives (20.8), and completes the proof of Theorem 20.1. ■

The conditional independence relation in Equation (20.7) can easily be verified from the causal graph associated with the model. Using a graphical interpretation of counterfactuals (Pearl 2000, p. 214-5), this relation reads:

$$(Y \perp\!\!\!\perp Z|W)_{G_{\underline{XZ}}} \quad (20.11)$$

In words, W d -separates Y from Z in the graph formed by deleting all (solid) arrows emanating from X and Z .

Figure 20.1(a) illustrates a typical graph associated with estimating the direct effect of X on Y . The identifying subgraph is shown in Figure 20.1(b), and illustrates how W d -separates Y from Z . The separation condition in (20.11) is somewhat stronger than (20.7), since the former implies the latter for every pair of values, x and x^* , of X (see (Pearl 2000, p. 214)). Likewise, condition (20.7) can be relaxed in several ways. However, since assumptions of counterfactual independencies can be meaningfully substantiated only when cast in structural form (Pearl 2000, p. 244–5), graphical conditions will be the target of our analysis.

The identification of the natural direct effect from *nonexperimental* data requires stronger conditions. From Equation (20.8) we see that it is sufficient to identify the conditional probabilities of two counterfactuals: $P(Y_{xz} = y|W = w)$ and $P(Z_{x^*} = z|W = w)$, where W is any set of covariates that satisfies Equation (20.7) (or (20.11)). This yields the following criterion for identification:

Theorem 20.2 Nonexperimental identification

The average natural direct-effect $NDE(x, x^; Y)$ is identifiable in nonexperimental studies if there exists a set W of covariates, nondescendants of X or Z , such that, for all values z and x we have:*

- (i) $Y_{xz} \perp\!\!\!\perp Z_{x^*}|W$
- (ii) $P(Y_{xz} = y|W = w)$ is identifiable
- (iii) $P(Z_{x^*} = z|W = w)$ is identifiable

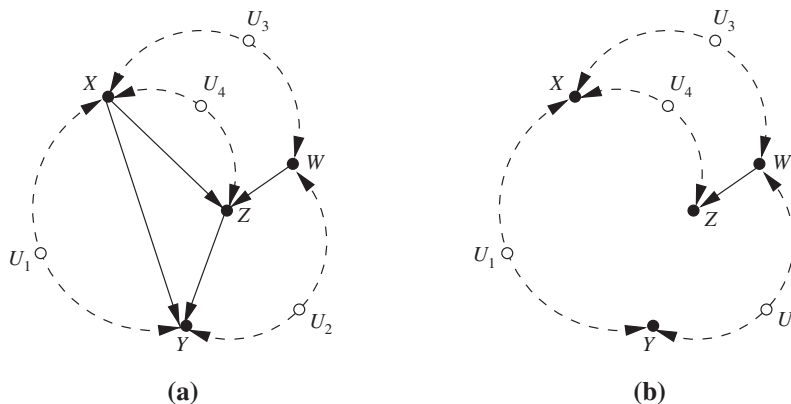


Figure 20.1 (a) A causal model with latent variables (U 's) where the natural direct effect can be identified in experimental studies. (b) The subgraph G_{XZ} illustrating the criterion of experimental identifiability (Equation 20.11): W d -separates Y from Z .

Moreover, if conditions (i)-(iii) are satisfied, the natural direct effect is given by (20.8).

Explicating these identification conditions in graphical terms (using Theorem 4.41 in (Pearl 2000)) yields the following corollary:

Corollary 20.1 Graphical identification criterion

The average natural direct-effect $NDE(x, x^*; Y)$ is identifiable in nonexperimental studies if there exist four sets of covariates, W_0, W_1, W_2 , and W_3 , such that

- (i) $(Y \perp\!\!\!\perp Z | W_0)_{G_{XZ}}$
- (ii) $(Y \perp\!\!\!\perp X | W_0, W_1)_{G_{XZ}}$
- (iii) $(Y \perp\!\!\!\perp Z | X, W_0, W_1, W_2)_{G_Z}$
- (iv) $(Z \perp\!\!\!\perp X | W_0, W_3)_{G_X}$
- (v) W_0, W_1 , and W_3 contain no descendant of X and W_2 contains no descendant of Z .

(Remark: G_{XZ} denotes the graph formed by deleting (from G) all arrows emanating from X or entering Z .)

As an example for applying these criteria, consider Figure 20.1(a), and assume that all variables (including the U 's) are observable. Conditions (i)-(iv) of Corollary 20.1 are satisfied if we choose:

$$W_0 = \{W\}, W_1 = \{U_1, U_2\}, W_2 = \emptyset \text{ and } W_3 = \{U_4\}$$

or, alternatively,

$$W_0 = \{U_2\}, W_1 = \{U_1\}, W_2 = \emptyset \text{ and } W_3 = \{U_3, U_4\}.$$

It is instructive to examine the form that expression (20.8) takes in Markovian models, (that is, acyclic models with independent error terms) where condition (20.7) is always satisfied with $W = \emptyset$, since Y_{xz} is independent of all variables in the model. In Markovian models, we also have the following three relationships:

$$P(Y_{xz} = y) = P(y|x, z) \quad (20.12)$$

since $X \cup Z$ is the set of Y 's parents,

$$P(Z_{x^*} = z) = \sum_s P(z|x^*, s)P(s), \quad (20.13)$$

$$P(Y_{x, Z_{x^*}} = y) = \sum_s \sum_z P(y|x, z)P(z|x^*, s)P(s) \quad (20.14)$$

where S stands for the parents of Z , excluding X , or any other set satisfying the back-door criterion (Pearl 2000, p. 79). This yields the following corollary of Theorem 20.1:

Corollary 20.2 *The average natural direct effect in Markovian models is identifiable from nonexperimental data, and it is given by*

$$NDE(x, x^*; Y) = \sum_s \sum_z [E(Y|x, z) - E(Y|x^*, z)]P(z|x^*, s)P(s) \quad (20.15)$$

where S stands for any set satisfying the back-door criterion between X and Z .

Equation (20.15) follows by substituting (20.14) into (20.6) and using the identity $E(Y_{x^*}) = E(Y_{x^* Z_{x^*}})$.

Further insight can be gained by examining simple Markovian models in which the effect of X on Z is not confounded, that is,

$$P(Z_{x^*} = z) = P(z|x^*). \quad (20.16)$$

In such models, a simple version of which is illustrated in Figure 20.2(b), Equation (20.13) can be replaced by Equation (20.16) and Equation (20.15) simplifies to

$$NDE(x, x^*; Y) = \sum_z [E(Y|x, z) - E(Y|x^*, z)]P(z|x^*). \quad (20.17)$$

This expression has a simple interpretation as a weighted average of the controlled direct effect $E(Y|x, z) - E(Y|x^*, z)$, where the intermediate value z is chosen according to its distribution under x^* .

20.3.5 Natural Indirect Effects: Formulation

As we discussed in Section 20.2.4, the prescriptive formulation of “controlled direct effect” has no parallel in indirect effects; we therefore use the descriptive formulation, and define *natural* indirect effects at both the unit and population levels. Lacking the controlled alternative, we will drop the title “natural” from discussions of indirect effects, unless it serves to convey a contrast.

Definition 20.6 **Unit-level indirect effect; qualitative**

An event $X = x$ is said to have an indirect effect on variable Y in situation $U = u$ if the following inequality holds

$$Y_{x^*}(u) \neq Y_{x^*,Z_x(u)}(u). \tag{20.18}$$

In words, the value of Y changes when we keep X fixed at its reference level $X = x^*$ and change Z to a new value, $Z_x(u)$, the same value that Z would attain under $X = x$.

Taking the difference between the two sides of Equation (20.18), we can define the unit level indirect effect as

$$NIE(x, x^*; Y, u) = Y_{x^*,Z_x(u)}(u) - Y_{x^*}(u) \tag{20.19}$$

and proceed to define its average in the population:

Definition 20.7 **Average indirect effect**

The average indirect effect of event $X = x$ on variable Y , denoted $NIE(x, x^*; Y)$, is defined as

$$NIE(x, x^*; Y) = E(Y_{x^*,Z_x}) - E(Y_{x^*}) \tag{20.20}$$

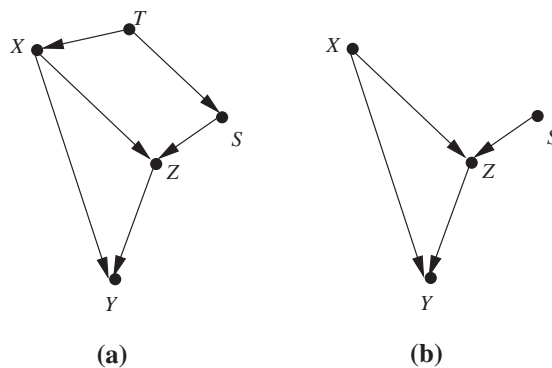


Figure 20.2 Simple Markovian models for which the natural direct effect is given by Equation (20.15) (for (a)) and Equation (20.17) (for (b)).

Comparing Equations (20.6) and (20.20), we see that the indirect effect associated with the transition from x^* to x is closely related to the natural direct effect associated with the reverse transition, from x to x^* . In fact, recalling that the difference $E(Y_x) - E(Y_{x^*})$ equals the total effect of $X = x$ on Y ,

$$TE(x, x^*; Y) = E(Y_x) - E(Y_{x^*}) \quad (20.21)$$

we obtain the following theorem:

Theorem 20.3 *The total, direct and indirect effects obey the following relationships*

$$TE(x, x^*; Y) = NIE(x, x^*; Y) - NDE(x^*, x; Y) \quad (20.22)$$

$$TE(x, x^*; Y) = NDE(x, x^*; Y) - NIE(x^*, x; Y) \quad (20.23)$$

In words, the total effect (on Y) associated with the transition from x^ to x is equal to the difference between the indirect effect associated with this transition and the (natural) direct effect associated with the reverse transition, from x to x^* .*

As strange as these relationships appear, they produce the standard, additive relation

$$TE(x, x^*; Y) = NIE(x, x^*; Y) + NDE(x, x^*; Y) \quad (20.24)$$

when applied to linear models. The reason is clear; in linear systems the effect of the transition from x^* to x is proportional to $x - x^*$, hence it is always equal and of opposite sign to the effect of the reverse transition. Thus, substituting in (20.22) (or (20.23)), yields (20.24).

20.3.6 Natural Indirect Effects: Identification

Equations (20.22) and (20.23) show that the indirect effect is identified whenever both the total and the (natural) direct effect are identified (for all x and x^*). Moreover, the identification conditions and the resulting expressions for indirect effects are identical to the corresponding ones for direct effects (Theorems 20.1 and 20.2), save for a simple exchange of the indices x and x^* . This is explicated in the following theorem.

Theorem 20.4 *If there exists a set W of covariates, nondescendants of X or Z , such that*

$$Y_{x^*z} \perp\!\!\!\perp Z_x | W \quad (20.25)$$

for all x and z , then the average indirect-effect is experimentally identifiable, and it is given by

$$NIE(x, x^*; Y) = \sum_{w,z} E(Y_{x^*z}|w)[P(Z_x = z|w) - P(Z_{x^*} = z|w)]P(w). \quad (20.26)$$

Moreover, the average indirect effect is identified in nonexperimental studies whenever the following expressions are identified for all z and w :

$$E(Y_{x^*z}|w), P(Z_x = z|w) \text{ and } P(Z_{x^*} = z|w),$$

with W satisfying Equation (20.25).

In the simple Markovian model depicted in Figure 20.2(b), Equation (20.26) reduces to

$$NIE(x, x^*; Y) = \sum_z E(Y|x^*, z)[P(z|x) - P(z|x^*)] \quad (20.27)$$

Contrasting Equation (20.27) with Equation (20.17), we see that the expression for the indirect effect fixes X at the reference value x^* , and lets z vary according to its distribution under the post-transition value of $X = x$. The expression for the direct effect fixes X at x , and lets z vary according to its distribution under the reference conditions $X = x^*$.

Applied to the sex discrimination example of Section 20.2.1, Equation (20.27) measures the expected change in male hiring, $E(Y_{x^*})$, if males were trained to acquire (in distribution) equal qualifications ($Z = z$) as those of females ($X = x$).

20.3.7 General Path-specific Effects

The analysis of the last section suggests that path-specific effects can best be understood in terms of a *path-deactivation process*, where a selected set of paths, rather than nodes, are forced to remain inactive during the transition from $X = x^*$ to $X = x$. In Figure 20.3, for example, if we wish to evaluate the effect of X on Y transmitted by the subgraph, $g : X \rightarrow Z \rightarrow W \rightarrow Y$, we cannot hold Z or W constant, for both must vary in the process. Rather, we isolate the desired effect by fixing the appropriate subset of arguments in each equation. In other words, we replace x with x^* in the equation for W , and replace z with $z^*(u) = Z_{x^*}(u)$ in the equation for Y . This amounts to creating a new model, in which each structural function f_i in M is replaced with a new function of a smaller set of arguments, since some of the arguments are replaced by constants. The following definition expresses this idea formally.

Definition 20.8 path-specific effect

Let G be the causal graph associated with model M , and let g be an edge-subgraph of G containing the paths selected for effect analysis. The g -specific effect of x on Y (relative to reference x^*) is defined as the total effect of x on Y in a modified model M_g^* formed as follows. Let each parent set PA_i in G be partitioned into two parts

$$PA_i = \{PA_i(g), PA_i(\bar{g})\} \quad (20.28)$$

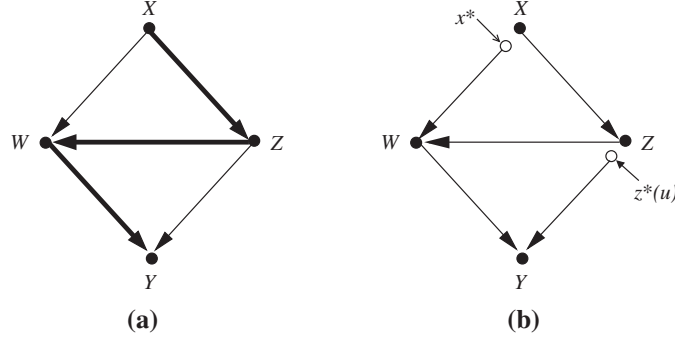


Figure 20.3 The path-specific effect transmitted through $X \rightarrow Z \rightarrow W \rightarrow Y$ (heavy lines) in (a) is equal to the total effect transmitted through the model in (b), treating x^* and $z^*(u)$ as constants. (By convention, u is not shown in the diagram.)

where $PA_i(g)$ represents those members of PA_i that are linked to X_i in g , and $PA_i(\bar{g})$ represents the complementary set, from which there is no link to X_i in g . We replace each function $f_i(pa_i, u)$ with a new function $f_i^*(pa_i, u; g)$, defined as

$$f_i^*(pa_i, u; g) = f_i(pa_i(g), pa_i^*(\bar{g}), u) \quad (20.29)$$

where $pa_i^*(\bar{g})$ stands for the values that the variables in $PA_i(\bar{g})$ would attain (in M and u) under $X = x^*$ (that is, $pa_i^*(\bar{g}) = PA_i(\bar{g})_{x^*}$). The g -specific effect of x on Y , denoted $SE_g(x, x^*; Y, u)_M$ is defined as

$$SE_g(x, x^*; Y, u)_M = TE(x, x^*; Y, u)_{M_g^*}. \quad (20.30)$$

We demonstrate this construction in the model of Figure 20.3 which stands for the equations:

$$\begin{aligned} z &= f_Z(x, u_Z) \\ w &= f_W(z, x, u_W) \\ y &= f_Y(z, w, u_Y) \end{aligned}$$

where u_Z , u_W , and u_Y are the components of u that enter the corresponding equations. Defining $z^*(u) = f_Z(x^*, u_Z)$, the modified model M_g^* reads:

$$\begin{aligned} z &= f_Z(x, u_Z) \\ w &= f_W(z, x^*, u_W) \text{ and} \\ y &= f_Y(z^*(u), w, u_Y) \end{aligned} \quad (20.31)$$

and our task amounts to computing the total effect of x on Y in M_g^* , or

$$TE(x, x^*; Y, u)_{M_g^*} = f_Y(z^*(u), f_W(f_Z(x, u_Z), x^*, u_W), u_Y) - Y_{x^*}(u) \quad (20.32)$$

It can be shown that the identification conditions for general path-specific effects are much more stringent than those of the direct and indirect effects. The path-specific effect shown in Figure 20.3, for example, is not identified even in Markovian models. Since direct and indirect effects are special cases of path-specific effects, the identification conditions of Theorems 20.2 and 20.3 raise the interesting question of whether a simple characterization exists of the class of subgraphs, g , whose path-specific effects are identifiable in Markovian models. I hope inquisitive readers will be able to solve this open problem.

20.4 Conclusions

This paper formulates a new definition of path-specific effects that is based on path switching, instead of variable fixing, and that extends the interpretation and evaluation of direct and indirect effects to nonlinear models. It is shown that, in nonparametric models, direct and indirect effects can be estimated consistently from both experimental and nonexperimental data, provided certain conditions hold in the causal diagram. Markovian models always satisfy these conditions. Using the new definition, the paper provides an operational interpretation of indirect effects, the policy significance of which was deemed enigmatic in recent literature.

On the conceptual front, the paper uncovers a class of nonstandard policy questions that cannot be formulated in the usual variable-fixing vocabulary and that can be evaluated, nevertheless, using the notions of direct and indirect effects. These policy questions concern redirecting the flow of influence in the system, and generally involve the deactivation of existing influences among specific variables. The ubiquity and manageability of such questions in causal modeling suggest that value-assignment manipulations, which control the outputs of the causal mechanism in the model, are less fundamental to the notion of causation than input-selection manipulations, which control the signals driving those mechanisms.

Acknowledgments

My interest in this topic was stimulated by Jacques Hagenars, who pointed out the importance of quantifying indirect effects in the social sciences (See <http://bayes.cs.ucla.edu/BOOK-2K/hagenars.html>.) Sol Kaufman, Sander Greenland, Steven Fienberg and Chris Hitchcock have provided helpful comments on the first draft

of this paper. This research was supported in parts by grants from NSF, ONR (MURI) and AFOSR.

References

- [Gastwirth, 1997] J.L. Gastwirth. Statistical evidence in discrimination cases. *Journal of the Royal Statistical Society, Series A*, 160(Part 2):289–303, 1997.
- [Goldberger, 1972] A.S. Goldberger. Structural equation models in the social sciences. *Econometrica: Journal of the Econometric Society*, 40:979–1001, 1972.
- [Hagenaars, 1993] J. Hagenaars. *Loglinear Models with Latent Variables*. Sage Publications, Newbury Park, CA, 1993.
- [Hesslow, 1976] G. Hesslow. Discussion: Two notes on the probabilistic approach to causality. *Philosophy of Science*, 43:290–292, 1976.
- [Lewis, 1973] D. Lewis. Counterfactuals and comparative probability. *Journal of Philosophical Logic*, 2, 1973.
- [Pearl, 2000] J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York, 2000.
- [Robins and Greenland, 1992] J.M. Robins and S. Greenland. Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, 3(2):143–155, 1992.
- [Rubin, 1974] D.B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66:688–701, 1974.



PART

CAUSALITY 2002–2020

21

Introduction by Judea Pearl

Paradoxes are the watchdogs of our hidden assumptions. Simpson's paradox, the topic of the first paper in this section (Chapter 22), has perplexed statisticians and philosophers for over a century and has escaped resolution because the assumptions involved were causal and inaccessible to conventional mathematics. In the first paper, I have used Simpson's paradox to demonstrate the limits of statistical methods, and why causal, rather than statistical considerations, are necessary for reaching a resolution and for answering the intricate questions that it raises. I consider this paradox to be the most compelling demonstration that our mind is governed by causal, rather than probabilistic, logic [Pearl 2014].

A popular belief in statistical circles has it that causal inference is a missing data problem. The second paper in this section (Chapter 23) argues for the converse: missing data is a causal problem, even in descriptive tasks. In other words, causal diagrams provide the natural way of modeling the mechanism causing missingness, and they permit us to determine, using graphical criteria, what quantities can be estimated consistently despite missingness. This work, by Karthika Mohan, opened our eyes to the possibility that many, if not all, so-called "statistical tasks" invoke informal causal reasoning and would therefore benefit from the formal methods offered by structural causal models [Mohan and Pearl 2014].

Another stalemate brought to a happy resolution was the problem of *External Validity*, which deals with generalizing empirical results across diverse environments. This age-old problem, the subject of Chapter 25 [Pearl and Bareinboim 2014], is at the heart of every scientific exploration since, invariably, laboratory findings are used in settings that are vastly different from the laboratory. Remarkably, despite efforts in psychology [Shadish et al. 2002] and economics [Manski 2007] the tasks of testing and establishing external validity have remained illusive, because statistics does not provide us with a language to encode disparities

and commonalities among environments or populations. “Selection diagrams,” described in Chapter 25, provide such encoding and enabled Elias Bareinboim and myself to reduce these tasks to symbolic exercise in *do*-calculus and decide what knowledge is required to take findings from experimental studies and generalize them to a new environment where no experiments are feasible. This technique has later led to the more general theory of Data Fusion [Bareinboim and Pearl 2016] which takes data from multiple sources, experimental as well as observational, and produces estimates of causal effects in yet a new environment, different from those studied. Data Fusion theory is one of the crown achievement of causal inference for it illuminates several key problems in machine learning, including “domain adaptation,” “robustness,” “transfer learning,” and more.

Selection bias, the topic of the third article in this section (Chapter 24), is another threat to validity that generations of experimentalists have bemoaned yet were unable to circumvent. This bias, created by selecting non-representative samples into the study, cannot be removed by randomization and can rarely be detected in the data. I was therefore delighted that the paper received the “Best Paper Award” at the 2014 Association for the Advancement of Artificial Intelligence (AAAI) Conference in Quebec. But more so, I was pleased to see a long-standing stalemate brought to a resolution by causal analysis [Bareinboim et al. 2014].

The fifth paper in this section, titled “Detecting latent heterogeneity,” (Chapter 26) was written in reaction to social scientists’ concerns with whether individuals differ in their response to a given treatment or program. It shows that, by combining observational and experimental data, it is possible to detect heterogeneity in the population even without observing the variables that are responsible for the differences. I labeled this counterintuitive result “a victory of formal counterfactual analysis,” and I am disappointed to see that it has received only 26 citations since its publication in [Pearl 2017]. This is the price one must pay for heroic victories.

References

- E. Bareinboim and J. Pearl. 2016. Causal inference and the data-fusion problem. *Proc Natl Acad Sci.* 113, 7345–7352.
- E. Bareinboim, J. Tian and J. Pearl. 2014. Recovering from selection bias in causal and statistical inference. *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*. AAAI Press, Palo Alto, CA, 2410–2416.
- C.F. Manski. 2007. *Identification of Prediction and Decision*. Harvard University Press, Cambridge, MA.
- K. Mohan and J. Pearl. 2014. Graphical models for recovering probabilistic and causal queries from missing data. *Advances of Neural Information Processing 27* (NIPS Proceedings), 1520–1528.

- J. Pearl. 2014. Comment: Understanding Simpson's Paradox. *Am. Stat.* 68, 1, 8–13.
- J. Pearl. 2017. Detecting latent heterogeneity. *Sociol. Methods Res.* 1–20.
- J. Pearl and E. Bareinboim. 2014. External validity: from *do*-calculus to transportability across populations. *Statistical Science*. 29, No. 4, 579–595. DOI: <https://doi.org/10.1214/14-STS486>.
- W.R. Shadish, T.D. Cook, and D.T. Campbell. 2002. Experimental and Quasi-experimental design for Generalized Causal Inference. Houghton-Mifflin, Boston, MA.

Comment: Understanding Simpson's Paradox

Judea Pearl

Simpson's paradox is often presented as a compelling demonstration of why we need statistics education in our schools. It is a reminder of how easy it is to fall into a web of paradoxical conclusions when relying solely on intuition, unaided by rigorous statistical methods. In recent years, ironically, the paradox assumed an added dimension when educators began using it to demonstrate the limits of statistical methods, and why causal, rather than statistical considerations are necessary to avoid those paradoxical conclusions (Wasserman 2004; Arah 2008; Pearl 2009, pp. 173–182).

My comments are divided into three parts. First, I will give a brief summary of the history of Simpson's paradox and how it has been treated in the statistical literature in the past century. Next, I will ask what is required to declare the paradox "resolved," and argue that modern understanding of causal inference has met those requirements. Finally, I will answer specific questions raised in Armistead's article and show how the resolution of Simpson's paradox can be taught for fun and progress.

22.1

The History

Simpson's paradox refers to a phenomenon whereby the association between a pair of variables (X, Y) reverses sign upon conditioning of a third variable, Z , regardless

This research was supported in parts by grants from NSF #IIS1249822 and #IIS1302448, and ONR #N00014-13-1-0153 and #N00014-10-1-0933. I appreciate the encouragement of Ronald Christensen, conversations with Miguel Hernán, and editorial comments by Madlyn Glymour.

Originally published in *The American Statistician*, 68(1):8-13, February 2014.

© Taylor & Francis, 2014. Republished with permission from Taylor & Francis.

Original DOI: [10.1080/00031305.2014.876829](https://doi.org/10.1080/00031305.2014.876829)

of the value taken by Z . If we partition the data into subpopulations, each representing a specific value of the third variable, the phenomenon appears as a sign reversal between the associations measured in the disaggregated subpopulations relative to the aggregated data, which describes the population as a whole.

Edward H. Simpson first addressed this phenomenon in a technical article in 1951, but [Karl Pearson et al.](#) in 1899 and [Udny Yule](#) in 1903 had mentioned a similar effect earlier. All three reported associations that disappear, rather than reversing signs upon aggregation. Sign reversal was first noted by [Cohen and Nagel](#) (1934) and then by [Blyth](#) (1972) who labeled the reversal “paradox,” presumably because the surprise that association reversal evokes among the unwary appears paradoxical at first.

Chapter 6 of my book *Causality* ([Pearl 2009](#), p. 176) remarks that, surprisingly, only two articles in the statistical literature attribute the peculiarity of Simpson’s reversal to causal interpretations. The first is [Pearson, Lee, and Bramley-Moore](#) (1899), in which a short remark warns us that correlation is not causation, and the second is [Lindley and Novick](#) (1981) who mentioned the possibility of explaining the paradox in “the language of causation” but chose not to do so “because the concept, although widely used, does not seem to be well defined” (p. 51). My survey further documents that, other than these two exceptions, the entire statistical literature from [Pearson, Lee, and Bramley-Moore](#) (1899) to the 1990s was not prepared to accept the idea that a statistical peculiarity, so clearly demonstrated in the data, could have causal roots.¹

In particular, the word “causal” does not appear in Simpson’s article, nor in the vast literature that followed, including [Blyth](#) (1972), who coined the term “paradox,” and the influential writings of [Agresti](#) (1983), [Bishop, Fienberg, and Holland](#) (1975), and [Whittemore](#) (1978).

What Simpson did notice though, was that depending on the story behind the data, the more “sensible interpretation” (his words) is sometimes compatible with the aggregate population, and sometimes with the disaggregated subpopulations. His example of the latter involves a positive association between treatment and survival both among males and females which disappears in the combined population. Here, his “sensible interpretation” is unambiguous: “The treatment can hardly be rejected as valueless to the race when it is beneficial when applied to males and to females.” His example of the former involved a deck of cards, in

1. This contrasts the historical account of [Hernán, Clayton, and Keiding](#) (2011) according to which “Such discrepancy [between marginal and conditional associations in the presence of confounding] had been already noted, formally described and explained in causal terms half a century before the publication of Simpson’s article...” Simpson and his predecessor did not have the vocabulary to articulate, let alone formally describe and explain causal phenomena.

which two independent face types become associated when partitioned according to a cleverly crafted rule (see [Hernán, Clayton, and Keiding 2011](#)). Here, claims Simpson, “it is the combined table which provides what we would call the sensible answer.” This key observation remained unnoticed until [Lindley and Novick \(1981\)](#) replicated it in a more realistic example which gave rise to reversal. The idea that statistical data, however large, are insufficient for determining what is “sensible,” and that it must be supplemented with extra-statistical knowledge to make sense was considered heresy in the 1950s.

[Lindley and Novick \(1981\)](#) elevated Simpson’s paradox to new heights by showing that there was no statistical criterion that would warn the investigator against drawing the wrong conclusions or indicate which data represented the correct answer. First they showed that reversal may lead to difficult choices in critical decision-making situations:

The apparent answer is, that when we know that the gender of the patient is male or when we know that it is female we do not use the treatment, but if the gender is unknown we should use the treatment! Obviously that conclusion is ridiculous. ([Novick 1983](#), p. 45)

Second, they showed that, with the very same data, we should consult either the combined table or the disaggregated tables, depending on the context. Clearly, when two different contexts compel us to take two opposite actions based on the same data, our decision must be driven not by statistical considerations, but by some additional information extracted from the context.

Third, they postulated a scientific characterization of the extra-statistical information that researchers take from the context, and which causes them to form a consensus as to which table gives the correct answer. That Lindley and Novick opted to characterize this information in terms of “exchangeability” rather than causality is understandable;² the state of causal language in the 1980s was so primitive that they could not express even the simple yet crucial fact that gender is not affected by the treatment.³ What is important though, is that the example they used to demonstrate that the correct answer lies in the aggregated data, had a totally different causal structure than the one where the correct answer lies in the disaggregated data. Specifically, the third variable (Plant Height) was affected

2. Lindley later regretted that choice ([Pearl 2009](#), p. 384), and indeed, his treatment of exchangeability was guided exclusively by causal considerations ([Meek and Glymour 1994](#)).

3. Statistics teachers would enjoy the challenge of explaining how the sentence “treatment does not change gender” can be expressed mathematically. Lindley and Novick tried, unsuccessfully of course, to use conditional probabilities.

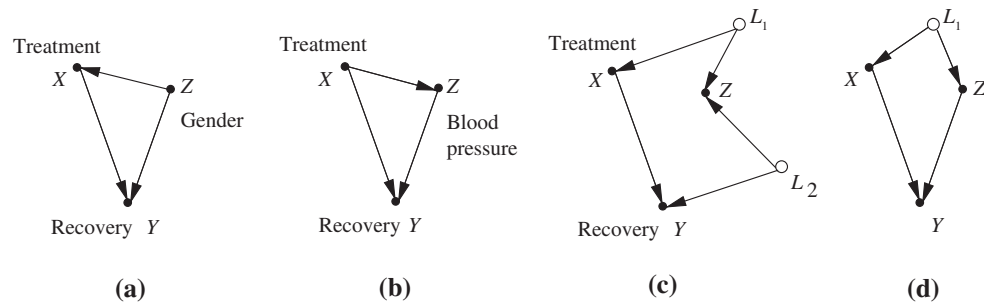


Figure 22.1 Graphs demonstrating the insufficiency of chronological information. In models (c) and (d), Z may occur before or after the treatment, yet the correct answer remains invariant to this timing: we should not condition on Z in model (c), and we should condition on Z in model (d). In both models, Z is not affected by the treatment.

by the treatment (Plant Color) as opposed to gender which is a pre-treatment confounder. (See an isomorphic model in Figure 22.1(b), where blood-pressure replacing plant-height.⁴)

More than 30 years have passed since the publication of Lindley and Novick's article, and the face of causality has changed dramatically. Not only do we now know which causal structures would support Simpson's reversals, we also know which structure places the correct answer with the aggregated data or with the disaggregated data. Moreover, the criterion for predicting where the correct answer lies (and, accordingly, where human consensus resides) turns out to be rather insensitive to temporal information, nor does it hinge critically on whether or not the third variable is affected by the treatment. It involves a simple graphical condition called "back-door" (Pearl 1993) which traces paths in the causal diagram and assures that all spurious paths from treatment to outcome are intercepted by the third variable. This will be demonstrated in the next section, where we argue that, armed with these criteria, we can safely proclaim Simpson's paradox "resolved."

22.2 A Paradox Resolved

Any claim to a resolution of a paradox, especially one that has resisted a century of attempted resolution must meet certain criteria. First and foremost, the solution must explain why people consider the phenomenon surprising or unbelievable.

4. Interestingly, Simpson's examples also had different causal structure; in the former, the third variable (gender) was a common cause of the other two, whereas in the latter, the third variable (paint on card) was a common effect of the other two (Hernán, Clayton, and Keiding 2011). Yet, although this difference changed Simpson's intuition of what is "more sensible," it did not stimulate his curiosity as a fundamental difference, worthy of scientific exploration.

Second, the solution must identify the class of scenarios in which the paradox may surface and distinguish it from scenarios where it will surely not surface. Finally, in those scenarios where the paradox leads to indecision, we must identify the correct answer, explain the features of the scenario that lead to that choice, and prove mathematically that the answer chosen is indeed correct. The next three subsections will describe how these three requirements are met in the case of Simpson's paradox and, naturally, will proceed to convince readers that the paradox deserves the title "resolved."

22.2.1 Simpson's Surprise

In explaining the surprise, we must first distinguish between "Simpson's reversal" and "Simpson's paradox;" the former being an arithmetic phenomenon in the calculus of proportions, the latter a psychological phenomenon that evokes surprise and disbelief. A full understanding of Simpson's paradox should explain why an innocent arithmetic reversal of an association, albeit uncommon, came to be regarded as "paradoxical," and why it has captured the fascination of statisticians, mathematicians, and philosophers for over a century (though it was first labeled "paradox" by [Blyth 1972](#)).

The arithmetics of proportions has its share of peculiarities, no doubt, but these tend to become objects of curiosity once they have been demonstrated and explained away by examples. For instance, naive students of probability may expect the average of a product to equal the product of the averages but quickly learn to guard against such expectations, given a few counterexamples. Likewise, students expect an association measured in a mixture distribution to equal a weighted average of the individual associations. They are surprised, therefore, when ratios of sums, $(a + b)/(c + d)$, are found to be ordered differently than individual ratios, a/c and b/d .⁵ Again, such arithmetic peculiarities are quickly accommodated by seasoned students as reminders against simplistic reasoning.

In contrast, an arithmetic peculiarity becomes "paradoxical" when it clashes with deeply held convictions that the peculiarity is impossible, and this occurs when one takes seriously the causal implications of Simpson's reversal in decision-making contexts. Reversals are indeed impossible whenever the third variable, say age or gender, stands for a pretreatment covariate because, so the reasoning goes, no drug can be harmful to both males and females yet beneficial to the population as a whole. The universality of this intuition reflects a deeply held and valid conviction that such a drug is physically impossible. Remarkably, such impossibility can

5. In Simpson's paradox, we witness the simultaneous orderings: $(a1 + b1)/(c1 + d1) > (a2 + b2)/(c2 + d2)$, $(a1/c1) < (a2/c2)$, and $(b1/d1) < (b2/d2)$.

be derived mathematically in the calculus of causation in the form of a “sure-thing” theorem (Pearl 2009, p. 181):

An action A that increases the probability of an event B in each subpopulation (of C) must also increase the probability of B in the population as a whole, provided that the action does not change the distribution of the subpopulations.⁶

Thus, regardless of whether effect size is measured by the odds ratio or other comparisons, regardless of whether Z is a confounder or not, and regardless of whether we have the correct causal structure on hand, our intuition should be offended by any effect reversal that appears to accompany the aggregation of data.

I am not aware of another condition that rules out effect reversal with comparable assertiveness and generality, requiring only that Z not be affected by our action, a requirement satisfied by all treatment-independent covariates Z . Thus, it is hard, if not impossible, to explain the surprise part of Simpson's reversal without postulating that human intuition is governed by causal calculus together with a persistent tendency to attribute causal interpretation to statistical associations.

22.2.2 Which Scenarios Invite Reversals?

Attending to the second requirement, we need first to agree on a language that describes and identifies the class of scenarios for which association reversal is possible. Since the notion of “scenario” connotes a process by which data is generated, a suitable language for such a process is a causal diagram, as it can simulate any data-generating process that operates sequentially along its arrows. For example, the diagram in Figure 22.1(a) can be regarded as a blueprint for a process in which $Z = \text{Gender}$ receives a random value (male or female) depending on the gender distribution in the population. The treatment is then assigned a value (treated or untreated) according to the conditional distribution $P(\text{treatment} \mid \text{male})$ or $P(\text{treatment} \mid \text{female})$. Finally, once gender and treatment receive their values, the outcome process (recovery) is activated and assigns a value to Y using the conditional distribution $P(Y = y \mid X = x, Z = z)$. All these local distributions can be estimated from the data. Thus, the scientific content of a given scenario can be encoded in the form of a directed acyclic graph (DAG), capable of simulating a set of data-generating processes compatible with the given scenario.

The theory of graphical models (Pearl 1988; Lauritzen 1996) can tell us, for a given DAG, whether Simpson's reversal is realizable or logically impossible in the

6. The no-change provision is probabilistic; it permits the action to change the classification of individual units so long as the relative sizes of the subpopulations remain unaltered.

simulated scenario. By a logical impossibility, we mean that for every scenario that fits the DAG structure, there is no way to assign processes to the arrows and generate data that exhibit association reversal as described by Simpson.

For example, the theory immediately tells us that all structures depicted in Figure 22.1 can exhibit reversal, while in Figure 22.2, reversal can occur in (a), (b), and (c), but not in (d), (e), or (f). That Simpson's paradox can occur in each of the structures in Figure 22.1 follows from the fact that the structures are observationally equivalent; each can emulate any distribution generated by the others. Therefore, if association reversal is realizable in one of the structures, say (a), it must be realizable in all structures. The same consideration applies to graphs (a), (b), and (c) of Figure 22.2, but not to (d), (e), or (f) which are where the X, Y association is collapsible over Z .

22.2.3 Making the Correct Decision

We now come to the hardest test of having resolved the paradox: proving that we can make the correct decision when reversal occurs. This can be accomplished either mathematically or by simulation. Mathematically, we use an algebraic method called “*do*-calculus” (Pearl 2009, p. 85–89) which is capable of determining, for any given model structure, the causal effect of one variable on another

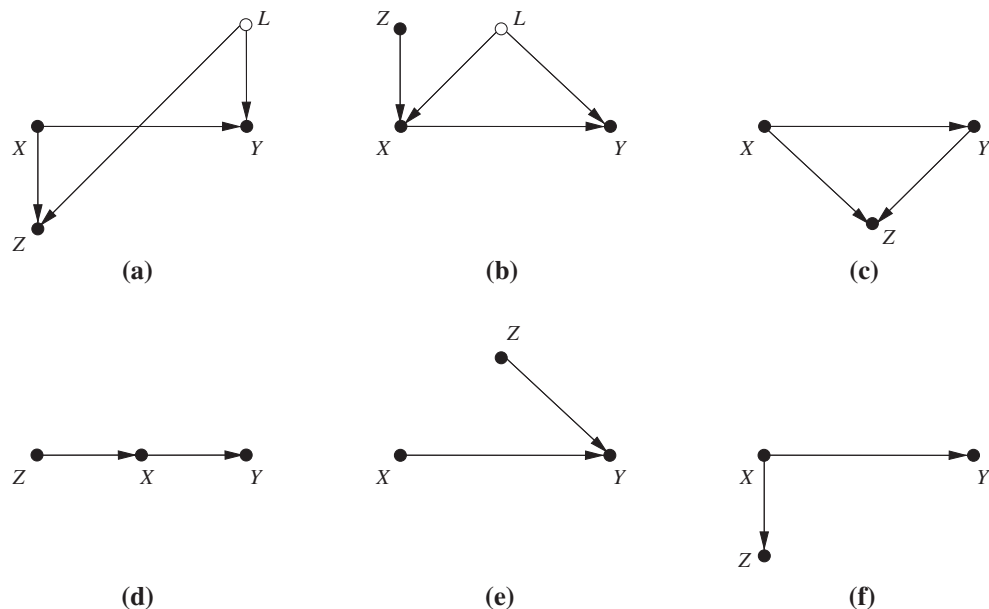


Figure 22.2 Simpson's reversal can be realized in models (a), (b), and (c) but not in (d), (e), or (f).

and which variables need to be measured to make this determination.⁷ Compliance with *do*-calculus should then constitute a proof that the decisions we made using graphical criteria is correct. Since some readers of this article may not be familiar with the *do*-calculus, simulation methods may be more convincing. Simulation “proofs” can be organized as a “guessing game,” where a “challenger” who knows the model behind the data dares an analyst to guess what the causal effect is (of X on Y) and checks the answer against the gold standard of a randomized trial, simulated on the model. Specifically, the “challenger” chooses a scenario (or a “story” to be simulated), and a set of simulation parameters such that the data generated would exhibit Simpson’s reversal. He then reveals the scenario (not the parameters) to the analyst. The analyst constructs a DAG that captures the scenario and guesses (using the structure of the DAG), whether the correct answer lies in the aggregated or disaggregated data. Finally, the “challenger” simulates a randomized trial on a fictitious population generated by the model, estimates the underlying causal effect, and checks the result against the analyst’s guess.

For example, the back-door criterion instructs us to guess that in Figure 22.1, in models (b) and (c) the correct answer is provided by the aggregated data, while in structures (a) and (d) the correct answer is provided by the disaggregated data. We simulate a randomized experiment on the (fictitious) population to determine whether the resulting effect is positive or negative, and compare it with the associations measured in the aggregated and disaggregated population. Remarkably, our guesses should prove correct regardless of the parameters used in the simulation model, as long as the structure of the simulator remains the same.⁸ This explains how people form a consensus about which data is “more sensible” (Simpson 1951) prior to actually seeing the data.

This is a good place to explain how the back-door criterion works, and how it determines where the correct answer resides. The principle is simple: the paths connecting X and Y are of two kinds, causal and spurious. Causative associations are carried by the causal paths, namely, those tracing arrows directed from X to Y . The other paths carry spurious associations and need to be blocked by conditioning on an appropriate set of covariates. All paths containing an arrow into X are spurious paths and need to be intercepted by the chosen set of covariates.

When dealing with a singleton covariate Z , as in the Simpson’s paradox, we need to merely ensure that

7. When such determination cannot be made from the given graph, as is the case in Figure 22.2(b), the *do*-calculus alerts us to this fact.

8. By “structure” we mean the list of variables that need be consulted in computing each variable V_i in the simulation.

1. Z is not a descendant of X , and
2. Z blocks every path that ends with an arrow into X .

(Extensions for descendants of X are given in Pearl (2009, p. 338), Shpitser, VanderWeele, and Robins (2010), and Pearl and Paz (2013)).

The operation of “blocking” requires a special handling of “collider” variables, which behave oppositely to arrow-emitting variables. The latter block the path when conditioned on, while the former block the path when they and all their descendants are not conditioned on. This special handling of “colliders” reflects a general phenomenon known as Berkson’s paradox (Berkson 1946), whereby observations on a common consequence of two independent causes render those causes dependent. For example, the outcomes of two independent coins are rendered dependent by the testimony that at least one of them is a tail.

Armed with this criterion we can determine, for example, that in Figures 22.1(a) and (d), if we wish to correctly estimate the effect of X on Y , we need to condition on Z (thus blocking the back-door path $X \leftarrow Z \rightarrow Y$). We can similarly determine that we should not condition on Z in Figures 22.1(b) and (c). The former because there are no back-door paths requiring blockage, and the latter because the back-door path $X \leftarrow \circ \rightarrow Z \leftarrow \circ \rightarrow Y$ is blocked when Z is not conditioned on. The correct decisions follow from this determination; when conditioning on Z is required, the Z -specific data carry the correct information. In Figure 22.2(c), for example, the aggregated information carries the correct information because the spurious (noncausal) path $X \rightarrow Z \leftarrow Y$ is blocked when Z is not conditioned on. The same applies to Figures 22.2(a) and 22.1(c).

Finally, we should remark that in certain models the correct answer may not lie in either the disaggregated or the aggregated data. This occurs when Z is not sufficient to block an active back-door path as in Figure 22.2(b); in such cases, a set of additional covariates may be needed, which takes us beyond the scope of this note.

The model in Figure 22.3 presents opportunities to simulate successive reversals, which could serve as an effective (and fascinating) instruction tool for introductory statistics classes. Here, we see that to block the only unblocked back-door path $X \leftarrow Z_1 \rightarrow Z_3 \rightarrow Y$, we need to condition on Z_1 . This means that, if the simulation machine is set to generate association reversal, the correct answer will reside in the disaggregated, Z_1 -specific data. If we further condition on a second variable, Z_2 , the back-door path $X \leftarrow \circ \rightarrow Z_2 \leftarrow Z_3 \rightarrow Y$ will become unblocked, and a bias will be created, meaning that the correct answer lies with the aggregated data. Upon further conditioning on Z_3 the bias is removed and the correct answer returns to the disaggregated, Z_3 -specific data.

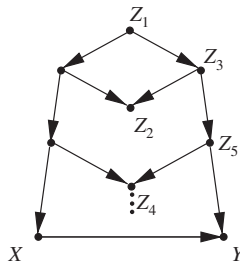


Figure 22.3 A multistage Simpson's paradox machine. Cumulative conditioning in the order $(Z_1, Z_2, Z_3, Z_4, Z_5)$ creates reversal at each stage, with the correct answers alternating between disaggregated and aggregated data.

Note that in each stage, we can set the numbers in the simulation machine so as to generate association reversal between the preconditioning and post-conditioning data. Note further that at any stage of the process we can check where the correct answer lies by subjecting the population generated to a hypothetical randomized trial.

22.3 Armistead's Critique

Armistead does not disagree with the technical points presented above and rightly so; they are backed by sound mathematical proofs. The main point of contention seems to be whether the disaggregated data are still valuable, when the correct answer lies with the aggregated data (as in Figures 22.1(a) and (c)). On this issue, Armistead says:

Whether causal or not, third variables can convey critical information about a first order relationship, study design, and previously unobserved variables. Any conditioning on nontrivial third variable that produces Simpson's Paradox should be carefully examined before either the aggregated or the disaggregated findings are accepted, regardless of whether the variable is thought to be causal.

I agree with the general thrust of this paragraph. Every variable can indeed "convey critical information" if such information is needed for answering the investigator's research question. But in our examples, we asked not whether the third variable conveys information about study design or other interesting subjects; we asked whether it would help us estimate the total effect of X on Y . In the context of this query, the answer is: NO; the aggregated (or disaggregated) findings can be accepted without further examination.

When we endeavor to ask other queries, other than total treatment effects, intermediate variables can of course provide useful information. For example, when we ask about the role of blood pressure in mediating the effect of treatment on recovery (as in Section 22.4) a whole set of mediation analytic techniques can be brought to bear on the question (e.g., VanderWeele 2009; Imai, Keele, and Yamamoto 2010; Pearl 2013) which aims to assess direct and indirect effects as formulated in Pearl (2001) and Robins and Greenland (1992). If, on the other hand, we ask questions about how the third variable (e.g., blood pressure) can help estimate treatment effects in the presence of unmeasured confounders, another set of tools is brought into consideration (see Pearl 1995). But when our query is “Which drug is more effective?” (assuming no unmeasured confounders, as in Figure 22.1(b)), the answer is unequivocal: “Ignore blood pressure.”

Finally, I also agree with the spirit of Armistead’s statement:

Any conditioning on nontrivial third variable that produces Simpson Paradox should be carefully examined before either the aggregated or the disaggregated findings are accepted, regardless of whether the variable is thought to be causal.

I must point out, however, that we can do better than “carefully examine” the third variable. Modern tools of causal analysis now permit us to determine mathematically whether the aggregated or disaggregated findings should be accepted.⁹ Specifically, in the blood-pressure example, mathematical analysis dictates that the aggregated findings give the correct answer to our specific research question, which is precisely what “careful examination” aims to accomplish. Armistead is correct in stating that this holds regardless of whether one categorizes “blood pressure” as causal or noncausal variable; what matters is the causal relationships of the third variable to other variables in the analysis, as portrayed in the diagram. Indeed, in Figure 22.1(c), for example, the third variable Z is not affected by the treatment, and still, it should not be controlled for; the aggregated finding should be accepted.

22.4 Conclusions

I hope that playing the multistage Simpson’s guessing game (Figure 22.3) would convince readers that we now understand most of the intricacies of Simpson’s paradox, and we can safely title it “resolved.”

9. Expressions such as “should be carefully examined” were used by statisticians in the precausal era to convey helplessness in handling causal questions.

References

- Agresti, A. (1983), "Fallacies, Statistical," in *Encyclopedia of Statistical Science* (vol. 3), eds. S. Kotz and N. Johnson, New York: Wiley, pp. 24–28. [400]
- Arah, O. (2008), "The Role of Causal Reasoning in Understanding Simpson's Paradox, Lord's Paradox, and the Suppression Effect: Covariate Selection in the Analysis of Observational Studies," *Emerging Themes in Epidemiology*, 4, DOI:10.1186/1742-7622-5-5. Available at <http://www.ete-online.com/content/5/1/5>. [399]
- Berkson, J. (1946), "Limitations of the Application of Fourfold Table Analysis to Hospital Data," *Biometrics Bulletin*, 2, 47–53. [407]
- Bishop, Y., Fienberg, S., and Holland, P. (1975), *Discrete Multivariate Analysis: Theory and Practice*, Cambridge, MA: MIT Press. [400]
- Blyth, C. (1972), "On Simpson's Paradox and the Sure-Thing Principle," *Journal of the American Statistical Association*, 67, 364–366. [400,403]
- Cohen, M., and Nagel, E. (1934), *An Introduction to Logic and the Scientific Method*, New York: Harcourt, Brace and Company. [400]
- Hernán, M., Clayton, D., and Keiding, N. (2011), "The Simpson's Paradox Unraveled," *International Journal of Epidemiology*, 40, 780–785. DOI:10.1093/ije/dyr041. [400,401,402]
- Imai, K., Keele, L., and Yamamoto, T. (2010), "Identification, Inference, and Sensitivity Analysis for Causal Mediation Effects," *Statistical Science*, 25, 51–71. [409]
- Lauritzen, S. (1996), *Graphical Models* (reprinted 2004 with corrections), Oxford: Clarendon Press. [404]
- Lindley, D., and Novick, M. (1981), "The Role of Exchangeability in Inference," *The Annals of Statistics*, 9, 45–58. [400,401]
- Meek, C., and Glymour, C. (1994), "Conditioning and Intervening," *British Journal of Philosophy Science*, 45, 1001–1021. [401]
- Novick, M. (1983), "The Centrality of Lord's Paradox and Exchangeability for all Statistical Inference," in *Principles of Modern Psychological Measurement*, eds. H. Wainer and S. Messick, Hillsdale, NJ: Earlbaum, pp. 41–53. [401]
- Pearl, J. (1988), *Probabilistic Reasoning in Intelligent Systems*, San Mateo, CA: Morgan Kaufmann. [404]
- (1993), "Comment: Graphical Models, Causality, and Intervention," *Statistical Science*, 8, 266–269. [402]
- (1995), "Causal Diagrams for Empirical Research," *Biometrika*, 82, 669–710. [409]
- (2001), "Direct and Indirect Effects," in *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, San Francisco, CA: Morgan Kaufmann, pp. 411–420. [409]
- (2009), *Causality: Models, Reasoning, and Inference* (2nd ed.), New York: Cambridge University Press. [399,400,401,404,405,407]
- (2013), "Interpretation and Identification of Causal Mediation," *Psychological Methods*. Since publishing, in *Psychological Methods*, 19, 459–481, 2014. [409]

- Pearl, J., and Paz, A. (2013), "Confounding Equivalence in Causal Inference," Technical Report no. R-343w, Department of Computer Science, University of California, Los Angeles, CA. Revised and submitted, October 2013, available at http://ftp.cs.ucla.edu/pub/stat_ser/r343w.pdf. Since publishing, in *Journal of Causal Inference*, 2, 75–93, April 2014. [407]
- Pearson, K., Lee, A., and Bramley-Moore, L. (1899), "Genetic (Reproductive) Selection: Inheritance of Fertility in Man, and of Fecundity in Thoroughbred Racehorses," *Philosophical Transactions of the Royal Society of London, Series A*, 192, 257–330. [400]
- Robins, J., and Greenland, S. (1992), "Identifiability and Exchangeability for Direct and Indirect Effects," *Epidemiology*, 3, 143–155. [409]
- Shpitser, I., VanderWeele, T., and Robins, J. (2010), "On the Validity of Covariate Adjustment for Estimating Causal Effects," in *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*, Corvallis, OR: AUAI, pp. 527–536. [407]
- Simpson, E. (1951), "The Interpretation of Interaction in Contingency Tables," *Journal of the Royal Statistical Society, Series B*, 13, 238–241. [406]
- VanderWeele, T. (2009), "Marginal Structural Models for the Estimation of Direct and Indirect Effects," *Epidemiology*, 20, 18–26. [409]
- Wasserman, L. (2004), *All of Statistics: A Concise Course in Statistical Inference*, New York: Springer. [399]
- Whittemore, A. (1978), "Collapsibility of Multidimensional Contingency Tables," *Journal of the Royal Statistical Society, Series B*, 40, 328–340. [400]
- Yule, G. (1903), "Notes on the Theory of Association of Attributes in Statistics," *Biometrika*, 2, 121–134. [400]

Graphical Models for Recovering Probabilistic and Causal Queries from Missing Data

Karthika Mohan* and Judea Pearl

Abstract

We address the problem of deciding whether a causal or probabilistic query is estimable from data corrupted by missing entries, given a model of missingness process. We extend the results of Mohan et al. (2013) by presenting more general conditions for recovering probabilistic queries of the form $P(y|x)$ and $P(y, x)$ as well as causal queries of the form $P(y|do(x))$. We show that causal queries may be recoverable even when the factors in their identifying estimands are not recoverable. Specifically, we derive graphical conditions for recovering causal effects of the form $P(y|do(x))$ when Y and its missingness mechanism are not d -separable. Finally, we apply our results to problems of attrition and characterize the recovery of causal effects from data corrupted by attrition.

23.1 Introduction

All branches of experimental science are plagued by missing data. Improper handling of missing data can bias outcomes and potentially distort the conclusions

*University of California, Los Angeles

Originally published in M. Welling, Z. Ghahramani, C. Cortes, and N. Lawrence (eds.), *Advances of Neural Information Processing 27* (NIPS Proceedings), 1520–1528, 2014. Republished with permission.

drawn from a study. Therefore, accurate diagnosis of the causes of missingness is crucial for the success of any research. We employ a formal representation called ‘Missingness Graphs’ (*m*-graphs, for short) to explicitly portray the missingness process as well as the dependencies among variables in the available dataset (Mohan et al. 2013). Apart from determining whether recoverability is feasible namely, whether there exists any theoretical impediment to estimability of queries of interest, *m*-graphs can also provide a means for communication and refinement of assumptions about the missingness process. Furthermore, *m*-graphs permit us to detect violations in modeling assumptions even when the dataset is contaminated with missing entries (Mohan and Pearl 2014).

In this paper, we extend the results of Mohan et al. (2013) by presenting general conditions under which probabilistic queries such as joint and conditional distributions can be recovered. We show that causal queries of the type $P(y|do(x))$ can be recovered even when the associated probabilistic relations such as $P(y, x)$ and $P(y|x)$ are not recoverable. In particular, causal effects may be recoverable even when Y is not separable from its missingness mechanism. Finally, we apply our results to recover causal effects when the available dataset is tainted by attrition.

This paper is organized as follows. Section 23.2 provides an overview of missingness graphs and reviews the notion of recoverability i.e., obtaining consistent estimates of a query, given a dataset and an *m*-graph. Section 23.3 refines the sequential factorization theorem presented in Mohan et al. (2013) and extends its applicability to a wider range of problems in which missingness mechanisms may influence each other. In Section 23.4, we present general algorithms to recover joint distributions from the class of problems for which sequential factorization theorem fails. In Section 23.5, we introduce new graphical criteria that preclude recoverability of joint and conditional distributions. In Section 23.6, we discuss recoverability of causal queries and show that unlike probabilistic queries, $P(y|do(x))$ may be recovered even when Y and its missingness mechanism (R_y) are not d -separable. In Section 23.7, we demonstrate how we can apply our results to problems of attrition in which missingness is a severe obstacle to sound inferences. Related works are discussed in Section 23.8 and conclusions are drawn in Section 23.9. Proofs of all theoretical results in this paper are provided in Appendix 23.A.

23.2 Missingness Graph and Recoverability

Missingness graphs as discussed below was first defined in Mohan et al. (2013) and we adopt the same notations. Let $G(\mathbb{V}, E)$ be the causal DAG where $\mathbb{V} = V \cup U \cup$

$V^* \cup \mathbb{R}$. V is the set of observable nodes. Nodes in the graph correspond to variables in the data set. U is the set of unobserved nodes (also called latent variables). E is the set of edges in the DAG. We use bi-directed edges as a shorthand notation to denote the existence of a U variable as common parent of two variables in $V \cup \mathbb{R}$. V is partitioned into V_o and V_m such that $V_o \subseteq V$ is the set of variables that are observed in all records in the population and $V_m \subseteq V$ is the set of variables that are missing in at least one record. Variable X is termed as *fully observed* if $X \in V_o$, *partially observed* if $X \in V_m$ and *substantive* if $X \in V_o \cup V_m$. Associated with every partially observed variable $V_i \in V_m$ are two other variables R_{v_i} and V_i^* , where V_i^* is a proxy variable that is actually observed, and R_{v_i} represents the status of the causal mechanism responsible for the missingness of V_i^* ; formally,

$$v_i^* = f(r_{v_i}, v_i) = \begin{cases} v_i & \text{if } r_{v_i} = 0 \\ m & \text{if } r_{v_i} = 1 \end{cases} \quad (23.1)$$

V^* is the set of all proxy variables and \mathbb{R} is the set of all causal mechanisms that are responsible for missingness. R variables may not be parents of variables in $V \cup U$. We call this graphical representation **Missingness Graph** (or *m-graph*). An example of an *m-graph* is given in Figure 23.1 (a). We use the following shorthand. For any variable X , let X' be a shorthand for $X = 0$. For any set $W \subseteq V_m \cup V_o \cup R$, let W_r , W_o and W_m be the shorthand for $W \cap R$, $W \cap V_o$ and $W \cap V_m$ respectively. Let R_w be a shorthand for $R_{V_m \cap W}$ i.e., R_w is the set containing missingness mechanisms of all

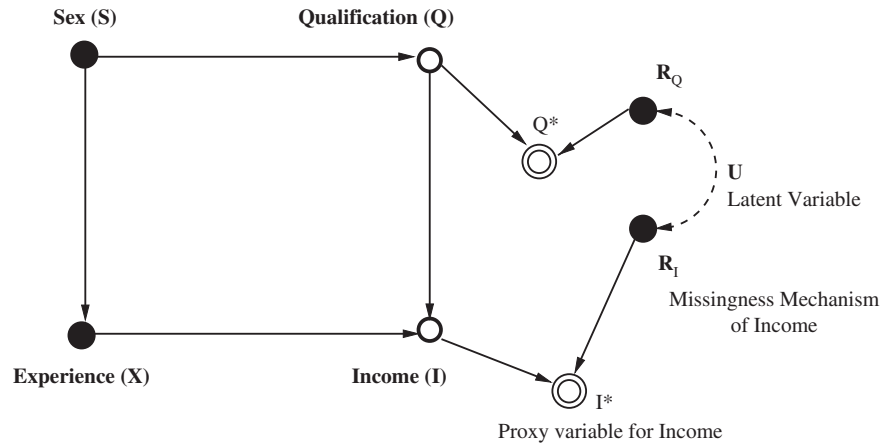


Figure 23.1 Typical *m-graph* where $V_o = \{S, X\}$, $V_m = \{I, Q\}$, $V^* = \{I^*, Q^*\}$, $R = \{R_i, R_q\}$ and U is the latent common cause. Members of V_o and V_m are represented by full and hollow circles respectively. The associated missingness process and assumptions are elaborated in Appendix 23.A.1.

partially observed variables in W . Note that R_w and W_r are not the same. $G_{\underline{X}}$ and $G_{\overline{X}}$ represent graphs formed by removing from G all edges leaving and entering X , respectively.

A *manifest distribution* $P(V_o, V^*, R)$ is the distribution that governs the available dataset. An *underlying distribution* $P(V_o, V_m, R)$ is said to be compatible with a given manifest distribution $P(V_o, V^*, R)$ if the latter can be obtained from the former using Equation 23.1. Manifest distribution P_m is compatible with a given underlying distribution P_u if $\forall X, X \subseteq V_m$ and $Y = V_m \setminus X$, the following equality holds true.

$$P_m(R'_x, R_y, X^*, Y^*, V_o) = P_u(R'_x, R_y, X, V_o)$$

where R'_x denotes $R_x = 0$ and R_y denotes $R_y = 1$. Refer Appendix 23.A.2 for an example.

23.2.1 Recoverability

Given a manifest distribution $P(V^*, V_o, R)$ and an m -graph G that depicts the missingness process, query Q is recoverable if we can compute a consistent estimate of Q as if no data were missing. Formally,

Definition 23.1 Recoverability (Mohan et al. 2013)

Given a m -graph G , and a target relation Q defined on the variables in V , Q is said to be recoverable in G if there exists an algorithm that produces a consistent estimate of Q for every dataset D such that $P(D)$ is (1) compatible with G and (2) strictly positive¹ i.e., $P(V_o, V^*, \mathbb{R}) > 0$.

For an introduction to the notion of recoverability, see Pearl and Mohan (2013) and Mohan et al. (2013).

23.3 Recovering Probabilistic Queries by Sequential Factorization

Mohan et al. (2013) (Theorem 23.4) presented a sufficient condition for recovering probabilistic queries such as joint and conditional distributions by using ordered factorizations. However, the theorem is not applicable to certain classes of problems such as those in longitudinal studies in which edges exist between R variables. General ordered factorization defined below broadens the concept of ordered factorization (Mohan et al. 2013) to include the set of R variables. Subsequently, the modified theorem (stated below as Theorem 23.1) will permit us to handle cases in which R variables are contained in separating sets that d -separate

1. An extension to datasets that are not strictly positive is sometimes feasible (Mohan et al. 2013).

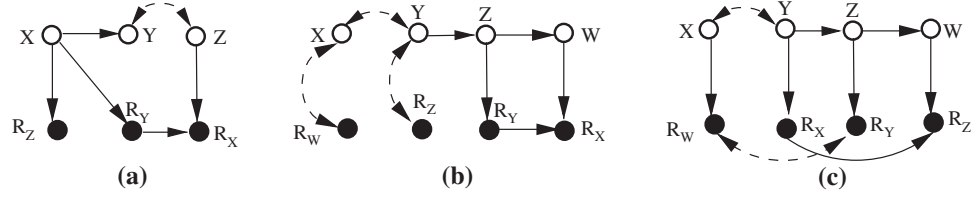


Figure 23.2 (a) m -graph in which $P(V)$ is recoverable by the sequential factorization (b) & (c): m -graphs for which no admissible sequence exists.

partially observed variables from their respective missingness mechanisms (example: $X \perp\!\!\!\perp R_x | R_y$ in Figure 23.2 (a)).

Definition 23.2 General Ordered factorization

Given a graph G and a set O of ordered $V \cup R$ variables $Y_1 < Y_2 < \dots < Y_k$, a general ordered factorization relative to G , denoted by $f(O)$, is a product of conditional probabilities $f(O) = \prod_i P(Y_i | X_i)$ where $X_i \subseteq \{Y_{i+1}, \dots, Y_n\}$ is a minimal set such that $Y_i \perp\!\!\!\perp (\{Y_{i+1}, \dots, Y_n\} \setminus X_i) | X_i$ holds in G .

Theorem 23.1 Sequential Factorization

A sufficient condition for recoverability of a relation Q defined over substantive variables is that Q be decomposable into a general ordered factorization, or a sum of such factorizations, such that every factor $Q_i = P(Y_i | X_i)$ satisfies, (1) $Y_i \perp\!\!\!\perp (R_{y_i}, R_{x_i}) | X_i \setminus \{R_{y_i}, R_{x_i}\}$, if $Y_i \in (V_o \cup V_m)$ and (2) $R_z \perp\!\!\!\perp R_{x_i} | X_i$ if $Y_i = R_z$ for any $Z \in V_m$, $Z \notin X_i$ and $X_r \cap R_{x_m} = \emptyset$.

An ordered factorization that satisfies the condition in Theorem 23.1 is called an *admissible sequence*.

The following example illustrates the use of Theorem 23.1 for recovering the joint distribution. Additionally, it sheds light on the need for the notion of *minimality* in Definition 23.2.

Example 23.1 We are interested in recovering $P(X, Y, Z)$ given the m -graph in Figure 23.2 (a). We discern from the graph that Definition 23.2 is satisfied because: (1) $P(Y | X, Z, R_y) = P(Y | X, Z)$ and (X, Z) is a minimal set such that $Y \perp\!\!\!\perp (\{X, Z, R_y\} \setminus (X, Z)) | (X, Z)$, (2) $P(X | R_y, Z) = P(X | R_y)$ and R_y is the minimal set such that $X \perp\!\!\!\perp (\{R_y, Z\} \setminus R_y) | R_y$ and (3) $P(Z | R_y) = P(Z)$ and \emptyset is the minimal set such that $Z \perp\!\!\!\perp R_y | \emptyset$. Therefore, the order $Y < X < Z < R_y$ induces a general ordered factorization $P(X, Y, Z, R_y) = P(Y | X, Z)P(X | R_y)P(Z)P(R_y)$. We now rewrite $P(X, Y, Z)$ as follows:

$$P(X, Y, Z) = \sum_{R_y} P(Y, X, Z, R_y) = P(Y | X, Z)P(Z) \sum_{R_y} P(X | R_y)P(R_y)$$

Since $Y \perp\!\!\!\perp R_y | X, Z$, $Z \perp\!\!\!\perp R_z$, $X \perp\!\!\!\perp R_x | R_y$, by Theorem 23.1 we have,

$$P(X, Y, Z) = P(Y|X, Z, R'_x, R'_y, R'_z)P(Z|R'_z) \sum_{R_y} P(X|R'_x, R_y)P(R_y)$$

Indeed, Equation 23.1 permits us to rewrite it as:

$$P(X, Y, Z) = P(Y^*|X^*, Z^*, R'_x, R'_y, R'_z)P(Z^*|R'_z) \sum_{R_y} P(X^*|R'_x, R_y)P(R_y)$$

$P(X, Y, Z)$ is recoverable because every term in the right hand side is consistently estimable from the available dataset.

Had we ignored the minimality requirement in Definition 23.2 and chosen to factorize $Y < X < Z < R_y$ using the chain rule, we would have obtained: $P(X, Y, Z, R_y) = P(Y|X, Z, R_y)P(X|Z, R_y)P(Z|R_y)P(R_y)$ which is not admissible since $X \perp\!\!\!\perp (R_z, R_x) | Z$ does not hold in the graph. In other words, existence of one admissible sequence based on an order O of variables does not guarantee that every factorization based on O is admissible; it is for this reason that we need to impose the condition of minimality in Definition 23.2.

The recovery procedure presented in Example 23.1 requires that we introduce R_y into the order. Indeed, there is no ordered factorization over the substantive variables $\{X, Y, Z\}$ that will permit recoverability of $P(X, Y, Z)$ in Figure 23.2 (a). This extension of Mohan et al. (2013) thus permits the recovery of probabilistic queries from problems in which the missingness mechanisms interact with one another.

23.4 Recoverability in the Absence of an Admissible Sequence

Mohan et al. (2013) presented a theorem (refer to Appendix 23.A.4) that stated the necessary and sufficient condition for recovering the joint distribution for the class of problems in which the parent set of every R variable is a subset of $V_o \cup V_m$. In contrast to Theorem 23.1, their theorem can handle problems for which no admissible sequence exists. The following theorem gives a generalization and is applicable to any given semi-Markovian model (for example, m -graphs in Figure 23.2 (b) & (c)). It relies on the notion of collider path and two new subsets, R^{part} : the partitions of R variables and $Mb(R^{(i)})$: substantive variables related to $R^{(i)}$, which we will define after stating the theorem.

Theorem 23.2 *Given an m -graph G in which no element in V_m is either a neighbor of its missingness mechanism or connected to its missingness mechanism by a collider path, $P(V)$ is recoverable if no $Mb(R^{(i)})$ contains a partially observed variable X such that $R_x \in R^{(i)}$ i.e., $\forall i$,*

$R^{(i)} \cap R_{Mb(R^{(i)})} = \emptyset$. Moreover, if recoverable, $P(V)$ is given by,

$$P(V) = \frac{P(V, R = 0)}{\prod_i P(R^{(i)} = 0 | Mb(R^{(i)}), R_{Mb(R^{(i)})} = 0)}$$

In Theorem 23.2:

- (i) collider path p between any two nodes X and Y is a path in which every intermediate node is a collider. Example, $X \rightarrow Z \leftarrow \dots \rightarrow Y$.
- (ii) $R^{part} = \{R^{(1)}, R^{(2)}, \dots, R^{(N)}\}$ are partitions of R variables such that for every element R_x and R_y , belonging to distinct partitions, the following conditions hold true: (i) R_x and R_y are not neighbors and (ii) R_x and R_y are not connected by a collider path. In Figure 23.2 (b): $R^{part} = \{R^{(1)}, R^{(2)}\}$ where $R^{(1)} = \{R_w, R_z\}$, $R^{(2)} = \{R_x, R_y\}$.
- (iii) $Mb(R^{(i)})$ is the Markov blanket of $R^{(i)}$ comprising of all substantive variables that are either neighbors or connected to variables in $R^{(i)}$ by a collider path (Richardson 2003). In Figure 23.2 (b): $Mb(R^{(1)}) = \{X, Y\}$ and $Mb(R^{(2)}) = \{Z, W\}$.

Appendix 23.A.6 demonstrates how Theorem 23.2 leads to the recoverability of $P(V)$ in Figure 23.2, to which theorems in Mohan et al. (2013) do not apply.

The following corollary yields a sufficient condition for recovering the joint distribution from the class of problems in which no bi-directed edge exists between variables in sets R and $V_o \cup V_m$ (for example, the m -graph described in Figure 23.2 (c)). These problems form a subset of the class of problems covered in Theorem 23.2. Subset $Pa^{sub}(R^{(i)})$ used in the corollary is the set of all substantive variables that are parents of variables in $R^{(i)}$. In Figure 23.2 (b): $Pa^{sub}(R^{(1)}) = \emptyset$ and $Pa^{sub}(R^{(2)}) = \{Z, W\}$.

Corollary 23.1 *Let G be an m -graph such that (i) $\forall X \in V_m \cup V_o$, no latent variable is a common parent of X and any member of R , and (ii) $\forall Y \in V_m$, Y is not a parent of R_y . If $\forall i$, $Pa^{sub}(R^{(i)})$ does not contain a partially observed variables whose missing mechanism is in $R^{(i)}$ i.e., $R^{(i)} \cap R_{Pa^{sub}(R^{(i)})} = \emptyset$, then $P(V)$ is recoverable and is given by,*

$$P(v) = \frac{P(R = 0, V)}{\prod_i P(R^{(i)} = 0 | Pa^{sub}(R^{(i)}), R_{Pa^{sub}(R^{(i)})} = 0)}.$$

23.5 Non-recoverability Criteria for Joint and Conditional Distributions

Up until now, we dealt with sufficient conditions for recoverability. It is important however to supplement these results with criteria for non-recoverability in order to

alert the user to the fact that the available assumptions are insufficient to produce a consistent estimate of the target query. Such criteria have not been treated formally in the literature thus far. In the following theorem we introduce two graphical conditions that preclude recoverability.

Theorem 23.3 Non-recoverability of $P(V)$

Given a semi-Markovian model G , the following conditions are necessary for recoverability of the joint distribution:

- (i) $\forall X \in V_m$, X and R_x are not neighbors and
- (ii) $\forall X \in V_m$, there does not exist a path from X to R_x in which every intermediate node is both a collider and a substantive variable.

In the following corollary, we leverage Theorem 23.3 to yield necessary conditions for recovering conditional distributions.

Corollary 23.2 Non-recoverability of $P(Y|X)$

Let X and Y be disjoint subsets of substantive variables. $P(Y|X)$ is non-recoverable in m -graph G if one of the following conditions is true:

- (1) Y and R_y are neighbors
- (2) G contains a collider path p connecting Y and R_y such that all intermediate nodes in p are in X .

23.6 Recovering Causal Queries

Given a causal query and a causal Bayesian network a complete algorithm exists for deciding whether the query is identifiable or not (Shpitser and Pearl 2006). Obviously, a query that is not identifiable in the substantive model is not recoverable from missing data. Therefore, a necessary condition for recoverability of a causal query is its identifiability which we will assume in the rest of our discussion.

Definition 23.3 Trivially Recoverable Query

A causal query Q is said to be trivially recoverable given an m -graph G if it has an estimand (in terms of substantive variables) in which every factor is recoverable.

Classes of problems that fall into the MCAR (Missing Completely At Random) and MAR (Missing At Random) category are much discussed in the literature (Rubin 1976) because in such categories probabilistic queries are recoverable by graph-blind algorithms. An immediate but important implication of trivial recoverability is that if data are MAR or MCAR and the query is identifiable, then it is also recoverable by model-blind algorithms.

Example 23.2 In the gender wage-gap study example in Figure 23.1 (a), the effect of sex on income, $P(I|do(S))$, is identifiable and is given by $P(I|S)$. By Theorem 23.2, $P(S, X, Q, I)$ is recoverable. Hence $P(I|do(S))$ is recoverable.

23.6.1 Recovering $P(y|do(z))$ when Y and R_y are inseparable

The recoverability of $P(V)$ hinges on the separability of a partially observed variable from its missingness mechanism (a condition established in Theorem 23.3). Remarkably, causal queries may circumvent this requirement. The following example demonstrates that $P(y|do(z))$ is recoverable even when Y and R_y are not separable.

Example 23.3 Examine Figure 23.3. By backdoor criterion, $P(y|do(z)) = \sum_w P(y|z, w)P(w)$. One might be tempted to conclude that the causal relation is non-recoverable because $P(w, z, y)$ is non-recoverable (by Theorem 23.2) and $P(y|z, w)$ is not recoverable (by Corollary 23.2). However, $P(y|do(z))$ is recoverable as demonstrated below:

$$P(y|do(z)) = P(y|do(z), R'_y) = \sum_w P(y|do(z), w, R'_y)P(w|do(z), R'_y) \quad (23.2)$$

$$P(y|do(z), w, R'_y) = P(y|z, w, R'_y) \quad (\text{by Rule-2 of do-calculus (Pearl 2009)}) \quad (23.3)$$

$$P(w|do(z), R'_y) = P(w|R'_y) \quad (\text{by Rule-3 of do-calculus}) \quad (23.4)$$

Substituting (23.3) and (23.4) in (23.2) we get:

$$P(y|do(z)) = \sum_w P(y|z, w, R'_y)P(w|R'_y) = \sum_w P(y^*|z, w, R'_y)P(w|R'_y)$$

The recoverability of $P(y|do(z))$ in the previous example follows from the notion of d^* -separability and dormant independence (Shpitser and Pearl 2008).

Definition 23.4 d^* -separation (Shpitser and Pearl 2008)

Let G be a causal diagram. Variable sets X, Y are d^* -separated in G given Z, W (written $X \perp_w Y|Z$), if we can find sets Z, W , such that $X \perp Y|Z$ in $G_{\overline{w}}$, and $P(y, x|z, do(w))$ is identifiable.

Definition 23.5 Inducing path (Verma and Pearl 1991)

A path p between X and Y is called an inducing path if every node on the path is a collider and an ancestor of either X or Y .

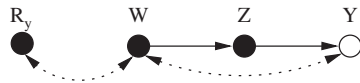


Figure 23.3 m -graph in which Y and R_y are not separable but still $P(Y|do(Z))$ is recoverable.

Theorem 23.4 Given an m -graph in which $|V_m| = 1$ and Y and R_y are connected by an inducing path, $P(y|do(x))$ is recoverable if there exists Z, W such that $Y \perp_w R_y|Z$ and for $W = W \setminus X$, the following conditions hold:

- (1) $Y \perp\!\!\!\perp W_1|X, Z$ in $G_{\bar{X}, W_1}$ and
- (2) $P(W_1, Z|do(X))$ and $P(Y|do(W_1), do(X), Z, R'_y)$ are identifiable.

Moreover, if recoverable then,

$$P(y|do(x)) = \sum_{W_1, Z} P(Y|do(W), do(X), Z, R'_y)P(Z, W_1|do(X)).$$

We can quickly conclude that $P(y|do(z))$ is recoverable in the m -graph in Figure 23.3 by verifying that the conditions in Theorem 23.4 hold in the m -graph.

23.7 Attrition

Attrition (i.e., participants dropping out from a study/experiment), is a ubiquitous phenomenon, especially in longitudinal studies. In this section, we shall discuss a special case of attrition called ‘Simple Attrition’ (for an in-depth treatment see Garcia 2013). In this problem, a researcher conducts a randomized trial, measures a set of variables (X, Y, Z) and obtains a dataset where outcome (Y) is corrupted by missing values (due to attrition). Clearly, due to randomization, the effect of treatment (X) on outcome (Y), $P(y|do(x))$, is identifiable and is given by $P(Y|X)$. We shall now demonstrate the usefulness of our previous discussion in recovering $P(y|do(x))$. Typical attrition problems are depicted in Figure 23.4. In Figure 23.4 (b) we can apply Theorem 23.1 to recover $P(y|do(x))$ as given below: $P(Y|X) = \sum_Z P(Y^*|X, Z, R'_y)P(Z|X)$. In Figure 23.4 (a), we observe that Y and R_y are connected by a collider path. Therefore by Corollary 23.2, $P(Y|X)$ is not recoverable; hence $P(y|do(x))$ is also not recoverable.

23.7.1 Recovering Joint Distributions under Simple Attrition

The following theorem yields the *necessary and sufficient* condition for recovering joint distributions from semi-Markovian models with a single partially observed variable i.e., $|V_m| = 1$ which includes models afflicted by simple attrition.



Figure 23.4 (a) m -graphs in which $P(y|do(x))$ is not recoverable (b) m -graphs in which $P(y|do(x))$ is recoverable.

Theorem 23.5 *Let $Y \in V_m$ and $|V_m| = 1$. $P(V)$ is recoverable in m -graph G if and only if Y and R_y are not neighbors and Y and R_y are not connected by a path in which all intermediate nodes are colliders. If both conditions are satisfied, then $P(V)$ is given by, $P(V) = P(Y|V_O, R_y = 0)P(V_O)$.*

23.7.2 Recovering Causal Effects under Simple Attrition

Theorem 23.6 *$P(y|do(x))$ is recoverable in the simple attrition case (with one partially observed variable) if and only if Y and R_y are neither neighbors nor connected by an inducing path. Moreover, if recoverable,*

$$P(Y|X) = \sum_z P(Y^*|X, Z, R'_y)P(Z|X) \quad (23.5)$$

where Z is the separating set that d -separates Y from R_y .

23.8 Related Work

Deletion based methods such as listwise deletion that are easy to understand as well as implement, guarantee consistent estimates only for certain categories of missingness such as MCAR (Rubin 1976). Maximum Likelihood method is known to yield consistent estimates under MAR assumption; expectation maximization algorithm and gradient based algorithms are widely used for searching for ML estimates under incomplete data (Lauritzen 1995; Dempster et al. 1977; Darwiche 2009; Koller and Friedman 2009). Most work in machine learning assumes MAR and proceeds with ML or Bayesian inference. However, there are exceptions such as recent work on collaborative filtering and recommender systems which develop probabilistic models that explicitly incorporate missing data mechanism (Marlin et al. 2011; Marlin and Zemel 2009; Marlin et al. 2007).

Other methods for handling missing data can be classified into two: (a) Inverse Probability Weighted Methods and (b) Imputation based methods (Rothman et al. 2008). Inverse Probability Weighting methods analyze and assign weights to complete records based on estimated probabilities of completeness (van der Laan and Robins 2003; Robins et al. 1994). Imputation based methods substitute a reasonable guess in the place of a missing value (Allison 2002) and Multiple Imputation (Little and Rubin 2002) is a widely used imputation method.

Missing data is a special case of coarsened data and data are said to be coarsened at random (CAR) if the coarsening mechanism is only a function of the observed data (Heitjan and Rubin 1991). Robins and Rotnitzky (1992) introduced

a methodology for parameter estimation from data structures for which full data has a non-zero probability of being fully observed and their methodology was later extended to deal with censored data in which complete data on subjects are never observed (van der Laan and Robins 1998).

The use of graphical models for handling missing data is a relatively new development. Daniel et al. (2012) used graphical models for analyzing missing information in the form of missing cases (due to sample selection bias). Attrition is a common occurrence in longitudinal studies and arises when subjects drop out of the study (Twisk and de Vente 2002; Shadish 2002) and Garcia (2013) analysed the problem of attrition using causal graphs. Thoemmes and Rose (2013) and Thoemmes and Mohan (2015) cautioned the practitioner that contrary to popular belief, not all auxiliary variables reduce bias. Both Garcia (2013) and Thoemmes and Rose (2013) associate missingness with a single variable and interactions among several missingness mechanisms are unexplored.

Mohan et al. (2013) employed a formal representation called Missingness Graphs to depict the missingness process, defined the notion of recoverability and derived conditions under which queries would be recoverable when datasets are categorized as Missing Not At Random (MNAR). Tests to detect misspecifications in the m -graph are discussed in Mohan and Pearl (2014).

23.9 Conclusion

Graphical models play a critical role in portraying the missingness process, encoding and communicating assumptions about missingness and deciding recoverability given a dataset afflicted with missingness. We presented graphical conditions for recovering joint and conditional distributions and sufficient conditions for recovering causal queries. We exemplified the recoverability of causal queries of the form $P(y|do(x))$ despite the existence of an inseparable path between Y and R_y , which is an insurmountable obstacle to the recovery of $P(Y)$. We applied our results to problems of attrition and presented necessary and sufficient graphical conditions for recovering causal effects in such problems.

Acknowledgments

This paper has benefited from discussions with Ilya Shpitser. This research was supported in parts by grants from NSF #IIS1249822 and #IIS1302448, and ONR #N00014-13-1-0153 and #N00014-10-1-0933.

References

P.D. Allison. *Missing data: Series: Quantitative applications in the social sciences*, Thousand Oaks, CA: Sage Publications, 2002.

- R.M. Daniel, M.G. Kenward, S.N. Cousens, and B.L. De Stavola. Using causal diagrams to guide analysis in missing data problems. *Statistical Methods in Medical Research*, 21(3):243–256, 2012.
- A. Darwiche. *Modeling and reasoning with Bayesian networks*. Cambridge University Press, 2009.
- A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1): 1–38, 1977.
- F. M. Garcia. Definition and diagnosis of problematic attrition in randomized controlled experiments. Working paper, April 2013. Available at SSRN: <http://ssrn.com/abstract=2267120>.
- D.F. Heitjan and D.B. Rubin. Ignorability and coarse data. *The Annals of Statistics*, 19(4): 2244–2253, 1991.
- D. Koller and N. Friedman. *Probabilistic graphical models: Principles and techniques*. Cambridge, MA: The MIT Press, 2009.
- S L Lauritzen. The EM algorithm for graphical association models with missing data. *Computational Statistics & Data Analysis*, 19(2):191–201, 1995.
- R.J.A. Little and D.B. Rubin. *Statistical analysis with missing data*. Hoboken, NJ: Wiley, 2002.
- B.M. Marlin and R.S. Zemel. Collaborative prediction and ranking with non-random missing data. In *Proceedings of the third ACM conference on Recommender systems*, 5–12. ACM, 2009.
- B.M. Marlin, R.S. Zemel, S. Roweis, and M. Slaney. Collaborative filtering and the missing at random assumption. *Proceedings of the Twenty-Third Conference on Uncertainty in Artificial Intelligence*, AUAI Press, 267–275, 2007.
- B.M. Marlin, R.S. Zemel, S.T. Roweis, and M. Slaney. Recommender systems: missing data and statistical model estimation. *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence*, AAAI Press, 2686–2691, 2011.
- K. Mohan and J. Pearl. On the testability of models with missing data. *Proceedings of AIS-TAT*, 2014.
- K. Mohan, J. Pearl, and J. Tian. Graphical models for inference with missing data. In C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems 26*, 1277–1285. 2013.
- J. Pearl. *Causality: Models, reasoning and inference*. Cambridge Univ Press, New York, 2nd edition, 2009.
- J. Pearl and K. Mohan. Recoverability and testability of missing data: Introduction and summary of results. Technical Report R-417, UCLA, 2013. Available at http://ftp.cs.ucla.edu/pub/stat_ser/r417.pdf.
- T. Richardson. Markov properties for acyclic directed mixed graphs. *Scandinavian Journal of Statistics*, 30(1):145–157, 2003.
- J.M. Robins and A. Rotnitzky. Recovery of information and adjustment for dependent censoring using surrogate markers. In N.P. Jewell, K. Kietz, and V.T. Farewell (Eds.), *AIDS Epidemiology*, Boston, MA: Birkhäuser, New York, NY: Springer 297–331, 1992.

- J.M. Robins, A. Rotnitzky, and L. P. Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866, 1994.
- K.J. Rothman, S. Greenland, and T.L. Lash. *Modern epidemiology*. Philadelphia, PA: Lippincott Williams & Wilkins, 2008.
- D.B. Rubin. Inference and missing data. *Biometrika*, 63:581–592, 1976.
- W.R. Shadish. Revisiting field experimentation: field notes for the future. *Psychological methods*, 7(1):3, 2002.
- I. Shpitser and J. Pearl. Identification of conditional interventional distributions. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, Cambridge, MA: AUAI Press, 437–444, 2006.
- I. Shpitser and J. Pearl. Dormant independence. *Proceedings of the Twenty Third Conference on the Association for the Advancement of Artificial Intelligence*, Menlo Park, CA: AAAI Press, 1081–1087, 2008.
- F. Thoemmes and K. Mohan. Graphical representation of missing data problems. *Structural Equation Modeling: A Multidisciplinary Journal*, 2015.
- F. Thoemmes and N. Rose. Selection of auxiliary variables in missing data problems: Not all auxiliary variables are created equal. Technical Report R-002, Cornell University, 2013.
- J. Twisk and W. de Vente. Attrition in longitudinal studies: How to deal with missing data. *Journal of clinical epidemiology*, 55(4):329–337, 2002.
- M.J. van der Laan and J.M. Robins. Locally efficient estimation with current status data and time-dependent covariates. *Journal of the American Statistical Association*, 93(442): 693–701, 1998.
- M.J. van der Laan and J.M. Robins. *Unified methods for censored longitudinal data and causality*. New York, NY: Springer Verlag, 2003.
- T.S Verma and J. Pearl. Equivalence and synthesis of causal models. In *Proceedings of the Sixth Conference in Artificial Intelligence*, AUAI Press, 220–227, 1991.

23.A

23.A.1

Appendix

Missingness Process in Figure 23.1

Figure 23.1 Missingness Graph depicting the missingness process in a hypothetical (job-specific) gender wage gap study that measured the variables: sex (S), work experience (X), qualification (Q) and income (I). Fully observed and partially observed variables are represented by filled and hollow nodes respectively. While sex and work experience were found to be fully observed in all records i.e., $V_o = \{S, X\}$, qualification and income were found to be missing in some of the records i.e., $V_m = \{Q, I\}$. R_Q and R_I denote the causes of missingness of Q and I respectively and are assumed to be independent of S, Q, I and X . The assumptions in the model are: (1) women are likely to be less qualified and experienced than

men, (2) income is determined by qualification and job experience of the candidate, and (3) missingness in Q and I are correlated, caused by unobserved common factors such as laziness or resistance to respond.

23.A.2 Testing Compatibility between Underlying and Manifest Distributions

Example 23.4 Let the incomplete dataset contain two partially observed variables, Z and W . The tests for compatibility between manifest distribution: $P_m(Z^*, W^*, R_z, R_w)$ and the underlying distribution: $P_u(Z, W, R_z, R_w)$ are:

Case-1: Let $X = \{Z, W\}$, then $Y = V_m \setminus X = \{\}$

$$P_m(Z^* = z, W^* = w, R_z = 0, R_w = 0) = P_u(Z = z, W = w, R_z = 0, R_w = 0) \forall z, w$$

Case-2: Let $X = \{Z\}$, then $Y = \{W\}$

$$P_m(Z^* = z, W^* = m, R_z = 0, R_w = 1) = \sum_w P_u(Z = z, w, R_z = 0, R_w = 1) \forall z$$

Case-3: Let $X = \{W\}$, then $Y = \{Z\}$

$$P_m(Z^* = m, W^* = w, R_z = 1, R_w = 0) = \sum_z P_u(z, W = w, R_z = 1, R_w = 0) \forall w$$

Case-4: Let $X = \{\}$, then $Y = \{Z, W\}$

$$P_m(Z^* = m, W^* = m, R_z = 1, R_w = 1) = \sum_{z,w} P_u(z, w, R_z = 1, R_w = 1)$$

23.A.3 Proof of Theorem 23.1

Proof. follows from Theorem 23.1 in Mohan et al. (2013) (restated below as Theorem 23.7) noting that ordered factorization is one specific form of decomposition. ■

Theorem 23.7 Mohan et al. (2013)

A query Q defined over variables in $V_o \cup V_m$ is recoverable if it is decomposable into terms of the form $Q_j = P(S_j|T_j)$ such that T_j contains the missingness mechanism $R_v = 0$ of every partially observed variable V that appears in Q_j .

23.A.4 Recovering $P(V)$ when Parents of R belong to $V_o \cup V_m$

Theorem 23.8 Recoverability of the Joint $P(V)$ (Mohan et al. 2013)

Given a m -graph G with no edges between the R variables and no latent variables as parents of R variables, a necessary and sufficient condition for recovering the joint

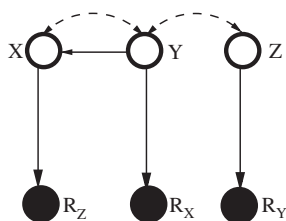


Figure 23.5 m -graph in which joint distribution is recoverable.

distribution $P(V)$ is that no variable X be a parent of its missingness mechanism R_X . Moreover, when recoverable, $P(V)$ is given by

$$P(v) = \frac{P(R = 0, v)}{\prod_i P(R_i = 0 | pa_{r_i}^o, pa_{r_i}^m, R_{pa_{r_i}^m} = 0)} \quad (23.6)$$

where $Pa_{r_i}^o \subseteq V_o$ and $Pa_{r_i}^m \subseteq V_m$ are the parents of R_i .

Example 23.5 We wish to recover $P(X, Y, Z)$ from the m -graph in Figure 23.1 (a). An enumeration of various orderings will reveal that none of the orders are admissible. Nevertheless, using Theorem 23.8, we can recover the joint probability as given below:

$$P(X, Y, Z) = \frac{P(R'_x, R'_y, R'_z, X, Y, Z)}{P(R'_z | X, R'_x) P(R'_x | Y, R'_y) P(R'_y | Z, R'_z)}$$

23.A.5 Proof of Theorem 23.2

Proof.

$$\begin{aligned} P(V) &= \frac{P(R = 0, V)}{P(R = 0 | V)} \\ &= \frac{P(R = 0, V)}{P(R^{(1)} = 0, R^{(2)} = 0, \dots, R^N = 0 | V)} \end{aligned}$$

$Mb(R^{(i)})$ d -separates $R^{(i)}$ from all variables that are not in $R^{(i)} \cup Mb(R^{(i)})$ i.e., $R^{(i)} \perp\!\!\!\perp (\{R, V\} - \{R^{(i)}, Mb(R^{(i)})\}) | Mb(R^{(i)})$. Hence,

$$P(V) = \frac{P(R = 0, V)}{\prod_i P(R^{(i)} = 0 | Mb(R^{(i)}))}$$

Using $R^{(i)} \cap R_{Mb(R^{(i)})} = \emptyset$ and $R^{(i)} \perp\!\!\!\perp (\{R, V\} - \{R^{(i)}, Mb(R^{(i)})\}) | Mb(R^{(i)})$ we get,

$$P(V) = \frac{P(R = 0, V)}{\prod_i P(R^{(i)} = 0 | Mb(R^{(i)}), R_{Mb(R^{(i)})} = 0)}$$

Now we can directly apply Equation 23.1 and express $P(V)$ in terms of quantities estimable from the available dataset. Therefore, $P(V)$ is recoverable. ■

23.A.6 Example: Recoverability by Theorem 23.2

Example 23.6 $P(X, Y, Z, W)$ is the query of interest and Figure 23.2 (b) depicts the missingness process and identifies the sets R^{part} and $Mb(R^{(i)})$. A quick inspection reveals that no admissible sequence exists. However, notice that $CI_1 : R^{(1)} \perp\!\!\!\perp (R^{(2)}, Mb(R^{(2)})) | Mb(R^{(1)})$ and $CI_2 : R^{(2)} \perp\!\!\!\perp (R^{(1)}, Mb(R^{(1)})) | Mb(R^{(2)})$ hold in the m -graph. We exploit these independencies to recover the joint distribution as detailed below:

$$\begin{aligned} P(X, Y, Z, W) &= \frac{P(R = 0, X, Y, Z, W)}{P(R = 0 | X, Y, Z, W)} = \frac{P(R = 0, X, Y, Z, W)}{P(R^{(1)} = 0, R^{(2)} = 0 | X, Y, Z, W)} \\ &= \frac{P(R = 0, X, Y, Z, W)}{P(R^{(1)} = 0 | X, Y, R^{(2)} = 0) P(R^{(2)} = 0 | Z, W, R^{(1)} = 0)} \quad (\text{Using } CI_1 \text{ and } CI_2) \end{aligned}$$

$$P(V) = \frac{P(R = 0, X^*, Y^*, Z^*, W^*)}{P(R_w = 0, R_z = 0 | X^*, Y^*, R_x = 0, R_y = 0) P(R_x = 0, R_y = 0 | Z^*, W^*, R_z = 0, R_w = 0)}$$

(By Equation 23.1)

23.A.7 Proof of Corollary 23.1

Proof.

$$P(V) = \frac{P(R = 0, V)}{P(R = 0 | V)} = \frac{P(R = 0, V)}{P(R^{(1)}, R^{(2)}, \dots, R^N | V)}.$$

Since $Pa^{sub}(R^{(i)}) \subseteq V$ d -separates R_i from all the other variables in $(V \cup R) \setminus (R^{(i)} \cup Pa^{sub}(R^{(i)}))$, we get

$$P(V) = \frac{P(R = 0, V)}{\prod_i P(R^{(i)} = 0 | Pa^{sub}(R^{(i)}))}.$$

Using $R^{(i)} \cap R_{Pa^{sub}(R^{(i)})} = \emptyset$ and $R^{(i)} \perp\!\!\!\perp (\{R, V\} - \{R^{(i)}, Pa^{sub}(R^{(i)})\}) | Pa^{sub}(R^{(i)})$ we get,

$$P(V) = \frac{P(R = 0, V)}{\prod_i P(R^{(i)} = 0 | Pa^{sub}(R^{(i)}), R_{Pa^{sub}(R^{(i)})} = 0)}.$$

■

23.A.8 Proof of Theorem 23.3

We will be using the following lemma (stated and proved in Mohan et al. (2013) (Supplementary materials)) in our proof.

Lemma 23.1 *If a target relation Q is not recoverable in m -graph G , then Q is not recoverable in the graph G' resulting from adding a single edge to G .*

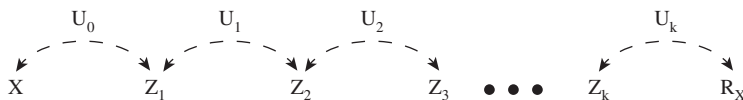


Figure 23.6 An m -graph in which $P(X, Z)$ is not-recoverable where $Z = \{Z_1, Z_2, \dots, Z_k\}$. X is partially observed, all Z variables are fully observed, parents of Z_i are U_{i-1} and U_i , parent of X is U_0 and parent of R_x is U_k .

Proof. Non-recoverability of $P(V)$ when X is a parent of R_x has been proved in Mohan et al. (2013). We will now prove non-recoverability of $P(X)$ and hence $P(V)$ when X and R_x have a latent parent.

M_1 and M_2 are two models in which variables U, X and R_x are binary and U is a fair coin. In M_1 , $X = 0$ and $R_x = u$ and in M_2 , $X = u$ and $R_x = u$. Notice that although the two models agree on the manifest distribution, they disagree on the query $P(X)$. Hence $P(X)$ is non-recoverable in $X \leftarrow U \rightarrow R_x$. Using Lemma 23.1, we can conclude that $P(V)$ is non-recoverable in any m -graph in which X and R_x are connected by a bi-directed edge.

Given the m -graph in Figure 23.6 we will now prove that $P(X, Z_1, Z_2, \dots, Z_k)$ is non-recoverable. Let M_3 and M_4 be two models such that all the variables are binary, all the U variables are fair coins, $X = U_0$, $R_x = U_k$ and $Z_i = U_{i-1} \oplus U_i$, $1 \leq i < k$. In M_3 , $Z_k = U_{k-1}$ and in M_4 , $Z_k = U_{k-1} \oplus U_k$. Both models yield the same manifest distribution. However, they disagree on the query $P(X, Z_1, Z_2, \dots, Z_k)$. For instance, in M_3 , $P(X = 0, Z = 0, R_x = 1) > 0$ where as in M_4 , $P(X = 0, Z = 0, R_x = 1) = 0$. Therefore in M_4 , $P(X = 0, Z = 0) = P(X = 0, Z = 0, R_x = 0)$ and in M_3 , $P(X = 0, Z = 0) = P(X = 0, Z = 0, R_x = 0) + P(X = 0, Z = 0, R_x = 1)$. Hence in the m -graph in Figure 23.6, the joint distribution $P(X, Z)$ is non-recoverable. Using Lemma 23.1, we can conclude that joint distribution is non-recoverable in any m -graph which has a bi-directed path from any partially observed variable X to its missingness mechanism R_x . ■

23.A.9 Proof of Corollary 23.2

Proof. Let $|V_m| = 1$ and $Y_1 \in Y$ be the only partially observed variable. Let G' be the subgraph containing all variables in $X \cup Y \cup \{R_{y_1}, Y_1^*\}$. We know that if (1) or (2) are true, then, (i) $P(X, Y)$ is not recoverable in G' and (ii) $P(X)$ is recoverable in G' . Therefore, $P(Y|X) = \frac{P(Y, X)}{P(X)}$ is not recoverable in G' and hence by Lemma 23.1, not recoverable in G . ■

23.A.10 Proof of Theorem 23.4

Proof. $P(Y|do(X)) = \sum_{z, w'} P(Y|Z, W', do(X))P(Z, W'|do(X))$.

If condition 1 holds, then by Rule-2 of *do*-calculus (Pearl 2009) we have:

$$P(Y|Z, W', do(X)) = P(Y|Z, do(X), do(W')).$$

Since $Y \perp_w R_y|Z$,

$$\begin{aligned} P(Y|Z, do(X), do(W')) &= P(Y|Z, do(X), do(W'), R'_y) \\ &= P(Y^*|Z, do(X), do(W'), R'_y). \end{aligned}$$

Therefore, $P(y|do(x))$ is recoverable. ■

23.A.11 Proof of Theorem 23.5

Proof.

(Sufficiency) Whenever (1) and (2) are satisfied, $Y \perp R_y|V_o$ holds. Hence, $P(V)$ which may be written as $P(Y|V_o)P(V_o)$ can be recovered as $P(Y^*|V_o, R_y = 0)P(V_o)$.

(Necessity) Follows from Theorem 23.2. ■

23.A.12 Proof of Theorem 23.6

Proof.

(Sufficiency) Under simple attrition, all paths to R_y from Y containing X are blocked by X . Therefore, when both conditions specified in the theorem are satisfied, it implies that Y and R_y are separable. Given that Z is any separator between Y and R_y , $P(Y|X)$ may be recovered as $\sum_z P(Y^*|X, Z, R'_y)P(Z|X)$.

(Necessity) Follows from Theorem 23.2. ■

Recovering from Selection Bias in Causal and Statistical Inference

Elias Bareinboim*, Jin Tian[†], and Judea Pearl

Abstract

Selection bias is caused by preferential exclusion of units from the samples and represents a major obstacle to valid causal and statistical inferences; it cannot be removed by randomized experiments and can rarely be detected in either experimental or observational studies. In this paper, we provide complete graphical and algorithmic conditions for recovering conditional probabilities from selection biased data. We also provide graphical conditions for recoverability when unbiased data is available over a subset of the variables. Finally, we provide a graphical condition that generalizes the backdoor criterion and serves to recover causal effects when the data is collected under preferential selection.

24.1 Introduction

Selection bias is induced by preferential selection of units for data analysis, usually governed by unknown factors including treatment, outcome, and their consequences, and represents a major obstacle to valid causal and statistical inferences. It cannot be removed by randomized experiments and can rarely be detected in

*University of California, Los Angeles; [†]Iowa State University

Originally published in Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence Palo Alto, CA: AAAI Press, 2410-2416, 2014. “Best Paper Award.”

Copyright 2020, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved. Republished with permission from AAAI.

either experimental or observational studies.¹ For instance, in a typical study of the effect of training program on earnings, subjects achieving higher incomes tend to report their earnings more frequently than those who earn less. The data-gathering process in this case will reflect this distortion in the sample proportions and, since the sample is no longer a faithful representation of the population, biased estimates will be produced regardless of how many samples were collected.

This preferential selection challenges the validity of inferences in several tasks in AI (Cooper 1995; Elkan 2001; Zadrozny 2004; Cortes et al. 2008) and Statistics (Whittemore 1978; Little and Rubin 1986; Jewell 1991; Kuroki and Cai 2006) as well as in the empirical sciences (e.g., Genetics (Pirinen, Donnelly, and Spencer 2012; Mefford and Witte 2012), Economics (Heckman 1979; Angrist 1997), and Epidemiology (Robins 2001; Glymour and Greenland 2008)).

To illuminate the nature of preferential selection, consider the data-generating model in Figure 24.1(a) in which X represents an action, Y represents an outcome, and S represents a binary indicator of entry into the data pool ($S = 1$ means that the unit is in the sample, $S = 0$ otherwise). If our goal is to compute the population-level conditional distribution $P(y|x)$, and the samples available are collected under selection, only $P(y, x|S = 1)$ is accessible for use.² Given that in principle these two distributions are just loosely connected, the natural question to ask is under what conditions $P(y|x)$ can be recovered from data coming from $P(y, x|S = 1)$. In this specific example, both action and outcome affect the entry in the data pool, which will be shown not to be recoverable (see Corollary 24.1) – i.e., there is no method capable of unbiasedly estimating the population-level distribution using data gathered under this selection process.

The bias arising from selection differs fundamentally from the one due to *confounding*, though both constitute threats to the validity of causal inferences. The former bias is due to treatment or outcome (or ancestors) affecting the inclusion of the subject in the sample (Figure 24.1(a)), while the latter is the result of treatment X and outcome Y being affected by a common omitted variables U (Figure 24.1(b)). In both cases, we have unblocked extraneous “flow” of information between treatment and outcome, which appear under the rubric of “spurious correlation,” since it is not what we seek to estimate.

It is instructive to understand selection graphically, as in Figure 24.1(a). The preferential selection that is encoded through conditioning on S creates spurious

1. Remarkably, there are special situations in which selection bias can be detected even from observations, as in the form of a non-chordal undirected component (Zhang 2008).

2. In a typical AI task such as classification, we could have X being a collection of features and Y the class to be predicted, and $P(y|x)$ would be the classifier that needs to be trained.

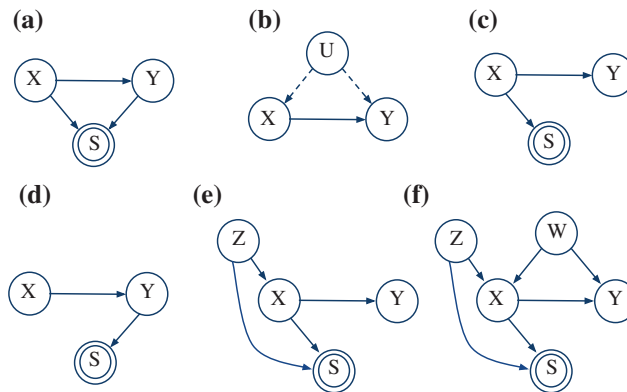


Figure 24.1 (a,b) Simplest examples of selection and confounding bias, respectively. (c,d) Treatment-dependent and outcome-dependent studies under selection, $Q = P(y|x)$ is recoverable in (c) but not in (d). (e,f) Treatment-dependent study where selection is also affected by driver of treatment Z (e.g., age); Q is recoverable in (e) but not in (f).

association between X and Y through two mechanisms. First, given that S is a collider, conditioning on it induces spurious association between its parents, X and Y (Pearl 1988). Second, S is also a descendant of a “virtual collider” Y , whose parents are X and the error term U_Y (also called “hidden variable”) which is always present, though often not shown in the diagram.³

24.1.1 Related Work and Our Contributions

There are three sets of assumptions that are enlightening to acknowledge if we want to understand the procedures available in the literature for treating selection bias – qualitative assumptions about the selection mechanism, parametric assumptions regarding the data-generating model, and quantitative assumptions about the selection process.

In the data-generating model in Figure 24.1(c), the selection of units to the sample is treatment-dependent, which means that it is caused by X , but not Y . This case has been studied in the literature and $Q = P(y|x)$ is known to be non-parametrically recoverable from selection (Greenland and Pearl 2011). Alternatively, in the data-generating model in Figure 24.1(d), the selection is caused by Y (outcome-dependent), and Q is not recoverable from selection (formally shown later on), but is the odds ratio (Cornfield 1951; Whittemore 1978; Geng 1992; Didelez, Kreiner, and Keiding 2010). As mentioned earlier, Q is also not recoverable in the graph in Figure 24.1(a). By and large, the literature is concerned with treatment-dependent or outcome-dependent selection, but selection might be

3. See (Pearl 2000, pp. 339-341) and (Pearl 2013) for further explanations of this bias mechanism.

caused by multiple reasons and embedded in more intricate realities. For instance, a driver of the treatment Z (e.g., age, sex, socio-economic status) may also be causing selection, see Figure 24.1(e,f). As it turns out, Q is recoverable in Figure 24.1(e) but not in (f), so different qualitative assumptions need to be modelled explicitly since each topology entails a different answer for recoverability.

The second assumption is related to the parametric form used by recoverability procedures. For instance, one variation of the selection problem was studied in Econometrics, and led to the celebrated method developed by [James Heckman \(1979\)](#). His two-step procedure removes the bias by leveraging the assumptions of linearity and normality of the data-generating model. A graph-based parametric analysis of selection bias is given in ([Pearl 2013](#)).

The final assumption is about the probability of being selected into the sample. In many settings in Machine learning and Statistics ([Elkan 2001](#); [Zadrozny 2004](#); [Smith and Elkan 2007](#); [Storkey 2009](#); [Hein 2009](#); [Cortes et al. 2008](#)), it is assumed that this probability, $P(S = 1 | Pa_s)$, can be modelled explicitly, which often is an unattainable requirement for the practitioner (e.g., it might be infeasible to assess the differential rates of how salaries are reported).

Our treatment differs fundamentally from the current literature regarding these assumptions. First, we do not constrain the type of data-generating model as outcome- or treatment-dependent, but we take arbitrary models (including these two) as input, in which a node S indicates selection for sampling. Second, we do not make parametric assumptions (e.g. linearity, normality, monotonicity) but operate non-parametrically based on causal graphical models ([Pearl 2000](#)), which is more robust, less prone to model misspecifications. Third, we do not rely on having the selection's probability $P(S = 1 | Pa_s)$, which is not always available in practice. Our work hinges on exploiting the qualitative knowledge encoded in the data-generating model for yielding recoverability. This knowledge is admittedly a demanding requirement for the scientist, but we now understand formally its necessity for *any* approach to recoverability – any procedure aiming for recoverability, implicitly or explicitly, relies on this knowledge ([Pearl 2000](#)).⁴

The analysis of selection bias requires a formal language within which the notion of data-generating model is given precise characterization, and the qualitative assumptions regarding how the variables affect selection can be encoded explicitly. The advent of causal diagrams ([Pearl 1995](#); [Spirtes, Glymour, and Scheines 2000](#); [Pearl 2000](#); [Koller and Friedman 2009](#)) provides such a language and renders the formalization of the selection problem possible.

4. A trivial instance of this necessity is Figure 24.1(c,d) where the odds ratio is recoverable, yet $P(y|x)$ is recoverable in 24.1(c) but not in (d).

Using this language, (Bareinboim and Pearl 2012) provided a complete treatment for selection relative to the OR.⁵ We generalize their treatment considering the estimability of conditional distributions and address three problems:

1. **Selection without external data:** The dataset is collected under selection bias, $P(\mathbf{v} | S = 1)$; under which conditions is $P(\mathbf{y} | \mathbf{x})$ recoverable?
2. **Selection with external data:** The dataset is collected under selection bias, $P(\mathbf{v} | S = 1)$, but there are unbiased samples from $P(\mathbf{t})$, for $\mathbf{T} \subseteq \mathbf{V}$; under which conditions is $P(\mathbf{y} | \mathbf{x})$ recoverable?
3. **Selection in causal inferences:** The data is collected under selection bias, $P(\mathbf{v} | S = 1)$, but there are unbiased samples from $P(\mathbf{t})$, for $\mathbf{T} \subseteq \mathbf{V}$; under which conditions is the interventional distribution $P(\mathbf{y} | do(\mathbf{x}))$ estimable?

We provide graphical and algorithmic conditions for these problems without resorting to parametric assumptions nor selection probabilities. Furthermore, the solution for selection without external data is complete, in the sense that whenever a quantity is said not to be recoverable by our conditions, there exists no procedure that are able to recover it (without adding assumptions). In estimating the effects of interventions, we generalize the *backdoor criterion* for when data is collected under selection.

24.2 Recoverability without External Data

We first introduce the formal notion of recoverability for conditional distributions when data is under selection.⁶

Definition 24.1 s-Recoverability

Given a causal graph G_s augmented with a node S encoding the selection mechanism (Bareinboim and Pearl 2012), the distribution $Q = P(\mathbf{y} | \mathbf{x})$ is said to be *s-recoverable* from selection biased data in G_s if the assumptions embedded in the causal model renders Q expressible in terms of the distribution under selection bias $P(\mathbf{v} | S = 1)$. Formally, for every two probability distributions P_1 and P_2 compatible with G_s , $P_1(\mathbf{v} | S = 1) = P_2(\mathbf{v} | S = 1) > 0$ implies $P_1(\mathbf{y} | \mathbf{x}) = P_2(\mathbf{y} | \mathbf{x})$.⁷

5. The odds ratio (OR) is a commonly used measure of association and has the form $(P(\mathbf{y} | \mathbf{x})P(\bar{\mathbf{y}} | \bar{\mathbf{x}}))/(P(\bar{\mathbf{y}} | \mathbf{x})P(\mathbf{y} | \bar{\mathbf{x}}))$. The symmetric form of the OR allows certain derivations.

6. This definition generalizes G -admissibility given in (Bareinboim and Pearl 2012).

7. We follow the conventions given in (Pearl 2000). We use typical graph notation with families (e.g., children, parents, ancestors). We denote variables by capital letters and their realized values by small letters. We use bold to denote sets of variables. We denote the set of all variables by \mathbf{V} , except for the selection mechanism S .

Consider the graph G_s in Figure 24.1(c) and assume that our goal is to establish s -recoverability of $Q = P(y|x)$. Note that by d -separation (Pearl 1988), X separates Y from S , (or $(Y \perp\!\!\!\perp S|X)$), so we can write $P(y|x) = P(y|x, S = 1)$. This is a very special situation since these two distributions can be arbitrarily distant from each other, but in this specific case G_s constrains Q in such a way that despite the fact that data was collected under selection and our goal is to answer a query about the overall population, there is no need to resort to additional data external to the biased study.

Now we want to establish whether Q is s -recoverable in the graph G_s in Figure 24.1(d). In this case, S is not d -separated from Y if we condition on X , so $(S \perp\!\!\!\perp Y|X)$ does not hold in at least one distribution compatible with G_s , and the identity $P(y|x) = P(y|x, S = 1)$ is not true in general. One may wonder if there is another way to s -recover Q in G_s , but this is not the case as formally shown next. That is, the assumptions encoded in G_s imply a universal impossibility; no matter how many samples of $P(x, y|S = 1)$ are accumulated or how sophisticated the estimation technique is, the estimator of $P(y|x)$ will never converge to its true value.

Lemma 24.1 $P(y|x)$ is not s -recoverable in Figure 24.1(d).

Proof. We construct two causal models such that P_1 is compatible with the graph G_s in Figure 24.1(d) and P_2 with the subgraph $G_2 = G_s \setminus \{Y \rightarrow S\}$. We will set the parameters of P_1 through its factors and then computing the parameters of P_2 by enforcing $P_2(\mathbf{V}|S = 1) = P_1(\mathbf{V}|S = 1)$. Since $P_2(\mathbf{V}|S = 1) = P_2(\mathbf{V})$, we will be enforcing $P_1(\mathbf{V}|S = 1) = P_2(\mathbf{V})$. Recoverability should hold for any parametrization, so we assume that all variables are binary. Given a Markovian causal model (Pearl 2000), P_1 can be parametrized through its factors in the decomposition over observables, $P_1(X), P_1(Y|X), P_1(S = 1|Y)$, for all X, Y .

We can write the conditional distribution in the second causal model as follows:

$$P_2(y|x) = P_1(y|x, S = 1) = \frac{P_1(y, x, S = 1)}{P_1(x, S = 1)} \quad (24.1)$$

$$= \frac{P_1(S = 1|y)P_1(y|x)}{P_1(S = 1|y)P_1(y|x) + P_1(S = 1|\bar{y})P_1(\bar{y}|x)}, \quad (24.2)$$

where the first equality, by construction, should be enforced, and the second and third by probability axioms. The other parameters of P_2 are free and can be chosen to match P_1 .

Finally, set the distribution of every family in P_1 but selection variable equal to $1/2$, and set the distribution $P_1(S = 1|y) = \alpha, P_1(S = 1|\bar{y}) = \beta$, for $0 < \alpha, \beta < 1$ and $\alpha \neq \beta$. This parametrization reduces Equation (24.2) to $P_2(y|x) = \alpha/(\alpha + \beta)$ and $P_1(y|x) = 1/2$, the result follows. ■

Corollary 24.1 $P(y|x)$ is not s -recoverable in Figure 24.1(a).

The corollary follows immediately noting that lack of s -recoverability with a subgraph (Figure 24.1(d)) precludes s -recoverability with the graph itself since the extra edge can be inactive in a compatible parametrization (Pearl 1988) (the converse is obviously not true). Lemma 24.1 is significant because Figure 24.1(d) can represent a study design that is typically used in empirical fields known as case-control studies. The result is also theoretically instructive since Figure 24.1(d) represents the smallest graph structure that is not s -recoverable, and its proof will set the tone for more general and arbitrary structures that we will be interested in (see Theorem 24.1).

Furthermore, consider the graph in Figure 24.1(e) in which the independence $(S \perp\!\!\!\perp Y|X)$ holds, so we can also recover Q from selection $(P(y|x, S=1) = P(y|x))$. However, $(S \perp\!\!\!\perp Y|X)$ does not hold in Figure 24.1(f) – there is an open path passing through X 's ancestor W (i.e., $S \leftarrow Z \rightarrow X \leftarrow W \rightarrow Y$) – and the natural question that arises is whether Q is recoverable in this case. It does not look obvious whether the absence of an independence precludes s -recoverability since there are other possible operators in probability theory that could be used leading to the s -recoverability of Q . To illustrate this point, note that it is not the case in causal inference that the inapplicability of the backdoor criterion (Pearl 2000, Chapter 3), which is also an independence constraint, implies the impossibility of recovering certain effects.

Remarkably, the next result states that the lack of this independence indeed precludes s -recoverability, i.e., the probe of one separation test in the graph is sufficient to evaluate whether a distribution is or is not s -recoverable.

Theorem 24.1 *The distribution $P(y|\mathbf{x})$ is s -recoverable from G_s if and only if $(S \perp\!\!\!\perp Y|\mathbf{X})$.*⁸

In words, Theorem 24.1 provides a powerful test for s -recoverability without external data, which means that when it disavows s -recoverability, there exists no procedure that would be capable of recovering the distribution from selection bias (without adding assumptions). Its sufficiency part is immediate, but the proof of necessity is somewhat involved since we need to show that for *all* graphical structures in which the given d -separation test fails, each of these structures does not allow for s -recoverability (i.e., a counter-example can always be produced showing agreement on $P(\mathbf{v}|S=1)$ and disagreement on $P(y|\mathbf{x})$).

The next corollary provides a test for s -recoverability of broader joint distributions (including Y alone):

8. Please refer to the Appendix 2 in the full report for the proofs (Bareinboim, Tian, and Pearl 2014).

Corollary 24.2 Let $\mathbf{Z} = \text{An}(S) \setminus \text{An}(Y)$ including S , and $\mathbf{A} = \text{Pa}(\mathbf{Z}) \cap (\text{An}(Y) \setminus \{Y\})$. $P(Y, \text{An}(Y) \setminus (\mathbf{A} \setminus \{Y\}) | \mathbf{A})$ is s -recoverable if and only if Y is not an ancestor of S .

This result can be embedded as a step reduction in an algorithm to s -recover a collection of distributions in the form of the corollary. We show such algorithm in (Bareinboim, Tian, and Pearl 2014).⁹ The main idea is to traverse the graph in a certain order s -recovering all joint distributions with the form given in the corollary (updating S along the way). If the algorithm exits with failure, it means that the distributions of its predecessors are not s -recoverable.

24.3 Recoverability with External Data

A natural question that arises is whether additional measurements in the population level over certain variables can help recovering a given distribution. For example, $P(\text{age})$ can be estimated from census data which is not under selection bias.

To illustrate how this problem may arise in practice, consider Figure 24.2 and assume that our goal is to s -recover $Q = P(y|x)$. It follows immediately from Theorem 24.1 that Q cannot be s -recovered without additional assumptions. Note, however, that the parents of the selection node $\text{Pa}_S = \{W_1, W_2\}$ separates S from all other nodes in the graph, which indicates that it would be sufficient for recoverability to measure $\mathbf{T} = \{W_1, W_2\} \cup \{X\}$ from external sources. To witness, note that after conditioning Q on W_1 and W_2 , we obtain:

$$\begin{aligned} P(y|x) &= \sum_{w_1, w_2} P(y|x, w_1, w_2)P(w_1, w_2|x) \\ &= \sum_{w_1, w_2} P(y|x, w_1, w_2, S=1)P(w_1, w_2|x), \end{aligned} \quad (24.3)$$

where the last equality follows from $(Y \perp\!\!\!\perp S | X, W_1, W_2)$. That is, Q can be s -recovered and is a combination of two different types of data; the first factor comes from biased data under selection, and the second factor is available from external data collected over the whole population.

Our goal is to understand the interplay between measurements taken over two types of variables, $\mathbf{M}, \mathbf{T} \subseteq V$, where \mathbf{M} are variables collected under selection bias, $P(\mathbf{M}|S=1)$, and \mathbf{T} are variables collected in the population-level, $P(\mathbf{T})$. In other words, we want to understand when (and how) can this new piece of evidence $P(\mathbf{T})$ together with the data under selection ($P(\mathbf{M}|S=1)$) help in extending the

9. This listing is useful when one needs to examine properties of the collection of distributions, analogously to the list of all backdoor admissible sets by (Textor and Liskiewicz 2011).

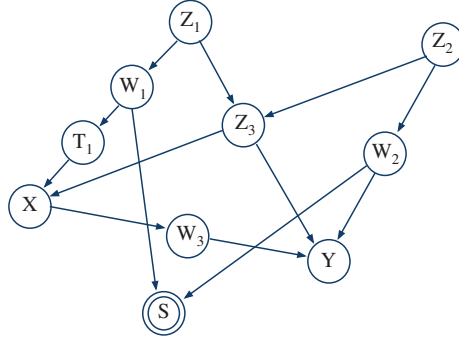


Figure 24.2 Causal model in which $Q = P(y|x)$ is not recoverable without external data (Theorem 24.1), but it is recoverable if measurements on the set $\mathbf{Pa}_s = \{W_1, W_2\}$ are taken (Theorem 24.2). Alternatively, even if not all parents of S are measured, any set including $\{W_2, Z_3\}$ would yield recoverability of Q .

treatment of the previous section for recovering the true underlying distribution $Q = P(y|x)$.¹⁰

Formally, we need to redefine s -recoverability for accommodating the availability of data from external sources.

Definition 24.2 s-Recoverability

Given a causal graph G_s augmented with a node S , the distribution $Q = P(y|x)$ is said to be s -recoverable from selection bias in G_s with external information over $\mathbf{T} \subseteq \mathbf{V}$ and selection biased data over $\mathbf{M} \subseteq \mathbf{V}$ (for short, s -recoverable) if the assumptions embedded in the causal model render Q expressible in terms of $P(\mathbf{m}|S = 1)$ and $P(\mathbf{t})$, both positive. Formally, for every two probability distributions P_1 and P_2 compatible with G_s , if they agree on the available distributions, $P_1(\mathbf{m}|S = 1) = P_2(\mathbf{m}|S = 1) > 0$, $P_1(\mathbf{t}) = P_2(\mathbf{t}) > 0$, they must agree on the query distribution, $P_1(y|x) = P_2(y|x)$.

The observation leading to Equation (24.3) provides a simple condition for s -recoverability when we can choose the variables to be collected. Let \mathbf{Pa}_s be the parent set of S . If measurements on the set $\mathbf{T} = \mathbf{Pa}_s \cup \{X\}$ can be taken without selection, we can write $P(y|x) = \sum_{\mathbf{pa}_s} P(y|x, \mathbf{pa}_s, S = 1)P(\mathbf{pa}_s|x)$, since S is separated from all nodes in the graph given its parent set. This implies s -recoverability where we have a mixture in which the first factor is obtainable from the biased data and the second from external sources.

This solution is predicated on the assumption that \mathbf{Pa}_s can be measured in the overall population, which can be a strong requirement, and begs a generalization

10. This problem subsumes the one given in the previous section since when $\mathbf{T} = \emptyset$, the two problems coincide. We separate them since they come in different shades in the literature and also just after solving the version without external data we can aim to solve its more general version; we discuss more about this later on.

to when part of \mathbf{Pa}_s is not measured. For instance, what if in Figure 24.2 W_1 cannot be measured? Would other measurements over a different set of variables also entail s -recoverability?

This can be expressed as a requirement that subsets of \mathbf{T} and \mathbf{M} can be found satisfying the following criterion:

Theorem 24.2 *If there is a set \mathbf{C} that is measured in the biased study with $\{\mathbf{X}, Y\}$ and in the population level with \mathbf{X} such that $(Y \perp\!\!\!\perp S \mid \{\mathbf{C}, \mathbf{X}\})$, then $P(y \mid \mathbf{x})$ is s -recoverable as*

$$P(y \mid \mathbf{x}) = \sum_{\mathbf{c}} P(y \mid \mathbf{x}, \mathbf{c}, S = 1)P(\mathbf{c} \mid \mathbf{x}). \quad (24.4)$$

In the example in Figure 24.2, it is trivial to confirm that any (pre-treatment) set \mathbf{C} containing W_2 and Z_3 would satisfy the conditions of the theorem. In particular, $\{W_2, Z_3\}$ is such a set, and it allows us to s -recover Q without measuring W_1 ($W_1 \in \mathbf{Pa}_s$) through Equation (24.4). Note, however, that the set $\mathbf{C} = \{W_2, Z_1, Z_2\}$ is not sufficient for s -recoverability. It fails to satisfy the separability condition of the theorem since conditioning on $\{X, W_2, Z_1, Z_2\}$ leaves an unblocked path between S and Y (i.e., $S \leftarrow W_1 \rightarrow T_1 \rightarrow X \leftarrow Z_3 \rightarrow Y$).

It can be computationally difficult to find a set satisfying the conditions of the theorem since this could imply a search over a potentially exponential number of subsets. Remarkably, the next result shows that the existence of such a set can be determined by a single d -separation test.

Theorem 24.3 *There exists some set $\mathbf{C} \subseteq \mathbf{T} \cap \mathbf{M}$ such that $Y \perp\!\!\!\perp S \mid \{\mathbf{C}, \mathbf{X}\}$ if and only if the set $(\mathbf{C}' \cup \mathbf{X})$ d -separates S from Y where $\mathbf{C}' = [(\mathbf{T} \cap \mathbf{M}) \cap \text{An}(Y \cup S \cup \mathbf{X})] \setminus (Y \cup S \cup \mathbf{X})$.*

In practice, we can restrict ourselves to minimal separators, that is, looking only for minimal set $\mathbf{C} \subseteq \mathbf{T} \cap \mathbf{M}$ such that $(Y \perp\!\!\!\perp S \mid \{\mathbf{C}, \mathbf{X}\})$. The algorithm for finding minimal separators has been given in (Acid and de Campos 1996; Tian, Paz, and Pearl 1998).

Despite the computational advantages given by Theorem 24.3, Theorem 24.2 still requires the existence of a separator \mathbf{C} measured in both the biased study (\mathbf{M}) and in the overall population (\mathbf{T}), and it is natural to ask whether this condition can be relaxed. Assume that all we have is a separator $\mathbf{C} \subseteq \mathbf{M}$, but \mathbf{C} (or some of its elements) is not measured in population \mathbf{T} , and therefore $P(\mathbf{c} \mid \mathbf{x})$ in Equation (24.4) still needs to be s -recovered. We could s -recover $P(\mathbf{c} \mid \mathbf{x})$ in the spirit of Theorem 24.2 as

$$P(\mathbf{c} \mid \mathbf{x}) = \sum_{\mathbf{c}_1} P(\mathbf{c} \mid \mathbf{x}, \mathbf{c}_1, S = 1)P(\mathbf{c}_1 \mid \mathbf{x}), \quad (24.5)$$

if there exists a set $C_1 \subseteq \mathbf{M} \cap \mathbf{T}$ such that $(S \perp\!\!\!\perp C_1 | \mathbf{X}, C_1)$. Now if this fails in that we can only find a separator $C_1 \subseteq \mathbf{M}$ not measured in \mathbf{T} , we can then attempt to recover $P(\mathbf{c}_1 | \mathbf{x})$ in the spirit of Theorem 24.2 by looking for another separator C_2 , and so on. At this point, it appears that Theorem 24.2 can be extended.

We further extend this idea by considering other possible probabilistic manipulations and embed them in a recursive procedure. For $\mathbf{W}, \mathbf{Z} \subseteq \mathbf{M}$, consider the problem of recovering $P(\mathbf{w} | \mathbf{z})$ from $P(\mathbf{t})$ and $P(\mathbf{m} | S = 1)$, and define procedure $RC(\mathbf{w}, \mathbf{z})$ as follows:

1. If $\mathbf{W} \cup \mathbf{Z} \subseteq \mathbf{T}$, then $P(\mathbf{w} | \mathbf{z})$ is s -recoverable.
2. If $(S \perp\!\!\!\perp \mathbf{W} | \mathbf{Z})$, then $P(\mathbf{w} | \mathbf{z})$ is s -recoverable as $P(\mathbf{w} | \mathbf{z}) = P(\mathbf{w} | \mathbf{z}, S = 1)$.
3. For minimal $\mathbf{C} \subseteq \mathbf{M}$ such that $(S \perp\!\!\!\perp \mathbf{W} | (\mathbf{Z} \cup \mathbf{C}))$, $P(\mathbf{w} | \mathbf{z}) = \sum_{\mathbf{c}} P(\mathbf{w} | \mathbf{z}, \mathbf{c}, S = 1)P(\mathbf{c} | \mathbf{z})$. If $\mathbf{C} \cup \mathbf{Z} \subseteq \mathbf{T}$, then $P(\mathbf{w} | \mathbf{z})$ is s -recoverable. Otherwise, call $RC(\mathbf{c}, \mathbf{z})$.
4. For some $\mathbf{W}' \subset \mathbf{W}$, $P(\mathbf{w} | \mathbf{z}) = P(\mathbf{w}' | \mathbf{w} \setminus \mathbf{w}', \mathbf{z})P(\mathbf{w} \setminus \mathbf{w}' | \mathbf{z})$. Call $RC(\mathbf{w}', \{\mathbf{w} \setminus \mathbf{w}'\} \cup \mathbf{z})$ and $RC(\mathbf{w} \setminus \mathbf{w}', \mathbf{z})$.
5. Exit with FAIL (to s -recover $P(\mathbf{w} | \mathbf{z})$) if for a singleton \mathbf{W} , none of the above operations are applicable.

Now, we define recoverability based on this procedure:

Definition 24.3 We say that $P(\mathbf{w} | \mathbf{z})$ is C -recoverable if and only if it is recovered by the procedure $RC(\mathbf{w}, \mathbf{z})$.

Remarkably, the manipulations considered in $RC()$ are not actually more powerful than Theorem 24.2, as shown next.

Theorem 24.4 For $\mathbf{X} \subseteq \mathbf{T}$, $Y \notin \mathbf{T}$, $Q = P(\mathbf{y} | \mathbf{x})$ is C -recoverable if and only if it is recoverable by Theorem 24.2, that is, if and only if there exists a set $\mathbf{C} \subseteq \mathbf{T} \cap \mathbf{M}$ such that $(Y \perp\!\!\!\perp S | \{\mathbf{C}, \mathbf{X}\})$ (where \mathbf{C} could be empty). If s -recoverable, $P(\mathbf{y} | \mathbf{x})$ is given by $P(\mathbf{y} | \mathbf{x}) = \sum_{\mathbf{c}} P(\mathbf{y} | \mathbf{x}, \mathbf{c}, S = 1)P(\mathbf{c} | \mathbf{x})$.

This result suggests that the constraint between measurement sets cannot be relaxed through ordinary decomposition and Theorem 24.2 captures the bulk of s -recoverable relations. (See proof in (Bareinboim, Tian, and Pearl 2014).) Importantly, this does not constitute a proof of necessity of Theorem 24.2.

Now we turn our attention to some special cases that appear in practice. Note that, so far, we assumed X being measured in the overall population, but in some scenarios Y 's prevalence might be available instead. So, assume $Y \in \mathbf{T}$ but some variables in \mathbf{X} are not measured in the population-level. Let $\mathbf{X}^0 = \mathbf{X} \cap \mathbf{T}$ and

$\mathbf{X}^m = \mathbf{X} \setminus \mathbf{X}^0$, we have

$$P(y | \mathbf{x}) = \frac{P(\mathbf{x}^m | y, \mathbf{x}^0)p(y | \mathbf{x}^0)}{\sum_y P(\mathbf{x}^m | y, \mathbf{x}^0)p(y | \mathbf{x}^0)} \quad (24.6)$$

Therefore, $P(y | \mathbf{x})$ is recoverable if $P(\mathbf{x}^m | y, \mathbf{x}^0)$ is recoverable. We could use the previous results to recover $P(\mathbf{x}^m | y, \mathbf{x}^0)$. In particular, Theorems 24.2 and 24.3 lead to:

Corollary 24.3 $P(y | \mathbf{x})$ is recoverable if there exists a set $\mathbf{C} \subseteq \mathbf{T} \cap \mathbf{M}$ (\mathbf{C} could be empty) such that $(\mathbf{X}^m \perp\!\!\!\perp S | \{\mathbf{C} \cup \mathbf{Y} \cup \mathbf{X}^0\})$. If recoverable, $P(y | \mathbf{x})$ is given by Equation (24.6) where

$$P(\mathbf{x}^m | y, \mathbf{x}^0) = \sum_c P(\mathbf{x}^m | y, \mathbf{x}^0, \mathbf{c}, S = 1)P(c | y, \mathbf{x}^0) \quad (24.7)$$

Corollary 24.4 $P(y | \mathbf{x})$ is recoverable via Corollary 24.3 if and only if the set $(\mathbf{C}' \cup \mathbf{Y} \cup \mathbf{X}^0)$ d -separates S from \mathbf{X}^m where $\mathbf{C}' = [(\mathbf{T} \cap \mathbf{M}) \cap \text{An}(\mathbf{Y} \cup \mathbf{S} \cup \mathbf{X})] \setminus (\mathbf{Y} \cup \mathbf{S} \cup \mathbf{X})$.

For example, in Figure 24.2, assuming $\mathbf{M} = \{X, Y, W_1, W_3, Z_3\}$ and $\mathbf{T} = \{Y, W_1, W_3, Z_3\}$, we have $S \perp\!\!\!\perp X | \{Y, W_1, W_3, Z_3\}$, therefore we can s -recover

$$P(x | y) = \sum_{w_1, w_3, z_3} P(x | y, w_1, w_3, z_3, S = 1)P(w_1, w_3, z_3 | y), \quad (24.8)$$

as well as $P(y | \mathbf{x})$ by substituting back Equation (24.8) in Equation (24.6).

Furthermore, it is worth examining when no data is gathered over \mathbf{X} or \mathbf{Y} in the population level. In this case, $P(y | \mathbf{x})$ may be recoverable through $P(\mathbf{x}, y)$, as shown in the sequel.

Corollary 24.5 $P(y | \mathbf{x})$ is recoverable if there exists a set $\mathbf{C} \subseteq \mathbf{T} \cap \mathbf{M}$ such that $(\{Y\} \cup \mathbf{X} \perp\!\!\!\perp S | \mathbf{C})$. If recoverable, $P(y, \mathbf{x})$ is given by $P(y, \mathbf{x}) = \sum_c P(y, \mathbf{x} | \mathbf{c}, S = 1)P(\mathbf{c})$.

For instance, $P(x, y)$ is s -recoverable in Figure 24.2 if $\mathbf{T} \cap \mathbf{M}$ contains $\{W_2, T_1, Z_3\}$ or $\{W_2, T_1, Z_1\}$ (without $\{X, Y\}$).

24.4 Recoverability of Causal Effects

We now turn our attention to the problem of estimating causal effects from selection biased data.¹¹

Our goal is to recover the effect of X on Y , $P(y | do(x))$ given the structure of G_s . Consider the graph G_s in Figure 24.3(a), in which X and Y are not confounded, hence, $P(y | do(x)) = P(y | x)$ and, based on Theorem 24.1, we conclude that $P(y | do(x))$

11. We assume the graph G_s represents a causal model, as defined in (Pearl 2000; Spirtes, Glymour, and Scheines 2000).

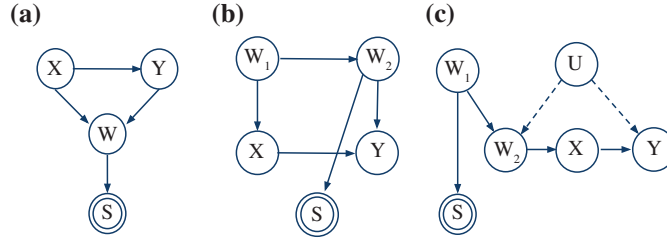


Figure 24.3 (a) Causal diagram in which $(S \perp\!\!\!\perp Y \mid \{X, W\})$ but $P(y \mid do(x))$ is not s -backdoor admissible. (b) $P(y \mid do(x))$ is s -recoverable through $\mathbf{T} = \{W_2\}$ but not $\{W_1\}$. (c) $\{W_2\}$ does not satisfy the s -backdoor criterion but $P(y \mid do(x))$ is still recoverable.

is not recoverable in G_s . Figure 24.3(b) and 24.3(c), on the other hand, contains covariates W_1 and W_2 that may satisfy conditions similar to those in Theorem 24.1 that would render $P(y \mid do(x))$ recoverable. These conditions, however, need to be strengthened significantly, to account for possible confounding between X and Y which, even in the absence of selection bias, might require adjustment for admissible covariates, namely, covariates that satisfy the backdoor condition (Pearl 1993). For example, $\{W_2\}$ satisfies the backdoor condition in both Figure 24.3(b) and (c), while $\{W_1\}$ satisfies this condition in (b) but not in (c).

Definition 24.4 below extends the backdoor condition to selection bias problems by identifying a set of covariates \mathbf{Z} that accomplishes two functions. Conditions (i) and (ii) assure us that \mathbf{Z} is backdoor admissible (Pearl and Paz 2013),¹² while conditions (iii) and (iv) act to separate S from Y , so as to permit recoverability from selection bias.

Definition 24.4 Selection-backdoor criterion

Let a set \mathbf{Z} of variables be partitioned into $\mathbf{Z}^+ \cup \mathbf{Z}^-$ such that \mathbf{Z}^+ contains all non-descendants of X and \mathbf{Z}^- the descendants of X . \mathbf{Z} is said to satisfy the selection backdoor criterion (s -backdoor, for short) relative to an ordered pair of variables (X, Y) and an ordered pair of sets (\mathbf{M}, \mathbf{T}) in a graph G_s if \mathbf{Z}^+ and \mathbf{Z}^- satisfy the following conditions:

- (i) \mathbf{Z}^+ blocks all back door paths from X to Y ;
- (ii) X and \mathbf{Z}^+ block all paths between \mathbf{Z}^- and Y , namely, $(\mathbf{Z}^- \perp\!\!\!\perp Y \mid X, \mathbf{Z}^+)$;
- (iii) X and \mathbf{Z} block all paths between S and Y , namely, $(Y \perp\!\!\!\perp S \mid X, \mathbf{Z})$;
- (iv) $\mathbf{Z} \cup \{X, Y\} \subseteq \mathbf{M}$, and $\mathbf{Z} \subseteq \mathbf{T}$.

12. These two conditions extend the usual backdoor criterion (Pearl 1993) to allow descendants of X to be part of \mathbf{Z} .

Consider Figure 24.3(a) where $\mathbf{Z}^- = \{W\}$, $\mathbf{Z}^+ = \{\}$ and \mathbf{Z}^- is *not* separated from Y given $\{\mathbf{X}\} \cup \mathbf{Z}^+$ in G_s , which means that condition (ii) of the s -backdoor is violated. So, despite the fact that the relationship between X and Y is unconfounded and ($Y \perp\!\!\!\perp S \mid \{W, X\}$), it is improper to adjust for $\{W\}$ when computing the target effect.

For the admissible cases, we are ready to state a sufficient condition that guarantees proper identifiability and recoverability of causal effects under selection bias:

Theorem 24.5 Selection-backdoor adjustment

If a set \mathbf{Z} satisfies the s -backdoor criterion relative to the pairs (X, Y) and (\mathbf{M}, \mathbf{T}) (as given in Definition 24.2), then the effect of X on Y is identifiable and s -recoverable and is given by the formula

$$P(y \mid do(x)) = \sum_{\mathbf{z}} P(y \mid x, \mathbf{z}, S = 1)P(\mathbf{z}) \quad (24.9)$$

Interestingly, X does not need to be measured in the overall population when the s -backdoor adjustment is applicable, which contrasts with the expression given in Theorem 24.2 where both X and \mathbf{Z} (equivalently \mathbf{C}) are needed.

Consider Figure 24.3(b) and assume our goal is to establish $Q = P(y \mid do(x))$ when external data over $\{W_2\}$ is available in both studies. Then, $\mathbf{Z} = \{W_2\}$ is s -backdoor admissible and the s -backdoor adjustment is applicable in this case. However, if $\mathbf{T} = \{W_1\}$, $\mathbf{Z} = \{W_1\}$ is backdoor admissible, but it is *not* s -backdoor admissible since condition (iii) is violated (i.e., $(S \perp\!\!\!\perp Y \mid \{W_1, X\})$ does not hold in G_s). This is interesting since the two sets $\{W_1\}$ and $\{W_2\}$ are c -equivalent (Pearl and Paz 2013), having the same potential for bias reduction in the general population. To understand why c -equivalence is not sufficient for s -recoverability, note that despite the equivalence for adjustment, $\sum_{w_1} P(y \mid x, w_1)P(w_1) = \sum_{w_2} P(y \mid x, w_2)P(w_2)$, the r.h.s. is obtainable from the data, while the l.h.s. is not.

Now we want to recover $Q = P(y \mid do(x))$ in Figure 24.3(c) (U is a latent variable) with $\mathbf{T} = \{W_2\}$. Condition (iii) of the s -backdoor fails since $(S \perp\!\!\!\perp Y \mid \{X, W_2\})$ does not hold. Alternatively, if we discard W_2 and consider the null set for adjustment ($\mathbf{Z} = \{\}$), condition (i) fails since there is an open backdoor path from X to Y ($X \leftarrow W_2 \leftarrow U \rightarrow Y$). Despite the inapplicability of the s -backdoor, $P(y \mid do(x))$ is still s -recoverable since, using do -calculus, we can show that $Q = P(y \mid do(x), S = 1)$, which reduces to $\sum_{w_2} P(y \mid x, w_2, S = 1)P(w_2 \mid S = 1)$, both factors s -recoverable without the need for external information.

The reliance on the do -calculus in recovering causal effects is expected since even when selection bias is absent, there exist identifiability results beyond the backdoor. Still, this criterion, which is generalized by the s -backdoor criterion,

is arguably the most used method for identifiability of causal effects currently available in the literature.

24.5 Conclusions

We provide conditions for recoverability from selection bias in statistical and causal inferences applicable for arbitrary structures in non-parametric settings. Theorem 24.1 provides a complete characterization of recoverability when no external information is available. Theorem 24.2 provides a sufficient condition for recoverability based on external information; it is optimized by Theorem 24.3 and strengthened by Theorem 24.4. Verifying these conditions takes polynomial time and could be used to decide what measurements are needed for recoverability. Theorem 24.5 further gives a graphical condition for recovering causal effects, which generalizes the backdoor adjustment. Since selection bias is a common problem across many disciplines, the methods developed in this paper should help to understand, formalize, and alleviate this problem in a broad range of data-intensive applications. This paper complements another aspect of the generalization problem in which causal effects are transported among differing environments (Bareinboim and Pearl 2013a; 2013b).

Acknowledgments

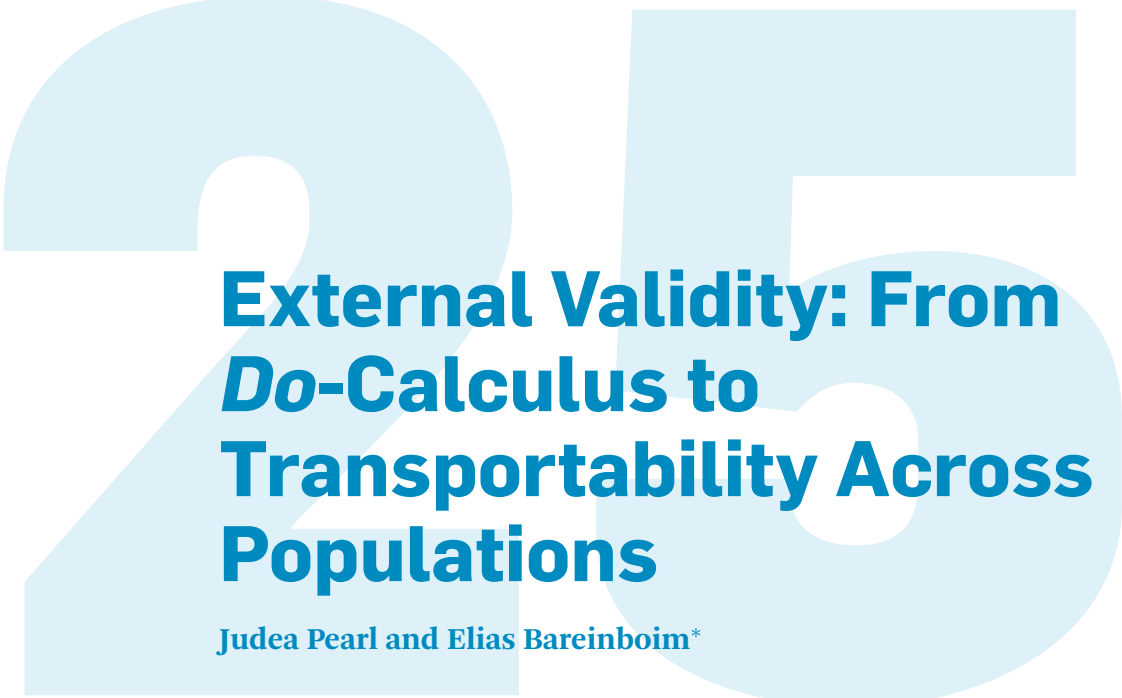
The authors would like to thank the reviewers for their comments that help improve the manuscript. This research was supported in parts by grants from NSF #IIS-1249822 and #IIS-1302448, and ONR #N00014-13-1-0153 and #N00014-10-1-0933.

References

- Acid, S., and de Campos, L. 1996. An algorithm for finding minimum d-separating sets in belief networks. In *Proceedings of the 12th Annual Conference on Uncertainty in Artificial Intelligence*, 3–10. San Francisco, CA: Morgan Kaufmann.
- Angrist, J. D. 1997. Conditional independence in sample selection models. *Economics Letters* 54(2):103–112.
- Bareinboim, E., and Pearl, J. 2012. Controlling selection bias in causal inference. In Girolami, M., and Lawrence, N., eds., *Proceedings of The Fifteenth International Conference on Artificial Intelligence and Statistics (AISTATS 2012)*, 100–108. JMLR (22).
- Bareinboim, E., and Pearl, J. 2013a. Meta-transportability of causal effects: A formal approach. In *Proceedings of The Sixteenth International Conference on Artificial Intelligence and Statistics (AISTATS 2013)*, 135–143. JMLR (31).
- Bareinboim, E., and Pearl, J. 2013b. Causal transportability with limited experiments. In desJardins, M., and Littman, M. L., eds., *Proceedings of The Twenty-Seventh Conference on Artificial Intelligence (AAAI 2013)*, 95–101.

- Bareinboim, E.; Tian, J.; and Pearl, J. 2014. Recovering from selection bias in causal and statistical inference. Technical Report R-425, Cognitive Systems Laboratory, Department of Computer Science, UCLA. Also in Carla E. Brodley and Peter Stone (Eds.) *Proceedings of the Twenty-eighth AAAI Conference on Artificial Intelligence*, Palo Alto, CA: AAAI Press, 2410–2416, 2014, “Best Paper Award.”
- Cooper, G. 1995. Causal discovery from data in the presence of selection bias. *Artificial Intelligence and Statistics* 140–150.
- Cornfield, J. 1951. A method of estimating comparative rates from clinical data; applications to cancer of the lung, breast, and cervix. *Journal of the National Cancer Institute* 11:1269–1275.
- Cortes, C.; Mohri, M.; Riley, M.; and Rostamizadeh, A. 2008. Sample selection bias correction theory. In *Proceedings of the 19th International Conference on Algorithmic Learning Theory*, ALT '08, 38–53. Berlin, Heidelberg: Springer-Verlag.
- Didelez, V.; Kreiner, S.; and Keiding, N. 2010. Graphical models for inference under outcome-dependent sampling. *Statistical Science* 25(3):368–387.
- Elkan, C. 2001. The foundations of cost-sensitive learning. In *Proceedings of the 17th International Joint Conference on Artificial Intelligence - Volume 2*, IJCAI'01, 973–978. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.
- Geng, Z. 1992. Collapsibility of relative risk in contingency tables with a response variable. *Journal Royal Statistical Society* 54(2):585–593.
- Glymour, M., and Greenland, S. 2008. Causal diagrams. In Rothman, K.; Greenland, S.; and Lash, T., eds., *Modern Epidemiology*. Philadelphia, PA: Lippincott Williams & Wilkins, 3rd edition. 183–209.
- Greenland, S., and Pearl, J. 2011. Adjustments and their consequences – collapsibility analysis using graphical models. *International Statistical Review* 79(3):401–426.
- Heckman, J. 1979. Sample selection bias as a specification error. *Econometrica* 47:153–161.
- Hein, M. 2009. Binary classification under sample selection bias. In Candela, J.; Sugiyama, M.; Schwaighofer, A.; and Lawrence, N., eds., *Dataset Shift in Machine Learning*. Cambridge, MA: MIT Press. 41–64.
- Jewell, N. P. 1991. Some surprising results about covariate adjustment in logistic regression models. *International Statistical Review* 59(2):227–240.
- Koller, D., and Friedman, N. 2009. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press.
- Kuroki, M., and Cai, Z. 2006. On recovering a population covariance matrix in the presence of selection bias. *Biometrika* 93(3):601–611.
- Little, R. J. A., and Rubin, D. B. 1986. *Statistical Analysis with Missing Data*. New York, NY, USA: John Wiley & Sons, Inc.
- Mefford, J., and Witte, J. S. 2012. The covariate’s dilemma. *PLoS Genet* 8(11):e1003096.
- Pearl, J., and Paz, A. 2013. Confounding equivalence in causal equivalence. In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence (UAI*

- 2010), 433–441. Corvallis, OR: AUAI. Also: Technical Report R-343w, Cognitive Systems Laboratory, Department of Computer Science, UCLA.
- Pearl, J. 1988. *Probabilistic Reasoning in Intelligent Systems*. San Mateo, CA: Morgan Kaufmann.
- Pearl, J. 1993. Aspects of graphical models connected with causality. In *Proceedings of the 49th Session of the International Statistical Institute*, 391–401.
- Pearl, J. 1995. Causal diagrams for empirical research. *Biometrika* 82(4):669–710.
- Pearl, J. 2000. *Causality: Models, Reasoning, and Inference*. New York: Cambridge University Press. Second ed., 2009.
- Pearl, J. 2013. Linear models: A useful “microscope” for causal analysis. *Journal of Causal Inference* 1:155–170.
- Pirinen, M.; Donnelly, P.; and Spencer, C. 2012. Including known covariates can reduce power to detect genetic effects in case-control studies. *Nature Genetics* 44:848–851.
- Robins, J. 2001. Data, design, and background knowledge in etiologic inference. *Epidemiology* 12(3):313–320.
- Smith, A. T., and Elkan, C. 2007. Making generative classifiers robust to selection bias. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '07, 657–666. New York, NY, USA: ACM.
- Spirtes, P.; Glymour, C.; and Scheines, R. 2000. *Causation, Prediction, and Search*. Cambridge, MA: MIT Press, 2nd edition.
- Storkey, A. 2009. When training and test sets are different: characterising learning transfer. In Candela, J.; Sugiyama, M.; Schwaighofer, A.; and Lawrence, N., eds., *Dataset Shift in Machine Learning*. Cambridge, MA: MIT Press. 3–28.
- Textor, J., and Liskiewicz, M. 2011. Adjustment criteria in causal diagrams: An algorithmic perspective. In Pfeffer, A., and Cozman, F., eds., *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence (UAI 2011)*, 681–688. AUAI Press.
- Tian, J.; Paz, A.; and Pearl, J. 1998. Finding minimal separating sets. Technical Report R-254, University of California, Los Angeles, CA.
- Whittemore, A. 1978. Collapsibility of multidimensional contingency tables. *Journal of the Royal Statistical Society, Series B* 40(3):328–340.
- Zadrozny, B. 2004. Learning and evaluating classifiers under sample selection bias. In *Proceedings of the Twenty-first International Conference on Machine Learning, ICML '04*, 114–. New York, NY, USA: ACM.
- Zhang, J. 2008. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence* 172:1873–1896.



External Validity: From *Do*-Calculus to Transportability Across Populations

Judea Pearl and Elias Bareinboim*

Abstract

The generalizability of empirical findings to new environments, settings or populations, often called “external validity,” is essential in most scientific explorations. This paper treats a particular problem of generalizability, called “transportability,” defined as a license to transfer causal effects learned in experimental studies to a new population, in which only observational studies can be conducted. We introduce a formal representation called “selection diagrams” for expressing knowledge about differences and commonalities between populations of interest and, using this representation, we reduce questions of transportability to symbolic derivations in the *do*-calculus. This reduction yields graph-based procedures for deciding, prior to observing any data, whether causal effects in the target population can be inferred from experimental findings in the study population. When the answer is affirmative, the procedures identify what experimental and observational findings need be obtained from the two populations, and how they can be combined to ensure bias-free transport.

Key words and phrases

Experimental design, generalizability, causal effects, external validity.

*University of California, Los Angeles

Originally published in *Statistical Science* 2014, Vol. 29, No. 4, 579–595

© Institute of Mathematical Statistics, 2014. Reprinted with the permission of the Institute of Mathematical Statistics.

Original DOI: [10.1214/14-STS486](https://doi.org/10.1214/14-STS486)

25.1 Introduction: Threats vs. Assumptions

Science is about generalization, and generalization requires that conclusions obtained in the laboratory be transported and applied elsewhere, in an environment that differs in many aspects from that of the laboratory.

Clearly, if the target environment is arbitrary, or drastically different from the study environment, nothing can be transferred and scientific progress will come to a standstill. However, the fact that most studies are conducted with the intention of applying the results elsewhere means that we usually deem the target environment sufficiently similar to the study environment to justify the transport of experimental results or their ramifications.

Remarkably, the conditions that permit such transport have not received systematic formal treatment. In statistical practice, problems related to combining and generalizing from diverse studies are handled by methods of meta analysis (Glass, 1976; Hedges and Olkin, 1985; Owen, 2009), or hierarchical models (Gelman and Hill, 2007), in which results of diverse studies are pooled together by standard statistical procedures (e.g., inverse-variance reweighting in meta-analysis, partial pooling in hierarchical modelling) and rarely make explicit distinction between experimental and observational regimes; performance is evaluated primarily by simulation.

To supplement these methodologies, our paper provides theoretical guidance in the form of limits on what can be achieved in practice, what problems are likely to be encountered when populations differ significantly from each other, what population differences can be circumvented by clever design and what differences constitute theoretical impediments, prohibiting generalization by any means whatsoever.

On the theoretical front, the standard literature on this topic, falling under rubrics such as “external validity” (Campbell and Stanley, 1963, Manski, 2007), “heterogeneity” (Höfler, Gloster and Hoyer, 2010), “quasi-experiments” (Shadish, Cook and Campbell, 2002, Chapter 3; Adelman, 1991),¹ consists primarily of “threats,” namely, explanations of what may go wrong when we try to transport results from one study to another while ignoring their differences. Rarely do we find an analysis of “licensing assumptions,” namely, formal conditions under which the

1. Manski (2007) defines “external validity” as follows: “An experiment is said to have “external validity” if the distribution of outcomes realized by a treatment group is the same as the distribution of outcome that would be realized in an actual program.” Campbell and Stanley (1963), page 5, take a slightly broader view: ““External validity” asks the question of generalizability: to what populations, settings, treatment variables, and measurement variables can this effect be generalized?”

transport of results across differing environments or populations is licensed from first principles.²

The reasons for this asymmetry are several. First, threats are safer to cite than assumptions. He who cites “threats” appears prudent, cautious and thoughtful, whereas he who seeks licensing assumptions risks suspicions of attempting to endorse those assumptions.

Second, assumptions are self-destructive in their honesty. The more explicit the assumption, the more criticism it invites, for it tends to trigger a richer space of alternative scenarios in which the assumption may fail. Researchers prefer, therefore, to declare threats in public and make assumptions in private.

Third, whereas threats can be communicated in plain English, supported by anecdotal pointers to familiar experiences, assumptions require a formal language within which the notion “environment” (or “population”) is given precise characterization, and differences among environments can be encoded and analyzed.

The advent of causal diagrams (Wright, 1921; Heise, 1975; Davis, 1984; Verma and Pearl, 1988; Spirtes, Glymour and Scheines, 1993; Pearl, 1995) together with models of interventions (Haavelmo, 1943; Strotz and Wold, 1960) and counterfactuals (Neyman, 1923; Rubin, 1974; Robins, 1986; Balke and Pearl, 1995) provides such a language and renders the formalization of transportability possible.

Armed with this language, this paper departs from the tradition of communicating “threats” and embarks instead on the task of formulating “licenses to transport,” namely, assumptions that, if they held true, would permit us to transport results across studies.

In addition, the paper uses the inferential machinery of the *do*-calculus (Pearl, 1995; Koller and Friedman, 2009; Huang and Valtorta, 2006; Shpitser and Pearl, 2006) to derive algorithms for deciding whether transportability is feasible and how experimental and observational findings can be combined to yield unbiased estimates of causal effects in the target population.

The paper is organized as follows. In Section 25.2, we review the foundations of structural equations modelling (SEM), the question of identifiability and the *do*-calculus that emerges from these foundations. (This section can be skipped

2. Hernán and VanderWeele (2011) studied such conditions in the context of compound treatments, where we seek to predict the effect of one version of a treatment from experiments with a different version. Their analysis is a special case of the theory developed in this paper (Petersen, 2011). A related application is reported in Robins, Orellana and Rotnitzky (2008) where a treatment strategy is extrapolated between two biological similar populations under different observational regimes.

by readers familiar with these concepts and tools.) In Section 25.3, we motivate the question of transportability through simple examples, and illustrate how the solution depends on the causal story behind the problem. In Section 25.4, we formally define the notion of transportability and reduce it to a problem of symbolic transformations in *do*-calculus. In Section 25.5, we provide a graphical criterion for deciding transportability and estimating transported causal effects. We conclude in Section 25.6 with brief discussions of related problems of external validity, these include statistical transportability and meta-analysis.

25.2 Preliminaries: The Logical Foundations of Causal Inference

The tools presented in this paper were developed in the context of non-parametric Structural Equations Models (SEM), which is one among several approaches to causal inference and goes back to (Haavelmo, 1943; Strotz and Wold, 1960). Other approaches include, for example, potential-outcomes (Rubin, 1974), Structured Tree Graphs (Robins, 1986), decision analytic (Dawid, 2002), Causal Bayesian Networks (Spirtes, Glymour and Scheines, 2000; Pearl, 2000, Chapter 1; Bareinboim, Brito and Pearl, 2012), and Settable Systems (White and Chalak, 2009). We will first describe the generic features common to all such approaches, and then summarize how these features are represented in SEM.³

25.2.1 Causal Models as Inference Engines

From a logical viewpoint, causal analysis relies on causal assumptions that cannot be deduced from (nonexperimental) data. Thus, every approach to causal inference must provide a systematic way of encoding, testing and combining these assumptions with data. Accordingly, we view causal modeling as an inference engine that takes three inputs and produces three outputs. The inputs are:

- I-1. A set A of qualitative causal *assumptions* which the investigator is prepared to defend on scientific grounds, and a model M_A that encodes these assumptions mathematically. (In SEM, M_A takes the form of a diagram or a set of unspecified functions.) A typical assumption is that no direct effect exists between a pair of variables (known as exclusion restriction), or that an

3. We use the acronym SEM for both parametric and nonparametric representations though, historically, SEM practitioners preferred the former (Bollen and Pearl, 2013). Pearl (2011) has used the term Structural Causal Models (SCM) to eliminate this confusion. While comparisons of the various approaches lie beyond the scope of this paper, we nevertheless propose that their merits be judged by the extent to which each facilitates the functions described below.

omitted factor, represented by an error term, is independent of other such factors observed or unobserved, known as well as unknown.

- I-2. A set Q of *queries* concerning causal or counterfactual relationships among variables of interest. In linear SEM, Q concerned the magnitudes of structural coefficients but, in general, Q may address causal relations directly, for example:

Q_1 : What is the effect of treatment X on outcome Y ?

Q_2 : Is this employer practicing gender discrimination?

In principle, each query $Q_i \in Q$ should be “well defined,” that is, computable from any fully specified model M compatible with A . (See Definition 25.1 for formal characterization of a model, and also Section 25.2.4 for the problem of identification in partially specified models.)

- I-3. A set D of experimental or non-experimental *data*, governed by a joint probability distribution presumably consistent with A .

The outputs are:

- O-1. A set A^* of statements which are the logical implications of A , separate from the data at hand. For example, that X has no effect on Y if we hold Z constant, or that Z is an instrument relative to $\{X, Y\}$.
- O-2. A set C of data-dependent *claims* concerning the magnitudes or likelihoods of the target queries in Q , each contingent on A . C may contain, for example, the estimated mean and variance of a given structural parameter, or the expected effect of a given intervention. Auxiliary to C , a causal model should also yield an estimand $Q_i(P)$ for each query in Q , or a determination that Q_i is not identifiable from P (Definition 25.2).
- O-3. A list T of testable statistical implications of A (which may or may not be part of O-2), and the degree $g(T_i)$, $T_i \in T$, to which the data agrees with each of those implications. A typical implication would be a conditional independence assertion, or an equality constraint between two probabilistic expressions. Testable constraints should be read from the model M_A (see Definition 25.3), and used to confirm or disconfirm the model against the data.

The structure of this inferential exercise is shown schematically in Figure 25.1. For a comprehensive review on methodological issues, see Pearl (2009a, 2012a).

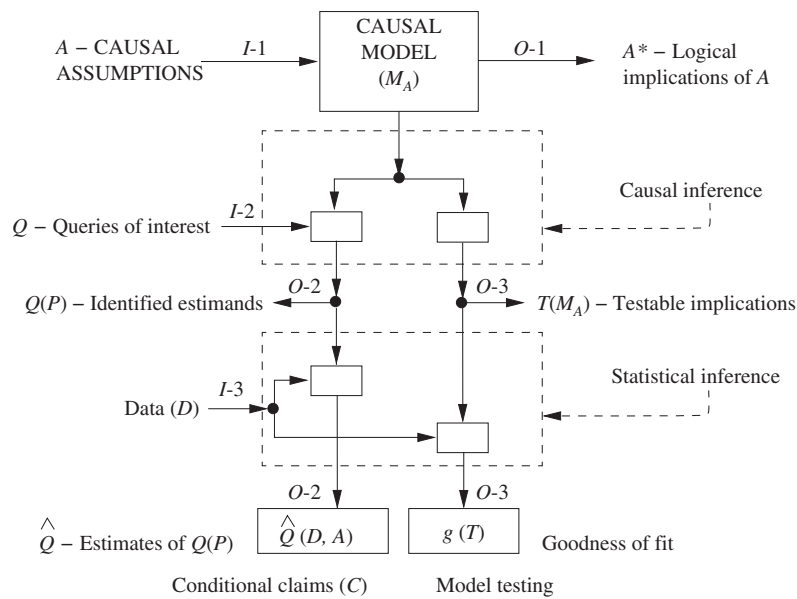


Figure 25.1 Causal analysis depicted as an inference engine converting assumptions (A), queries (Q), and data (D) into logical implications (A^*), conditional claims (C), and data-fitness indices ($g(T)$).

25.2.2 Assumptions in Nonparametric Models

A structural equation model (SEM) M is defined as follows.

Definition 25.1 Structural equation model (Pearl, 2000, page 203)

1. A set U of background or exogenous variables, representing factors outside the model, which nevertheless affect relationships within the model.
2. A set $V = \{V_1, \dots, V_n\}$ of endogenous variables, assumed to be observable. Each of these variables is functionally dependent on some subset PA_i of $U \cup V$.
3. A set F of functions $\{f_1, \dots, f_n\}$ such that each f_i determines the value of $V_i \in V$, $v_i = f_i(pa_i, u)$.
4. A joint probability distribution $P(u)$ over U .

A simple SEM model is depicted in Figure 25.2(a), which represents the following three functions:

$$\begin{aligned}
 z &= f_Z(u_Z), \\
 x &= f_X(z, u_X), \\
 y &= f_Y(x, u_Y),
 \end{aligned}
 \tag{25.1}$$

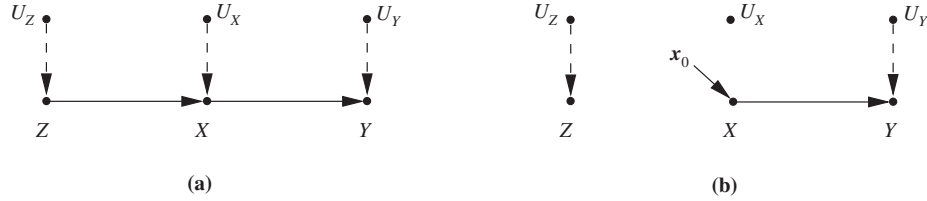


Figure 25.2 The diagrams associated with (a) the structural model of Equation (25.1) and (b) the modified model of Equation (25.2), representing the intervention $do(X = x_0)$.

where in this particular example, U_Z , U_X and U_Y are assumed to be jointly independent but otherwise arbitrarily distributed. Whenever dependence exists between any two exogenous variables, a bidirected arrow will be added to the diagram to represent this dependence (e.g., Figure 25.4).⁴ Each of these functions represents a causal process (or mechanism) that determines the value of the left variable (output) from the values on the right variables (inputs), and is assumed to be invariant unless explicitly intervened on. The absence of a variable from the right-hand side of an equation encodes the assumption that nature ignores that variable in the process of determining the value of the output variable. For example, the absence of variable Z from the arguments of f_Y conveys the empirical claim that variations in Z will leave Y unchanged, as long as variables U_Y and X remain constant.

It is important to distinguish between a *fully specified model* in which $P(U)$ and the collection of functions F are specified and a *partially specified model*, usually in the form of a diagram. The former entails one and only one observational distribution $P(V)$; the latter entails a set of observational distributions $P(V)$ that are compatible with the graph (those that can be generated by specifying $\langle F, P(u) \rangle$).

25.2.3 Representing Interventions, Counterfactuals and Causal Effects

This feature of invariance permits us to derive powerful claims about causal effects and counterfactuals, even in nonparametric models, where all functions and distributions remain unknown. This is done through a mathematical operator called $do(x)$, which simulates physical interventions by deleting certain functions from the model, replacing them with a constant $X = x$, while keeping the rest of the model unchanged (Haavelmo, 1943; Strotz and Wold, 1960; Pearl, 2014). For example, to emulate an intervention $do(x_0)$ that sets X to a constant x_0 in model M of Figure 25.2(a), the equation for x in Equation (25.1) is replaced by $x = x_0$, and we

4. More precisely, the absence of bidirected arrows implies marginal independences relative of the respective exogenous variables. In other words, the set of all bidirected edges constitute an i-map of $P(U)$ (Richardson, 2003).

obtain a new model, M_{x_0} ,

$$\begin{aligned} z &= f_Z(u_Z), \\ x &= x_0, \\ y &= f_Y(x, u_Y), \end{aligned} \tag{25.2}$$

the graphical description of which is shown in Figure 25.2(b).

The joint distribution associated with this modified model, denoted $P(z, y | do(x_0))$ describes the postintervention distribution of variables Y and Z (also called “controlled” or “experimental” distribution), to be distinguished from the preintervention distribution, $P(x, y, z)$, associated with the original model of Equation (25.1). For example, if X represents a treatment variable, Y a response variable, and Z some covariate that affects the amount of treatment received, then the distribution $P(z, y | do(x_0))$ gives the proportion of individuals that would attain response level $Y = y$ and covariate level $Z = z$ under the hypothetical situation in which treatment $X = x_0$ is administered uniformly to the population.⁵

In general, we can formally define the postintervention distribution by the equation

$$P_M(y | do(x)) = P_{M_x}(y). \tag{25.3}$$

In words, in the framework of model M , the postintervention distribution of outcome Y is defined as the probability that model M_x assigns to each outcome level $Y = y$. From this distribution, which is readily computed from any fully specified model M , we are able to assess treatment efficacy by comparing aspects of this distribution at different levels of x_0 .⁶

25.2.4 Identification, d -Separation and Causal Calculus

A central question in causal analysis is the question of *identification* of causal queries (e.g., the effect of intervention $do(X = x_0)$) from a combination of data and a partially specified model, for example, when only the graph is given and neither the functions F nor the distribution of U . In linear parametric settings, the question of identification reduces to asking whether some model parameter, β , has a unique

5. Equivalently, $P(z, y | do(x_0))$ can be interpreted as the joint probability of $(Z = z, Y = y)$ under a randomized experiment among units receiving treatment level $X = x_0$. Readers versed in potential-outcome notations may interpret $P(y | do(x), z)$ as the probability $P(Y_x = y | Z_x = z)$, where Y_x is the potential outcome under treatment $X = x$.

6. Counterfactuals are defined similarly through the equation $Y_x(u) = Y_{M_x}(u)$ (see Pearl, 2009b, Chapter 7), but will not be needed for the discussions in this paper.

solution in terms of the parameters of P (say the population covariance matrix). In the nonparametric formulation, the notion of “has a unique solution” does not directly apply since quantities such as $Q(M) = P(y | do(x))$ have no parametric signature and are defined procedurally by simulating an intervention in a causal model M , as in Equation (25.2). The following definition captures the requirement that Q be estimable from the data:

Definition 25.2 Identifiability

A causal query $Q(M)$ is identifiable, given a set of assumptions A , if for any two (fully specified) models, M_1 and M_2 , that satisfy A , we have⁷

$$P(M_1) = P(M_2) \implies Q(M_1) = Q(M_2). \quad (25.4)$$

In words, the functional details of M_1 and M_2 do not matter; what matters is that the assumptions in A (e.g., those encoded in the diagram) would constrain the variability of those details in such a way that equality of P 's would entail equality of Q 's. When this happens, Q depends on P only, and should therefore be expressible in terms of the parameters of P .

When a query Q is given in the form of a *do*-expression, for example, $Q = P(y | do(x), z)$, its identifiability can be decided systematically using an algebraic procedure known as the *do*-calculus (Pearl, 1995). It consists of three inference rules that permit us to map interventional and observational distributions whenever certain conditions hold in the causal diagram G .

The conditions that permit the application these inference rules can be read off the diagrams using a graphical criterion known as *d*-separation (Pearl, 1988).

Definition 25.3 *d*-separation

A set S of nodes is said to block a path p if either

1. p contains at least one arrow-emitting node that is in S , or
2. p contains at least one collision node that is outside S and has no descendant in S .

If S blocks *all* paths from set X to set Y , it is said to “*d*-separate X and Y ,” and then, it can be shown that variables X and Y are independent given S , written $X \perp\!\!\!\perp Y | S$.⁸

7. An implication similar to (25.4) is used in the standard statistical definition of parameter identification, where it conveys the uniqueness of a parameter set θ given a distribution P_θ (Lehmann and Casella, 1998). To see the connection, one should think about the query $Q = P(y | do(x))$ as a function $Q = g(\theta)$ where θ is the pair $F \cup P(u)$ that characterizes a fully specified model M .

8. See Hayduk, Cummings, and Stratkoter (2003), Glymour and Greenland (2008) and Pearl (2009b), pages 335, for a gentle introduction to *d*-separation.

D -separation reflects conditional independencies that hold in any distribution $P(v)$ that is compatible with the causal assumptions A embedded in the diagram. To illustrate, the path $U_Z \rightarrow Z \rightarrow X \rightarrow Y$ in Figure 25.2(a) is blocked by $S = \{Z\}$ and by $S = \{X\}$, since each emits an arrow along that path. Consequently we can infer that the conditional independencies $U_Z \perp\!\!\!\perp Y | Z$ and $U_Z \perp\!\!\!\perp Y | X$ will be satisfied in any probability function that this model can generate, regardless of how we parameterize the arrows. Likewise, the path $U_Z \rightarrow Z \rightarrow X \leftarrow U_X$ is blocked by the null set $\{\emptyset\}$, but it is not blocked by $S = \{Y\}$ since Y is a descendant of the collision node X . Consequently, the marginal independence $U_Z \perp\!\!\!\perp U_X$ will hold in the distribution, but $U_Z \perp\!\!\!\perp U_X | Y$ may or may not hold.⁹

25.2.5 The Rules of *do*-Calculus

Let X, Y, Z and W be arbitrary disjoint sets of nodes in a causal DAG G . We denote by $G_{\bar{X}}$ the graph obtained by deleting from G all arrows pointing to nodes in X . Likewise, we denote by $G_{\underline{X}}$ the graph obtained by deleting from G all arrows emerging from nodes in X . To represent the deletion of both incoming and outgoing arrows, we use the notation $G_{\bar{X}\underline{X}}$.

The following three rules are valid for every interventional distribution compatible with G :

Rule 25.1 Insertion/deletion of observations

$$P(y | do(x), z, w) = P(y | do(x), w) \quad \text{if } (Y \perp\!\!\!\perp Z | X, W)_{G_{\bar{X}}}. \quad (25.5)$$

Rule 25.2 Action/observation exchange

$$P(y | do(x), do(z), w) = P(y | do(x), z, w) \quad \text{if } (Y \perp\!\!\!\perp Z | X, W)_{G_{\bar{X}\underline{Z}}}. \quad (25.6)$$

Rule 25.3 Insertion/deletion of actions

$$P(y | do(x), do(z), w) = P(y | do(x), w) \quad \text{if } (Y \perp\!\!\!\perp Z | X, W)_{G_{\bar{X}\underline{Z}(W)}}, \quad (25.7)$$

where $Z(W)$ is the set of Z -nodes that are not ancestors of any W -node in $G_{\bar{X}}$.

To establish identifiability of a query Q , one needs to repeatedly apply the rules of *do*-calculus to Q , until the final expression no longer contains a *do*-operator;¹⁰ this renders it estimable from nonexperimental data. The *do*-calculus was proven

9. This special handling of collision nodes (or *colliders*, e.g., $Z \rightarrow X \leftarrow U_X$) reflects a general phenomenon known as *Berkson's paradox* (Berkson, 1946), whereby observations on a common consequence of two independent causes render those causes dependent. For example, the outcomes of two independent coins are rendered dependent by the testimony that at least one of them is a tail.

10. Such derivations are illustrated in graphical details in Pearl (2009b), page 87.

to be complete for the identifiability of causal effects in the form $Q = P(y | do(x), z)$ (Shpitser and Pearl, 2006; Huang and Valtorta, 2006), which means that if Q cannot be expressed in terms of the probability of observables P by repeated application of these three rules, such an expression does not exist. In other words, the query is not estimable from observational studies without making further assumptions, for example, linearity, monotonicity, additivity, absence of interactions, etc.

We shall see that, to establish transportability, the goal will be different; instead of eliminating *do*-operators from the query expression, we will need to separate them from a set of variables S that represent disparities between populations.

25.3 Inference Across Populations: Motivating Examples

To motivate the treatment of Section 25.4, we first demonstrate some of the subtle questions that transportability entails through three simple examples, informally depicted in Figure 25.3.

Example 25.1 Consider the graph in Figure 25.3(a) that represents cause-effect relationships in the pretreatment population in Los Angeles. We conduct a randomized trial in Los Angeles and estimate the causal effect of exposure X on outcome Y for every age group $Z = z$.^{11,12} We now wish to generalize the results to the population of New York City (NYC), but data alert us to the fact that the study distribution $P(x, y, z)$ in LA is significantly different from the one in NYC (call the latter $P^*(x, y, z)$). In particular, we notice that the average age in NYC is significantly higher than that in LA. How are we to estimate the causal effect of X on Y in NYC, denoted $P^*(y | do(x))$?

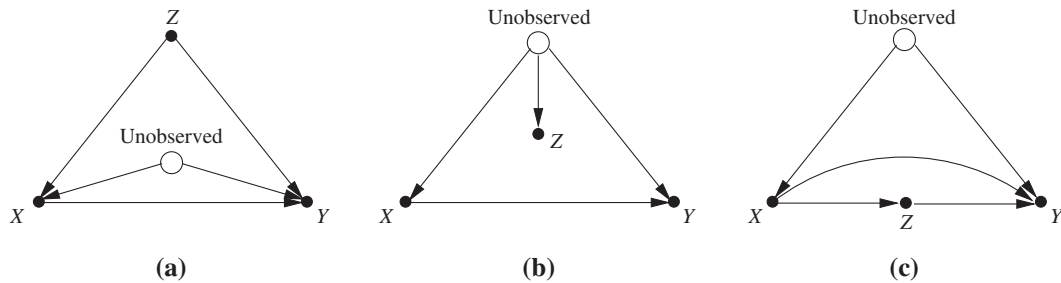


Figure 25.3 Causal diagrams depicting Examples 25.1–25.3. In (a) Z represents “age.” In (b), Z represents “linguistic skills” while age (in hollow circle) is unmeasured. In (c), Z represents a biological marker situated between the treatment (X) and a disease (Y).

11. Throughout the paper, each graph represents the causal structure of the population prior to the treatment, hence X stands for the level of treatment taken by an individual out of free choice.
 12. The arrow from Z to X represents the tendency of older people to seek treatment more often than younger people, and the arrow from Z to Y represents the effect of age on the outcome.

Our natural inclination would be to assume that age-specific effects are invariant across cities and so, if the LA study provides us with (estimates of) age-specific causal effects $P(y | do(x), Z = z)$, the overall causal effect in NYC should be

$$P^*(y | do(x)) = \sum_z P(y | do(x), z)P^*(z). \quad (25.8)$$

This *transport formula* combines experimental results obtained in LA, $P(y | do(x), z)$, with observational aspects of NYC population, $P^*(z)$, to obtain an experimental claim $P^*(y | do(x))$ about NYC.¹³

Our first task in this paper will be to explicate the assumptions that renders this extrapolation valid. We ask, for example, what must we assume about other confounding variables beside age, both latent and observed, for Equation (25.8) to be valid, or, would the same transport formula hold if Z was not age, but some proxy for age, say, language proficiency? More intricate yet, what if Z stood for an exposure-dependent variable, say hypertension level, that stands between X and Y ?

Let us examine the proxy issue first.

Example 25.2 Let the variable Z in Example 25.1 stand for subject's language proficiency, and let us assume that Z does not affect exposure (X) or outcome (Y), yet it correlates with both, being a proxy for age which is not measured in either study [see Figure 25.3(b)]. Given the observed disparity $P(z) \neq P^*(z)$, how are we to estimate the causal effect $P^*(y | do(x))$ for the target population of NYC from the z -specific causal effect $P(y | do(x), z)$ estimated at the study population of LA?

The inequality $P(z) \neq P^*(z)$ in this example may reflect either age difference or differences in the way that Z correlates with age. If the two cities enjoy identical age distributions and NYC residents acquire linguistic skills at a younger age, then since Z has no effect whatsoever on X and Y , the inequality $P(z) \neq P^*(z)$ can be ignored and, intuitively, the proper transport formula would be

$$P^*(y | do(x)) = P(y | do(x)). \quad (25.9)$$

If, on the other hand, the conditional probabilities $P(z | \text{age})$ and $P^*(z | \text{age})$ are the same in both cities, and the inequality $P(z) \neq P^*(z)$ reflects genuine age differences, Equation (25.9) is no longer valid, since the age difference may be a critical factor in determining how people react to X . We see, therefore, that the choice of the proper

13. At first glance, Equation (25.8) may be regarded as a routine application of “standardization” or “recalibration”—a statistical extrapolation method that can be traced back to a century-old tradition in demography and political arithmetic (Westergaard, 1916; Yule, 1934; Lane and Nelder, 1982). On a second thought it raises the deeper question of why we consider age-specific effects to be invariant across populations. See discussion following Example 25.2.

transport formula depends on the causal context in which population differences are embedded.

This example also demonstrates why the invariance of Z -specific causal effects should not be taken for granted. While justified in Example 25.1, with $Z = \text{age}$, it fails in Example 25.2, in which Z was equated with “language skills.” Indeed, using Figure 25.3(b) for guidance, the Z -specific effect of X on Y in NYC is given by:

$$\begin{aligned} P^*(y | do(x), z) &= \sum_{\text{age}} P^*(y | do(x), z, \text{age}) P^*(\text{age} | do(x), z) \\ &= \sum_{\text{age}} P^*(y | do(x), \text{age}) P^*(\text{age} | z) \\ &= \sum_{\text{age}} P(y | do(x), \text{age}) P^*(\text{age} | z). \end{aligned}$$

Thus, if the two populations differ in the relation between age and skill, that is,

$$P(\text{age} | z) \neq P^*(\text{age} | z)$$

the skill-specific causal effect would differ as well.

The intuition is clear. A NYC person at skill level $Z = z$ is likely to be in a totally different age group from his skill-equals in Los Angeles and, since it is age, not skill that shapes the way individuals respond to treatment, it is only reasonable that Los Angeles residents would respond differently to treatment than their NYC counterparts at the very same skill level.

The essential difference between Examples 25.1 and 25.2 is that age is normally taken to be an exogenous variable (not assigned by other factors in the model) while skills may be indicative of earlier factors (age, education, ethnicity) capable of modifying the causal effect. Therefore, conditional on skill, the effect may be different in the two populations.

Example 25.3 Examine the case where Z is a X -dependent variable, say a disease bio-marker, standing on the causal pathways between X and Y as shown in Figure 25.3(c). Assume further that the disparity $P(z | x) \neq P^*(z | x)$ is discovered and that, again, both the average and the z -specific causal effect $P(y | do(x), z)$ are estimated in the LA experiment, for all levels of X and Z . Can we, based on information given, estimate the average (or z -specific) causal effect in the target population of NYC?

Here, Equation (25.8) is wrong because the overall causal effect (in both LA and NYC) is no longer a simple average of the z -specific causal effects. The correct

weighting rule is

$$\begin{aligned} P^*(y | do(x)) \\ = \sum_z P^*(y | do(x), z)P^*(z | do(x)), \end{aligned} \quad (25.10)$$

which reduces to (25.8) only in the special case where Z is unaffected by X . Equation (25.9) is also wrong because we can no longer argue, as we did in Example 25.2, that Z does not affect Y , hence it can be ignored. Here, Z lies on the causal pathway between X and Y so, clearly, it affects their relationship. What then is the correct transport formula for this scenario?

To cast this example in a more realistic setting, let us assume that we wish to use Z as a “surrogate endpoint” to predict the efficacy of treatment X on outcome Y , where Y is too difficult and/or expensive to measure routinely (Prentice, 1989; Ellenberg and Hamilton, 1989). Thus, instead of considering experimental and observational studies conducted at two different locations, we consider two such studies taking place at the same location, but at different times. In the first study, we measure $P(y, z | do(x))$ and discover that Z is a good surrogate, namely, knowing the effect of treatment on Z allows prediction of the effect of treatment on the more clinically relevant outcome (Y) (Joffe and Greene, 2009). Once Z is proclaimed a “surrogate endpoint,” it invites efforts to find direct means of controlling Z . For example, if cholesterol level is found to be a predictor of heart diseases in a long-run trial, drug manufacturers would rush to offer cholesterol-reducing substances for public consumption. As a result, both the prior $P(z)$ and the treatment-dependent probability $P(z | do(x))$ would undergo a change, resulting in $P^*(z)$ and $P^*(z | do(x))$, respectively.

We now wish to reassess the effect of the drug $P^*(y | do(x))$ in the new population and do it in the cheapest possible way, namely, by conducting an observational study to estimate $P^*(z, x)$, acknowledging that confounding exists between X and Y and that the drug affects Y both directly and through Z , as shown in Figure 25.3(c).

Using a graphical representation to encode the assumptions articulated thus far, and further assuming that the disparity observed stems only from a difference in people’s susceptibility to X (and not due to a change in some unobservable confounder), we will prove in Section 25.5 that the correct transport formula should be

$$P^*(y | do(x)) = \sum_z P(y | do(x), z)P^*(z | x), \quad (25.11)$$

which is different from both (25.8) and (25.9). It calls instead for the z -specific effects to be reweighted by the conditional probability $P^*(z|x)$, estimated in the target population.¹⁴

To see how the transportability problem fits into the general scheme of causal analysis discussed in Section 25.2.1 (Figure 25.1), we note that, in our case, the data comes from two sources, experimental (from the study) and non-experimental (from the target), assumptions are encoded in the form of selection diagrams, and the query stands for the causal effect (e.g., $P^*(y|do(x))$). Although this paper does not discuss the goodness-of-fit problem, standard methods are available for testing the compatibility of the selection diagram with the data available.

25.4 Formalizing Transportability

25.4.1 Selection Diagrams and Selection Variables

The pattern that emerges from the examples discussed in Section 25.3 indicates that transportability is a causal, not statistical notion. In other words, the conditions that license transport as well as the formulas through which results are transported depend on the causal relations between the variables in the domain, not merely on their statistics. For instance, it was important in Example 25.3 to ascertain that the change in $P(z|x)$ was due to the change in the way Z is affected by X , but not due to a change in confounding conditions between the two. This cannot be determined solely by comparing $P(z|x)$ and $P^*(z|x)$. If X and Z are confounded [e.g., Figure 25.6(e)], it is quite possible for the inequality $P(z|x) \neq P^*(z|x)$ to hold, reflecting differences in confounding, while the way that Z is affected by X (i.e., $P(z|do(x))$) is the same in the two populations—a different transport formula will then emerge for this case.

Consequently, licensing transportability requires knowledge of the mechanisms, or processes, through which population differences come about; different localization of these mechanisms yield different transport formulae. This can be seen most vividly in Example 25.2 [Figure 25.3(b)] where we reasoned that no reweighting is necessary if the disparity $P(z) \neq P^*(z)$ originates with the way language proficiency depends on age, while the age distribution itself remains the same. Yet, because age is not measured, this condition cannot be detected in the probability distribution P , and cannot be distinguished from an alternative

14. Quite often the possibility of running a second randomized experiment to estimate $P^*(z|do(x))$ is also available to investigators, though at a higher cost. In such cases, a transport formula would be derivable under more relaxed assumptions, for example, allowing for X and Z to be confounded.

condition,

$$P(\text{age}) \neq P^*(\text{age}) \quad \text{and} \quad P(z|\text{age}) = P^*(z|\text{age}),$$

one that may require reweighting according to Equation (25.8). In other words, every probability distribution $P(x, y, z)$ that is compatible with the process of Figure 25.3(b) is also compatible with that of Figure 25.3(a) and, yet, the two processes dictate different transport formulas.

Based on these observations, it is clear that if we are to represent formally the differences between populations (similarly, between experimental settings or environments), we must resort to a representation in which the causal mechanisms are explicitly encoded and in which differences in populations are represented as local modifications of those mechanisms.

To this end, we will use causal diagrams augmented with a set, S , of “selection variables,” where each member of S corresponds to a mechanism by which the two populations differ, and switching between the two populations will be represented by conditioning on different values of these S variables.¹⁵

Intuitively, if $P(v|do(x))$ stands for the distribution of a set V of variables in the experimental study (with X randomized) then we designate by $P^*(v|do(x))$ the distribution of V if we were to conduct the study on population Π^* instead of Π . We now attribute the difference between the two to the action of a set S of selection variables, and write^{16,17}

$$P^*(v|do(x)) = P(v|do(x), s^*).$$

The selection variables in S may represent all factors by which populations may differ or that may “threaten” the transport of conclusions between populations. For

15. Disparities among populations or subpopulations can also arise from differences in design; for example, if two samples are drawn by different criteria from a given population. The problem of generalizing between two such subpopulations is usually called *sampling selection bias* (Heckman, 1979; Hernán, Hernández-Díaz and Robins, 2004; Cole and Stuart, 2010; Pearl, 2013; Bareinboim, Tian and Pearl, 2014). In this paper, we deal only with nature-induced, not man-made disparities.

16. Alternatively, one can represent the two populations’ distributions by $P(v|do(x), s)$, and $P(v|do(x), s^*)$, respectively. The results, however, will be the same, since only the location of S enters the analysis.

17. Pearl (1993, 2009b, page 71), Spirtes, Glymour and Scheines (1993) and Dawid (2002), for example, use conditioning on auxiliary variables to switch between experimental and observational studies. Dawid (2002) further uses such variables to represent changes in parameters of probability distributions.

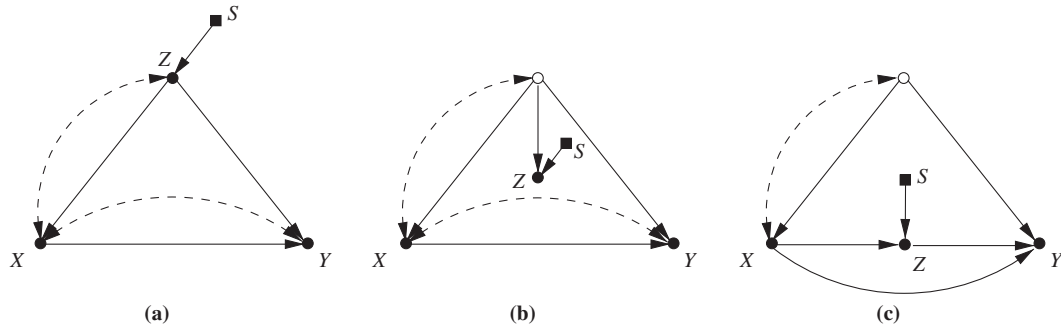


Figure 25.4 Selection diagrams depicting specific versions of Examples 25.1–25.3. In (a), the two populations differ in age distributions. In (b), the populations differ in how Z depends on age (an unmeasured variable, represented by the hollow circle) and the age distributions are the same. In (c), the populations differ in how Z depends on X . In all diagrams, dashed arcs (e.g., $X \leftarrow \text{-----} \rightarrow Y$) represent the presence of latent variables affecting both X and Y .

example, in Figure 25.4(a) the age disparity $P(z) \neq P^*(z)$ discussed in Example 25.1 will be represented by the inequality

$$P(z) \neq P(z|s),$$

where S stands for all factors responsible for drawing subjects at age $Z = z$ to NYC rather than LA.

Of equal importance is the absence of an S variable pointing to Y in Figure 25.4(a), which encodes the assumption that age-specific effects are invariant across the two populations.

This graphical representation, which we will call “selection diagrams” is defined as follows:¹⁸

Definition 25.4 Selection diagram

Let $\langle M, M^* \rangle$ be a pair of structural causal models (Definition 25.1) relative to domains $\langle \Pi, \Pi^* \rangle$, sharing a causal diagram G . $\langle M, M^* \rangle$ is said to induce a selection diagram D if D is constructed as follows:

1. Every edge in G is also an edge in D ;

18. The assumption that there are no structural changes between domains can be relaxed starting with $D = G^*$ and adding S -nodes following the same procedure as in Definition 25.4, while enforcing acyclicity. In extreme cases in which the two domains differ in causal directionality (Spirtes, Glymour and Scheines, 2000, pages 298–299), acyclicity cannot be maintained. This complication as well as one created when G is an edge-super set of G^* require a more elaborated graphical representation and lie beyond the scope of this paper.

2. D contains an extra edge $S_i \rightarrow V_i$ whenever there might exist a discrepancy $f_i \neq f_i^*$ or $P(U_i) \neq P^*(U_i)$ between M and M^* .

In summary, the S -variables locate the *mechanisms* where structural discrepancies between the two populations are suspected to take place. Alternatively, the absence of a selection node pointing to a variable represents the assumption that the mechanism responsible for assigning value to that variable is the same in the two populations. In the extreme case, we could add selection nodes to all variables, which means that we have no reason to believe that the populations share any mechanism in common, and this, of course would inhibit any exchange of information among the populations. The invariance assumptions between populations, as we will see, will open the door for the transport of some experimental findings.

For clarity, we will represent the S variables by squares, as in Figure 25.4, which uses selection diagrams to encode the three examples discussed in Section 25.3. (Besides the S variables, these graphs also include additional latent variables, represented by bidirected edges, which makes the examples more realistic.) In particular, Figure 25.4(a) and 25.4(b) represent, respectively, two different mechanisms responsible for the observed disparity $P(z) \neq P^*(z)$. The first [Figure 25.4(a)] dictates transport formula (25.8), while the second [Figure 25.4(b)] calls for direct, unadjusted transport (25.9). This difference stems from the location of the S variables in the two diagrams. In Figure 25.4(a), the S variable represents unspecified factors that cause age differences between the two populations, while in Figure 25.4(b), S represents factors that cause differences in reading skills (Z) while the age distribution itself (unobserved) remains the same.

In this paper, we will address the issue of transportability assuming that scientific knowledge about invariance of certain mechanisms is available and encoded in the selection diagram through the S nodes. Such knowledge is, admittedly, more demanding than that which shapes the structure of each causal diagram in isolation. It is, however, a prerequisite for any attempt to justify transfer of findings across populations, which makes selection diagrams a mathematical object worthy of analysis.

25.4.2 Transportability: Definitions and Examples

Using selection diagrams as the basic representational language, and harnessing the concepts of intervention, *do*-calculus, and identifiability (Section 25.2), we can now give the notion of transportability a formal definition.

Definition 25.5 Transportability

Let D be a selection diagram relative to domains $\langle \Pi, \Pi^* \rangle$. Let $\langle P, I \rangle$ be the pair of observational and interventional distributions of Π , and P^* be the observational

distribution of Π^* . The causal relation $R(\Pi^*) = P^*(y | do(x), z)$ is said to be transportable from Π to Π^* in D if $R(\Pi^*)$ is uniquely computable from P, P^*, I in any model that induces D .

Two interesting connections between identifiability and transportability are worth noting. First, note that all identifiable causal relations in D are also transportable, because they can be computed directly from P^* and require no experimental information from Π . Second, note that given causal diagram G , one can produce a selection diagram D such that identifiability in G is equivalent to transportability in D . First set $D = G$, and then add selection nodes pointing to all variables in D , which represents that the target domain does not share any mechanism with its counterpart—this is equivalent to the problem of identifiability because the only way to achieve transportability is to identify R from scratch in the target population.

While the problems of identifiability and transportability are related, proofs of nontransportability are more involved than those of nonidentifiability for they require one to demonstrate the nonexistence of two competing models compatible with D , agreeing on $\{P, P^*, I\}$, and disagreeing on $R(\Pi^*)$.

Definition 25.5 is declarative, and does not offer an effective method of demonstrating transportability even in simple models. Theorem 25.1 offers such a method using a sequence of derivations in *do*-calculus.

Theorem 25.1 *Let D be the selection diagram characterizing two populations, Π and Π^* , and S a set of selection variables in D . The relation $R = P^*(y | do(x), z)$ is transportable from Π to Π^* if the expression $P(y | do(x), z, s)$ is reducible, using the rules of *do*-calculus, to an expression in which S appears only as a conditioning variable in *do*-free terms.*

Proof. Every relation satisfying the condition of Theorem 25.1 can be written as an algebraic combination of two kinds of terms, those that involve S and those that do not. The former can be written as P^* -terms and are estimable, therefore, from observations on Π^* , as required by Definition 25.5. All other terms, especially those involving *do*-operators, do not contain S ; they are experimentally identifiable therefore in Π . ■

This criterion was proven to be both sufficient and necessary for causal effects, namely $R = P^*(y | do(x))$ (Bareinboim and Pearl, 2012). Theorem 25.1, though procedural, does not specify the sequence of rules leading to the needed reduction when such a sequence exists. Bareinboim and Pearl (2013b) derived a complete procedural solution for this, based on graphical method developed in (Tian and Pearl, 2002;

Shpitser and Pearl, 2006). Despite its completeness, however, the procedural solution is not trivial, and we take here an alternative route to establish a simple and transparent procedure for confirming transportability, guided by two recognizable subgoals.

Definition 25.6 Trivial transportability

A causal relation R is said to be *trivially transportable* from Π to Π^* , if $R(\Pi^*)$ is identifiable from (G^*, P^*) .

This criterion amounts to an ordinary test of identifiability of causal relations using graphs, as given by Definition 25.2. It permits us to estimate $R(\Pi^*)$ directly from observational studies on Π^* , un-aided by causal information from Π .

Example 25.4 Let R be the causal effect $P^*(y | do(x))$ and let the selection diagram of Π and Π^* be given by $X \rightarrow Y \leftarrow S$, then R is trivially transportable, since $R(\Pi^*) = P^*(y | x)$.

Another special case of transportability occurs when a causal relation has identical form in both domains—no recalibration is needed.

Definition 25.7 Direct transportability

A causal relation R is said to be *directly transportable* from Π to Π^* , if $R(\Pi^*) = R(\Pi)$.

A graphical test for direct transportability of $R = P^*(y | do(x), z)$ follows from *do*-calculus and reads: $(S \perp\!\!\!\perp Y | X, Z)_{G_{\bar{X}}}$; in words, X blocks all paths from S to Y once we remove all arrows pointing to X and condition on Z . As a concrete example, this test is satisfied in Figure 25.4(a) and, therefore, the z -specific effects is the same in both populations; it is directly transportable.

Remark The notion of “external validity” as defined by Manski (2007) (footnote 1) corresponds to Direct Transportability, for it requires that R retains its validity without adjustment, as in Equation (25.9). Such conditions preclude the use of information from Π^* to recalibrate R .

Example 25.5 Let R be the causal effect of X on Y , and let D have a single S node pointing to X , then R is directly transportable, because causal effects are independent of the selection mechanism (see Pearl, 2009b, pages 72 and 73).

Example 25.6 Let R be the z -specific causal effect of X on Y $P^*(y | do(x), z)$ where Z is a set of variables, and P and P^* differ only in the conditional probabilities $P(z | pa(Z))$ and $P^*(z | pa(Z))$ such that $(Z \perp\!\!\!\perp Y | pa(Z))$, as shown in Figure 25.4(b). Under these conditions, R is not directly transportable. However, the $pa(Z)$ -specific causal effects $P^*(y | do(x), pa(Z))$ are directly transportable, and so is $P^*(y | do(x))$. Note that, due to the confounding arcs, none of these quantities is identifiable.

25.5 Transportability of Causal Effects—A Graphical Criterion

We now state and prove two theorems that permit us to decide algorithmically, given a selection diagram, whether a relation is transportable between two populations, and what the transport formula should be.

Theorem 25.2 *Let D be the selection diagram characterizing two populations, Π and Π^* , and S the set of selection variables in D . The strata-specific causal effect $P^*(y | do(x), z)$ is transportable from Π to Π^* if Z d -separates Y from S in the X -manipulated version of D , that is, Z satisfies $(Y \perp\!\!\!\perp S | Z, X)_{D_{\bar{X}}}$.*

Proof.

$$P^*(y | do(x), z) = P(y | do(x), z, s^*).$$

From Rule 25.1 of do -calculus we have: $P(y | do(x), z, s^*) = P(y | do(x), z)$ whenever Z satisfies $(Y \perp\!\!\!\perp S | Z, X)$ in $D_{\bar{X}}$. This proves Theorem 25.2. ■

Definition 25.8 S-admissibility

A set T of variables satisfying $(Y \perp\!\!\!\perp S | T, X)$ in $D_{\bar{X}}$ will be called S -admissible (with respect to the causal effect of X on Y).

Corollary 25.1 *The average causal effect $P^*(y | do(x))$ is transportable from Π to Π^* if there exists a set Z of observed pretreatment covariates that is S -admissible. Moreover, the transport formula is given by the weighting of Equation (25.8).*

Example 25.7 The causal effect is transportable in Figure 25.4(a), since Z is S -admissible, and in Figure 25.4(b), where the empty set is S -admissible. It is also transportable by the same criterion in Figure 25.5(b), where W is S -admissible, but not in Figure 25.5(a) where no S -admissible set exists.

Corollary 25.2 *Any S variable that is pointing directly into X as in Figure 25.6(a), or that is d -separated from Y in $D_{\bar{X}}$ can be ignored.*

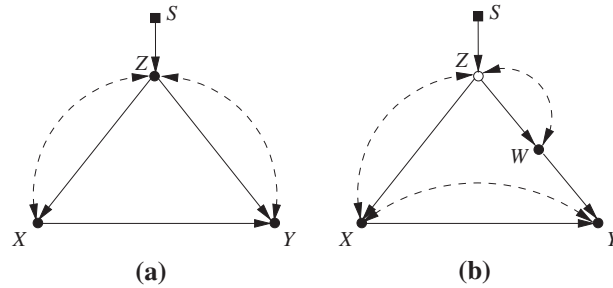


Figure 25.5 Selection diagrams illustrating S -admissibility. (a) Has no S -admissible set while in (b), W is S -admissible.

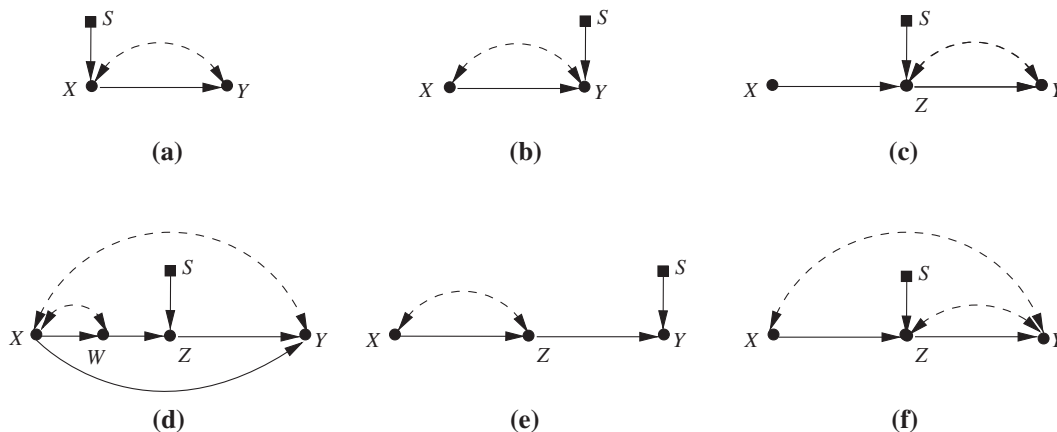


Figure 25.6 Selection diagrams illustrating transportability. The causal effect $P(y | do(x))$ is (trivially) transportable in (c) but not in (b) and (f). It is transportable in (a), (d), and (e) (see Corollary 25.2).

This follows from the fact that the empty set is S -admissible relative to any such S variable. Conceptually, the corollary reflects the understanding that differences in propensity to receive treatment do not hinder the transportability of treatment effects; the randomization used in the experimental study washes away such differences.

We now generalize Theorem 25.2 to cases involving treatment-dependent Z variables, as in Figure 25.4(c).

Theorem 25.3 *The average causal effect $P^*(y | do(x))$ is transportable from Π to Π^* if either one of the following conditions holds:*

1. $P^*(y | do(x))$ is trivially transportable;
2. There exists a set of covariates, Z (possibly affected by X) such that Z is S -admissible and for which $P^*(z | do(x))$ is transportable;
3. There exists a set of covariates, W that satisfy $(X \perp\!\!\!\perp Y | W)_{D_{\overline{X}(W)}}$ and for which $P^*(w | do(x))$ is transportable.

Proof. 1. Condition 1 entails transportability.

2. If condition 2 holds, it implies

$$P^*(y | do(x)) = P(y | do(x), s) \tag{25.12}$$

$$= \sum_z P(y | do(x), z, s) P(z | do(x), s) \tag{25.13}$$

$$= \sum_z P(y | do(x), z) P^*(z | do(x)). \tag{25.14}$$

We now note that the transportability of $P(z | do(x))$ should reduce $P^*(z | do(x))$ to a star-free expression and would render $P^*(y | do(x))$ transportable.

3. If condition 3 holds, it implies

$$P^*(y | do(x)) = P(y | do(x), s) \quad (25.15)$$

$$= \sum_w P(y | do(x), w, s) P(w | do(x), s) \quad (25.16)$$

$$= \sum_w P(y | w, s) P^*(w | do(x)) \quad (25.17)$$

(by Rule 25.3 of do -calculus)

$$= \sum_w P^*(y | w) P^*(w | do(x)). \quad (25.18)$$

We similarly note that the transportability of $P^*(w | do(x))$ should reduce $P(w | do(x), s)$ to a star-free expression and would render $P^*(y | do(x))$ transportable. This proves Theorem 25.3. ■

Example 25.8 To illustrate the application of Theorem 25.3, let us apply it to Figure 25.4(c), which corresponds to the surrogate endpoint problem discussed in Section 25.3 (Example 25.3). Our goal is to estimate $P^*(y | do(x))$ —the effect of X on Y in the new population created by changes in how Z responds to X . The structure of the problem permits us to satisfy condition 2 of the Theorem 25.3, since Z is S -admissible and $P^*(z | do(x))$ is trivially transportable. The former can be seen from $(S \perp\!\!\!\perp Y | X, Z)_{G_{\bar{X}}}$, hence $P^*(y | do(x), z) = P(y | do(x), z)$; the latter can be seen from the fact that X and Z are unconfounded, hence $P^*(z | do(x)) = P^*(z | x)$. Putting the two together, we get

$$P^*(y | do(x)) = \sum_z P(y | do(x), z) P^*(z | x), \quad (25.19)$$

which proves Equation (25.11).

Remark The test entailed by Theorem 25.3 is recursive, since the transportability of one causal effect depends on that of another. However, given that the diagram is finite and acyclic, the sets Z and W needed in conditions 2 and 3 of Theorem 25.3 would become closer and closer to X , and the iterative process will terminate after a finite number of steps. This occurs because the causal effects $P^*(z | do(x))$ (likewise, $P^*(w | do(x))$) is trivially transportable and equals $P(z)$ for any Z node that is not a descendant of X . Thus, the need for reiteration applies only to those members of Z that lie on the causal pathways from X to Y . Note further that the analyst need not terminate the procedure upon satisfying the conditions of Theorem 25.3. If one wishes to reduce the number of experiments, it can continue until no further reduction is feasible.

Example 25.9 Figure 25.6(d) requires that we invoke both conditions of Theorem 25.3, iteratively. To satisfy condition 2, we note that Z is S -admissible, and we need to prove the transportability of $P^*(z | do(x))$. To do that, we invoke condition 3 and note that W d -separates X from Z in D . There remains to confirm the transportability of $P^*(w | do(x))$, but this is guaranteed by the fact that the empty set is S -admissible relative to W , since $(W \perp\!\!\!\perp S)$. Hence, by Theorem 25.2 (replacing Y with W) $P^*(w | do(x))$ is transportable, which bestows transportability on $P^*(y | do(x))$. Thus, the final transport formula (derived formally in 25.A) is:

$$P^*(y | do(x)) = \sum_z P(y | do(x), z) \cdot \sum_w P(w | do(x)) P^*(z | w). \quad (25.20)$$

The first two factors of the expression are estimable in the experimental study, and the third through observational studies on the target population. Note that the joint effect $P(y, w, z | do(x))$ need not be estimated in the experiment; a decomposition that results in decrease of measurement cost and sampling variability.

A similar analysis proves the transportability of the causal effect in Figure 25.6(e) (see Pearl and Bareinboim, 2011). The model of Figure 25.6(f), however, does not allow for the transportability of $P^*(y | do(x))$ as witnessed by the absence of S -admissible set in the diagram, and the inapplicability of condition 3 of Theorem 25.3.

Example 25.10 To illustrate the power of Theorem 25.3 in discerning transportability and deriving transport formulae, Figure 25.7 represents a more intricate selection diagram, which requires several iteration to discern transportability. The transport formula for this diagram is given by (derived formally in 25.A):

$$P^*(y | do(x)) = \sum_z P(y | do(x), z) \cdot \sum_w P^*(z | w) \sum_t P(w | do(x), t) P^*(t). \quad (25.21)$$

The main power of this formula is to guide investigators in deciding what measurements need be taken in both the experimental study and the target population. It asserts, for example, that variables U and V need not be measured. It likewise asserts that the W -specific causal effects need not be estimated in the experimental study and only the conditional probabilities $P^*(z | w)$ and $P^*(t)$ need be estimated in the target population. The derivation of this formulae is given in 25.A.

Despite its power, Theorem 25.3 is not complete, namely, it is not guaranteed to approve all transportable relations or to disapprove all nontransportable ones. An example of the former is contrived in Bareinboim and Pearl (2012), where an alternative, necessary and sufficient condition is established in both graphical and algorithmic form. Theorem 25.3 provides, nevertheless, a simple and powerful method of establishing transportability in practice.

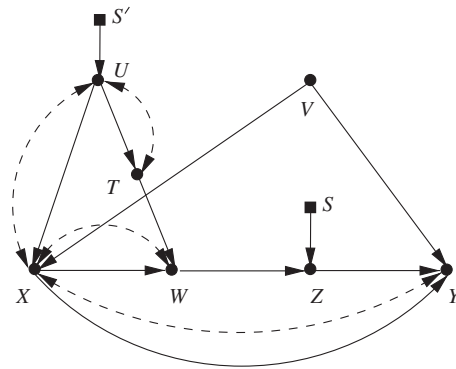


Figure 25.7 Selection diagram in which the causal effect is shown to be transportable in multiple iterations of Theorem 25.3 (see 25.A).

25.6 Conclusions

Given judgements of how target populations may differ from those under study, the paper offers a formal representational language for making these assessments precise and for deciding whether causal relations in the target population can be inferred from those obtained in an experimental study. When such inference is possible, the criteria provided by Theorems 25.2 and 25.3 yield transport formulae, namely, principled ways of calibrating the transported relations so as to properly account for differences in the populations. These transport formulae enable the investigator to select the essential measurements in both the experimental and observational studies, and thus minimize measurement costs and sample variability.

The inferences licensed by Theorem 25.2 and 25.3 represent worst case analysis, since we have assumed, in the tradition of nonparametric modeling, that every variable may potentially be an effect-modifier (or moderator). If one is willing to assume that certain relationships are noninteractive, or monotonic as is the case in additive models, then additional transport licenses may be issued, beyond those sanctioned by Theorems 25.2 and 25.3.

While the results of this paper concern the transfer of causal information from experimental to observational studies, the method can also benefit in transporting statistical findings from one observational study to another (Pearl and Bareinboim, 2011). The rationale for such transfer is two-fold. First, information from the first study may enable researchers to avoid repeated measurement of certain variables in the target population. Second, by pooling data from both populations, we increase the precision in which their commonalities are estimated and, indirectly, also increase the precision by which the target relationship is

transported. Substantial reduction in sampling variability can be thus achieved through this decomposition (Pearl, 2012b).

Clearly, the same data-sharing philosophy can be used to guide Meta-Analysis (Glass, 1976; Hedges and Olkin, 1985; Rosenthal, 1995; Owen, 2009), where one attempts to combine results from many experimental and observational studies, each conducted on a different population and under a different set of conditions, so as to construct an aggregate measure of effect size that is “better,” in some formal sense, than any one study in isolation. While traditional approaches aim to average out differences between studies, our theory exploits the commonalities among the populations studied and the target population. By pooling together commonalities and discarding areas of disparity, we gain maximum use of the available samples (Bareinboim and Pearl, 2013c).

To be of immediate use, our method relies on the assumption that the analyst is in possession of sufficient background knowledge to determine, at least qualitatively, where two populations may differ from one another. This knowledge is not vastly different from that required in any principled approach to causation in observational studies, since judgement about possible effects of omitted factors is crucial in any such analysis. Whereas such knowledge may only be partially available, the analysis presented in this paper is nevertheless essential for understanding what knowledge is needed for the task to succeed and how sensitive conclusions are to knowledge that we do not possess.

Real-life situations will be marred, of course, with additional complications that were not addressed directly in this paper; for example, measurement errors, selection bias, finite sample variability, uncertainty about the graph structure and the possible existence of unmeasured confounders between any two nodes in the diagram. Such issues are not unique to transportability; they plague any problem in causal analysis, regardless of whether they are represented formally or ignored by avoiding formalism. The methods offered in this paper are representative of what theory permits us to do in ideal situations, and the graphical representation presented in this paper makes the assumptions explicit and transparent. Transparency is essential for reaching tentative consensus among researchers and for facilitating discussions to distinguish that which is deemed plausible and important from that which is negligible or implausible.

Finally, it is important to mention two recent extensions of the results reported in this article. Bareinboim and Pearl (2013a) have addressed the problem of transportability in cases where only a limited set of experiments can be conducted at the source environment. Subsequently, the results were generalized to the problem of “meta-transportability,” that is, pooling experimental results from multiple and disparate sources to synthesize a consistent estimate of a causal relation at yet

another environment, potentially different from each of the former (Bareinboim and Pearl, 2013c). It is shown that such synthesis may be feasible from multiple sources even in cases where it is not feasible from any one source in isolation.

25.A Appendix

Derivation of the transport formula for the causal effect in the model of Figure 25.6(d), [Equation (25.20)]:

$$\begin{aligned}
 P^*(y | do(x)) &= P(y | do(x), s) \\
 &= \sum_z P(y | do(x), s, z) P(z | do(x), s) \\
 &= \sum_z P(y | do(x), z) P(z | do(x), s) \\
 &\quad \text{(2nd condition of Theorem 25.3, } S\text{-admissibility of } Z \text{ of } CE(X, Y)) \\
 &= \sum_z P(y | do(x), z) \cdot \sum_w P(z | do(x), w, s) P(w | do(x), s) \\
 &= \sum_z P(y | do(x), z) \cdot \sum_w P(z | w, s) P(w | do(x), s) \\
 &\quad \text{(3rd condition of Theorem 25.3, } (X \perp\!\!\!\perp Z | W, S)_{D_{\overline{X}(w)}}) \\
 &= \sum_z P(y | do(x), z) \cdot \sum_w P(z | w, s) P(w | do(x)) \\
 &\quad \text{(2nd condition of Theorem 25.3, } S\text{-admissibility} \\
 &\quad \text{of the empty set } \{\} \text{ of } CE(X, W)) \\
 &= \sum_z P(y | do(x), z) \cdot \sum_w P^*(z | w) P(w | do(x)). \tag{25.A.1}
 \end{aligned}$$

Derivation of the transport formula for the causal effect in the model of Figure 25.7, [Equation (25.21)]:

$$\begin{aligned}
 P^*(y | do(x)) &= P(y | do(x), s, s') \\
 &= \sum_z P(y | do(x), s, s', z) P(z | do(x), s, s') \\
 &= \sum_z P(y | do(x), z) P(z | do(x), s, s') \\
 &\quad \text{(2nd condition of Theorem 25.3, } S\text{-admissibility of } Z \text{ of } CE(X, Z)) \\
 &= \sum_z P(y | do(x), z) \cdot \sum_w P(z | do(x), s, s', w) P(w | do(x), s, s') \\
 &= \sum_z P(y | do(x), z) \cdot \sum_w P(z | s, s', w) P(w | do(x), s, s')
 \end{aligned}$$

$$\begin{aligned}
& \text{(3rd condition of Theorem 25.3, } (X \perp\!\!\!\perp Z \mid W, S, S')_{D_{\overline{X(W)}}}) \\
&= \sum_z P(y \mid do(x), z) \sum_w P(z \mid s, s', w) \cdot \sum_t P(w \mid do(x), s, s', t) P(t \mid do(x), s, s') \\
&= \sum_z P(y \mid do(x), z) \sum_w P(z \mid s, s', w) \cdot \sum_t P(w \mid do(x), t) P(t \mid do(x), s, s') \\
& \text{(2nd condition of Theorem 25.3, } S\text{-admissibility of } T \text{ on } CE(X, W)) \\
&= \sum_z P(y \mid do(x), z) \sum_w P(z \mid s, s', w) \cdot \sum_t P(w \mid do(x), t) P(t \mid s, s') \\
& \text{(1st condition of Theorem 25.3/Rule 25.3 of } do\text{-calculus,} \\
& \text{(} X \perp\!\!\!\perp T \mid S, S')_D) \\
&= \sum_z P(y \mid do(x), z) \sum_w P^*(z \mid w) \cdot \sum_t P(w \mid do(x), t) P^*(t). \tag{25.A.2}
\end{aligned}$$

Acknowledgments

This paper benefited from discussions with Onyebuchi Arah, Stuart Baker, Sander Greenland, Michael Hoefler, Marshall Joffe, William Shadish, Ian Shrier and Dylan Small. We are grateful to two anonymous referees for thorough reviews of this manuscript and for suggesting a simplification in the transport formula of Example 25.10. This research was supported in parts by NIH Grant #1R01 LM009961-01, NSF Grant #IIS-0914211 and ONR grant #N000-14-09-1-0665.

References

- ADELMAN, L. (1991). Experiments, quasi-experiments, and case studies: A review of empirical methods for evaluating decision support systems. *IEEE Transactions on Systems, Man and Cybernetics* 21 293–301.
- BALKE, A. and PEARL, J. (1995). Counterfactuals and policy analysis in structural models. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence* (P. Besnard and S. Hanks, eds.) 11–18. Morgan Kaufmann, San Francisco, CA.
- BAREINBOIM, E., BRITO, C. and PEARL, J. (2012). Local characterizations of causal Bayesian networks. In *Graph Structures for Knowledge Representation and Reasoning. Lecture Notes in Artificial Intelligence* 7205 1–17. Springer, Berlin.
- BAREINBOIM, E. and PEARL, J. (2012). Transportability of causal effects: Completeness results. In *Proceedings of the Twenty-Sixth National Conference on Artificial Intelligence* 698–704. AAAI Press, Menlo Park, CA.
- BAREINBOIM, E. and PEARL, J. (2013a). Causal transportability with limited experiments. In *Proceedings of the Twenty-Seventh National Conference on Artificial Intelligence* 95–101. AAAI Press, Menlo Park, CA.

- BAREINBOIM, E. and PEARL, J. (2013b). A general algorithm for deciding transportability of experimental results. *J. Causal Inference* **1** 107–134.
- BAREINBOIM, E. and PEARL, J. (2013c). Meta-transportability of causal effects: A formal approach. In *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics (AISTATS 2013)*. *J. Mach. Learn. Res.* **31** 135–143.
- BAREINBOIM, E., TIAN, J. and PEARL, J. (2014). Recovering from selection bias in causal and statistical inference. In *Proceedings of The Twenty-Eighth Conference on Artificial Intelligence* (C. E. Brodley and P. Stone, eds.) 2410–2416. AAAI Press, Menlo Park, CA.
- BERKSON, J. (1946). Limitations of the application of fourfold table analysis to hospital data. *Biometrics* **2** 47–53.
- BOLLEN, K. A. and PEARL, J. (2013). Eight myths about causality and structural equation models. In *Handbook of Causal Analysis for Social Research* (S. L. Morgan, ed.) Chapter 15. Springer, New York.
- CAMPBELL, D. and STANLEY, J. (1963). *Experimental and Quasi-Experimental Designs for Research*. Wadsworth, Chicago.
- COLE, S. R. and STUART, E. A. (2010). Generalizing evidence from randomized clinical trials to target populations: The ACTG 320 trial. *Am. J. Epidemiol.* **172** 107–115.
- DAVIS, J. A. (1984). Extending Rosenberg’s technique for standardizing percentage tables. *Social Forces* **62** 679–708.
- DAWID, A. P. (2002). Influence diagrams for causal modelling and inference. *Internat. Statist. Rev.* **70** 161–189.
- ELLENBERG, S. S. and HAMILTON, J. M. (1989). Surrogate endpoints in clinical trials: Cancer. *Stat. Med.* **8** 405–413.
- GELMAN, A. and HILL, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models. Analytical Methods for Social Research*. Cambridge Univ. Press, New York.
- GLASS, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher* **5** 3–8.
- GLYMOUR, M. M. and GREENLAND, S. (2008). Causal diagrams. In *Modern Epidemiology*, 3rd ed. (K. J. Rothman, S. Greenland and T. L. Lash, eds.) 183–209. Lippincott Williams & Wilkins, Philadelphia, PA.
- HAAVELMO, T. (1943). The statistical implications of a system of simultaneous equations. *Econometrica* **11** 1–12.
- HAYDUK, L., CUMMINGS, G., STRATKOTTER, R., NIMMO, M., GRYGORYEV, K., DOSMAN, D., GILLESPIE, M., PAZDERKA-ROBINSON, H. and BOADU, K. (2003). Pearl’s *d*-separation: One more step into causal thinking. *Struct. Equ. Model.* **10** 289–311.
- HECKMAN, J. J. (1979). Sample selection bias as a specification error. *Econometrica* **47** 153–161.
- HEDGES, L. V. and OLKIN, I. (1985). *Statistical Methods for Meta-Analysis*. Academic Press, Orlando, FL.
- HEISE, D. R. (1975). *Causal Analysis*. Wiley, New York.

- HERNÁN, M. A., HERNÁNDEZ-DÍAZ, S. and ROBINS, J. M. (2004). A structural approach to selection bias. *Epidemiology* **15** 615–625.
- HERNÁN, M. A. and VANDERWEELE, T. J. (2011). Compound treatments and transportability of causal inference. *Epidemiology* **22** 368–377.
- HÖFLER, M., GLOSTER, A. T. and HOYER, J. (2010). Causal effects in psychotherapy: Counterfactuals counteract overgeneralization. *Psychotherapy Research* **20** 668–679. DOI:[10.1080/10503307.2010.501041](https://doi.org/10.1080/10503307.2010.501041)
- HUANG, Y. and VALTORTA, M. (2006). Pearl's calculus of intervention is complete. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence* (R. Dechter and T. S. Richardson, eds.) 217–224. AUAI Press, Corvallis, OR.
- JOFFE, M. M. and GREENE, T. (2009). Related causal frameworks for surrogate outcomes. *Biometrics* **65** 530–538.
- KOLLER, D. and FRIEDMAN, N. (2009). *Probabilistic Graphical Models: Principles and Techniques. Adaptive Computation and Machine Learning*. MIT Press, Cambridge, MA.
- LANE, P. W. and NELDER, J. A. (1982). Analysis of covariance and standardization as instances of prediction. *Biometrics* **38** 613–621.
- LEHMANN, E. L. and CASELLA, G. (1998). *Theory of Point Estimation*, 2nd ed. Springer, New York.
- MANSKI, C. (2007). *Identification for Prediction and Decision*. Harvard Univ. Press, Cambridge, MA.
- NEYMAN, J. (1923). Sur les applications de la thar des probabilités aux expériences Agaricales: Essay des principe. English translation of excerpts by D. Dabrowska and T. Speed in *Statist. Sci.* **5** (1990) 463–472.
- OWEN, A. B. (2009). Karl Pearson's meta-analysis revisited. *Ann. Statist.* **37** 3867–3892.
- PEARL, J. (1988). *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, San Mateo, CA.
- PEARL, J. (1993). Graphical models, causality, and intervention. *Statist. Sci.* **8** 266–273.
- PEARL, J. (1995). Causal diagrams for empirical research. *Biometrika* **82** 669–710.
- PEARL, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge Univ. Press, Cambridge.
- PEARL, J. (2009a). Causal inference in statistics: An overview. *Stat. Surv.* **3** 96–146.
- PEARL, J. (2009b). *Causality: Models, Reasoning, and Inference*, 2nd ed. Cambridge Univ. Press, Cambridge.
- PEARL, J. (2011). The structural theory of causation. In *Causality in the Sciences* (P. McKay Illari, F. Russo and J. Williamson, eds.) 697–727. Clarendon Press, Oxford.
- PEARL, J. (2012a). The causal foundations of structural equation modeling. In *Handbook of Structural Equation Modeling* (R. H. Hoyle, ed.). Guilford Press, New York.
- PEARL, J. (2012b). Some thoughts concerning transfer learning, with applications to meta-analysis and data-sharing estimation. Technical Report R-387, Cognitive Systems Laboratory, Dept. Computer Science, UCLA.

- PEARL, J. (2013). Linear models: A useful “microscope” for causal analysis. *J. Causal Inference* **1** 155–170.
- PEARL, J. (2014). Trygve Haavelmo and the emergence of causal calculus. *Econometric Theory, Special Issue on Haavelmo Centennial*. Published online: 10 June 2014. DOI:[10.1017/S0266466614000231](https://doi.org/10.1017/S0266466614000231).
- PEARL, J. and BAREINBOIM, E. (2011). Transportability across studies: A formal approach. Technical Report R-372, Cognitive Systems Laboratory, Dept. Computer Science, UCLA.
- PETERSEN, M. L. (2011). Compound treatments, transportability, and the structural causal model: The power and simplicity of causal graphs. *Epidemiology* **22** 378–381.
- PRENTICE, R. L. (1989). Surrogate endpoints in clinical trials: Definition and operational criteria. *Stat. Med.* **8** 431–440.
- RICHARDSON, T. (2003). Markov properties for acyclic directed mixed graphs. *Scand. J. Stat.* **30** 145–157.
- ROBINS, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure period—Application to control of the healthy worker survivor effect. *Math. Modelling* **7** 1393–1512.
- ROBINS, J., ORELLANA, L. and ROTNITZKY, A. (2008). Estimation and extrapolation of optimal treatment and testing strategies. *Stat. Med.* **27** 4678–4721.
- ROSENTHAL, R. (1995). Writing meta-analytic reviews. *Psychological Bulletin* **118** 183–192.
- RUBIN, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educational Psychology* **66** 688–701.
- SHADISH, W. R., COOK, T. D. and CAMPBELL, D. T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*, 2nd ed. Houghton-Mifflin, Boston.
- SHPITSER, I. and PEARL, J. (2006). Identification of conditional interventional distributions. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence* (R. Dechter and T. S. Richardson, eds.) 437–444. AAAI Press, Corvallis, OR.
- SPIRITES, P., GLYMOUR, C. and SCHEINES, R. (1993). *Causation, Prediction, and Search. Lecture Notes in Statistics* **81**. Springer, New York.
- SPIRITES, P., GLYMOUR, C. and SCHEINES, R. (2000). *Causation, Prediction, and Search*, 2nd ed. *Adaptive Computation and Machine Learning*. MIT Press, Cambridge, MA.
- STROTZ, R. H. and WOLD, H. O. A. (1960). Recursive vs. nonrecursive systems: An attempt at synthesis. *Econometrica* **28** 417–427.
- TIAN, J. and PEARL, J. (2002). A general identification condition for causal effects. In *Proceedings of the Eighteenth National Conference on Artificial Intelligence* 567–573. AAAI Press/The MIT Press, Menlo Park, CA.
- VERMA, T. and PEARL, J. (1988). Causal networks: Semantics and expressiveness. In *Proceedings of the Fourth Workshop on Uncertainty in Artificial Intelligence* 352–359. Mountain View, CA. Also in *Uncertainty in AI* **4** (1990) (R. Shachter, T. S. Levitt, L. N. Kanal and J. F. Lemmer, eds.) 69–76. North-Holland, Amsterdam.
- WESTERGAARD, H. (1916). Scope and method of statistics. *Publ. Amer. Statist. Assoc.* **15** 229–276.

- WHITE, H. and CHALAK, K. (2009). Settable systems: An extension of Pearl's causal model with optimization, equilibrium, and learning. *J. Mach. Learn. Res.* **10** 1759–1799.
- WRIGHT, S. (1921). Correlation and causation. *J. Agricultural Research* **20** 557–585.
- YULE, G. U. (1934). On some points relating to vital statistics, more especially statistics of occupational mortality. *J. Roy. Statist. Soc.* **97** 1–84.

Detecting Latent Heterogeneity

Judea Pearl

Abstract

We address the task of determining, from statistical averages alone, whether a population under study consists of several subpopulations, unknown to the investigator, each responding to a given treatment markedly differently. We show that such determination is feasible in three cases: (1) randomized trials with binary treatments, (2) models where treatment effects can be identified by adjustment for covariates, and (3) models in which treatment effects can be identified by mediating instruments. In each of these cases, we provide an explicit condition which, if confirmed empirically, proves that treatment effect is not uniform but varies appreciably across individuals.

Keywords

heterogeneity, treatment on the treated, negative selection, effect modification, variable-effect bias

26.1 Introduction

Many social and health researchers are concerned with “the problem of heterogeneity,” namely, the presence of idiosyncratic groups that react differently to treatment or policies (Angrist 1998; Angrist and Krueger 1999; Elwert and Winship 2010;

Originally published in *Sociological Methods & Research* 1-20, 2015.

© The Author(s). Republished with permission.

Reprints and permission: sagepub.com/journalsPermissions.nav

Original DOI:[10.1177/0049124115600597](https://doi.org/10.1177/0049124115600597) smr.sagepub.com

Heckman and Robb 1985; Heckman, Urzua, and Vytlačil 2006; Morgan and Winship 2007, 2015; Morgan and Todd 2008; Winship and Morgan 1999; Xie, Brand, and Jann 2012). The reason is obvious. Health scientists need to know whether an approved drug is uniformly beneficial or kills some and saves more. Social scientists need to know whether those who have access to a program benefit most from the program; the alternative calls for revising recruiting policies (Brand and Xie 2010).

Heterogeneity also introduces bias if one ventures to estimate average effects using linear or constant-effect models. Indeed, the bulk of the literature on this topic is concerned with demonstrating or minimizing this bias. Such bias is of no concern, however, to students of nonparametric models where heterogeneity is assumed a priori within the model, thus protecting analysts from ever drawing conclusions that heterogeneity could invalidate.

Instead, nonparametric analysis concerns the detection of heterogeneity, if such exists, and locating its boundaries as narrowly as possible, within the granularity of the model. A straightforward way of assessing heterogeneity is to estimate the “interaction” or “effect modifying” capacity of various features of units (VanderWeele and Robins 2007). This amounts to estimating and comparing c -specific, or “conditional” effects, where c stands for a set of baseline covariates that characterize the units (Shpitser and Pearl 2006).

This article shows, however, that, under certain conditions, it is possible to assess the degree of heterogeneity in the population even without knowing the covariates C that make units differ in their response to treatment. We call this type of exogeneity “latent.”

The second section of this article will describe covariate-specific methods of detecting heterogeneity and will summarize the capabilities and limitations of these methods. The third section defines a latent heterogeneity that produces differences between treated and untreated units. The fourth section will identify three settings in which this type of heterogeneity can be detected and assessed from empirical data. These include

1. randomized trials with binary treatments (Detecting Heterogeneity in Randomized Trials subsection),
2. covariate adjustment (Detecting Heterogeneity Through Adjustment subsection), and
3. mediating instrumental variables (Detecting Heterogeneity Through Mediating Instruments subsection).

The fifth section presents a numerical example involving enrollment disparity in a job training program, where individuals possessing an unusual talent (a latent

characteristics) have higher propensity to enroll in the program and are less likely to benefit from it. The section shows how the tests developed in Detecting Heterogeneity in Randomized Trials and Detecting Heterogeneity Through Adjustment subsections can be used to detect such unusual characteristic and to assess its prevalence in the population.

Finally, Appendix A demonstrates the detection of a more drastic type of heterogeneity, where the population is composed of two distinct subpopulations undetected by any observed characteristics, only through their behavior under both observational and experimental studies (Pearl 2013).¹ Appendix B will illustrate how structural models facilitate the evaluation of counterfactuals in general and heterogeneity in particular.

26.2 Covariate-Induced Heterogeneity

If we can measure any characteristic C of individuals, a straightforward way of searching for heterogeneity is to determine if people having this characteristic respond differently from those not having it. There can of course be many group differences that escape measurement, this is unavoidable, but finding an observed characteristic accompanied by unusual effect size gives us a definitive warning that heterogeneity exists, and that its magnitude is at least equal to that found by examining C .

Formally, we can cast these considerations as follows.

26.2.1 Assessing Covariate-Induced Heterogeneity

Let C stand for any measured baseline covariate, and let $E(Y_1 - Y_0|C = c)$ stand for the causal effect² in stratum $C = c$ of C . If $E(Y_1 - Y_0|C = c)$ is identifiable (for all c), we can then estimate the effect difference:

$$D(c_i, c_j) = |E(Y_1 - Y_0|C = c_i) - E(Y_1 - Y_0|C = c_j)|, \quad (26.1)$$

for any two strata c_i and c_j of C . $D(c_i, c_j)$ gives the extent to which the effect size in group $C = c_i$ differs from that of group $C = c_j$. Further generalizing to all pairs (c_i, c_j) , we get a lower bound LB on the heterogeneity between any two labeled groups in the population:

1. This example is taken from Pearl (2013).

2. In this section, we assume a binary treatment variable $X = (0, 1)$ and an outcome variable Y with two potential outcomes, Y_0 and Y_1 , designating the hypothetical values of Y under treatment conditions $X = 0$ and $X = 1$, respectively. The logic of potential outcomes (Rosenbaum and Rubin 1983) and its equivalence to structural equations were established in (Simon and Rescher 1966; Balke and Pearl 1994a, b; Galles and Pearl 1998; Halpern 1998; Pearl 2015).

$$LB = \max_{(c_i, c_j)} D(c_i, c_j). \quad (26.2)$$

This bound extends, of course, to the case where C is a vector of measured covariates and c_i, c_j are any two instantiations of the variables in that vector. If we remove the requirement of identifiability, LB represents the best measure of heterogeneity in the population, given the crudeness of our measurements. When the identifiability requirement is imposed, LB represents the best assessment of heterogeneity, given both the crudeness of measurements and the opacity of non-experimental data. The two main problems in computing the lower bound in Equation (26.2) are, first, to find a C for which the c -specific effect is identifiable and, second, to perform the maximization in Equation (26.2) over all pairs (i, j) and all vectors C .

26.2.2 Special Cases

Three special cases of estimable covariate-based heterogeneity are worth mentioning.

C is admissible. If C is admissible,³ the c -specific effect is identified through

$$E(Y_1 - Y_0|C = c) = E(Y|X = 1, C = c) - E(Y|X = 0, C = c),$$

and $D(c_i, c_j)$ is estimable by simple regression.

C is part of an admissible set. Assume C in itself is not admissible, but we can observe a set S of covariates such that $S \cup C$ is admissible (as in Figure 26.1b and c). In such a case, the c -specific effect is still identifiable with⁴:

$$E(Y_1 - Y_0|C = c) = \sum_s [E(Y|X = 1, S = s, C = c) - E(Y|X = 0, S = s, C = c)]P(s|c).$$

Figure 26.1 depicts four models in which the c -specific effect is identifiable and two models in which it is not identifiable.

3. By “admissible,” we mean a set C of covariates that satisfy the backdoor criterion (Pearl 1993; Pearl 2009:79-81) in the causal diagram and thus permits the identification of the average causal effect by controlling for C . Admissibility entails the conditional independence $(Y_x \perp\!\!\!\perp X|C)$, sometimes called “conditional ignorability” (Rosenbaum and Rubin 1983). The backdoor criterion provides a scientific basis and a transparent test for conditional ignorability-type claims, which many researchers entrust to intuition.

4. In practice, the summation over S can be prohibitive, and propensity score weighting can be used over the unit interval $0 \leq PS \leq 1$ (Brand and Xie 2010).

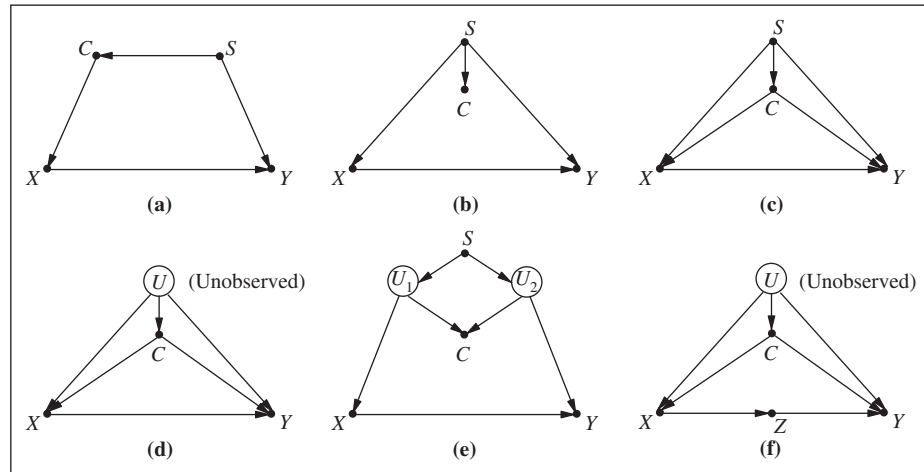


Figure 26.1 Models (a), (b), and (c) permit the identification of the c -specific effect of X on Y (by adjustment). Model (d) does not permit this identification, lacking an admissible set. Model (e) does not permit the identification of c -specific effects, even though S is admissible. Model (f) permits the identification using measurement of Z though no admissible set exists (U, U_1 and U_2 are unobserved).

Identification in the absence of admissible sets. If C is not part of an admissible set, the c -specific effect cannot be identified by adjustment. A typical example is given in Figure 26.1d. Since U is unobserved, the confounding path $X \leftarrow U \rightarrow Y$ remains open even if we adjust for C . However, the measurement of other variables in the model may nevertheless permit the identification of $E(Y_1 - Y_0 | C = c)$ by other methods, and the bound LB can be estimated accordingly. An example is given in Figure 26.1f, where $E(Y_1 - Y_0 | C = c)$ is identifiable through the front-door estimator (Pearl 1995, see also Detecting Heterogeneity Through Mediating Instruments subsection) by virtue of measuring an intermediate variable Z . A complete characterization of models that permit the identification of c -specific effects is given by Shpitser and Pearl (2006).

C excluded from all admissible sets. An intriguing pattern of heterogeneity is described in Figure 26.1e. Here S is an admissible set, but if we add C to S , admissibility is destroyed. This occurs because C is a collider, so conditioning on C would open the path $X \leftarrow U_1 \rightarrow C \leftarrow U_2 \rightarrow Y$ in violation of the backdoor condition. This means that, even if C is observed, we cannot identify the c -specific effects (of X on Y) and, therefore, we cannot assess whether units falling in different strata of C differ in their response to X . Adjustment for c_i or c_j , be it with or without S , would tell us nothing about the causal effects in those strata and would thus

prevent us from using the comparisons described in the subsection on Assessing Covariate-Induced Heterogeneity, Equation (26.1).

Note that model (e) is statistically indistinguishable from model (c), implying that no statistical test, however clever, can determine whether a given set $\{S,C\}$ of covariates is admissible. This includes sensitivity analysis, which is often presumed to provide evidence for ignorability or admissibility.

26.3 Latent Heterogeneity between the Treated and Untreated

So far, the aim of the analysis has been to find two subgroups $C = c_i$ and $C = c_j$ with unequal effect sizes, where C was an observed baseline characteristic of individuals. In this section, we abandon this requirement and seek “latent heterogeneity,” namely, heterogeneity that is not present in any baseline covariate but stems from unknown origin and manifests itself in effect differences between the treated and untreated groups.

26.3.1 Two Types of Confounding

The potential for detecting such heterogeneity was unveiled in the analyses of [Winship and Morgan \(1999\)](#) and [Xie et al. \(2012\)](#) who decomposed the *average treatment effect ATE* into several components⁵:

$$\begin{aligned} ATE &= E(Y_1 - Y_0) = E(Y|X = 1) - E(Y|X = 0) \\ &\quad - [E(Y_0|X = 1) - E(Y_0|X = 0)] - (ETT - ETU)/P(X = 0), \end{aligned}$$

where ETT and ETU are the average effect of treatment on the treated and untreated respectively,⁶ that is:

$$\begin{aligned} ETT &= E(Y_1 - Y_0|X = 1), \\ ETU &= E(Y_1 - Y_0|X = 0). \end{aligned}$$

They observed that the bias,

$$Bias = E(Y|X = 1) - E(Y|X = 0) - ATE,$$

5. This decomposition follows from the consistency rule: $E(Y_1|X = 1) = E(Y|X = 1)$, $E(Y_0|X = 0) = E(Y|X = 0)$. It was first proposed in sociology by [Winship and Morgan \(1999:667\)](#) in a paper that raised awareness for the importance of treatment-effect heterogeneity. Emphasis on ETT and ETU was introduced earlier in econometrics by Heckman and his coworkers ([Heckman 1992](#); [Heckman and Robb 1986](#)).

6. [Xie et al. \(2012\)](#) used D for treatment and $TT - TUT$ instead of $ETT - ETU$. In contrast, [Morgan and Winship \(2015\)](#) use $ATT - ATC$. Here, we use X for treatment, consistent with theoretical analyses in [Shpitser and Pearl \(2009\)](#), where the acronym ETT was used, and a necessary and sufficient condition for identifying ETT was developed.

is made up of two components with distinct characteristics. The first is $[E(Y_0|X = 1) - E(Y_0|X = 0)]$ and the second is $ETT - ETU$. The former is not a causal effect but merely a difference in output (Y) between two groups under the same “no-treatment” regime. The latter, on the other hand, represents difference in treatment effects of two groups, the treated and the untreated, and would be nonzero only if the two groups respond differently to treatment, thus exhibiting heterogeneity.⁷

Xie et al. called the former type-I bias and the latter type-II bias, whereas Morgan and Winship (2007:46-48) called them *baseline bias* and *differential treatment effect bias*. We will shorten the labels to read *baseline* and *variable-effect* biases, respectively. To understand the two types of biases, think about two groups, one with high Y that is aggressively selected for treatment, and one with low Y , which is rarely selected for treatment. There will definitely be a bias in estimating ATE , even if all units have the same treatment effect. Now think about two other groups, both achieving the same Y under no treatment, but one is sensitive to X and one is not. If the second is more likely to select treatment, a bias is generated solely by the sensitivity difference between the two groups.

26.3.2 Separating Fixed-Effect from Variable-Effect Bias

To convince ourselves that baseline and variable-effect biases, as defined earlier, indeed capture fixed-effect and variable-effect subpopulations, respectively, we evaluate their corresponding expressions in a linear model with an interaction term. The model is shown in Figure 26.2 and represents the structural equations:

$$\begin{aligned}
 y &= \beta x + \gamma z + \delta xz + \varepsilon_1 \\
 x &= \alpha z + \varepsilon_2 \\
 z &= \varepsilon_3,
 \end{aligned}$$

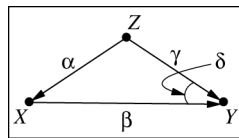


Figure 26.2 A linear model with interaction, demonstrating baseline and variable-effect biases. The former is proportional to $\gamma\alpha$ and independent of δ ; the latter is proportional to $\delta\alpha$ and independent of γ , reflecting effect variability.

7. Heckman et al. (2006) called this difference *essential heterogeneity*.

where the disturbances ε_1 , ε_2 , and ε_3 are assumed to be mutually independent. Indeed, for variable-effect bias, we obtain⁸:

$$ETT - ETU = \alpha\delta(x' - x)^2,$$

whereas for baseline bias, we have:

$$E(Yx|X = x') - E(Yx|X = x) = \gamma\alpha(x' - x).$$

(x and x' are two arbitrary levels of the treatment.) This is exactly the decomposition we expect; the former captures the bias introduced through the interaction term δ (representing variable effect), whereas the latter represents the bias that would prevail in the linear (or fixed-effect) case, without that interaction.

Note also the $ETT - ETU$ vanishes when $\alpha = 0$. Thus, not every effect heterogeneity is detected through the difference $ETT - ETU$. When interactions are strong (i.e., high δ) we certainly have appreciable heterogeneity between units with high Z and units with low Z . However, this heterogeneity will remain undetected, and it will not be revealed through the difference $ETT - ETU$, unless Z also affects the treatment assignment X .

26.4 Three Ways of Detecting Heterogeneity

The interesting feature in the preceding analysis is that the decomposition into fixed-effect and variable-effect components can be defined counterfactually, without resorting to a specific model or a specific covariate set. This means that whenever we can identify ETT and ETU , we can also obtain an indication of heterogeneity, regardless of whether we can name or observe the covariates responsible for the heterogeneity. Moreover, even in cases where auxiliary measurements are needed for identifying ETT and ETU , the graphical theory of ETT (Shpitser and Pearl 2009) can guide us in the assessment of heterogeneity by (26.1) selecting the right set of measurements and (26.2) obtaining the right estimands for ETT and ETU .

The three classical cases where ETT can be identified are as follows:

1. The treatment is binary, and $E(Y_1)$ and $E(Y_0)$ are identifiable by some method (e.g., randomized trials).
2. The treatment is arbitrary, and $E(Y_x)$ is identifiable (for all x) by adjustment for an admissible set of covariates.
3. ATE is identified through mediating instruments.

The following subsections deal separately with each of these cases.

8. These expressions follow directly from the structural definition of counterfactuals (Pearl 2009:98) as defined in Equation (26.12). A complete derivation is given in Appendix B.

26.4.1 Detecting Heterogeneity in Randomized Trials

It is well known that, when treatment is binary, ETT and ETU are identified whenever $E(Y_0)$ and $E(Y_1)$ are identified (Pearl 2009:396-97). Moreover, the relation between these quantities is given by:

$$\begin{aligned} ETT &= E(Y_1 - Y_0|X = 1) \\ &= E(Y|X = 1) - [E(Y_0) - E(Y|X = 0)(1 - p)]/p \\ ETU &= E(Y_1 - Y_0|X = 0) \\ &= [E(Y_1) - E(Y|X = 1)p]/(1 - p) - E(Y|X = 0), \end{aligned}$$

where $p = P(X = 1)$.⁹

We conclude that in a (binary) randomized clinical trial, where $E(Y_0)$ and $E(Y_1)$ are estimable empirically, the difference $ETT - ETU$ is estimable as well and is given by:

$$ETT - ETU = [E(Y|X = 1) - E(Y_1)]/(1 - p) + [E(Y|X = 0) - E(Y_0)]/p. \quad (26.3)$$

Likewise, the size of the baseline bias is identifiable from clinical trials and is given by:

$$E(Y_0|X = 1) - E(Y_0|X = 0) = [E(Y_0) - E(Y|X = 0)]/p. \quad (26.4)$$

This means that, based on pretrial and posttrial data, we can estimate the heterogeneity bias that exists in the population prior to randomization, and we can accomplish this without measuring any covariate whatsoever.

This result might appear surprising at first; how can we possibly detect the existence of individual variations among units when we have only population data? Upon further reflection, however, we note that $ETT - ETU$ does not represent the degree of heterogeneity in the population but rather that portion of heterogeneity that exhibits preferential selection to treatment. Additionally, we are not entirely justified in claiming that we accomplish this assessment without measuring *any* covariate. The treatment itself serves as a measured covariate in our case, since it is a proxy for those factors that affect the choice of treatment.

While these explanations mitigate the surprise, the point remains that effect heterogeneity is not entirely shielded from empirical scrutiny, even when we only have population data. Whenever experimental findings reveal a nonzero

9. These expressions can readily be derived by noting that $E(Y_0|X = 0) = E(Y|X = 0)$ and writing: $E(Y_0) = E(Y_0|X = 1)p + E(Y|X = 0)(1 - p)$. For nonbinary treatments, ETT is not expressible in terms of $E(Y_0)$ and $E(Y_1)$.

$ETT - ETU$, one can categorically state that heterogeneity exists in the population, that is, there exist at least two groups whose treatment effects differ from one another.

The analysis also tells us which combination of observational and experimental data would compel us to conclude that the population consists of at least two disparate groups. In particular, Equation (26.3) implies that whenever we observe the inequality:

$$P(X = 1)[E(Y|X = 1) - E(Y_1)] \neq P(X = 0)[E(Y|X = 0) - E(Y_0)], \quad (26.5)$$

we can be assured that the population is marred by heterogeneity, and, in such cases, a systematic exploration may be undertaken to unveil its underlying sources. This is not a trivial result by any means; it is in fact counterintuitive and should be considered a victory of formal counterfactual analysis. The fifth section presents numerical examples of such findings and Appendix A provides an example where Equation (26.5) returns equality despite rampant heterogeneity.

Sander Greenland suggested (personal communication, January 24, 2015) that heterogeneity in randomized trials is related to the issue debated by Fisher versus Neyman about the appropriate nulls to test. Fisher advocated the strict (point) null $Y_1 = Y_0$ for all units (which led to his famous exact test); in contrast, Neyman advocated the much weaker mean null $E(Y_1) = E(Y_0)$, which allows arbitrarily extensive heterogeneity, ostensibly on the grounds that nothing finer could be discerned in a randomized experiment (Greenland 1991).

Equation (26.5) casts this debate in a new setting. While Fisher's exact null cannot be distinguished from Neyman's mean null in a pure randomized experiment, such distinction is feasible when we have a combination of randomized and observational data. In fact, the inequality in Equation (26.5) can be regarded as a testable condition for rejecting Fisher's null hypothesis.

The fifth section and Appendix A present models where Neyman's mean null holds, $E(Y_1) = E(Y_0)$, as well as inequality in Equation (26.5), thus rejecting Fisher's sharp null. The same test can be applied when the outcome distribution under treatment is identical to the outcome distribution for control, a case where conventional approaches to testing heterogeneity fail (Ding 2014; Greenland 1991).

26.4.2 Detecting Heterogeneity Through Adjustment

The second case where ETT and ETU are identified is when an admissible set Z of covariates can be measured, yielding (see note 2) the adjustment formula:

$$E(Y_x) = \sum_z E(Y|x, z)P(z), \quad (26.6)$$

where x is any treatment level, not necessarily one or zero. It can be further shown that if Z is admissible, the expression for $E(Y_x|x')$ can be identified as well (Shpitser and Pearl 2009), and is given by:

$$E(Y_x|x') = \sum_z E(Y|x, z)P(z|x'). \quad (26.7)$$

(Shpitser and Pearl 2009). It is almost the same as the adjustment Equation (26.6), save for using $P(z|x')$ as a weighting function, instead of $P(z)$.¹⁰

Accordingly, we can write the difference $ETT - ETU$ as:

$$\begin{aligned} ETT - ETU &= E(Y_{x'} - Y_x|X = x') - E(Y_{x'} - Y_x|X = x) \\ &= \sum_z [E(Y|X = x', z) - E(Y|X = x, z)][P(z|X = x') - P(z|X = x)] \end{aligned} \quad (26.8)$$

and thus establish an explicit and general formula for the detectable part of variable-effect heterogeneity.¹¹

When the set Z is large, the estimation of Equation (26.8) can be enhanced using propensity score adjustment. But aside from providing a powerful estimation method in sparse data studies, the use of propensity scores does not add to the discussion of identification (Pearl 2009:348-52).

An objection might be raised to classifying the heterogeneity detected by Equation (26.8) as latent when, in fact, it could only be uncovered using a set Z of observed covariates. The justification rests on the realization that the treated-untreated heterogeneity, $ETT - ETU$, is a property of the population, not of the set Z chosen to uncover it. Z serves merely as an auxiliary tool for uncovering $ETT - ETU$; it does not affect its value. Moreover, $ETT - ETU$ represents a new species of heterogeneity, unrelated to those induced by the strata of Z (see the subsection on Special Cases). To witness, Equation (26.8) shows that the heterogeneity between the treated and untreated groups may be many times larger than that induced by any two strata of Z . For a trivial, albeit contrived example, let Z take on integer values $z = 1, 2, \dots, k$, and let:

$$E(Y|X = x', z) - E(Y|X = x, z),$$

10. This difference accounts for the modified Horvitz-Thompson weights required for estimating ETT and ETU by regression (Morgan and Winship 2015:231).

11. Morgan and Todd (2008) recognized the fact that ETT and ETU are estimable (using weighted regression) whenever conditional ignorability holds. Equation (26.8) extends their analysis by providing an explicit formula for $ETT - ETU$, applicable whenever a set Z of covariates is observed that is deemed admissible for identifying ATE . (Note that identifying ATE , in itself, is insufficient.) Brand and Halaby (2005) used bootstrapping methods to determine whether the difference between the ETT and the ETU is significant.

be positive for even values of z and negative for odd values. If we now let the difference $P(z|X = x') - P(z|X = x)$ be positive for even values and negative for odd values of z , $ETT - ETU$ increases indefinitely as k increases, while the effect difference between any two strata of Z remains bounded. We also note, somewhat counter-intuitively, that the treated-untreated heterogeneity ($ETT - ETU$) vanishes within each stratum $Z = z$ of an admissible set Z , while the overall difference $ETT - ETU$ need not be zero. The reason is that ETT and ETU invoke different weighing functions in averaging over the values of z ; $P(z|X = x')$ is invoked in the former and $P(z|X = x)$ in the latter.¹²

26.4.3 Detecting Heterogeneity Through Mediating Instruments

Identification by adjustment requires modeling assumptions that researchers may not be prepared to make. Attempting to circumvent this requirement, some researchers have advocated the use of instrumental variables, which appears to require milder assumptions (Angrist and Pischke 2010; Pearl 2015). Aside from the fact that good instruments are hard to come by and that the choice of instruments often requires strong modeling assumptions, identification through instruments suffers from a fundamental limitation in that it is effective only in linear (or pseudo-linear) models, and in nonparametric models, can only identify local effects, sometimes called *LATE* (Angrist, Imbens, and Rubin 1996; Brand and Thomas 2013).

Fortunately, the use of mediating instruments overcomes these limitations and identifies causal effects in nonparametric models even in the presence of unknown confounders. The method of mediating instruments, also known as “the front-door criterion” (Pearl 1995) is depicted in Figure 26.3 and assumes the availability

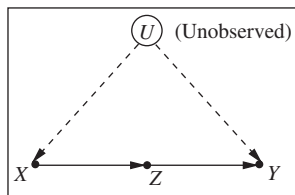


Figure 26.3 A model in which variable Z acts as a mediating instrument for identifying the causal effect of X on Y in the presence of unknown or unobserved confounders (U).

12. This is an interesting variant of Simpson’s paradox that surfaces when the aggregation of data results in sign reversal of all statistical associations (Blyth 1972; Simpson 1951). However, in the standard exposition of Simpson’s paradox, the signs of all causal effects remain unaltered (Pearl 2009:180-82; 2014). Here we witness a causal, not associational relationship that is present in the combined population and is absent in each and every subpopulation.

of covariates Z that intercept all directed paths from treatment (X) to outcome (Y).¹³ Moreover, the graphical theory of *ETT* teaches us that both *ETT* and *ETU* are identifiable in the model of Figure 26.3 and can be obtained from the estimand:

$$E(Y_x|X = x') = \sum_z E(Y|z, x')P(z|x), \quad (26.9)$$

where x and x' are any two levels of the treatment (Shpitser and Pearl 2009).

Remarkably, this expression is almost identical to the one obtained through adjustment for confounders Z , Equation (26.7), save for exchanging x and x' . Moreover, and in contrast to identification by randomized experiment, this estimand remains valid for nonbinary treatments as well.

Accordingly, the estimand for the heterogeneous component of the bias becomes identical to that of Equation (26.8):

$$\begin{aligned} ETT - ETU &= E(Y_{x'} - Y_x|X = x') - E(Y_{x'} - Y_x|X = x) \\ &= \sum_z [E(Y|X = x', z) - E(Y|X = x, z)][P(z|X = x') - P(z|X = x)], \end{aligned} \quad (26.10)$$

with $X = x'$ representing the treatment level received and $X = x$ a comparison reference. Likewise, the expression for the baseline component of the bias becomes:

$$E(Y_{x'}|X = x') - E(Y_x|X = x) = \sum_z [E(Y|z, x') - E(Y|z, x)]P(z|x). \quad (26.11)$$

We are now in possession of simple expressions for both the heterogeneous and homogeneous parts of the bias. These expressions enable us to decompose the bias into its heterogeneous and homogeneous parts without any reference to the latent confounders (U), which may remain unknown or unnamed. Whereas detection by randomized trials requires physical control, and is limited to binary treatments, and detection through ordinary adjustment requires an admissible set of deconfounders, the method of mediating instruments gives us a general way of assessing the impact of homogeneous versus heterogeneous mechanisms on the observed bias without knowing the actual mechanisms involved.

26.5 Example: Heterogeneity in Recruitment

A government is funding a job training program aimed at getting jobless people back into the workforce. A pilot randomized experiment shows that the program

13. For application of the front-door criterion in the social sciences, see Chalak and White (2012) and Morgan and Winship (2007, 2015).

is effective; a higher percentage of people were hired among those trained than among the untrained. As a result, the program is approved, and a recruitment effort is launched to encourage enrollment among the unemployed.

A study conducted a year later reveals that the hiring rate among the trained is even higher than in the randomized study. Still, critics claim that the program is a waste of tax payers' money because, while the program was somewhat successful in the experimental study, where participants were chosen at random, there is no proof that the program accomplishes its mission among those recruited for enrollment. Those enrolled, so the critics say, are more intelligent, more resourceful, and more socially connected than the eligibles who did not enroll, and would have found a job regardless of the training. The population is not homogeneous, the critics claim; the informed who are first to enroll draw little benefit from the program, while the weak and uninformed who could truly benefit from it were not aggressively recruited.

In order to assess the extent to which the $ETT - ETU$ test can detect the presence of such heterogeneity, we will simulate the hiring process assuming two types of individuals, "informed" and "uninformed." Let $Z = 1$ stand for the class of informed individuals, for whom the chances of hiring after training is only 10 percentage higher than without training, 0.9 versus 0.8. Let $Z = 0$ stand for the class of uninformed individuals, for whom the chances of hiring after training are 70 percent higher than without training, 0.8 versus 0.1. We will assume that the propensity for enrollment among the informed, q_2 , is higher than that among the uninformed, q_1 , that is, $q_2 - q_1 = P(X = 1|Z = 1) - P(X = 1|Z = 0) > 0$.

Since we are dealing with a binary treatment, we can assess the magnitude of $ETT - ETU$ using Equation (26.3) without measuring any covariates. We rely solely on $\{E(Y_1), E(Y_0)\}$, which are estimable from the experimental study, and $\{E(Y|X = 1), E(Y|X = 0)\}$, which are estimable from the observational study, and reflect the current recruitment policy. The plots in Figure 26.4 depict the difference $ETT - ETU$ as a function of r , the percentage of informed individuals in the population, with each curve representing a fixed enrollment disparity $q_2 - q_1$.

In generating these plots, we assume a model similar in structure to the one in Figure 26.2, with Z being the only confounder between X and Y . We further assume the following parameters:

$$E[Y|X = 1, Z = 1] = 0.9$$

$$E[Y|X = 0, Z = 1] = 0.8$$

$$E[Y|X = 1, Z = 0] = 0.8$$

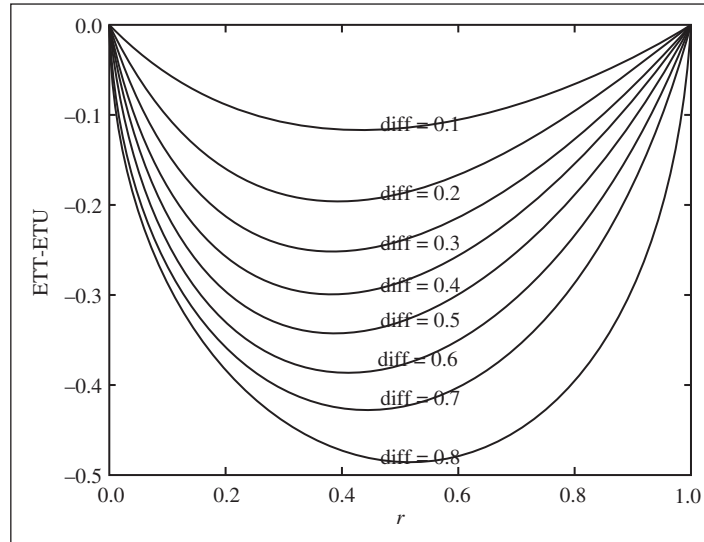


Figure 26.4 $ETT - ETU$ versus r for different levels of enrollment disparity, $q_2 - q_1$.

$$E[Y|X = 0, Z = 0] = 0.1$$

$$q_1 = P(X = 1|Z = 0) = 0.1.$$

We see that $ETT - ETU$ is negative, indicating loss of opportunity due to misdirected recruiting policy, with those in the program benefitting less from it than (potentially) those who are not in it. The higher the enrollment discrepancy $q_2 - q_1$ between the informed and the uninformed, the more negative the difference $ETT - ETU$.

We further see that the difference $ETT - ETU$ becomes zero when the population becomes homogeneous, at $r = 0$ or $r = 1$, with the slopes at these two points measuring the sensitivity of program effectiveness to the presence of heterogeneous individuals. Plots such as those in Figure 26.2 provide valuable information about the nature and magnitude of the heterogeneity observed. For example, if in a randomized experiment we observe the difference $ETT - ETU = -0.3$ (through Equation (26.3)), we can then infer that, if the propensity difference $q_2 - q_1$ is lower than 0.5, the proportion r must lie between 0.20 and 0.62. The larger the difference $q_2 - q_1$, the wider the bounds for r .

26.6 Conclusions

This article explores ways of uncovering the presence of effect heterogeneity without knowing the factors that may produce it. This possibility was shown to be

realizable in the three most common designs in which the *ATE* can be estimated: (1) randomized experiments, (2) covariate adjustment, and (3) mediating instruments. The only exceptions in these three designs are randomized experiments with nonbinary treatments and models in which *ATE* is identified and *ETT* is not. Such models can be recognized using the graphical theory of *ETT* (Shpitser and Pearl 2009), which provides a complete set of conditions for the identification of *ETT* and *ETU* from modeling assumptions.

In all three cases that allow for the detection of latent heterogeneity, we have derived explicit conditions that, if observed in practice, behoove us to conclude that subpopulations exist that differ in their response to treatment. These conditions can also serve to assess, albeit roughly (in the form of lower bounds), the magnitude of the heterogeneity detected.

Acknowledgments

I am indebted to Jennie Brand and Stephen Morgan for calling my attention to the sociological literature on heterogeneity and commenting on earlier versions of the manuscript. Subsequently, this article benefitted from discussions with Felix Elwert and Sander Greenland. I thank Ang Li for generating the plots of Figure 26.4.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported in parts by grants from NSF #IIS-1249822 and ONR #N00014-13-1-0153 and #N00014-10-1-0933.

References

- Angrist, J. D., G. Imbens, and D. Rubin. 1996. "Identification of Causal Effects Using Instrumental Variables (with Comments)." *Journal of the American Statistical Association* 91: 444-72.
- Angrist, J. D. 1998. "Estimating the Labor Market on Voluntary Military Service Using Social Security Date on Military Applicants." *Econometrica* 66:249-88.
- Angrist, J. D. and A. B. Krueger. 1999. "Handbook of Labor Economics." Pp. 1277-366 in *Causality: Statistical Perspectives and Applications*, 1st ed., vol. 3, edited by O. Ashenfelter and D. Card. Amsterdam, the Netherlands: Elsevier.

- Angrist, J. D. and J.-S. Pischke. 2010. "The Credibility Revolution in Empirical Economics: How Better Research Design Is Taking the Con out of Econometrics." *Journal of Economic Perspectives* 24:3-30.
- Balke, A. and J. Pearl. 1994a. "Counterfactual Probabilities: Computational Methods, Bounds, and Applications." Pp. 46-54 in *Uncertainty in Artificial Intelligence 10*, edited by R. L. de Mantaras and D. Poole. San Mateo, CA: Morgan Kaufmann.
- Balke, A. and J. Pearl. 1994b. "Probabilistic Evaluation of Counterfactual Queries." Pp. 230-37 in *Proceedings of the Twelfth National Conference on Artificial Intelligence*, vol. I, edited by B. Hayes-Roth and R. E. Korf. Menlo Park, CA: MIT Press.
- Blyth, C. 1972. "On Simpson's Paradox and the Sure-thing Principle." *Journal of the American Statistical Association* 67:364-66.
- Brand, J. E. and C. N. Halaby. 2005. "Regression and Matching Estimates of the Effects of Elite College Attendance on Educational and Career Achievement." *Social Science Research* 35:749-70.
- Brand, J. E. and J. S. Thomas. 2013. "Causal Effect Heterogeneity." Pp. 189-213 in *Handbook of Causal Analysis for Social Research*, chap. 11, edited by S. L. Morgan. Dordrecht, the Netherlands: Springer.
- Brand, J. E. and Y. Xie 2010. "Who Benefits Most from College? Evidence for Negative Selection in Heterogeneous Economic Returns to Higher Education." *American Sociological Review* 75:273-302.
- Chalakov, K. and H. White. 2012. "An Extended Class of Instrumental Variables for the Estimation of Causal Effects." *Canadian Journal of Economics* 44:1-31.
- Ding, P. 2014. "A Paradox from Randomization-based Causal Inference." Tech. rep., Harvard University, Cambridge, MA. arXiv:1402.0142v3.
- Elwert, F. and C. Winship. 2010. "Effect Heterogeneity and Bias in Main-effects-only Regression Models." Pp. 327-36 in *Heuristics, Probability and Causality: A Tribute to Judea Pearl*, edited by R. Dechter, H. Geffner, and J. Halpern. Milton Keynes, U.K.: College Publications.
- Galles, D. and J. Pearl. 1998. "An Axiomatic Characterization of Causal Counterfactuals." *Foundation of Science* 3:151-82.
- Greenland, S. 1991. "On the Logical Justification of Conditional Tests for Two-by-two Contingency Tables." *The American Statistician* 45:248-51.
- Halpern, J. 1998. "Axiomatizing Causal Reasoning." Pp. 202-10 in *Uncertainty in Artificial Intelligence*, edited by G. Cooper and S. Moral. San Francisco, CA: Morgan Kaufmann; *Journal of Artificial Intelligence Research* 12:17-37, 2000.
- Heckman, J. 1992. "Randomization and Social Policy Evaluation." Pp. 201-30 in *Evaluations: Welfare and Training Programs*, edited by C. Manski and I. Garfinkle. Cambridge, MA: Harvard University Press.
- Heckman, J. and R. Robb. 1985. "Alternative Methods for Evaluating the Impact of Interventions." Pp. 156-245 in *Longitudinal Analysis of Labor Market Data*, edited by J. Heckman and B. Singer. New York: Cambridge University Press.

- Heckman, J. and R. Robb. 1986. "Alternative Methods for Solving the Problem of Selection Bias in Evaluating the Impact of Treatments on Outcomes." Pp. 63-107 in *Drawing Inference from Self Selected Samples*, edited by H. Wainer. New York: Springer-Verlag.
- Heckman, J., S. Urzua, and E. Vytlacil. 2006. "Understanding Instrumental Variables in Models with Essential Heterogeneity." *The Review of Economics and Statistics* 88:389-432.
- Morgan, S. L. and C. Winship. 2007. *Counterfactuals and Causal Inference: Methods and Principles for Social Research (Analytical Methods for Social Research)*. New York: Cambridge University Press.
- Morgan, S. L. and C. Winship. 2015. *Counterfactuals and Causal Inference: Methods and Principles for Social Research (Analytical Methods for Social Research)*. 2nd ed. New York: Cambridge University Press.
- Morgan, S. L. and J. J. Todd. 2008. "A Diagnostic Routine for the Detection of Consequential Heterogeneity of Causal Effects." *Sociological Methodology* 38:231-81.
- Pearl, J. 1993. "Comment: Graphical Models, Causality, and Intervention." *Statistical Science* 8:266-69.
- Pearl, J. 1995. "Causal Diagrams for Empirical Research." *Biometrika* 82:669-710.
- Pearl, J. 2009. *Causality: Models, Reasoning, and Inference*. 2nd ed. New York: Cambridge University Press.
- Pearl, J. 2013. "The Curse of Free-will and the Paradox of Inevitable Regret." *Journal of Causal Inference* 1:255-57.
- Pearl, J. 2014. "Understanding Simpson's Paradox." *The American Statistician* 68:8-13.
- Pearl, J. 2015. "Trygve Haavelmo and the Emergence of Causal Calculus." *Econometric Theory* 31:152-79. Special issue on Haavelmo Centennial.
- Rosenbaum, P. and D. Rubin. 1983. "The Central Role of Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70:41-55.
- Shpitser, I. and J. Pearl. 2006. "Identification of Conditional Interventional Distributions." Pp. 437-44 in *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, edited by R. Dechter and T. Richardson. Corvallis, OR: AUAI Press.
- Shpitser, I. and J. Pearl. 2009. "Effects of Treatment on the Treated: Identification and Generalization." Pp. 514-21 in *Proceedings of the Twenty-fifth Conference on Uncertainty in Artificial Intelligence*, edited by J. Bilmes and A. Ng. Corvallis, OR: AUAI Press.
- Simon, H. and N. Rescher. 1966. "Cause and Counterfactual." *Philosophy and Science* 33: 323-40.
- Simpson, E. 1951. "The Interpretation of Interaction in Contingency Tables." *Journal of the Royal Statistical Society, Series B* 13:238-41.
- VanderWeele, T. and J. Robins. 2007. "Four Types of Effect Modification: A Classification Based on Directed Acyclic Graphs." *Epidemiology* 18:561-68.
- Winship, C. and S. L. Morgan. 1999. "The Estimation of Causal Effects from Observational Data." *Annual Review of Sociology* 25:659-706.

Xie, Y., J. E. Brand, and B. Jann. 2012. "Estimating Heterogeneous Treatment Effects with Observational Data." *Sociological Methodology* 42:314-47.

Author Biography

Judea Pearl is Chancellor's professor of computer science and statistics at UCLA. He is a graduate of the Technion, Israel, and joined the faculty of UCLA in 1970, where he currently directs the Cognitive Systems Laboratory and conducts research in artificial intelligence, human cognition and philosophy of science. Pearl has authored three books, (*Heuristics* (1983), *Probabilistic Reasoning* (1988) and *Causality* (2000, 2009)) and winner of the London School of Economics Lakatos Award. He is a member of the National Academy of Sciences, and a fellow of the cognitive science society and the Association for the Advancement of Artificial Intelligence. In 2012, he won the Technion's Harvey Prize and the ACM Alan Turing Award.

26.A Appendix A (An Extreme Case of Latent Heterogeneity)¹⁴

The example below demonstrates a case in which the bias is zero, the average causal effect is zero and, yet, heterogeneity is high and can be detected by Equation (26.5), using no modeling assumptions.

A study was conducted to determine which of two schools, *A* or *B*, has a more effective educational program. 200 randomly selected students underwent a randomized trial and were randomly assigned to the two schools, 100 to each. Another group of 200 (randomly selected) students were allowed to choose schools on their own; 100 selected *A* and 100 *B*. After a year of study, students were tested in a uniform, state run exam, and data showed the following:

100% of the *A*-choosing students failed the state exam

100% of the *B*-choosing students failed the state exam

50% of the *A*-randomized students failed the state exam

50% of the *B*-randomized students failed the state exam

It appears that, when given a choice, students tend to pick the school that is worse for them, which is strange but explainable. Suppose school *A* deemphasized math and *B* deemphasized history, while the state exam demands proficiency in both math and history. If students choose schools by the area of their strength, then free choice amounts to a license to neglect one of the required subjects,

14. This example is taken from (Pearl, 2012).

which is a ticket to failure. Random assignment would force at least 50% of the students to study an area of weakness, which may explain the 50% success rate in the randomized groups.

From the data available, and letting $X = 1$ and $X = 0$ stand for “School A chosen” and “School B chosen,” respectively, we can infer the following findings:

$$p = \frac{1}{2}, \quad E(Y|X = 1) = 0, \quad E(Y|X = 0) = 0$$

$$E(Y_1) = \frac{1}{2}, \quad E(Y_0) = \frac{1}{2}$$

Accordingly we have:

$$\begin{aligned} \text{Bias} &= E[(Y|X = 1) - E(Y|X = 0)] - [E(Y_1) - E(Y_0)] \\ &= 0 - 0 - \left(\frac{1}{2} - \frac{1}{2}\right) \\ &= 0 \end{aligned}$$

$$\begin{aligned} \text{Baseline Bias} &= E(Y_0|X = 1) - E(Y_0|X = 0) \\ &= [E(Y_0) - E(Y|X = 0)]/p \\ &= \left(\frac{1}{2} - 0\right)2 \\ &= 1 \end{aligned}$$

$$\begin{aligned} \text{Variable-effect Bias} &= (ETT - ETU)(1 - p) \\ &= [E(Y|X = 1) - E(Y_1)]p/(1 - p) + [E(Y|X = 0) - E(Y_0)] \\ &= \left(0 - \frac{1}{2}\right) + \left(0 - \frac{1}{2}\right) \\ &= -1 \end{aligned}$$

We conclude that a substantial effect-heterogeneity exists in the population. In fact, the bias is composed of two components of equal magnitude and opposite sign. This result is not surprising given that our population is composed indeed of two distinct subpopulations, indexed by school preference, which have two different treatment effects. Those who prefer school B have clearly different benefit from A vs. B as compared to those who prefer school A; the former would pass the exam, the latter would fail.

It is also interesting, at this point, to examine models in which latent heterogeneity is rampant, yet remains undetected by the difference $ETT - ETU$. Such models are discussed in (Pearl, 2009, pp. 35–6), which can be adapted to the story above by assuming that Z (students school preference) is totally independent of X

(the school actually attended). In such an environment, the two groups will still exhibit the disparate treatment effects, but the difference $ETT - ETU$ will be zero, because the relationship between X and Y is not confounded.

26.B Appendix B (Assessing Heterogeneity in Structural Equation Models)

In this Appendix, I first define counterfactuals in terms of structural equation models, and then illustrate how this definition facilitates the detection of heterogeneity in the linear model discussed in Section 26.3.2. The definition is fundamental to the understanding of counterfactuals in general, and for that reason, I will first introduce the method and then solve the example in minute details. The solution will demonstrate the role of structural models in defining and evaluating counterfactuals.

26.B.1 The Structural Origin of Counterfactuals

At the center of the definition lies a model M consisting of a set of equations that represents the investigator's perception of reality. M consists of two sets of variables, U and V (exogenous and endogenous), and a set F of equations that determine how values are assigned to each variable $V_i \in V$. Thus for example, the equation

$$v_i = f_i(v, u)$$

describes a physical process by which Nature *examines* the current values, v and u , of all variables in V and U and accordingly *assigns* variable V_i the value $v_i = f_i(v, u)$. The variables in U are considered "exogenous," namely, background conditions for which no explanatory mechanism is encoded in model M . Every instantiation $U = u$ of the exogenous variables corresponds to defining a "unit," or a "situation" in the model, and uniquely determines the values of all variables in V . Therefore, if we assign a probability $P(u)$ to U , it defines a probability function $P(v)$ on V . The probabilities on U and V can best be interpreted as the proportion of the population with a particular combination of values on U and/or V .

The basic counterfactual entity in structural models is the sentence: " Y would be y had X been x in situation $U = u$," denoted $Y_x(u) = y$, where Y and X are any variables in V . The key to interpreting counterfactuals is to treat the subjunctive phrase "had X been x " as an instruction to make a minimal modification in the current model, so as to ensure the antecedent condition $X = x$. Such a minimal modification amounts to replacing the equation for X with a constant x , which may be thought of as an external intervention $do(X = x)$, not necessarily by a human

experimenter, that imposes the condition $X = x$. This replacement permits the constant x to differ from the actual value of X (namely $f_x(v, u)$) without rendering the system of equations inconsistent, thus allowing all variables, exogenous as well as endogenous, to serve as antecedents.

Letting M_x stand for a modified version of M , with the equation(s) of X replaced by $X = x$, the formal definition of the counterfactual $Y_x(u)$ reads:

$$Y_x(u) \triangleq Y_{M_x}(u). \quad (26.12)$$

In words: The counterfactual $Y_x(u)$ in model M is defined as the solution for Y in the “surgically modified” submodel M_x .

This definition, first proposed in (Balke and Pearl, 1994a, b) was recently dubbed the “First Law of causal inference” (Pearl, 2015) due to its universality, and because it treats counterfactuals as an intrinsic property of reality rather than a byproduct of a specific experimental design. Simon and Rescher (1966) came close to this definition but, lacking the “wiping out” operator, could not reconcile the contradiction that evolves when an observation $X = x'$ clashes with the antecedent $X = x$ of the counterfactual Y_x . Galles and Pearl (1998) and Halpern (1998) have given a complete axiomatization of structural counterfactuals, embracing both recursive and non-recursive models (see also Pearl, 2009, Chapter 7). They showed that the axioms governing recursive structural counterfactuals are identical to those used in the potential outcomes framework, hence the two systems are logically identical – a theorem in one is a theorem in the other. This means that relying on structural models as a basis for counterfactuals does not impose additional assumptions beyond those routinely invoked by potential outcome practitioners.

$P(u)$ induces a well defined probability distribution on V , $P(v)$. As such, it not only defines the probability of any single counterfactual, also assigns joint distribution of all conceivable counterfactuals, including those that may not be observed. Thus the probability of the Boolean combination, “ $Y_x = y$ AND $Z_{x'} = z$ ” for variables Y and Z in V and two different values of X , x and x' , is well-defined even though it is impossible for both outcomes to be simultaneously observed as $X = x$ and $X = x'$ cannot be concurrently true.

In general, the probability of the counterfactual sentence $P(Y_x = y|e)$, where e is any information about an individual, can be computed by the 3-step process:

Step 1 (abduction): Update the probability $P(u)$ to obtain $P(u|e)$.

Step 2 (action): Replace the equations corresponding to variables in set X by the equations $X = x$.

Step 3 (prediction): Use the modified model to compute the probability of $Y = y$.

In temporal metaphors, Step 1 explains the past (U) in light of the current evidence e ; Step 2 bends the course of history (minimally) to comply with the hypothetical antecedent $X = x$; finally, Step 3 predicts the future (Y) based on our new understanding of the past and our newly established condition, $X = x$.

26.B.2 Illustration

To demonstrate the power of this definition, let us compute the latent heterogeneity $ETT - ETU$ for the interaction model discussed in Section 26.3.2. The model (shown in Figure 26.2) represents the structural equation model:

$$\begin{aligned} M : \quad Y &= \beta X + \gamma Z + \delta XZ + \epsilon_1 \\ X &= \alpha Z + \epsilon_2 \\ Z &= \epsilon_3. \end{aligned}$$

The modified model M_x , representing the intervention $X = x$, is given by

$$\begin{aligned} M_x : \quad Y &= \beta X + \gamma Z + \delta xZ + \epsilon_1 \\ X &= x \\ Z &= \epsilon_3. \end{aligned}$$

Let $X = x$ represent the treatment administered and $X = x'$ the level that X attains under natural, no-treatment conditions. We first compute the conditional counterfactual $E(Y_x|X = x')$ which appears in the expressions of ETT and ETU

$$\begin{aligned} ETT &= E[Y_x - Y_{x'}|X = x] \\ ETU &= E[Y_x - Y_{x'}|X = x']. \end{aligned}$$

Since Y_x is equal to the solution for Y in the mutilated model M_x , we have

$$\begin{aligned} E[Y_x|X = x'] &= E[\beta x + \gamma Z + \delta xZ + \epsilon_1|X = x'] \\ &= \beta x + (\gamma + \delta x)E[Z|X = x'] \end{aligned}$$

where we make use of the orthogonality assumption $\epsilon_1 \perp\!\!\!\perp X$. Further assuming standardized variables (i.e., zero mean and unit variance) we have $E[Z|X = x'] = \alpha x'$, which leads to

$$E[Y_x|X = x'] = \beta x + (\gamma + \delta x)\alpha x'.$$

Accordingly, the effect of treatment on the treated is given by

$$\begin{aligned} ETT &= E[Y_x - Y_{x'}|X = x] \\ &= E[Y|X = x] - E[Y_{x'}|X = x] \end{aligned}$$

$$\begin{aligned} &= \beta x + \alpha \gamma x + \alpha \delta x^2 - [\beta x' + (\gamma + \delta x')\alpha x] \\ &= (\beta + \alpha \delta x)(x - x'). \end{aligned}$$

In a similar fashion we obtain

$$\begin{aligned} ETU &= E[Y_x - Y_{x'} | X = x'] \\ &= (\beta + \alpha \delta x')(x - x') \end{aligned}$$

and finally:

$$ETT - ETU = \alpha \delta (x - x')^2,$$

which confirms the result stated in Section [26.3.2](#).

PART

**CONTRIBUTED
ARTICLES**



On Pearl's Hierarchy and the Foundations of Causal Inference

Elias Bareinboim (Columbia University),
Juan D. Correa (Columbia University),
Duligur Ibeling (Stanford University),
Thomas Icard (Stanford University)

Abstract

Cause-and-effect relationships play a central role in how we perceive and make sense of the world around us, how we act upon it, and ultimately, how we understand ourselves. Almost two decades ago, computer scientist Judea Pearl made a breakthrough in understanding causality by discovering and systematically studying the “Ladder of Causation,” a framework that highlights the distinct roles of seeing, doing, and imagining. In honor of this landmark discovery, we name this the Pearl Causal Hierarchy (PCH). In this chapter, we develop a novel and comprehensive treatment of the PCH through two complementary lenses: one logical-probabilistic and another inferential-graphical. Following Pearl's own presentation of the hierarchy, we begin by showing how the PCH organically emerges from a well-specified collection of causal mechanisms (a structural causal model, or SCM). We then turn to the logical lens. Our first result, the Causal Hierarchy Theorem (CHT), demonstrates that the three layers of the hierarchy almost always separate in a measure-theoretic sense. Roughly speaking, the CHT says that data at one layer virtually always underdetermines information at higher layers. As in most practical settings the scientist does not have access to the precise form of the underlying causal mechanisms—only to data generated by them with respect to some of the PCH's layers—this motivates us to study inferences within the PCH through the graphical lens. Specifically, we explore a set of methods known as causal inference that enable inferences bridging the PCH's layers given a partial

specification of the SCM. For instance, one may want to infer what would happen had an intervention been performed in the environment (second-layer statement) when only passive observations (first-layer data) are available. We introduce a family of graphical models that allows the scientist to represent such a partial specification of the SCM in a cognitively meaningful and parsimonious way. Finally, we investigate an inferential system known as do-calculus, showing how it can be sufficient, and in many cases necessary, to allow inferences across the PCH's layers. We believe that connecting with the essential dimensions of human experience as delineated by the PCH is a critical step toward creating the next generation of artificial intelligence (AI) systems that will be safe, robust, human-compatible, and aligned with the social good.

27.1 Introduction

Causal information is deemed highly valuable and desirable along many dimensions of the human endeavor, including science, engineering, business, and law. The ability to learn, process, and leverage causal information is arguably a distinctive feature of *Homo sapiens* when compared to other species, perhaps one of the hallmarks of human intelligence [Penn and Povinelli 2007]. Pearl argued for the centrality of causal reasoning eloquently in his most recent book [Pearl and Mackenzie 2018, p. 1], for instance: “Some tens of thousands of years ago, humans began to realize that certain things cause other things and that tinkering with the former can change the latter... From this discovery came organized societies, then towns and cities, and eventually the science and technology-based civilization we enjoy today. All because we asked a simple question: Why?”

Given the centrality of causation throughout so many aspects of human experience, we would naturally like to have a formal framework for encoding and reasoning with cause-and-effect relationships. Interestingly, the 20th century saw other instances in which an intuitive, ordinary concept underwent mathematical formalization before entering engineering practice. As an especially notable example, it may be surprising to readers outside computer science and related disciplines to learn that the notion of *computation* itself was only semi-formally understood up until the 1920s. Following the seminal work of mathematician and philosopher Alan Turing, among others, multiple breakthroughs ensued, including the very emergence of the modern computer, passing through the theory and foundations of computer science, and culminating in the rich and varied technological advances we enjoy today.

We feel it is appropriate in this special edition dedicated to Judea Pearl, a Turing awardee himself, to recognize a similar historical development in the discipline of causality. The subject was studied in a semi-formal way for centuries

[[Hume 1739, 1748](#), [von Wright 1971](#), [Mackie 1980](#)], to cite a few prominent references, and Pearl, his collaborators, and many others helped to understand and formalize this notion. Following this precise mathematization, we now see a blossoming of developments and rapid expansion toward applications.

What was the crucial development that spawned such dramatic progress on this centuries-old problem? One critical insight, tracing back at least to the British empiricist philosophers, is that the causal mechanisms behind a system under investigation are not generally observable, but they do produce observable traces (“data,” in modern terminology).¹ That is, “reality” and the data generated by it are fundamentally distinct. This dichotomy has been prominent at least since Pearl’s seminal *Biometrika* paper [[Pearl 1995](#)], and received central status and comprehensive treatment in his longer treatise [[Pearl 2000](#)]. This insight naturally leads to two practical desiderata for any proper framework for causal inference, namely:

1. The causal mechanisms underlying the phenomenon under investigation should be accounted for—indeed, formalized—in the analysis.
2. This collection of mechanisms (even if mostly unobservable) should be formally tied to its output: the generated phenomena and corresponding datasets.

This intuitive picture is illustrated in Figure 27.1(a). One of the main goals of this chapter is to make this distinction crisp and unambiguous, translating these two desiderata into a formal framework, and uncovering its consequences for the practice of causal inference.

Regarding the first requirement, the underlying reality (“ground truth”) that is our putative target can be naturally represented as a collection of causal mechanisms in the form of a mathematical object called a *structural causal model* (SCM) [[Pearl 1995, 2000](#)], to be introduced in Section 27.2. In many practical settings, it may be challenging, even impossible, to determine the specific form of the underlying causal mechanisms, especially when high-dimensional, complex phenomena are involved and humans are present in the loop.² Nevertheless, we ordinarily

1. For instance, Locke famously argued that when we observe data, we cannot “so much as guess, much less know, their manner of production” [[Locke 1690](#), Essay IV]. Hume maintained a similarly skeptical stance, stating that “nature has kept us at a great distance from all her secrets, and has afforded only the knowledge of a few superficial qualities of objects; while she conceals from us those powers and principles, on which the influence of these objects entirely depends” [[Hume 1748](#), section 4.16]. See [de Pierris \[2015\]](#) for a discussion.

2. At the same time, many of the natural sciences, most prominently physics and chemistry, will often purport to determine the underlying causal mechanisms quite precisely, under strict experimental conditions.

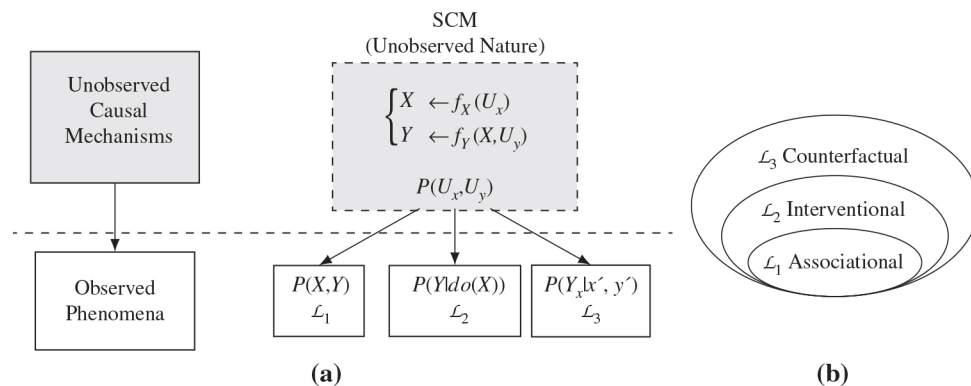


Figure 27.1 (a) Collection of causal mechanisms (or SCM) generating certain observed phenomena (qualitatively different probability distributions). (b) PCH's containment structure.

presume that these causal mechanisms are there regardless of our practical ability to discover their form, shape, and specific details.

Regarding the second requirement, Pearl further noted something very basic and fundamental, namely, that each collection of causal mechanisms (i.e., SCM) induces a causal hierarchy (or “ladder of causation”), which highlights qualitatively different aspects of the underlying reality. We fondly name this the Pearl Causal Hierarchy (PCH, for short), for he was the first to identify and study it systematically [Pearl 1995, 2000, Pearl and Mackenzie 2018]. The hierarchy consists of three layers (or “rungs”) encoding different concepts: the associational, the interventional, and the counterfactual, corresponding roughly to the ordinary human activities of seeing, doing, and imagining, respectively [Pearl and Mackenzie 2018, chapter 27]. Knowledge at each layer allows reasoning about different classes of causal concepts, or “queries.” Layer 1 deals with purely “observational,” factual information. Layer 2 encodes information about what would happen, hypothetically speaking, were some intervention to be performed, namely, effects of actions. Finally, Layer 3 involves queries about what would have happened, counterfactually speaking, had some intervention been performed, given that something else in fact occurred (possibly conflicting with the hypothetical intervention). The hierarchy establishes a useful classification of concepts that might be relevant for a given task, thereby also classifying formal frameworks in terms of the questions that they are able to represent and, ideally, answer.

27.1.1 Roadmap of the Chapter

Against this background, we start in Section 27.2 by showing how the PCH naturally emerges from an SCM, formally characterizing the layers by means of symbolic

logical languages, each of which receives a straightforward interpretation in an SCM. Thus, as soon as one admits that a domain of interest can be represented by an SCM (whether or not we, as an epistemological matter, know much about it), the hierarchy of causal concepts already exists.³ In Section 27.3, we prove that the PCH is strict for almost-all SCMs (Theorem 27.1), in a technical sense of “almost-all” (Figure 27.1(b)).⁴ It follows (Corollary 27.1) that it is *generically impossible* to draw higher-layer inferences using only lower-layer information, a result known informally in the field under the familiar adage: “no causes-in, no causes-out” [Cartwright 1989].

In the second part of the chapter (Section 27.4), we acknowledge that in many practical settings our ability to interact with (observe and experiment on) the phenomenon of interest is modest at best, and inducing a reasonable, fully specified SCM is essentially hopeless.⁵ Virtually all approaches to causal inference, therefore, set for themselves a more restricted target, operating under the less-stringent condition that only partial knowledge of the underlying SCM is available. The problem of causal inference is thus to perform inferences across layers of the hierarchy from a partial understanding of the SCM. Technically speaking, if one has Layer-1 type of data, for example, collected through random sampling, and aims to infer the effect of a new intervention (Layer-2 type of query), we show that the problem is not always solvable.

Departing from these impossibility results, we develop a framework that can parsimoniously and efficiently encode knowledge (viz., structural constraints) necessary to perform this general class of inferences. In particular, we move beyond Layer-1 type constraints (conditional independences) and investigate structural constraints that live in Layer 2. We use these constraints to define a new family

3. This is despite skepticism that has been expressed in the literature about meaningfulness of one layer of the hierarchy or another; cf., for example, Maudlin [2019] on Layer 2, and Dawid [2000] on Layer 3.

4. Hierarchies abound in logic and computer science, particularly those pertaining to computational resources, with prominent examples being the Chomsky–Schützenberger hierarchy [Chomsky 1959] and its probabilistic variant (see Icard [2020]), or the polynomial time complexity hierarchy [Stockmeyer 1977]. Such hierarchies delimit what can be computed given various bounds on computational resources. Perhaps surprisingly, the Pearl hierarchy is orthogonal to these hierarchies. If one’s representation language is only capable of encoding queries at a given layer, no amount of time or space for computation—and no amount of data either—will allow making inferences at higher layers.

5. Of course, if we have been able to induce the structural mechanisms themselves—as may be feasible in some of the sciences, for example, molecular biology or Newtonian physics—we can simply “read off” any causal information we like by computing it directly or, for instance, by simulating the corresponding mechanisms.

of graphical models called *causal Bayesian networks* (CBNs), which are composed of a pair, a graphical model, and a collection of observational and interventional distributions. Against this backdrop, we provide a novel proof of *do-calculus* [Pearl 1995] based strictly on Layer 2 semantics. We then show how the graphical structure bridges the layers of the PCH; one may be able to draw inferences at a higher layer from a combination of partial knowledge of the underlying structural model, in the form of a causal graph, and data at lower layers. We conclude and summarize this chapter in Section 27.5.

27.1.2 Notation

We now introduce the notation used throughout this chapter. Single random variables are denoted by (non-boldface) uppercase letters X and the range (or possible values) of X is written as $\text{Val}(X)$. Lowercase x denotes a particular element in this range, $x \in \text{Val}(X)$. Boldfaced uppercase \mathbf{X} denotes a collection of variables, $\text{Val}(\mathbf{X})$ their possible joint values, and boldfaced lowercase \mathbf{x} a particular joint realization $\mathbf{x} \in \text{Val}(\mathbf{X})$. For example, two independent fair coin flips are represented by $\mathbf{X} = \{X_1, X_2\}$, $\text{Val}(X_1) = \text{Val}(X_2) = \{0, 1\}$, $\text{Val}(\mathbf{X}) = \{(0, 0), \dots, (1, 1)\}$, with $P(x_1) = P(x_2) = \sum_{x_2} P(x_1, x_2) = \sum_{\mathbf{x}(X_1)=x_1} P(\mathbf{x}) = 1/2$.

27.2 Structural Causal Models and the Causal Hierarchy

We build on the language of SCMs to describe the collection of mechanisms underpinning a phenomenon of interest. Essentially, any causal inference can be seen as an inquiry about these mechanisms or their properties, in some way or another. We will generally dispense with the distinction between the underlying system and its SCM.

Each SCM naturally defines a qualitative hierarchy of concepts, described as the “ladder of causation” in Pearl and Mackenzie [2018], which we have been calling the PCH (Figure 27.1). Following Pearl’s presentation, we label the layers (or rungs, or levels) of the hierarchy *associational*, *interventional*, and *counterfactual*. The concepts of each layer can be described in a formal language and correspond roughly to distinct notions within human cognition. Each of these allows one to articulate, with mathematical precision, qualitatively different types of questions regarding the observed variables of the underlying system; for some examples, see Table 27.1.

SCMs provide a flexible formalism for data-generating models, subsuming virtually all of the previous frameworks in the literature. In the sequel, we formally define SCMs and then show how a fully specified model underpins the concepts in the PCH.

Table 27.1 Pearl’s Causal Hierarchy

	Layer (Symbolic)	Typical Activity	Typical Question	Example	Machine Learning
\mathcal{L}_1	Associational $P(y x)$	Seeing	What is? How would seeing X change my belief in Y ?	What does a symptom tell us about the disease?	Supervised/ Unsupervised Learning
\mathcal{L}_2	Interventional $P(y do(x),c)$	Doing	What if? What if I do X ?	What if I take aspirin, will my headache be cured?	Reinforcement Learning
\mathcal{L}_3	Counterfactual $P(y_x x',y')$	Imagining	Why? What if I had acted differently?	Was it the aspirin that stopped my headache?	

Definition 27.1 Structural Causal Model (SCM)

An SCM \mathcal{M} is a 4-tuple $\langle \mathbf{U}, \mathbf{V}, \mathcal{F}, P(\mathbf{U}) \rangle$, where

- \mathbf{U} is a set of background variables, also called exogenous variables, that are determined by factors outside the model;
- \mathbf{V} is a set $\{V_1, V_2, \dots, V_n\}$ of variables, called endogenous, that are determined by other variables in the model—that is, variables in $\mathbf{U} \cup \mathbf{V}$;
- \mathcal{F} is a set of functions $\{f_1, f_2, \dots, f_n\}$ such that each f_i is a mapping from (the respective domains of) $U_i \cup Pa_i$ to V_i , where $U_i \subseteq \mathbf{U}$, $Pa_i \subseteq \mathbf{V} \setminus V_i$, and the entire set \mathcal{F} forms a mapping from \mathbf{U} to \mathbf{V} . That is, for $i = 1, \dots, n$, each $f_i \in \mathcal{F}$ is such that

$$v_i \leftarrow f_i(pa_i, u_i), \tag{27.1}$$

that is, it assigns a value to V_i that depends on (the values of) a select set of variables in $\mathbf{U} \cup \mathbf{V}$; and

- $P(\mathbf{U})$ is a probability function defined over the domain of \mathbf{U} . ■

Each SCM can be seen as partitioning the variables involved in the phenomenon into sets of exogenous (unobserved) and endogenous (observed) variables, respectively, \mathbf{U} and \mathbf{V} . The exogenous ones are determined “outside” of the model and their associated probability distribution, $P(\mathbf{U})$, represents a summary of the

state of the world outside the phenomenon of interest. In many settings, these variables represent the *units* involved in the phenomenon, which correspond to elements of the population under study, for instance, patients, students, and customers. Naturally, their randomness (encoded in $P(\mathbf{U})$) induces variations in the endogenous set \mathbf{V} .

Inside the model, the value of each endogenous variable V_i is determined by a causal process, $v_i \leftarrow f_i(pa_i, u_i)$, that maps the exogenous factors U_i and a set of endogenous variables Pa_i (so-called parents) to V_i . These causal processes—or mechanisms—are assumed to be invariant unless explicitly intervened on (as defined later in the section).⁶ Together with the background factors, they represent the data-generating process according to which Nature assigns values to the endogenous variables in the study.

Henceforth, we assume that \mathbf{V} and its domain are finite,⁷ and that the model is acyclic (sometimes known as *recursive*).⁸ A structural model is *Markovian* if the exogenous parent sets U_i, U_j are independent whenever $i \neq j$. Here, we will allow for the sharing of exogenous parents and for arbitrary dependences among the exogenous variables, which means that, in general, the SCM need not be Markovian. This wider class of models is called *semi-Markovian*. For concreteness, we provide a simple SCM next.

Example 27.1 Consider a game of chance described through the SCM $\mathcal{M}^1 = \langle \mathbf{U} = \{U_1, U_2\}, \mathbf{V} = \{X, Y\}, \mathcal{F}, P(U_1, U_2) \rangle$, where

$$\mathcal{F} = \begin{cases} X & \leftarrow U_1 + U_2 \\ Y & \leftarrow U_1 - U_2 \end{cases}, \quad (27.2)$$

and $P(U_i = k) = 1/6$, $i = 1, 2$, $k = 1, \dots, 6$. In other words, this structural model represents the setting in which two dice are rolled but only the sum (X) and the difference (Y) of their values are observed. Here, $\text{Val}(X) = \{2, \dots, 12\}$ and $\text{Val}(Y) = \{-5, \dots, 0, \dots, 5\}$. ■

6. It is possible to conceive an SCM as “a high-level abstraction of an underlying system of differential equations” [Schölkopf 2019], which under relatively mild conditions is attainable [Rubenstein et al. 2017].

7. Much of the theory of SCMs extends straightforwardly to the infinitary setting [Ibeling and Icard 2019].

8. An SCM \mathcal{M} is said to be recursive if there exists a “temporal” order over the functions in \mathcal{F} such that for every pair $f_i, f_j \in \mathcal{F}$, if $f_i < f_j$ in the order, we have that f_i does not have V_j as an argument. In particular, this implies that choosing a unit \mathbf{u} uniquely fixes the values of all variables in \mathbf{V} . For $\mathbf{Y} \subseteq \mathbf{V}$, we write $\mathbf{Y}(\mathbf{u})$ to denote the solution of \mathbf{Y} given unit \mathbf{u} . For a more comprehensive discussion, see Galles and Pearl [1998] and Halpern [1998, 2000].

27.2.1 Pearl Hierarchy, Layer 1—Seeing

Layer 1 of the hierarchy (Table 27.1) captures the notion of “seeing,” that is, observing a certain phenomenon unfold, and perhaps making inferences about it. For instance, if we observe a certain symptom, how will this change our belief in the disease? An SCM gives natural valuations for quantities of this kind (cf. equation (7.2) in Pearl [2000]), as shown next.

Definition 27.2 Layer 1 Valuation—“Observing”

An SCM $\mathcal{M} = \langle \mathbf{U}, \mathbf{V}, \mathcal{F}, P(\mathbf{U}) \rangle$ defines a joint probability distribution $P^{\mathcal{M}}(\mathbf{V})$ such that for each $\mathbf{Y} \subseteq \mathbf{V}$:⁹

$$P^{\mathcal{M}}(\mathbf{y}) = \sum_{\{\mathbf{u} \mid \mathbf{Y}(\mathbf{u})=\mathbf{y}\}} P(\mathbf{u}), \quad (27.3)$$

where $\mathbf{Y}(\mathbf{u})$ is the solution for \mathbf{Y} after evaluating \mathcal{F} with $\mathbf{U} = \mathbf{u}$. ■

This evaluation is graphically depicted in Figure 27.2(i), which represents a mapping from the external and unobserved state of the system (distributed as $P(\mathbf{U})$), to an observable state (distributed as $P(\mathbf{V})$). For concreteness, let us consider Example 27.1 again. Let the dice (exogenous variables) be $\langle U_1 = 1, U_2 = 1 \rangle$, then $\mathbf{V} = \{X, Y\}$ attain their values through \mathcal{F} as $X = 1 + 1 = 2$ and $Y = 1 - 1 = 0$. As $P(U_1 = 1, U_2 = 1) = 1/36$ and $\langle U_1 = 1, U_2 = 1 \rangle$ is the only configuration capable of producing the observed behavior $\langle X = 2, Y = 0 \rangle$, it follows that $P(X = 2, Y = 0) = 1/36$. More interestingly, consider the different dice (exogenous) configurations $\langle U_1, U_2 \rangle = \{\langle 1, 1 \rangle, \langle 2, 2 \rangle, \langle 3, 3 \rangle, \langle 4, 4 \rangle, \langle 5, 5 \rangle, \langle 6, 6 \rangle\}$, which are all compatible with $\langle Y = 0 \rangle$. As each of the \mathbf{U} 's realization happens with probability $1/36$, the event of the difference between the first and second dice being zero ($Y = 0$) occurs with probability $1/6$. Finally, what is the probability of the difference of the two dice being zero ($Y = 0$) if we know that their sum is two, that is, $P(Y = 0 \mid X = 2)$? The answer is one as the only event compatible with $\langle X = 2, Y = 0 \rangle$ is $\langle U_1 = 1, U_2 = 1 \rangle$. Without any evidence, the event ($Y = 0$) happens with probability $1/6$, yet if we know that $X = 2$, the event becomes certain (probability 1).

Many tasks throughout data sciences can be seen as evaluating the probability of certain events occurring. In the context of modern machine learning, for example, one could observe a certain collection of pixels, or features, with the goal of predicting whether it contains a dog or a cat. Consider a slightly more involved example that appears in the context of medical decision-making.

9. We will typically omit the superscript on $P^{\mathcal{M}}$ whenever there is no room for confusion, thus using P for both the distribution $P(\mathbf{U})$ on exogenous variables and the distributions $P(\mathbf{Y})$ on endogenous variables induced by the SCM.

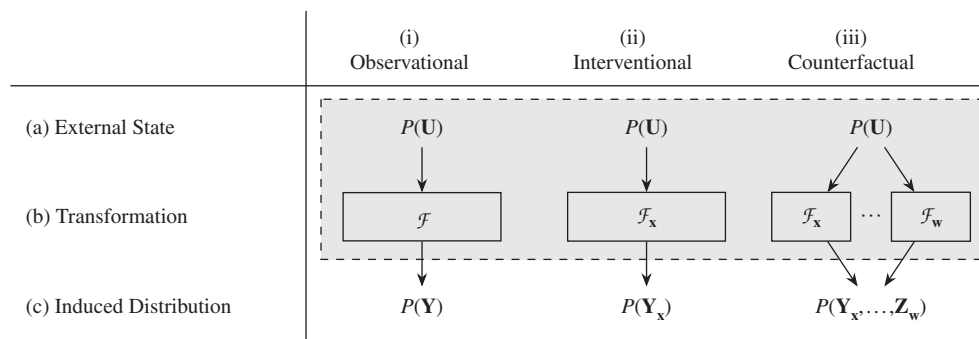


Figure 27.2 Given an SCM's initial state (i.e., population) (a), we show the different functional transformations (b) and the corresponding induced distribution (c) of each layer of the hierarchy. (i) represents the transformation (i.e., \mathcal{F}) from the natural state of the system ($P(\mathbf{U})$) to an observational world, (ii) to an interventional world (i.e., with modified mechanisms \mathcal{F}_x), and (iii) to multiple counterfactual worlds (i.e., with multiple modified mechanisms).

Example 27.2 The SCM $\mathcal{M}^2 = \langle \mathbf{V} = \{X, Y, Z\}, \mathbf{U} = \{U_r, U_x, U_y, U_z\}, \mathcal{F} = \{f_x, f_y, f_z\}, P(U_r, U_x, U_y, U_z) \rangle$, where \mathcal{F} will be specified below. The endogenous variables \mathbf{V} represent, respectively, a certain treatment X (e.g., drug), an outcome Y (survival), and the presence or not of a symptom Z (hypertension). The exogenous variable U_r represents whether the person has a certain natural resistance to the disease, and U_x, U_y, U_z are sources of variations outside the model affecting X, Y, Z , respectively. In this population, units with resistance ($U_r = 1$) are likely to survive ($Y = 1$) regardless of the treatment received. Whenever the symptom is present ($Z = 1$), physicians try to counter it by prescribing this drug ($X = 1$). While the treatment ($X = 1$) helps resistant patients (with $U_r = 1$), it worsens the situation for those who are not resistant ($U_r = 0$). The form of the underlying causal mechanisms is:

$$\mathcal{F} = \begin{cases} Z & \leftarrow \mathbb{1}_{\{U_r=1, U_z=1\}} \\ X & \leftarrow \mathbb{1}_{\{Z=1, U_x=1\}} + \mathbb{1}_{\{Z=0, U_x=0\}} \\ Y & \leftarrow \mathbb{1}_{\{X=1, U_r=1\}} + \mathbb{1}_{\{X=0, U_r=1, U_y=1\}} + \mathbb{1}_{\{X=0, U_r=0, U_y=0\}} \end{cases} . \quad (27.4)$$

Finally, all the exogenous variables are binary with $P(U_r = 1) = 0.25$, $P(U_z = 1) = 0.95$, $P(U_x = 1) = 0.9$, and $P(U_y = 1) = 0.7$.

Recall that Definition 27.2 (Equation 27.3) induces a mapping between $P(\mathbf{U})$ and $P(\mathbf{V})$, such that a query $P(Y = 1 | X = 1)$ can be evaluated from \mathcal{M} as:

$$P(Y = 1 | X = 1) = \frac{P(Y = 1, X = 1)}{P(X = 1)} = \frac{\sum_{\{\mathbf{u} | Y(\mathbf{u})=1, X(\mathbf{u})=1\}} P(\mathbf{u})}{\sum_{\{\mathbf{u} | X(\mathbf{u})=1\}} P(\mathbf{u})} = \frac{0.215}{0.29} = 0.7414, \quad (27.5)$$

which is just the ratio between the sum of the probabilities of the events in the space of \mathbf{U} consistent with the events $\langle Y = 1, X = 1 \rangle$ and $\langle X = 1 \rangle$. This means that the probability of survival given that one took the drug is higher than chance. Similarly, one could obtain other probabilistic expressions such as $P(Y = 1 | X = 0) = 0.3197$ or $P(Z = 1) = 0.2375$. One may be tempted to believe at this point that the drug has a positive effect upon comparing the probabilities $P(Y = 1 | X = 0)$ and $P(Y = 1 | X = 1)$. We shall discuss this issue next. ■

27.2.2 Pearl Hierarchy, Layer 2—Doing

Layer 2 of the hierarchy (Table 27.1) allows one to represent the notion of “doing,” that is, intervening (acting) in the world to bring about some state of affairs. For instance, if a physician gives a drug to her patient, would the headache be cured? A modification of an SCM gives natural valuations for quantities of this kind, as defined next.

Definition 27.3 Submodel—“Interventional SCM”

Let \mathcal{M} be a causal model, \mathbf{X} a set of variables in \mathbf{V} , and \mathbf{x} a particular realization of \mathbf{X} . A submodel $\mathcal{M}_{\mathbf{x}}$ of \mathcal{M} is the causal model

$$\mathcal{M}_{\mathbf{x}} = \langle \mathbf{U}, \mathbf{V}, \mathcal{F}_{\mathbf{x}}, P(\mathbf{U}) \rangle, \quad \text{where } \mathcal{F}_{\mathbf{x}} = \{f_i : V_i \notin \mathbf{X}\} \cup \{\mathbf{X} \leftarrow \mathbf{x}\}. \quad (27.6)$$

In other words, performing an external intervention (or action) is modeled through the replacement of the original (natural) mechanisms associated with some variables \mathbf{X} with a constant \mathbf{x} , which is represented by the *do*-operator.^{10,11} The impact of the intervention on an outcome variable Y is called *potential response* (cf. definition (7.1.4) in Pearl [2000]).

Definition 27.4 Potential Response

Let \mathbf{X} and \mathbf{Y} be two sets of variables in \mathbf{V} , and \mathbf{u} be a unit. The potential response

10. The idea of representing intervention through the modification of equations in a structural system appears to have first emerged in the context of Econometrics, see Haavelmo [1943], Marschak [1950], and Simon [1953]. It was then made more explicit and called “wiping out” by Strotz and Wold [1960].

11. Pearl credits his realization on the connection of this operation with graphical models to a lecture of Peter Spirtes at the International Congress on Logic, Methodology and Philosophy of Science (Uppsala, Sweden, 1991), in his words [Pearl 2000, p. 104]: “In one of his slides, Peter illustrated how a causal diagram would change when a variable is manipulated. To me, that slide of Spirtes’s—when combined with the deterministic structural equations—was the key to unfolding the manipulative account of causation (...).”

$\mathbf{Y}_x(\mathbf{u})$ is defined as the solution for \mathbf{Y} of the set of equations \mathcal{F}_x with respect to SCM \mathcal{M} (for short, $\mathbf{Y}_{\mathcal{M}_x}(\mathbf{u})$). That is, $\mathbf{Y}_x(\mathbf{u}) = \mathbf{Y}_{\mathcal{M}_x}(\mathbf{u})$. ■

An SCM gives valuation for interventional quantities (equation 7.3 Pearl [2000]) as follows:

Definition 27.5 Layer 2 Valuation—“Intervening”

An SCM $\mathcal{M} = \langle \mathbf{U}, \mathbf{V}, \mathcal{F}, P(\mathbf{U}) \rangle$ induces a family of joint distributions over \mathbf{V} , one for each intervention \mathbf{x} . For each $\mathbf{Y} \subseteq \mathbf{V}$:

$$P^{\mathcal{M}}(\mathbf{y}_x) = \sum_{\{\mathbf{u} \mid \mathbf{Y}_x(\mathbf{u}) = \mathbf{y}\}} P(\mathbf{u}). \quad (27.7)$$

The *potential response* expresses causal effects, and over a probabilistic setting it induces random variables. Specifically, Y_x denotes a random variable induced by averaging the potential response $Y_x(\mathbf{u})$ over all \mathbf{u} according to $P(\mathbf{U})$.¹² Further, note that this procedure disconnects X from any other source of “natural” variation when it follows the original function f_x (e.g., the observed (Pa_x) or unobserved (U_x) parents). This means that the variations of Y in this world would be due to changes in X (say, from 0 to 1) that occurred externally, from outside the modeled system.¹³ This, in turn, guarantees that they will be *causal*. To see why, note that all variations of X that may have an effect on Y can only be realized through variables of which X is an argument, as X itself is a constant, not affected by other variables. Indeed, the notion of an *average causal effect* can be formally written as $E(Y_{X=1}) - E(Y_{X=0})$.¹⁴

The distribution $P(\mathbf{Y}_x)$ defined in Equation (27.7) is often written $P(\mathbf{Y} \mid do(\mathbf{x}))$, and we henceforth adopt this notation in the context of PCH's second layer.¹⁵

12. The notation $Y_x(u)$ is borrowed from the potential-outcome framework of Neyman [1923] and Rubin [1974]. See Pearl [2000, section 7.4.4] for a more detailed comparison; see also Pearl and Bareinboim [2019].

13. For a discussion of what it means for these changes to arise “from outside” the system, see, for example, Woodward [2003]. Of course, in many settings this simply means the intervention is performed deliberately by an *agent* outside the system, for example, in typical reinforcement learning applications [Sutton and Barto 2018].

14. This difference and the corresponding expected values are sometimes taken as the definition of “causal effect,” see Rosenbaum and Rubin [1983]. In the structural account of causation pursued here, this quantity is not a primitive but derivable from the SCM, as all others within the PCH. To witness, note $Y_{X=1} \leftarrow f_Y(1, \varepsilon_Y)$ when $do(X = 1)$.

15. This allows researchers to use the syntax to immediately distinguish statements that are amenable to some sort of experimentation, at least in principle, from other counterfactuals that may be empirically unrealizable.

Example 27.3 Example 27.1 continued

Let us consider the same dice game but now the observer decides to misreport the sum of the two dice as 2, which can be written as submodel $\mathcal{M}_{X=2}$:

$$\mathcal{F}_{X=2} = \begin{cases} X & \leftarrow 2 \\ Y & \leftarrow U_1 - U_2, \end{cases}, \quad (27.8)$$

while $P(\mathbf{U})$ remains invariant. It can be immediately seen that $Y_{X=2}(u_1, u_2)$ is the same as $Y(u_1, u_2)$; in other words, misreporting the sum of the two dice will of course not change their difference. This, in turn, entails the following probabilistic invariance,

$$P(Y = 0 | do(X = 2)) = P(Y = 0). \quad (27.9)$$

In fact, the distribution of Y when X is fixed to two remains the same as before (i.e., $P(Y = 0 | do(X = 2)) = 1/6$). We saw in the first part of the example that knowing that the sum was two meant that, with probability one, their difference had to be zero (i.e., $P(Y = 0 | X = 2) = 1$). On the other hand, intervening on X will not change Y 's distribution (Equation 27.9); as we say, X does not have a *causal effect* on Y . ■

Example 27.4 Example 27.2 continued

Consider now that a public health official performs an intervention by giving the treatment to all patients regardless of the symptom (Z). This means that the function f_X would be replaced by the constant 1. In other words, patients do not have an option of deciding their own treatment but are compelled to take the specific drug.¹⁶ This is represented through the new modified set of mechanisms,

$$\mathcal{F}_{X=1} = \begin{cases} Z & \leftarrow \mathbb{1}_{\{U_r=1, U_z=1\}} \\ X & \leftarrow \mathbb{1} \\ Y & \leftarrow \mathbb{1}_{\{X=1, U_r=1\}} + \mathbb{1}_{\{X=0, U_r=1, U_y=1\}} + \mathbb{1}_{\{X=0, U_r=0, U_y=0\}} \end{cases}, \quad (27.10)$$

and where the distribution of exogenous variables remains the same. Note that the potential response $Y_{X=1}(\mathbf{u})$ represents the survival of patient \mathbf{u} had they been treated, while the random variable $Y_{X=1}$ describes the average population survival

16. This physical procedure is the very basis for the discipline of experimental design [Fisher 1936], which is realized through randomization of the treatment assignment in a sample of the population. In practice, the function of X, f_X , is replaced with an alternative source of randomness that is uncorrelated with any other variable in the system.

had everyone been given the treatment. Notice that for those patients who naturally received treatment ($X \leftarrow f_x(\mathbf{U}) = 1$), the natural outcome $Y(\mathbf{u})$ is equal to $Y_{X=1}(\mathbf{u})$. For this intervened model, $Y_{X=1}(\mathbf{u})$ is equal to 1 in every event where $U_r = 1$, regardless of U_z , U_x , and U_y . Then

$$P(Y = 1 | do(X = 1)) = \sum_{\{\mathbf{u} | Y_{X=1}(\mathbf{u})=1\}} P(\mathbf{u}) = \sum_{\{u_r | Y_{X=1}(u_r)=1\}} P(u_r) = P(U_r = 1) = 0.25. \quad (27.11)$$

Similarly, one can evaluate $P(Y = 1 | do(X = 0))$, which is equal to 0.4. This may be surprising as from the perspective of Layer 1, $P(Y = 1 | X = 1) - P(Y = 1 | X = 0) = 0.43 > 0$, which appears to suggest that taking the drug is helpful, having a positive effect on recovery. On the other hand, interventionally speaking, $P(Y = 1 | do(X = 1)) - P(Y = 1 | do(X = 0)) = -0.15 < 0$, which means that the drug has a negative (average) effect in the population. ■

The evaluation of an interventional distribution is a function of the modified system \mathcal{M}_x that reflects \mathcal{F}_x , which follows from the replacement of \mathbf{X} , as illustrated in Figure 27.2(ii). Even though computing observational and interventional distributions is immediate from a fully specified SCM, the distinction between Layer 1 (seeing) and Layer 2 (doing) is a central topic in causal inference, as discussed more substantively in Section 27.4.

27.2.3 Pearl Hierarchy, Layer 3—Imagining Counterfactual Worlds

Layer 3 of the hierarchy (Table 27.1) allows operationalizing the notion of “imagination” (and the closely related activities of retrospection, prospection, and other forms of “modal” reasoning), that is, thinking about alternative ways the world could be, including ways that might conflict with how the world, in fact, currently is. For instance, if the patient took the aspirin and the headache was cured, would the headache still be gone had they not taken the drug? Or, if an individual ended up getting a great promotion, would this still be the case had they not earned a PhD? What if they had a different gender? Obviously, in this world, the person has a particular gender, has a PhD, and ended up getting the promotion, so we would need a way of conceiving and grounding these alternative possibilities to evaluate such scenarios. In fact, no experiment in the world (Layer 2) will be sufficient to answer this type of question in general, despite their ubiquity in human discourse, cognition, and decision-making. Fortunately, the meaning of every term in the counterfactual layer (\mathcal{L}_3) can be directly determined from a fully specified SCM, as described in the sequel:

Definition 27.6 Layer 3 Valuation

An SCM $\mathcal{M} = \langle \mathbf{U}, \mathbf{V}, \mathcal{F}, P(\mathbf{U}) \rangle$ induces a family of joint distributions over counterfactual events $\mathbf{Y}_x, \dots, \mathbf{Z}_w$, for any $\mathbf{Y}, \mathbf{Z}, \dots, \mathbf{X}, \mathbf{W} \subseteq \mathbf{V}$:

$$P^{\mathcal{M}}(\mathbf{y}_x, \dots, \mathbf{z}_w) = \sum_{\{\mathbf{u} \mid \mathbf{Y}_x(\mathbf{u})=\mathbf{y}, \dots, \mathbf{Z}_w(\mathbf{u})=\mathbf{z}\}} P(\mathbf{u}). \quad (27.12)$$

■

Note that the left-hand side (LHS) of Equation (27.12) contains variables with different subscripts, which, syntactically, encode different counterfactual “worlds.”

Example 27.5 Example 27.2 continued

As there is a group of patients who did not receive the treatment and died ($X = 0, Y = 0$), one may wonder whether these patients would have been alive ($Y = 1$) had they been given the treatment ($X = 1$). In the language of Layer 3, this question is written as $P(Y_{X=1} = 1 \mid X = 0, Y = 0)$. This is a non-trivial question as these individuals did not take the drug and are already deceased in the actual world (as displayed after the conditioning bar, $X = 0, Y = 0$); the question is about an unrealized world and how these patients would have reacted had they been submitted to a different course of action (formally written before the conditioning bar, $Y_{X=1} = 1$). In other words, did they die because of the lack of treatment? Or would this fatal unfolding of events happen regardless of the treatment? Unfortunately, there is no conceivable experiment in which we could draw samples from $P(Y_{X=1} = 1 \mid X = 0, Y = 0)$, as these patients cannot be resuscitated and submitted to the alternative condition. This is the very essence of counterfactuals.

For simplicity, note that $P(Y_{X=1} = 1 \mid X = 0, Y = 0)$ can be written as the ratio $P(Y_{X=1} = 1, X = 0, Y = 0) / P(X = 0, Y = 0)$, where the denominator is trivially obtainable as it only involves observational probabilities (about one specific world, the factual one). The numerator, $P(Y_{X=1} = 1, X = 0, Y = 0)$, refers to two different worlds, which requires us to climb up to the third layer in order to formally specify the quantity of interest. Using the procedure dictated in Equation (27.12), we obtain

$$\begin{aligned} P(Y_{X=1} = 1 \mid X = 0, Y = 0) &= \frac{P(Y_{X=1} = 1, X = 0, Y = 0)}{P(X = 0, Y = 0)} \\ &= \frac{\sum_{\{\mathbf{u} \mid Y_{X=1}(\mathbf{u})=1, X(\mathbf{u})=0, Y(\mathbf{u})=0\}} P(\mathbf{u})}{\sum_{\{\mathbf{u} \mid X(\mathbf{u})=0, Y(\mathbf{u})=0\}} P(\mathbf{u})} = 0.0217. \end{aligned}$$

This evaluation is shown step by step in [Bareinboim et al. \[2020, appendix D\]](#), but we emphasize here that the expression in the numerator involves evaluating multiple worlds simultaneously (in this case, one factual and one related to

intervention $do(X = 1)$), as illustrated in Figure 27.2(iii). The conclusion following from this counterfactual analysis is clear: even if we had given the treatment to everyone who did not survive, only around 2% would have survived. In other words, the drug would not have prevented their deaths. Another aspect of this situation worth examining is whether the treatment would have been harmful for those who did not get it and still survived, formally written in Layer 3 language as $P(Y_{X=1} = 1 | X = 0, Y = 1)$. Following the same procedure, we find that this quantity is 0.1079, which means that about 90% of such people would have died had they been given the treatment. While a uniform policy over the entire population would be catastrophic (as shown in Example 27.4), the physicians knew what they were doing in this case and were effective in choosing the treatment for the patients who could benefit more from it. ■

There are many other counterfactual quantities implied by a structural model, for example, the previous two quantities can be combined to form the *probability of necessity and sufficiency* (PNS) [Pearl 2000, chapter 9], written as $P(y_x, y_{x'})$. The PNS encodes the extent to which a certain treatment to a particular outcome would be both necessary and sufficient. This quantity addresses a quintessential “why” question, where one wants to understand what caused a given event. Still in the purview of Layer 3, some critical applications demand that counterfactuals be nested inside other counterfactuals. For instance, consider the quantity $Y_{x, M_{x'}}$ that represents the counterfactual value of Y had X been x , and M had whatever value it would have taken had X been x' . In other words, for Y the value of X is x , while for M the value of X is x' . This type of nested counterfactuals allow us to write contrasts such as $P[Y_{x, M_x} - Y_{x, M_{x'}}]$, the so-called *indirect effect* on Y when X changes from x' to x [Pearl 2001]. The use of nested counterfactuals led to a very natural and general treatment of direct, indirect, and spurious effects, including a precise understanding of their relationship in non-linear systems [Pearl 2012, VanderWeele 2015, Zhang and Bareinboim 2018].

27.3 Pearl Hierarchy—A Logical Perspective

We have seen that each layer of the PCH corresponds to a different intuitive notion in human cognition: seeing, acting, and imagining. Table 27.1 presents characteristic questions associated with each of the layers. Layer 1 concerns questions like, “How likely is Y given that I observe X ?” Layer 2 asks hypothetical (“conditional”) questions such as, “How likely *would* Y be if one were to make X happen?” Layer 3 takes us further, allowing questions like, “Given that I observed X and Y , how likely would Y have been if X' had been true instead of X ?”

What does the difference among these questions amount to, given that an SCM answers all of them? Implicit in our presentation was a series of increasingly complex symbolic languages (Definitions 27.2, 27.5, and 27.6). Each type of question above can be phrased in one of these languages, the analysis of which reveals a logical perspective on PCH. We begin our analysis by isolating the syntax of these systems. We define languages \mathcal{L}_1 , \mathcal{L}_2 , \mathcal{L}_3 , each based on polynomials built over basic probability terms $P(\alpha)$. The only differences among them are the terms $P(\alpha)$ allowed: as we go up in the PCH, increasingly complex expressions α are allowed in the probability terms. In particular, \mathcal{L}_1 is just a familiar probabilistic logic (see Fagin et al. [1990]).

Definition 27.7 Symbolic Languages \mathcal{L}_1 , \mathcal{L}_2 , \mathcal{L}_3

Let variables \mathbf{V} be given and $\mathbf{X}, \mathbf{Y}, \mathbf{Z} \subseteq \mathbf{V}$. Each language \mathcal{L}_i , $i = 1, 2, 3$, consists of (Boolean combinations of) inequalities between polynomials over terms $P(\alpha)$, where $P(\alpha)$ is an \mathcal{L}_i term, defined as follows:

- \mathcal{L}_1 terms are those of the form $P(\mathbf{Y} = \mathbf{y})$, encoding the probability that \mathbf{Y} take on values \mathbf{y} ;
- \mathcal{L}_2 terms additionally include probabilities of *conditional* expressions, $P(\mathbf{Y}_x = \mathbf{y})$, giving the probability that variables \mathbf{Y} *would* take on values \mathbf{y} , were \mathbf{X} to have values \mathbf{x} ;
- \mathcal{L}_3 terms encode probabilities over *conjunctions* of conditional (that is, \mathcal{L}_2) expressions, $P(\mathbf{Y}_x = \mathbf{y}, \dots, \mathbf{Z}_w = \mathbf{z})$, symbolizing the joint probability that all of these conditional statements hold simultaneously. ■

For concreteness, a typical \mathcal{L}_1 sentence might be $P(X = 1, Y = 1) = P(X = 1)P(Y = 1)$. The \mathcal{L}_1 conjunction over all such combinations

$$\begin{aligned} P(X = 1, Y = 1) &= P(X = 1)P(Y = 1) \wedge P(X = 1, Y = 0) = P(X = 1)P(Y = 0) \\ &\wedge P(X = 0, Y = 1) = P(X = 0)P(Y = 1) \wedge P(X = 0, Y = 0) = P(X = 0)P(Y = 0) \end{aligned} \quad (27.13)$$

would express that X and Y are probabilistically independent if X and Y are binary variables. Of course, we would ordinarily write this simply as $P(X, Y) = P(X)P(Y)$.

In \mathcal{L}_2 we have sentences like $P(Y_{X=1} = 1) = 3/4$, which intuitively expresses that the probability of Y taking on value 1 were X to take on value 1 is $3/4$.¹⁷ As before, we could also write this as $P(Y = 1 | do(X = 1)) = 3/4$. Finally, \mathcal{L}_3 allows

17. These “conditional” expressions such as $Y_{X=1} = 1$ are familiar from the literature in conditional logic. In David Lewis’s early work on counterfactual conditionals, $Y_{X=1} = 1$ would have been written $X = 1 \boxrightarrow Y = 1$ (see Lewis [1973]). More recently, some authors have used notation from dynamic logic, $[X = 1]Y = 1$, with the same interpretation over SCMs (see, e.g., Halpern [2000]). For more discussion on the connection between the present SCM-based interpretation

statements about joint probabilities over conditional terms with possibly inconsistent subscripts (also known as antecedents in logic). For instance, $P(y_x, y'_x) \geq P(y|x) - P(y|x')$ is a statement expressing a lower bound on the PNS.¹⁸

Definition 27.7 gives the formal structure (*syntax*) of $\mathcal{L}_1, \mathcal{L}_2, \mathcal{L}_3$, but not their interpretation or meaning (*semantics*). In fact, we have already specified their meaning in SCMs via Definitions 27.2, 27.5, and 27.6. Specifically, let Ω denote the set of all SCMs over endogenous variables \mathbf{V} . Then each $\mathcal{M} \in \Omega$ assigns a real number to $P(\alpha)$ for all α at each layer, namely the value $P^{\mathcal{M}}(\alpha) \in [0, 1]$. Given such numbers, arithmetic and logic suffice to finish evaluating these languages. Thus, in each SCM \mathcal{M} , every sentence of our languages, such as Equation (27.13), comes out true or false.¹⁹ At this stage, we are ready to formally define the PCH:

Definition 27.8 Pearl Causal Hierarchy (PCH)

Let \mathcal{M}^* be a fully specified SCM. The collection of observational, interventional, and counterfactual distributions induced by \mathcal{M}^* , as delineated by languages $\mathcal{L}_1, \mathcal{L}_2, \mathcal{L}_3$ (syntax) and following Definitions 27.2, 27.5, and 27.6 (semantics), is called the Pearl Causal Hierarchy. ■

In summary, as soon as we have an SCM, the PCH is thereby well defined, in the sense that this SCM provides valuations for any conceivable quantity in these languages $\mathcal{L}_1, \mathcal{L}_2, \mathcal{L}_3$ (of associations, interventions, and counterfactuals, respectively). It therefore makes sense to ask about properties of the hierarchy for any given SCM, as well as for the class Ω of all SCMs. One substantive question is whether the PCH can be shown strict.

If we take \mathcal{L}_1 terms to involve a tacit empty intervention, that is, that $P(\mathbf{y})$ means $P(\mathbf{y}_\emptyset)$, then the formal syntax of this series of languages clearly forms a strict hierarchy $\mathcal{L}_1 \subsetneq \mathcal{L}_2 \subsetneq \mathcal{L}_3$: there are patently \mathcal{L}_2 terms that do not appear in \mathcal{L}_1 (e.g.,

and Lewis's "system-of-spheres" interpretation, we refer readers to Pearl [2000, sections 7.4.1–7.4.3] and Briggs [2012], Halpern [2013], and Zhang [2013]. A third interpretation is over (probabilistic) "simulation" programs, which under suitable conditions are equivalent to SCMs—see Ibeling and Icard [2018, 2019, 2020].

18. For details of this bound and the assumptions guaranteeing it, see Pearl [2000, theorem 9.2.10]. Formally speaking, statements such as this one involving conditional probabilities are shorthand for polynomial inequalities; in this case the polynomial inequality is $P(y_x, y'_x)P(x)P(x') + P(x', y)P(x) \geq P(x, y)P(x')$.

19. Building on the classic axiomatization for (finite) *deterministic* SCMs [Galles and Pearl 1998, Halpern 2000], the probabilistic logical languages $\mathcal{L}_1, \mathcal{L}_2, \mathcal{L}_3$ were axiomatized over probabilistic SCMs in Ibeling and Icard [2020]. The work presented in this chapter—including Definition 27.8 and Theorem 27.1 (below)—can be cast in axiomatic terms, although these results do not depend in any direct way on questions of axiomatization.

$P(y_x)$), and \mathcal{L}_3 terms that do not appear in \mathcal{L}_2 (e.g., $P(y_x, y'_{x'})$). One has the impression that each layer of the Pearl hierarchy is somehow richer or more expressive than those below it, capable of encoding information about an underlying ground truth that surpasses what lower layers can possibly express. Is this an illusion, the mere appearance of complexity, or are the concepts expressed by the layers in some way fundamentally distinct?²⁰ The sense of strictness that we would like to understand concerns the fundamental issue of logical *expressiveness*. If each language did not expressively exceed its predecessors, then in some sense our talk of causation and imagination would be no more than mere figures of speech, being fully reducible to lower layers.

What would it mean for the layers of the hierarchy *not* to be distinct? Toward clarifying this, let us call the set of all layer i (\mathcal{L}_i) statements that come out true according to some $\mathcal{M} \in \Omega$ the \mathcal{L}_i -theory of \mathcal{M} . We shall write $\mathcal{M} \sim_i \mathcal{M}'$ for $\mathcal{M}, \mathcal{M}' \in \Omega$ to mean that their \mathcal{L}_i -theories coincide, that is, that $\mathcal{M}, \mathcal{M}'$ agree on all layer i statements. Intuitively, $\mathcal{M} \sim_i \mathcal{M}'$ says that \mathcal{M} and \mathcal{M}' are indistinguishable given knowledge only of \mathcal{L}_i .

For the remainder of this section assume that the true data-generating process \mathcal{M}^* is fixed. Suppose we had that $\mathcal{M}^* \sim_2 \mathcal{M}$ implies $\mathcal{M}^* \sim_3 \mathcal{M}$ for any other SCM $\mathcal{M} \in \Omega$; that is, any SCM \mathcal{M} which agrees with \mathcal{M}^* on all \mathcal{L}_2 valuations also agrees on all of the \mathcal{L}_3 valuations.²¹ This would mean that the collection of \mathcal{L}_2 facts *fully determines* all of the \mathcal{L}_3 facts. More colloquially, if this happens, it means that we can answer any \mathcal{L}_3 question—including any counterfactual question, for example, the exact value of $P(y_x | y'_{x'})$ —merely from \mathcal{L}_2 information. For instance, simply construct any SCM \mathcal{M} with the right \mathcal{L}_2 valuation (i.e., such that $\mathcal{M} \sim_2 \mathcal{M}^*$) and read off the \mathcal{L}_3 facts from \mathcal{M} .²² In this case it would not matter that \mathcal{M} is not the true data-generating process, as any differences would not be visible even at \mathcal{L}_3 . This can

20. As a rough analogy, consider the ordinary concepts of “cardinality of the integers,” “cardinality of the rational numbers,” and “cardinality of the real numbers.” One’s first intuition may be that these are three distinct notions, and moreover that they form a kind of hierarchy: there are *strictly more* rational numbers than integers, and strictly more real numbers than rational numbers. Of course, in this instance the intuition can be vindicated in the second case but dismissed as an illusion in the first. (See, e.g., Cantorian arguments from any basic textbook in logic or CS.)

21. For readers familiar with causal inference, this can be seen as a generalization of the notion of identifiability (e.g., see Pearl [2000, definition 3.2.3]), where P represents all quantities in layer i , Q all quantities in layer j , and the set of features F_M is left unrestricted (all in the notation of Pearl [2000]). This more relaxed notion has a long history in mathematical logic, namely, Padoa’s method in the theory of definability [Beth 1956].

22. Alternatively, given the completeness results in Ibeling and Icard [2020], one could axiomatically derive any \mathcal{L}_3 statement from appropriate \mathcal{L}_2 statements.

happen in exceptional circumstances, for instance, if the functional relationship is deterministic.

An additional motivation for understanding when layers of the PCH might collapse comes from the observation that, at least in some notable cases, adding syntactic complexity does not genuinely increase expressivity. As an example, we could extend the language \mathcal{L}_3 to allow more complex expressions. We discussed nested counterfactuals earlier in this chapter (Section 27.2), namely, statements such as $P(Y_{x,Z_{x'}})$, which can also be given a natural interpretation in SCMs. Such notions are of significant interest, but it can be shown that any such statement is systematically reducible to a Layer 3 statement. (See Bareinboim et al. [2020, appendix B] for details.) That is, for any statement φ involving nested counterfactual expressions, there is an \mathcal{L}_3 statement ψ such that φ and ψ hold in exactly the same models.²³ Such a result shows that adding nested counterfactuals, while providing a useful shorthand, would not allow us to say anything about the world above and beyond what we can say in \mathcal{L}_3 . Does something similar happen with Layers 1, 2, and 3 themselves? How often might an \mathcal{L}_3 -theory completely reduce to an \mathcal{L}_2 -theory, or an \mathcal{L}_2 -theory reduce to an \mathcal{L}_1 -theory?

In light of the foregoing, we can say exactly what it means for the PCH to collapse in a given SCM \mathcal{M}^* . Note that the quantification here is over the class of all SCMs in Ω , that is, all SCMs with the same set of endogenous (i.e., observable) variables as \mathcal{M}^* :

Definition 27.9 Collapse relative to \mathcal{M}^*

Layer j of the causal hierarchy *collapses* to Layer i , with $i < j$, relative to $\mathcal{M}^* \in \Omega$ if $\mathcal{M}^* \sim_i \mathcal{M}$ implies that $\mathcal{M}^* \sim_j \mathcal{M}$ for all $\mathcal{M} \in \Omega$.²⁴ ■

The significance of the possibility of collapse cannot be overstated. To the extent that Layer 2 collapses to Layer 1, this would imply that we can draw all possible causal conclusions from mere correlations. Likewise, if Layer 3 collapses to Layer 2, this means that we could make statements about any counterfactual merely by conducting controlled experiments.

Our main result can then be stated (first, informally) as:

Theorem 27.1 Causal Hierarchy Theorem (CHT), informal version

The PCH almost never collapses. That is, for almost any SCM, the layers of the hierarchy remain distinct. ■

23. In logic, we would say that nested counterfactuals are thus *definable* in \mathcal{L}_3 (see, e.g., Beth [1956]).

24. Equivalently, there does not exist $\mathcal{M} \in \Omega$ such that $\mathcal{M}^* \sim_i \mathcal{M}$ but $\mathcal{M}^* \not\sim_j \mathcal{M}$. In other words, every layer j query can be answered with suitable layer i data.

What does *almost-never* mean? Here is an analogy. Suppose (fully specified) SCMs are drawn at random from Ω . Then, the probability that we draw an SCM relative to which PCH collapses is 0. This holds regardless of the distribution on SCMs, so long as it is smooth.

The CHT says that there will typically be causal questions that one cannot answer with knowledge and/or data restricted to a lower layer.²⁵ This can be seen as the formal grounding for the intuition behind the PCH discussed in [Pearl and Mackenzie \[2018, chapter 27\]](#):

Corollary 27.1 To answer questions at Layer i , one needs knowledge at Layer i or higher.

With this intuitive understanding of the CHT, we now state the formal version and offer an outline of the main arguments used in the proof. In order to state the theorem, note that \sim_3 is an equivalence relation on Ω , inducing \mathcal{L}_3 -equivalence classes of SCMs. Under a suitable encoding, this space of equivalence classes can be seen as a convex subset of $[0, 1]^K$, for $K \in \mathbb{N}$. This means we can put a natural (uniform) *measure* on the space of (equivalence classes) of SCMs. The theorem then states (for the complete proof and further details, we refer readers to [Bareinboim et al. \[2020, appendix A\]](#)):

Theorem 27.1 **CHT, formal version**

With respect to the Lebesgue measure over (a suitable encoding of \mathcal{L}_3 -equivalence classes of) SCMs, the subset in which any PCH collapse occurs is measure zero. ■

It bears emphasis that the CHT is a theory-neutral result in the sense that it makes only minimal assumptions and only presupposes the existence of a temporal ordering of the structural mechanisms—an assumption made to obtain unique valuations via [Definitions 27.2, 27.5, and 27.6](#).

In the remainder of this section, we would like to discuss the basic idea behind the CHT proof. There are essentially two parts to the argument: one showing that \mathcal{L}_2 almost never collapses to \mathcal{L}_1 , and the second showing that \mathcal{L}_3 almost never collapses to \mathcal{L}_2 . In both parts it suffices to identify some simple property of SCMs that we can show is *typical*, and moreover sufficient to ensure non-collapse.

In fact, Layer 2 never collapses to Layer 1: for any SCM \mathcal{M}^* there is always another SCM \mathcal{M} with the same \mathcal{L}_1 -theory but a different \mathcal{L}_2 -theory. In case there

25. The investigation of the next section will be on conditions that could allow causal inferences from lower-level data combined with graphical assumptions of the underlying SCM; see, for example, [Bareinboim and Pearl \[2016\]](#). Another common thread in the literature is structural learning: adopting arguably mild assumptions of minimality (e.g., faithfulness) one can often discover fragments of the underlying causal diagram (Layer 2) from observational data (Layer 1) [[Spirtes et al. 2001](#), [Peters et al. 2017](#)].

is any non-trivial dependence in \mathcal{M}^* , we can construct a second model \mathcal{M} with a single exogenous variable U and all endogenous variables depending only on U , such that $\mathcal{M}^* \sim_1 \mathcal{M}$ (cf. Suppes and Zanotti [1981]). On the other hand, if \mathcal{M}^* has no variable depending on any other, it is possible to induce such a dependence that, nonetheless, does not show up at Layer 1. (For full details of the argument, see Bareinboim et al. [2020, appendix A]).

The case of Layers 2 and 3 is slightly more subtle. The reason is that adding or removing arguments in the underlying functional relationships usually changes the corresponding causal effect. Here we need to show that the equations of the true \mathcal{M}^* can be perturbed in a way that it does not affect any \mathcal{L}_2 facts but does change some joint probabilities over combinations of potential responses. It turns out there are many ways to accomplish this goal; however, for the CHT we need a systematic method. One possibility—again, informally speaking—is to take two exogenous variable settings that witness two different values for some potential response, and swap these values with some sufficiently small probability (see Bareinboim et al. [2020, appendix A]). For this to work, essentially all we need is for there to be at least some non-trivial probabilistic relationship between variables. This property is quite obviously typical of SCMs. We illustrate this method with our running Example 27.2 (Example 27.7 below).

Turning now to these examples, we start with a variation of a classic construction presented by Pearl himself [Pearl 2000, section 1.4.4]. The example has been used to demonstrate the inadequacy of (causal) Bayesian networks (discussed further in the next section) for encoding counterfactual information. Here we use it to illustrate a more abstract lesson, namely, that knowing the values of higher-layer expressions generically requires knowing progressively more about the underlying SCM (Corollary 27.1).

Example 27.6 Let $\mathcal{M}^* = \langle \mathbf{U} = \{U_1, U_2\}, \mathbf{V} = \{X, Y\}, \mathcal{F}^*, P(U) \rangle$, where

$$\mathcal{F}^* = \begin{cases} X & \leftarrow U_1 \\ Y & \leftarrow U_2 \end{cases} . \quad (27.14)$$

and U_1, U_2 are binary with $P(U_1 = 1) = P(U_2 = 1) = 1/2$. Let the variable X represent whether the patient received treatment and Y whether they recovered. Evidently, $P^{\mathcal{M}^*}(x, y) = 1/4$ for all values of X, Y . In particular X, Y are independent. Now, suppose that we just observed samples from $P^{\mathcal{M}^*}$ and were confident, statistically speaking, that X, Y are probabilistically independent. Would we be justified in concluding that X has no causal effect on Y ? If the actual mechanism happened to be \mathcal{M}^* , then this would certainly be the case. However, this Layer 1 data is equally consistent with other SCMs in which Y depends strongly on X . Let \mathcal{M} be just like \mathcal{M}^* ,

except with mechanisms:

$$\mathcal{F} = \begin{cases} X & \leftarrow \mathbb{1}_{U_1=U_2} \\ Y & \leftarrow U_1 + \mathbb{1}_{X=1, U_1=0, U_2=1} \end{cases} . \quad (27.15)$$

Then $P^{\mathcal{M}^*}(X, Y) = P^{\mathcal{M}}(X, Y)$, yet $P^{\mathcal{M}^*}(Y = 1 | do(X = 1)) = 1/2$ as X does not affect Y in \mathcal{M}^* , while $P^{\mathcal{M}}(Y = 1 | do(X = 1)) = 3/4$. If \mathcal{M} were the actual mechanisms, assigning the treatment would actually improve the chance of survival. Thus, just as one cannot infer causation from correlation, one cannot always expect to infer correlation from causation.

Having internalized this lesson that correlation and causation are distinct, one might perform a randomized controlled trial and discover that all causal effects in this setting trivialize; in particular, $P(Y | do(X)) = P(Y)$ —the treatment does not affect the chance of survival at all. Suppose we observe patient S , who took the treatment and died. We might well like to know whether S 's death occurred *because of* the treatment, *in spite of* the treatment, or *regardless of* the treatment. This is a quintessentially counterfactual question: given that S took the treatment and died, what is the probability that S *would have* survived had they not been treated? We write this as $P(Y_{X=0} = 1 | X = 1, Y = 0)$, as discussed in Example 27.4. Can we infer anything about this expression from Layer 2 information (in this case, that all causal effects trivialize)? We cannot, as shown by other variations of \mathcal{M}^* , say \mathcal{M}' such that

$$\mathcal{F}' = \begin{cases} X & \leftarrow U_1 \\ Y & \leftarrow XU_2 + (1 - X)(1 - U_2) \end{cases} . \quad (27.16)$$

Like \mathcal{M} , this model reveals a dependence of Y on X . However, this is not at all visible at Layer 1 or at Layer 2; all causal effects trivialize in \mathcal{M}' as well. The dependence only becomes visible at Layer 3. In \mathcal{M}^* , we have $P^{\mathcal{M}^*}(Y_{X=0} = 1 | X = 1, Y = 0) = 0$, whereas in \mathcal{M}' we have the exact opposite pattern, $P^{\mathcal{M}'}(Y_{X=0} = 1 | X = 1, Y = 0) = 1$. These two models thus make diametrically opposed predictions about whether S *would have* survived had they not taken the treatment. In other words, the best *explanation* for S 's death may be completely different depending on whether the world is like \mathcal{M}^* or \mathcal{M}' . In \mathcal{M}^* , S would have died anyway, while in \mathcal{M}' , S would actually have survived, if only they had not been given the treatment. Needless to say, such matters can be of fundamental importance for critical practical questions, such as determining who or what is to blame for S 's death. ■

The CHT tells us that the failure of collapse witnessed in Example 27.6 is typical. However, it is worth seeing further examples to appreciate the many ways we can

take an SCM \mathcal{M}^* and find an alternative SCM \mathcal{M} that agrees at all lower layers but disagrees at higher layers.

We discuss two quite different strategies in the next example. To show that Layer 2 does not collapse to Layer 1, we actually *eliminate* the functional dependence of one variable on another—all probabilistic dependence patterns are due to common causes. More interestingly, we employ a very general method to show that Layer 3 does not collapse to Layer 2, whose efficacy is proven systematically in Bareinboim et al. [2020, lemma 2].

Example 27.7 Example 27.2 continued

For the SCM $\mathcal{M}^* = \mathcal{M}^2$ of Example 27.2, consider another model \mathcal{M} with the equation for Y replaced by a new equation $Y \leftarrow \mathbb{1}_{\{U_r=1, U_x=1, U_z=1\}} + \mathbb{1}_{\{U_r=1, U_x=0, U_z=0\}} + \mathbb{1}_{\{U_r=1, U_x=0, U_y=1, U_z=1\}} + \mathbb{1}_{\{U_r=1, U_x=1, U_y=1, U_z=0\}} + \mathbb{1}_{\{U_r=1, U_x=1, U_y=0\}}$, and everything else unchanged. It is then easy to check that $\mathcal{M}^* \sim_1 \mathcal{M}$. However, Y now no longer shows a functional dependence on X : the probabilistic dependence of Y on X is due to the common causes U_x, U_z, U_r . While in Example 27.4 we saw that $P^{\mathcal{M}^*}(Y|X) \neq P^{\mathcal{M}^*}(Y|do(X))$, here we have $P^{\mathcal{M}}(Y|X) = P^{\mathcal{M}}(Y|do(X))$. In other words, even though X does exert a causal influence on Y (assuming \mathcal{M}^* is the true data-generating process), we would not be able to infer this from observational data alone.

To show that Layer 3 does not collapse to Layer 2, consider a third model \mathcal{M}' , in which X, Y, Z all share one exogenous parent U , with $\text{Val}(U) = \{0, 1\}^4 \cup \{u_1^*, u_2^*\}$. The probability of a quadruple $\langle u_r, u_z, u_x, u_y \rangle$ in this model is simply given by the product from model \mathcal{M}^* — $P(U_r = u_r) \cdot P(U_z = u_z) \cdot P(U_x = u_x) \cdot P(U_y = u_y)$ —with one exception: for the two quadruples, $\langle 1, 1, 1, 0 \rangle$ and $\langle 1, 1, 0, 0 \rangle$, we subtract $\varepsilon = .005$ from these probabilities, and redistribute the remaining mass so that u_1^* and u_2^* each receive probability ε . This produces a proper distribution $P'(U)$. We will continue to write, for example, $U_r = u$ simply to mean that $U \neq u_1^*, u_2^*$ and the first coordinate of U is u , and similarly for U_z, U_x, U_y . The mechanisms are now:

$$\mathcal{F}' = \begin{cases} Z \leftarrow \mathbb{1}_{\{U_r=1, U_z=1\}} + \mathbb{1}_{U \in \{u_1^*, u_2^*\}} \\ X \leftarrow \mathbb{1}_{\{Z=1, U_x=1\}} + \mathbb{1}_{\{Z=0, U_x=0\}} + \mathbb{1}_{U=u_2^*} \\ Y \leftarrow \mathbb{1}_{\{X=1, U_r=1\}} + \mathbb{1}_{\{X=0, U_r=1, U_y=1\}} + \mathbb{1}_{\{X=0, U_r=0, U_y=0\}} + \mathbb{1}_{\{X=1, U \in \{u_1^*, u_2^*\}\}} \end{cases} \quad (27.17)$$

To check that the joint distributions $P^{\mathcal{M}^*}(X, Y, Z)$ and $P^{\mathcal{M}'}(X, Y, Z)$ are the same, note that the two models coincide at all exogenous settings with the exception of the two quadruples $\langle 1, 1, 1, 0 \rangle$ and $\langle 1, 1, 0, 0 \rangle$. In the first we have $Z = X = Y = 1$, and the ε -loss in probability for this possibility is corrected by the fact that $X(u_2^*) = Y(u_2^*) = Z(u_2^*) = 1$ and $P'(u_2^*) = \varepsilon$. Similarly for $\langle 1, 1, 0, 0 \rangle$ and the state $Z = 1, X = Y = 0$, which results when $U = u_1^*$. To show that $\mathcal{M}^* \sim_2 \mathcal{M}'$ is also straightforward.

However, consider the \mathcal{L}_3 expression $Y_{Z=1} = 1, Y_{Z=0} = 1$, which says that the patient would survive no matter whether hypertension was induced or prevented. For both exogenous settings $\langle 1, 1, 1, 0 \rangle$ and $\langle 1, 1, 0, 0 \rangle$, this expression is false, yet in setting u_2^* the expression is true. Hence, $P^{\mathcal{M}'}(Y_{Z=1} = 1, Y_{Z=0} = 1) = P^{\mathcal{M}^*}(Y_{Z=1} = 1, Y_{Z=0} = 1) + \varepsilon$. ■

While collapse of the layers is possible if \mathcal{M}^* is exceptional, the CHT shows that this is the exception indeed. Typical cases are similar to Examples 27.6 and 27.7, each showing a different way of perturbing an SCM to obtain a second SCM revealing non-collapse. In fact, a typical data-generating process \mathcal{M}^* encodes rich information at all three layers, and even small changes to the mechanisms in \mathcal{M}^* can have substantial impact on quantities across the hierarchy. Critically, such differences will often be visible only at higher layers in the PCH.

The lesson learned from the CHT is clear—as the layers of PCH come apart in the generic case and one cannot make inferences at one layer given knowledge at lower layers (e.g., using observational data to make interventional claims), some additional assumptions are logically necessary if one wants in general to do *causal inference*.

27.4 Pearl Hierarchy—A Graphical Perspective

All conceivable quantities from any layer of the PCH—associational, interventional, and counterfactual—are immediately computable once the fully specified SCM is known. Unfortunately, in most practical settings, it's usually hard to determine the structural model at this level of precision, and the CHT severely curtails the ability to “climb up” the PCH via lower-level data. Learning about cause-and-effect relationships is arguably one of the main goals found throughout the sciences. After all, how could causal inferences be performed?

The recognition that there are mechanisms underlying the phenomena of interest, but that we usually cannot determine them precisely, gives rise to the discipline of *causal inference* [Pearl 2000]. Virtually every approach to causal inference works under the stringent condition that only partial knowledge of the underlying SCM is available. One pervasive task is to determine the effect of an intervention—what would happen with Y were X to be intervened on and set to x , $P(Y | do(X = x))$ —from observational data, $P(X, Y)$. This constitutes a cross-layer inference where the goal is to use data from layer \mathcal{L}_1 to try to make an inference about an \mathcal{L}_2 quantity, given a partial specification of the underlying SCM (see Figure 27.3 [a–d]).

In this section, we investigate the question of what type of causal knowledge could be (1) intuitively meaningful, (2) possibly available, and (3) powerful enough to encode constraints that would allow cross-layer inferences, *as if* the SCM were

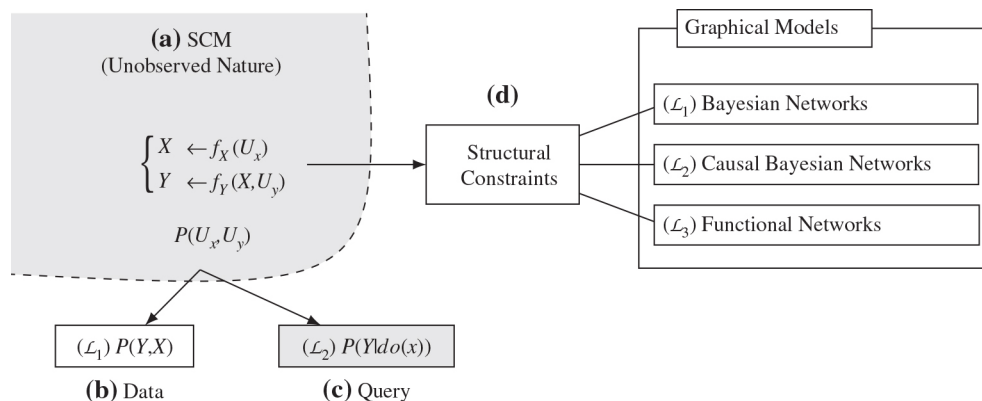


Figure 27.3 Example of Prototypical Causal Inference—on top the SCM itself, representing the unobserved collection of mechanisms and corresponding uncertainty (a); at the bottom, the different probability distributions entailed by the model (b, c); on the right side, the graphical model representing the specific constraints of the SCM (d).

itself available. A key observation useful to answer this question is that each SCM imprints specific “marks” on the distributions it generates, depicted generically in the schema in Figure 27.3(d) as *structural constraints*.

One first attempt to solve this task could be to leverage \mathcal{L}_1 -constraints, those imprinted on the observed \mathcal{L}_1 data by the unknown SCM, to make inferences about the target \mathcal{L}_2 -quantity. This is especially appealing considering that \mathcal{L}_1 data is often readily available. The signature type of constraint for \mathcal{L}_1 distributions is known as *conditional independence*, and *Bayesian Networks* (BNs) are among the most prominent formal models used to encode this type of knowledge. The example below shows that \mathcal{L}_1 constraints (and BNs) alone are insufficient to support causal reasoning in general.

Example 27.8 Let \mathcal{M}^1 and \mathcal{M}^2 be two SCMs such that $\mathbf{V} = \{X, Z, Y\}$, $\mathbf{U} = \{U_x, U_z, U_y\}$, and the structural mechanisms are, respectively,

$$\mathcal{F}_1 = \begin{cases} X \leftarrow U_x \\ Z \leftarrow X \oplus U_z \\ Y \leftarrow Z \oplus U_y \end{cases}, \quad \mathcal{F}_2 = \begin{cases} X \leftarrow Z \oplus U_x \\ Z \leftarrow Y \oplus U_z \\ Y \leftarrow U_y \end{cases}, \quad (27.18)$$

where \oplus is the logical *xor* operator. Further, the distributions of the exogenous variables are $P^1(U_x = 1) = P^2(U_y = 1) = 1/2$, $P^1(U_z = 1) = P^2(U_x = 1) = a$, and $P^1(U_y = 1) = P^2(U_z = 1) = b$, for some $a, b \in (0, 1)$. It can immediately be seen (via Definition 27.2 and Equation (27.3)) that both models generate the same

observational distribution,

$$\begin{aligned}
 P^{1,2}(X = 0, Z = 0, Y = 0) &= P^{1,2}(X = 1, Z = 1, Y = 1) = (1 - a)(1 - b)/2, \\
 P^{1,2}(X = 0, Z = 0, Y = 1) &= P^{1,2}(X = 1, Z = 1, Y = 0) = (1 - a)b/2, \\
 P^{1,2}(X = 0, Z = 1, Y = 1) &= P^{1,2}(X = 1, Z = 0, Y = 0) = a(1 - b)/2, \\
 P^{1,2}(X = 0, Z = 1, Y = 0) &= P^{1,2}(X = 1, Z = 0, Y = 1) = ab/2.
 \end{aligned} \tag{27.19}$$

We further compute the effect of the intervention $do(x)$ (via Definition 27.5 and Equation 27.7),

$$P^1(Y = 1 | do(X = 1)) = ab + (1 - a)(1 - b), \quad P^2(Y = 1 | do(X = 1)) = 1/2, \tag{27.20}$$

which are different for most values a, b . The models \mathcal{M}^1 and \mathcal{M}^2 naturally induce BNs \mathcal{G}^1 and \mathcal{G}^2 , respectively; see Figure 27.4(a) and (b).²⁶ In terms of \mathcal{L}_1 -constraints, \mathcal{G}^1 and \mathcal{G}^2 both imply that X is independent of Y given Z (for short, $X \perp Y | Z$) and nothing more.²⁷ This means that \mathcal{G}^1 and \mathcal{G}^2 are equivalent through the lens of \mathcal{L}_1 , while the original \mathcal{M}^1 and \mathcal{M}^2 generate different answers to \mathcal{L}_2 queries, as shown in Equation (27.20). ■

The main takeaway from the example is that from only the distribution $P(\mathbf{V})$ and the qualitative (conditional independence) constraints implied by it, it is impossible to tell whether the underlying reality corresponds to \mathcal{M}^1 , \mathcal{M}^2 , or any other SCM inducing the same $P(\mathbf{V})$, while each such model could entail a different causal effect. This suggests that, in general, causal inference cannot be carried out with mere \mathcal{L}_1 objects—the observational distribution, its constraints, and corresponding models (BNs). This result can be seen as a graphical instantiation of Corollary 27.1 and is schematically summarized in Figure 27.4.

27.4.1 Causal Inference via \mathcal{L}_2 -constraints—Markovian Causal Bayesian Networks

Having witnessed the impossibility of performing causal inference from \mathcal{L}_1 constraints, we come back to the original question—what kind of structural constraints (Figure 27.3(d)) imprinted by the underlying SCM could license causal

26. This construction follows from the order in which the functions are determined in the SCM, systematized in Definition 24 [Bareinboim et al. 2020, appendix C]. This procedure is guaranteed to produce BNs that are compatible with the independence constraints implied by the SCM in \mathcal{L}_1 [Bareinboim et al. 2020, theorem 8, appendix C].

27. We refer readers to Bareinboim et al. [2020, appendix C], for more details on a criterion called *d-separation* [Pearl 1988], which is the tool used for reading these constraints off from the graphical model.

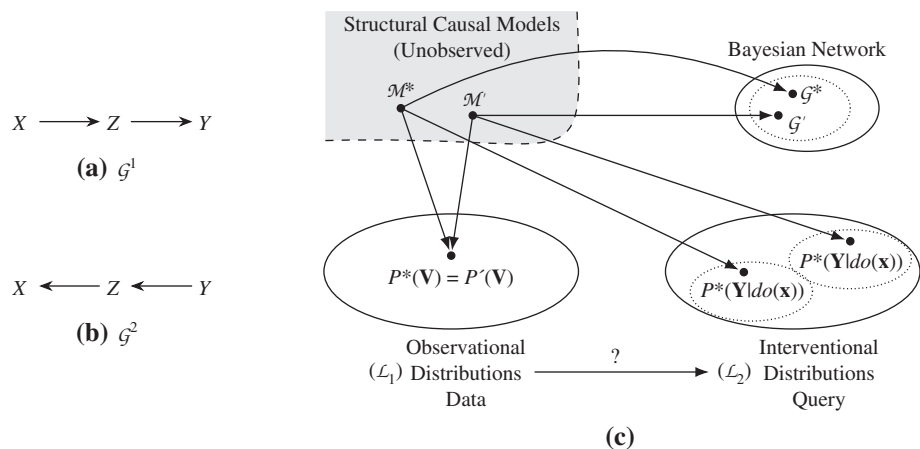


Figure 27.4 Two causal diagrams encoding knowledge about the causal mechanisms governing three observable variables X , Z , and Y . In (a) X is an argument to f_Z , and Z an argument to f_Y . In (b) the opposite is true. In (c), schema representing the impossibility of identifying causal queries from \mathcal{L}_1 data, constraints, and graphical models.

inferences? To answer this question, it is instructive to compare more closely the effect of an intervention $X = 1$ in the two SCMs from Example 27.8. First, note that the function f_Y does not depend on X in the submodel $\mathcal{M}_{X=1}^2$ (constructed following Definition 27.3); so, probabilistically, Y will not depend on X . This implies the following relationship between distributions,

$$P^2(Y = 1 | do(X = 1)) = P(Y = 1), \tag{27.21}$$

In contrast, note that (i) f_Y does take into account the value of X in $\mathcal{M}_{X=1}^1$, and (ii) Y responds (or varies) in the same way when X takes a particular value, be it naturally (as in \mathcal{M}^1) or due to an intervention (as in $\mathcal{M}_{X=1}^1$). These facts can be formally written as

$$P^1(Y = 1 | do(X = 1)) = P(Y = 1 | X = 1). \tag{27.22}$$

The exact computation of Equations (27.21) and (27.22) follows immediately from Definitions 27.2 and 27.5. Remarkably, the intuition behind these equalities does not arise from the particular form of the underlying functions, the exogenous variables, or their distribution, but from structural properties of the model. In particular, they are determined by qualitative functional dependences among the variables: what variable is an argument to the function of the other.

Technically, these equalities can be seen as constraints (not conditional independences) and can be pieced together and given a graphical interpretation.

Consider again Equation (27.21) as an example, which says that variable X does not have an effect on Y (doing X does not change the marginal distribution of Y), which graphically would entail that X is not an ancestor of Y in \mathcal{G}^2 . While true in \mathcal{M}^2 , it certainly does not hold in \mathcal{M}^1 , nor, consequently, in \mathcal{G}^1 . Even though \mathcal{G}^1 and \mathcal{G}^2 are graphically equivalent with respect to \mathcal{L}_1 , and could be used interchangeably for probabilistic reasoning, they are, interventionally speaking, very distinct objects.

These constraints encode one of the fundamental intuitions we have about causality, namely, the asymmetry that a cause may change its effect but not the other way around. Our goal henceforth will be to systematically incorporate these constraints into a new family of graphical models with arrows carrying causal meaning and supporting \mathcal{L}_2 -types of inferences. First, we introduce a procedure that returns a new graphical model following the intuition behind the constraints discussed so far, and then show how it relates to the collection of interventional distributions (\mathcal{L}_2 -valuations) entailed by the SCM.

Definition 27.10 Causal Diagram (Markovian Models)

Consider a Markovian SCM $\mathcal{M} = \langle \mathbf{U}, \mathbf{V}, \mathcal{F}, P(\mathbf{U}) \rangle$. Then, \mathcal{G} is said to be a *causal diagram* (of \mathcal{M}) if constructed as follows:

1. add a vertex for every endogenous variable in the set \mathbf{V} ,
2. add an edge ($V_j \rightarrow V_i$) for every $V_i \in \mathbf{V}$ if V_j appears as an argument of $f_i \in \mathcal{F}$.

■

The procedure encapsulated in Definition 27.10 is central to the elicitation of the knowledge necessary to perform causal inference (Figure 27.3(d)). Intuitively, \mathcal{G} has an arrow from A to B ($A \rightarrow B$) if B “listens” to the value of A ; functionally, A appears as an argument of the mechanism of B . The importance of this notion has been emphasized in the literature by Pearl: “This listening metaphor encapsulates the entire knowledge that a causal network conveys; the rest can be derived, sometimes by leveraging data” [Pearl and Mackenzie 2018, p. 129]. This construction produces a coarsening of the underlying SCM such that the arguments of the functions are preserved while their particular forms are discarded.²⁸

The assumptions that the causal diagram encodes about the SCM impose constraints not only over the \mathcal{L}_1 -distribution P but also over all the interventional (\mathcal{L}_2) distributions as encapsulated in the following definition [Bareinboim et al. 2012].

Definition 27.11 Causal Bayesian Network (CBN)-Markovian

Let \mathbf{P}_* be the collection of all interventional distributions $P(\mathbf{V} | do(\mathbf{x}))$, $\mathbf{x} \subseteq \mathbf{V}$,

²⁸ Given the lack of constraints over the form and shape of the underlying functions and distribution of the exogenous variables, these models are usually called *non-parametric* in the causal inference literature.

$\mathbf{x} \in \text{Val}(\mathbf{X})$, including the null intervention, $P(\mathbf{V})$, where \mathbf{V} is the set of observed variables. A directed acyclic graph \mathcal{G} is called a CBN for \mathbf{P}_* if for all $\mathbf{X} \subseteq \mathbf{V}$, the following conditions hold:

- (i) [Markovian] $P(\mathbf{V} | do(\mathbf{x}))$ is Markov relative to \mathcal{G} .
- (ii) [Missing-link] For every $V_i \in \mathbf{V}$, $V_i \notin \mathbf{X}$ such that there is no arrow from \mathbf{X} to V_i in \mathcal{G} :

$$P(v_i | do(pa_i), do(\mathbf{x})) = P(v_i | do(pa_i)). \quad (27.23)$$

- (iii) [Parents do/see] For every $V_i \in \mathbf{V}$, $V_i \notin \mathbf{X}$:

$$P(v_i | do(\mathbf{x}), do(pa_i)) = P(v_i | do(\mathbf{x}), pa_i). \quad (27.24)$$

■

The first condition requires the graph to be *Markov relative*²⁹ to every interventional distribution $P(\mathbf{V} | do(\mathbf{X} = \mathbf{x}))$, which holds if every variable is independent of its non-descendants given its parents.³⁰ The second condition, missing-link, encapsulates the type of constraint exemplified by Equation (27.21): after fixing the parents of a variable by intervention, the corresponding function should be insensitive to any other intervention elsewhere in the system. In other words, the parents Pa_i *interventionally* shield V_i from interventions ($do(\mathbf{X})$) on other variables. Finally, the third condition, parents do/see, encodes the intuition behind Equation (27.22): whether the function f_i takes the value of its arguments following an intervention ($do(Pa_i = pa_i)$) or by observation (conditioned on $Pa_i = pa_i$), the same behavior for V_i is observed.

Some observations follow immediately from these conditions. First, and perhaps not surprisingly, a CBN encodes stronger assumptions about the world than a BN. In fact, all the content of a BN is encapsulated in condition (i) of a CBN (Definition 27.11) with respect to the observational (null intervention) distribution $P(\mathbf{V})$ (\mathcal{L}_1). A CBN encodes additional constraints on interventional distributions (\mathcal{L}_2) beyond conditional independence, involving different interventions such as those represented in conditions (ii) and (iii).

29. This notion is also known in the literature as *compatibility* or *i-mapness* [Pearl 1988, Koller and Friedman 2009], which is usually encoded in the decomposition of $P(\mathbf{v})$ as $\prod_i P(v_i | pa_i)$ in the Markovian case.

30. In some accounts of causation, this condition is known as the *causal Markov condition* (CMC), and is usually phrased in terms of “causal” parents. We invite the reader to check that conditions (ii) and (iii) are in no way implied by (i). One could in fact see Definition 27.11 as offering a precise characterization of what CMC formally means.

Second, readers familiar with graphical models will be quick to point out that the knowledge encoded in these models is not in the presence but in the absence of the arrows; each missing arrow makes a claim about a certain type of invariance. In the context of BNs (\mathcal{L}_1), each missing arrow corresponds to a conditional independence, a probabilistic type of invariance.³¹ On the other hand, each missing arrow in a CBN represents an \mathcal{L}_2 -type constraint, for example, the lack of a direct effect, as encoded in Definition 27.11 through condition (ii). This new family of constraints closes a long-standing semantic gap, from a graphical model’s perspective, rendering the causal interpretation of the graphical model totally unambiguous.

Before proving that this graphical model encapsulates all the probabilistic and causal constraints required for reasoning in \mathcal{L}_2 , we show next that the \mathcal{L}_2 -empirical content of an SCM—that is, the collection of observational and interventional distributions (Definition 27.5)—indeed matches the content of the CBN (Definition 27.10), as defined above.

Theorem 27.2 \mathcal{L}_2 -Connection—SCM-CBN (Markovian)

The causal diagram \mathcal{G} induced by the SCM \mathcal{M} (following the constructive procedure in Definition 27.10) is a CBN for $\mathbf{P}_*^{\mathcal{M}}$ —the collection of observational and experimental distributions induced by \mathcal{M} . ■

For the complete proof, see Bareinboim et al. [2020, appendix D]. As this result demonstrates, CBNs serve as proxies for SCMs in terms of the observed \mathcal{L}_2 distributions. In practice, whenever the SCM is not fully known and the collection of interventional distributions is not available, this duality suggests that a CBN can act as a basis for causal reasoning. To ground this point, we go back to our task of inferring the interventional distribution, $P(\mathbf{Y} | do(\mathbf{X} = \mathbf{x}))$, from a combination of the observational distribution, $P(\mathbf{V})$, and the qualitative knowledge of the SCM encoded in the causal diagram \mathcal{G} . A remarkable result that holds in Markovian models is that causal inference is always possible, that is, any interventional distribution is computable from \mathcal{L}_1 -data.

Theorem 27.3 Truncated Factorization Product (Markovian)

Let the graphical model \mathcal{G} be a CBN for the set of interventional distributions \mathbf{P}_* . For any $\mathbf{X} \subseteq \mathbf{V}$, the interventional (\mathcal{L}_2) distribution $P(\mathbf{V} | do(\mathbf{x}))$ is identifiable through the truncated factorization product, namely,

$$P(\mathbf{v} | do(\mathbf{x})) = \prod_{\{i | V_i \in \mathbf{X}\}} P(v_i | pa_i) \Big|_{\mathbf{X}=\mathbf{x}}. \quad (27.25)$$

31. One can show that there always exists a separator, in the d -separation sense, between non-adjacent nodes.

In other words, the interventional distribution in the LHS of Equation (27.25) can be expressed as the product given in the right-hand side (RHS) involving only \mathcal{L}_1 -quantities, where the factors relative to the intervened variables are removed, hence the name *truncated factorization product* (see Pearl [2000, equation 1.37]).³² Obviously, any marginal distribution of interest can be obtained by summing out the irrelevant factors, including the causal effect of X on Y .

27.4.2 Causal Inference via \mathcal{L}_2 -constraints—Semi-Markovian Causal Bayes Networks

The treatment provided for the Markovian case turned out to be simple and elegant, yet surprisingly powerful. The causal graph is a perfect surrogate for the SCM in the sense that all \mathcal{L}_2 quantities (causal effects) are computable from \mathcal{L}_1 -type of data (observational) and the constraints in \mathcal{G} . A “model-theoretic” way of understanding this result is that all the SCMs that induce the same causal diagram and generate the same observational distribution will also generate the same set of experimental distributions, immediately computable via the truncated product (Theorem 27.3). This is a quite remarkable result as we moved from a model based on \mathcal{L}_1 -structural constraints (e.g., a Bayes net) such that no causal inference was permitted, to a model encoding \mathcal{L}_2 -constraints (a causal Bayes net) such that any conceivable cross-layer inference is immediately allowed.

In light of these results, one may be tempted to surmise that causal inference is a solved problem. This could not be farther from the truth, unfortunately. The assumption that all the relevant factors about the phenomenon under investigation are measured and represented in the causal diagram (i.e., Markovianity holds) is often too stringent, and violated in most real-world scenarios. This means that the aforementioned results are usually not applicable in practice. Departing from this observation, our goal is to understand the principles that allow cross-layer inferences when the Markov condition does not hold, which entails incorporating unobserved confounders as a building block of \mathcal{L}_2 -graphical models. We start by investigating the reasons the machinery developed so far is insufficient to accommodate such cases.

Example 27.9 Example 27.1 revisited

Recall the two-dice game where the endogenous variables X and Y (the sum and difference of two dice, respectively) do not functionally depend on each other, despite their strong association. One could attempt to model such a setting with

32. The truncated formula is also known as the “manipulation theorem” [Spirtes et al. 2001] or G-computation formula [Robins 1986, p. 1423]. For further details, we refer readers to Pearl [2000, section 3.6.4].

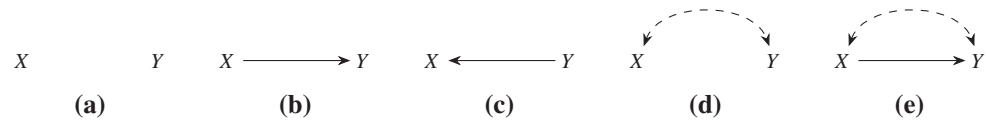


Figure 27.5 The diagram in (a) implies that neither X nor Y is an argument to the function of the other. In (b, c) one endogenous variable causes the other. In (d) there is no causal relationship yet the functions share exogenous arguments, as encoded through the bidirected arrow. In (e) both types of influence are encoded.

the graphical structure shown in Figure 27.5(a), somewhat naively, trying to avoid a directed arrow between X and Y . As previously noted, if the sum of the dice is equal to two ($X = 2$), one could, with probability one, infer that the two dice obtained the same value ($Y = 0$). The hypothesized graphical model, however, forces the two variables to be independent, which would rule out the possibility of performing such an inference.

Upon recognition of such impropriety, one could reconsider adding an arrow from X to Y (or Y to X) so as to leverage the valuable information shared across the observed variables, as shown in Figure 27.5(b). We previously learned, on the other hand, that reporting that the sum of the dice is 2 does not change their difference, formally, $P(Y | do(X = 2)) = P(Y)$ must hold in this setting (Equation 27.9). Obviously, this would be violated were the world to mirror this graphical structure. To witness, consider the alternative SCM \mathcal{M}' where the function for X is identical and $Y \leftarrow (X - 2U_2)$. We can verify that $P(X, Y)$ is the same as in \mathcal{M}^1 , while the causal effect of X on Y is non-zero. ■

The recognition that certain dependencies among endogenous variables cannot be *explained* by other variables inside the model (but also cannot be ignored) led Pearl to introduce a new type of arrow to account for these relationships. The new arrows are dashed and bidirected. In the example above, variables X and Y are correlated due to the existence of two common exogenous variables, $\{U_1, U_2\}$, which are arguments of both f_X and f_Y . We will usually refer to these variables as U_{xy} since, *a priori*, we will neither know, nor want to assume, their particular form, dimensionality, or distribution. This new type of arrow will allow for the probabilistic dependence between them, ($X \perp\!\!\!\perp Y$), while being neutral with respect to their interventional invariance. That is, it would accept constraints such as $P(Y | do(X)) = P(Y)$ and $P(X | do(Y)) = P(X)$. See Figure 27.5(d) for a graphical example.

In practice, some variables may be related through both sources of variations—one exogenous, not explained by the variables in the model, and another endogenous, causally explained by the relationships between the variables in the model,

as shown in Figure 27.5(e). Due to the unobserved confounder U_{xy} , the equality $P(Y | do(x)) = P(Y | x)$ will not, in general, hold. In other words, Y 's distribution will be different depending on whether we observe $X = x$ or intervene and $do(X = x)$. Fundamentally, this will translate into a violation of the constraint encoded in Equation (27.22) and, more generally, in condition (iii) of the definition of CBNs (Definition 27.11).

Our goal, henceforth, will be to cope with the complexity arising due to violations of Markovianity. One particular implication of these violations is the widening of the empirical content carried by the CBN versus its underlying SCM, as shown in the next example.

Example 27.10 Consider two SCMs \mathcal{M}^* and \mathcal{M}' such that $\mathbf{V} = \{X, Y\}$, $\mathbf{U} = \{U_{xy}, U_y\}$, the structural mechanisms are $\mathcal{F} = \{X \leftarrow U_{xy}, Y \leftarrow (X \oplus U_y) \text{ if } X = U_{xy}, \delta \text{ otherwise}\}$, where $\delta = 0$ for \mathcal{M}^* and $\delta = 1$ for \mathcal{M}' . The exogenous distributions of both models, $P^*(\mathbf{U})$ and $P'(\mathbf{U})$, are the same and given by $P(U_{xy} = 1) = 1/2$, $P(U_y = 1) = 3/4$, and they both follow the diagram shown in Figure 27.5(e). It is easy to verify that both models induce the same $P(\mathbf{V})$, while $P^*(Y = 1 | do(X = 1)) = 1/8 \neq 5/8 = P'(Y = 1 | do(X = 1))$. ■

Remarkably, this is our first encounter with a situation in which a causal diagram—encoding all the \mathcal{L}_2 -structural invariances of the underlying SCM \mathcal{M}^* —is too weak, incapable of answering the intended cross-layer inference—computing $P(Y | do(x))$ from the corresponding \mathcal{L}_1 -distribution, $P(X, Y)$. There exists at least one other SCM \mathcal{M}' that shares the same set of structural features, in the form of the constraints encoded in the causal diagram, but generates a different answer for the causal effect. In other words, one cannot commit and make a claim about the target effect as there are multiple, unobserved SCMs compatible with the given diagram and observational data.

Whenever the causal effect is not uniquely computable from the constraints embedded in the graphical model, we say that it is non-identifiable from \mathcal{G} (to be formally defined later on). More generally, we would like to understand under what conditions an interventional distribution can be computed from the observational one, given the structural constraints encoded in the causal diagram. First, we supplement the Markovian construction of CBNs, given in Definition 27.10, to formally account for the existence of unobserved confounders.

Definition 27.12 Causal Diagram (Semi-Markovian Models)

Consider an SCM $\mathcal{M} = \langle \mathbf{U}, \mathbf{V}, \mathcal{F}, P(\mathbf{U}) \rangle$. Then, \mathcal{G} is said to be a *causal diagram* (of \mathcal{M}) if constructed as follows:

- (1) add a vertex for every endogenous variable in the set \mathbf{V} ,

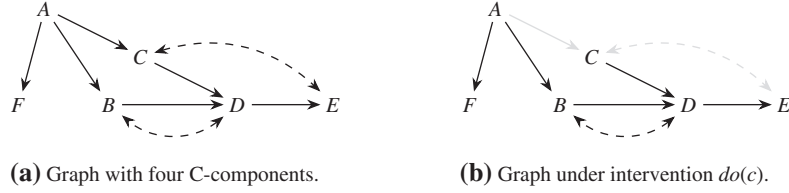


Figure 27.6 Causal diagram with bidirected arrows and its mutilated counterpart under $do(c)$.

- (2) add an edge $(V_j \rightarrow V_i)$ for every $V_i, V_j \in \mathbf{V}$ if V_j appears as an argument of $f_i \in \mathcal{F}$.
- (3) add a bidirected edge $(V_j \leftrightarrow V_i)$ for every $V_i, V_j \in \mathbf{V}$ if the corresponding $U_i, U_j \subset \mathbf{U}$ are correlated or the corresponding functions f_i, f_j share some $U \in \mathbf{U}$ as an argument. ■

Following this procedure, each SCM \mathcal{M} induces a unique causal diagram. Furthermore, each bidirected arrow encodes unobserved confounding in \mathcal{G} . They indicate correlation between the unobserved parents of the endogenous variables at the endpoints of such edges.

27.4.2.1 Revisiting Locality in Semi-Markovian Models

Graphical models provide a transparent and systematic way of encoding structural constraints about the underlying SCM (Figure 27.3(d)). In practice, these constraints follow from the autonomy of the structural mechanisms [Aldrich 1989, Pearl 2000], which materializes as local relationships in the causal diagram. In Markovian models, these local constraints appear in the form of family relationships, for example, (1) each variable V_i is independent of its non-descendants given its parents Pa_i , or (2) each variable is invariant to interventions in other variables once its parents are held constant (following Definition 27.11). The local nature of these relations leads to a parsimonious factorization of the joint probability distribution, and translates into desirable sample and computational complexity properties.

On the other hand, the family relations in semi-Markovian models are less well-behaved and the boundaries of influence among the variables are usually less local. To witness, consider Figure 27.6(a), and note that, where $Pa_d = \{B, C\}$ and the remaining $NDesc_d = \{A, F\}$, $D \perp\!\!\!\perp NDesc_d \mid Pa_d$ does not hold as D and A are connected through the open path $D \leftarrow B \leftarrow A$. We introduce below a construct called *confounded component* [Tian and Pearl 2002b] to restore and help to make sense of modularity in these models.

Definition 27.13 Confounded Component

Let $\{C_1, C_2, \dots, C_k\}$ be a partition over the set \mathbf{V} . C_i is said to be a confounded component (C-component) of \mathcal{G} if there exists a path made of bidirected edges between V_i and V_j , for every $V_i, V_j \in C_i$ in \mathcal{G} , and C_i is maximal. ■

This construct represents clusters of variables that share the same exogenous variations regardless of their directed connections. The causal diagram in Figure 27.6(a) has two bidirected edges indicating the presence of unobserved confounders affecting the pairs (B, D) and (C, E) and contains four C-components, namely, $C_1 = \{A\}$, $C_2 = \{B, D\}$, $C_3 = \{C, E\}$, and $C_4 = \{F\}$. Similarly, each causal diagram in Figure 27.5(a–c) contains two C-components, $C_1 = \{X\}$ and $C_2 = \{Y\}$, while each in Figure 27.5(d, e) contains one C-component, $C_1 = \{X, Y\}$.

Our goal is to understand the boundaries of influence among variables in semi-Markovian models as the parents of a node no longer shield it from its non-descendants, and this condition is a basic building block in the construction of Markovian models. Consider again the graph in Figure 27.6(a) and the node E and its only parent D . If we condition on D , E will not be independent of its non-descendants in the graph. Obviously, E is automatically connected to its bidirected neighbors, so it cannot be separated from C . Further, upon conditioning on the parent D , the collider through C is opened up as D is its descendant (i.e., $E \leftarrow \text{---} \rightarrow C \leftarrow A$ carries correlation given D). In this case, the ancestors and descendants of C also become correlated with E , which is now connected to every other variable in the graph (A, F, B) . Further, note that by conditioning on C itself, its descendants will be independent of E but its ancestors and ancestors' descendants will still be connected. In this graph, E is connected to all other nodes upon conditioning on its observed parent D and C-component neighbor C , that is, A, B, F . Then, we also need to condition on the parents of C (i.e., A) to render its other ancestors and their descendants (i.e., F) independent of E .

Putting these observations together, for each endogenous variable V_i , we need to condition on its parents, the variables in the same C-component that precede it, and the parents of the latter so as to shield V_i from the other non-descendants in the graph. Such a maximal set is formally defined as Pa_i^+ as follows. Let $<$ be a topological order V_1, \dots, V_n of the variables \mathbf{V} in \mathcal{G} ,³³ and let $\mathcal{G}(V_i)$ be the subgraph of \mathcal{G} composed only of variables in V_1, \dots, V_i . Given $\mathbf{X} \subseteq \mathbf{V}$, let $Pa^1(\mathbf{X}) = \mathbf{X} \cup \{Pa(X) : X \in \mathbf{X}\}$; further, let $C(V_i)$ be the C-component of V_i in $\mathcal{G}(V_i)$. Then define $Pa_i^+ = Pa^1(\{V \in C(V_i) : V \leq V_i\}) \setminus \{V_i\}$. For instance, in Figure 27.6(a), $Pa_e^+ = \{D, C, A\}$ and $Pa_d^+ = \{B, C, A\}$.

33. That is, an order on the nodes (endogenous variables) \mathbf{V} such that if $V_j \rightarrow V_i \in \mathcal{G}$, then $V_j < V_i$.

Akin to the concept of *Markov relative*, a causal diagram also imposes factorization constraints over the observational distribution in semi-Markovian CBNs, as shown next.

Definition 27.14 Semi-Markov Relative

A distribution P is said to be *semi-Markov relative* to a graph \mathcal{G} if for any topological order $<$ of \mathcal{G} , P factorizes as

$$P(\mathbf{v}) = \prod_{v_i \in \mathbf{V}} P(v_i | pa_i^+), \quad (27.26)$$

where pa_i^+ is defined using $<$. ■

Not only is the joint observational distribution related to a causal graph, but so are the \mathcal{L}_2 -distributions $P(\cdot | do(\mathbf{x}))$ under an intervention $do(\mathbf{X} = \mathbf{x})$. The corresponding graph is $\mathcal{G}_{\bar{\mathbf{x}}}$, where the incoming arrows toward \mathbf{X} are cut, and the semi-Markovian factorization is

$$P_{\mathbf{x}}(\mathbf{v}) = \prod_{v_i \in \mathbf{V}} P_{\mathbf{x}}(v_i | pa_i^{\mathbf{x}+}), \quad (27.27)$$

where $pa_i^{\mathbf{x}+}$ is constructed as pa_i^+ but according to $\mathcal{G}_{\bar{\mathbf{x}}}$.

Example 27.11 Factorization implied by the semi-Markov condition

Let $P(A, B, C, D, E, F)$ be a distribution semi-Markov relative to the diagram \mathcal{G} in Figure 27.6(a). One topological order of \mathcal{G} is $A < B < C < D < E < F$, which implies that $P(a, b, c, d, e, f) = P(a)P(b|a)P(c|a)P(d|b, c, a)P(e|d, c, a)P(f|a)$. In contrast, an application of the chain rule yields: $P(a, b, c, d, e, f) = P(a)P(b|a)P(c|b, a)P(d|b, c, a)P(e|d, c, b, a)P(f|e, d, c, b, a)$.

A comparison of the two previous factorizations highlights some of the independence constraints implied by the semi-Markov condition, for instance, $(C \perp\!\!\!\perp B | A)$, $(E \perp\!\!\!\perp B | D, C, A)$, and $(F \perp\!\!\!\perp E, D, C, B | A)$. The same applies to interventional distributions. First, let $P_c(A, B, C, D, E, F)$ be semi-Markov relative to $\mathcal{G}_{\bar{c}}$ (Figure 27.6(b)). Then, note that $P_c(A, B, C, D, E, F)$ factorizes as $P_c(a) P_c(b|a)P_c(c) P_c(d|b, c, a)P_c(e|d)P_c(f|a)$. This distribution satisfies the same conditional independence constraints as $P(A, B, C, D, E, F)$, but also additional ones such as $(E \perp\!\!\!\perp A | D)$. This constraint holds true as $(C \leftarrow \dots \rightarrow E)$ is absent in $\mathcal{G}_{\bar{c}}$. The extended parents in both distributions are $pa_e^+ = \{A, C, D\}$ and $pa_e^{c+} = \{D\}$. ■

27.4.2.2 CBNs with Latent Variables—Putting All the Pieces Together

The constructive procedure described in Definition 27.12 produces a coarsening of the underlying SCM such that (1) the arguments of the functions are preserved

while their particular forms are discarded, and (2) the relationships between the exogenous variables are preserved while their precise distribution is discarded.³⁴ The pair $(\mathcal{G}, \mathbf{P}_*)$ consisting of a causal diagram \mathcal{G} , constructed through such a procedure, and the collection of interventional (\mathcal{L}_2) distributions, \mathbf{P}_* , will be called a CBN if it satisfies the definition below. This substitutes for Definition 27.11 in semi-Markovian models, and is similar to the way that constraints on a (observational) probability distribution (viz., conditional independencies) are captured by graphical constraints in a BN and the additional missing-link and do-see constraints are encoded in the Markov-CBNs (Definition 27.11).

Definition 27.15 Causal Bayesian Network (CBN)-Semi-Markovian

Let \mathbf{P}_* be the collection of all interventional distributions $P(\mathbf{V} | do(\mathbf{x}))$, $\mathbf{X} \subseteq \mathbf{V}$, $\mathbf{x} \in \text{Val}(\mathbf{X})$, including the null intervention, $P(\mathbf{V})$, where \mathbf{V} is the set of observed variables. A graphical model with directed and bidirected edges \mathcal{G} is a CBN for \mathbf{P}_* if for every intervention $do(\mathbf{X} = \mathbf{x})$, $\mathbf{X} \subseteq \mathbf{V}$, the following conditions hold:

- (i) [Semi-Markovian] $P(\mathbf{V} | do(\mathbf{x}))$ is semi-Markov relative to $\mathcal{G}_{\bar{\mathbf{x}}}$.
- (ii) [Missing directed-link] For every $V_i \in \mathbf{V} \setminus \mathbf{X}$, $\mathbf{W} \subseteq \mathbf{V} \setminus (Pa_i^{\mathbf{x}^+} \cup \mathbf{X} \cup \{V_i\})$:

$$P(v_i | do(\mathbf{x}), pa_i^{\mathbf{x}^+}, do(\mathbf{w})) = P(v_i | do(\mathbf{x}), pa_i^{\mathbf{x}^+}), \quad (27.28)$$

- (iii) [Missing bidirected-link] For every $V_i \in \mathbf{V} \setminus \mathbf{X}$, let $Pa_i^{\mathbf{x}^+}$ be partitioned into two sets of confounded and unconfounded parents, Pa_i^c and Pa_i^u in $\mathcal{G}_{\bar{\mathbf{x}}}$. Then

$$P(v_i | do(\mathbf{x}), pa_i^c, do(pa_i^u)) = P(v_i | do(\mathbf{x}), pa_i^c, pa_i^u). \quad (27.29)$$

■

The first condition requires each interventional distribution to factorize in a semi-Markovian fashion relative to the corresponding interventional graph $\mathcal{G}_{\bar{\mathbf{x}}}$, as discussed in Example 27.11. The remaining conditions give semantics for the missing directed and bidirected links in the model, which encode the lack of direct effect and of unobserved confounders between the corresponding variables, respectively. Specifically, the missing directed-link condition (ii) states that under any intervention $do(\mathbf{X} = \mathbf{x})$, conditioning on the set of augmented parents $Pa_i^{\mathbf{x}^+}$ renders V_i invariant to an intervention on other variables \mathbf{W} —in other words, \mathbf{W} has no direct effect on V_i . For instance, note that for $V_i = D$ in Figure 27.6(a), $P(d | do(f, e), b, c, a) = P(d | b, c, a)$ as well as $P(d | do(b, c), do(a, f, e)) = P(d | do(b, c))$.

34. Given the lack of constraints over the form and shape of the underlying functions and distribution of the exogenous variables, it is possible to non-parametrically write one in terms of the other.

Further, the missing bidirected-link condition relaxes the stringent parents do/see condition in Markovian CBNs (Definition 27.11(iii)). Note that the do/see condition does not hold due to the unobserved correlation between certain endogenous variables, for instance, both $P(d | do(b)) = P(d | b)$ and $P(e | do(d)) = P(e | d)$ do not hold in Figure 27.6(a).³⁵ Still, given the set of extended parents of V_i , observations and interventions on parents not connected via a bidirected path (i.e., Pa_i^u) yield the same distribution. For instance, $P(e | do(a, d), c) = P(e | a, d, c)$, where $Pa_e^u = \{A, D\}$, $Pa_e^c = \{C\}$; also, $P(d | do(b, a, c)) = P(d | do(b), a, c)$, where $Pa_d^u = \{A, C\}$, $Pa_d^c = \{B\}$. There exists no unobserved confounding in Markovian models, so $Pa_i^u = Pa_i$, which means that the condition is enforced for all parents.

Finally, the causal diagram \mathcal{G} constructed from the SCM and the set of interventional distributions \mathbf{P}_* can be formally connected through the following result:

Theorem 27.4 \mathcal{L}_2 -Connection—SCM-CBN (Semi-Markovian)

The causal diagram \mathcal{G} induced by the SCM \mathcal{M} (following the constructive procedure in Definition 27.12) is a CBN for $\mathbf{P}_*^{\mathcal{M}}$. ■

One could take an axiomatic view of CBNs and consider alternative constructions that satisfy their conditions, detached from the structural semantics (similarly to the Markovian case). We provide in Bareinboim et al. [2020, appendix D] a procedure called CONSTRUCTCBN (see Theorem 10) that constitutes such an alternative. It can be seen as the experimental-stochastic counterpart of the SCM-functional Definition 27.12. We show in the next section that CBNs can act as a basis for causal inference regardless of their underlying generating model.

27.4.2.3 Cross-layer Inferences through CBNs with Latent Variables

The causal diagram associated with a CBN will sometimes be a proper surrogate for the SCM, and allow one to compute the effect of interventions *as if* the fully specified SCM were available. Unfortunately, in some other cases, it will be insufficient, as evident from the discussion in Example 27.10. We introduce next the notion of identifiability [Pearl 2000, p. 77] to more visibly capture each of these instances.

Definition 27.16 Effect Identifiability

The causal effect of an action $do(\mathbf{X} = \mathbf{x})$ on a set of variables \mathbf{Y} given a set of observations on variables $\mathbf{Z} = \mathbf{z}$, $P(\mathbf{Y} | do(\mathbf{x}), \mathbf{z})$, is said to be identifiable from P and \mathcal{G} if for every two models $\mathcal{M}^{(1)}$ and $\mathcal{M}^{(2)}$ with causal diagram \mathcal{G} , $P^{(1)}(\mathbf{v}) = P^{(2)}(\mathbf{v}) > 0$ implies $P^{(1)}(\mathbf{Y} | do(\mathbf{x}), \mathbf{z}) = P^{(2)}(\mathbf{Y} | do(\mathbf{x}), \mathbf{z})$. ■

35. To see why this is the case in the last expression, first let U_d be any exogenous argument to f_D . Now note that $P(e | do(d))$ does not depend on U_d , while $P(e | d)$ does due to the path $U_d \rightarrow D \leftarrow C \leftarrow \dots \rightarrow E$.

This formalizes the very natural type of cross-layer inference we have discussed in Figure 27.3, namely: given qualitative assumptions encoded in the causal diagram \mathcal{G} , one would like to establish whether the interventional distribution (\mathcal{L}_2 -quantity) $P(\mathbf{Y} | do(\mathbf{x}), \mathbf{z})$ is inferable from the observational one (\mathcal{L}_1 -data). We introduce next a set of inference rules known as *do-calculus* [Pearl 1995] developed to answer this question.^{36,37}

Theorem 27.5 Do-Calculus

Let \mathcal{G} be a CBN for \mathbf{P}_* , then \mathbf{P}_* satisfies the Do-Calculus rules according to \mathcal{G} . Namely, for any disjoint sets $\mathbf{X}, \mathbf{Y}, \mathbf{Z}, \mathbf{W} \subseteq \mathbf{V}$ the following three rules hold:

$$\text{Rule 1 } P(\mathbf{y} | do(\mathbf{x}), \mathbf{z}, \mathbf{w}) = P(\mathbf{y} | do(\mathbf{x}), \mathbf{w}) \quad \text{if } (\mathbf{Y} \perp\!\!\!\perp \mathbf{Z} | \mathbf{X}, \mathbf{W}) \text{ in } \mathcal{G}_{\bar{\mathbf{X}}}. \quad (27.30)$$

$$\text{Rule 2 } P(\mathbf{y} | do(\mathbf{x}), do(\mathbf{z}), \mathbf{w}) = P(\mathbf{y} | do(\mathbf{x}), \mathbf{z}, \mathbf{w}) \quad \text{if } (\mathbf{Y} \perp\!\!\!\perp \mathbf{Z} | \mathbf{X}, \mathbf{W}) \text{ in } \mathcal{G}_{\bar{\mathbf{XZ}}}. \quad (27.31)$$

$$\text{Rule 3 } P(\mathbf{y} | do(\mathbf{x}), do(\mathbf{z}), \mathbf{w}) = P(\mathbf{y} | do(\mathbf{x}), \mathbf{w}) \quad \text{if } (\mathbf{Y} \perp\!\!\!\perp \mathbf{Z} | \mathbf{X}, \mathbf{W}) \text{ in } \mathcal{G}_{\bar{\mathbf{XZ}}(\mathbf{w})}, \quad (27.32)$$

where a graph $\mathcal{G}_{\bar{\mathbf{XZ}}}$ is obtained from \mathcal{G} by removing the arrows incoming to \mathbf{X} and outgoing from \mathbf{Z} , and $\mathbf{Z}(\mathbf{W})$ is the set of \mathbf{Z} -nodes non-ancestors of \mathbf{W} in the corresponding graph. ■

These rules can be seen as a tool that allows one to navigate in the space of interventional distributions, jumping across unrealized worlds, and licensed by the invariances encoded in the causal graph. Specifically, rule 1 can be seen as an extension of the d-separation criterion for reading conditional independences under a fixed intervention $do(\mathbf{X} = \mathbf{x})$ from the graph denoted $\mathcal{G}_{\bar{\mathbf{X}}}$. Furthermore, rules 2 and 3 entail constraints among distributions under different interventions. Rule 2 permits the *exchange* of a $do(\mathbf{z})$ operator with an observation of $\mathbf{Z} = \mathbf{z}$, capturing situations when intervening and observing \mathbf{Z} influence the set of variables \mathbf{Y} indistinguishably. Rule 3 licenses the *removal* or *addition* of an intervention from

36. The do-calculus can be seen as an inference engine that allows the local constraints encoded in the CBN, in terms of the family relationships, to be translated and combined to generate (global) constraints involving other variables.

37. The duality between local and global constraints is a central theme in probabilistic reasoning, where the family factorization dictated by the graphical model is local while d-separation is global, allowing one to read off non-trivial constraints implied by the model [Pearl 1988, Lauritzen 1996]. The graphical model could be seen as a basis, that is, a parsimonious encoder of exponentially many conditional independences. In causal inference, do-calculus can be seen as a generalization of d-separation to generate global, interventional-type of constraints.

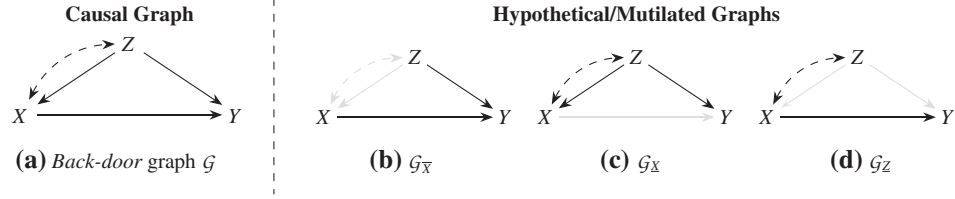


Figure 27.7 (a) Graph representing a model where the query $P(y | do(x))$ is identifiable. The query can be derived using do-calculus rules licensed by graphs (b), (c), and (d).

a probability expression, recognizing situations where $do(z)$ has no effect whatsoever on \mathbf{Y} . A more detailed discussion of do-calculus can be found in Pearl [2000, chapter 3].³⁸

We have previously shown that in simple settings causal inference is unattainable with only \mathcal{L}_1 -data, and that knowledge conveniently encoded in the form of a causal diagram is required. Next, we show how the knowledge from the diagram together with the inference rules of do-calculus allows for the identification of the query $P(y | do(x))$ in the context of the model represented in Figure 27.7(a). First, we start with the target query and then apply do-calculus:

$$P(y | do(x)) = \sum_z P(y | do(x), z)P(z | do(x)) \quad \text{Summing over } Z \quad (27.33)$$

$$= \sum_z P(y | do(x), z)P(z) \quad \text{Rule 3: } (Z \perp\!\!\!\perp X)_{\mathcal{G}_{\bar{X}}} \quad (27.34)$$

$$= \sum_z P(y | x, z)P(z) \quad \text{Rule 2: } (Y \perp\!\!\!\perp X | Z)_{\mathcal{G}_{\underline{X}}}. \quad (27.35)$$

Each step above is accompanied by the corresponding probability axiom or rule, supported by the licensing graphs $\mathcal{G}_{\bar{X}}$ and $\mathcal{G}_{\underline{X}}$ (Figure 27.7(b) and (c), respectively). As desired, the RHS of Equation (27.35) is a function of $P(\mathbf{V})$, hence, estimable from \mathcal{L}_1 -data. This means that no matter the functional form of the endogenous variables or the distribution over the exogenous ones, for all SCMs compatible with the graph in Figure 27.7(a), the causal effect of X on Y will always be equal to Equation (27.35). This can be seen as an instance of the back-door criterion [Pearl 1993], and the particular function in Equation (27.35) is known as adjustment (for \mathbf{Z}).

The importance of the back-door criterion stems from the fact that adjustment is a very common technique used to identify causal effects in the sciences. While the adjustment expression has been used since much earlier than the discovery of

38. Interestingly, the do-calculus theorem (Theorem 27.5) as stated here was derived entirely within the domain of CBNs and Layer 2 constraints, which contrasts with the traditional proposition ([Pearl 1995, theorem 27.3]) based on Layer 3 facts.

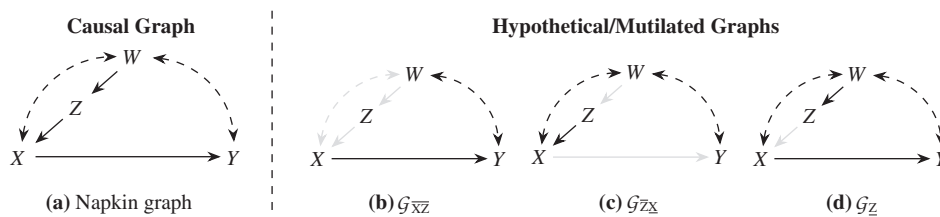


Figure 27.8 Napkin graph (a) and derived graphs used to identify $P(y | do(x))$.

the criterion itself [Pearl 1993], the back-door is the first to provide a transparent way one could judge the plausibility of the assumptions required to map \mathcal{L}_1 -data to an \mathcal{L}_2 -quantity based on a model of the world.³⁹

For the effect of Z on X , $P(X | do(z))$, in the same graph (Figure 27.7(a)), there exists no set Z that can be used to identify the effect by adjustment. Note that in the graph where the arrows outgoing from Z are cut (Figure 27.7(d)), Z and X cannot be separated due to the existence of the latent path, $Z \leftarrow \dots \rightarrow X$. More strongly, $P(X | do(z))$ is not identifiable from the observational distribution by any other means. We leave as an exercise the construction of a counter-example based on Example 27.10's proof. Broadly, the effect of a certain intervention may or may not be identifiable, depending on the particular causal diagram and the topological relations between treatment, outcome, and latent variables.

Finally, there are involved scenarios that are somewhat surprising as they go beyond some of the intuitions discussed in the examples above; see diagram in Figure 27.8(a). The task is to identify the effect of X on Y , $P(Y | do(x))$, from $P(W, Z, X, Y)$. It is obvious that the effect cannot be identified by the back-door criterion, and in \mathcal{G}_X , conditioning on $\{Z\}, \{W\}, \{Z, W\}$ leaves the back-door path $X \leftarrow \dots \rightarrow W \leftarrow \dots \rightarrow Y$ opened. After all, one may be tempted to believe that the effect of X on Y is not identifiable in this case. Contrary to this intuition, consider the following derivation in do-calculus:

$$P(y | do(x)) = P(y | do(x), do(z)) \quad \text{Rule 3: } (Y \perp\!\!\!\perp Z | X)_{\mathcal{G}_{\overline{XZ}}} \quad (27.36)$$

$$= P(y | do(z), x) \quad \text{Rule 2: } (Y \perp\!\!\!\perp X)_{\mathcal{G}_{\overline{ZX}}} \quad (27.37)$$

$$= \frac{P(y, x | do(z))}{P(x | do(z))} \quad \text{Def. of cond. probability.} \quad (27.38)$$

The rules used in each step and the licensing graphs are shown in Figure 27.8(b)–(c). At this point, the back-door adjustment (similar to Equations (27.33)–(27.35))

39. The back-door criterion provides a formal and transparent condition to judge the validity of a condition called *conditional ignorability* [Imbens and Rubin 2015]; see further details in Pearl [2000, section 11.3.2].

can be applied to solve for both factors in Equation (27.38). To witness, note that in the numerator, $P(y, x | do(z))$, $\{W\}$ is back-door admissible with respect to $(Z, \{Y, X\})$, as $(Y, X \perp\!\!\!\perp Z | W)_{\mathcal{G}_Z}$, as shown in Figure 27.8(d). The denominator follows by marginalizing Y out. Putting these two results together and replacing it back into Equation (27.38) lead to:

$$P(y | do(x)) = \frac{\sum_w P(y, x | z, w)P(w)}{\sum_w P(x | z, w)P(w)}. \quad (27.39)$$

The RHS of Equation (27.39) is expressible in terms of $P(\mathbf{V})$, which means that for any SCM compatible with the graph, the causal effect will always be the same, regardless of the details of the underlying mechanisms and distribution over the exogenous variables. The expression shown in Equation (27.39) is a ratio following from the application of the back-door criterion twice.

The problem of deciding identifiability, also known as non-parametric identification, has been extensively studied in the literature. There are a number of conditions that have been proposed to solve this problem, including Galles and Pearl [1995], Pearl and Robins [1995], Kuroki and Miyakawa [1999], and Spirtes et al. [2001]. The do-calculus provides a general mathematical treatment for non-parametric identification [Pearl 1995]. It has been made systematic and shown to be complete for the task of identification from a combination of observations and experiments [Tian and Pearl 2002a, Huang and Valtorta 2006, Shpitser and Pearl 2006, Bareinboim and Pearl 2012, Lee et al. 2019]. In other words, given a causal diagram \mathcal{G} and a collection of observational and experimental distributions, the target effect of \mathbf{X} on \mathbf{Y} given a set of covariates \mathbf{Z} , $P(y | do(x), z)$, is identifiable if and only if there exists a sequence of application of the rules of do-calculus that reaches an estimand in terms of the available distributions.

27.5 Conclusions

We investigated a mathematical structure called the PCH, which was discovered by Judea Pearl when studying the conditions under which some types of causal explanations can be inferred from data [Pearl 2000, Pearl and Mackenzie 2018]. The PCH is certainly one of the most productive conceptual breakthroughs in the science of causal inference over the last decades. It highlights and formalizes the distinct roles of some basic human capabilities—*seeing*, *doing*, and *imagining*—spanning cognition, AI, and scientific discovery. The structure is pervasive in the empirical world: as long as a complex system can be described as a collection of causal mechanisms—that is, an SCM (Definition 27.1)—the hierarchy relative to the modeled phenomena emerges (Definition 27.8).

The main contribution of this chapter is a detailed analysis of the PCH through different perspectives: one semantical (Section 27.2), another logical-probabilistic (Section 27.3), and another inferential-graphical (Section 27.4). These complementary approaches elucidate the PCH from different angles, ranging from when one knows everything about a specific SCM (semantical), to talking about classes of SCMs in general (probabilistic), and ending with one SCM that is particular to the environment of interest but which is not fully observed (graphical). We hope these distinct angles provide a powerful tool for studying causation across different research communities, with far-reaching implications for scientific practice in a wide range of data-driven fields. For instance, we expect these results to underpin the next generation of AI systems, which should be data-efficient, explainable, and aligned with society's goals.

Acknowledgments

This work has benefited immensely from conversations with David Blei, Carlos Cinelli, Paul Daniell, Philip Dawid, Heyang Gong, David Kinney, Sanghack Lee, Judea Pearl, Jin Tian, Yuhao Wang, and Junzhe Zhang. We are grateful to the Editors, Professors Rina Dechter, Hector Geffner, and Joe Halpern, for the opportunity to contribute to this volume. Elias Bareinboim and Juan Correa were partially supported by grants from NSF IIS-1704352 and IIS-1750807 (CAREER). Duligur Ibeling was supported by the NSF Graduate Research Fellowship (DGE-1656518). Thomas Icard was partially supported by the Center for the Study of Language and Information.

References

- J. Aldrich. 1989. Autonomy. *Oxford Econ. Pap.* 41, 15–34. DOI: <https://doi.org/10.1093/oxfordjournals.oep.a041889>.
- E. Bareinboim and J. Pearl. 2012. Causal inference by surrogate experiments: z-Identifiability. In N. d. F. Murphy and Kevin (Eds.), *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence*. AUAI Press, 113–120.
- E. Bareinboim and J. Pearl. 2016. Causal inference and the data-fusion problem. *Proc. Natl. Acad. Sci.* 113, 27, 7345–7352. DOI: <https://doi.org/10.1073/pnas.1510507113>.
- E. Bareinboim, C. Brito, and J. Pearl. 2012. Local characterizations of causal Bayesian networks. In M. Croitoru, S. Rudolph, N. Wilson, J. Howse, and O. Corby (Eds.), *Graph Structures for Knowledge Representation and Reasoning*. Springer Berlin Heidelberg, Berlin, Heidelberg, 1–17. DOI: https://doi.org/10.1007/978-3-642-29449-5_1.
- E. Bareinboim, J. D. Correa, D. Ibeling, and T. Icard. 2020. *On Pearl's Hierarchy and the Foundations of Causal Inference*. Technical Report R-60, Causal AI Lab, Columbia University.
- E. W. Beth. 1956. On Padoa's method in the theory of definition. *J. Symb. Log.* 2, 1, 194–195. DOI: <https://doi.org/10.2307/2268764>.

- R. Briggs. 2012. Interventionist counterfactuals. *Philos. Stud.* 160, 1, 139–166. DOI: <https://doi.org/10.1007/s11098-012-9908-5>.
- N. Cartwright. 1989. *Nature's Capacities and Their Measurement*. Clarendon Press, Oxford. DOI: <https://doi.org/10.1093/0198235070.001.0001>.
- N. Chomsky. 1959. On certain formal properties of grammars. *Inf. Control.* 2, 137–167. [https://doi.org/10.1016/S0019-9958\(59\)90362-6](https://doi.org/10.1016/S0019-9958(59)90362-6).
- A. P. Dawid. 2000. Causal inference without counterfactuals (with comments and rejoinder). *J. Am. Stat. Assoc.* 95, 450, 407–448. DOI: <https://doi.org/10.1080/01621459.2000.10474210>.
- R. Fagin, J. Y. Halpern, and N. Megiddo. 1990. A logic for reasoning about probabilities. *Inf. Comput.* 87, 1/2, 78–128. DOI: [https://doi.org/10.1016/0890-5401\(90\)90060-U](https://doi.org/10.1016/0890-5401(90)90060-U).
- R. A. Fisher. 1936. Design of experiments. *Br. Med. J.* 1, 3923, 554. DOI: <https://doi.org/10.1136/bmj.1.3923.554-a>.
- D. Galles and J. Pearl. 1995. Testing identifiability of causal effects. In P. Besnard and S. Hanks (Eds.), *Uncertainty in Artificial Intelligence 11*. Morgan Kaufmann, San Francisco, 185–195.
- D. Galles and J. Pearl. 1998. An axiomatic characterization of causal counterfactuals. *Found. Sci.* 3, 1, 151–182. DOI: <https://doi.org/10.1023/A:1009602825894>.
- T. Haavelmo. 1943. The statistical implications of a system of simultaneous equations. *Econometrica* 11, 1, 1. DOI: <https://doi.org/10.2307/1905714>.
- J. Y. Halpern. 1998. Axiomatizing causal reasoning. In G. F. Cooper and S. Moral (Eds.), *Uncertainty in Artificial Intelligence*. Cornell University, Morgan Kaufmann, San Francisco, CA, 202–210.
- J. Y. Halpern. 2000. Axiomatizing causal reasoning. *J. Artif. Intell. Res.* 12, 317–337. DOI: <https://doi.org/10.1613/jair.648>.
- J. Y. Halpern. 2013. From causal models to counterfactual structures. *Rev. Symb. Logic.* 6, 2, 305–322. DOI: <https://doi.org/10.1017/S1755020312000305>.
- Y. Huang and M. Valtorta. 2006. Identifiability in causal Bayesian networks: A sound and complete algorithm. In *Proceedings of the Twenty-First National Conference on Artificial Intelligence (AAAI 2006)*. AAAI Press, Menlo Park, CA, 1149–1156.
- D. Hume. 1739. *A Treatise of Human Nature*. Oxford University Press, Oxford.
- D. Hume. 1748. *An Enquiry Concerning Human Understanding*. Open Court Press, LaSalle.
- D. Ibeling and T. Icard. 2018. On the conditional logic of simulation models. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*. 1868–1874. DOI: <https://doi.org/10.24963/ijcai.2018/258>.
- D. Ibeling and T. Icard. 2019. On open-universe causal reasoning. In *Proceedings of the 35th Conference on Uncertainty in Artificial Intelligence*.
- D. Ibeling and T. Icard. 2020. Probabilistic reasoning across the causal hierarchy. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*. DOI: <https://doi.org/10.1609/aaai.v34i06.6577>.

- T. Icard. 2020. Calibrating generative models: The probabilistic Chomsky–Schützenberger hierarchy. *J. Math. Psychol.* 95. DOI: <https://doi.org/10.1016/j.jmp.2019.102308>.
- G. W. Imbens and D. B. Rubin. 2015. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, Cambridge, MA. DOI: <https://doi.org/10.1017/CBO9781139025751>.
- D. Koller and N. Friedman. 2009. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press.
- M. Kuroki and M. Miyakawa. 1999. Identifiability criteria for causal effects of joint interventions. *J. R. Stat. Soc.* 29, 105–117. DOI: <https://doi.org/10.14490/jjss1995.29.105>.
- S. L. Lauritzen. 1996. *Graphical Models*. Clarendon Press, Oxford.
- S. Lee, J. D. Correa, and E. Bareinboim. 2019. General identifiability with arbitrary surrogate experiments. In *Proceedings of the Thirty-Fifth Conference Annual Conference on Uncertainty in Artificial Intelligence*. AUAI Press, in press, Corvallis, OR.
- D. Lewis. 1973. *Counterfactuals*. Harvard University Press, Cambridge, MA.
- J. Locke. 1690. *An Essay Concerning Human Understanding*. London, Thomas Basset.
- J. L. Mackie. 1980. *The Cement of the Universe: A Study of Causation*. Clarendon Press, Oxford. DOI: <https://doi.org/10.1093/0198246420.001.0001>.
- J. Marschak. 1950. Statistical inference in economics. In T. Koopmans (Ed.), *Statistical Inference in Dynamic Economic Models*. Wiley, New York, 1–50.
- T. Maudlin. 2019. The why of the world. *Boston Review*. <https://bostonreview.net/science-nature/tim-maudlin-why-world>. Accessed February 10, 2020.
- J. Neyman. 1923. On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Stat. Sci.* 5, 4, 465–480. DOI: <https://doi.org/10.1214/ss/1177012031>.
- J. Pearl. 1988. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, San Mateo, CA.
- J. Pearl. 1993. Aspects of graphical models connected with causality. In *Proceedings of the 49th Session of the International Statistical Institute*, 1 (August), 399–401.
- J. Pearl. 1995. Causal diagrams for empirical research. *Biometrika* 82, 4, 669–688. DOI: <https://doi.org/10.1093/biomet/82.4.669>.
- J. Pearl. 2000. *Causality: Models, Reasoning, and Inference*. (2nd. ed.). Cambridge University Press, NY. DOI: <https://doi.org/10.1017/S0266466603004109>.
- J. Pearl. 2001. Bayesianism and causality, or, why I am only a half-Bayesian. In *Foundations of Bayesianism, Applied Logic Series, Volume 24*. Kluwer Academic Publishers, 19–36. DOI: https://doi.org/10.1007/978-94-017-1586-7_2.
- J. Pearl. 2012. The mediation formula: A guide to the assessment of causal pathways in nonlinear models. In C. Berzuini, P. Dawid, and L. Bernardinelli (Eds.), *Causality: Statistical Perspectives and Applications*, John Wiley and Sons, Ltd, Chichester, UK, 151–179. DOI: <https://doi.org/10.1002/9781119945710.ch12>.

- J. Pearl and E. Bareinboim. 2019. A note on “generalizability of study results.” *J. Epidemiol.* 30, 186–188. DOI: <https://doi.org/10.1097/EDE.0000000000000939>.
- J. Pearl and D. Mackenzie. 2018. *The Book of Why*. Basic Books, New York.
- J. Pearl and J. M. Robins. 1995. Probabilistic evaluation of sequential plans from causal models with hidden variables. In *Uncertainty in Artificial Intelligence 11*. Morgan Kaufmann, 444–453.
- D. C. Penn and D. J. Povinelli. 2007. Causal cognition in human and nonhuman animals: A comparative, critical review. *Annu. Rev. Psychol.* 58, 97–118. DOI: <https://doi.org/10.1146/annurev.psych.58.110405.085555>.
- J. Peters, D. Janzing, and B. Schölkopf. 2017. *Elements of Causal Inference: Foundations and Learning Algorithms*. The MIT Press.
- G. de Pierris. 2015. *Ideas, Evidence, and Method: Hume’s Skepticism and Naturalism concerning Knowledge and Causation*. Oxford University Press. DOI: <https://doi.org/10.1093/acprof:oso/9780198716785.001.0001>.
- J. M. Robins. 1986. A new approach to causal inference in mortality studies with a sustained exposure period—applications to control of the healthy workers survivor effect. *Math. Model.* 7, 1393–1512. DOI: [https://doi.org/10.1016/0270-0255\(86\)90088-6](https://doi.org/10.1016/0270-0255(86)90088-6).
- P. R. Rosenbaum and D. B. Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70, 1, 41–55. DOI: <https://doi.org/10.1093/biomet/70.1.41>.
- P. K. Rubenstein, S. Weichwald, S. Bongers, J. M. Mooij, D. Janzing, M. Grosse-Wentrup, and B. Schölkopf. 2017. Causal consistency of structural equation models. In *Proceedings of the 33rd Conference on Uncertainty in Artificial Intelligence (UAI)*.
- D. B. Rubin. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* 66, 5, 688–701. DOI: <https://doi.org/10.1037/h0037350>.
- B. Schölkopf. 2019. Causality for machine learning. *arXiv preprint arXiv:1911.10500*.
- I. Shpitser and J. Pearl. 2006. Identification of joint interventional distributions in recursive semi-Markovian causal models. In *Proceedings of the Twenty-First AAAI Conference on Artificial Intelligence*. 2, 1219–1226.
- H. A. Simon. 1953. Causal ordering and identifiability. In W. C. Hood and T. C. Koopmans (Eds.), *Studies in Econometric Method*, Wiley and Sons, Inc., New York, 49–74. DOI: https://doi.org/10.1007/978-94-010-9521-1_5.
- P. Spirtes, C. N. Glymour, and R. Scheines. 2001. *Causation, Prediction, and Search*. (2nd ed.). MIT Press.
- L. J. Stockmeyer. 1977. The polynomial-time hierarchy. *Theor. Comput. Sci.* 3, 1–22. DOI: [https://doi.org/10.1016/0304-3975\(76\)90061-X](https://doi.org/10.1016/0304-3975(76)90061-X).
- R. H. Strotz and H. O. A. Wold. 1960. Recursive versus nonrecursive systems: An attempt at synthesis. *Econometrica* 28, 417–427. DOI: <https://doi.org/10.2307/1907731>.
- P. Suppes and M. Zanotti. 1981. When are probabilistic explanations possible? *Synthese* 48, 191–199. DOI: <https://doi.org/10.1007/BF01063886>.

- R. S. Sutton and A. G. Barto. 2018. *Reinforcement Learning: An Introduction*. (2nd. ed.) The MIT Press.
- J. Tian and J. Pearl. 2002a. A general identification condition for causal effects. In *Proceedings of the Eighteenth National Conference on Artificial Intelligence (AAAI 2002)*. 567–573.
- J. Tian and J. Pearl. 2002b. On the testable implications of causal models with hidden variables. In *Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence*. 519–527.
- T. VanderWeele. 2015. *Explanation in Causal Inference: Methods for Mediation and Interaction*. Oxford University Press.
- J. Woodward. 2003. *Making Things Happen*. Oxford University Press, New York. DOI: <https://doi.org/10.1093/0195155270.001.0001>.
- G. H. von Wright. 1971. *Explanation and Understanding*. Cornell University Press. DOI: https://doi.org/10.1007/978-94-010-1823-4_15.
- J. Zhang. 2013. A Lewisian logic of causal counterfactuals. *Minds Mach.* 23, 77–93. DOI: <https://doi.org/10.1007/s11023-011-9261-z>.
- J. Zhang and E. Bareinboim. 2018. Fairness in decision-making—the causal explanation formula. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*. 2037–2045.

The Tale Wags the DAG

Philip Dawid (University of Cambridge)

*The glass is falling hour by hour, the glass will fall forever,
But if you break the bloody glass you won't hold up the weather.¹*

Abstract

In this chapter, I review a number of applications of directed acyclic graphs (DAGs), which have played a major role in Judea Pearl's fundamental contributions to probabilistic and causal reasoning.

28.1 Introduction

Over many years and through numerous influential publications, Judea Pearl has popularized and developed the use of graphical representations, particularly directed acyclic graphs (DAGs), to display and manipulate properties of probabilistic and causal systems. There are, however, several ways of doing this, with differences in the sort of problem represented, the detailed graphical structure, and the intended interpretation. Here I will survey a variety of DAG models, examining their relationships and differences. I emphasize in particular the specific tale a DAG is intended to tell, and examine how, and how well, it tells its tale. I hang these tales on Pearl's metaphor of "the ladder of causation," with its three rungs telling tales of association, causation, and imagination.

In Section 28.2, I briefly outline the ladder of causation and introduce DAGs. Section 28.3 gives some necessary, purely mathematical, notation and theory of DAGs. In Section 28.4, I describe how a DAG can be used, either as it stands or through an elaboration that introduces auxiliary "error variables," to model and manipulate conditional independence properties of a joint probability distribution. Section 28.5 moves up one rung, to consider how a DAG can be used to

1. From "Bagpipe Music" in *Collected Poems*, by Louis MacNeice, published by Faber and Faber. © Estate of Louis MacNeice, reprinted by permission of David Higham.

represent causal properties. Again, this can be in its raw form, or by elaborations to include error variables and/or non-stochastic regime indicator variables. I argue for the value of regime indicators, but the irrelevance of error variables, for telling causal tales. Error (or “background”) variables are, however, vital for Section 28.6, which discusses the strengths and limitations of DAG models to relate the actual world with unrealized parallel universes, and so address such problems as the attribution of blame or responsibility.

The material covered in this chapter has been described in more detail in a number of previous articles: Constantinou and Dawid et al. [2017], Dawid [2000, 2002, 2007a, 2007b, 2010a, 2010b, 2015]; however, the organization is new.

28.2 The Ladder of Causation

In *The Book of Why* [Pearl and Mackenzie 2018]—a comprehensive and fascinating overview of his approach and contributions to causal inference—Judea Pearl has made vividly explicit the progressive 3-fold nature of the subject, using the metaphor of the “ladder of causation.” This ladder has three rungs: from the bottom, “Seeing,” “Doing,” and “Imagining.” As we climb the ladder, we meet increasingly complex problems and increasingly sophisticated methods for tackling them. This logical upwards journey also describes Pearl’s own progression as he developed his causal understandings and contributions—a journey I have followed, in Pearl’s footsteps, hoping to pick up a few crumbs here and there. It has been a wonderfully educational and fruitful climb—though I have sometimes taken some off-piste paths of my own, and have not found it easy to join Pearl at the very summit.

Pearl’s approach is largely centered around the representation of an applied problem by means of a DAG, as exemplified in Figure 28.1. A DAG consists of a set of nodes, with arrows between some of them, it being impossible to return to one’s starting point by following the arrows. In applications, the nodes of such a graph will correspond to (observable or unobservable) random quantities of interest.

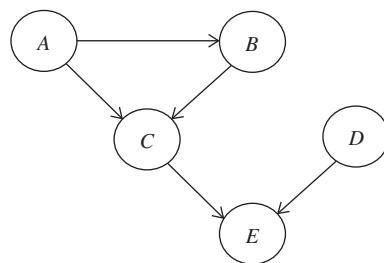


Figure 28.1 A DAG.

However, there is no simple interpretation of the arrows. An attempt to supply an interpretation of the arrows may be termed a “DAG semantics.”

In this chapter, I shall consider the distinctions and connections between the DAG semantics relevant to each of the three rungs of the ladder of causation.

28.3 Ground Level: Syntax

It is first necessary to introduce some purely syntactical terminology and concepts associated with a DAG \mathcal{D} . Let \mathcal{V} be the set of nodes of \mathcal{D} . For $V, W \in \mathcal{V}$, if there is an arrow in \mathcal{D} from V to W , then V is called a *parent* of W , and W is a *child* of V . The set of parents of V is denoted by $\text{pa}(V)$, and the set of its children by $\text{ch}(V)$. This genealogical metaphor is extended, in obvious ways, to define *ancestor* and *descendant*; $\text{de}(V)$ and $\text{an}(V)$ are the sets of descendants and ancestors of V , respectively. Nodes V and W are *married* if there is an arrow in either direction between them. They are *partners* if they have a common child. A configuration of two unmarried partners with their common child is termed an *immorality*.

We now introduce a somewhat complex, but fundamental, graph-separation property, \mathcal{D} -separation, that may hold between subsets $\mathcal{S}, \mathcal{T}, \mathcal{U}$ of \mathcal{V} . When it holds, we will say that \mathcal{S} and \mathcal{T} are \mathcal{D} -separated by \mathcal{U} , and write $\mathcal{S} \perp_{\mathcal{D}} \mathcal{T} \mid \mathcal{U}$. (We also write $\mathcal{S} \perp_{\mathcal{D}} \mathcal{T}$ when \mathcal{U} is empty.) We may drop the subscript \mathcal{D} when the relevant DAG \mathcal{D} is clear from the context.

There are two different, but logically equivalent, ways, *d-separation* and *moralization*, to describe the property $\mathcal{S} \perp_{\mathcal{D}} \mathcal{T} \mid \mathcal{U}$. These are as follows.

***d*-separation** [Verma and Pearl 1990] A *trail* in \mathcal{D} is a sequence of distinct nodes such that each adjacent pair is married. A trail π from V to W is said to be *blocked* by $\mathcal{U} \subseteq \mathcal{V}$ if it contains a node Z such that *either*

- $Z \in \mathcal{U}$ and the arrows of π do not meet head-to-head at Z ; or
- Z and all its descendants are not in \mathcal{U} , and the arrows of π do meet head-to-head at Z .

Then $\mathcal{S} \perp_{\mathcal{D}} \mathcal{T} \mid \mathcal{U}$ if all trails in \mathcal{D} from a node in \mathcal{S} to a node in \mathcal{T} are blocked by \mathcal{U} .

Moralization [Lauritzen et al. 1990] This criterion involves three successive steps:

1. **Ancestral graph** Form a new DAG \mathcal{D}' by removing from \mathcal{D} any node which is neither in $\mathcal{S} \cup \mathcal{T} \cup \mathcal{U}$ nor is an ancestor of a node in this set, together with any edges in or out of such nodes.
2. **Moralization** In \mathcal{D}' , marry any unmarried partners by adding an undirected line between them. Then remove any remaining arrowheads.

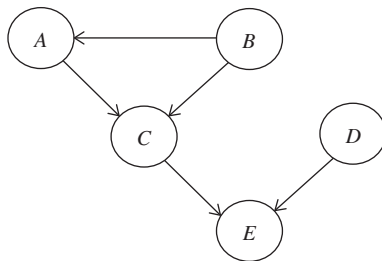


Figure 28.2 Another DAG.

3. **Separation** In the undirected graph so constructed, look for a path which joins a node in S to one in T but does not intersect U . Then $S \perp_{\mathcal{D}} T | U$ if there is no such path.

Applying either of these criteria to the DAG of Figure 28.1, we find that, for example, $(D, E) \perp_{\mathcal{D}} (A, B) | C$.

It should be emphasized that the arrows of \mathcal{D} enter the \mathcal{D} -separation criterion only very indirectly. In particular, no direct interpretation should be attached to the direction of an arrow. Indeed, different DAGs, with some arrows pointing in opposite directions, can determine the identical relation $\perp_{\mathcal{D}}$, in which case they are termed *Markov equivalent*. A necessary and sufficient condition for this is that both DAGs would look the same if the arrowheads were removed, and that they have identical immoralities [Frydenberg 1990]. In particular, the DAG of Figure 28.2 is Markov equivalent to that of Figure 28.1—which shows that no simple conclusion can be drawn from the direction of the arrow between A and B .

28.4 Rung 1: Seeing

This rung of the “ladder of causation” in fact has no connection whatsoever with causality. Rather, a DAG \mathcal{D} is used to represent, and facilitate working with, a joint probability distribution P for the set \mathcal{V} of random variables associated with its nodes. An early and influential approach, albeit limited to a subclass of DAGs, was presented in Pearl [1988]. A thorough development may be found in Cowell et al. [1999].

28.4.1 Qualitative Structure

More precisely, the DAG encodes certain qualitative aspects of a distribution: the independencies and conditional independencies [Dawid 1979] between sets of variables. For $S, T, U \subseteq \mathcal{V}$, we write $S \perp_P T | U$ to denote the probabilistic independence, under P , of S and T , given U (or just $S \perp_P T$ if U is empty). Again, we can

drop the subscript P when it is clear from the context. We say that \mathcal{D} represents P if:

$$\mathcal{S} \perp_{\mathcal{D}} \mathcal{T} | \mathcal{U} \Rightarrow \mathcal{S} \perp_P \mathcal{T} | \mathcal{U}. \quad (28.1)$$

A representation is *faithful* if the reverse implication also holds. We point out that not every probabilistic distribution has a faithful DAG representation.

As an example of criterion 28.1, if the DAG of Figure 28.1 (or, equivalently, Figure 28.2) represents P , then $(D, E) \perp_P (A, B) | C$.

Note that, on the left-hand side of criterion 28.1, \mathcal{S} , \mathcal{T} , and \mathcal{U} are considered as purely geometric objects, being sets of nodes of \mathcal{D} ; whereas on the right-hand side they are interpreted as the associated sets of random variables, with probability distribution governed by P . Correspondingly, the \mathcal{D} -separation relation $\perp_{\mathcal{D}}$ on the left-hand side is a purely geometric concept relating to the DAG \mathcal{D} , while the conditional independence relation \perp_P on the right-hand side is a purely probabilistic concept relating to the distribution P . Criterion 28.1 thus sets up a correspondence between these two different universes, and constitutes the *probabilistic semantics* for interpreting a DAG; a DAG endowed with this semantics is a *probabilistic DAG*.

It can be shown that \mathcal{D} represents P if, for each $V \in \mathcal{V}$,

$$V \perp_P \text{nd}(V) | \text{pa}(V), \quad (28.2)$$

where $\text{nd}(V)$ denotes the non-descendants of V .

28.4.2 Quantitative Structure

There will be many distributions P represented by the same probabilistic DAG \mathcal{D} . To specify any one of these, it is enough to specify the *parent-child distributions*: that is, for each $V \in \mathcal{V}$, the conditional distributions of V given $\text{pa}(V)$. This is generally much more economical than specifying directly the joint probabilities for all the variables in \mathcal{V} .

There exist elegant algorithms, taking account of the structure of \mathcal{D} , that streamline computation of the marginal distribution over a set of variables $\mathcal{S} \subseteq \mathcal{V}$, and of the conditional distribution of \mathcal{S} , given observations on some other set of variables \mathcal{T} [Cowell et al. 1999]. The latter solves the “seeing” problem: How should I update my uncertainty about \mathcal{U} , when I have observed the values of the variables in \mathcal{T} ? These algorithms are incorporated into sophisticated software environments such as HUGIN (<https://www.hugin.com>) or GENIE (<https://www.bayesfusion.com/genie/>).

28.4.3 Empirical Assessment

If we can gather data from a joint distribution P , we can check to see whether or not it can be represented by a given DAG \mathcal{D} , by testing whether it satisfies all the conditional independencies implied by criterion 28.1 (it is enough to test those of 28.2). When it is, we can use the data to estimate the parent–child distributions, and so reconstruct the whole distribution P .

Alternatively, if we can assume that P is represented by some DAG (perhaps in a given restricted set of DAGs), we can try and identify such a DAG using data generated from P . This process can again be based on tests of conditional independence, or alternatively on techniques of model comparison. The former approach is termed “constraint-based,” and the latter, “score-based.” Both methods are typically marketed as aiming at “causal discovery”—though no causal interpretation of the DAGs is involved.

28.4.4 Functional DAGs

The DAG \mathcal{D}^f of Figure 28.3 has a special structure in which all stochasticity is confined to the “error variables” ($\varepsilon_A, \varepsilon_B, \varepsilon_C, \varepsilon_D, \varepsilon_E$), while each of the “domain variable” (A, B, C, D, E) is specified as a non-random function of its parents (indicated by the solid-headed arrows). This is equivalent to a (recursive) non-parametric structural equation model, with functional relations and with independent error variables:

$$A = f_A(\varepsilon_A)$$

$$B = f_B(A, \varepsilon_B)$$

$$C = f_C(A, B, \varepsilon_C)$$

$$D = f_D(\varepsilon_D)$$

$$E = f_E(C, D, \varepsilon_E).$$

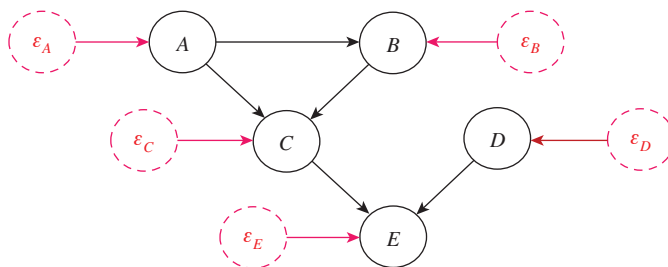


Figure 28.3 A functional DAG.

28.4.5 Downsizing and Upsizing

It is straightforward to show that, when we consider \mathcal{D}^f -separation relationships involving only the domain variables, these agree with the \mathcal{D} -separation relationships of Figure 28.1. The qualitative structure of any probabilistic DAG \mathcal{D} can be reproduced by “downsizing” from a suitable functional DAG \mathcal{D}^f . In fact, it is easy to show that the same holds for the quantitative structure: by suitable choices for the error variables and their distributions, and for the internal functional relationships, in \mathcal{D}^f , we can reproduce any distribution P for the domain variables that is represented by \mathcal{D} . Pearl and others often move backward and forwards between a functional and a purely probabilistic representation of a domain distribution, with many researchers apparently treating the functional version as more fundamental: this may reflect a background in “hard science” disciplines such as physics, where functional dependencies are more familiar than probabilistic ones.

However, at the quantitative level, “upsizing” from \mathcal{D} to \mathcal{D}^f is far from being uniquely determined.

Example 28.1 Consider the simple probabilistic DAG of Figure 28.4, and its functional version in Figure 28.5. Suppose that, in Figure 28.4, $X \in \{0, 1\}$ is a binary coin flip, Y is continuous, and the conditional distribution of Y given $X = x$ is normal, with mean x and variance 1. In its functional elaboration, Figure 28.5, we take ε_X to be binary with $\text{pr}(\varepsilon_X = 1) = 0.5$, and $X = \varepsilon_X$. We take as ε_Y the 2-component vector (E_0, E_1) , having a bivariate normal distribution, with mean-vector $(0, 1)$, both variances 1, and correlation ρ : the function $Y = f_Y(X, \varepsilon_Y)$ is given by $Y = E_X$. This functional model reproduces the probabilistic model for (X, Y) . But note that this will be so for any value of $\rho \in [-1, 1]$, so that there is no unique upsizing of quantitative properties of Figure 28.4 to Figure 28.5.



Figure 28.4 A simple probabilistic DAG.

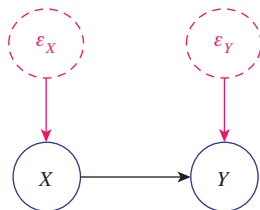


Figure 28.5 A simple functional DAG.

If we want to use a DAG to compute with, or otherwise work with, the distribution P for the domain variables, it makes no difference whether we use the probabilistic DAG or some (any) functional upsizing of it. However, there is no advantage, either for specification or for computation, in using a functional DAG. And if we are tempted to use a functional DAG to make an inference that depends essentially, directly or indirectly, on the error variables, we must pay due attention to non-uniqueness issues, which could lead to different conclusions, depending on how these are resolved.

28.4.6 Empirical Assessment

Suppose we can gather data on the domain variables from some distribution P represented by Figure 28.1. Given enough data we could essentially identify P , and so the full quantitative structure of Figure 28.1. However, since upsizing is non-unique, from such empirical data we could never hope to determine the full quantitative structure of Figure 28.3. The same holds for any probabilistic DAG and its functional version: while the probabilistic DAG can be fully identified with only domain data, some aspects of the functional model will remain forever unidentifiable.

28.5 Rung 2: Doing

We have emphasized that, in the above “seeing” interpretation of a DAG, the interpretation of the direction of the arrows in a DAG is indirect: the conditional independencies of P , as represented by the graphical separation property $\perp_{\mathcal{D}}$, are determined, in a rather obscure way, by the overall structure of the DAG. Indeed, since arrows have direction, whereas the property they represent, $\mathcal{S} \perp_{\mathcal{P}} \mathcal{T} | \mathcal{U}$, is entirely symmetric as between \mathcal{S} and \mathcal{T} , it should be obvious that the relationship must be very indirect.

Nevertheless, on eyeballing a probabilistic DAG such as that of Figure 28.1, it is hard to resist the temptation to endow each arrow with a meaning: this is the fallacy of “reification,” which regards all properties of a representation as necessarily relating to the thing represented: it is as if we expected the contour lines on a map to be visible on the ground.

One tempting misinterpretation of a probabilistic DAG is to regard the arrows as expressing relationships of cause and effect. Thus, we might interpret Figure 28.1 as representing that A causes B , that A and B jointly cause C , that C has no effect on D , and that C and D jointly cause E . Of course, all this presupposes some understanding, typically implicit, of what is meant by “cause”—a task that philosophers have struggled with for millennia. Furthermore, the Markov equivalence of Figures 28.1

and 28.2 shows that such an interpretation cannot be consistently applied without further qualification, since no observational data could distinguish between them, but according to the above causal interpretation they disagree as to whether A causes B , or B causes A . In Dawid [2010a, 2010b], I have considered and criticized various popular ways in which probabilistic DAGs are endowed with additional causal meaning. Many of the so-called “causal discovery” methods mentioned in Section 28.4.3, which look for a DAG representing a probability distribution on the basis of samples from it, claim to have identified cause–effect relationships; but such interpretations can be problematic.

28.5.1 Intervention DAGs

Nevertheless, with due care we *can* impose a clear and meaningful causal interpretation on a DAG. A way to do this was signposted by Spirtes et al. [2000], and has been intensively explored by Judea Pearl, whose book [Pearl 2009] presents a thorough account.

First, we need a clear interpretation of the term “cause.” We do this in terms of a primitive concept of “intervention”: some actual or conceptual means by which an external agent can control the value of a variable—for example, by setting a dial. Such an intervention is a *cause*, and we will generally be interested in its *effects*, in terms of the resulting probability distribution over other, unintervened, variables.

Pearl represents the intervention that sets the value of a variable X to x by $do(X = x)$; an alternative notation, which I shall use, is $X \leftarrow x$. As Pearl emphasizes, we must not confuse $\text{pr}(Y = y | X = x)$ and $\text{pr}(Y = y | X \leftarrow x)$. For example, if X is the reading on a barometer and Y is the air pressure, then *seeing* $X = x$ would indicate that Y is likely to be close to x ; but *doing* $X \leftarrow x$, by moving the indicator hand on the barometer, would give no information about Y .²

“Causal inference” can be regarded as the attempt to infer the effects of interventions (“doing”) on the basis of purely observational (“seeing”) data. Now in general there need be no relationship whatsoever between these different behaviors—when we kick a system, it may behave in ways that could not be guessed at by merely observing it. So no progress is possible unless we start with some plausible assumptions, relating the distinct regimes of seeing and doing, in order to extract causal conclusions—“No Causes In, No Causes Out” [Cartwright 1994]. A very general algebraic framework in which such assumptions can be formalized and manipulated is the “decision-theoretic” approach of Dawid [2015], grounded

2. —as noted by Louis MacNeice at the top of this chapter.

in a generalized concept of conditional independence [Constantinou and Dawid 2017]. Pearl’s approach can be interpreted as a specialization of this, targeted on DAG representations.

Consider again the DAG of Figure 28.1. Instead of regarding this as representing just one “seeing” distribution for the variables, we could also try to relate it to various “doing” regimes. For example, we might believe that the probabilistic dependence of C on the values of its parents A and B would be the same, both in the seeing regime and in a doing regime where the values of A and/or B and/or both are set by external intervention.

In Pearl’s account, a DAG is taken to represent the additional assumptions that, for any node $V \in \mathcal{V}$, its distribution, given the values of its parents, is the same, no matter which nodes in the system (excepting V itself) are set by external intervention. This supplies the DAG with a new *causal semantics*, and I term a DAG endowed with this semantics an *intervention* or *Pearlian* DAG. The extra conditions imply that the same qualitative DAG (albeit with a different, but related, quantitative structure) represents the probabilistic structure, both of the seeing regime and of any doing regime arising from interventions that set the values of any set of the variables. The causal semantics of an intervention DAG supports a rich calculus, developed by Pearl and his students and collaborators, whereby it is possible to interrogate an intervention DAG to determine just what causal conclusions can be deduced from the basic assumptions and observational data, and how.

We note that, when reinterpreted as intervention DAGs, Figures 28.1 and 28.2 are no longer equivalent. Under the former, an intervention on B will have no effect on A , which will retain its marginal seeing distribution, $p(a)$. Under the latter, A will respond to an intervention $B \leftarrow b$ by assuming its conditional seeing distribution, $p(a|b)$.

28.5.2 Augmented DAGs

Now it can be confusing to have different semantic interpretations—probabilistic or causal—of what looks like the same DAG. One road out of the confusion is to modify the DAG itself to make the additional causal assumptions explicit. As an example of how to do this, Figure 28.6 elaborates Figure 28.1 into an “augmented DAG,” having an additional *regime indicator* F_V associated with each original node V [Spirtes et al. 2000]. The square outlines are to indicate that these variables are not random, but rather serve to specify which interventions are being considered. The values associated with F_V are those of V , together with an additional value \emptyset . The interpretation is as follows. If $F_V = v$, for some value v of V , that corresponds to making the intervention $V \leftarrow v$. If $F_V = \emptyset$, that corresponds to leaving V alone.

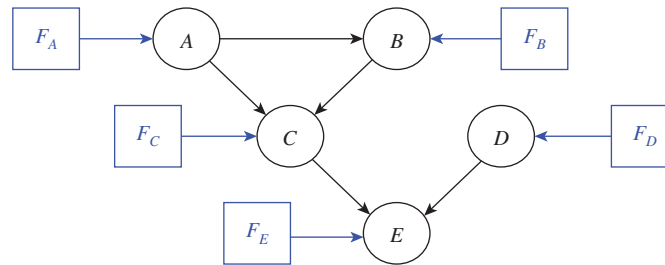


Figure 28.6 An augmented DAG.

Not only does the augmented DAG explicitly show the role of interventions, it also encodes, by applying criterion 28.1 to the augmented DAG, the Pearlian assumptions that relate the different regimes, which are left implicit in Figure 28.1 considered as an intervention DAG. For example, in Figure 28.6 we have $C \perp_{\mathcal{D}} (F_A, F_B) | (A, B, F_C)$. Applying criterion 28.1 yields $C \perp\!\!\!\perp (F_A, F_B) | (A, B, F_C)$. This makes sense even though the F_V nodes are not random: it says that the distribution of C , for given values of (F_A, F_B, A, B, F_C) , in fact depends only on the values of (A, B, F_C) . In particular, consider taking $F_C = \emptyset$, so that C is not intervened upon. Then the above expression represents the property that, once we know the values of A and B , the distribution of C will be the same, no matter what the values of (F_A, F_B) are—that is, no matter whether A and B arise naturally or are set by intervention. In this manner the desired causal interpretation of a DAG is fully represented by the \mathcal{D} -separation property applied to its augmented version. Furthermore, results relating seeing and doing, such as the *back-door* criterion or, more generally, the axioms of Pearl’s *do-calculus*, are trivial applications of ordinary \mathcal{D} -separation applied to the augmented DAG [Dawid 2002]—no new contortions, such as restricting attention to back-door paths, are needed. The augmentation device also serves to make clear the *causal* inequivalence of the Markov equivalent DAGs of Figures 28.1 and 28.2. Their augmented versions, which explicitly encode the causal assumptions otherwise left implicit, are no longer Markov equivalent since, for example, the immorality $A \rightarrow B \leftarrow F_B$ in Figure 28.6 is not preserved when we reverse the arrow from A to B .

Although Pearl initially made use of augmented DAGs, he later dropped these in favor of the leaner unaugmented picture. In my view this was a retrograde step, with no advantages to offset its disadvantages.

28.5.3 Empirical Assessment

An intervention or augmented DAG represents not just the unperturbed distribution for the system but also the many distributions that arise from imposing

interventions, at any set of nodes. Moreover, it relates these different distributions in highly constrained ways: for any variable V that is not intervened on, its parent–child distributions should be the same under all the possible regimes. Such a DAG thus typically encodes a very large body of assumptions about how the world behaves. In principle, these assumptions are testable, if we can gather data under all the regimes; and, when they hold, we can estimate (even from the unperturbed regime alone) the parent–child distributions, and thus the full quantitative structure. In practice it will often be impossible to conduct the required experiments to verify that a DAG is correctly specified, and recourse must be had to reasoned arguments and justifications (e.g., based on scientific understandings) for the selected structure.

28.5.4 Downsizing and Upsizing

Any intervention DAG can be downsized to a probabilistic DAG, representing only how the system behaves when unperturbed. When data are only available from the unperturbed regime, we might find a probabilistic DAG representing the inferred distribution, and then be tempted to upsize this to the full Pearlian DAG, now also describing interventional behaviors. But without further evidence from interventional studies, or good arguments based on scientific understanding, there is no good reason to expect this interpretation to be valid—simply identifying a probabilistic DAG and assuming that it would remain valid under the much more restricted causal semantics can be a very dangerous step to take. There is the added problem that observational data cannot distinguish between Markov equivalent DAGs—but these will upsize to inequivalent intervention DAGs.

28.5.5 Functional Intervention DAGs

Similarly to Section 28.4.4, we can elaborate an intervention DAG by attaching error variables and functional relationships as in Figure 28.3. The causal semantics now requires the persistence of all parent–child functional relationships under interventions at domain nodes, as well as invariance of the distributions of the error variables. Again, this kind of representation is often favored by those whose discipline encourages a deterministic view of the laws of nature.

Again, if we were further to augment a functional intervention DAG with regime indicators, as in Figure 28.7, straightforward application of \mathcal{D} -separation would automatically and explicitly represent the otherwise implicit causal semantics. But this is never done.

We can downsize from a functional intervention (or augmented functional) DAG to obtain a regular intervention (or augmented) DAG, thus moving from a

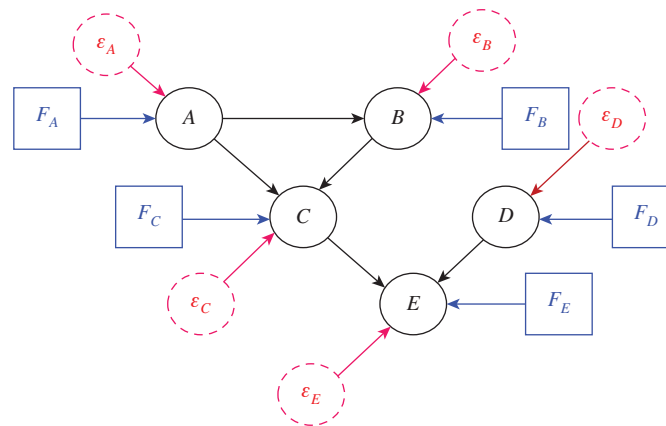


Figure 28.7 An augmented functional DAG.

model of a world governed by deterministic relationships to one where uncertainty is taken seriously; and any regular intervention (or augmented) DAG, with its full causal semantics, can be so obtained; but again, upsizing from the regular to the functional version is non-unique. And once again, there is nothing to be gained from working with the functional form when only domain variables are being considered. Pearl's initial approach to causal DAGs [Pearl 1993] was based on the purely probabilistic interpretation of Section 28.5.1; Pearl [1995] translated this into functional language, but no essential use was made of the additional structure.

28.6 Rung 3: Imagining

As we step up to the third rung of Pearl's ladder of causation, we leave the empirical world behind and enter the world of pure imagination. Thus, in the simple story of Example 28.1, let us consider the DAG of Figure 28.4 as an intervention DAG. Let $X \leftarrow 1$ [resp., 0] encode that an individual Ann is [resp., is not] treated with aspirin for her headache, and let Y be the negative log-duration in minutes of her headache. In fact, Ann was treated with the aspirin and her headache went away in 15 minutes. Did the aspirin help her?

To answer this question we must consider what might have happened if Ann had not been treated with the aspirin—how long might her headache have lasted then? If it would have lasted longer than 15 minutes, then taking the aspirin has indeed helped her.

However, this comparator situation is purely imaginary, since Ann was in fact treated, and there is no way of knowing, at any rate through any direct empirical

evidence, what would have happened in the counterfactual³ case that she was not. Instead, we have to imagine what would be the case in such counterfactual circumstances. To what extent might we gain some assistance from a formal representation of the problem?

Pearl’s approach is based on a functional representation of a problem, such as pictured in Figures 28.3 or 28.5. On the lower rungs of the ladder there was no advantage in using a functional rather than a probabilistic representation, but here it is essential. What we previously termed “error variables,” which we did not need to think much about, are now regarded as “background variables”: real-world variables that are unobserved, but, in conjunction with the specified functional relations, determine all the domain variables.

The functional DAG is now endowed with semantics that incorporates, and significantly extends, the causal semantics of an intervention DAG so as to support counterfactual reasoning. Specifically, we suppose that the background variables and functional forms are common to all the parallel (factual or counterfactual) universes we wish to consider, while the domain variables can be different in different universes. Clearly, this must be the case for any variable imagined to have some counterfactual value, and these differences will be propagated through the system to other variables.

Conditioning on known observations and actions in the factual universe, we can update the joint distribution of the background variables; we can then feed the revised distribution into the model of a counterfactual universe to predict what would have happened there. Thus in the story of Example 28.1, we know, from observations in the factual universe, that $f_Y(1, \varepsilon_Y) = -\log 15 = \lambda$, say. With this knowledge we can update the distribution of the background terms. In the imagined comparator universe, with $X \leftarrow 0$, we are interested in the implied response, $f_Y(0, \varepsilon_Y)$ —whose updated distribution can now be computed.

Appealing though this ploy is, it has many problems. For starters, when and how might one justify one’s model, with its background variables and functions constant across universes? There are also choices to be made as to just which variables are regarded as staying constant—why only the background variables and not some domain variables? In any problem of non-trivial complexity there will be various different stories that could be told, relating different universes in different ways. For example, there have been lawsuits by various states against tobacco companies, claiming that, if the companies had publicized their knowledge of the dangers of smoking when they first knew of them, many lives could have been saved. Damages

3. Because counter to known facts. The term counterfactual is often misused for other, less problematic concepts—see Dawid [2007b].

are sought for the additional costs placed on health services—meaning the excess cost in the actual world, over that of an imagined world in which they had made their knowledge public. But how should we imagine that world? One could reasonably argue that in such a world, by giving up smoking, people would have lived longer than they actually did. Then the actual (non-)actions of the tobacco companies might well have *decreased* the cost to the health services. But what seems to be required for the case at hand is to imagine a world where people had the same lifetimes, but were healthier, that is, to regard lifetimes, as well as background variables, as constant across universes.

More technically, but crucially, this approach falls foul of the non-uniqueness of the functional representation of an intervention DAG. Recall that in Example 28.1, consistent with the known properties of the domain variables in the factual world, we could take ε_Y to be a vector (E_0, E_1) , having a bivariate normal distribution with completely arbitrary correlation ρ ; and then take the function $f_Y(X, \varepsilon_Y)$ as E_X . Then on observing $Y = 1$ in the factual world, we learn $E_1 = \lambda$. Conditioning on this in the bivariate distribution of ε_Y , the updated distribution for E_0 is normal with mean $\rho(\lambda - 1)$ and variance $1 - \rho^2$. Moreover, in the counterfactual comparator universe, $Y = E_1$, so Y would likewise be assigned this distribution. The problem is that ρ is arbitrary and cannot be identified from any empirical evidence, so this approach does not supply an answer to our question, unless supplemented with further, non-empirically justified, stories about the relationship between real and imaginary universes.

However, this does not mean that nothing at all can be said about counterfactual variables. In the above example we must have $\rho \in [-1, 1]$, so we can infer that the counterfactual mean of Y lies between $\lambda - 1$ and $1 - \lambda$. Similar inequalities can be obtained for problems with binary responses, and these can often be improved by taking account of information on other variables in the system or data generated under different regimes [Tian and Pearl 2000, Dawid 2011, Dawid et al. 2016, 2017, 2019]. However, although there are very special circumstances where exact identification becomes possible [Tian and Pearl 2000], typically this is not the case, and we are left with a rather fuzzy answer to our question. But after all, imagination should never be too tightly constrained.

28.7 Conclusion

I hope I have shown the richness and variety of uses for a DAG. However, sometimes a DAG is used without adequate attention to its interpretation. Judea Pearl's clarity of understanding and exposition of the various applications of DAGs, at all levels of the ladder of causation, should serve as a valuable corrective and guide for all of us.

References

- N. Cartwright. 1994. *Nature's Capacities and Their Measurement*. Clarendon Press, Oxford. DOI: <https://doi.org/10.1093/0198235070.001.0001>.
- P. Constantinou and A. P. Dawid. 2017. Extended conditional independence and applications in causal inference. *Ann. Stat.* 45, 2618–2653. DOI: <https://doi.org/10.1214/16-AOS1537>.
- R. G. Cowell, A. P. Dawid, S. L. Lauritzen, and D. J. Spiegelhalter. 1999. *Probabilistic Networks and Expert Systems*. Springer, New York.
- A. P. Dawid. 1979. Conditional independence in statistical theory (with discussion). *J. R. Stat. Soc. Series B* 41, 1–31. <https://www.jstor.org/stable/2984718>.
- A. P. Dawid. 2000. Causal inference without counterfactuals (with discussion). *J. Am. Stat. Assoc.* 95, 407–448. DOI: <https://doi.org/10.2307/2669377>.
- A. P. Dawid. 2002. Influence diagrams for causal modelling and inference. *Int. Stat. Rev.* 70, 161–189. Corrigenda, *ibid.*, 437. DOI: <https://doi.org/10.2307/1403901>.
- A. P. Dawid. 2007a. Fundamentals of statistical causality. Research Report 279, Department of Statistical Science, University College London, 94 pp. https://www.ucl.ac.uk/drupal/site_statistics/sites/statistics/files/migrated-files/rr279.pdf.
- A. P. Dawid. 2007b. Counterfactuals, hypotheticals and potential responses: A philosophical examination of statistical causality. In F. Russo and J. Williamson (eds.), *Causality and Probability in the Sciences*, volume 5 of *Texts in Philosophy*. College Publications, London, 503–532.
- A. P. Dawid. 2010a. Beware of the DAG! In I. Guyon, D. Janzing, and B. Schölkopf (eds.), *Proceedings of the NIPS 2008 Workshop on Causality*, volume 6 of *Journal of Machine Learning Research Workshop and Conference Proceedings*. 59–86. <http://tinyurl.com/33va7tm>.
- A. P. Dawid. 2010b. Seeing and doing: The Pearl synthesis. In R. Dechter, H. Geffner, and J. Y. Halpern (eds.), *Heuristics, Probability and Causality: A Tribute to Judea Pearl*. College Publications, London, 296–311.
- A. P. Dawid. 2011. The role of scientific and statistical evidence in assessing causality. In R. Goldberg, (ed.), *Perspectives on Causation*. Hart Publishing, Oxford, 133–147.
- A. P. Dawid. 2015. Statistical causality from a decision-theoretic perspective. *Ann. Rev. Stat. Appl.* 2, 273–303. DOI: <https://doi.org/10.1146/annurev-statistics-010814-020105>.
- A. P. Dawid, R. Murtas, and M. Musio. 2016. Bounding the probability of causation in mediation analysis. In *Topics on Methodological and Applied Statistical Inference*. Springer, 75–84. DOI: https://doi.org/10.1007/978-3-319-44093-4_8.
- A. P. Dawid, M. Musio, and R. Murtas. 2017. The probability of causation. *Law Probab Risk.* 16, 163–179. DOI: <https://doi.org/10.1093/lpr/mgx012>.
- A. P. Dawid, M. Humphreys, and M. Musio. 2019. Bounding causes of effects with mediators. <http://arxiv.org/abs/1907.00399>.
- M. Frydenberg. 1990. The chain graph Markov property. *Scand. J. Stat.* 17, 333–353. <https://www.jstor.org/stable/4616181>.

- S. L. Lauritzen, A. P. Dawid, B. N. Larsen, and H.-G. Leimer. 1990. Independence properties of directed Markov fields. *Networks* 20, 491–505. DOI: <https://doi.org/10.1002/net.3230200503>.
- J. Pearl. 1988. *Probabilistic Inference in Intelligent Systems*. Morgan Kaufmann Publishers, San Mateo, CA. DOI: <https://doi.org/10.1016/C2009-0-27609-4>.
- J. Pearl. 1993. Aspects of graphical models connected with causality. In *Proceedings of the 49th Session of the International Statistical Institute*. 391–401.
- J. Pearl. 1995. Causal diagrams for empirical research (with discussion). *Biometrika* 82, 669–710. DOI: <https://doi.org/10.2307/2337329>.
- J. Pearl. 2009. *Causality: Models, Reasoning, and Inference* (2nd. ed.). Cambridge University Press, Cambridge. DOI: <https://doi.org/10.1017/CBO9780511803161>.
- J. Pearl and D. Mackenzie. 2018. *The Book of Why*. Basic Books, New York.
- P. Spirtes, C. Glymour, and R. Scheines. 2000. *Causation, Prediction, and Search* (2nd. ed.). Springer-Verlag, New York.
- J. Tian and J. Pearl. 2000. Probabilities of causation: Bounds and identification. *Ann. Math. Artif. Intell.* 28, 287–313. DOI: <https://doi.org/10.1023/A:1018912507879>.
- T. Verma and J. Pearl. 1990. Causal networks: Semantics and expressiveness. In R. D. Shachter, T. S. Levitt, L. N. Kanal, and J. F. Lemmer (eds.), *Uncertainty in Artificial Intelligence 4*. North-Holland, Amsterdam, 69–76. DOI: <https://doi.org/10.1016/B978-0-444-88650-7.50011-1>.

Instrumental Variables with Treatment-induced Selection: Exact Bias Results

Felix Elwert (University of Wisconsin–Madison),
Elan Segarra (University of Wisconsin–Madison)

Instrumental variable (IV) estimation suffers selection bias when the analysis conditions on the treatment. Judea Pearl's [2000, p. 248] early graphical definition of instrumental variables explicitly prohibited conditioning on the treatment. Nonetheless, the practice remains common. In this chapter, we derive exact analytic expressions for IV selection bias across a range of data-generating models, and for various selection-inducing procedures. We present four sets of results for linear models. First, IV selection bias depends on the conditioning procedure (covariate adjustment vs. sample truncation). Second, IV selection bias due to covariate adjustment is the limiting case of IV selection bias due to sample truncation. Third, in certain models, the IV and ordinary least squares (OLS) estimators under selection bound the true causal effect in large samples. Fourth, we characterize situations where IV remains preferred to OLS despite selection on the treatment. These results broaden the notion of IV selection bias beyond sample truncation, replace prior simulation findings with exact analytic formulas, and enable formal sensitivity analyses.

29.1 Introduction

Instrumental variable (IV) analysis is a popular approach for identifying causal effects when the treatment is confounded by omitted variables. IV analysis rests on two main assumptions: that the instrument is associated with the treatment

(“relevance”), and that the instrument is associated with the outcome only via the effect of treatment on the outcome (“exclusion”). The exclusion assumption is the sticking point of many empirical applications because it requires theoretical justification and is testable only to a very limited degree [e.g., [Balke and Pearl 1997](#), [Richardson and Robins 2010](#)].

One type of exclusion violation that has recently gained attention is selection bias [e.g., [Mogstad and Wiswall 2012](#), [Engberg et al. 2014](#), [Swanson et al. 2015](#), [Ertefaie et al. 2016](#), [Canan et al. 2017](#), [Hughes et al. 2019](#)]. We say that IV analysis suffers selection bias when conditioning (rather than not conditioning) on some variable violates the exclusion assumption. One particularly important case is *treatment-induced IV selection bias*: whenever treatment is confounded by unobservables, conditioning on a variable that has been affected by treatment induces bias. [Judea Pearl \[2000, p. 248\]](#) recognized this problem and presented the first definition of instrumental variables that outright prohibits conditioning on variables affected by treatment. Despite Pearl’s warning, however, conditioning on such “descendants” of treatment remains common in IV analysis.

Past research on treatment-induced IV selection bias [e.g., [Swanson et al. 2015](#), [Canan et al. 2017](#), [Gkatzionis and Burgess 2019](#), [Hughes et al. 2019](#)] is limited in two respects. First, it has focused on IV selection bias induced by sample truncation, which occurs when observations are excluded from the sample.¹ This focus neglects that other conditioning procedures, such as covariate adjustment, can also induce selection bias. Second, in situations where consistent estimators are not readily available, the literature characterizes the size and sign of IV selection bias by simulation. Without analytic bias expressions, however, it is unclear which stylized facts resulting from simulation studies hold generically.

This chapter makes two main contributions. First, we derive analytic expressions for treatment-induced IV selection bias for a range of different data-generating models. Second, we compare the biases resulting from two different selection-inducing conditioning procedures: sample truncation and covariate adjustment. For tractability, we focus on linear models with homogeneous (constant) effects and normal errors.

We highlight several results. First, the selection procedure matters. Within a given data-generating model, selection by truncation and selection by covariate

1. Some studies have proposed corrections, bounds, or sensitivity analyses for IV selection bias in certain truncation scenarios (e.g., [Mogstad and Wiswall \[2012\]](#), [Engberg et al. \[2014\]](#), [Canan et al. \[2017\]](#), [Vansteelandt et al. \[2018\]](#), [Gkatzionis and Burgess \[2019\]](#), [Hughes et al. \[2019\]](#)). These approaches often rely on knowing the selection probability of both the observed and the truncated observations.

adjustment introduce quantitatively different biases into IV analysis. Second, selection bias by adjustment is the limiting case of selection bias by truncation. Third, in certain models, the IV and OLS estimators with selection bound the true causal effect in large samples. Fourth, our analytic bias expressions characterize the models in which IV is less biased than OLS, which obtains when treatment does not exert an extreme effect on selection.

The rest of the chapter proceeds as follows. Section 29.2 reviews basic facts about directed acyclic graphs for linear models. Section 29.3 defines instrumental variables in econometric and graphical notation. Section 29.4 describes conditions under which selection violates the IV exclusion assumption and defines IV estimation under selection by truncation and covariate adjustment. Section 29.5 presents analytic expressions for the bias in IV and OLS estimators over a range of models with treatment-induced selection by truncation and by covariate adjustment. Section 29.6 concludes.

29.2 Causal Graphs

The challenge of selection bias in IV analysis is transparently communicated with graphical causal models [Pearl 2009, Maathuis et al. 2018]. Here, we review the basics. A *causal graph* represents the structural equations of the data-generating model. Causal graphs consist of *nodes* representing variables and *directed edges* representing direct causal effects. Causal graphs are assumed explicitly to display the observed and unobserved common causes of all variables. By convention, causal graphs do not explicitly display the idiosyncratic shocks that affect individual variables.

Throughout, we assume that the causal graphs represent linear data-generating models with homogeneous effects and normally distributed errors.² Without loss of generality, we further assume that all variables are standardized to have mean zero and unit variance. The direct causal effect of one variable on another variable in such models is given by its *path parameter*, which is bounded by $[-1,1]$. For example, the causal graph in Figure 29.1(a) represents the linear structural equations model given in Figure 29.1(b), with path parameters $\pi, \beta, \gamma, \delta_1$, and δ_2 . For each variable $V \in \{Z, U, T, S, Y\}$ the idiosyncratic shocks are marginally independent and normally distributed, $\varepsilon_V \sim N(0, \sigma_V^2)$, with variance σ_V^2 scaled so that each $V \sim N(0, 1)$. Since U is unobserved, the *structural error* term on Y in econometric terminology is $\omega_Y = \delta_2 U + \varepsilon_Y$. Notice that T is correlated with the structural error, $Cov(T, \omega_Y) \neq 0$, because both depend on the unobserved confounder, U .

2. Some results do not rest on the joint normality assumption, but our results on IV selection bias with truncation do.

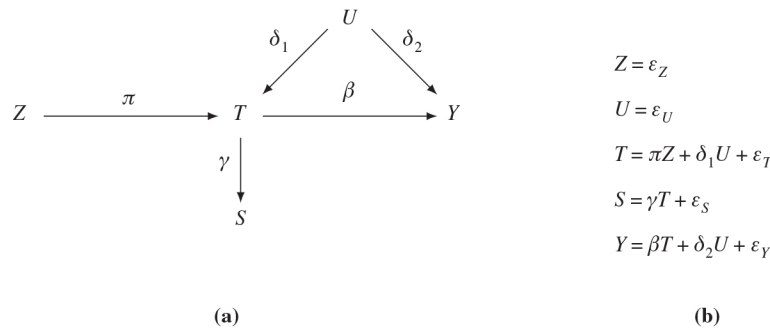


Figure 29.1 IV scenario where the selection variable is a function of treatment alone, equivalently displayed as a causal graph (a) and as a linear structural equations model (b).

Under mild conditions to avoid knife-edge cases, simple rules determine the covariance structure of data generated by a model [Pearl 2009]. The notions of paths, collider variables, and descendants play a central role in these rules. A *path* is an acyclic sequence of adjacent arrows between two variables, regardless of the direction of the arrows. In a *causal path* from treatment to outcome, all arrows point toward the outcome. In a *non-causal*, or *spurious*, path between treatment and outcome, at least one arrow points away from the outcome. A variable is called a *collider* with respect to a specific path if it receives two inbound arrows on the path. For example, T is a collider on the path $Z \rightarrow T \leftarrow U \rightarrow Y$. The *descendant set* of a variable contains all variables directly and indirectly caused by it, for example, $\text{desc}(T) = \{S, Y\}$ in Figure 29.1(a).

Two variables are statistically independent if all paths between them are closed; and two variables are statistically associated if there is at least one open path between them [Verma and Pearl 1988]. A path is *closed* (does not transmit association) iff either (a) it contains a collider and neither the collider nor any of its descendants are conditioned on, or (b) it contains a non-collider that is conditioned by exact stratification. A path is *open* (does transmit association) if it is not closed [Pearl 1988]. Importantly, when a path contains only one collider, then conditioning on this collider, or any of its descendants, opens this path.

The marginal covariance between two variables in a linear model with standardized variables is given by Wright's [1934] rule as the sum of the product of the path parameters on the open paths that connect the variables. For example, the marginal covariance between Z and Y in Figure 29.1(a) is $\text{Cov}(Y, Z) = \pi\beta$ because the path $Z \rightarrow T \rightarrow Y$ is the only open path (the other path, $Z \rightarrow T \leftarrow U \rightarrow Y$, is closed by the unconditioned collider T). The conditional covariance

between variables A and B , after adjusting for some covariate C , is $Cov(A, B | C) = Cov(A, B) - Cov(A, C)Cov(B, C)$. The novel bias results in this chapter hinge on deriving conditional covariances when *truncating* the sample as a function of C .

29.3 Instrumental Variables

Let T be the treatment variable of interest, Y be the outcome, Z be the candidate instrumental variable, and \mathbf{X} be a set of covariates. Econometrically, an instrumental variable is defined by two assumptions.

Definition 29.1 A variable, Z , is called an instrumental variable for the causal effect of T on Y , β , if, conditional on the set of covariates \mathbf{X} (which may be empty),

E1: Z is associated with T , $Cov(Z, T | \mathbf{X}) \neq 0$,

E2: Z is not associated with the structural error term, ω_Y , on Y , $Cov(Z, \omega_Y | \mathbf{X}) = 0$.

Assumption E1 is called *relevance*, and Assumption E2 is called *exclusion*. Pearl [2001] provides a graphical definition.

Definition 29.2 A variable, Z , is called an instrumental variable for the causal effect of T on Y , β , if, conditional on the set of covariates \mathbf{X} (which may be empty),

G1: There is at least one open path from Z to T conditional on \mathbf{X} ,

G2: \mathbf{X} does not contain descendants of Y , $\mathbf{X} \cap desc(Y) = \emptyset$,

G3: There is no open path from Z to Y conditional on \mathbf{X} , other than those paths that terminate in a causal path from T to Y .

Assumption G1 defines *relevance*, and Assumptions G2 and G3 together define *exclusion*. We say that a candidate instrumental variable is “valid” if it is relevant and excluded, and “invalid” otherwise. For example, in Figure 29.1(a), Z is a valid instrument without conditioning on S , since Z is relevant (associated with T) by the open path $Z \rightarrow T$, and Z is excluded (unassociated with the structural error term on Y) since the only open path from Z to Y , $Z \rightarrow T \rightarrow Y$, terminates in the causal effect of T on Y . When Z is a valid instrumental variable, then the standard IV estimator, given by the sample analog of

$$\beta_{IV} = \frac{Cov(Y, Z | \mathbf{X})}{Cov(T, Z | \mathbf{X})},$$

is consistent for the causal effect of T on Y in linear and homogeneous models. The numerator of this estimator is called the *reduced form* and the denominator is called the *first stage*. The behavior of this IV estimator is the focus of this chapter. For simplicity, we will henceforth write β_{IV} and β_{OLS} to refer to the probability lim-

its (as the sample size tends to infinity) of the standard IV and OLS estimators, respectively.

29.4 Selection Bias in IV: Qualitative Analysis

We say that the IV estimator suffers selection bias when conditioning on some variable violates the exclusion assumption. For example, conditioning on a variable that opens a path between Z and Y that does not terminate in the causal effect of T on Y violates exclusion both in the sense of G3 and E2. Hughes et al. [2019] catalogue several models in which selection violates exclusion.

We focus on the IV selection bias that results from conditioning on a descendant of T , $S \in \text{desc}(T)$. For example, in Figure 29.1(a), conditioning on S invalidates the use of Z as an instrumental variable because T is the only collider variable on the path $Z \rightarrow T \leftarrow U \rightarrow Y$, and conditioning on S as the descendant of the collider T opens this path. The association “transmitted” by this open path overtly violates the exclusion condition G3 and similarly violates the exclusion condition E2 since ω_Y is a function of U . This rationalizes why Pearl’s [2000, p. 248] early graphical IV definition outright rules out conditioning on descendants of treatment.

Since conditioning on a variable can result from many different procedures during data collection or data analysis, selection bias in IV analysis can result from many different procedures as well. Analysts should be aware, however, that different ways of conditioning on a variable may induce quantitatively different selection biases. In this chapter, we contrast selection bias resulting from two empirically common conditioning procedures: sample truncation and covariate adjustment.

Truncation occurs when observations are preferentially excluded from the sample [Bareinboim et al. 2014], for example, due to attrition or listwise deletion of missing data. Write $R=1$ for retained observations and $R=0$ for excluded (truncated) observations. Let S be the (possibly latent) continuous variable that determines truncation. We distinguish between interval truncation and point truncation. *Interval truncation* restricts the sample to observations with a range of values of S , for example, $R = \mathbf{1}(S \geq s_0)$ or $R = \mathbf{1}(s_1 \geq S \geq s_0)$, where $\mathbf{1}(\cdot)$ is the indicator function. A limiting case of interval truncation is *point truncation*, where the sample is restricted to units with a single value of S , $R = \mathbf{1}(S = s_0)$. The truncated IV estimator is given by

$$\beta_{IV|tr} = \frac{\text{Cov}(Z, Y | R = 1)}{\text{Cov}(Z, T | R = 1)}.$$

With truncation (as opposed to censoring) the analyst does not have access to the truncated observations, cannot estimate the probability of truncation, and hence cannot use inverse-probability weights to correct for truncation [Canan et al. 2017,

Gkatzionis and Burgess 2019]. In Figure 29.1(a), a truncated sample would involve the empiricist observing $\{Z, T, Y\}$ only for units with $R = 1$.

Although selection can also occur due to *covariate adjustment* for S , this procedure has received less attention in the literature on IV selection bias. With covariate adjustment the analyst observes $\{Z, T, S, Y\}$ for all units. Adjustment involves first exactly stratifying on S , computing the estimator within each stratum, and then averaging across the marginal distribution of S . Thus, the IV estimator under adjustment on S is given by

$$\beta_{IV|Adj} = \int \frac{\text{Cov}(Z, Y | S = s)}{\text{Cov}(Z, T | S = s)} f_S(s) ds,$$

where $f_S(s)$ is the marginal distribution of S . In linear models, controlling for a variable as a main effect in OLS or 2SLS amounts to covariate adjustment on the variable [Angrist and Pischke 2008].

Next, we analytically characterize selection bias in IV analysis and OLS regression for various data-generating models and provide intuition.

29.5 Selection Bias in IV: Quantitative Analysis

This section derives exact analytic expressions for selection bias across a range of common data-generating models. For each model, we contrast the selection bias for the IV and the OLS estimators, resulting from two different conditioning strategies. First, we present the selection bias resulting from covariate adjustment on S . Next, we newly derive the selection bias from interval truncation on S , $R = \mathbf{1}(S \geq s_0)$. We assume a probit link between S and the binary selection indicator, R .³ Since IV analysis suffers small-sample bias regardless of selection, we study its large-sample behavior (asymptotic bias).

29.5.1 Selection as a Function of Treatment Alone

Consider the most basic scenario of IV selection bias in Figure 29.1(a). As stated above, Z in this model is a valid instrumental variable for the causal effect of T on Y , β , if the analysis does not condition on S . Conditioning on S , however, invalidates Z as an instrumental variable because S is a descendant of T , and T is a collider on the path $Z \rightarrow T \leftarrow U \rightarrow Y$. Conditioning on S opens this path, which induces an association between Z and Y via U and hence violates the exclusion condition.

Proposition 29.1 gives the selection bias in the standard IV estimator when the analysis adjusts for S .

3. Numerical simulations in prior work have assumed logit selection [Canan et al. 2017, Gkatzionis and Burgess 2019, Hughes et al. 2019]. Switching to probit selection captures the same intuition but gains analytic tractability.

Proposition 29.1 In a linear and homogeneous model with normal errors represented by Figure 29.1(a) and covariate adjustment on S , the standard instrumental variable estimator converges in probability to

$$\beta_{IV|Adj} = \beta - \delta_1 \delta_2 \frac{\gamma^2}{1 - \gamma^2}.$$

The proof follows from regression algebra and Wright's rule [Wright 1934]. The magnitude of selection bias due to covariate adjustment in the IV estimator depends on two components. First, selection bias increases with the strength of unobserved confounding between T and Y via U , $\delta_1 \delta_2$ (which corresponds to the path $Z \rightarrow T \leftarrow U \rightarrow Y$ that is opened by conditioning on S , less the first stage $Z \rightarrow T$). Second, selection bias increases with the effect of the treatment T on the selection variable, S , γ . When $\gamma = 0$, S contains no information about the collider T , conditioning on S does not open the path $Z \rightarrow T \leftarrow U \rightarrow Y$, and selection bias is zero. By contrast, as $|\gamma| \rightarrow 1$, the magnitude of the bias increases without bound because adjusting for S increasingly amounts to adjusting for the collider T itself, while at the same time reducing the first stage. (If the analysis directly adjusted for T , then the first stage would go to zero and the IV estimator would not be defined.)

Proposition 29.2 derives the IV selection bias due to interval truncation on S .

Proposition 29.2 In a linear and homogeneous model with normal errors represented by Figure 29.1(a) and truncation on S , $R = \mathbf{1}(S \geq s_0)$, the standard instrumental variable estimator converges in probability to

$$\beta_{IV|Tr} = \beta - \delta_1 \delta_2 \frac{\psi \gamma^2}{1 - \psi \gamma^2}, \quad \text{where } \psi = \frac{\phi(s_0)}{1 - \Phi(s_0)} \left(\frac{\phi(s_0)}{1 - \Phi(s_0)} - s_0 \right),$$

and $\phi(\cdot)$ and $\Phi(\cdot)$ are the standard normal pdf and cdf, respectively.

Proposition 29.2 (proved in Appendix 29.A.1) illustrates that IV selection bias due to truncation (Proposition 29.2) differs from IV selection bias due to adjustment (Proposition 29.1) only in that truncation deflates the contribution of the effect of T on S , γ , by the factor $\psi \in (0, 1)$. Since ψ is the derivative of the standard normal hazard function, it monotonically increases with the *severity of truncation*, $Pr(R = 0) = \Phi(s_0)$, as shown in Figure 29.2(a). Hence, interval truncation leads to less IV selection bias than covariate adjustment in Figure 29.1(a),

Corollary 29.1 In a linear and homogeneous model with normal errors represented by Figure 29.1(a), the magnitude of IV-adjustment bias is weakly larger than that of IV-truncation bias: $|\beta_{IV|Adj} - \beta| \geq |\beta_{IV|Tr} - \beta|$.

Corollary 29.1 makes intuitive sense. Adjustment involves first exactly stratifying and then averaging across strata defined by $S = s$. Exact stratification on S uses

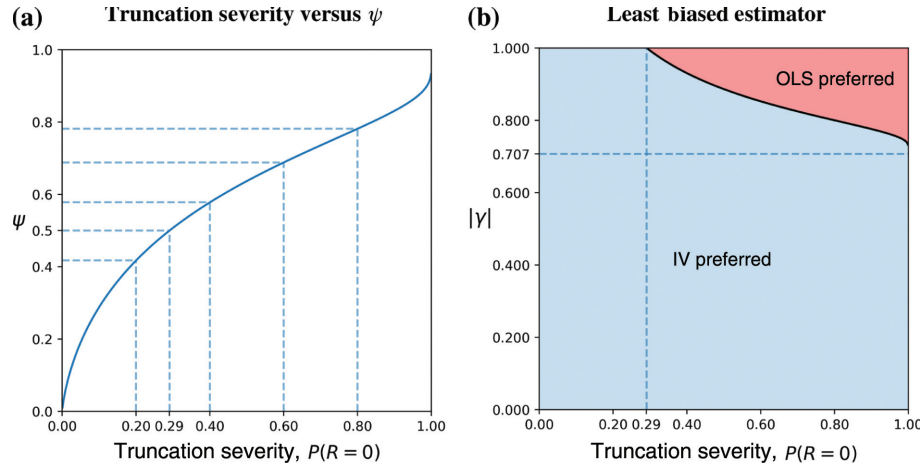


Figure 29.2 (a) ψ monotonically increases with truncation severity. (b) Whether OLS or IV is less biased under selection depends on truncation severity and the effect of T on S , $|\gamma|$.

all information about T that is contained in S , hence opening the biasing path as much as conditioning on S possibly can. By contrast, interval truncation amounts to imprecise stratification on S (retaining observations across a range of values on S , but not exactly stratifying on any particular value), hence opening the biasing path less.

Of some methodological interest, we further note, in Figure 29.1(a), that IV selection bias by truncation converges on IV selection bias by covariate adjustment as the severity of truncation increases to shrink the remaining sample to a single point. Proposition 29.3 states that this observation is true for all models, not only for Figure 29.1(a).

Proposition 29.3 In a linear and homogeneous model with normal errors, selection bias in the standard instrumental variable estimator due to covariate adjustment is the limiting case of selection bias due to point truncation,

$$\lim_{s_0 \rightarrow \infty} \beta_{IV|Tr} = \beta_{IV|Adj}. \tag{29.1}$$

This proposition makes intuitive sense. Covariate adjustment involves exact stratification on $S = s$, which defines point truncation. Since the probability limits of all s -stratum specific estimators are identical in linear Gaussian models, selection bias by adjustment equals selection bias by point truncation. The proof in Appendix 29.A.2 formalizes this intuition.

Proposition 29.2 helps inform empirical choices in practice. When selection is unavoidable (e.g., because the data were truncated during data collection), should

analysts choose IV or OLS? Figure 29.2(b) shows that the IV estimator is preferred to OLS, with respect to bias, for most combinations of γ and truncation severity. Since OLS bias (with or without truncation) only depends on unobserved confounding, that is, $\beta_{OLS|Tr} - \beta = \delta_1\delta_2$, the difference in magnitude between the OLS and IV biases with truncation is given by

$$|\beta_{OLS|Tr} - \beta| - |\beta_{IV|Tr} - \beta| = |\delta_1\delta_2| \frac{1 - 2\psi\gamma^2}{1 - \psi\gamma^2}.$$

Hence, the IV estimator is preferred when $\psi\gamma^2 \leq \frac{1}{2}$. Specifically, when fewer than 29.1% of observations are truncated (corresponding to $\psi \leq 0.5$), IV is preferred regardless of the effect of T on S , γ . Conversely, when $|\gamma| < \sqrt{0.5} \approx 0.707$, no amount of truncation makes OLS preferable over IV. Recalling that γ cannot exceed 1 in magnitude, the selection variable S would have to be an extraordinarily strong proxy for T to make IV more biased than OLS at any level of truncation.

Perhaps most useful for practice, we note that selection bias (by truncation or adjustment) in Figure 29.1(a) is proportional to the negative of OLS confounding bias. Therefore, the OLS and IV estimators under selection bound the true causal effect.

Corollary 29.2 In a linear and homogeneous model with normal errors represented by Figure 29.1(a), the OLS estimator and the instrumental variable estimator with selection bound the causal effect of T on Y , β ,

$$\begin{aligned} \beta_{IV|Tr} &\leq \beta \leq \beta_{OLS}, & \text{when } \delta_1\delta_2 > 0, \\ \beta_{IV|Tr} &\geq \beta \geq \beta_{OLS}, & \text{when } \delta_1\delta_2 < 0. \end{aligned}$$

The fact that the IV selection bias has the opposite sign of the OLS selection bias in Figure 29.1(a) is owed to linearity and homogeneity: in linear and homogeneous models, conditioning on a collider or its descendant reverses the sign of the product of the path parameters for the associated path. For example, if all path parameters along the biasing path $Z \rightarrow T \leftarrow U \rightarrow Y$ are positive, then conditioning on $S \in \text{desc}(T)$ will induce a negative association along this path. Since the IV bias hinges on conditioning on S , the selection bias would be negative. By contrast, OLS bias in Figure 29.1(a) does not hinge on conditioning on S and instead results from confounding along $T \leftarrow U \rightarrow Y$. Therefore, OLS bias would be positive.

29.5.2 Selection as a Function of a Mediator

Next, consider models in which the selection variable, S , is a mediator of the effect of treatment on the outcome, as in the causal graphs in Figure 29.3(a) and (b).

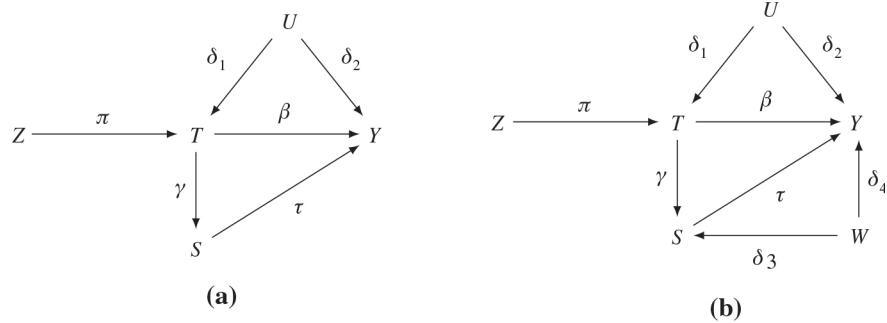


Figure 29.3 IV scenarios where the selection variable is both a descendant of treatment and a mediator.

These situations are worth investigating for two reasons: first, empiricists are often interested in the direct causal effect of T on Y , which necessitates conditioning on S ; second, they result in qualitatively different bias representations.

Suppose that the analyst is interested in the direct causal effect of T on Y , β , in the model of Figure 29.3(a). The bias in the IV and OLS estimators under interval truncation and adjustment for S is given in Proposition 29.4.

Proposition 29.4 In a linear and homogeneous model with normal errors represented by Figure 29.3(a), the standard instrumental variable estimator with selection on S , converges in probability to

$$\beta_{IV|S} = \beta - \delta_1\delta_2 \frac{\psi\gamma^2}{1 - \psi\gamma^2} + \gamma\tau \frac{1 - \psi}{1 - \psi\gamma^2},$$

and the OLS estimator with selection on S converges in probability to

$$\beta_{OLS|S} = \beta + \delta_1\delta_2 + \gamma\tau \frac{1 - \psi}{1 - \psi\gamma^2},$$

where

$$\psi = \begin{cases} \frac{\phi(s_0)}{1 - \Phi(s_0)} \left(\frac{\phi(s_0)}{1 - \Phi(s_0)} - s_0 \right) & \text{with truncation on } S, R = \mathbf{1}(S \geq s_0) \\ 1 & \text{with adjustment on } S \end{cases}.$$

All bias expressions in Proposition 29.4 have a straightforward graphical interpretation. With *adjustment* on S , the indirect causal path $T \rightarrow S \rightarrow Y$ is completely blocked, because S is a non-collider on this path. Hence, the bias in the IV and OLS estimators with adjustment on S equals the IV and OLS adjustment biases

in Figure 29.1(a), where S was not a mediator. With adjustment on S , IV is biased by selection, whereas OLS is biased by confounding; IV selection bias will generally be smaller in magnitude than OLS confounding bias (unless the effect of T on S is very large); and IV and OLS with adjustment bound the true direct causal effect.

With *truncation* on S , however, the indirect path $T \rightarrow S \rightarrow Y$ is not completely blocked and hence contributes a new term to both IV and OLS bias. For both IV and OLS, this term equals the strength of the partially blocked indirect path, $\gamma\tau$, deflated by the multiplier $0 \leq (1 - \psi)/(1 - \psi\gamma^2) \leq 1$. The size of the multiplier depends both on the truncation severity, ψ , and on the effect of T on S , γ , but in opposite directions. As γ is fixed and truncation increases, $\psi \rightarrow 1$, the analysis conditions ever more precisely on an ever smaller range of values of S ; hence the indirect path is increasingly blocked, and both the multiplier and the bias term tend to 0. By contrast, when ψ is fixed and the effect of T on S increases, $|\gamma| \rightarrow 1$, the information about T contained in S increases, the multiplier tends to 1, and the path is increasingly opened.

By Proposition 29.3, it remains true in Figure 29.3(a) that IV selection bias due to adjustment is the limiting case of IV selection bias due to point truncation. However, it is no longer necessarily true that IV with adjustment is more biased than IV with truncation. The bias ordering now depends on the signs and relative sizes of the two additive bias term (representing the biasing paths $T \leftarrow U \rightarrow Y$ and $T \rightarrow S \rightarrow Y$), and on how well the indirect path $T \rightarrow S \rightarrow Y$ is closed by truncation. Hence, when selection is made on a mediator of the treatment effect, selection bias by adjustment could be larger or smaller in magnitude than selection bias by truncation. Bounding the true causal effect also becomes more difficult. With truncation on S , IV and OLS with selection do not necessarily bound the true direct causal effect.

The analysis is further complicated when the effect of S on Y is confounded by some unobserved variable, W , as in Figure 29.3(b). This situation is arguably more realistic than the model in Figure 29.3(a), because mediators in observational studies are expected to be confounded. Here, conditioning on S (by adjustment or truncation) in IV analysis opens a new path, $Z \rightarrow T \rightarrow S \leftarrow W \rightarrow Y$, which violates the exclusion assumption; and in OLS it opens $T \rightarrow S \leftarrow W \rightarrow Y$, which biases OLS regression. The resulting bias expressions are the same as those in Proposition 29.4 with an additional bias term, $-\gamma\delta_3\delta_4\frac{\psi}{1-\psi\gamma^2}$. Once more, IV selection bias due to adjustment is the limiting case of IV selection bias due to point truncation. However, no pair of estimators (among $\beta_{IV|Tr}$, $\beta_{IV|Adj}$, $\beta_{OLS|Tr}$, $\beta_{OLS|Adj}$) can be relied on to bound the true direct causal effect in the model of Figure 29.3(b).

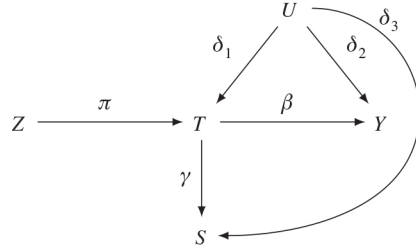


Figure 29.4 IV scenario where the selection variable is both a descendant of the treatment and the unobserved confounder.

29.5.3 Selection on Treatment and the Unobserved Confounder

Finally, we consider situations where the selection variable, S , is also a descendant of the unobserved U that confounds the effect of treatment on the outcome (Figure 29.4).

Proposition 29.5 In a linear and homogeneous model with normal errors represented by Figure 29.4, the standard instrumental variable estimator with selection on S converges in probability to

$$\beta_{IV|S} = \beta - \delta_1 \delta_2 \frac{\psi \gamma^2}{1 - \psi \gamma (\gamma + \delta_1 \delta_3)} - \gamma \delta_3 \delta_2 \frac{\psi}{1 - \psi \gamma (\gamma + \delta_1 \delta_3)},$$

and the OLS estimator with selection on S converges in probability to

$$\beta_{OLS|S} = \beta + \delta_1 \delta_2 \frac{1 - \psi (\gamma^2 + \gamma \delta_1 \delta_3 + \delta_3^2)}{1 - \psi \gamma (\gamma + \delta_1 \delta_3)^2} - \gamma \delta_3 \delta_2 \frac{\psi}{1 - \psi \gamma (\gamma + \delta_1 \delta_3)^2},$$

where

$$\psi = \begin{cases} \frac{\phi(s_0)}{1 - \Phi(s_0)} \left(\frac{\phi(s_0)}{1 - \Phi(s_0)} - s_0 \right) & \text{with truncation on } S, R = \mathbf{1}(S \geq s_0) \\ 1 & \text{with adjustment on } S \end{cases}.$$

Three points stand out about selection bias in Figure 29.4. First, when S is a descendant of both T and U , conditioning on S opens a new path, $T \rightarrow S \leftarrow U \rightarrow Y$, which biases IV and OLS with adjustment or truncation on S .

Second, in contrast to models considered previously, the bias term associated with each biasing path ($T \leftarrow U \rightarrow Y$ and $T \rightarrow S \leftarrow U \rightarrow Y$) is now a function of the path parameters of both paths. In other words, the path-specific biases interact. Pearl's graphical causal models provide intuition for this interaction. Consider, for example, the second bias term. First, conditioning on S opens the

path $T \rightarrow S \leftarrow U \rightarrow Y$. Hence, the bias term depends on $\gamma\delta_3\delta_2$. Second, conditioning on S also absorbs variance from U (a non-collider on $T \rightarrow S \leftarrow U \rightarrow Y$) because S is a descendant of U along the path $U \rightarrow T \rightarrow S$. Hence, the bias term also depends on δ_1 .

Third, the direction of the interaction, and hence the overall bias, depends on the specific parameter values. This makes the bias order of these estimators fairly unpredictable and prevents generic recommendations for or against any one estimator. This ambiguity provides additional motivation for using exact bias formulas for sensitivity analysis.

29.6 Conclusion

Conditioning on the wrong variable can induce selection bias in IV analysis. When consistent estimators are not available, analysts should gauge the bias in their estimators by principled speculation or formal sensitivity analysis. To enable this work, we have derived analytic expressions for IV selection biases that have previously been characterized only by simulation.

Our analysis specifically focused on scenarios where selection is a function of a confounded treatment. Judea Pearl's [2000] graphical IV criterion specifically prohibited conditioning on a descendant of treatment. But the practice appears to remain common, thereby calling for formal analysis. Our analytic expressions present asymptotic IV selection bias in terms of substantively interpretable standardized path parameters for Gaussian models. Empowered by Pearl's graphical causal models, we further provided intuition by decomposing the bias into terms that map onto the paths in the data-generating model that are opened (or closed) by selection. Leveraging prior knowledge or principled theory, analysts may use our bias expressions to conduct formal sensitivity analyses by populating the free parameters to derive the size of the bias. Even with partial information our expressions may provide informative bounds on the bias.

We present three broad conclusions. First, in the models we investigated, IV selection bias depends on three ingredients: (i) the strength of each biasing path in the model, (ii) the effect of treatment on the selection variable, $|\gamma|$; and (iii) truncation severity, ψ , i.e., the share of the full sample excluded from the analysis by truncation. The magnitude of the bias term associated with each biasing path increases with the strength of the path, with $|\gamma|$, and with truncation severity, ψ , if selection is made on a collider or descendant of a collider on the path; and the magnitude of the bias term increases with the strength of the path and with $|\gamma|$, but decreases with ψ , if selection is made on a non-collider on the biasing path.

Second, the sign and magnitude of IV selection bias depend on the selection procedure: in all linear Gaussian IV models, the bias induced by covariate

adjustment is the limiting case of bias induced by point truncation. This does not mean that adjustment bias is always larger than truncation bias, only that adjustment bias equals truncation bias if truncation had reduced the sample to a single point.

Third, rather usefully, in some models (where selection is only a function of treatment and the selection variable is not a mediator) IV and OLS suffer selection biases of opposite signs, such that these estimators bound the true causal effect. In the same models, unless the effect of treatment on selection is very large, IV with selection suffers less bias than OLS with or without selection.

29.A Appendix

29.A.1 Proof of Truncation Bias Expressions

We derive the bias under truncation by leveraging a result from [Tallis 1965].

Lemma 29.1 Let $V \in \mathbb{R}^k$ follow a multivariate normal distribution, $V \sim N(0, \Sigma)$, and define the truncated random vector $\tilde{V} = \{v \in V : c'v \geq p\}$ with $p \in \mathbb{R}$, $c \in \mathbb{R}^k$, and $|c| = 1$. Then the expectation and variance of the truncated random vector are given by

$$\begin{aligned} E[\tilde{V}] &= \Sigma c \kappa^{-1} \lambda\left(\frac{p}{\kappa}\right) \\ \text{Var}(\tilde{V}) &= \Sigma - \Sigma c c' \Sigma \kappa^{-2} \psi \end{aligned}$$

where $\kappa = (c' \Sigma c)^{-1/2}$, $\lambda(x) = \frac{\phi(x)}{1 - \Phi(x)}$ is the hazard function of the standard normal distribution, and

$$\psi = \lambda\left(\frac{p}{\kappa}\right) \left(\lambda\left(\frac{p}{\kappa}\right) - \frac{p}{\kappa} \right).$$

Using properties of the standard normal hazard function it can be shown that ψ is in fact the derivative of the hazard function.

Proof of Proposition 29.2. Consider the model described by Figure 29.1(a). Since the idiosyncratic shocks are all normally distributed, all variables in the model are normally distributed. Specifically for vectors $V = [Z \ U \ T \ S \ Y]'$ and $\varepsilon = [\varepsilon_Z \ \varepsilon_U \ \varepsilon_T \ \varepsilon_S \ \varepsilon_Y]'$, the standardized⁴ model has the reduced form $V = \Gamma \varepsilon$,

4. Standardization implies non-unit variance for some of the shocks. For example, when $\text{Var}(T) = 1$, then ε_T is $\text{Var}(\varepsilon_T) = 1 - \pi^2 - \delta_1^2$.

where $\varepsilon \sim N(0, \Sigma_\varepsilon)$ and

$$\Gamma = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ \pi & \delta_1 & 1 & 0 & 0 \\ \gamma\pi & \gamma\delta_1 & \gamma & 1 & 0 \\ \beta\pi & \beta\delta_1 + \delta_2 & \beta & 0 & 1 \end{bmatrix}$$

$$\Sigma_\varepsilon = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 - \pi^2 - \delta_1^2 & 0 & 0 \\ 0 & 0 & 0 & 1 - \gamma^2 & 0 \\ 0 & 0 & 0 & 0 & 1 - \beta^2 - \delta_2^2 - 2\beta\delta_1\delta_2 \end{bmatrix}.$$

Since this implies that $V \sim N(0, \Gamma\Sigma_\varepsilon\Gamma')$, our truncation scenario, $R = \mathbf{1}(S \geq s_0)$, allows for direct application of Lemma 29.1 to derive the covariance matrix of the truncated distribution, $\tilde{V} = V|S \geq s_0$. For Lemma 29.1, $c = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 \end{bmatrix}'$, $p = s_0$, and $\Sigma = \Gamma\Sigma_\varepsilon\Gamma'$. This implies $\kappa = 1$ and thus

$$\text{Var}(\tilde{V}) = \Gamma\Sigma_\varepsilon\Gamma' - \Gamma\Sigma_\varepsilon\Gamma'cc'\Gamma\Sigma_\varepsilon\Gamma'\psi \quad \text{where} \quad \psi = \lambda(s_0)(\lambda(s_0) - s_0).$$

Finally, the IV estimand with truncation is given by the ratio of the truncated covariance between instrument and outcome and the truncated covariance between instrument and treatment. After some enjoyable algebra, we evaluate $\text{Var}(\tilde{V})$, extract the relevant covariances, and obtain

$$\beta_{IV|tr} = \frac{\text{Cov}(Z, Y|S \geq s_0)}{\text{Cov}(Z, T|S \geq s_0)} = \frac{\beta\pi - \psi\gamma\pi(\beta\gamma + \gamma\delta_1\delta_2)}{\pi - \psi\gamma^2\pi} = \beta - \delta_1\delta_2 \frac{\psi\gamma^2}{1 - \psi\gamma^2}.$$

■

The proofs of Propositions 29.4 and 29.5 proceed analogously, using the appropriate reduced form matrix, Γ , for each scenario.

29.A.2 Proof of Adjustment as Point Truncation (Proposition 29.3)

Proof. Define the stratum-specific IV estimator when $S = s$ as

$$\beta_{IV|s}(s) = \frac{\text{Cov}(Z, Y|S = s)}{\text{Cov}(Z, T|S = s)}$$

Notice that $\beta_{IV|s}(s)$ is the IV estimator under point truncation (i.e., the limit of the interval truncated estimator as the interval collapses to a point).

In a homogeneous linear model with normal errors, $V = \begin{bmatrix} Z & U & T & S & Y \end{bmatrix}'$ will follow a multivariate normal distribution. Multivariate normal distributions have the useful property that their conditional distributions have constant covariances across the conditioning level. Hence, for all $V_1, V_2, V_3 \in \{Z, U, T, S, Y\}$ and $v_0, v_1 \in \mathbb{R}$, we have that

$$\text{Cov}(V_1, V_2 | V_3 = v_0) = \text{Cov}(V_1, V_2 | V_3 = v_1).$$

It follows that $\beta_{IV|S}(s_0) = \beta_{IV|S}(s_1)$ for any $s_0, s_1 \in \mathbb{R}$. Since the stratum-specific IV estimator is constant across strata of S , this implies that the IV estimator under adjustment on S is the same as any stratum-specific IV estimator. ■

References

- J. D. Angrist and J.-S. Pischke. 2008. *Mostly Harmless Econometrics*. Princeton University Press, Princeton, NJ. DOI: <https://doi.org/10.1515/9781400829828>.
- A. Balke and J. Pearl. 1997. Bounds on treatment effects from studies with imperfect compliance. *J. Am. Stat. Assoc.* 92, 439, 1171–1176. DOI: <https://doi.org/10.1080/01621459.1997.10474074>.
- E. Bareinboim, J. Tian, and J. Pearl. 2014. Recovering from selection bias in causal and statistical inference. In *Proceedings of the 28th AAAI Conference on Artificial Intelligence*, Quebec, Canada. 2410–2416.
- C. Canan, C. Lesko, and B. Lau. 2017. Instrumental variable analyses and selection bias. *Epidemiology* 28, 3, 396–398. DOI: <https://doi.org/10.1097/EDE.0000000000000639>.
- J. Engberg, D. Epple, J. Imbrogno, H. Sieg, and R. Zimmer. 2014. Evaluating education programs that have lotteried admission and selective attrition. *J. Labor Econ.* 32, 1, 27–63. DOI: <https://doi.org/10.1086/671797>.
- A. Ertefaie, D. Small, J. Flory, and S. Hennessy. 2016. Selection bias when using instrumental variable methods to compare two treatments but more than two treatments are available. *Int. J. Biostat.* 12, 1, 219–232. DOI: <https://doi.org/10.1515/ijb-2015-0006>.
- A. Gkatzionis and S. Burgess. 2019. Contextualizing selection bias in Mendelian randomization: how bad is it likely to be? *Int. J. Epidemiol.* 48, 3, 691–701. DOI: <https://doi.org/10.1093/ije/dyy202>.
- R. A. Hughes, N. M. Davies, G. Davey Smith, and K. Tilling. 2019. Selection bias when estimating average treatment effects using one-sample instrumental variable analysis. *Epidemiology* 30, 3, 350–357. DOI: <https://doi.org/10.1097/EDE.0000000000000972>.
- M. Maathuis, M. Drton, S. Lauritzen, and M. Wainwright (Eds.). 2018. *Handbook of Graphical Models* (1st. ed.). CRC Press, Boca Raton, FL. DOI: <https://doi.org/10.1201/9780429463976>.
- M. Mogstad and M. Wiswall. 2012. Instrumental variables estimation with partially missing instruments. *Econ. Lett.* 114, 2, 186–189. DOI: <https://doi.org/10.1016/j.econlet.2011.10.013>.

- J. Pearl. 1988. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers Inc., San Francisco, CA.
- J. Pearl. 2000. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York.
- J. Pearl. 2001. *Parameter Identification: A New Perspective* (second draft). Technical Report R-276. UCLA Cognitive Systems Laboratory.
- J. Pearl. 2009. *Causality* (2nd. ed.). Cambridge University Press, New York.
- T. S. Richardson and J. M. Robins. 2010. Analysis of the binary instrumental variable model. In *Heuristics, Probability and Causality: A Tribute to Judea Pearl*. College Publications, 415–444.
- S. A. Swanson, J. M. Robins, M. Miller, and M. A. Hernán. 2015. Selecting on treatment: A pervasive form of bias in instrumental variable analyses. *Am. J. Epidemiol.* 284, 1–7. DOI: <https://doi.org/10.1093/aje/kwu284>.
- G. M. Tallis. 1965. Plane truncation in normal populations. *J. R. Stat. Soc.* 27, 2, 301–307. DOI: <https://doi.org/10.1111/j.2517-6161.1965.tb01497.x>.
- S. Vansteelandt, S. Walter, and E. Tchetgen Tchetgen. 2018. Eliminating survivor bias in two-stage instrumental variable estimators. *Epidemiology* 29, 4, 536–541. DOI: <https://doi.org/10.1097/EDE.0000000000000835>.
- T. Verma and J. Pearl. 1988. Causal networks: Semantics and expressiveness. In *Proceedings of the Fourth Annual Conference on Uncertainty in Artificial Intelligence*. North-Holland Publishing Co, Amsterdam, Netherlands, 69–78. DOI: <https://doi.org/10.1016/B978-0-444-88650-7.50011-1>.
- S. Wright. 1934. The method of path coefficients. *Ann. Math. Stat.* 5, 3, 161–215. DOI: <https://doi.org/10.1214/aoms/1177732676>.

Causal Models and Cognitive Development

Alison Gopnik (University of California at Berkeley)

Pearl's work has had an important influence on the field of cognitive development. In particular, in hundreds of empirical studies, causal models, combining ideas about probability, intervention, and counterfactuals, have turned out to play an essential role in children's everyday knowledge. Even very young children learn such models from data in the ways that Pearl suggested. New frontiers in the project of understanding children's causal learning include sampling, active search and experimentation, and combining causal models with deep learning and deep reinforcement learning techniques.

In the year 2000, more than 20 years ago, my graduate students and I made a weekly trek across the campus and up the hill to the computer science department. We were there as part of a reading group discussing a brand-new book, *Causality* by Judea Pearl. Those students went on to become distinguished faculty, and 20 years later, they and *their* students, and many other psychologists, are still working on problems that were inspired by that book and those conversations. So am I.

Why would developmental psychologists, usually found sitting in tiny chairs opposite toddlers in preschools, immerse themselves in a volume of statistics and equations? The book, and Pearl's work, in general, speaks to a foundational problem that is at the core of the study of cognitive development. Cognitive development and machine learning belong to the same natural category, along with the philosophy of science, epistemology, and vision science, even if they live in opposite corners of the campus. (And all these disciplines are in a different natural category than sociologically closer ones like adult cognitive psychology, cognitive neuroscience, and philosophy of mind.)

Developmental psychology, machine learning, and philosophy of science might seem like strange bedfellows, but they are all trying to solve the same

problem—sometimes called the problem of induction. How can we know anything about the world around us? After all, the information that reaches us from that world is just a stream of photons at our retinas and disturbances of air at our ear drums. And yet we come to know about people and poodles, tables and toys, quarks and quasars. How is this possible? We seem to have abstract, hierarchical, structured representations of the world around us, and those representations allow us to make wide-ranging generalizations and predictions. And yet, we also seem to somehow construct those representations from data that is concrete, messy, and particular.

Going back to Plato and Aristotle, there have been two basic approaches to solving this problem. The nativist option is simply to deny that the abstract representations *are* derived from the data. Instead, they are there innately, from a past life or in the mind of God, for Plato and Descartes, because of evolution for more recent thinkers. The other, empiricist, option is to deny that the abstract representations exist—simply combine enough statistical data and you can do all the same inferential work. This approach goes all the way back to Aristotle and Locke but also underpins many of the most recent approaches to machine learning.

For people who actually study the development of human knowledge, whether as developmental psychologists or philosophers of science, these alternatives have always seemed unsatisfying. When we actually look in detail at the progress of children's thinking, or the progress of science, we do, in fact, see both abstract representations and qualitative changes in those representations in the light of new evidence.

In the past, Jean Piaget, the great founder of cognitive development, argued for “constructivism” as an alternative to nativism and empiricism, and philosophers like Carnap and Kuhn, who were actually both influenced by developmental psychology as well as the history of science, articulated similar ideas. In the 1980s, a number of psychologists including me, Susan Carey, Henry Wellman, and Susan Gelman, articulated the “theory theory”—the idea that children's conceptual development could be understood by analogy to scientific theory formation, explicitly connecting conceptual development and scientific theory change [Carey 1985, Gopnik 1988, Wellman and Gelman 1992]. “Theory theory” researchers could qualitatively describe children's representations as theories and chart the changes in those representations as children learned more. The research program made a great deal of empirical progress, describing the development of intuitive physics [Smith et al. 1985], biology [Carey 1985] and especially intuitive psychology or “theory of mind” [Gopnik and Wellman 1994].

The problem, though, was that there was no computational way of characterizing the constructive process that was responsible for theory formation and

change either in childhood or in science. The overarching faith of cognitive science is that the mind is a computational system instantiated in the brain. In some areas of cognitive science, particularly vision science, we really had begun to redeem that faith and solve the problem of induction computationally and even neurally. Building on 100 years of perception and psychophysics, vision scientists could begin to describe how the visual system recovers information about objects and space from the light patterns on the retina, and computer vision systems could start to instantiate those ideas (e.g., Marr [1982]). There has been remarkable progress on this project since; although, of course, there is still much work to be done.

But doing the same thing for theories, whether these were the everyday theories of childhood or the theories of formal science, seemed like an impossibly forbidding task. Indeed, in the early 1990s there was a kind of nihilism about solving such problems, reflected in both the philosophy of science on the one hand, and in statistics, on the other hand. The slogans of the time were “there is a logic of confirmation but no logic of discovery” and “no causation from correlation.”

This was where Pearl’s work came in. Although theories involve many kinds of representations, certainly causal representations are crucial, both in everyday cognition and in science. And for a long time, going back at least to David Hume, causal knowledge was one of the canonical cases of the problem of induction. As Hume pointed out, it seemed impossible to see how simply observing the constant conjunction of two events could lead you to a causal conclusion, and yet such inferences are ubiquitous in both everyday cognition and in science. The pessimism about scientific induction very much extended to causation. Bertrand Russell famously said, “The law of causality, I believe, like much that passes muster among philosophers, is a relic of a bygone age, surviving, like the monarchy, only because it is erroneously supposed to do no harm.”

In the 1990s, two developments coming from very different directions restored the reign of causality and articulated a computational account of causal inference in the form of “causal Bayes nets.” One set of developments came from Pearl’s initial work on expert reasoning [Pearl 1988]. Initially, Pearl’s project was to find a way a computer could generate complex judgments and predictions about conditional dependencies in the way that experts, like doctors, do (if, but only if, the patient has a fever and green phlegm as well as a cough, and tests for viruses are negative, antibiotics will help). It turned out that the best way to do this was to equip the system with causal models, integrating ideas about probability, intervention, and counterfactual inference. In parallel, the philosophers of science Peter Spirtes, Richard Scheines, and Clark Glymour at Carnegie Mellon University formulated very similar mathematical ideas [Spirtes et al. 1993]. Moreover, philosophers like

James Woodward, working in a more traditional philosophical framework, used these ideas to characterize the very nature of causation [Woodward 2003].

The central idea was to use graphical models as a way of representing the relations among variables in a causal system, and to systematically relate those relations to the conditional probability of the variables. Within this system it was possible to define the effects of interventions (in what Pearl called the “do-calculus”) as well as counterfactuals. The distinction between associations and predictions, on the one hand, and interventions and counterfactuals on the other hand, is the crucial distinction that separates mere correlation from causation. I might notice correlations both between having yellow nicotine-stained fingers and getting lung cancer, and between smoking and getting lung cancer, and make the appropriate prediction that someone who has yellow fingers or who smokes is more likely to have cancer. But an intervention to wash the yellow off a patient’s fingers won’t have any effect on the probability of cancer, while an intervention to stop smoking will. Similarly, the counterfactual that if the patient had washed his hands he wouldn’t have gotten cancer is false while the similar counterfactual about smoking is true.

The entire causal Bayes net system allowed for interwoven inferences about probability, intervention, and counterfactuals in a way that captures many of the central elements of causation both in everyday life and in science. If you knew the causal structure you could make accurate predictions, interventions, and counterfactual inferences, and significantly, the formalism naturally distinguished between these different kinds of inferences.

The Bayes net formalism also had important implications for causal learning and the problem of causal induction. The formalism made systematic connections between the structure of the causal graphs and data about the conditional probability of events and the outcomes of interventions. This meant that in principle the inferences could be reversed—if we knew about the conditional probability of variables and the outcomes of interventions on those variables, we could accurately infer the causal structure. And this, in turn, suggested a computational solution to the classic problem of theory induction. Scientists had always used evidence from statistics (i.e., patterns of conditional probability) and experiments (i.e., systematic interventions on variables) to infer causal structure. But thanks to the new work on causal Bayes nets we could begin to explain mathematically how and why this actually worked.

Initially, nobody thought of these systems as potential models of everyday human cognition, let alone children’s cognition. (I have an email exchange with Clark Glymour from 1989, where I suggested children might be doing something similar to Bayes nets, his initial response was that these systems were precisely designed to do things that humans couldn’t). By the late 1990s though, this idea

had come to seem more appealing, and at least worth testing. Even if children couldn't make inferences about complex systems with hundreds of variables, would they use the same basic principles to uncover causal structure?

At the time essentially all of the work of children's causal reasoning, and adults', for that matter, fell into one of two camps, either researchers who emphasized the role of reasoning about physical mechanisms in causal understanding, or those who saw causal reasoning as merely an extension of simple association. The combination of graphical models, probability, intervention, and counterfactuals was an entirely new way of approaching the subject.

Glymour and I decided to test whether children might do something like causal Bayes net inference with a new method—the blicket detector—a machine that lights up and plays music when you put some things on it and not others. The first question, which we tested with my student David Sobel, one of the participants in the *Causality* reading group and now at Brown, was whether children could make any causal inferences with this method (they could) [Gopnik and Sobel 2000]. By 2000, we realized that we could use simple methods like this to test more complex inferences, of the sort that Pearl described. In particular, could children use conditional probability and intervention to make inferences? (they could) [Gopnik et al. 2001]. After one of the reading group meetings, my student Laura Schulz, now at the Massachusetts Institute of Technology, raced excitedly down the hill from the computer science department to the hardware store, where she constructed a toy with two gears and a switch to test whether children could infer different causal structures (chains vs. common causes, for instance) from the pattern of interventions and answer counterfactual questions (they could) [Schulz et al. 2007]. By 2004, we had shown that preschoolers could determine the direction of causal arrows, infer unobserved variables, and design novel interventions, and that they did so in a way that fit much more naturally with Pearl's and Spirtes, Scheines, and Glymour's ideas than any of the traditional views of causal knowledge [Glymour 2002, Gopnik et al. 2004]. In 2005, the McDonnell Foundation funded a large interdisciplinary grant combining developmental psychologists, philosophers, and computationalists to work more on these ideas [see Gopnik and Schulz 2007, Gopnik 2012, Gopnik and Wellman 2012].

Over the next 10 years, this work continued and expanded. Although the causal Bayes net formalism was particularly elegantly designed and relatively easy to implement, it was to begin with, at least, rather limited in scope. The causal graphs were limited to describing systems of variables at a single level of description. A number of psychologists and cognitive scientists, notably Josh Tenenbaum, Tom Griffiths, and Noah Goodman, who were all involved in the McDonnell collaborative, argued for a much more expansive and general version of the project that

Pearl started, including a wide range of probabilistic generative models with different kinds of logical structure and including hierarchical as well as single-level models [see Griffiths and Tenenbaum 2007, 2009, Griffiths et al. 2010, Goodman et al. 2011, review in Tenenbaum et al. 2011].

This became an important and pervasive movement within cognitive science. It is often described as the “Bayesian” approach but this is something of a misnomer. The Bayesian part of the idea is simply this. If you have a probabilistic generative model, like a causal Bayes net, and can therefore systematically predict the probability of a pattern of evidence given that model, then you can invert this inference in a Bayesian way to infer the probability of the model given the evidence. But all the work is done by the specifics of the generative model, how well it is linked to the data, and how feasible it is to perform the Bayesian inversion and solve the search problems that result. Causal Bayes nets were and remain one of the best examples of how a probabilistic generative model could actually work.

Fei Xu, another developmental psychologist who pioneered the idea of probabilistic generative models [Xu and Tenenbaum 2007], came up with the term “rational constructivism” [Xu and Kushnir 2012], which is perhaps the best way of describing the enterprise. I suspect that the popularity of the Bayesian terminology partly reflects a principle I call The Tyranny of the Euphonious Monosyllable—if Kolmogorov had discovered Bayes’ rule it wouldn’t have taken off as a descriptor. But it certainly could, and perhaps should, be called Pearl-y Cognitive Science instead.

Further work in my lab and others over the next 15 years showed that very young children could make Pearl-y causal inferences across a wide range of domains, including “theory of mind.” Tamar Kushnir, now at Duke, yet another student who had been part of the Pearl reading group, showed that even 18-month-olds could use Pearl-y methods to infer other people’s preferences and desires [Kushnir et al. 2010]. One interesting body of work has argued that children use something like an intuitive utility calculus—a representation of the causal relationships between goals and actions—to understand other people [Hamlin et al. 2013, Lucas et al. 2014]. Kushnir and I and others showed that children and even infants were remarkably skilled at tracking and using conditional probabilities [Saffran et al. 1996, Kushnir and Gopnik 2005, Xu and Garcia 2008]. We and others also showed that children were not limited to making inferences about specific causal relationships. Instead, they could also infer quite abstract features of causal structure, such as whether causal structures were disjunctive or conjunctive. In fact, in some circumstances they could do this better than adults [Dewar and Xu 2010,

[Lucas et al. 2014](#), [Gopnik et al. 2017](#)]. Moreover, we recently showed that low-income children in Peru and in Head Start programs in the USA were just as good at making these inferences as the usual middle-class American samples [[Wente et al. 2019](#)].

In short, across what are now hundreds of studies from dozens of labs with thousands of children, it turns out that if you give children a particular pattern of data they can infer which causal structure was most likely to have generated that data, and can design new interventions and counterfactuals on that basis, in precisely the way that Pearl described.

So far, this is a largely triumphal story. But as always in science, advances lead to new problems and much of the most interesting recent work in cognitive science focuses on those problems.

One of the strengths of probabilistic generative models such as Pearl's is precisely that they are probabilistic. Earlier attempts to solve the problem of induction, such as Noam Chomsky's theory of how children infer grammars from linguistic data, were deterministic. Either a grammar was supported by the data or it wasn't. This also meant that induction was radically underdetermined—there was almost never a way of definitely ruling a grammar in or out given the data, and that led to Chomsky's nativist conclusions. The Bayesian probabilistic model approach in contrast, considers a wide range of hypotheses and tries to determine how likely each hypothesis is given the data and your prior knowledge.

But there's a catch. The catch is that for the Bayesian inversion trick to work you need to have some way of searching among the possible hypotheses and testing them against the data. Even for a relatively restricted set of representations like simple causal graphs with a limited number of variables, this problem quickly becomes untenable—there are simply too many possibilities to consider. And as the range of representations we consider becomes more abstract and complex, as with hierarchical Bayes nets, for example, or “language of thought” probabilistic logics, the search problem just becomes hairier.

Much of the exciting recent work in cognitive science, following up on Pearl's work, tries to find solutions to the search problem. Two approaches are especially interesting and exciting. First, in the computational literature the search problem is often solved by some form of sampling, randomly but systematically testing some hypotheses rather than others (e.g., [Roberts and Casella \[1999\]](#)). At least in “asymptopia,” as one statistician calls it, these sampling methods can approximate full Bayesian inference. My collaborator Tom Griffiths and I and a number of others have shown that both adults and children show the signatures of this kind of sampling [[Vul and Pashler 2008](#), [Denison et al. 2012](#), [Ullman et al. 2012](#), [Bonawitz](#)

[et al. 2014](#)]. How these sampling measures could be extended and how randomness and systematicity are combined are fascinating directions for the future.

Active learning is an even more interesting and underexplored way of solving the search problem. The relationship between causal structure and intervention means that interventions can be deliberately designed to reveal causal structure, as in scientific experiments. In the early work on causal Bayes nets the assumption was that systems passively absorbed patterns of data and matched them against the potential graphical structures. When we began our first *blicket* detector experiments, I remember remarking that one of the big advantages of working with computers over kids was that computers weren't constantly trying to grab the blocks and try them on the machine! That observation has turned into a very productive research program, particularly as pursued by Laura Schulz and her student and my post-doc Elizabeth Bonawitz, now at Harvard. Schulz and Bonawitz have shown that children's spontaneous play often involves active interventions that are designed to resolve causal ambiguities and recover causal models [[Schulz and Bonawitz 2007](#), [Schulz et al. 2008](#), [Schulz 2012](#)]. The philosopher of science Frederick Eberhardt, now at the California Institute of Technology, another product of the McDonnell collaborative, has pursued a similar project in the context of science—systematically using the formalism to describe how experiments can reveal causal structure [[Eberhardt and Scheines 2007](#)].

A final frontier is the integration of causal inference and the more empiricist and statistical forms of learning, such as “deep learning” and “deep reinforcement” learning that have led to the very recent renaissance of artificial intelligence (AI), and were the subject of the 2018 Turing prize. Although these techniques have turned out to be surprisingly effective, they are beginning to come up against significant limitations. In particular, they allow only limited kinds of generalizations, and they require very large data sets and supervised forms of learning.

Increasingly, AI researchers are turning back to combine the neural network techniques with Pearl's work on causal models and the empirical work in cognitive development to try to design systems that have the power and flexibility of children's learning. For example, causality and cognitive development both play a central role in the recent DARPA machine common sense program, which we are part of at Berkeley.

Perhaps it is symbolic that the Berkeley Artificial Intelligence Research unit, of which I am now a member, just moved into the same building as the Developmental Psychology group. Both geographically and intellectually, the distance between the two fields is beginning to disappear. We very much have Judea Pearl to thank for that.

References

- E. Bonawitz, S. Denison, T. Griffiths, and A. Gopnik. 2014. Probabilistic models, learning algorithms, response variability: Sampling in cognitive development. *Trends Cogn. Sci.* 18, 497–500. DOI: <https://dx.doi.org/10.1016/j.tics.2014.06.006>.
- S. Carey. 1985. *Conceptual Change in Childhood*. MIT Press, Cambridge, MA.
- S. Denison, E. Bonawitz, A. Gopnik, and T. L. Griffiths. 2012. Rational variability in children's causal inferences: The Sampling Hypothesis. *Cognition* 126, 285–300. DOI: <https://doi.org/10.1016/j.cognition.2012.10.010>.
- K. M. Dewar and F. Xu. 2010. Induction, overhypothesis, and the origin of abstract knowledge. *Psychol. Sci.* 21, 12, 1871–1877. DOI: <https://dx.doi.org/10.1177/0956797610388810>.
- F. Eberhardt and R. Scheines. 2007. Interventions and causal inference. *Philos. Sci.* 74, 5, 981–995. DOI: <https://doi.org/10.1086/525638>.
- C. N. Glymour. 2002. *The Mind's Arrows: Bayes Nets and Graphical Causal Models in Psychology*. MIT Press, Cambridge, MA. DOI: <https://doi.org/10.1093/mind/112.446.340>.
- N. D. Goodman, T. D. Ullman, and J. B. Tenenbaum. 2011. Learning a theory of causality. *Psychol. Rev.* 118, 1, 110–119. DOI: <https://doi.org/10.1037/a0021336>.
- A. Gopnik. 1988. Conceptual and semantic development as theory change: The case of object permanence. *Mind Lang.* 3, 3, 197–216. DOI: <https://doi.org/10.1111/j.1468-0017.1988.tb00143.x>.
- A. Gopnik. 2012. Scientific thinking in very young children: Theoretical advances, empirical discoveries and policy implications. *Science* 337, 6102, 1623–1627. DOI: <https://doi.org/10.1126/science.1223416>.
- A. Gopnik and H. M. Wellman. 1994. The theory theory. In L. Hirschfeld and S. Gelman (Eds.), *Domain Specificity in Cognition and Culture*. Cambridge University Press, New York, 257–293. DOI: <https://doi.org/10.1017/CBO9780511752902.011>.
- A. Gopnik and D. Sobel. 2000. Detecting blickets: How young children use information about novel causal powers in categorization and induction. *Child Dev.* 71, 5, 1205–1222. DOI: <https://doi.org/10.1111/1467-8624.00224>.
- A. Gopnik and L. Schulz. (Eds.). 2007. *Causal Learning: Psychology, Philosophy, and Computation*. Oxford University Press, Inc., New York. DOI: <https://doi.org/10.1093/acprof:oso/9780195176803.001.0001>.
- A. Gopnik and H. M. Wellman. 2012. Reconstructing constructivism: Causal models, Bayesian learning mechanisms, and the theory theory. *Psychol. Bull.* 138, 6, 1085–1108. DOI: <https://doi.org/10.1037/a0028044>.
- A. Gopnik, D. M. Sobel, L. E. Schulz, and C. Glymour. 2001. Causal learning mechanisms in very young children: Two-, three-, and four-year-olds infer causal relations from patterns of variation and covariation. *Dev. Psychol.* 37, 5, 620–629. DOI: <https://doi.org/10.1037/0012-1649.37.5.620>.
- A. Gopnik, C. Glymour, D. M. Sobel, L. E. Schulz, T. Kushnir, and D. Danks. 2004. A theory of causal learning in children: Causal maps and Bayes nets. *Psychol. Rev.* 111, 1, 3–32. DOI: <https://doi.org/10.1037/0033-295X.111.1.3>.

- A. Gopnik, S. O'Grady, C. G. Lucas, T. L. Griffiths, A. Wente, S. Bridgers, R. Aboody, H. Fung, and R. E. Dahl. 2017. Changes in cognitive flexibility and hypothesis search across human life history from childhood to adolescence to adulthood. *Proc. Natl. Acad. Sci. U S A* 114, 30, 7892–7899. DOI: <https://doi.org/10.1073/pnas.1700811114>.
- T. Griffiths and J. B. Tenenbaum. 2007. Two proposals for causal grammars. In A. Gopnik & L. Schulz (Eds.), *Causal Learning: Psychology, Philosophy, and Computation*. Oxford University Press, Inc., New York, 323–345. DOI: <https://doi.org/10.1093/acprof:oso/9780195176803.003.0021>.
- T. L. Griffiths and J. B. Tenenbaum. 2009. Theory-based causal induction. *Psychol. Rev.* 116, 4, 661–716. DOI: <https://doi.org/10.1037/a0017201>.
- T. L. Griffiths, N. Chater, C. Kemp, A. Perfors, and J. B. Tenenbaum. 2010. Probabilistic models of cognition: Exploring representations and inductive biases. *Trends Cogn. Sci.* 14, 8, 357–364. DOI: <https://doi.org/10.1016/j.tics.2010.05.004>.
- K. Hamlin, T. Ullman, J. Tenenbaum, N. Goodman, and C. Baker. 2013. The mentalistic basis of core social cognition: Experiments in preverbal infants and a computational model. *Dev. Sci.* 16, 2, 209–226. DOI: <https://doi.org/10.1111/desc.12017>.
- T. S. Kuhn. 1962. *The Structure of Scientific Revolutions*. University of Chicago Press, Chicago. DOI: <https://doi.org/10.1119/1.1969660>.
- T. Kushnir and A. Gopnik. 2005. Young children infer causal strength from probabilities and interventions. *Psychol. Sci.* 16, 9, 678–683. DOI: <https://doi.org/10.1111/j.1467-9280.2005.01595.x>.
- T. Kushnir, F. Xu, and H. M. Wellman. 2010. Young children use statistical sampling to infer the preferences of other people. *Psychol. Sci.* 21, 8, 1134–1140. DOI: <https://doi.org/10.1177/0956797610376652>.
- C. G. Lucas, T. L. Griffiths, F. Xu, C. Fawcett, A. Gopnik, T. Kushnir, L. Markson, and J. Hu. 2014. The child as econometrician: A rational model of preference understanding in children. *PLoS One* 9, e92160. DOI: <https://doi.org/10.1371/journal.pone.0092160>.
- C. Lucas, S. Bridgers, T. Griffiths, and A. Gopnik. 2014. When children are better (or at least more open-minded) learners than adults: Developmental differences in learning the forms of causal relationships. *Cognition*. 131, 2, 284–299. DOI: <https://doi.org/10.1016/j.cognition.2013.12.010>.
- D. Marr. 1982. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. W.H. Freeman, San Francisco. DOI: <https://doi.org/10.7551/mitpress/9780262514620.001.0001>.
- J. Pearl. 1988. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufman, San Mateo, CA.
- J. Pearl. 2000. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York. DOI: <https://doi.org/10.1017/S0266466603004109>.
- C. P. Robert and G. Casella. 1999. *Monte Carlo Statistical Methods*. Springer-Verlag, New York.

- J. R. Saffran, R. N. Aslin, and R. N. Newport. 1996. Statistical learning by 8-month-old infants. *Science* 274, 5294, 1926–1928. DOI: <https://doi.org/10.1126/science.274.5294.1926>.
- L. E. Schulz. 2012. Origins of inquiry: Inductive inference and exploration in young children. *Trends Cogn. Sci.* 16, 7, 382–389. DOI: <https://doi.org/10.1016/j.tics.2012.06.004>.
- L. E. Schulz and E. B. Bonawitz. 2007. Serious fun: Preschoolers engage in more exploratory play when evidence is confounded. *Dev. Psychol.* 43, 4, 1045–1050. DOI: <https://doi.org/10.1037/0012-1649.43.4.1045>.
- L. E. Schulz, A. Gopnik, and C. Glymour. 2007. Preschool children learn about causal structure from conditional interventions. *Dev. Sci.* 10, 3, 322–332. DOI: <https://doi.org/10.1111/j.1467-7687.2007.00587.x>.
- L. E. Schulz, H. R. Standing, and E. B. Bonawitz. 2008. Word, thought, and deed: The role of object categories in children's inductive inferences and exploratory play. *Dev. Psychol.* 44, 5, 1266–1276. DOI: <https://doi.org/10.1037/0012-1649.44.5.1266>.
- C. Smith, S. Carey, and M. Wiser. 1985. On differentiation: A case study of the development of the concepts of size, weight, and density. *Cognition* 21, 3, 177–237. DOI: [https://doi.org/10.1016/0010-0277\(85\)90025-3](https://doi.org/10.1016/0010-0277(85)90025-3).
- P. Spirtes, C. Glymour, and R. Scheines. 1993. *Causation, Prediction and Search, Springer Lecture Notes in Statistics* (2nd. ed.). MIT Press, Cambridge, MA, 2000.
- J. B. Tenenbaum, C. Kemp, T. L. Griffiths, and N. D. Goodman. 2011. How to grow a mind: Statistics, structure, and abstraction. *Science* 331, 6022, 1279–1285. DOI: <https://doi.org/10.1126/science.1192788>.
- T. D. Ullman, N. D. Goodman, and J. B. Tennenbaum. 2012. Theory acquisition as stochastic search. *Cogn. Dev.* 27, 4, 455–480. DOI: <https://doi.org/10.1016/j.cogdev.2012.07.005>.
- E. Vul and H. Pashler. 2008. Measuring the crowd within: Probabilistic representations within individuals. *Psychol. Sci.* 19, 7, 645–647. DOI: <https://doi.org/10.1111/j.1467-9280.2008.02136.x>.
- H. M. Wellman and S. A. Gelman. 1992. Cognitive development: Foundational theories of core domains. *Ann. Rev. Psychol.* 43, 337–375. DOI: <https://doi.org/10.1146/annurev.ps.43.020192.002005>.
- A. Wente, K. Kimura, C. Walker, N. Banerjee, M. Fernández Flecha, B. MacDonald, C. Lucas, and A. Gopnik. 2019. Causal learning across culture and SES. *Child Dev.* 90, 859–875. DOI: <https://doi.org/10.1111/cdev.12943>.
- J. Woodward. 2003. *Making Things Happen: A Theory of Causal Explanation*. Oxford University Press, New York. DOI: <https://doi.org/10.1093/0195155270.001.0001>.
- F. Xu and V. Garcia. 2008. Intuitive statistics by 8-month-old infants. *Proc. Natl. Acad. Sci. USA* 105, 13, 5012–5015. DOI: <https://doi.org/10.1073/pnas.0704450105>.
- F. Xu and J. B. Tenenbaum. 2007. Word learning as Bayesian inference. *Psychol. Rev.* 114, 2, 245–272. DOI: <https://doi.org/10.1037/0033-295X.114.2.245>.
- F. Xu and T. Kushnir. 2012. (Vol. Ed.), *Rational Constructivism in Cognitive Development*. Academic Press, Elsevier Inc., Waltham, MA, 161–189.

The Causal Foundations of Applied Probability and Statistics

Sander Greenland (University of California, Los Angeles)

Abstract

Statistical science (as opposed to mathematical statistics) involves far more than probability theory, for it requires realistic causal models of data generators—even for purely descriptive goals. Statistical decision theory requires more causality: rational decisions are actions taken to minimize costs while maximizing benefits, and thus require explication of causes of loss and gain. Competent statistical practice thus integrates logic, context, and probability into scientific inference and decision using narratives filled with causality. This reality was seen and accounted for intuitively by the founders of modern statistics but was not well recognized in the ensuing statistical theory (which focused instead on the causally inert properties of probability measures). Nonetheless, both statistical foundations and basic statistics can and should be taught using formal causal models. The causal view of statistical science fits within a broader information-processing framework that illuminates and unifies frequentist, Bayesian, and related probability-based foundations of statistics. Causality theory can thus be seen as a key component connecting computation to contextual information, not “extra-statistical” but instead essential for sound statistical training and applications.

The only immediate utility of all the sciences is to teach us how to control and regulate future events through their causes. – Hume [1748]

31.1 Introduction: Scientific Inference is a Branch of Causality Theory

I will argue that realistic and thus scientifically relevant statistical theory is best viewed as a subdomain of causality theory, not a separate entity or an extension of probability. In particular, the application of statistics (and indeed most technology) must deal with causation if it is to represent adequately the underlying reality of how we came to observe what was seen—that is, the causal network leading to the data.¹ The network we deploy for analysis incorporates whatever time-order and independence assumptions we use for interpreting observed associations, whether those assumptions are derived from background (contextual) or design information [Pearl 1995, 2009, Robins 2001]. In making this case, I will invoke Pearl’s own arguments (e.g., as in Pearl [2009], Wasserstein [2018]) to deduce that statistics should integrate causal networks into its basic teachings and indeed into its entire theory, starting with the probability and bias models that are used to build up statistical methods and interpret their outputs.

Every real data analysis has a causal component comprising the causal network assumed to have created the dataset. Decision analysis has a further causal component showing the effects of decisions. Although these causal components are usually left implicit, a primary purpose of design strategies is to rule out alternative causal explanations for observations. Consider one of the most advanced research projects of all time, the search for the Higgs boson. Almost all statistical attention focused on the one-sided 5-sigma detection criterion [Lamb 2012], roughly equivalent to an α -level of 0.0000003, or requiring at least $-\log_2(0.0000003) = 22$ bits of information against the null [Greenland 2019] to declare detection. Yet the causal component is just as important: it includes every attempt to eliminate explanations for such extreme deviations other than the Higgs boson, for example, the painstaking checks of equipment are actions taken to block the mechanisms that could cause anything near that deviation (other than the Higgs mechanism itself).

Thus, because statistical analyses need a causal skeleton to connect to the world, causality is not extra-statistical but instead is a logical antecedent of real-world inferences. Claims of random or “ignorable” or “unbiased” sampling or allocation are justified by causal actions to block (“control”) unwanted causal effects

1. This view arguably applies even when dealing with quantum phenomena, at least in the QBist view [Mermin 2016]. In that view, the laws of quantum mechanics describe how equipment settings causally affect individual perceptions, where the latter become formalized as coherent predictive bets or frequency claims about subsequent observations under those settings (in contrast to other theories that treat quantum probabilities as properties of the environment). Such a controversial view is, however, unnecessary for the everyday applications of probability and causation that typify most of science and technology, and so will not be pursued here.

on the sample patterns. Without such actions of causal blocking, independence can only be treated as a subjective exchangeability assumption whose justification requires detailed contextual information about the absence of factors capable of causally influencing both selection (including selection for treatment) and outcomes [Greenland 1990]. Otherwise, it is essential to consider pathways for the causation of biases (non-random, systematic errors) and their interactions [Pearl 1995, Greenland et al. 1999, Maclure and Schneeweiss 2001, Hernán et al. 2004, Greenland 2010a, 2012a, 2021].

The remainder of the present chapter elaborates on the following points: probability is inadequate as a foundation for applied statistics because competent statistical practice integrates logic, context, and probability into scientific inference and decision, using causal narratives to explain diverse data [Greenland et al. 2004]. Thus, given the absence of elaborated causality discussions in statistics textbooks and coursework, we should not be surprised at the widespread misuse and misinterpretation of statistical methods and results. This is why incorporation of causality into introductory statistics is needed as urgently as other far more modest yet equally resisted reforms involving shifts in labels and interpretations for P-values and interval estimates.²

As a preliminary, consider that the Merriam-Webster Online Dictionary [2019] defines statistics as “a branch of mathematics dealing with data collection, organization, analysis, interpretation and presentation.” Many working statisticians (including me) regard the “branch of mathematics” portion as abjectly wrong, akin to calling physics, computer science, or any other heavily mathematical field a branch of mathematics. But we can fix that by replacing “branch of mathematics” with “science” to obtain

Statistics is the science of data collection, organization, analysis, interpretation and presentation, often in the service of decision analysis.

The amended definition makes no explicit mention of *either* probability or causation, but it is implicitly causal throughout, describing a sequence of actions with at least partial time ordering, each of which is capable of affecting subsequent actions: study design affects actions during data collection (e.g., restrictions on selection); these actions along with events during data collection (e.g., censoring) affect the data that result; these actions and events affect (or should affect) the study description and the data analysis; and the analysis results will affect the presentation. Overall, the presumed causal structure of this sequence supplies

2. Such as replacement of misleading terms like “statistical significance” and “confidence” by more modest terms like “compatibility” [Greenland 2017b, 2019, Amrhein et al. 2019, McShane et al. 2019, Wasserstein et al. 2019, Greenland and Rafi 2020, Rafi and Greenland 2020].

the basis for a justifiable interpretation of the study. Thus, whether answering the most esoteric scientific questions or the most mundane administrative ones, and whether the question is descriptive, causal, or purely predictive, causal reasoning will be crucially involved (albeit often hidden to ill effect in equations and assumptions used to get the “results”).

31.2 Causality is Central Even for Purely Descriptive Goals

As Pearl has often noted, causal descriptions encode the information and goals that lead to concerns about associations [Pearl 2009]. Consider survey statistics, in which the target question is not itself causal, merely descriptive, such as the proportion of voters who would vote for a given candidate. A competent survey researcher will be concerned about what characteristics C will affect both survey participation ($S = 1$) and voting intent V . Using square brackets to indicate that the observations are conditioned on $S = 1$, this concern is encoded in the diagram

$$[S = 1] \leftarrow C \rightarrow V,$$

in which we can see bias in the sampling estimator for the preference distribution $\Pr(V = v)$ will be induced by the selection on S . If instead we said only that the concern is about characteristics that are *associated* with both participation and preference (as in $S \leftrightarrow C \leftrightarrow V$), we would obscure the contextual basis for the concern.

To paraphrase Pearl, statistical analysis without causality is like medicine without physiology. As an example, if we see a difference in ethnic distributions (C) between our survey and population demographic data, we should be concerned about mis-estimating (say) the proportion of Trump voters in the target population. This concern is not because “white ethnicity is *associated* with voting for Trump” as some academic descriptions would have it, but because we expect that being a white male causes sympathy (or prevents antipathy) for Trump’s pronouncements relative to being black. That expectation arises from a simple causal relation encoded in $C \rightarrow S$, which creates the concern about only seeing preferences of those in the survey, that is, seeing only $\Pr(V = v | S = 1)$.

When survey methods attempt to adjust for the difference by reweighting the sample using the target-population ethnicity distribution, that adjustment can be seen as an attempt to counterbalance the $C \rightarrow S$ arrow in the mechanism generating the sample. This added computation in producing a reweighted sample is traditionally treated as a purely numeric artifice, but is also a causal process: someone must physically obtain target-weight data and program the reweighting to create the adjusted estimate. It is misleading to describe this action as “simulating

removal of an arrow”; it is instead the addition to the data generator of a weighting intervention W in a new causal pathway within

$$[S = 1] \leftarrow C \rightarrow V \leftarrow W \leftarrow C.$$

W is engineered to (hopefully) balance out the bias from conditioning on selection $[S = 1]$. Note that C appears twice in this diagram to allow it to be written in one line; writing it twice separates the initial effect of C on voter preferences (V) and sample formation (participation S) from its later effect on the analysis weighting W .

31.3 The Strength of Probabilistic Independence Demands Physical Independence

By data generator, I do *not* mean some abstract structural equation but rather the entire set of actual physical mechanisms that produce our observations. Even in the simplest games of chance, it is the *physical* (mechanical, causal) independence of coin tosses which licenses our teaching that betting systems for toss sequences will fail to beat simple expectations based on the frequency of heads observed so far. A causal diagram for a sequence of independent identically distributed (i.i.d.) tosses with outcome indicators Y_1, \dots, Y_N would thus show these N indicators as N isolated (unconnected) nodes.³ More generally, every missing arrow implies an independence assumption, and such an assumption is really a large *set* of assumptions on the joint distribution of the data Y_1, \dots, Y_N .

One way to measure the information in or logical strength of an independence assumption is by the number of logically independent constraints it imposes (equivalent to the number of parameters whose value it specifies, or the number of dimensions or degrees of freedom it removes from further consideration). Allowing for any possible dependency pattern (as suggested by “non-parametric”) among the Y_1, \dots, Y_N yields a measure of order N factorial; even if we count only pairwise dependencies, the number of patterns is of order N^2 (see Appendix 31.A). Either way, when described honestly, an i.i.d. assumption is not one assumption but rather a *set* of assumptions that grows far faster than the number of observations N . The amount of deductive (digital, syntactical) information in this assumption set is thus beyond anything data frequencies alone could contain; only contextual (background and design) information can supply enough information to warrant such a large set of assumptions.

3. A Bayes network would generalize this diagram to show an exchangeable sequence with a node representing the single-toss probability feeding into the Y_n .

This enormous logical content of random sampling and randomization illustrates why they are such powerful investigative tools: only the physical act of blocking all causal effects on selection or treatment can provide deductive justification for the entire set of assumptions corresponding to “independence.”

31.4 The Superconducting Supercollider of Selection

In human field studies, realistic causal diagrams should always have a selection (sampling) indicator node S as shown as part of the data-generating process. This node may be influenced by (and perhaps even influence) study variables. By definition, only those with $S = 1$ are observed; thus S will always be conditioned on. If S is affected by more than one variable it will be a conditioned collider and thus a potential bias source under ordinary graphical rules [Greenland 2010a, 2012a]. Most basic causal-diagram introductions (including those I helped write) can be faulted for not emphasizing this fact. We can now fault statistics education for the same reason in that the “ignorability” of selection under random sampling has led to forgettability of the physical selection mechanism in settings where it is not random in any mechanical sense and thus not ignorable in any practical sense.

An important point for graphically representing these problems is that not all of what is known as selection bias arises from S being a collider.⁴ For example, classical selection bias requires no collider in the causal graph of data collection. Consider the earlier voting-survey graph $[S = 1] \leftarrow C \rightarrow V$; the bias here corresponds to classical confounding, as it comes from an open back-door path connecting V to S via a shared cause (the causal fork at C). As with confounding, a solution is to condition (stratify) on C , which allows identification of C -conditional voter intentions.

Unlike in classical confounding, however, conditioning is only a partial solution: in the example, the goal is to recover the marginal (C -unconditional) distribution $\Pr(V = v)$ of V in the targeted S -unconditional population. Unfortunately, that V marginal is not identified if the graph is the only information available on the target population. This identification is achieved in classical demographic and epidemiologic standardization⁵ by averaging the observed C -conditionals $\Pr(V = v|C = c, S = 1)$ over the C distribution of the target population, $\Pr(C = c)$; this procedure assumes, however, that V is independent of selection given C , so that $\Pr(V = v|C = c, S = 1) = \Pr(V = v|C = c)$, as implied by $S \leftarrow C \rightarrow V$.

4. This point is contrary to Hernán et al. [2004]; see Hernán [2017] for a reconciliation.

5. Not to be confused with “standardization” as in dividing a variable by its standard deviation, which damages comparisons of estimates both within and across studies [Greenland et al. 1986, 1991].

A parallel example of selection bias without collider bias arises in studying the effect of a treatment X on an outcome Y when C is a modifier of the treatment effect, as in

$$[S = 1] \leftarrow C \rightarrow Y \leftarrow X$$

[Hernán 2017]: C is independent of treatment X , and Y is independent of selection S given C , but the $S \leftarrow C \rightarrow Y$ path still can bias the estimated marginal $X \rightarrow Y$ effect given the conditioning on selection ($S = 1$); this bias would become intractable if selection (observation) affected the targeted effects (as in $S \rightarrow Y \leftarrow X$).

31.5 Data and Algorithms are Causes of Reported Results

The causal sequence continues once the data are collected: a statistical procedure is a data-processing algorithm whose flow chart can be viewed as a causal diagram showing how each computational step determines the next. Usually, each node is a deterministic function of its parents, but may include simulations (as in bootstrap and Markov chain Monte Carlo procedures) that may result in stochastic conditional branches. Finally, the outputs of the algorithm cause researchers and readers to interpret and report the study in particular ways, whether mechanically (e.g., in misreports of “no association” because a P -value exceeded 0.05) or informally, and can strongly affect whether and where the results are published.

Given the causal nature of data generation, calling causal models “extra-statistical” is a misleading characterization of both causality and statistics: valid statistical analysis is causal to the core; hence, *realistic statistical analysis is a subset of causal analysis*. Not even “extra-distributional” is correct because the core problem is about factors producing (causing) differences in distributions of those targeted (e.g., voters, patients with a given indication for treatment) and those observed (e.g., survey responders, patients in a trial). Without a causal model for deducing the assumed data distribution from the entire physical data generator, we have no basis for claiming our probability calculations are connected to our target or the world beyond our immediate data.

To summarize so far: taking off from the epilogue of Pearl [2009], statistics as conceived and practiced competently is about laying out the causal sequences leading from data to inferences (perceptions) and decisions. Within this sequence, a statistical analysis algorithm or protocol is a causal submodel for how that data will be processed into outputs. Those outputs will then be interpreted as statements connecting the target population to our data under our causally derived sampling model, with the connections established via open paths in the causal diagram between the target and the data, including connections passing through the

ever-present selection node S . Probability plays a central role in terms of formalizing the expected behaviors (propensities) of the data generator under different hypotheses; but that formalization is physically justified only when it is deduced from the causal structure of the generator.

31.6 Getting Causality into Statistics by Putting Statistics into Causal Terms from the Start

Labeling causation as “extra-statistical” creates an excuse to continue to ignore causality theory in statistical teaching and methods research, and stay within the insufficient descriptions of acausal probability theory as the only formal foundation of statistics. This leads to bad practice, such as confusing probabilities of group events with probabilities of individual events within a group. Examples of such confusion [Greenland and Robins 1988, Robins and Greenland 1989, Greenland et al. 2020] may help statisticians recognize causality as an essential component that distinguishes application-relevant statistical theory from acausal probability and its extensions in mathematical statistics. Again, sound applications also need detailed causal explanations of how the data were generated, including the physical mechanisms that led to being in different comparison groups and to inclusion in the dataset ($S = 1$).

These causal explanations provide the contextual justifications for the probability models used in the analysis, displaying information about study features that physically constrain data generation. One teaching implication is that students must master causal thinking before they can master real-world statistical inference; thus, basic logic and its causal extensions should be covered from the start of introductory statistics, *before* probability and statistics. But the curriculum for doing so is in its infancy. I used this sequencing in my UCLA courses; however, all incoming students had at least basic statistics, and most also had research methods courses in which at least informal ideas of causality were covered. Thus, the students needed retraining to remove common misconceptions about the implications (or lack thereof) of various statistical results for causal questions.

Students had no trouble mastering the idea of associations passing through causal forks (such as $X \leftarrow C \rightarrow Y$) or mediators (such as $X \rightarrow M \rightarrow Y$); in fact, their entire intuition for bias and adjustment came from these two cases. On the other hand, their intuitions for paths through colliders (such as $X \rightarrow S \leftarrow Y$) were backward, as should be no surprise: collider bias is by definition the negative or inverse of confounding because collider bias arises from conditioning (on colliders), whereas confounding is removed by conditioning (on shared causes). Hence, for

absolute measures, confounding bias equals the unconditional association minus a conditional association, whereas collider bias equals a conditional association minus the unconditional association.

Again, this view applies not only for causal research questions but also for descriptive survey research. In all real settings in which perfection is unattainable, researchers should try to understand causes of non-response, loss, missing data, misreporting, and other sources of uncertainty and inferential distortion,⁶—for example, by placing these bias sources in a causal diagram to guide study design and interpretation. Only then can they begin to master the far more subtle notions of probabilistic inference from incomplete observations.

31.7 Causation in 20th-century Statistics

Statistical foundation debates raged throughout the last century but focused exclusively on prioritization of logical criteria such as internal coherence (no violations of the axioms of probability theory) versus self-calibration (meeting select frequency criteria over data sequences generated by the distribution used to derive the data-processing algorithm). Yet formal causal modeling is as old as the modern statistical foundations laid down by Fisher, Neyman, DeFinetti, and many others in the first half of the 20th century. Although [Neyman \[1923\]](#) went largely unnoticed, potential-outcome (“counterfactual”) models entered prestigious statistics journals by the 1930s and had an ongoing presence before their broad uptake began in the 1980s (e.g., [Welch \[1937\]](#), [Wilk \[1955\]](#), [Copas \[1973\]](#)). Even without such formalisms, the probability models on which statistical procedures were based were supposed to be frequency summaries of causal mechanisms with certain physical independencies built in by design; these independencies made the mechanisms “ignorable” [[Rubin 1978](#)], a misleading term because the data-generating mechanism should always be described in detail, never ignored. Such mechanisms include random sampling, which makes selection *S* an unaffected (exogenous) node, and random allocation, which makes treatment assignment an unaffected node.

Statistical developments in the 20th century were concerned foremost with causal inferences derived from physical randomization, whether by nature, as in genetic recombination, or by design. Fisher was often quite straightforward in his causal descriptions and how he regarded causal inference about treatment effects

6. These include bad research practices such as “P-hacking”: Searching out analyses that give P-values above *or* below a threshold for “significance” [[Greenland 2017a](#), [2017b](#), [2019](#), [Amrhein et al. 2019](#)].

as the central goal of scientific experimentation in the life sciences. By the mid-1930s, he had laid out potential outcomes clearly enough (even if only verbally) to see the distinction between the sharp null of no effect on any unit (used to derive randomization tests) and Neyman’s weak null of no effect on the mean [Greenland 1991]. His *Design of Experiments* [Fisher 1935] gives primacy to experimental action (design) over mathematics, as seen in Section 2 of his introduction to the first edition, in which he states

“I have assumed, as the experimenter always does assume, that it is possible to draw valid inferences from the results of experimentation; **that it is possible to argue from consequences to causes**, from observations to hypotheses; as a statistician would say, from a sample to the population from which the sample was drawn, or, as a logician might put it, from the particular to the general.”

His ensuing verbal descriptions were soon formalized by others into a clear potential-outcome model form, where *for each unit* explicit counterfactual (unobserved) treatment assignments lead to possibly distinct outcomes (e.g., see Welch [1937, pp. 22–23]).

Nonetheless, the statistical theory that dominated subsequent advanced teaching and methods research became an extension of measure-theoretic probability, a development decried by those who followed Fisher in emphasizing the importance of context [Box 1990]. It is thus somewhat ironic that Fisher’s downfall (as manifested in his defense of smoking against charges of carcinogenicity) was his inability to neutrally synthesize all available evidence sources, particularly in mishandling sources of information not derived from physical randomization. This failing can be viewed as one of being unable to form realistic models for confounding effects coupled with (or perhaps caused by) personal wishes for vindication of his own smoking habit [Stolley 1991]. These sorts of “human factors” are themselves extraneous causes of what gets reported and publicized, and thus need to be accounted for in any realistic model for literature analysis [Greenland 2012b, 2017a, 2017b].

31.8 Causal Analysis versus Traditional Statistical Analysis

In applied statistics, assumptions are made to simplify modeling effort, which like everything else is resource constrained. For example, the standard modeling assumption “linear in the natural parameter” is rarely, if ever, deduced from anything; instead, standard statistical methods treat it as certainly true provided there is no evidence to the contrary (even if there is little evidence to judge its accuracy or practical impact). This convention is based on the ease of use of such models,

especially their transparency and computational stability relative to intrinsically non-linear models, along with the idea that basic linear trend components are sometimes the only components that are needed or that can be stably estimated from available data.

A retreat from causal to convenience justification is only to be expected when applications involve complexities beyond complete formal (algorithmic) modeling capacities, as in biology, medicine, and social sciences. In such applications, all models are wrong at some practical level of analysis, and are often wrong in very consequential ways *even when they are useful for improving predictions of yet-unseen events such as treatment effects*. The classic epidemiologic example is malaria, a disease whose name means “bad air” in the parent Italian. Before modern times, social groups noted that malaria rates were higher near swamps and attributed that to toxic effects on the air from the swamps, as suggested by the foul smell associated with swamps. This wrong theory (causal-system model) of

$$\begin{array}{l} \text{swamp} \rightarrow \text{toxic air} \rightarrow \text{malaria} \\ \text{housing location} \rightarrow \text{toxic air} \rightarrow \text{malaria} \end{array}$$

led to successful interventions such as draining swamps and building elevated houses, even though it missed the actual causal structure of

$$\begin{array}{l} \text{swamp} \rightarrow \text{mosquito exposure} \rightarrow \text{malaria} \\ \text{housing location} \rightarrow \text{mosquito exposure} \rightarrow \text{malaria} \end{array}$$

which predicted the same intervention effects. To explain these successes of the wrong model, we may note that the swamp intervention tested only the swamp \rightarrow malaria effect while the housing intervention tested only the housing \rightarrow malaria effect. Both interventions left wide open the identity of the intermediates (and thus specifics of the mechanism for intervention), yet were taken to demonstrate the (in fact untested) pathway of toxic air.

Such examples show that causal theories can include important mistakes even while successfully predicting intervention effects, and show why those theories should not be taken as true because of such successes (even in a world where causal laws are stable and thus inductive reasoning is justified). They instead need ongoing novel tests (not just “replication”) before basing actions on pathways that have not yet been tested by experiments. The enhanced risk of error for a mechanistic causal theory over a mere predictive/associative theory is not a disadvantage; however, it reflects the greater specificity, greater logical content, and hence greater testability of such theories, properties that are often promoted as hallmarks of good scientific theories [Popper 1962].

That such a theory can pass apparently strong experimental tests yet be erroneous in important ways (as in the malaria example) is one reason pragmatic analysts reject notions of “experimental support” for scientific (real-world) causal theories. Other theories (including many never imagined) may pass the same experimental test, so at most we can only say an experiment supports the broad class of theories that predict results close to what was observed. Put another way: an intervention experiment provides evidence only on *classes* of mechanisms (those whose diagrams have directed paths from the observed intervention to the observed outcome), not specific mechanisms, and thus leaves open many details of intervention effects.

That caution applies even more strongly in passive observational (non-experimental) studies, especially when their data are “analyzed” (summarized) by statistics based on randomization assumptions. In that case one can view a conventional interval estimate as a blur around the point estimate, indicating irreducible uncertainty about the behavior of the data generator. But any inferential connection of these summaries to a targeted treatment effect should be mediated by explicit causal models; specifically, extraction of information about the target effect (e.g., in the form of credible uncertainty intervals for the target) requires causal models for physical data generation that include non-random variation (bias) sources beyond the treatment [Greenland 1990, 2012a, Greenland et al. 1999, Maclure and Schneeweiss 2001, Robins 2001, Hernán et al. 2004, Glymour and Greenland 2008]. It also requires recognition that effects cannot always be identified by observed associations, and that some effects cannot be statistically identified at all, even from randomized trials [Kaufman 2009, Robins and Richardson 2011].

31.9 Relating Causality to Traditional Statistical Philosophies and “Objective” Statistics

As has been long and widely emphasized in various terms (e.g., Cox [1978], Box [1980, 1990], Rubin [1984], Good [1992], Barnard [1996], Chatfield [2002], Kelly and Glymour [2004], Greenland [2006, 2010b], Senn [2011], Gelman and Shalizi [2013]), frequentism and Bayesianism are incomplete both as learning theories and as philosophies of statistics in the pragmatic sense that each alone are insufficient for all sound applications. Notably, causal justifications are the foundation for classical frequentism, which demands that all model constraints be deduced from real mechanical constraints on the physical data-generating process. Nonetheless, it seems modeling analyses in health, medical, and social sciences rarely have such physical justification.

Beyond graphs, causality theory formalizes design information (such as randomization and matching) by the constraints that information places on the distributions of unobserved variables (e.g., [Greenland \[1990\]](#), [Pearl \[1995\]](#), [Robins \[2001\]](#), [Hernán and Robins \[2020\]](#)). Use of that information is especially important when the modeled data generator is not fully understood as a coherent whole—a problem long recognized and discussed at length in the literature on model uncertainty (e.g., [Leamer \[1978\]](#), [Box \[1980\]](#)). The deficiency of strict, coherent (operational subjective) Bayesianism is its assumption that all aspects of this uncertainty have been captured by the prior and likelihood, thus excluding the possibility of model misspecification [[Leamer 1978](#), [Box 1980](#), [Senn 2011](#)]. DeFinetti himself was aware of this limitation:

“...everything is based on distinctions which are themselves uncertain and vague, and which we conventionally translate into terms of certainty only because of the logical formulation...In the mathematical formulation of any problem it is necessary to base oneself on some appropriate idealizations and simplification. This is, however, a disadvantage; it is a distorting factor which one should always try to keep in check, and to approach circumspectly. It is unfortunate that the reverse often happens. One loses sight of the original nature of the problem, falls in love with the idealization, and then blames reality for not conforming to it.” [[DeFinetti 1975](#), p. 279]⁷

By asking for physically causal justifications of the data distributions employed in statistical analyses (whether those analyses are labeled frequentist or Bayesian), we may minimize the excessive certainty imposed by simply assuming a probability model and proceeding as if that idealization were a known fact.

DeFinetti was of course writing in support of a contentious, purely subjective view of probability, and the utility of the entire “subjective”/“objective” distinction in statistics has been questioned [[Gelman and Hennig 2017](#)]. Nonetheless, many statisticians assign primacy to “objective” model components (those derivable from observed mechanisms, such as random-number generators). What supports a claim that a variable is “completely random” (fully randomized) in an objective frequency sense? Modern causality theory can identify this randomness with the assumption that the variable is exogenous or instrumental, in that its causes affect the system under study only through the variable [[Pearl 2009](#)]. Again, in “objective” theory this sharp, strong assumption is *deduced* from the

7. I am indebted to Stephen Senn for reminding me of this and other remarkable passages in DeFinetti.

physical data-generating mechanism, not from observed frequencies or other purely associational information.

Consider “fair” coin tossing, in which the influence of the person tossing (who might be a magician) is blocked by having them throw the coin against a wall and then step back before the bounce and landing, thus blocking skilled tossing and other trickery as influences of the outcome. Then, even under classical deterministic mechanics, the functional complexity of the relation of the outcome to the initial toss is transcomputable or chaotic. This type of complexity forces our predictions to rely on distributions that arise as attractors of statistical behavior (e.g., laws of large numbers, central-limit effects), instead of deterministic mathematical models. In doing so we assume a certain causal stability across trials whose consequences are summarized in our models. Such a stability assumption needs justification based on direct observation (the physical mechanism is unchanging) and thus is objective; without that, causal stability is an underived (and usually implicit) assumption and thus is not objective in this sense.

In this way, the traditional “objective”/“subjective” distinction in statistical methods resides within causality theory, not in the “frequentist” versus “Bayesian” distinction (which are both vague labels for highly heterogeneous collections of statistical tools and philosophies, as [Good \[1971\]](#) explained for Bayes). The core idea behind “objective” statistics is that one demands that each distribution used in the statistical processing of the data be derivable from a verifiable physical (causal) mechanism. That demand can be made regardless of whether that processing is labeled “frequentist,” “Bayesian,” “likelihoodist,” or something else—a view which does not exclude Bayesian methods but does reject mere expressions of opinions as priors for those methods [[von Mises 1981](#)].

31.10 Discussion

Judea Pearl has been a celebrated promoter of causal models over pure probability, especially for encoding the background (contextual) information in a problem [[Pearl 1995, 2001, 2009](#)]. At times, however, he has referred to causality as “extra-statistical,” a label which ignores the realities that any applied statistician must face in practice. Those realities make causality integral to statistics; yet, by calling causality “extra-statistical,” we absolve those bearing the professional label “statistician” of any responsibility to understand, let alone teach, causality theory. Fortunately, many younger statisticians have a keen interest in causal models as tools to create better statistical science. To encourage this trend, we should include causal models from the start of statistical training as an integral component of study design and data analysis, in addition to complementary presentation of frequentist and Bayesian ideas.

As a less-often stated yet even more fundamental need, basic statistics should begin with the elements of deductive logic. When I was teaching statistical foundations and principles, most students I encountered (including statistics majors) had neither studied nor fully understood basic logical principles, and thus were prone to naïve fallacies in verbal arguments. Thus, the topic sequence in my class covered logic as a foundation for causal thinking, followed by causality theory as a foundation for probability and association explanation. This material was contrasted to their previous instruction, which typically involved rote application of mysterious descriptions and formulas for statistical comparisons and regressions. Students were always delighted to at last see applied statistics as the coordinated merging of the three essentials of logic, causation, and probability to provide a transparent foundation for sound study design, analysis, and interpretation.

Admittedly, traditionally trained statisticians may be too firmly wedded to probabilistic foundations to ever concede this causal primacy, and some radical subjective Bayesians reject causality altogether (e.g., [Lad \[2006\]](#)). Nonetheless, probabilists curious about the causal approach may more easily conceive the unification of causality and probability within information theory, which can serve as an overarching framework for statistical modeling and inference (I have found that an information framework even helps students correctly understand P-values [[Greenland 2019](#), [Rafi and Greenland 2020](#)]). Causal diagrams then provide an intuitive representation of information flows as time-sequential functional relations across event sequences.

31.11 Conclusion

Statistical science (as opposed to mathematical statistics) involves far more than data—it requires realistic *causal* models for the generation of that data and the deduction of their empirical consequences. Evaluating the realism of those models in turn requires immersion in the subject matter (context) under study. Decisions further require explication of the various pathways by which those decisions would cause gains (benefits) and losses (costs). Bringing these causal elements to the foreground is essential for sound teaching and applications of statistics.

31.A Appendix

31.A.1 A Counting Measure for the Logical Content of a Finite Exchangeability Assumption

For any formal deductive system and set of constraints A in the system, define A as logically minimal if it satisfies the joint deductive independence condition: for

any pair (B,C) of disjoint non-empty subsets of A, C cannot be deduced from B. We may then define the logical-content measure $v(G)$ of an arbitrary set of constraints G in the system as the largest cardinality $|A|$ among minimal subsets A of G; $v(G)$ may be infinite if G is infinite.

Now consider the common statistical assumption that the observations Y_1, \dots, Y_N are independent identically distributed conditional on any model m in a set M . Then, given a prior distribution on M , the Y_1, \dots, Y_N are unconditionally exchangeable; that is, every one of the $N!$ permutations of indices in the joint distribution leaves that distribution unchanged. Exchangeability is logically equivalent to $N!-1$ independent constraints, one for each non-null permutation; denoting the set of these constraints by G , with no further constraint we have $v(G) = |G| = N!-1$. By imposing further constraints on the joint distribution, we may reduce $v(G)$ considerably. Nonetheless, even with the extreme simplification of multivariate normality we get $v(G)$ of order N^2 (since exchangeability requires homogeneous variances and homogeneous covariances), and thus still entails far more constraints than there are observations N .

Acknowledgments

I am grateful to Steve Cole, Joseph Halpern, Jay Kaufman, Blakeley McShane, and Sherrilyn Roush for their helpful comments on the drafts.

References

- V. Amrhein, D. Trafimow, and S. Greenland. 2019. Inferential statistics as descriptive statistics: There is no replication crisis if we don't expect replication. *Am. Stat.* 73, supplement 1, 262–270. Open access at www.tandfonline.com/doi/pdf/10.1080/00031305.2018.1543137. DOI: <https://doi.org/10.1080/00031305.2018.1543137>.
- G. A. Barnard. 1996. Fragments of a statistical autobiography. *Student* 1, 257–268.
- G. E. P. Box. 1980. Sampling and Bayes inference in scientific modeling and robustness (with discussion). *J. R. Stat. Soc Ser. A.* 143, 383–430. DOI: <https://doi.org/10.2307/2982063>.
- G. E. P. Box. 1990. Comment: “The unity and diversity of probability” by Glen Schafer. *Stat. Sci.* 5, 448–449.
- C. Chatfield. 2002. Confessions of a pragmatic statistician. *The Statistician* 51, 1, 1–20. DOI: <https://doi.org/10.1111/1467-9884.00294>.
- J. B. Copas. 1973. Randomization models for matched and unmatched 2×2 tables. *Biometrika* 60, 467–476. DOI: <https://doi.org/10.2307/2334995>.
- D. R. Cox. 1978. Foundations of statistical inference: The case for eclecticism. *Aust. J. Stat.* 20, 1, 43–59. DOI: <https://doi.org/10.1111/j.1467-842X.1978.tb01094.x>.
- B. DeFinetti. 1975. *Theory of Probability*, Vol. 2. Wiley, New York.
- R. A. Fisher. 1935. *The Design of Experiments*. Oliver & Boyd, Edinburgh.

- A. Gelman and C. Shalizi. 2013. Philosophy and the practice of Bayesian statistics (with discussion). *Br. J. Math. Stat. Psychol.* 66, 8–80. DOI: <https://doi.org/10.1111/j.2044-8317.2011.02037.x>.
- A. Gelman and C. Hennig. 2017. Beyond subjective and objective in statistics (with discussion). *J. R. Stat. Soc. Ser. A.* 180, 4, 967–1033. DOI: <https://doi.org/10.1111/rssa.12276>.
- M. M. Glymour and S. Greenland. 2008. Causal diagrams. Ch. 12 In K. J. Rothman, S. Greenland, and T. Lash (Eds.), *Modern Epidemiology* (3rd. ed). Lippincott Williams & Wilkins, Philadelphia, 32–50.
- I. J. Good. 1971. 46,656 varieties of Bayesians (letter). *The American Statistician*, 25, 62–63. Reprinted as Ch. 3 In I. J. Good (Ed.). 1983. *Good Thinking*. University of Minnesota Press, Minneapolis, 20–21.
- I. J. Good. 1992. The Bayes/non-Bayes compromise: A brief review. *J. Am. Stat. Assoc.* 87, 597–606.
- S. Greenland. 1990. Randomization, statistics, and causal inference. *Epidemiology* 1, 421–429. DOI: <https://doi.org/10.1097/00001648-199011000-00003>.
- S. Greenland. 1991. On the logical justification of conditional tests for two-by-two-contingency tables. *Am. Stat.* 45, 248–251.
- S. Greenland. 2006. Bayesian perspectives for epidemiologic research: I. Foundations and basic methods. *Int J Epidemiol.* 35, 765–778. Reprinted with edits as. 2008. Bayesian Analysis, Ch. 18 In K. J. Rothman, S. Greenland, and T. Lash (Eds.), *Modern Epidemiology* (3rd. ed.). Lippincott Williams & Wilkins, Philadelphia, 32–50.
- S. Greenland. 2010a. Overthrowing the tyranny of null hypotheses hidden in causal diagrams. Ch. 22 In R. Dechter, H. Geffner, and J. Y. Halpern (Eds.), *Heuristics, Probabilities, and Causality: A Tribute to Judea Pearl*. College Publications, London, 365–382.
- S. Greenland. 2010b. Comment: The need for syncretism in applied statistics. *Stat. Sci.* 25, 2, 158–161. DOI: <https://doi.org/10.1214/10-STS308A>.
- S. Greenland. 2012a. Causal inference as a prediction problem: Assumptions, identification, and evidence synthesis. Ch. 5 In C. Berzuini, A. P. Dawid, and L. Bernardinelli (Eds.), *Causal Inference: Statistical Perspectives and Applications*. John Wiley and Sons, Chichester, UK, 43–58. DOI: <https://doi.org/10.1002/9781119945710.ch5>.
- S. Greenland. 2012b. Transparency and disclosure, neutrality and balance: Shared values or just shared words? *J. Epidemiol. Community Health* 66, 967–970. DOI: <https://doi.org/10.1136/jech-2011-200459>.
- S. Greenland. 2017a. For and against methodology: Some perspectives on recent causal and statistical inference debates. *Eur. J. Epidemiol.* 32, 3–20. DOI: <https://doi.org/10.1007/s10654-017-0230-6>.
- S. Greenland. 2017b. The need for cognitive science in methodology. *Am. J. Epidemiol.* 186, 639–645. DOI: <https://doi.org/10.1093/aje/kwx259>.
- S. Greenland. 2019. Some misleading criticisms of P-values and their resolution with S-values. *Am. Stat.* 73, supplement 1, 106–114. Open access at www.tandfonline.com/doi/pdf/10.1080/00031305.2018.1529625. DOI: <https://doi.org/10.1080/00031305.2018.1529625>.

- S. Greenland. 2021. Dealing with the inevitable deficiencies of bias analysis – and all analyses. *Am. J. Epidemiol.* in press.
- S. Greenland and Z. Rafi. 2020. To aid statistical inference, emphasize unconditional descriptions of statistics. <http://arxiv.org/abs/1909.08583>.
- S. Greenland, M. P. Fay, and E. H. Brittain, et al. 2020. On causal inferences for personalized medicine: How hidden causal assumptions led to erroneous causal claims about the D-value. *Am. Stat.* 74, 243–248. DOI: <https://doi.org/10.1080/00031305.2019.1575771>.
- S. Greenland, M. Gago-Dominguez, and J. E. Castellao. 2004. The value of risk-factor (“black-box”) epidemiology (with discussion). *Epidemiology* 15, 519–535. DOI: <https://doi.org/10.1097/01.ede.0000134867.12896.23>.
- S. Greenland, M. Maclure, J. J. Schlesselman, C. Poole, and H. Morgenstern. 1991. Standardized regression coefficients: A further critique and a review of alternatives. *Epidemiology* 2, 387–392.
- S. Greenland, J. Pearl, and J. M. Robins. 1999. Causal diagrams for epidemiologic research. *Epidemiology* 10, 37–48.
- S. Greenland and J. M. Robins. 1988. Conceptual problems in the definition and interpretation of attributable fractions. *Am. J. Epidemiol.* 128, 1185–1197. DOI: <https://doi.org/10.1093/oxfordjournals.aje.a115073>.
- S. Greenland, J. J. Schlesselman, and M. H. Criqui. 1986. The fallacy of employing standardized regression coefficients and correlations as measures of effect. *Am. J. Epidemiol.* 123, 203–208. DOI: <https://doi.org/10.1093/oxfordjournals.aje.a114229>.
- M. A. Hernán. 2017. Selection bias without colliders. *Am. J. Epidemiol.* 185, 11, 1048–1050. DOI: <https://doi.org/10.1093/aje/kwx077>.
- M. A. Hernán, S. Hernández-Díaz, and J. M. Robins. 2004. A structural approach to selection bias. *Epidemiology* 15, 5, 615–625. DOI: <https://doi.org/10.1097/01.ede.0000135174.63482.43>.
- M. A. Hernán and J. M. Robins. 2020. *Causal Inference: What If*. Chapman & Hall, New York, to appear.
- D. Hume. 1748. *An Enquiry Concerning Human Understanding*. Oxford Univ. Press, Oxford, 2007 printing, 56.
- J. Kaufman. 2009. Gilding the black box. *Int. J. Epidemiol.* 38, 845–847. DOI: <https://doi.org/10.1093/ije/dyp323>.
- K. T. Kelly and C. Glymour. 2004. Why probability does not capture the logic of scientific justification. In C. Hitchcock (Ed.), *Contemporary Debates in Philosophy of Science*. Blackwell, Malden, MA, 94–114.
- F. Lad. 2006. Objective Bayesian statistics: Do you buy it? Should we sell it? *Bayesian Anal.* 1, 3, 441–444.
- E. Lamb. 2012. 5 sigma – what’s that? *Scientific American Observations*, posted July 17, 2012 at <https://blogs.scientificamerican.com/observations/five-sigmawhats-that/>, viewed June 2, 2019.
- E. E. Leamer. 1978. *Specification Searches*. Wiley, New York.

- M. M. Maclure and S. Schneeweiss. 2001. Causation of bias: The episcopo. *Epidemiology* 12, 1, 114–122. DOI: <https://doi.org/10.1097/00001648-200101000-00019>.
- B. B. McShane, D. Gal, A. Gelman, C. Robert, and J. L. Tackett. 2019. Abandon statistical significance. *Am. Stat.* 73, 235–45. DOI: <https://doi.org/10.1080/00031305.2018.1527253>.
- N. D. Mermin. 2016. Why QBism is not the Copenhagen interpretation and what John Bell might have thought of it. In R. Bertlmann and A. Zeilinger A (Eds.), *Quantum [Un]Speakables II*. Springer, Cham. First Online 16 November 2016. DOI: https://doi.org/10.1007/978-3-319-38987-5_4.
- Merriam-Webster Dictionary. 2019. “Statistics.” <https://www.merriam-webster.com/dictionary/statistics>, accessed May 16, 2019.
- J. Neyman. 1923. Sur les applications de la thar des probabilités aux expériences Agaricales: Essay des principe. [English translation of excerpts (1990) by D. Dabrowska and T. Speed, *Stat. Sci.* 5, 463–472.].
- J. Pearl. 1995. Causal diagrams for empirical research. *Biometrika* 82, 669–688. DOI: <https://doi.org/10.1093/biomet/82.4.669>.
- J. Pearl. 2001. Bayesianism and causality, or, why I am only a half-Bayesian. In D. Corfield and J. Williamson (Eds.), *Foundations of Bayesianism*. Kluwer Applied Logic Series, 24. Kluwer Academic Publishers, 19–36.
- J. Pearl. 2009. *Causality: Models, Reasoning and Inference*. (2nd. ed). Cambridge University Press, New York.
- K. R. Popper. 1962. *Conjectures and Refutations*. Basic Books, New York.
- Z. Rafi and S. Greenland. 2020. Semantic and cognitive tools to aid statistical inference: Replace confidence and significance by compatibility and surprise. *BMC Med. Res. Methodol.* 20, 244. DOI: <https://doi.org/10.1186/s12874-020-01105-9>, <https://bmcmmedresmethodol.biomedcentral.com/articles/10.1186/s12874-020-01105-9>, updates at <http://arxiv.org/abs/1909.08579>.
- J. M. Robins. 2001. Data, design, and background knowledge in etiologic inference. *Epidemiology* 12, 313–320. DOI: <https://doi.org/10.1097/00001648-200105000-00011>.
- J. M. Robins and S. Greenland. 1989. The probability of causation under a stochastic model for individual risks. *Biometrics* 45, 1125–1138. (Erratum: 1991, 48, 824). DOI: <https://doi.org/10.2307/2531765>.
- J. M. Robins and T. S. Richardson. 2011. Alternative graphical causal models and the identification of direct effects. Ch. 6 In P. Shrouf, K. Keyes, and K. Ornstein (Eds.), *Causality and Psychopathology: Finding the Determinants of Disorders and their Cures*. Oxford University Press, Oxford, 1–52.
- D. B. Rubin. 1978. Bayesian inference for causal effects: The role of randomization. *Ann. Stat.* 6, 34–58. DOI: <https://doi.org/10.1214/aos/1176344064>.
- D. B. Rubin. 1984. Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Ann. Stat.* 12, 1151–1172. DOI: <https://doi.org/10.1214/aos/1176346785>.
- S. Senn. 2011. You may believe you are a Bayesian but you are probably wrong. *RMM* 2, 48–66.

- P. Stolley. 1991. When genius errs: R. A. Fisher and the lung cancer controversy. *Am. J. Epidemiol.* 133, 416–425. DOI: <https://doi.org/10.1093/oxfordjournals.aje.a116055>.
- R. von Mises. 1981. *Probability, Statistics and Truth*, 2nd rev. (English ed.). Dover, New York.
- R. L. Wasserstein. 2018. Turing Award winner, longtime ASA member publishes *The Book of Why* (interview with Judea Pearl). *Amstat News* Aug. 2018, 12–14.
- R. L. Wasserstein, A. L. Schirm, and N. A. Lazar. 2019. Moving to a world beyond “ $p < 0.05$.” *The Am. Stat.* 73, 1–19. DOI: <https://doi.org/10.1080/00031305.2019.1583913>.
- B. L. Welch. 1937. On the z-test in randomized blocks and Latin squares. *Biometrika* 29, 1–2, 21–52. DOI: <https://doi.org/10.2307/2332405>.
- M. B. Wilk. 1955. The randomization analysis of a generalized randomized block design. *Biometrika* 42, 70–79. DOI: <https://doi.org/10.2307/2333423>.

Pearl on Actual Causation

Christopher Hitchcock (California Institute of Technology)

Abstract

This chapter surveys Judea Pearl's work on actual causation. After briefly introducing the concept of actual causation, it presents the structural equation framework used by Pearl to analyze actual causation. Earlier definitions of actual causation are presented to illustrate some of the difficulties involved in analyzing this concept. One of Pearl's definitions of actual causation is presented in detail, and its strengths and weaknesses are examined. The chapter concludes with reflections on Pearl's contributions to the topic.

32.1 Introduction

Judea Pearl offered three different but closely related definitions of *actual causation* using the formalism of *structural equation models*. The first appeared in chapter 10 of *Causality: Models, Reasoning, and Inference* [Pearl 2000, 2009]; the others appeared in a series of papers co-authored with Joseph Halpern [Halpern and Pearl 2001a, 2001b, 2005a, 2005b]. Pearl's definitions are based on the *but-for* definition of causation used in common law, and build on important earlier work by the philosopher David Lewis [Lewis 1973, 1986]. Pearl's definitions have been very influential and have inspired a number of further attempts to refine the definition within the same formalism; an incomplete selection includes Blanchard and Schaffer [2017], Beckers and Vennekens [2017, 2018], Fenton-Glynn [2017], Gallow [2021], Glymour and Wimberly [2007], Hall [2007], Halpern [2008, 2016], Halpern and Hitchcock [2015], Hitchcock [2001, 2007], Menzies [2004, 2017], and Woodward [2003, chapter 2]. This chapter will provide an introduction to the topic.

32.2 Actual Causation

We may illustrate the concept of actual causation with a traditional example. Billy and Suzy are throwing stones. Suzy throws her stone at the window, it hits the

window, and the window breaks. We would naturally summarize this episode in one of the following ways:

- Suzy's throw caused the window to break
- Suzy caused the window to break by throwing a stone at it
- The window broke because Suzy threw a stone at it

These statements describe a relation of actual causation between two events: Suzy throwing a stone and the window breaking. We may make the following generalizations about relations of actual causation:

- They relate particular events, rather than types or properties. In our example, it is a particular throw, of a particular stone, by a particular girl, at a particular time and place that causes a particular window to break at a particular time and place. The statements of actual causation listed above say nothing about the efficacy of throws or rocks in general, nor about the causes of broken windows in general.
- They depend on how events actually play out. Suzy *might* not have thrown, her throw *might* not have hit, Billy *might* have thrown the stone that broke the window; but as things actually happened, it was Suzy's throw that caused the window to break.
- Claims of actual causation are typically (but not always) made after the fact. Before Suzy throws, it may be hard to predict whether she will throw or whether her aim will be true. After the fact, it is relatively easy to judge that Suzy's throw caused the window to break.
- Relations of actual causation are particularly relevant to judgments of moral responsibility and legal liability. We would hold Suzy morally responsible for the broken window and require her parents to pay for its replacement (Suzy is still a minor).

This is not a rigorous or complete definition, but it provides some indication of the target of analysis.

32.3 Causal Models and But-for Causation

One of Pearl's many innovations was introducing the use of *structural equation models* (SEMs) to represent the causal structure of a situation such as the one described in the vignette about Billy and Suzy. SEMs have been widely used in a number of fields, including agronomy, econometrics, and epidemiology, and Pearl has a great deal to say about their use in these areas as well. I will not attempt to

provide a rigorous presentation of this formalism, but will introduce it by means of examples in the hopes of making it intuitive. These examples will serve three purposes. They will help us to introduce the formalism; they will provide test cases for theories of actual causation; and they will demonstrate some of the problems facing earlier accounts. While I will present two previous definitions of actual causation to set the stage for Pearl's definitions, my formulations of them will be anachronistic—they will be couched within the formalism developed later by Pearl.

All examples will involve Billy and Suzy throwing stones at a window, and we will make the following assumptions throughout: (1) whenever Billy or Suzy throws a rock, their aim is true and they throw with sufficient force to shatter the window; (2) the window does not break spontaneously, or due to any other cause not explicitly mentioned. We will represent various scenarios using *variables* with the following interpretations:

- ST —Suzy throws her rock
- SF —Suzy's rock flies through the air toward the window
- SH —Suzy's rock hits the window
- BT —Billy throws his rock
- BH —Billy's rock hits the window
- BB —Billy blocks Suzy's rock
- WB —the window breaks

Each variable takes the value 1 if the relevant event occurs, and 0 if it does not. We might think of these as propositions that can be true or false, rather than variables. But the variables in a SEM need not be binary—we could, for example, have a variable representing the velocity of Suzy's rock—but we will restrict ourselves to binary variables for simplicity. An assignment of a value to a variable corresponds to a particular event; for example, $ST = 1$ corresponds to Suzy's throwing her rock at a particular time and place. These will be the candidates for causes and effects.

Example 32.1 Suzy throws her rock at the window, which breaks. (Billy has not yet arrived.)

In this little story, it would be natural to judge that Suzy's throw caused the window to break. We can model this very simple example as follows:

Model 32.1 $M_{32.1}$

- $ST = 1$
- $WB = ST$

The first equation tells us that $ST = 1$, that is, that Suzy throws her rock. In this model, ST is an *exogenous* variable; its value is determined by factors that are not explicitly modeled.¹ The second equation tells us how the value of WB depends upon the value of ST . Specifically, it tells us that if and only if $ST = 1$ (Suzy throws), then $WB = 1$ (the window breaks). However, this equation is different from a normal logical biconditional in that it matters which variable we put on the left-hand side. The equations in a causal model are *structural* equations, meaning that they encode information about causal structure. This model is *acyclic*, meaning that the equations can be ordered so that each variable appears on the left-hand side of an equation before it appears on the right. Variables that are introduced earlier in this ordering will be said to be *upstream*, and those that appear later are *downstream*. In $M_{32.1}$, ST is upstream of WB , and WB is downstream of ST . We will only consider acyclic models in what follows. In an acyclic SEM, the values of the exogenous variables uniquely determine the values of all of the endogenous variables via the equations. (Probability can be added to the models, but we will skip this complication.) Thus, in $M_{32.1}$, WB will take the value 1, which we can write $M_{32.1} \models WB = 1$. One basic criterion of adequacy for a causal model is that it entail values of the variables corresponding to events that actually occurred in the situation or story being modeled. (For this reason, we will sometimes refer to the values that variables take in a given model as the *actual* values of the variables in that model.)

If we want to know what *would* have happened if Suzy had *not* thrown, we remove the original equation for ST and replace it with the imposed value $ST = 0$.

Model 32.1.1 $M_{32.1.1} = M_{32.1}[ST \leftarrow 0]$

- $ST \not= 1, ST = 0$
- $WB = ST$

The notation $M_{32.1}[ST \leftarrow 0]$ indicates that the new model is formed by starting with $M_{32.1}$, striking out the equation for ST , and replacing it with the setting $ST = 0$. Setting the value of a variable in this way is called an *intervention*. We can now compute from the resulting equations that $WB = 0$. We have thus verified the following *counterfactual*: If Suzy hadn't thrown her rock, the window would not have broken. The breaking of the window *counterfactually depends* upon Suzy's throw. A second basic

1. I am oversimplifying the treatment of exogenous variables. In Pearl's various formulations, exogenous variables do not represent factors that form part of the scenario. Thus the full model would treat ST as an endogenous variable whose value is determined by one or more exogenous variables. Pearl then distinguishes between the model proper, and a specific setting of the exogenous variables. I am combining both of these together in what I am calling a model.

condition of adequacy for a causal model is that it entail only counterfactuals that are true in the situation or story being modeled.

This leads us to a first attempt to define actual causation:

Definition 32.1 But-for.

If X and Y are distinct variables in the causal model M , then $X = x$ is an actual cause of $Y = y$ in M just in case:

1. $M \models X = x, Y = y$
2. There exist values $x' \neq x$ of X and $y' \neq y$ of Y such that $M[X \leftarrow x'] \models Y = y'$

This is the *but-for* definition of causation that is frequently used in common law. It tells us that $X = x$ is a cause of $Y = y$ just in case (1) these are the actual values of these variables, and (2) if X had taken some other value, Y would not have been equal to y . To simplify the later exposition, let us say that $X = x$ is a *but-for* cause of $Y = y$ in model M just in case Definition 32.1 rules that $X = x$ is an actual cause of $Y = y$ in model M . In Example 32.1, as modeled by $M_{32.1}$, if ST had not been equal to 1, WB would not have been equal to 1. In the language of common law, the window would not have broken *but for* Suzy's throw. In Example 32.1, Definition 32.1 gives the intuitively correct answer. We may also model the scenario described in Example 32.1 by interpolating variables between ST and WB :

Model 32.1.2 $M_{32.1.2}$

- $ST = 1$
- $SF = ST$
- $SH = SF$
- $WB = SH$

This model tells us that whether the window breaks counterfactually depends upon whether Suzy's stone hits it, which depends upon whether Suzy's rock is flying through the air, which depends upon whether she threw it.

It is helpful, but not strictly necessary, to represent the structure of a causal model with a directed graph. We draw an arrow from X to Y just in case X appears on the right-hand side of the equation for Y . The graph for $M_{32.1.2}$ is shown in Figure 32.1.

Like $M_{32.1}$, $M_{32.1.2}$ also implies that if Suzy had not thrown, the window would not have shattered, as the reader can verify by replacing the first equation with

$$ST \longrightarrow SF \longrightarrow SH \longrightarrow WB$$

Figure 32.1 Directed graph of $M_{32.1.2}$.

$ST = 0$. Suppose now that we want to evaluate the counterfactual situation where Suzy's rock does not hit the window. Following our procedure, we produce the new model:

Model 32.1.3 $M_{32.1.3} = M_{32.1.2}[SH \leftarrow 0]$

- $ST = 1$
- $SF = ST$
- ~~$SH = SF$~~ , $SH = 0$
- $WB = SH$

Note that we *replace* the equation for SH , rather than just plugging in the value 0 for SH in the original equations. This reflects the idea that when we intervene to set $SH = 0$ we override the previously existing causal structure and impose the value 0 on SH . This is similar to Lewis's idea that we should think of the antecedent of a counterfactual being made true by a small miracle [Lewis 1979]. We represent this graphically by “breaking the arrow” into SH (Figure 32.2).

When we evaluate the new system of equations, we get $WB = 0$ (the window wouldn't have broken), but ST and SF remain unchanged (Suzy still would have thrown, and her rock still would have flown through the air). What this example shows is that counterfactuals do not *backtrack* (in the terminology of Lewis [1979]). A hypothetical change introduced through an intervention may lead to changes in the values of *downstream* variables, but it will not lead to any changes in the values of *upstream* variables. The relation of counterfactual dependence is asymmetric (in acyclic models).

The asymmetry of counterfactual dependence is a good thing for Definition 32.1: it means that Definition 32.1 does *not* have the consequence that Suzy's rock hitting the window caused her to throw it. More generally, if $X = x$ is an actual cause of $Y = y$, then $Y = y$ will not be an actual cause of $X = x$. Thus Definition 32.1 can capture the intuitive idea that causation is an asymmetric relation.

Two further points about counterfactuals: First, we can readily extend our procedure for evaluating counterfactuals to cases where we intervene on multiple variables. We replace the equations for all of the variables on which we intervene.



Figure 32.2 Directed graph of $M_{32.1.3} = M_{32.1.2}[SH \leftarrow 0]$.

Second, if we intervene to set one or more variables to their actual values in the model, all other variables will take their actual values.² That is:

Fact 32.1 If $M \models \vec{X} = \vec{x}, \vec{Y} = \vec{y}$, then $M[\vec{X} \leftarrow \vec{x}] \models \vec{Y} = \vec{y}$.³

32.4 Pre-emption and Lewis

It has been known since at least 1925 that Definition 32.1 is inadequate [McLaughlin 1925]. In particular, it fails in cases of *pre-emption*. Here is an illustration:

Example 32.2 Billy and Suzy are holding their stones, ready to throw. Billy decides to let Suzy throw first. Suzy throws her rock, which shatters the window. If Suzy hadn't thrown her rock, Billy would have thrown his rock at the window.

In this example, the window's breaking does not counterfactually depend upon Suzy's throw. If Suzy hadn't thrown, Billy's rock would have broken the window. Nonetheless, it is natural to judge that Suzy's throw caused the window to shatter. This is called a case of *pre-emption* because Suzy pre-empted Billy by throwing first.

Here is a simple and natural causal model for Example 32.2:

Model 32.2 $M_{32.2}$

- $ST = 1$
- $BT = \neg ST$
- $WB = ST \vee BT$

The second equation tells us that Billy would throw just in case Suzy doesn't. The third equation says that the window would break just in case either Suzy or Billy throws. This model is pictured in Figure 32.3. Note that the arrow from ST to BT indicates that the first variable influences the second, but it does not tell us what the direction of influence is. That is, the arrow does not tell us whether the equation is $BT = \neg ST$ or $BT = ST$ —whether Suzy's throw causes Billy's throw or prevents it. Thus, the equations of the model contain strictly more information than the corresponding graph. The graph does help us to see that ST influences WB via two different routes: one direct and one via BT .

2. Note, however, that not all propositions remain true in the new model that results from such an intervention. In particular, some counterfactuals may change in truth value. See, for example, Briggs [2012] for discussion.

3. I am using $\vec{X} = \vec{x}$ as a fairly intuitive shorthand. If $\vec{X} \equiv (X_1, \dots, X_n)$ is an ordered set of variables, and $\vec{x} \equiv (x_1, \dots, x_n)$ is an ordered set of values, then $\vec{X} = \vec{x}$ abbreviates the conjunction of propositions $X_i = x_i$ for $i = 1, \dots, n$.

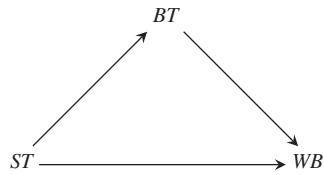


Figure 32.3 Directed graph of $M_{32.2}$.

The reader can check that in $M_{32.2}$, $BT = 0$ (Billy doesn't throw) and $WB = 1$ (the window breaks). However, if Suzy hadn't thrown ($ST = 0$), then Billy would have thrown ($BT = 1$) and the window would have shattered anyway ($WB = 1$).

Lewis [1973] introduced a counterfactual theory of causation that improves upon the simple *but-for* definition. Lewis argued that causation is a *transitive* relation. If $X = x$ is an actual cause of $Y = y$ and $Y = y$ is an actual cause of $Z = z$, then $X = x$ should be an actual cause of $Z = z$. Definition 32.1 does not have this consequence since the relation of counterfactual dependence is not transitive (as we shall see in a moment). Lewis took counterfactual dependence to be *sufficient* for causation, but not necessary. $X = x$ can be an actual cause of $Z = z$ in the absence of counterfactual dependence if there is a suitable chain of counterfactual dependence.

Definition 32.2 Lewis

If X and Z are distinct variables in the causal model M , then $X = x$ is an actual cause of $Z = z$ in M just in case:

- There exists a sequence of variables $X \equiv Y_1, Y_2, \dots, Y_{n-1}, Y_n \equiv Z$ such that: $Y_i = y_i$ is a *but-for* cause of $Y_{i+1} = y_{i+1}$ for all $i = 1, \dots, n - 1$.

Note that this entails that $M \models X = x, Z = z, Y_i = y_i$ for all i . *But-for* causation is a special case where $n = 2$.

Lewis's definition doesn't yield the intuitive result that $ST = 1$ is an actual cause of $WB = 1$ in $M_{32.2}$, but it does give this result in a slightly different model of Example 32.2, in which an additional variable is interpolated:

Model 32.2.1 $M_{32.2.1}$

- $ST = 1$
- $BT = \neg ST$
- $SF = ST$
- $WB = SF \vee BT$

(See Figure 32.4.) In this model, $ST = 1$ is a *but-for* cause of $SF = 1$ (if Suzy hadn't thrown, her rock wouldn't have flown through the air); and $SF = 1$ is a *but-for*

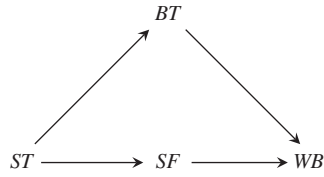


Figure 32.4 Directed graph of $M_{32.2.1}$.

cause of $WB = 1$ (if Suzy's rock hadn't been flying through the air, the window wouldn't have broken). Thus we have a chain of counterfactual dependence, and Definition 32.2 rules that $ST = 1$ is an actual cause of $WB = 1$. The first step of this chain, from ST to SF , is both intuitive, and easy to verify using model $M_{32.2.1}$. The second step, from SF to WB , is less intuitive. We will first use the model to evaluate what happens under the counterfactual supposition that $SF = 0$:

Model 32.2.2 $M_{32.2.2} = M_{32.2.1}[SF \leftarrow 0]$

- $ST = 1$
- $BT = \neg ST$
- $SF \neq ST, SF = 0$
- $WB = SF \vee BT$

In this model, $ST = 1$ (Suzy still throws), $BT = 0$ (Billy doesn't throw), $SF = 0$ (Suzy's rock does not fly through the air), and $WB = 0$ (the window remains intact). Since counterfactuals do not backtrack, if Suzy's rock hadn't flown she still would have thrown, and Billy still would have refrained from throwing. We are to imagine that Suzy's rock vanishes or disintegrates after leaving her hand, or something intervenes to knock it out of the air. Since Billy's throw was conditioned on Suzy's throw, and not on the flight of her rock, he would not throw in this situation.

One question this raises is whether $M_{32.2}$ or $M_{32.2.1}$ is the "right" model of Example 32.2. Definition 32.2 yields a definition of actual causation that is *model-relative*. But the hypothetical examples that are used to assess the adequacy of definitions of causation are presented in natural language; they don't wear a preferred model on their sleeve. This raises several questions: What makes one causal model rather than another the "right" model of a particular situation? Is there a uniquely correct causal model? If not, what makes a causal model *apt* for analysis? Halpern and Hitchcock [2010] and Blanchard and Schaffer [2017] provide some preliminary discussion of these issues. Given an analysis of actual causation, when are the verdicts of that analysis stable under additions to and deletions from a causal model? Is this a desirable feature of an analysis? Can this kind of stability be used

to motivate a particular analysis? Halpern [2016, chapter 4] and Gallow [2021] take up these issues. As we will see, model-relativity will be a recurring issue.

32.5 Intransitivity and Overdetermination

Despite achieving some success with Example 32.2, Lewis's definition faces problems with other examples. The first such example raises questions about Lewis's hypothesis that actual causation is transitive.

Example 32.3 Suzy throws her rock toward the window. Billy does not want the window to break, so he leaps into action and blocks Suzy's rock. The window remains intact.

We can model this example as follows:

Model 32.3 $M_{32.3}$

- $ST = 1$
- $BB = ST$
- $WB = ST \wedge \neg BB$

(See Figure 32.5.) The last equation says that the window will break just in case Suzy throws and Billy doesn't block her rock.

In this model, $ST = 1$ is a *but-for* cause of $BB = 1$: if Suzy hadn't thrown, Billy wouldn't have blocked her rock. Moreover, $BB = 1$ is a *but-for* cause of $WB = 0$: if Billy hadn't blocked Suzy's rock, the window would have broken. (Remember that counterfactuals do no backtrack, so if Billy hadn't blocked the rock, Suzy still would have thrown). We can verify this second counterfactual by intervening to set $BB = 0$.

Model 32.3.1 $M_{32.3.1} = M_{32.3}[BB \leftarrow 0]$

- $ST = 1$
- $\cancel{BB} = \cancel{ST}, BB = 0$
- $WB = ST \wedge \neg BB$

We can compute that $WB = 1$ in this model. Since there is a chain of counterfactual dependence from $ST = 1$ to $BB = 1$ to $WB = 0$, Definition 32.2 rules that $ST = 1$ is

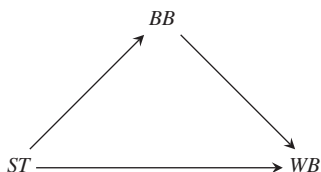


Figure 32.5 Directed graph of $M_{32.3}$.

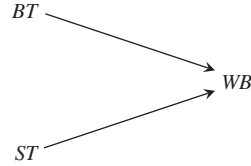


Figure 32.6 Directed graph of $M_{32.4}$.

an actual cause of $WB = 0$.⁴ But most people find this verdict unintuitive. Suzy's throw did not cause the window to remain intact (or prevent it from breaking). Lewis's definition gives the wrong answer. Moreover, this example is a counterexample to the transitivity of causation: Suzy's throw caused Billy to block her rock, and Billy's action caused the window to remain intact, but Suzy's throw did not cause the window to remain intact. This undermines one of the main motivations for moving from Definition 32.1 to Definition 32.2.

Lewis's definition also has trouble with causes of *symmetric overdetermination*:

Example 32.4 Billy and Suzy both throw their rocks at the window. The rocks hit the window simultaneously, and the window breaks.

Model 32.4 $M_{32.4}$

- $ST = 1$
- $BT = 1$
- $WB = ST \vee BT$

(See Figure 32.6.) The logical *or* in the last equation reflects the fact that either throw would be sufficient on its own to break the window.

$WB = 1$ does not counterfactually depend upon $ST = 1$: If Suzy hadn't thrown, the window still would have broken (because of Billy's throw). Nonetheless, most people judge that Suzy's throw and Billy's throw are both causes of the window breaking.⁵ I will leave it to the reader to verify that it does not help to interpolate variables such as SF or SH between ST and WB .

Here is another case of pre-emption that differs from Example 32.2⁶:

4. Interpolating a variable such as SF between ST and WB won't change this result.

5. Or perhaps they are parts of a joint cause. This is the verdict of one of the definitions of actual causation discussed in Halpern [2016].

6. This is an example of what Lewis [1986] calls *late pre-emption*; Example 32.2 is a case of *early pre-emption*. The nomenclature is not very intuitive. The key difference is that in early pre-emption the back-up process (Billy) is cut off before the effect (the window breaking) occurs; in late pre-emption the back-up process is still in progress when the effect occurs.

Example 32.5 Suzy throws her rock slightly before Billy does. Her rock hits the window and smashes it. Billy’s rock sails through the space where the window used to be.

Once again, it seems clear that Suzy’s throw caused the window to break; but the window would have broken if Suzy hadn’t thrown (due to Billy’s rock). And once again, interpolating variables does not solve the problem. Unlike Example 32.4, however, there is an asymmetry between Suzy’s throw and Billy’s throw: Suzy’s throw is a cause of the window breaking, but Billy’s is not.

How should we model Example 32.5? $M_{32.4}$, which we used to model Example 32.4, is minimally adequate in the sense that it correctly describes the values of the variables, and that it also entails only true counterfactuals. However, if a definition of actual causation is going to yield a different verdict about Examples 32.4 and 32.5, then we will need to model these cases differently. In particular, it is apparent that $M_{32.4}$ is *symmetric* between ST and BT . Any account of actual causation that rules that $ST = 1$ is an actual cause of $WB = 1$ in $M_{32.4}$ will also have to rule that $BT = 1$ is an actual cause. If we wish to rule that Susy’s throw is a cause of the window breaking in Example 32.5 while Billy’s throw is not, there will need to be a corresponding asymmetry in the causal model. A more adequate representation (from Halpern and Pearl [2001a]) would be:

Model 32.5 $M_{32.5}$

- $ST = 1$
- $SH = ST$
- $BT = 1$
- $BH = BT \wedge \neg SH$
- $WB = SH \vee BH$

(See Figure 32.7.) In this model, we can derive that $SH = 1$ (Suzy’s rock hits the bottle), while $BH = 0$ (Billy’s rock does not hit the bottle). This is an important asymmetry between Suzy’s throw and Billy’s throw that we might hope to exploit.

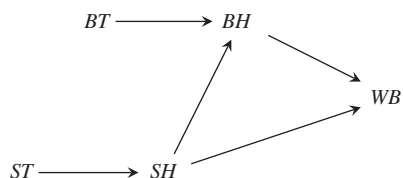


Figure 32.7 Directed graph of $M_{32.5}$.

32.6 Pearl's Definitions of Actual Causation

Pearl has given three different definitions of Actual Causation in his published work, in chapter 10 of Pearl [2000, 2009]⁷; and in a series of papers co-authored with Halpern [Halpern and Pearl 2001a, 2001b, 2005a, 2005b]. I will focus here on the definition from Halpern and Pearl [2001a].⁸

Definition 32.3 HP.

If X and Y are distinct variables in causal model M , then $X = x$ is an actual cause of $Y = y$ in M just in case:

1. $M \models X = x, Y = y$
2. There exists a partition (\vec{Z}, \vec{W}) of the variables in M , with $X \in \vec{Z}$, some setting x' of X , and some setting \vec{w}' of the variables in \vec{W} such that
 - (a) $M[X \leftarrow x', \vec{W} \leftarrow \vec{w}'] \models Y \neq y$
 - (b) $M[X \leftarrow x, \vec{W} \leftarrow \vec{w}', \vec{Z}' \leftarrow \vec{z}^*] \models Y = y$ for all $\vec{Z}' \subseteq \vec{Z}$ (where $M \models \vec{Z}' = \vec{z}^*$).⁹

Condition 1 is straightforward: it just says that x and y are the values that X and Y actually take in the model. Condition 2 requires some unpacking.

The variables in the causal model are split into two sets, \vec{W} , and \vec{Z} . We may think of \vec{Z} as making up the *causal process*. It will include X and Y , and may also include some of the variables that lie on causal paths between X and Y . The variables in \vec{W} may be thought of as being *off to the side*. While they may lie on *some* causal

7. A predecessor of this definition appears in a technical report [Pearl 1998].

8. Halpern and Pearl [2001a] and the postscript to chapter 10 of Pearl [2009] describe the reasons for preferring the definition of Halpern and Pearl [2001a] to that of Pearl [2000]. Halpern and Pearl were moved to modify their definition in light of a putative counterexample described in Hopkins and Pearl [2003], giving rise to the new definition presented in Halpern and Pearl [2005a]. However, I think that the earlier definition of Halpern and Pearl [2001a] can handle this example by using a more sophisticated model. This closely parallels the move from modeling Example 32.5 using $M_{32.4}$ to using $M_{32.5}$. The Hopkins–Pearl case is an example of pre-emption, and its structure is not adequately captured without adding one additional variable.

9. I have simplified this definition in a couple of ways. Halpern and Pearl [2001a] allow the effect to be an arbitrary Boolean combination of propositions about the values of variables in the model. They don't require that the cause and effect involve distinct variables, although they note the possibility of adding such a restriction. They also allow the cause to be a conjunction of assignments of values to variables, but add a third clause to the definition that imposes a minimality condition on the cause. It turns out that this minimality condition implies that causes always involve single variables. (This is not the case with the definition of Halpern and Pearl [2005a], however.)

path between X and Y , they are not part of the particular causal process that makes $X = x$ an actual cause of $Y = y$.¹⁰

Condition 2(a) says that $Y = y$ counterfactually depends upon $X = x$, not in the original model M but in the new model that results when we also set \vec{W} to \vec{w}' . The values \vec{w}' may be the actual values of \vec{W} , but they need not be.

Condition 2(b) is a restriction on the permissible settings $\vec{W} = \vec{w}'$. The condition tells us that the setting of $\vec{W} = \vec{w}'$ cannot interfere with the causal process \vec{Z} too much. Specifically, setting \vec{W} to \vec{w}' can't result in a different value of Y when X is set to its actual value, and when any members of \vec{Z} are set to their actual value.

When $X = x$ and $Y = y$ satisfy the conditions of Definition 32.3 in model M , we will say that $X = x$ is an *HP cause* of $Y = y$ in M . We may note the following two facts about Definition 32.3:

Fact 32.2 When $\vec{W} = \emptyset$, Definition 32.3 reduces to Definition 32.1.

Hence *but-for* causation is sufficient for *HP* causation.

Fact 32.3 When $M \models \vec{W} = \vec{w}'$, the setting $\vec{W} = \vec{w}'$ satisfies condition 2(b).

Fact 32.3 follows from Fact 32.1.

Let us now see how the Halpern–Pearl definition of actual causation handles our various examples.

Analysis of Example 32.1 Suzy throws her rock at the window, which breaks.

- $M_{32.1} : ST = 1, WB = ST$

We want to show that $ST = 1$ (Suzy's throw) is an actual cause of $WB = 1$ (the window breaking). Let $\vec{W} = \emptyset$. By Fact 32.2, Definition 32.3 now reduces to Definition 32.1. Since the but-for test rules that $ST = 1$ is an actual cause of $WB = 1$ in this simple example, the HP test does as well.

Analysis of Example 32.2 Billy decides to let Suzy throw first. Suzy throws her rock, which shatters the window. If Suzy hadn't thrown her rock, Billy would have thrown.

- $M_{32.2} : ST = 1, BT = \neg ST, WB = ST \vee BT$

We want to show that $ST = 1$ is an actual cause of $WB = 1$. Let $\vec{W} = (BT)$, and $\vec{w}' = (0)$. Since $BT = 0$ in $M_{32.2}$, Fact 32.3 implies that condition 2(b) is satisfied. To check condition 2(a):

- $M_{32.2}[ST \leftarrow 0, BT \leftarrow 0] : ST \neq 1, ST = 0, BT \neq \neg ST, BT = 0, WB = ST \vee BT$

10. Although there may be more than one way of dividing variables into sets such that Definition 32.3 is satisfied. Variables that are off to the side in one partition may be part of the causal process in another.

We can compute that $WB = 0$. This computation validates the counterfactual: “If Suzy didn’t throw, and Billy didn’t throw, the window would not have broken.” An equivalent counterfactual that more closely tracks the logic of the Definition 32.3 is: “Holding fixed that Billy didn’t throw, if Suzy hadn’t thrown, the window would not have broken.”

We may think of the analysis in this way: ST influences WB via two different causal pathways—one direct and one via BT (see Figure 32.3). By intervening to fix the value of BT at 0, we block the influence of ST on WB via the indirect path. When we “wiggle” ST , we prevent BT from “wiggling” with it. We thus isolate the influence of ST on WB along the direct path. It is in virtue of this influence that $ST = 1$ is an actual cause of $WB = 1$.

Analysis of Example 32.4 Billy and Suzy both throw their rocks at the window. The rocks hit the window simultaneously, and the window breaks.

- $M_{32.4} : ST = 1, BT = 1, WB = ST \vee BT$

We want to show that $ST = 1$ is an actual cause of $WB = 1$. Let $\vec{W} = (BT)$, and $\vec{w}' = (0)$. Since this is not the actual value of BT , we cannot rely on Fact 32.3 to guarantee that condition 2(b) is met. To check condition 2(b), we must set $ST = 1$ and $BT = 0$; and we must check that $WB = 1$ both when we set WB to 1, and when we leave WB alone. Obviously, if we set WB to 1, we will have $WB = 1$. So let us check the other case:

- $M_{32.4}[ST \leftarrow 1, BT \leftarrow 0] : ST \leftarrow 1, ST = 1, BT \leftarrow 0, BT = 0, WB = ST \vee BT$

We can compute that $WB = 1$, so condition 2(b) is met.

Let us now check condition 2(a).

- $M_{32.4}[ST \leftarrow 0, BT \leftarrow 0] : ST \leftarrow 0, ST = 0, BT \leftarrow 0, BT = 0, WB = ST \vee BT$

We can compute that $WB = 0$, so condition 2(a) is met.

Although the window’s breaking does not counterfactually depend upon Suzy’s throw in the actual situation, it does depend on her throw in the closely related situation where Billy does not throw. Changing whether Billy throws does not interfere sufficiently with the process connecting Suzy’s throw to the shattered window, so this is a legitimate situation in which to check for actual causation.

Analysis of Example 32.5 Suzy throws her rock slightly before Billy does. Her rock hits the window and smashes it.

We will analyze this example using the more sophisticated model $M_{32.5}$ (Figure 32.7).

- $M_{32.5} : ST = 1, SH = ST, BT = 1, BH = BT \wedge \neg SH, WB = SH \vee BH$

We first want to show that $ST = 1$ is a cause of $WB = 1$. We may choose $\vec{W} = (BH)$ with the setting $\vec{w}' = (0)$.¹¹ Since this is the actual value of BT , Fact 32.3 implies that condition 2(b) is satisfied. To check 2(a):

- $M_{32.5}[ST \leftarrow 0, BH \leftarrow 0]$:

$$ST = \cancel{1}, ST = 0, SH = ST, BT = 1, BH = \cancel{BT \wedge \neg SH}, BH = 0, WB = SH \vee BH$$

This implies $WB = 0$, so 2(a) is satisfied.¹² The analysis is similar to that in Example 32.2. By holding BH fixed at 0, we isolate the influence of Susy's throw along the path from ST to SH to WB .

We would also like to show that $BT = 1$ is not an actual cause of $WB = 1$. We will not go through all of the possible combinations, but let us see why the parallel strategy of choosing $\vec{W} = (SH)$ will not work. First, we could try the actual setting $SH = 1$. With this setting, condition 2(a) fails:

- $M_{32.5}[BT \leftarrow 0, SH \leftarrow 1]$:

$$ST = 1, SH = \cancel{ST}, SH = 1, BT = \cancel{1}, BT = 0, BH = BT \wedge \neg SH, WB = SH \vee BH$$

This model implies that $WB = 1$. When we fix SH at 1, WB does not counterfactually depend upon BT . So let us try instead the setting $SH = 0$. Since this is not the actual setting of SH , we will need to check whether this setting satisfies condition 2(b). We can show that it does not by choosing $\vec{Z}' = (BH)$. Since BH takes the value 0 in the actual model, we need to check:

- $M_{32.5}[BT \leftarrow 0, SH \leftarrow 0, BH \leftarrow 0]$:

$$ST = 1, SH = \cancel{ST}, SH = 0, BT = \cancel{1}, BT = 0, \\ BH = \cancel{BT \wedge \neg SH}, BH = 0, WB = SH \vee BH$$

In this model, $WB = 0$, violating 2(b). Setting $SH = 0$ is too big a change to the model. Thus, no setting for SH works.

The mathematically astute reader will notice that I have skipped Example 32.3. While Definition 32.3 yields the intuitively correct result when we use $M_{32.3}$, it yields the wrong result if we interpolate a variable.

11. There are other choices that will work: $BT = 0$, $BT = 1 \wedge BH = 0$, and $BT = 0 \wedge BH = 0$.

12. See Hall [2007] for criticism of this analysis of Example 32.5.

Analysis of Example 32.3 Suzy throws her rock toward the window. Billy does not want the window to break, so he blocks Suzy's rock. The window remains intact.

We will model this example as follows:

Model 32.3.2 $M_{32.3.2}$

- $ST = 1$
- $SF = ST$
- $BB = ST$
- $WB = SF \wedge \neg BB$

See Figure 32.8. Although the intuitive verdict is that Suzy's throw did not cause the window to remain intact, Definition 32.3 rules that $ST = 1$ is an actual cause of $WB = 1$. To see this, choose $\vec{w} = (SF)$ and $\vec{w}' = (1)$. Since this is the actual value of SF , Fact 32.3 implies that condition 2(b) is met. Checking 2(a):

- $M_{32.3.2}[ST \leftarrow 0, SF \leftarrow 1] : ST \neq 1, ST = 0, SF \neq ST, SF = 1, BB = ST, WB = SF \wedge \neg BB$

In this model, $WB = 1$. Since Suzy didn't throw, Billy didn't block. But Suzy's stone was flying through the air (at a point too late for Billy to block it) so the window broke.

Halpern and Pearl [2005a] address this problem by allowing causal models to include restrictions on interventions. That is, in addition to the structural equations, a causal model will also specify that certain combinations of values of variables are impermissible, and cannot be realized by interventions. For example, model $M_{32.3.2}$ might specify that one cannot simultaneously set $ST = 0$ and $SF = 1$ by intervention. Hitchcock [2001] notes that the counterfactual involved in this case is psychologically unnatural: We are to imagine that Suzy does not throw, lulling Billy into complacency; then somehow Suzy's rock appears mid-air flying toward the window, too late for Billy to block it. Hall [2007], Halpern [2008], Hitchcock [2007], Halpern and Hitchcock [2015], and Menzies [2017] try to resolve this kind of problem by appeal to considerations of *normality*: only combinations of settings

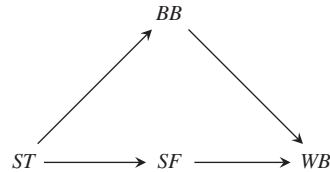


Figure 32.8 Directed graph of $M_{32.3.2}$.

that correspond to normal states can underwrite relations of actual causation. All of these approaches imply that actual causation depends on more than just the objective content of causal models.

This example also highlights the recurring problem of model-relativity.

32.7 Pearl's Achievement

We have highlighted a few of Pearl's accomplishments on the topic of actual causation. He has introduced the formalism of SEMs to the project of defining actual causation. And he has offered new definitions that have improved upon previous definitions and have inspired further developments by others. But none of Pearl's definitions perfectly capture judgments of actual causation, and—spoiler alert—neither do any of the definitions that have followed. So where does this leave us?

The situation is familiar in philosophy. In Plato's famous dialogues, Socrates asks his students: What is justice? What is piety? What is knowledge? His students propose definitions, and Socrates presents clever counterexamples to shoot them down. Two and a half millennia later, we are still shooting them down. This is not to say that we have not learned a great deal in the process, but philosophy has not converged on accepted definitions of any of these concepts.

The situation is no less frustrating for being familiar. And it seems particularly frustrating in the case of causation. We might suspect that a concept like *justice* is multi-faceted, and perhaps at least partly subjective; for this reason it might defy precise definition. But surely *causation* is not like this? Aren't causal relations part of the objective structure of the world? Don't we have well-defined empirical procedures, such as randomized controlled trials, for establishing causal claims?

Perhaps Pearl's most important contribution to our understanding of actual causation is indirect. Through his work, we better understand the place of actual causation in our conceptual economy. By setting his definitions of actual causation in the much broader context of causal modeling and causal inference, Pearl has shown us that *actual causation* is in fact a very specialized causal concept. The very fact that Pearl's first definition appears in the tenth and last chapter of [Pearl \[2000\]](#) tells us that there is a great deal one can say about causation without settling on a definition of actual causation.

This fact is hidden in our language. We say: "Suzy's throw *caused* the window to break." The verb suggests a fully general notion of causation: nothing indicates that a specialized causal notion—actual causation—is being invoked.

Moreover, Pearl's work helps us to see that actual causation is not just causation among particular events (as a number of philosophers have suggested). As we have seen without examples, we can construct causal models of particular situations

that capture aspects of their causal structure. These models do not, by themselves, tell us what the *actual causes* are. For example, one cannot simply inspect $M_{32.5}$ and read off that Suzy's throw is an actual cause of the window shattering. To make this judgment, we further need a definition of actual causation in terms of the underlying causal structure. But even without such a definition, we can use our causal models to evaluate counterfactuals and predict the effects of interventions. This tells us that there is causal structure among individual events that is *not* actual causation.

Once we recognize that actual causation is a specialized causal concept that exists as a kind of overlay on a more basic causal skeleton of causal structure, it becomes more palatable to admit that actual causation may be like justice: multi-faceted, partly subjective, impossible to define precisely. We may admit this without denying that there is objective causal structure in the world, the kind of structure that can be rigorously investigated by using formal methods and empirical investigation. This does not mean that attempts to define actual causation are pointless.¹³ For example, by embedding a concept of actual causation in a richer framework for investigating causation, we are better placed to ask and answer the question of why we have and use a notion of actual causation.¹⁴ But thanks to Pearl, we may be a bit more forgiving on ourselves if our definitions of actual causation come up short. Our understanding of causation in general does not hang in the balance.

References

- S. Beckers and J. Vennekens. 2017. The transitivity and asymmetry of actual causation. *Ergo* 4, 1, 17. DOI: <https://doi.org/10.3998/ergo.12405314.0004.001>.
- S. Beckers and J. Vennekens. 2018. A principled approach to defining actual causation. *Synthese* 195, 2, 835–862. DOI: <https://doi.org/10.1007/s11229-016-1247-1>.
- T. Blanchard and J. Schaffer. 2017. Cause without default. In *Making a Difference: Essays on the Philosophy of Causation*. Oxford University Press, Oxford, 175–214. DOI: <https://doi.org/10.1093/oso/9780198746911.001.0001>.
- R. Briggs. 2012. Interventionist counterfactuals. *Philos. Stud.* 160, 139–166. DOI: <https://doi.org/10.1007/s11098-012-9908-5>.
- L. Fenton-Glynn. 2017. A proposed probabilistic extension of the Halpern and Pearl definition of “actual cause.” *Br. J. Philos. Sci.* 68, 4, 1061–1124. DOI: <https://doi.org/10.1093/bjps/axv056>.
- D. Gallow. 2021. A Model-invariant Theory of Causation. *Philos. Rev.* 130, 45–96. DOI: <https://doi.org/10.1215/00318108-8699682>.

13. See Glymour et al. [2010] for skepticism on this score.

14. See Hitchcock [2017] for one attempt to do this.

- C. Glymour and F. Wimberly. 2007. Actual causes and thought experiments. In J. Campbell, M. O'Rourke, and H. Silverstein (Eds.), *Causation and Explanation*. MIT Press, Cambridge, MA, 43–67.
- C. Glymour, D. Danks, B. Glymour, F. Eberhardt, J. Ramsey, R. Scheines, P. Spirtes, C. M. Teng, and J. Zhang. 2010. Actual causation: A stone soup essay. *Synthese* 175, 169–192. DOI: <https://doi.org/10.1007/s11229-009-9497-9>.
- N. Hall. 2007. Structural equations and causation. *Philos. Stud.* 132, 109–136. DOI: <https://doi.org/10.1007/s11098-008-9216-2>.
- J. Y. Halpern. 2008. Defaults and normality in causal structures. In *Principles of Knowledge Representation and Reasoning: Proceedings Eleventh International Conference (KR '08)*. 198–208.
- J. Y. Halpern. 2016. *Actual Causality*. M.I.T. Press, Cambridge, MA. DOI: <https://doi.org/10.7551/mitpress/10809.001.0001>.
- J. Y. Halpern and C. Hitchcock. 2010. Actual causation and the art of modeling. In *Causality, Probability, and Heuristics: A Tribute to Judea Pearl*. College Publications, London, 383–406.
- J. Y. Halpern and C. Hitchcock. 2015. Graded causation and defaults. *Br. J. Philos. Sci.* 66, 413–457. DOI: <https://doi.org/10.1093/bjps/axt050>.
- J. Y. Halpern and J. Pearl. 2001a. Causes and explanations: A structural-model approach. Part I: Causes. In *Proceedings Seventeenth Conference on Uncertainty in Artificial Intelligence (UAI 2001)*. 194–202.
- J. Y. Halpern and J. Pearl. 2001b. Causes and explanations: A structural-model approach. Part II: Explanation. In *Proceedings Seventeenth International Joint Conference on Artificial Intelligence (IJCAI '01)*. 27–34.
- J. Y. Halpern and J. Pearl. 2005a. Causes and explanations: A structural-model approach. Part I: Causes. *Br. J. Philos. Sci.* 56, 4, 843–887. DOI: <https://doi.org/10.1093/bjps/axi147>.
- J. Y. Halpern and J. Pearl. 2005b. Causes and explanations: A structural-model approach. Part II: Explanations. *Br. J. Philos. Sci.* 56, 4, 889–911. DOI: <https://doi.org/10.1093/bjps/axi148>.
- C. Hitchcock. 2001. The intransitivity of causation revealed in equations and graphs. *J. Philos.* 98, 6, 273–299. DOI: <https://doi.org/10.2307/2678432>.
- C. Hitchcock. 2007. Prevention, preemption, and the principle of sufficient reason. *Philos. Rev.* 116, 495–532. DOI: <https://doi.org/10.1215/00318108-2007-012>.
- C. Hitchcock. 2017. Actual causation: What's the use? In H. Beebe, C. Hitchcock, and H. Price (Eds.), *Making a Difference: Essays on the Philosophy of Causation*. Oxford University Press, Oxford, 116–131. DOI: <https://doi.org/10.1093/oso/9780198746911.001.0001>.
- M. Hopkins and J. Pearl. 2003. Clarifying the usage of structural models for commonsense causal reasoning. In *Logical Formalizations of Commonsense Reasoning: Papers from the AAAI Spring Symposium*. AAAI Press, Menlo Park, CA, 83–89.
- D. Lewis. 1973. Causation. *J. Philos.* 70, 556–567. DOI: <https://doi.org/10.2307/2025310>.

- D. Lewis. 1979. Counterfactual dependence and time's arrow. *Noûs* 13, 455–476. DOI: <https://doi.org/10.2307/2215339>.
- D. Lewis. 1986. Causation. In *Philosophical Papers*, Vol. II. Oxford University Press, Oxford. Includes Postscripts A-E to “Causation,” 159–213. DOI: <https://doi.org/10.1093/0195036468.001.0001>.
- J. A. McLaughlin. 1925. Proximate cause. *Harv. L. Rev.* 39, 149–199. DOI: <https://doi.org/10.2307/1328484>.
- P. Menzies. 2004. Causal models, token causation, and processes. *Philos. Sci.* 71, 820–832. DOI: <https://doi.org/10.1086/425057>.
- P. Menzies. 2017. The problem of counterfactual isomorphs. In H. Beebe, C. Hitchcock, and H. Price (Eds.), *Making a Difference: Essays on the Philosophy of Causation*. Oxford University Press, Oxford, 153–172. DOI: <https://doi.org/10.1093/oso/9780198746911.001.0001>.
- J. Pearl. 1998. *On the Definition of Actual Cause*. Technical Report R-259, Department of Computer Science, University of California, Los Angeles, CA.
- J. Pearl. 2000. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, Cambridge. DOI: <https://doi.org/10.1017/S0266466603004109>.
- J. Pearl. 2009. *Causality: Models, Reasoning, and Inference* (2nd. ed). Cambridge University Press, Cambridge. DOI: <https://doi.org/10.1017/CBO9780511803161>.
- J. Woodward. 2003. *Making Things Happen: A Theory of Causal Explanation*. Oxford University Press, Oxford. DOI: <https://doi.org/10.1093/0195155270.001.0001>.

Causal Diagram and Social Science Research

Kosuke Imai (Harvard University)

It is a tremendous honor for me to contribute to the volume celebrating Judea Pearl's work. As the Turing Award signifies, Judea is no doubt one of the giants (along with Don Rubin and Jamie Robins) who created and developed the interdisciplinary field of causal inference methodology. Many of us have built and will continue to build our research on his foundational work. Personally, I had the pleasure of working with Judea as co-editor of *Journal of Causal Inference* over the last several years. I also learned a great deal from Judea's work on causal mediation in many occasions, including our lively exchanges in a journal [Imai et al. 2014, Pearl 2014]. In this chapter, I would like to briefly describe the impact Judea's work has had on social science research and then illustrate it with two examples from my own recent research. Finally, I will briefly discuss how Judea's work may advance the future of causal research in the social sciences.

33.1 Graphical Causal Models and Social Science Research

Judea Pearl's work on the use of graphical models for causal inference [Pearl 2000] has found many applications in the field of epidemiology. However, graphical causal models have not yet made their way into mainstream social science research. For example, as Judea himself acknowledges, many popular econometrics textbooks do not cover the graphical approach [Chen and Pearl 2013]. Although there exist some pedagogical work in sociology that introduces the graphical models framework [Elwert 2013, Morgan and Winship 2007], most social scientists exclusively rely on the potential outcomes framework in their teaching and research. Although it is always difficult to make a significant impact in another discipline, the absence of graphical causal models may come as a surprise given that econometrics and other social science methodology fields have a long tradition of

structural equation models, which can be represented by graphical causal models [Pearl 2015].

My own view is that graphical causal models have the potential to be applied in social science research that studies complex causal relationships. Social science has experienced the “causal inference revolution” over the last 30 years. As a result, researchers pay more attention to the issues of causal identification in order to distinguish causal relationships from associations. The potential outcomes framework has provided an intuitive and powerful way to formally conduct causal analyses. In many simple problems, it has provided the necessary tools and produced numerous methodological developments, from instrumental variables to regression discontinuity and difference-in-differences designs. However, researchers are beginning to study more complex causal relationships including spillover and carryover effects. I believe that graphical causal models can play an essential role in such studies by effectively communicating causal assumptions and allowing researchers to formally derive identification results. I will illustrate this point by briefly describing two recent examples from my own recent research [Imai and Kim 2019, Imai et al. 2020].

33.2 Two Applications of Graphical Causal Models

33.2.1 Causal Inference with Panel Data

Many social scientists rely upon linear regression models with fixed effects when estimating causal effects from panel data in observational studies. Suppose we have a simple random sample of N units, for each of which we observe a total of T repeated measurements. We use X_{it} to represent a binary treatment variable where it equals 1 if unit i receives the treatment at time t and equals zero otherwise. If we use Y_{it} to denote the outcome variable for unit i at time t , then a canonical linear regression model with unit fixed effects is given by,

$$Y_{it} = \alpha_i + \beta X_{it} + \varepsilon_{it} \quad (33.1)$$

where α_i represents the fixed effect for unit i and ε_{it} is the error term with $\mathbb{E}(\varepsilon_{it}) = 0$. Often, researchers also include a set of time-varying confounders Z_{it} as an attempt to adjust for them.

These and other related linear regression models with fixed effects are extremely popular among applied social scientists. The main reason for this popularity is that the fixed effects α_i can adjust for any unobserved time-invariant, unit-specific confounders U_i . Since most researchers worry about unobserved confounding, the inclusion of fixed effects gives them great comfort. However, most

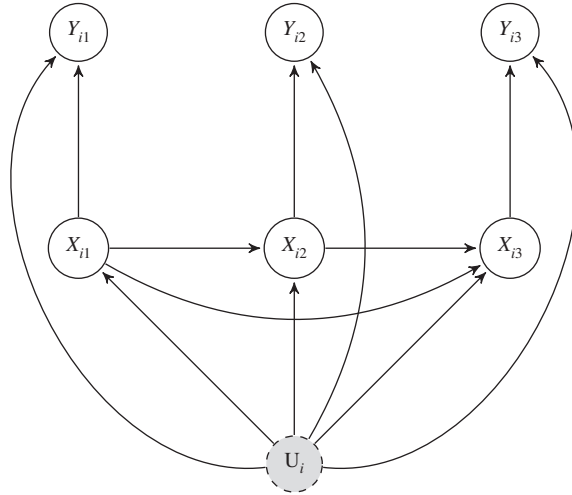


Figure 33.1 Directed acyclic graph for regression models with unit fixed effects based on three time periods. The model is given in Equation 33.1. The outcome and treatment variables for unit i at time t are denoted by Y_{it} and X_{it} , respectively. The unobserved time-invariant, unit-specific confounders are denoted by U_i . This figure is reproduced from figure 1 of Imai and Kim [2019].

textbooks describe the assumption of the model given in Equation 33.1 as the so-called strict exogeneity, which can be written as,

$$\mathbb{E}(\varepsilon_{it} | \alpha_i, X_{it}) = 0. \quad (33.2)$$

In my experience, most applied researchers fail to gain an intuitive understanding of this assumption. A part of the problem is that the assumption is stated in terms of error term.

In contrast, directed acyclic graphs (DAGs) can much more effectively communicate the causal assumptions behind these types of models. Figure 33.1 presents the causal DAG for the model given in Equation 33.1. We observe that the model assumes the absence of causal dynamics. In particular, there is no arrow from a past outcome to a future treatment, implying that the former does not causally affect the latter. In fact, using the DAG, it is straightforward to show that the existence of such an arrow makes it impossible to non-parametrically identify the average causal effect of X_{it} on Y_{it} . Most importantly, the DAG effectively highlights the fundamental tradeoff in causal inference for panel data, which is difficult to see in the standard statement of the identification assumption given in Equation 33.2. The ability to adjust for unobserved, time-invariant, and unit-specific confounders U_i comes with a cost: one must assume away dynamic causal relationships.

33.2.2 Causal Inference with Interference between Units

The second example, which is based on the randomized evaluation of the Indian National Health Insurance Program [Imai et al. 2020], also illustrates the potential use of graphical causal models in social science research as a tool to effectively communicate certain causal assumptions. Consider a two-stage randomized experiment [Hudgens and Halloran 2008] in which randomly selected villages are assigned to one of the two different treatment assignment mechanisms, called “High” and “Low.” If a village is assigned to the High mechanism, then 80% of its households are randomly assigned to the treatment group. On the other hand, if a village is randomly assigned to the Low mechanism, only 40% of its households are randomly assigned to the treatment group. We use the binary random variable Z_{ij} to denote whether household i in village j is assigned to the treatment group ($Z_{ij} = 1$) or the control group ($Z_{ij} = 0$).

In this experiment, there was a problem of non-compliance because we could only encourage, but not enforce, the random treatment assignment for ethical and logistical reasons. As a result, some households in the treatment group did not sign up for the insurance program while others in the control group ended up enrolling in it. Let D_{ij} represent the binary treatment receipt variable, which is equal to 1 if household i in village j actually received the treatment and is equal to 0 otherwise. To further complicate this evaluation project, people appear to have talked to each other within each village about the insurance program and as a result the treatment receipt of one household D_{ij} may have been affected by the treatment assignment of another household $Z_{i'j}$ within the same village. Moreover, researchers have hypothesized that there may exist a spillover effect of one’s treatment receipt D_{ij} on the outcome of another household $Y_{i'j}$. For example, if a large number of households enrolled in the insurance program, it may affect the healthcare utilization of another household who did not sign up for the program because of the overcrowding of local clinics.

Let’s assume the so-called partial interference assumption, which states that there exists no interference across villages; that is, households affect one another only within each village. In the potential outcomes framework, this means that the potential values of one’s treatment receipt and outcome depend on the treatment assignment vector, that is, $Y_{ij} = Y_{ij}(\mathbf{Z}_j)$ and $D_{ij}(\mathbf{Z}_j)$ where $\mathbf{Z}_j = (Z_{1j}, \dots, Z_{n_jj})$ is the vector of treatment assignments for n_j households in village j . When analyzing such a complex experiment, several assumptions are necessary to make progress. Imai et al. [2020] extend the exclusion restriction of the standard instrumental variables analysis [Angrist et al. 1996, Balke and Pearl 1997] and assume that

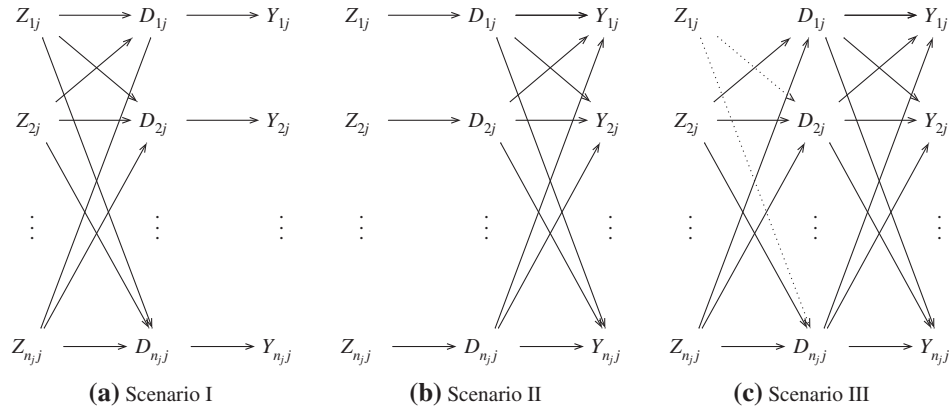


Figure 33.2 Three identification assumptions restricting interference. This figure is reproduced from figure 1 of Imai et al. [2020]. Scenario I assumes no spillover effect of the treatment receipt D on the outcome Y . Scenario II assumes no spillover effect of the treatment assignment Z on D . Finally, Scenario III assumes no spillover effect of Z on D (dotted arrows) among non-compliers whose own treatment assignment Z_{ij} does not affect their own treatment receipt D_{ij} .

the treatment receipt vector \mathbf{Z}_j affects the outcome Y_{ij} only through the treatment receipt vector $\mathbf{D}_j = (D_{1j}, \dots, D_{nj})$. Imai et al. [2020] then consider three additional restrictions on the patterns of interference for identifying causal effects. Although these assumptions can be expressed using the potential outcomes, the resulting notation is complex and makes it difficult to effectively communicate the intuitive ideas behind them.

Figure 33.2 illustrates the effectiveness of causal DAGs in this application. The first scenario in the left depicts the assumption of no spillover effect of the treatment receipt on the outcome. This assumption is represented by the absence of arrows from D_{ij} to $Y_{i'j}$ for $i \neq i'$. The second scenario in the middle represents the assumption of no spillover effect of the treatment assignment on the treatment receipt, which is indicated by the absence of arrows from Z_{ij} to $D_{i'j}$ for $i \neq i'$. Finally, the third scenario in the right illustrates the assumption of no spillover effect of Z on D among “non-compliers” (dotted arrows) whose own treatment assignment Z_{ij} does not affect their own treatment receipt D_{ij} (no arrow from Z_{ij} to D_{ij}). In addition, all three scenarios assume the aforementioned exclusion restriction as indicated by the absence of direct arrows from Z_{ij} to Y_{ij} . Thus, although there are other types of assumptions such as monotonicity that are difficult to represent in causal DAGs, they can visually illustrate many complex assumptions in an intuitive manner.

33.3 The Future of Causal Research in the Social Sciences

Over the last three decades, the Causal Revolution has swept through social sciences. Of course, the main goal of social science research has always been causal inference because social scientists are primarily concerned about the causes and consequences of policies and human behavior in the society. And yet, it was the formalization of causal language that has brought the explosion of methodological development and scientific applications. Judea Pearl has played a major role in this Causal Revolution and has transformed many scientific disciplines.

The first half of the Causal Revolution has focused upon simple settings, in which spillover and carryover effects are often assumed to be absent. However, in social sciences, human beings constantly interact with each other and as a result spillover effects are the rule rather than the exception. In addition, many social scientists collect repeated measurements and are beginning to conduct sequential experiments. More data on social networks and geographical information systems (GIS) are also becoming available to researchers. These developments call for new methodologies that can handle complex causal relationships across time and space. I expect our new methods to be built upon the foundation Judea has developed, and his impact in the social sciences will be felt for years to come.

References

- J. D. Angrist, G. W. Imbens, and D. B. Rubin. 1996. Identification of causal effects using instrumental variables (with discussion). *J. Am. Stat. Assoc.* 91, 434, 444–455. DOI: <https://doi.org/10.2307/2291629>.
- A. Balke and J. Pearl. 1997. Bounds on treatment effects from studies with imperfect compliance. *J. Am. Stat. Assoc.* 92, 1171–1176. DOI: <https://doi.org/10.1080/01621459.1997.10474074>.
- B. Chen and J. Pearl. 2013. Regression and causation: A critical examination of six econometrics textbooks. *Real-World Econ. Rev.* 65, 2–20.
- F. Elwert. 2013. *Handbook of Causal Analysis for Social Research*. Chapter Graphical Causal Models. Springer, Dordrecht, 245–273. DOI: <https://doi.org/10.1007/978-94-007-6094-3>.
- M. G. Hudgens and E. Halloran. 2008. Toward causal inference with interference. *J. Am. Stat. Assoc.* 103, 482 (June 2008), 832–842. DOI: <https://doi.org/10.1198/016214508000000292>.
- K. Imai, Z. Jiang, and A. Malai. 2020. Causal inference with interference and noncompliance in two-stage randomized experiments. *J. Am. Stat. Assoc.* DOI: <https://doi.org/10.1080/01621459.2020.1775612>.
- K. Imai, L. Keele, D. Tingley, and T. Yamamoto. 2014. Comment on Pearl: Practical implications of theoretical results for causal mediation analysis. *Psychol. Methods* 19, 4, 482–487. DOI: <https://doi.org/10.1037/met0000021>.

- K. Imai and I. S. Kim. 2019. When should we use linear unit fixed effects regression models for causal inference with longitudinal data? *Am. J. Pol. Sci.* 63, 2, 467–490. DOI: <https://doi.org/10.1111/ajps.12417>.
- S. L. Morgan and C. Winship. 2007. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. Cambridge University Press, New York. DOI: <https://doi.org/10.1017/CBO9780511804564>.
- J. Pearl. 2000. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York. DOI: <https://doi.org/10.1017/S026646660300410>.
- J. Pearl. 2014. Interpretation and identification of causal mediation. *Psychol. Methods* 19, 4, 459–481. DOI: <https://doi.org/10.1037/a0036434>.
- J. Pearl. 2015. Trygve Haavelmo and the emergence of causal calculus. *Econom. Theory* 31, 152–179. DOI: <https://doi.org/10.1017/S0266466614000231>.

Causal Graphs for Missing Data: A Gentle Introduction

Karthika Mohan (University of California)

In this chapter we describe how causal graphs can be used for processing missing data. In particular, we model the missingness process using causal graphs and present graph-based definitions of various missingness mechanisms. Given a graph and a target quantity to be estimated, we present various methods for determining if and how a consistent estimate of the quantity can be computed. We further present techniques for detecting misspecifications in the model. We demonstrate all of the above using toy examples and small graphs, thus making it easy to understand the various intricacies and nuances of graph-based missing data analysis.

34.1 Introduction

Consider the following problems: (i) estimating the average income of a population in which the wealthy are reluctant to reveal their income, (ii) estimating the causal effect of diet and stress on obesity, given a dataset in which teenagers left several questions unanswered, and (iii) making product recommendations using data in which customers rated items only when they loved it. The underlying common theme in (i), (ii), and (iii) above is the estimation of a desired target quantity given missing data, that is, data in which values of one or more variables are missing.

Problems caused due to missing data are notoriously complex, afflict all branches of empirical sciences, and could potentially bias the outcomes of studies. Much of the research on missing data has focused on identifying conditions (such as Missing at Random [MAR] and Missing Completely at Random [MCAR])

[Rubin 1976]) under which the causes of missingness can be ignored when estimating quantities of interest. A widely held belief is that when the underlying cause of missingness is not random (Missing Not at Random (MNAR) [Rubin 1976]), it is rarely possible to compute estimates with any degree of confidence (example 1.17 in Little and Rubin [2002]).

In this chapter we discuss the recent advances in missing data theory that facilitate processing of MNAR data (i.e., non-ignorable missingness); in particular, we focus on *recoverability* (i.e., computing consistent estimates of quantities of interest) and *testability* (i.e., developing tests to determine the compatibility of a model with the available data). The following section describes missingness graphs (m-graphs), which are causal graphs that encode the (causal and statistical) assumptions about the process that generated missing data.

34.2 Missingness Graphs

Let $G(\mathbf{V}, E)$ be the causal directed acyclic graph (DAG) where \mathbf{V} is the set of nodes and E is the set of edges. Nodes in the graph correspond to variables in the dataset and are partitioned into five categories, namely, $\mathbf{V} = V_o \cup V_m \cup U \cup V^* \cup R$ as described in Table 34.1. For example, in Figure 34.1(a), $V_o = \{G\}$, $V_m = \{I\}$, $R = \{R_I\}$, $V^* = \{I^*\}$, and $U = \{\}$.

Table 34.1 Notations

V_o	Set of all fully observed variables
V_m	Set of variables with missing values
U	Set of unobserved (latent) variables
R	Set of missingness mechanisms
V^*	Set of all proxy variables

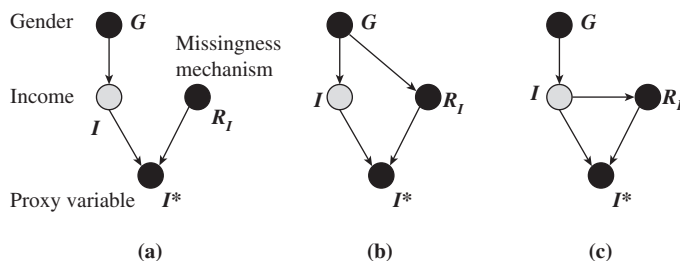


Figure 34.1 Causal graphs depicting various missingness mechanisms: (a) Missing Completely at Random (MCAR), (b) Missing at Random (MAR), and (c) Missing Not at Random (MNAR).

Every $X \in V_m$ is associated with two variables R_X and X^* , where X^* is the proxy variable that is actually observed and R_X represents the status of the mechanism responsible for the missing values in X^* ; formally,

$$x^* = f(r_x, x) = \begin{cases} x & \text{if } r_x = 0 \\ m & \text{if } r_x = 1 \end{cases} \quad (34.1)$$

Unless stated otherwise, it is assumed that no variable in $V_o \cup V_m \cup U$ is a child of an R variable. U is the set of unobserved nodes, also called latent variables. Two nodes X and Y can be connected by a directed edge, that is, $X \rightarrow Y$, indicating that X is a cause of Y , or by a bi-directed edge, $X \leftrightarrow Y$, denoting the existence of a latent variable $U_i \in U$ that is a parent of both X and Y . This graphical representation is called a *missingness graph* (or *m-graph*) [Mohan et al. 2013]. $P^*(V^*, V_o, R)$ is called the observed data distribution, and $P(V_m, V_o, R)$ is called the true distribution. Any given true and observed data distribution are said to be compatible if the latter can be constructed from the former by repeatedly applying Equation (34.1). Conditional independencies are read off the graph using the d-separation criterion [Pearl 2009]. For any binary variable X , x' and x denote $X = 0$ and $X = 1$, respectively.

34.2.1 Graphical Representation of Missingness Categories

The graphical model-based definitions of the various missingness mechanisms [Rubin 1976, Little and Rubin 2002] that can be used to effortlessly decide the missingness categories are as follows:

1. Data are MCAR if $V_m \cup V_o \cup U \perp\!\!\!\perp R$ holds in the m-graph. Example: m-graph in Figure 34.1(a) in which $\{G, I\} \perp\!\!\!\perp R_I$ holds. Essentially, parents of R variables can only be other R variables.
2. Data are MAR if $V_m \cup U \perp\!\!\!\perp R | V_o$ holds in the m-graph. Example: m-graph in Figure 34.1(b) in which $\{I\} \perp\!\!\!\perp R_I | G$ holds. For MAR to hold, no parent of any R variable should belong to $V_m \cup U$; put differently, parents of R variables may only belong to $V_o \cup R$.
3. Data that are not MAR fall under the MNAR category. Example: m-graph in Figure 34.1(c) in which $\{I\} \not\perp\!\!\!\perp R_I | G$. In this case at least one R variable will have a parent that is either a latent variable or a variable with missing values, that is, belonging to $V_m \cup U$.

Remark 34.1 Observe that the graphical condition for MCAR immediately satisfies that for MAR; if parents of R variables may only be other R variables, then they clearly cannot be in $V_m \cup U$. Thus any model that is MCAR is MAR as well; this also follows from the weak union graphoid axiom [Pearl 2009].

34.3 Recoverability

Let Q denote a quantity of interest such as the joint/conditional distribution and causal effect. Given an m-graph G , Q is recoverable if there exists an algorithm that can (asymptotically) compute the true value of Q as if no data were missing. In the remainder of this section, we exemplify various recoverability techniques and their intricacies using small graphs as favored and taught by Judea Pearl, and as seen in many of his publications.

34.3.1 Recoverability in MAR and MCAR Problems

Consider the problem of recovering the joint distribution $P(G, I)$ given the m-graph in Figure 34.1(a) and the observed data distribution in Table 34.2.

$$\begin{aligned} P(G, I) &= P(G, I | r'_I) \text{ (since } \{G, I\} \perp\!\!\!\perp R_I \text{ in the m-graph)} \\ &= P(G, I^* | r'_I) \text{ (using Equation 34.1)} \end{aligned} \quad (34.2)$$

The preceding equations demonstrated how $P(G, I)$, which is a function of the partially observed variable I and fully observed variable G , is transformed into one over variables in the observed data distribution, I^* and G . The final expression derived in Equation (34.2), $P(I^*, G | r'_I)$, is an estimand for $P(G, I)$, that is, it is an expression for $P(G, I)$ in terms of the available data that precisely defines what needs to be estimated. Recoverability is established once we derive an estimand. Note that the observed data distribution per se played no part in recoverability, which was established using assumptions in the m-graph ($\{G, I\} \perp\!\!\!\perp R_I$) and the missingness equation Equation (34.1). Thus, *recoverability is a property of the m-graph*.

34.3.1.1 Recoverability of Joint Distribution in MCAR and MAR Models

We shall now show that the joint distribution, $P(V_m, V_o)$, is recoverable in all MCAR and MAR m-graphs.

Table 34.2 Observed data distribution generated by the m-graph in Figure 34.1(a)

G	I^*	R_I	$P(G, I^*, R_I)$
M	H	0	p_1
M	L	0	p_2
F	H	0	p_3
F	L	0	p_4
M	M	1	p_5
F	M	1	p_6

G and I are binary variables that can take values Male (M) and Female (F), and High (H) and Low (L), respectively. P_i s denote probabilities such that $\sum_{i=1}^6 p_i = 1$.

Recoverability of joint distribution $P(V_o, V_m)$ in MCAR problems:

$$\begin{aligned} P(V_o, V_m) &= P(V_o, V_m | R = 0) \text{ (since } (V_m, V_o) \perp\!\!\!\perp R \text{ when MCAR holds in an m-graph)} \\ &= P(V_o, V^* | R = 0) \text{ (using Equation 34.1)} \end{aligned} \quad (34.3)$$

Recoverability of joint distribution $P(V_o, V_m)$ in MAR problems:

$$\begin{aligned} P(V_o, V_m) &= P(V_m | V_o) P(V_o) \\ &= P(V_m | V_o, R = 0) P(V_o) \text{ (since } V_m \perp\!\!\!\perp R | V_o \text{ when MAR holds in an m-graph)} \\ &= P(V^* | V_o, R = 0) P(V_o) \text{ (using Equation 34.1)} \end{aligned} \quad (34.4)$$

Equations (34.3) and (34.4) establish recoverability by presenting an estimand for the joint distribution.

34.3.1.2 Recoverability as a Guide for Estimation

Having established recoverability for all MAR and MCAR problems, we will now show how recoverability serves as a guide for estimation. We will exemplify estimation using deletion-based procedures.

The estimand in Equation 34.2 can be expressed as,

$$P(I^*, G | r'_I) = \frac{P(I^*, G, r'_I)}{P(r'_I)}$$

It licenses the estimation of $P(G, I)$ exclusively from cases/samples in which $V_m = \{I\}$ is always observed, that is, $R_I = 0$. This procedure is known as *listwise deletion* or *complete case analysis*. In order to estimate using this method we may only use the first four rows in Table 34.2 in which $R_I = 0$. Table 34.3 shows the joint distribution estimated in this manner. However, notice that the information contained in the last two rows of Table 34.2 in which $R_I = 1$ has been left unused, thus resulting in wastage of samples [McKnight et al. 2007, Enders 2010]. Hence this procedure, while convenient and fast to implement, is not recommended in practice even if it guarantees consistent estimates. We describe below an alternate procedure that utilizes samples more efficiently.

As stated in Remark 34.1, any model that is MCAR is also MAR; hence, any estimation algorithm applicable to MAR is applicable to MCAR as well. Thus, to recover $P(G, I)$ given the MCAR graph in Figure 34.1(a), we could apply Equation (34.4) to obtain:

$$P(G, I) = P(I^* | G, r'_I) P(G)$$

Table 34.3 Complete case analysis–based estimation of joint distribution given the m-graph in Figure 34.1(a) and the data in Table 34.2

G	I	$P(G, I)$
M	H	$\frac{p_1}{p_1+p_2+p_3+p_4}$
M	L	$\frac{p_2}{p_1+p_2+p_3+p_4}$
F	H	$\frac{p_3}{p_1+p_2+p_3+p_4}$
F	L	$\frac{p_4}{p_1+p_2+p_3+p_4}$

The estimand above dictates that we compute $P(I^*|G, r'_i)$ exclusively from samples in which I is observed and $P(G)$ from all samples, including those in which I is missing as shown in Table 34.4. Clearly, this utilizes data in a better manner compared to listwise deletion exemplified in Table 34.3. Efficient graph-based deletion procedures for MCAR and MAR that exploit available samples to a greater extent, thus yielding better quality estimates, are discussed in [Van den Broeck et al. \[2015\]](#).

34.3.2 Recoverability in MNAR Problems

In this subsection, we exemplify various recoverability techniques for MNAR using simple models.

34.3.2.1 Recovering $P(X, Y)$ Given the m-graph G in Figure 34.2(a)

G is one of the simplest examples of MNAR in which missingness in R_X is caused by Y , a variable with missing values. $V_m = \{X, Y\}$, $V_o = \{\}$ and due to the edge from Y to R_X , MAR does not hold, that is, $\{X, Y\} \not\perp\!\!\!\perp \{R_x, R_y\}$. Joint distribution $P(X, Y)$ is recoverable given G as shown below:

$$\begin{aligned}
 P(X, Y) &= P(X|Y)P(Y) \text{ (using chain rule)} \\
 &= P(X|Y, r'_x, r'_y)P(Y|r'_y) \text{ (since } X \perp\!\!\!\perp R_x, R_y|Y \text{ and } Y \perp\!\!\!\perp R_y \text{ hold in } G) \\
 &= P(X^*|Y^*, r'_x, r'_y)P(Y^*|r'_y) \text{ (using Equation 34.1)}
 \end{aligned}$$

We call the above technique sequential factorization [[Mohan and Pearl 2018](#)]. It is sensitive to the order of factorization. Had we factorized $P(X, Y)$ as $P(Y|X)P(X)$ in

Table 34.4 A deletion-based method with less sample wastage for estimating joint distribution given the m-graph in Figure 34.1(a) and the data in Table 34.2

G	I	$P(G, I)$
M	H	$\frac{p_1(p_1+p_2+p_5)}{p_1+p_2}$
M	L	$\frac{p_2(p_1+p_2+p_5)}{p_1+p_2}$
F	H	$\frac{p_3(p_3+p_4+p_6)}{p_3+p_4}$
F	L	$\frac{p_4(p_3+p_4+p_6)}{p_3+p_4}$

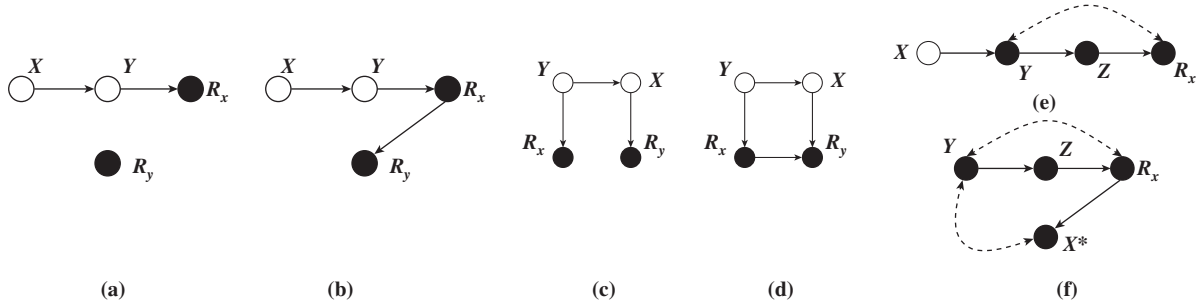


Figure 34.2 (a)–(e) m-graphs depicting MNAR missingness. Proxy variables have not been explicitly portrayed to keep the figures simple and clear. (f) Graph corresponding to m-graph (e) in which X is treated as a latent variable.

the first step, it would have been harder to establish recoverability. We further note that the estimand dictates that $P(X|Y)$ be estimated from samples in which both X and Y are observed and $P(Y)$ be estimated from samples in which Y is observed, regardless of the missingness status of X .

34.3.2.2 Recovering $P(X, Y)$ Given the m-graph in Figure 34.2(b)

For exactly the same reasons as those described in Section 34.3.2.1, this m-graph also depicts MNAR. However, notice that m-graphs in Figure 34.2(a) and (b) differ in the way the R variables are connected. An edge exists between the R variables in m-graph (b) whereas in (a) $R_x \perp\!\!\!\perp R_y$. We show below that this seemingly minor change results in a substantially different estimand (and estimation process).

$$\begin{aligned}
 P(X, Y) &= P(X|Y)P(Y) \\
 &= P(X|Y, r'_x, r'_y)P(Y) \text{ (since } X \perp\!\!\!\perp R_x, R_y|Y) \\
 &= P(X|Y, r'_x, r'_y) \sum_{R_x} P(Y|R_x)P(R_x) \\
 &= P(X|Y, r'_x, r'_y) \sum_{R_x} P(Y|R_x, r'_y)P(R_x) \text{ (since } Y \perp\!\!\!\perp R_y|R_x) \\
 &= P(X^*|Y^*, r'_x, r'_y) \sum_{R_x} P(Y^*|R_x, r'_y)P(R_x) \text{ (using Equation 34.1)}
 \end{aligned}$$

This example underscores the importance of modeling the causal relationship among R variables. For instance, had the m-graph been $X \rightarrow Y \rightarrow R_x \leftrightarrow R_y$, the estimand for $P(X, Y)$ would have been identical to the one derived in Section 34.3.2.1.

34.3.2.3 Recovering $P(X, Y)$ Given the m-graph in Figure 34.2(c)

The parents of both R variables in this m-graph are variables with missing values. Hence the m-graph depicts MNAR missingness. Recoverability of $P(X, Y)$ given

this m-graph is discussed in [Mohan et al. \[2013\]](#) and the recoverability procedure presented therein forms the basis for most recoverability methods for MNAR. In this subsection we present an alternate method that requires inspecting all missingness patterns one by one.

$$\begin{aligned} P(X, Y) &= \sum_{R_x, R_y} P(X, Y, R_x, R_y) \\ &= P(X, Y, r'_x, r'_y) + P(X, Y, r'_x, r_y) \\ &\quad + P(X, Y, R_x =, r'_y) + P(X, Y, r_x, r_y) \end{aligned}$$

To prove recoverability of $P(X, Y)$, we will show that each term in the sum is recoverable. It follows from Equation (34.1) that $P(X, Y, r'_x, r'_y) = P(X^*, Y^*, r'_x, r'_y)$ and hence $P(X, Y, r'_x, r'_y)$ is recoverable. We will now show that $P(X, Y, r_x, r'_y)$ is recoverable.

$$\begin{aligned} P(X, Y, r_x, r'_y) &= P(X|Y, r_x, r'_y)P(Y|r_x, r'_y)P(r_x, r'_y) \\ &= P(X|Y, r'_x, r'_y)P(Y|r_x, r'_y)P(r_x, r'_y) \text{ (since } X \perp\!\!\!\perp R_x|Y, R_y) \\ &= P(X^*|Y^*, r'_x, r'_y)P(Y^*|r_x, r'_y)P(r_x, r'_y) \text{ (using Equation 34.1)} \end{aligned}$$

In a similar manner we can show that $P(X, Y, r'_x, r_y) = P(Y^*|X^*, r'_x, r'_y)P(X^*|r'_x, r_y)P(r'_x, r_y)$ and hence recoverable. What remains to be shown is that $P(X, Y, r_x, r_y)$ is recoverable.

$$\begin{aligned} P(X, Y, r_x, r_y) &= P(X|Y, r_x, r_y)P(r_x|Y, r_y)P(Y, r_y) \\ &= P(X|Y, r'_x, r_y)P(r_x|Y, r_y)P(Y, r_y) \end{aligned} \tag{34.5}$$

$$\begin{aligned} &= \frac{P(X, Y, r'_x, r_y)}{P(Y, r'_x, r_y)}P(r_x|Y, r_y)P(Y, r_y) \\ &= \frac{P(Y|X, r'_x, r_y)P(X, r'_x, r_y)}{P(r'_x|Y, r_y)P(Y, r_y)}P(r_x|Y, r_y)P(Y, r_y) \\ &= \frac{P(Y^*|X^*, r'_x, r'_y)P(X^*, r'_x, r_y)}{P(r'_x|Y^*, r'_y)}P(r_x|Y^*, r'_y) \end{aligned} \tag{34.6}$$

In Equation (34.5), we replaced r_x with r'_x since $X \perp\!\!\!\perp R_x|Y, R_y$ holds in the graph. In Equation (34.6), we first cancelled out $P(Y, r_y)$ from the numerator and denominator, and then replaced r_y with r'_y in (i) $P(Y|X, r'_x, r_y)$ by applying $Y \perp\!\!\!\perp R_y|X, R_x$ and in (ii) $P(r'_x|Y, r_y)$ and $P(r_x|Y, r_y)$ by applying $R_x \perp\!\!\!\perp R_y|Y$. Finally, using Equation (34.1) we replaced Y with Y^* and X with X^* .

34.3.2.4 Recovering $P(X, Y)$ Given the m-graph in Figure 34.2(d)

The m-graph depicts MNAR for exactly the same reasons discussed in Section 34.3.2.4. Here we are recovering a conditional distribution as opposed to all previous examples of recoverability that discussed joint distributions.

$$\begin{aligned} P(X|Y) &= P(X|Y, r'_x) \text{ (since } X \perp\!\!\!\perp R_x|Y) \\ &= \frac{P(X, Y, r'_x)}{\sum_X P(X, Y, r'_x)} \end{aligned} \quad (34.7)$$

$$\begin{aligned} P(X, Y, r'_x) &= P(Y|X, r'_x)P(X, r'_x) \\ &= P(Y|X, r'_x, r'_y)P(X, r'_x) \text{ (since } Y \perp\!\!\!\perp R_y|X, R_x) \\ &= P(Y^*|X^*, r'_x, r'_y)P(X^*, r'_x) \end{aligned} \quad (34.8)$$

Substituting the right-hand side (RHS) of Equation (34.8) in the place of $P(X, Y, r'_x)$ in Equation (34.7), we get

$$P(X|Y) = \frac{P(Y^*|X^*, r'_x, r'_y)P(X^*, r'_x)}{\sum_X P(Y^*|X^*, r'_x, r'_y)P(X^*, r'_x)}$$

34.3.2.5 Recovering $P(X)$ Given the m-graph in Figure 34.2(e)

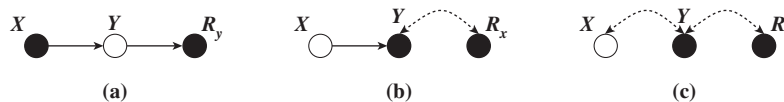
The dotted bi-directed edge indicates that there exists a latent variable that is a parent of both Y and R_x , and this makes the model MNAR. This graph is different from all the other m-graphs that we have examined thus far. Notice that here, although X and R_x are not connected by an edge, there exists no separating set that can separate them. This is because there are two paths between X and R_x ; on one path Y is a collider and Z , the descendant of a collider, and on the other path Y and Z are part of a chain. So, including Y or Z in the separating set will open the collider path, while excluding either one of them would leave the chain open. Interestingly, $P(X)$ is still recoverable as detailed below:

$$\begin{aligned} P(X) &= P(X|do(z)) \text{ (using rule 3 of do-calculus [Pearl 2009])} \\ &= P(X|do(z), r'_x) \text{ (using rule 1 of do-calculus [Pearl 2009])} \\ &= P(X^*|do(z), r'_x) \text{ (using Equation 34.1)} \end{aligned}$$

We have reduced the problem of recovering $P(X)$ to the problem of identifying the causal effect such that the causal query is defined over variables in the observed data distribution. Since the causal query is not a function of X , it can be identified using methods available in Shpitser and Pearl [2006] and the graph shown in Figure 34.2(f) in which X is treated as a latent variable.

Table 34.5 Observed data distribution $P(X^*, R_x)$ corresponding to the m-graph $X \rightarrow R_x$

X	R_x	$P(X^*, R_x)$
0	0	0.3
1	0	0.5
m	1	0.2

**Figure 34.3** m-graphs in which $P(X, Y)$ is not recoverable.

Finally, we note that although in this chapter we focus on discrete variables, recoverability techniques exist for continuous variables and have been discussed in Pearl [2013] and Mohan et al. [2018].

34.3.3 Non-recoverability

Consider the problem of recovering $P(X)$ given the m-graph $G : X \rightarrow R_x$. R_x is dependent on X and we have no additional information regarding this dependence. Table 34.5 presents a dataset generated by G . It could be that X is missing only when its value is 1 or it could be that X is missing only when its value is 0. In the former case $P(x') = 0.3$, whereas in the latter case $P(x') = 0.5$. Using the available information in G , it is not possible to find the (true) value of $P(X)$ even if we are given infinitely many samples, that is, $P(X)$ is non-recoverable. In fact, non-recoverability of $P(X)$ would persist even if G had more variables in it (formally proved in Mohan et al. [2013], Mohan and Pearl [2014a, 2014b]). In general, joint distribution is non-recoverable whenever there exists a variable X with missing values (i.e., $X \in V_m$) such that either:

1. X and R_x are neighbors or
2. X and R_x are connected by a path in which all intermediate nodes are colliders.

Thus, $P(X, Y)$ is non-recoverable in all the three m-graphs in Figure 34.3. However, in Figure 34.3(a) $P(X|Y)$ is recoverable, and in Figure 34.3(b) and (c) $P(X)$ is recoverable.

34.4 Testability

Testability when there is no missingness: When X and Y are fully observed variables, the independence statement $X \perp\!\!\!\perp Y$ is testable, that is, there exist distributions

over X and Y in which $X \perp\!\!\!\perp Y$ does not hold. Therefore, given a graph G and a distribution $P(X, Y)$, if the graph portrays $X \perp\!\!\!\perp Y$ and the claim does not hold in the distribution, then we can conclude that the graph and distribution are not compatible. Thus, under no missingness, d-separations serve as testable implications of a graphical model [Pearl 2009].

Non-testability under missingness: The simplest missing data distribution is $P(X^*, R_x)$, which is obtained when the substantive variable of interest is a single variable X . Let the query to be recovered be $P(X)$. As shown in the previous sections, recoverability of $P(X)$ hinges on $X \perp\!\!\!\perp R_x$; if it holds then $P(X)$ is recoverable, otherwise not. Given the decisive nature of this independence, can we test it?

$X \perp\!\!\!\perp R_x$ is testable only if it is refutable in all true distributions that are compatible with the observed data distribution. However, for any observed data distribution $P(X^*, R_x)$, there exists a true distribution $P'(X, R_x)$ in which $X \perp\!\!\!\perp R_x$ holds. It can be constructed as $P'(X, R_x) = P(X^* | R_x = 0)P(R_x)$. Hence the claim is not refutable. Put differently, independence claims between a variable and its mechanism are not testable [Mohan and Pearl 2014b].

Testable implications of m -graphs: d-separations that abide by the following syntactic rules are testable under missingness (X and Y are singletons) [Mohan and Pearl 2014a].

$$X \perp\!\!\!\perp Y | Z, R_x, R_y, R_z \quad (34.9)$$

$$X \perp\!\!\!\perp R_y | Z, R_x, R_z \quad (34.10)$$

$$R_x \perp\!\!\!\perp R_y | Z, R_z \quad (34.11)$$

Example of testability: Figure 34.2(a) encodes the conditional independence $X \perp\!\!\!\perp R_y | R_x$, which matches the syntactic rule 34.10 above when $Z = \{\}$. It follows from $X \perp\!\!\!\perp R_y | R_x$ that:

$$P(X | r_y, r'_x) = P(X | r'_y, r'_x)$$

Using Equation (34.1) we can rewrite the above as,

$$P(X^* | r_y, r'_x) = P(X^* | r'_y, r'_x)$$

The preceding claim, which is defined over X^*, R_x and R_y , is testable given the observed data distribution. If the claim is violated, we conclude that the model and data are not compatible. Note that this test not only detects incompatibility but also helps in locating the source of incompatibility.

On the indispensability of causal assumptions: Let $G_1 : X \perp\!\!\!\perp R_X$ and $G_2 : X \rightarrow R_X$. G_1 encodes the assumption $X \perp\!\!\!\perp R_X$, whereas G_2 does not. Since $X \perp\!\!\!\perp R_X$ is not testable, G_1 and G_2 are statistically indistinguishable, that is, any given observed data distribution $P(X^*, R_X)$ compatible with G_1 is also compatible with G_2 . However, they encode different causal assumptions. In G_1 where X does not cause its own missingness $P(X)$ is recoverable, whereas in G_2 where X causes its own missingness $P(X)$ is not recoverable. Thus, there exists no universal algorithm that can determine recoverability without examining the model and taking into account the embedded causal assumptions.

In conclusion, *missing data is a causal inference problem!*

References

- C. Enders. 2010. *Applied Missing Data Analysis*. Guilford Press, London, New York.
- R. Little and D. Rubin. 2002. *Statistical Analysis with Missing Data*. Wiley, New York.
- P. McKnight, K. McKnight, S. Sidani, and A. Figueredo. 2007. *Missing Data: A Gentle Introduction*. Guilford Press, London, New York.
- K. Mohan and J. Pearl. 2014a. On the testability of models with missing data. *Proceedings of AISTAT*. 33, 643–650.
- K. Mohan and J. Pearl. 2014b. Graphical models for recovering probabilistic and causal queries from missing data. In *Advances in Neural Information Processing Systems 27*. Curran Associates, Inc., Red Hook, NY, 1520–1528.
- K. Mohan and J. Pearl. 2018. Graphical models for processing missing data. *arXiv preprint arXiv:1801.03583*.
- K. Mohan, J. Pearl, and J. Tian. 2013. Graphical models for inference with missing data. In *Advances in Neural Information Processing Systems 26*. Curran Associates, Inc., Red Hook, New York, 1277–1285.
- K. Mohan, F. Thoenmes, and J. Pearl. 2018. Estimation with incomplete data: The linear case. In *IJCAI*. 5082–5088. DOI: <https://doi.org/10.24963/ijcai.2018/705>.
- J. Pearl. 2009. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York. DOI: <https://doi.org/10.1017/CBO9780511803161>.
- J. Pearl. 2013. Linear models: A useful “microscope” for causal analysis. *J. Causal Inference* 1, 1, 155–170. DOI: <https://doi.org/10.1515/jci-2013-0003>.
- D. Rubin. 1976. Inference and missing data. *Biometrika* 63, 581–592. DOI: <https://doi.org/10.2307/2335739>.
- I. Shpitser and J. Pearl. 2006. Identification of conditional interventional distributions. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*. 437–444.
- G. Van den Broeck, K. Mohan, A. Choi, A. Darwiche, and J. Pearl. 2015. An efficient method for Bayesian network parameter learning from incomplete data. In *Proceedings of the 31st Conference on Uncertainty in Artificial Intelligence*. 161–170.

A Note of Appreciation

Azaria Paz (The Technion)

As we all know, Judea's book, *Probabilistic Reasoning in Intelligent Systems*, 1988, set the foundations of the groundbreaking theory of "Causality."

Outlined in his *Probabilistic Reasoning* book, I found the following paragraph on page 132:

"The theory of graphoids was conceived in the summer of 1985 when Azaria Paz visited UCLA and he and I began collaborating on the problem of graphical representations."

Indeed, in the summer of 1985 I visited UCLA in order to attend a conference. I met Judea there for the first time and we had several scientific conversations and I was very much impressed by the new way of representing causal models by graphs.

In one of those discussions Judea mentioned to me an open problem. I took that problem as a challenge and after some time I managed to solve it. Judea liked my solution and thus a collaboration started between us that resulted, over the years, in six coauthored journal papers, four conference papers, and two technical reports. In addition, a friendship developed between us and I have been with him both in happy times, when he received the Turing Award and when he received honorary degree from the Technion, his alma mater, and also in tragic times.

I am very thankful to Judea for his friendship and for giving me the opportunity to contribute, even if only a small contribution, to the beautiful and useful theory he developed. As a token of thanks, I constructed two personal talismans for him shown below.

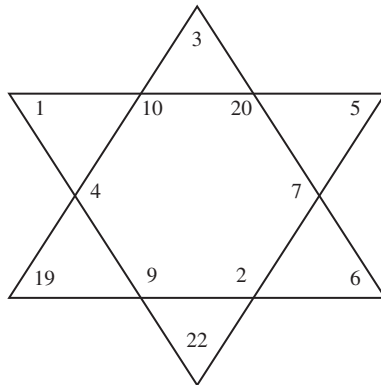
35.1 A Magic Square

9	7	20
23	12	1
4	17	15

Notice the following:

- 4 and 9 in the first column are the day and month of Judea's birth
- $4 + 15 = 19$ in the last row and $19 + 17 = 36$ in the last row together with 19 give 1936 the year when Judea was born.
- All rows, columns, and diagonals sum to 36.
- 7 and 20 in the first row point to the Turing Award since T is the 20th letter in the alphabet and it was awarded 7 years ago.
- Finally, there is a legend in the Jewish folklore claiming that there are 36 righteous men on behalf of which humankind exist. Is Judea one of them?

35.2 A Magic Shield of David



Notice the following:

- The sum of all 4 integers along any of the 6 edges of the two triangles is 36, which together with 19 on the bottom left corner of the upper triangle make 1936, the year when Judea was born.
- 4 and 9 in the left corner of the upper triangles is the date of Judea's birth.
- As in the magic square, 20 and 7 on the right edge of the upper triangle point to the Turing Award.

- d. Finally, 22 at the bottom of the shield is the number of letters in the Hebrew alphabet and 10 in the upper edge of the lower triangle is the number of digits. Together they make 32, which is a very important number in the Kabala, the mystical Jewish canon since it is connected to the word LEV, which means “heart” in Hebrew. And since $32 = 2^5$, it is believed by the Kabala that the world is 5-dimensional: The 3 space dimensions, the time dimension, and the “good and bad” dimension. All five dimensions are infinite in both directions.

Thanks for your friendship Judea, and I wish you many more years of scientific activity and good health.

Causal Models for Dynamical Systems

Jonas Peters (University of Copenhagen),
Stefan Bauer (MPI Tübingen),
Niklas Pfister (University of Copenhagen)

Abstract

A probabilistic model describes a system in its observational state. In many situations, however, we are interested in the system's response under interventions. The class of structural causal models provides a language that allows us to model the behavior under interventions. It can be taken as a starting point to answer a plethora of causal questions, including the identification of causal effects or causal structure learning. In this chapter, we provide a natural and straightforward extension of this concept to dynamical systems, focusing on continuous time models. In particular, we introduce two types of causal kinetic models that differ in how the randomness enters into the model: it may either be considered as observational noise or as systematic driving noise. In both cases, we define interventions and therefore provide a possible starting point for causal inference. In this sense, this chapter provides more questions than answers. The focus of the proposed causal kinetic models lies on the dynamics themselves rather than corresponding stationary distributions, for example. We believe that this is beneficial when the aim is to model the full-time evolution of the system and data are measured at different time points. Under this focus, it is natural to consider interventions in the differential equations themselves.

We wholeheartedly congratulate Judea Pearl on winning the Turing Award. His groundbreaking work has inspired much of our work, with this chapter being only one of several examples.

36.1 Introduction

In causality, we aim to understand how a system reacts under interventions, for example, in gene knock-out experiments. There are different interventions we can

perform (including none at all), and we therefore require different descriptions of the data-generating process. Some systems may be adequately described by deterministic equations, but if the system possesses observational noise, unobserved factors, or intrinsic randomness, data-generating processes are more appropriately modeled using the language of probability. In data-driven sciences, we are used to modeling the data-generating process with a single probability distribution, for example, using a multivariate Gaussian with a certain covariance matrix. As argued above, however, causal models come with a plethora of distributions: one distribution for each type of modeled intervention.

In general, the intervention distributions are not arbitrarily different as it would be meaningless to talk about a single underlying system otherwise. It is a key challenge to describe which parts of the distribution change and which parts remain invariant when considering different interventions. Many researchers from various disciplines engaged in this question and developed the fundamental assumptions that are often referred to as invariance, autonomy, or modularity [Wright 1921, Haavelmo 1944, Aldrich 1989, Hoover 1990, Pearl 2009, Richardson and Robins 2013, Imbens and Rubin 2015]. The concept of invariance relies heavily on what it means to intervene on a system, making a precise formulation of interventions crucial for causal modeling. Arguably one of the clearest formulation of interventions is Judea Pearl's *do*-formalism [Pearl 2009, chapter 36]. One starts with a fixed reference distribution called the observational distribution; one may think of it as describing the system in its natural state with no intervention being performed. The system and its corresponding distribution is assumed to have a modular structure, and performing a *do*-intervention means changing some of the modules. This process yields an intervention distribution, often denoted by a *do*(.) subscript. While this description can be made formal in various ways, we focus on one that is based on structural causal models (SCMs) [Wright 1921, Bollen 1989, Pearl 2009]. Usually, the formulation of SCMs include random variables. We believe, however, that the descriptive power of SCMs lies in their modular structure, which can be separated from randomness. We therefore introduce two different versions of SCMs: a deterministic version with measurement noise and a version containing random variables.

36.1.1 Structural Causal Models with Measurement Noise

A *deterministic SCM* over d variables x^1, \dots, x^d is a collection of d assignments

$$x^k := f^k(x^{\mathbf{PA}_k}), \quad k = 1, \dots, d, \quad (36.1)$$

where for any $k \in \{1, \dots, d\}$, $\mathbf{PA}_k \subseteq \{1, \dots, d\} \setminus \{k\}$ is called the set of direct parents of x^k , and f^k is a real-valued function. If $\mathbf{PA}_k = \emptyset$, then $f^k(x^{\mathbf{PA}_k})$ should be interpreted

as a constant. For each SCM, we obtain a corresponding graphical representation of the causal structure over the vertices¹ $(1, \dots, d)$ by drawing directed edges from \mathbf{PA}_k to k for all $k \in \{1, \dots, d\}$. We further assume that the system (Equation 36.1) is uniquely solvable, which may be the case, even if the graph contains directed cycles, such as $3 \rightarrow 1 \rightarrow 4 \rightarrow 3$. Each SCM then induces a state of the system characterized by a single point in \mathbb{R}^d . We will see in Section 36.1.3 that the modular structure of Equation (36.1) is key to the ability to serve as a causal model. The assignments in Equation (36.1) can be thought of as lines in a computer program that generate a specific state of the system. Interventions will be modeled as replacements of some of these lines.

We may now assume to obtain noisy observations of the system, for example, for each $k \in \{1, \dots, d\}$, we may have

$$X^k := x^k + \varepsilon^k, \quad (36.2)$$

where $\varepsilon^1, \dots, \varepsilon^d$ are jointly independent random variables. Instead of a single point, this model now induces a joint distribution over the observed random variables X^1, \dots, X^d .

36.1.2 Structural Causal Models with Driving Noise

More common than the above approach is the assumption that the randomness enters inside the structural assignments. Formally, a *stochastic SCM* over d random variables X^1, \dots, X^d is a collection of d assignments

$$X^k := f^k(X^{\mathbf{PA}_k}, \varepsilon^k), \quad k = 1, \dots, d, \quad (36.3)$$

together with a distribution over the noise variables $\varepsilon^1, \dots, \varepsilon^d$. As above, we obtain a corresponding graphical representation of the causal structure over the vertices $(1, \dots, d)$ by drawing directed edges from \mathbf{PA}_k to k for all $k \in \{1, \dots, d\}$. We further assume that the joint noise distribution is absolutely continuous with respect to a product measure and that it factorizes, that is, the noise components are assumed to be jointly independent. As before, we require the system (Equation 36.3) to be uniquely solvable, which is always satisfied if the graph is acyclic, for example. An SCM induces a unique joint distribution over the variables X^1, \dots, X^d (e.g., Bongers et al. [2016]), and an observed dataset may be modeled as a collection of independent identically distributed (i.i.d.) realizations from that distribution.

The two approaches described above serve different purposes. The model described in Equation (36.2) might be helpful when the underlying system is assumed to be deterministic and all randomness can be thought of as

1. By slight abuse of notation, we identify (x^1, \dots, x^d) with its indices $(1, \dots, d)$.

measurement noise, for example. While this might be a realistic assumption in many applications, the approach comes with various statistical difficulties, including the famous errors-in-variables problem [Carroll et al. 2006] and an increased difficulty when identifying parameters or causal structure from data [Zhang et al. 2018]. We speculate that this is one of the reasons why less work seems to be devoted to the first approach. The second approach is better understood but assumes that the noise is not purely measurement noise, but enters into the causal mechanism. Whether this assumption is reasonable depends on the application at hand.

36.1.3 Interventions

SCMs allow us to define *interventions*. For any $j \in \{1, \dots, d\}$, we can replace the corresponding assignments in Equation (36.1) or Equation (36.3). In the former case, we could replace the assignment with $X^j := \tilde{f}^j(X^{\text{PA}_j})$ and in the latter case with $X^j := \tilde{f}^j(X^{\text{PA}_j}, \tilde{\varepsilon}^j)$, for example. Usually, we restrict ourselves to interventions that yield a new SCM, so the interventions must respect unique solvability. If that is the case, the intervention induces a new state of the system that we denote by $do(X^j := \tilde{f}^j(x^{\text{PA}_j}))$ or $do(X^j := \tilde{f}^j(X^{\text{PA}_j}, \tilde{\varepsilon}^j))$, respectively. An intervention on one of the variables propagates through the system, possibly affecting many other variables that are graphical descendants of the targeted node. For the stochastic SCMs from Section 36.1.2, one may think about randomized experiments as a *do*-intervention and the well-known hard (or point) interventions $do(X^j := 4)$ appear as a special case. Pearl [2009] provides many insightful examples of SCMs and interventions throughout his book. Bongers et al. [2016] give measure theoretic details underlying the construction of SCMs. Below, we extend the concept of SCMs to dynamical systems and give a concrete example of an SCM, a graph, and interventions in that context (see Figure 36.2).

The above definition clarifies which parts of the distribution remain invariant under interventions. In the case of Section 36.1.2, each conditional distribution X^k , given $X^{\text{PA}_k} = x$, is determined by the structural assignment for X^k . Thus, two distributions induce the same conditionals $X^k | X^{\text{PA}_k} = x$ if one of the distributions is induced by an SCM and the other one is induced by the same SCM after intervening on a fixed $j \neq k$, for example.

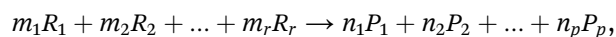
It may further be instructive to think about equivalence of two causal models. They may be called *observationally equivalent* if they induce the same observational distribution and *interventionally equivalent* if they induce the same observational distribution as well as the same intervention distributions (e.g., Peters et al. [2017, section 6.8]). One of the fundamental problems when learning causal structures from data is that two causal models may be observationally equivalent, but not interventionally equivalent.

36.1.4 Time-dependent Data

In many practical applications, an i.i.d. dataset does not provide an adequate description for the data sample at hand. In particular, the concepts above are lacking the notion of time. Different causal methodology and several extensions of SCMs have been proposed [Wiener 1956, Granger 1969, Schreiber 2000, White and Lu 2010, Hyttinen et al. 2013, Peters et al. 2013, Pfister et al. 2018b], mostly considering discrete time models such as vector autoregressive models [Lütkepohl 2007], for example. Peters et al. [2009], and Bauer et al. [2016] discuss the relation between causality and the arrow of time. Causal inference for longitudinal studies has been studied extensively too (e.g., Robins [1997], Aalen et al. [2008], Vanderweele [2015]), where the results are often formulated in the language of potential outcomes [Imbens and Rubin 2015] rather than SCMs. In this article, we focus on continuous time systems that are governed by ODEs. In particular, we propose a natural and straightforward extension of the notion of SCM to dynamical systems. The construction closely follows the existing ideas of SCMs and interventions. Similar constructions have been suggested elsewhere, and we try our best to provide the relevant references and point out existing differences. Parts of this book chapter are taken from Pfister et al. [2018a, 2019], where we focus on model selection and parameter inference.

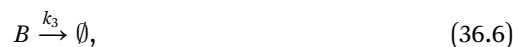
36.2 Chemical Reaction Networks and ODEs

In many natural sciences and even some social sciences, there are processes that can be modeled by a set of governing differential equations. Examples are found in diverse areas such as bioprocessing (e.g., Ogunnaike and Ray [1994]), economics (e.g., Zhang [2005]), genetics (e.g., Chen et al. [1999]), neuroscience (e.g., Friston et al. [2003]), or robotics (e.g., Murray [2017]). Below, we provide two examples that come from a subclass of dynamical models, namely those that are driven by chemical reactions and connect to ODE-based models by mass-action kinetics. The general principles, however, can readily be extended to more complex model classes. Formally, a general reaction (e.g., Wilkinson [2006]) takes the form



where r is the number of reactants and p is the number of products. Both R_i and P_j can be thought of as molecules and are often called species. The coefficients m_i and n_j are positive integers, called stoichiometries. We now provide two examples: (1) a famous and often-used model that describes the abundance of predators and prey, illustrating the law of mass-action kinetics, and (2) Michaelis–Menten kinetics, which results in nonlinear ODEs.

Lotka–Volterra model The Lotka–Volterra model [Lotka 1909] takes the form



where A and B describe abundance of prey and predators, respectively. In this model, the prey reproduce by themselves, but the predators require abundance of prey for reproduction. The coefficients k_1 , k_2 , and k_3 indicate the rates with which the reactions happen.

In mass-action kinetics [Waage and Guldberg 1864], one usually considers the concentration $[X]$ of a species X , the square parentheses indicating that one refers to the concentration rather than to the integer number of abundant species or molecules. The law of mass-action states that the instantaneous rate of each reaction is proportional to the product of each of its reactants raised to the power of its stoichiometry. For the Lotka–Volterra model this yields

$$\frac{d}{dt}[A] = k_1[A] - k_2[A][B] \quad (36.7)$$

$$\frac{d}{dt}[B] = k_2[A][B] - k_3[B]. \quad (36.8)$$

Figure 36.1 shows solutions for these differential equations for both an observational setting (left plot) with rates $k_1 = 0.1$, $k_2 = 0.05$, and $k_3 = 0.05$ and initial values $[A]_0 = 1$ and $[B]_0 = 1.5$, as well as an interventional setting (right plot) where we set $k_1 = 0.05$ and $[B]_0 = 2$. Even though Equations (36.7) and (36.8) contain

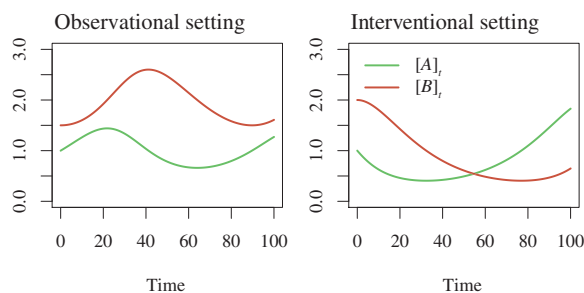
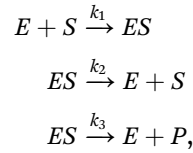


Figure 36.1 Example trajectories for the basic Lotka–Volterra model given in Equations (36.7) and (36.8). The left plot corresponds to observation setting with rates $k_1 = 0.1$, $k_2 = 0.05$, and $k_3 = 0.05$ and initial values $[A]_0 = 1$ and $[B]_0 = 1.5$, and the right plot to an intervention where we set $k_1 = 0.05$ and $[B]_0 = 2$.

interaction terms of the concentration of the different species, they are linear in the model parameters, a property that is exploited by many practical methods.

Michaelis–Menten kinetics In Michaelis–Menten kinetics [Michaelis and Menten 1913], the starting point is a specific enzyme reaction given by the equations



where the enzyme E binds to a substrate S and finally releases a product P . Under some simplifying assumptions regarding the relation of rates of the reactions, this yields the equation

$$\frac{d}{dt}[P] = c_1 \frac{[S]}{c_2 + [S]}, \quad (36.9)$$

where c_1, c_2 are constants. There are many reactions that can be described by this model; Michaelis and Menten [1913] used it to describe how the enzyme invertase catalyzes the hydrolysis of sucrose into glucose and fructose.

36.3 Causal Kinetic Models

We now define a causal model class for dynamical systems. The reader may think about the example of a Lotka–Volterra model, as described in Equations (36.7) and (36.8), or Michaelis–Menten kinetics (36.9), both of which fit into the general framework described below. In analogy to Sections 36.1.1 and 36.1.2, we first consider a deterministic version with measurement noise and secondly a version where the randomness enters inside the structural equations.

36.3.1 Causal Kinetic Models with Measurement Noise

We regard the following definition as a natural and straightforward extension of SCMs, even though we have not seen it in this form before. A *deterministic causal kinetic model* over processes $\mathbf{x} := (\mathbf{x}_t)_t := (x_t^1, \dots, x_t^d)_t$ is a collection of d ODEs and initial value assignments

$$\frac{d}{dt}x_t^1 := f^1(x_t^{\text{PA}_1}), \quad x_0^1 := \zeta_0^1,$$

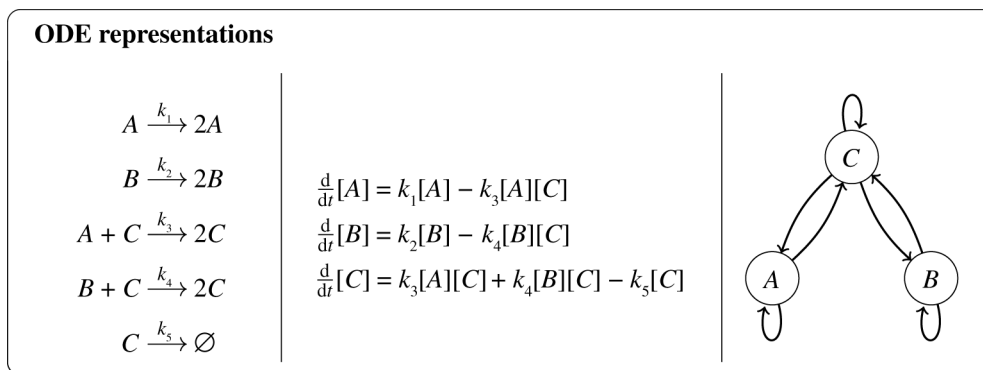


Figure 36.2 Illustration of different ODE representations: (chemical) reactions (left), ODE system derived by mass-action kinetics (middle), and corresponding graph (right).

$$\begin{aligned} \frac{d}{dt}x_t^2 &:= f^2(x_t^{\mathbf{PA}_2}), & x_0^2 &:= \xi_0^2, \\ &\vdots \\ \frac{d}{dt}x_t^d &:= f^d(x_t^{\mathbf{PA}_d}), & x_0^d &:= \xi_0^d. \end{aligned}$$

Here, for any $k \in \{1, \dots, d\}$, $\frac{d}{dt}x_t^k$ denotes the time derivative of the component x^k at time t and $\mathbf{PA}_k \subseteq \{1, \dots, d\}$ is called the set of direct parents of x^k (and may include x^k itself). We require that the system of initial value problems is uniquely solvable. For each causal kinetic model, we can obtain a corresponding graph over the vertices $\{1, \dots, d\}$ by drawing edges from \mathbf{PA}_k to k , for $k \in \{1, \dots, d\}$ (see Figure 36.2). If we consider the initial values as random variables, this induces a distribution over $\mathbf{x} = (\mathbf{x}_t)_t$.

Similarly, as in the case of deterministic SCMs, causal kinetic models are deterministic models describing an underlying causal structure. The observed data can then be modeled as noisy observations of the system, that is,²

$$\mathbf{X}_t = \mathbf{x}_t + \varepsilon_t, \tag{36.10}$$

where one may assume for simplicity that each noise component of ε_t is i.i.d., for example. This induces a distribution over $\mathbf{X} = (\mathbf{X}_t)_t$.

36.3.2 Causal Kinetic Models with Driving Noise

As for SCMs, the randomness might also be added directly into the structural assignments. This yields a more involved mathematical formulation though, since

2. Alternatively, one may add the noise variables only at observed time points.

the objects of interest are continuous time processes. We define a *stochastic causal kinetic model* over processes $\mathbf{X} := (\mathbf{X}_t)_t := (X_t^1, \dots, X_t^d)_t$ as a collection of d stochastic differential equations (SDEs) and initial value assignments

$$dX_t^k := f^k(X_t^{\text{PA}_k})dt + h^k(X_t^{\text{PA}_k})dW_t^k, \quad X_0^k := \xi_0^k, \quad (36.11)$$

where dW_t^k can be thought of as an independent white noise process and $W_t^k = \int_0^t dW_s^k$ as a Brownian motion.³ Again, we require that the SDEs in Equation (36.11) are uniquely solvable, which in the setting of SDEs becomes substantially harder to verify. The functions f^k are called drift coefficients and the functions h^k are called diffusion coefficients. Intuitively, it can be helpful to think about the change $X_{t+\Delta}^k - X_t^k$ as being normally distributed with expectation $f^k(X_t^{\text{PA}_k}) \cdot \Delta$ and variance $h^k(X_t^{\text{PA}_k})^2 \cdot \Delta$, where Δ is a small increment in time. In the most basic setting, h^k can be assumed to be constant, which results in an integrated equation of the form

$$X_t^k := \int_0^t f^k(X_s^{\text{PA}_k})ds + W_t^k.$$

In general, solving SDEs is a difficult problem, and numerical procedures often have slower rates compared with their deterministic counterparts [Han and Kloeden 2017]. We believe that despite these difficulties, SDE-based causal models may potentially prove useful in several areas of applications. There are some works that have made first attempts to circumvent the difficulties of models using SDEs by looking at random differential equations [Bauer et al. 2017, Bongers and Mooij 2018, Abbati et al. 2019], which still allow including randomness directly into the causal structure. As for SCMs, it depends on the application whether a causal model with measurement noise or the stochastic causal kinetic model is the more appropriate choice.

36.3.3 Interventions

An intervention on the system replaces some of the structural assignments. Interventions can change the dynamics of the process x^k , the initial values or both at the same time. This definition allows for several ways of manipulating the system, which may prove useful when modeling complex dynamical systems and their perturbations; some of the possibilities are discussed below. Formally, for a deterministic causal kinetic model over a process $(\mathbf{x}_t)_t$, an *intervention* on the process x^k

3. Readers who are not familiar with the formal definition of SDEs may think about them as a notational abbreviation for the integrated form, that is, $X_t^k := \int_0^t f^k(X_s^{\text{PA}_k})ds + \int_0^t h^k(X_s^{\text{PA}_k})dW_s^k$.

for $k \in \{1, \dots, d\}$ corresponds to replacing the k -th initial condition or the k -th ODE with

$$x_0^k := \xi \quad \text{or} \quad \frac{d}{dt}x_t^k := g(x_t^{\mathbf{PA}}),$$

respectively, where $\mathbf{PA} \subseteq \{1, \dots, d\}$ is the set of new parent components. In both cases, we still require that the system of initial value problems is uniquely solvable. The interventions are denoted by

$$do(x_0^k := \xi) \quad \text{and} \quad do\left(\frac{d}{dt}x_t^k := g(x_t^{\mathbf{PA}})\right),$$

respectively. The same definitions apply in the presence of observational noise ε_t , see Equation (36.10), where the noise is added after the system has been perturbed. For a stochastic causal kinetic model, we analogously define the interventions

$$do(x_0^k := \xi) \quad \text{and} \quad do(dX_t^k := g(X_t^{\mathbf{PA}}) + j(X_t^{\mathbf{PA}})dW_t^k).$$

While we regard both deterministic and stochastic causal kinetic models as potentially relevant for practical applications, we will, in the remainder of this chapter, focus on deterministic causal models.

If the ODE system is induced by a set of reactions, a natural class of interventions is described by replacing one (or some) of the reactions. In the Lotka–Volterra model from Section 36.2, changing the rate of the first reaction, (Equation 36.4), that is, changing k_1 to \tilde{k}_1 , say, yields a change of the assignment (Equation 36.7). Changing the rate of the second reaction (Equation 36.5), however, yields a change of both assignments, Equation (36.7) and Equation (36.8). In general, changing one of the reactions induces a change in differential equations for all variables that appear in the reaction (either on the left or on the right). The proposed framework additionally allows us to set a variable x^k to a constant value c by performing the interventions $do(x_0^k := c)$ and $do(\frac{d}{dt}x_t^k := 0)$. To obtain a softer version of this effect, we may also introduce a forcing term that “pulls” the variable x^k to a certain value c . Alternatively, one can keep the dependence of $\frac{d}{dt}x^k$ on x^ℓ intact but change the strength of this dependence, or even completely change the parent set.

We believe that in a system that is described well by a system of differential equations, it is most natural to formulate the interventions as differential equations too. Nevertheless, for a differentiable ζ , interventions of the form $x_t^k := \zeta(x_t^A)$ with $A \subseteq \{1, \dots, d\} \setminus \{k\}$ (e.g., Hansen and Sokol [2014]) and $x_t^k := \zeta(t)$ (e.g., Rubenstein et al. [2018]) are included in the above formalism as well. The intervention $do(x_t^k := \zeta(x_t^A))$ can be obtained by $do(\frac{d}{dt}x_t^k := \frac{d}{dt}\zeta(x_t^A))$ and $do(x_0^k := \zeta(x_0^A))$. Similarly, $do(x_t^k := \zeta(t))$ is realized by $do(\frac{d}{dt}x_t^k := \frac{d}{dt}\zeta(t))$ and $do(x_0^k := \zeta(0))$.

36.3.4 Other Causal Models for Dynamical Systems and Related Work

We introduced the formal framework of causal kinetic models that allows us to model dynamical systems with a set of differential equations and specify what we mean by intervening in the system. Several other useful proposals have been made that connect differential equations with causality. Here, we briefly review some of these suggestions and point out a few of the differences. In general, the attempts are tailored toward different goals.

[Mooij et al. \[2013\]](#), [Blom and Mooij \[2018\]](#), [Bongers and Mooij \[2018\]](#), and [Rubenstein et al. \[2018\]](#) consider (deterministic and random) ODEs. Their goal is to describe the asymptotic solution of such a system as a causal model. The authors consider interventions that fix the full-time trajectory of a variable to a pre-defined solution, for example, to a constant. [Mooij et al. \[2013\]](#) consider interventions on the ODE system itself. In that work, the authors are primarily interested in the equilibrium of the ODE system (assuming that it exists) and its relation to standard SCMs; they explicitly do not distinguish between interventions that yield the same equilibrium. These approaches may be particularly useful when the focus lies on the stationary solution, rather than the full dynamics. [Hansen and Sokol \[2014\]](#) consider SDEs, which contain ODEs as a special case, and introduce interventions, for which at any point in time the intervened variable can be written as a deterministic function of other variables. [Christiansen et al. \[2020\]](#) consider causal models for spatio-temporal data.

In practice, the application at hand determines which of the models and interventions are most appropriate for describing the real-world experiment. The structure of causal kinetic models closely follows the spirit of the SCMs described above. In particular, its modular structure once more highlights which parts remain invariant under interventions.

36.4 Challenges in Causal Inference for ODE-based Systems

Formalizing a causal model for dynamical systems can be taken as a starting point to conduct causal inference. Similarly to the i.i.d. case, we might be interested in adjustment results, do-calculus, the effect of hidden variables, or causal discovery (see, e.g., [Pearl \[2009\]](#)). To the best of our knowledge, for dynamical systems most of such questions are still open. Possible reasons are the difficulties that arise when working with dynamical systems, some of which we highlight below. (1) In the deterministic settings, solving a standard algebraic equation is easier than solving an algebraic equation involving differentials. (2) When adding observational noise, the induced distributions on the left-hand side of the structural assignments are more complicated in causal kinetic systems than in SCMs.

(3) Suppose that in the i.i.d. case (Equation 36.3) the noise variables are additive. If the parents of each variable (and therefore the structure of the whole system) are known, the causal mechanisms, that is, the functions f^k , can be estimated by nonlinear regression techniques. In contrast, in the case of dynamical systems the fitting process is much more involved, and many different methods have been suggested. This includes various versions of goodness-of-fit of the integrated system, nonlinear least squares methods, or gradient matching [Bard 1974, Benson 1979, Varah 1982, Ramsay et al. 2007, Calderhead et al. 2009, Oates et al. 2014, Dattner and Klaassen 2015, Macdonald and Husmeier 2015, Raue et al. 2015, Wenk et al. 2019]. (4) In the i.i.d. setting, Markov conditions connect properties of the graph, such as d -separation [Pearl 2009], with properties of the joint distribution, such as conditional independence [Lauritzen 1996]. For dynamical models, however, it is not apparent that conditional independence is the right notion. For specific model classes, there is interesting work exploiting the concept of local independence [Schweder 1970, Didelez 2000, 2008, Mogensen et al. 2018, Mogensen and Hansen 2019], with several questions still being open. Finally, (5), in most real-world systems not all relevant variables are observed, which means that they need to be modeled as hidden variables. While in the i.i.d. case there is some understanding of the effects of hidden variables on observed distributions, on the identification of causal effects, and on causal discovery (e.g., Verma and Pearl [1991], Spirtes et al. [2000], Richardson and Spirtes [2002], Hernán and Robins [2006], Silva et al. [2006], Pearl [2009], Hyttinen et al. [2012], Evans [2015], Richardson et al. [2017], Shpitser [2018]), more work is needed in the case of dynamical systems.

36.5 From Invariance to Causality and Generalizability

In many real-world systems the underlying structure is unknown and needs to be inferred from data. That is, for any k , we do not know which variables are contained in \mathbf{PA}_k . This setting is often referred to as structure learning or causal discovery [Spirtes et al. 2000, Pearl 2009]. To state the problem let us assume that the observed data consist of n repetitions of discrete time observations of each of the d variables \mathbf{x} , or its noisy version $\tilde{\mathbf{X}}$, on the time grid $\mathbf{t} = (t_1, \dots, t_L)$. By concatenating the time series for the d variables, one may represent the data by an $n \times (d \cdot L)$ matrix. Several methods have been suggested to solve this task (e.g., Oates et al. [2014], Raue et al. [2015], Mikkelsen and Hansen [2017]), most of which combine structure learning, that is, model selection, with a parameter inference step. Some methods [Oates et al. 2014] explicitly consider the causal nature of this problem. We briefly describe below, in a simplified setting, how it is possible to exploit the invariances induced by the underlying causal kinetic model for causal discovery. Assume there is a target process $y := x^1$, for which the parents are unknown and

of particular interest. In short, we assume that each of the n repetitions has been generated by a model of the form

$$\frac{d}{dt}y_t = f^y(\mathbf{x}_t^{\mathbf{PA}_y}), \quad (36.12)$$

for a fixed function f^y , possibly with additional measurement noise $\tilde{Y}_t = y_t + \varepsilon_t$. This assumption holds, for example, if the measurements stem from an underlying causal kinetic model under different interventional settings, in none of which the variable y has been intervened on. In practice, the right-hand side of Equation (36.12) is unknown, and the goal is thus to identify the causal predictors among the \mathbf{x} , that is, to infer both the parents \mathbf{PA}_y of y as well as the function f^y . In Pfister et al. [2018a] we propose a procedure that specifically exploits the invariance of Equation (36.12) to tackle the problem of structure learning. Each of the repetitions is assumed to be part of an environment or experimental condition. We suppose this assignment is known, for example, repetitions 1, ..., 6 are known to belong to experimental condition one, repetitions 7, ..., 19 to condition two, and all remaining repetitions to condition three. The method then outputs a ranking of models (or variables) by trading off the predictability and invariance of such models. In the i.i.d. case, trade-offs in a similar spirit have been suggested by Rojas-Carulla et al. [2018], Magliacane et al. [2018], and Rothenhäusler et al. [2021], for example.

The model in Equation (36.12) is valid independently of interventions on variables other than y and can thus be used for prediction in a new experimental setup, even if there are large perturbations on the predictors \mathbf{x} . As a consequence, the method proposed in Pfister et al. [2018a] outputs models that generalize better to unseen experiments, even when considering real data from large metabolic network experiments. This finding adds to a recent debate suggesting the addition of invariance as a fitting criteria to data science methodology [Schölkopf et al. 2012, Yu 2013, Bareinboim and Pearl 2016, Meinshausen et al. 2016, Peters et al. 2016, Yu and Kumbier 2019]. At its core lies the modularity of the structure of the causal model and its implied relation between causality and invariance.

36.6 Conclusions

We have discussed an extension of SCMs to systems that are governed by differential equations. As in the i.i.d. case, the models may be equipped with either measurement noise or driving noise, where the latter case uses the concept of SDEs. These two model classes, called causal kinetic models, may serve as a starting point for answering questions commonly asked in the field of causal inference.

Many of such questions are neither fully understood nor answered, and more work is needed to gather as much understanding as we have for i.i.d. data.

The mathematical complexity of the models poses a challenge when working with kinetic models. Some aspects of causal inference, however, may become easier. The concept of faithfulness suggests, for example, that in the i.i.d. setting a child of a random variable is predictive for its parent. This assumption seems less justified in the case of dynamical processes. Also, considering local independence and assuming causal sufficiency, Markov equivalence classes contain only a single directed acyclic graph [Mogensen and Hansen 2019]. Both of these points may prove to be useful for causal discovery. Furthermore, intervening on a set of differential equations usually affects the whole time trajectory. Relatively mild interventions may thus carry a lot of information about the causal structure. This may be particularly relevant when the available data are not yet sufficient to identify causal mechanisms, and additional data have to be collected. There is a close connection between experimentation and causal inference (e.g., Imai et al. [2013], Peters et al. [2016]); the selection of measurement readouts, time points, or intervention strategies could guide experimentation and has the potential to significantly reduce the number of complicated and expensive experiments.

While there are several differences to the i.i.d. case, causal kinetic models exhibit the same modularity as SCMs. As a consequence, invariance ideas can be exploited in a similar way as it is done in the i.i.d. case. This includes methods that trade off invariance and predictability to select models that may generalize better to unseen experiments.

Acknowledgments

The authors thank Søren Wengel Mogensen, Nikolaj Theodor Thams and Niels Richard Hansen for helpful discussions. JP was supported by a research grant (18968) from VILLUM FONDEN and Carlsberg Foundation.

References

- O. Aalen, O. Borgan, and H. Gjessing. 2008. *Survival and Event History Analysis: A Process Point of View*. Springer, New York. DOI: <https://doi.org/10.1007/978-0-387-68560-1>.
- G. Abbati, P. Wenk, S. Bauer, M. A. Osborne, A. Krause, and B. Schölkopf. 2019. ARS and MaRS—adversarial and MMD-minimizing regression for SDEs. In *Proceedings of the 29th International Conference on Machine Learning (ICML)*. DOI: <https://doi.org/10.3929/ETHZ-B-000385687>.
- J. Aldrich. 1989. Autonomy. *Oxf. Econ. Pap.* 41, 15–34. DOI: <https://doi.org/10.1093/oxfordjournals.oep.a041889>.
- Y. Bard. 1974. *Nonlinear Parameter Estimation*. Academic Press, New York.

- E. Bareinboim and J. Pearl. 2016. Causal inference and the data-fusion problem. *Proc. Natl. Acad. Sci.* 113, 27, 7345–7352. DOI: <https://doi.org/10.1073/pnas.1510507113>.
- S. Bauer, B. Schölkopf, and J. Peters. 2016. The arrow of time in multivariate time series. In *Proceedings of the 26th International Conference on Machine Learning (ICML)*. 2043–2051.
- S. Bauer, N. Gorbach, D. Miladinovic, and J. M. Buhmann. 2017. Efficient and flexible inference for stochastic systems. In *Advances in Neural Information Processing Systems (NIPS)*. 6988–6998. DOI: <https://doi.org/10.3929/ethz-b-000261734>.
- M. Benson. 1979. Parameter fitting in dynamic models. *Ecol. Model.* 6, 97–115. DOI: [https://doi.org/10.1016/0304-3800\(79\)90029-2](https://doi.org/10.1016/0304-3800(79)90029-2).
- T. Blom and J. M. Mooij. 2018. Generalized structural causal models. *ArXiv e-prints (1805.06539)*.
- K. A. Bollen. 1989. *Structural Equations with Latent Variables*. John Wiley & Sons, New York. DOI: <https://doi.org/10.1002/9781118619179>.
- S. Bongers and J. M. Mooij. 2018. From random differential equations to structural causal models: The stochastic case. *ArXiv e-prints (1803.08784)*.
- S. Bongers, P. Forré, J. Peters, and J. M. Mooij. 2016. Foundations of structural causal models with cycles and latent variables. *Ann. Stat. ArXiv e-prints (1611.06221v5)*.
- B. Calderhead, M. Girolami, and N. D. Lawrence. 2009. Accelerating Bayesian inference over nonlinear differential equations with Gaussian processes. In *Advances in Neural Information Processing Systems (NIPS)*. 217–224.
- R. J. Carroll, D. Ruppert, L. A. Stefanski, and C. M. Crainiceanu. 2006. *Measurement Error in Nonlinear Models: A Modern Perspective* (2nd. ed.). Chapman and Hall/CRC, Boca Raton, FL. DOI: <https://doi.org/10.1201/9781420010138>.
- T. Chen, H. L. He, and G. Church. 1999. Modeling gene expression with differential equations. In *Biocomputing'99*, World Scientific, 29–40. DOI: https://doi.org/10.1142/9789814447300_0004.
- R. Christiansen, M. Baumann, T. Kuemmerle, M. Mahecha, J. Peters. 2020. Towards causal inference for spatio-temporal data: conflict and forest loss in colombia. *ArXiv e-prints (2005.08639)*.
- I. Dattner and C. A. J. Klaassen. 2015. Optimal rate of direct estimators in systems of ordinary differential equations linear in functions of the parameters. *Electron. J. Stat.* 9, 2, 1939–1973. DOI: <https://doi.org/10.1214/15-EJS1053>.
- V. Didelez. 2000. *Graphical Models for Event History Analysis Based on Local Independence*. PhD thesis. Universität Dortmund.
- V. Didelez. 2008. Graphical models for marked point processes based on local independence. *J. R. Stat. Soc. Ser. B* 70, 1, 245–264. DOI: <https://doi.org/10.1111/j.1467-9868.2007.00634.x>.
- R. J. Evans. 2015. Margins of discrete Bayesian networks. *Ann. Stat.* 46, 6A, 2623–2656. DOI: <https://doi.org/https://doi.org/10.1214/17-AOS1631>.
- K. J. Friston, L. Harrison, and W. Penny. 2003. Dynamic causal modelling. *Neuroimage* 19, 4, 1273–1302. DOI: [https://doi.org/10.1016/S1053-8119\(03\)00202-7](https://doi.org/10.1016/S1053-8119(03)00202-7).

- C. W. J. Granger. 1969. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica* 37, 3, 424–438. DOI: <https://doi.org/10.2307/1912791>.
- T. Haavelmo. 1944. The probability approach in econometrics. *Econometrica* 12, supplement, S1–S115. DOI: <https://doi.org/10.2307/1906935>.
- X. Han and P. E. Kloeden. 2017. *Random Ordinary Differential Equations and Their Numerical Solution*. Springer. DOI: <https://doi.org/10.1007/978-981-10-6265-0>.
- N. R. Hansen and A. Sokol. 2014. Causal interpretation of stochastic differential equations. *Electro. J. Probab.* 19, 100, 1–24. DOI: <https://doi.org/10.1214/EJP.v19-2891>.
- M. A. Hernán and J. M. Robins. 2006. Instruments for causal inference: An epidemiologist's dream? *Epidemiology* 17, 360–372. DOI: <https://doi.org/10.1097/01.ede.0000222409.00878.37>.
- K. D. Hoover. 1990. The logic of causal inference. *Econ. Philos.* 6, 207–234. DOI: <https://doi.org/10.1017/S026626710000122X>.
- A. Hyttinen, F. Eberhardt, and P. O. Hoyer. 2012. Learning linear cyclic causal models with latent variables. *J. Mach. Learn. Res.* 13, 1, 3387–3439.
- A. Hyttinen, P. O. Hoyer, F. Eberhardt, and M. Järvisalo. 2013. Discovering cyclic causal models with latent variables: A general SAT-based procedure. In *Proceedings of the 29th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*. 301–310.
- K. Imai, D. Tingley, and T. Yamamoto. 2013. Experimental designs for identifying causal mechanisms. *J. R. Stat. Soc. Ser. A (Statistics in Society)* 176, 1, 5–51. DOI: <https://doi.org/10.1111/j.1467-985X.2012.01032.x>.
- G. W. Imbens and D. B. Rubin. 2015. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, New York. DOI: <https://doi.org/10.1017/CBO9781139025751>.
- S. Lauritzen. 1996. *Graphical Models*. Oxford University Press, New York. DOI: [https://doi.org/10.1002/\(SICI\)1097-0258\(19991115\)18:21<2983::AID-SIM198>3.0.CO;2-A](https://doi.org/10.1002/(SICI)1097-0258(19991115)18:21<2983::AID-SIM198>3.0.CO;2-A).
- A. J. Lotka. 1909. Contribution to the theory of periodic reactions. *J. Phys. Chem.* 14, 3, 271–274. DOI: <http://dx.doi.org/10.1021/j150111a004>.
- H. Lütkepohl. 2007. *New Introduction to Multiple Time Series Analysis*. Springer, Berlin, Germany. DOI: <http://dx.doi.org/10.1007/978-3-540-27752-1>.
- B. Macdonald and D. Husmeier. 2015. Gradient matching methods for computational inference in mechanistic models for systems biology: A review and comparative analysis. *Front. Bioeng. Biotechnol.* 3, 180. DOI: <http://dx.doi.org/10.3389/fbioe.2015.00180>.
- S. Magliacane, T. van Ommen, T. Claassen, S. Bongers, P. Versteeg, and J. M. Mooij. 2018. Domain adaptation by using causal inference to predict invariant conditional distributions. In *Advances in Neural Information Processing Systems 31*, 10846–10856.
- N. Meinshausen, A. Hauser, J. Mooij, J. Peters, P. Versteeg, and P. Bühlmann. 2016. Methods for causal inference from gene perturbation experiments and validation. *Proc. Natl. Acad. Sci.* 113, 27, 7361–7368. DOI: <http://dx.doi.org/10.1073/pnas.1510493113>.
- L. Michaelis and M. L. Menten. 1913. Die Kinetik der Invertinwirkung. *Biochem Z.* 49, 333–369. Translation available at <https://pubs.acs.org/doi/suppl/10.1021/bi201284u>.

- F. V. Mikkelsen and N. R. Hansen. 2017. Learning large scale ordinary differential equation systems. *ArXiv e-prints (1710.09308)*.
- S. W. Mogensen and N. R. Hansen. 2020. Markov equivalence of marginalized local independence graphs. *Ann. Stat.* 48, 1, 539–559. DOI: <https://doi.org/10.1214/19-AOS1821>.
- S. W. Mogensen, D. Malinsky, and N. R. Hansen. 2018. Causal learning for partially observed stochastic dynamical systems. In *Proceedings of the Thirty-Fourth Conference on Uncertainty in Artificial Intelligence (UAI)*. 350–360.
- J. M. Mooij, D. Janzing, and B. Schölkopf. 2013. From ordinary differential equations to structural causal models: The deterministic case. In *Proceedings of the 29th Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI)*. AUAI Press, Corvallis, Oregon, 440–448.
- R. Murray. 2017. *A Mathematical Introduction to Robotic Manipulation*. CRC Press. DOI: <https://doi.org/10.1201/9781315136370>.
- C. J. Oates, F. Dondelinger, N. Bayani, J. Korkola, J. W. Gray, and S. Mukherjee. 2014. Causal network inference using biochemical kinetics. *Bioinformatics* 30, 17, i468–i474. DOI: <http://dx.doi.org/10.1093/bioinformatics/btu452>.
- B. Ogunnaike and W. Ray. 1994. *Process Dynamics, Modeling, and Control*, Vol. 1. Oxford University Press, New York. DOI: <https://doi.org/10.1002/aic.690440523>.
- J. Pearl. 2009. *Causality: Models, Reasoning, and Inference*. (2nd. ed.). Cambridge University Press, New York. DOI: <http://dx.doi.org/10.1017/CBO9780511803161>.
- J. Peters, D. Janzing, A. Gretton, and B. Schölkopf. 2009. Detecting the direction of causal time series. In *Proceedings of the 26th International Conference on Machine Learning (ICML)*. ACM Press, 801–808. DOI: <https://doi.org/10.1145/1553374.1553477>.
- J. Peters, D. Janzing, and B. Schölkopf. 2013. Causal inference on time series using structural equation models. In *Advances in Neural Information Processing Systems 26 (NIPS)*. Curran Associates, Inc.
- J. Peters, P. Bühlmann, and N. Meinshausen. 2016. Causal inference using invariant prediction: Identification and confidence intervals. *J. R. Stat. Soc. Series B.* 78, 5, 947–1012. DOI: <https://doi.org/10.1111/rssb.121670>.
- J. Peters, D. Janzing, and B. Schölkopf. 2017. *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press, Cambridge, MA.
- N. Pfister, S. Bauer, and J. Peters. 2018a. Identifying causal structure in large-scale kinetic systems. *ArXiv e-prints (1810.11776)*.
- N. Pfister, P. Bühlmann, and J. Peters. 2018b. Invariant causal prediction for sequential data. *J. Am. Stat. Assoc.* 114, 1264–1276. DOI: <https://doi.org/10.1080/01621459.2018.1491403>.
- N. Pfister, S. Bauer, and J. Peters. 2019. Learning stable structures in kinetic systems: Benefits of a causal approach. *Proc. Natl. Acad. Sci. U.S.A.* 116, 25405–25411. DOI: <https://doi.org/10.1073/pnas.1905688116>.
- J. O. Ramsay, G. Hooker, D. Campbell, and J. Cao. 2007. Parameter estimation for differential equations: A generalized smoothing approach. *J. R. Stat. Soc. Series B* 69, 5, 741–796. DOI: <https://doi.org/10.1111/j.1467-9868.2007.00610.x>.

- A. Raue, B. Steiert, M. Schelker, C. Kreutz, T. Maiwald, H. Hass, J. Vanlier, C. Tönsing, L. Adlung, R. Engesser, W. Mader, T. Heinemann, J. Hasenauer, M. Schilling, T. Höfer, E. Klipp, F. Theis, U. Klingmüller, B. Schöberl, and J. Timmer. 2015. Data2dynamics: A modeling environment tailored to parameter estimation in dynamical systems. *Bioinformatics* 31, 21, 3558–3560. DOI: <https://doi.org/10.1093/bioinformatics/btv405>.
- T. Richardson and P. Spirtes. 2002. Ancestral graph Markov models. *Ann. Stat.* 30, 4, 962–1030. DOI: <https://doi.org/10.1214/aos/1031689015>.
- T. Richardson and J. M. Robins. 2013. Single world intervention graphs (SWIGs): A unification of the counterfactual and graphical approaches to causality. *Center for the Statistics and the Social Sciences, University of Washington Series. Working Paper 128, 30 April 2013*.
- T. Richardson, R. J. Evans, J. M. Robins, and I. Shpitser. 2017. Nested Markov properties for acyclic directed mixed graphs. *ArXiv e-prints (1701.06686)*.
- J. M. Robins. 1997. Causal inference from complex longitudinal data. In M. Berkane (Ed.), *Latent Variable Modeling and Applications to Causality*. Springer, New York, 69–117. DOI: https://doi.org/10.1007/978-1-4612-1842-5_4.
- M. Rojas-Carulla, B. Schölkopf, R. Turner, and J. Peters. 2018. Causal transfer in machine learning. *J. Mach. Learn. Res.* 19, 36, 1–34.
- D. Rothenhäusler, P. Bühlmann, N. Meinshausen, and J. Peters. 2021. Anchor regression: Heterogeneous data meets causality. *J. R. Stat. Soc. Series B.* 83, 215–246. DOI: <https://doi.org/10.1111/rssb.12398>.
- P. Rubenstein, S. Bongers, J. M. Mooij, and B. Schölkopf. 2018. From deterministic ODEs to dynamic structural causal models. In *Proceedings of the 34th Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI)*.
- B. Schölkopf, D. Janzing, J. Peters, E. Sgouritsa, K. Zhang, and J. M. Mooij. 2012. On causal and anticausal learning. In *Proceedings of the 29th International Conference on Machine Learning (ICML)*.
- T. Schreiber. 2000. Measuring information transfer. *Phys. Rev. Lett.* 85, 461–464. DOI: <https://doi.org/10.1103/PhysRevLett.85.461>.
- T. Schweder. 1970. Composable Markov processes. *J. Appl. Probab.* 7, 400–410. DOI: <https://doi.org/10.2307/3211973>.
- I. Shpitser. 2018. Identification in graphical causal models. In M. Maathuis, M. Drton, S. L. Lauritzen, and M. Wainwright (Eds.), *Handbook of Graphical Models*. CRC Press. DOI: <https://doi.org/10.1201/9780429463976>.
- R. Silva, R. Scheines, C. Glymour, and P. Spirtes. 2006. Learning the structure of linear latent variable models. *J. Mach. Learn. Res.* 7, 191–246.
- P. Spirtes, C. Glymour, and R. Scheines. 2000. *Causation, Prediction, and Search*. (2nd. ed.). MIT Press. DOI: <https://doi.org/10.1007/978-1-4612-2748-9>.
- T. J. Vanderweele. 2015. *Explanation in Causal Inference: Methods for Mediation and Interaction*. Oxford University Press, New York.

- J. M. Varah. 1982. A spline least squares method for numerical parameter estimation in differential equations. *SIAM J. Sci. Stat. Comput.* 3, 1, 28–46. DOI: <https://doi.org/10.1137/0903003>.
- T. Verma and J. Pearl. 1991. Equivalence and synthesis of causal models. In *Proceedings of the 6th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*. 255–270.
- P. Waage and C. M. Guldberg. 1864. Studier over affiniteten (in Danish). *Forhandlinger i Videnskabs-selskabet i Christiania*. 35–45.
- P. Wenk, A. Gotovos, S. Bauer, N. Gorbach, A. Krause, and J. M. Buhmann. 2019. Fast Gaussian process based gradient matching for parameter identification in systems of nonlinear ODEs. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS)*.
- H. White and X. Lu. 2010. Granger causality and dynamic structural systems. *J. Financ. Econom.* 8, 2, 193–243. DOI: <https://doi.org/10.1093/jjfinec/nbq006>.
- N. Wiener. 1956. The theory of prediction. In E. Beckenbach (Ed.), *Modern Mathematics for Engineers*. McGraw-Hill, New York.
- D. J. Wilkinson. 2006. *Stochastic Modelling for Systems Biology*. Chapman and Hall/CRC Mathematical and Computational Biology Series. Chapman & Hall/CRC.
- S. Wright. 1921. Correlation and causation. *J. Agric. Res.* 20, 557–585.
- B. Yu. 2013. Stability. *Bernoulli* 19, 4, 1484–1500. DOI: <https://doi.org/10.3150/13-BEJSP14>.
- B. Yu and K. Kumbier. 2019. Three principles of data science: Predictability, computability, and stability (pcs). *ArXiv e-prints (1901.08152)*.
- W.-B. Zhang. 2005. *Differential Equations, Bifurcations, and Chaos in Economics*, Vol. 68. World Scientific Publishing Company. DOI: <https://doi.org/10.1142/5827>.
- K. Zhang, M. Gong, J. Ramsey, K. Batmanghelich, P. Spirtes, and C. Glymour. 2018. Causal discovery with linear non-Gaussian models under measurement error: Structural identifiability results. In *Proceedings of the 34th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*. AUAI Press.

Probabilistic Programming Languages: Independent Choices and Deterministic Systems

David Poole (University of British Columbia),
Frank Wood (University of British Columbia)

Pearl [2000, p. 26] attributes to Laplace [1814] the idea of a probabilistic model as a deterministic system with stochastic inputs. Pearl defines causal models in terms of deterministic systems with stochastic inputs. In this chapter, we show how deterministic systems with (independent) probabilistic inputs are also the basis of modern probabilistic programming languages [van de Meent et al. 2018]. Probabilistic programs can be seen as consisting of independent choices (over which there are probability distributions) and deterministic programs that give the consequences of these choices. The work on developing such languages has gone in parallel with the development of causal models, and many of the foundations are remarkably similar. Most of the work in probabilistic programming languages has been in the context of specific languages. This chapter abstracts the work on probabilistic programming languages from specific languages and explains some design choices in the design of these languages.

Probabilistic programming languages owe their beginnings to the development of simulation languages such as Simula [Dahl and Nygaard 1966]. Simula was designed for discrete event simulations, and the built-in random number generator allowed for stochastic simulations. Probabilistic programming languages, as opposed to simulation languages, introduce one critical additional language syntactic feature and are interpreted entirely differently:

Conditioning: the ability to indicate, syntactically, that some variable values are observed

Inference: interpreting a probabilistic program means computing the posterior distribution of arbitrary variables conditioned on these observations. For the discrete case, the semantics can be seen in terms of rejection sampling: accept only the simulations that produce the observed values, but there are other semantics that have also been developed.

First, we explain how we can get from discrete Bayesian networks [Pearl 1988] to independent choices plus a deterministic system (by augmenting the set of variables). This can then be extended to allow for Turing-complete deterministic systems, which results in discrete probabilistic programming languages. We consider languages with continuous random variables later on.

Consider how to represent a Bayesian network in terms of a deterministic system with independent inputs. In essence, we construct a random variable for each free parameter of the original model. A deterministic system can be used to obtain the original variables from the new variables. In this new augmented model, there are two possible worlds structures: one in terms of the new independent random variables, and one effectively in the space of the original random variables. The dimensionality of the augmented space is the number of free parameters which is greater than the dimensionality of the original space (unless all variables were independent). However, the variables in the augmented worlds can be assumed to be independent, which is convenient.

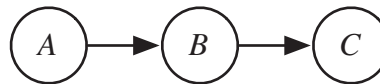
The original worlds can be obtained using abduction. Abduction is a form of reasoning characterized by “reasoning to the best explanation.” It is typically characterized by finding a minimal consistent set of assumables that imply some observation. Poole [1991, 1993b] gave an abductive characterization of a discrete probabilistic programming language, which gave a mapping between the independent possible world structure, and the descriptions of the worlds produced by abduction. Given an observation and query variables, an explanation is an assignment of values to a subset of the probabilistic choices that implies the observations and a value for the query variable. The set of all observations forms a sigma algebra that has the properties needed to compute conditional probabilities. This works even in cases where the deterministic system is a Turing machine, in which case there can be infinitely many possible worlds, as long as the algorithm halts with probability one. With continuous variables, the measure over the real variables and the measure over the paths induced by the program interact in complex ways [Staton et al. 2016, Heunen et al. 2017, Staton 2017, Scibior et al. 2018]. Here we will outline the measure theory induced by programming languages just for the discrete case.

There had been parallel developments in the development of causality [Pearl 2000], with causal models being deterministic systems with stochastic inputs. The augmented variables in the probabilistic programming languages are the variables needed for counterfactual reasoning.

37.1 Probabilistic Models and Deterministic Systems

In order to understand probabilistic programming languages, it is instructive to see how a discrete probabilistic model in terms of a Bayesian network (belief network) [Pearl 1988] can be represented as a deterministic system with probabilistic inputs.

Consider the following simple belief network, with Boolean random variables:



There are five free parameters to be assigned for this model; for concreteness assume the following values (where $A = \text{true}$ is written as a , and $A = \text{false}$ is written as $\neg a$, and similarly for the other variables):

$$P(a) = 0.1$$

$$P(b|a) = 0.8$$

$$P(b|\neg a) = 0.3$$

$$P(c|b) = 0.4$$

$$P(c|\neg b) = 0.75$$

To represent such a belief network in a probabilistic programming language, there are probabilistic inputs corresponding to the free parameters, and the programming language specifies what follows from them. For example, in Simula syntax [Dahl and Nygaard 1966], this could be represented as:

```

begin
  Boolean a, b, c;
  a := draw (0.1);
  if a then
    b := draw (0.8);
  else
    b := draw (0.3);
  if b then
    c := draw (0.4);

```

```

else
  c := draw (0.75);
end

```

where $draw(p)$ is a Simula system predicate that returns true with probability p ; each time it is called, there is an independent draw.

Probabilistic programming languages also have mechanisms for specifying observations, which Simula did not have. Suppose c was observed, and the query is for the posterior probability of b . The conditional probability $P(b|c)$ is the proportion of those runs with c true that also have b true. This could be computed using the Simula evaluator by doing rejection sampling: running the program many times, and rejecting those runs that do not assign c to true. Out of the non-rejected runs, it would return the proportion that have b true. Of course, conditioning does not need to be implemented that way; much of the development of probabilistic programming languages over the last 30 years is in devising more efficient ways to implement conditioning.

An equivalent model to the Simula program can be given in terms of logic. There can be five random variables, corresponding to the five independent draws, let's call them A , $Bifa$, $Bifna$, $Cifb$, $Cifnb$. These are independent with $P(a) = 0.1$, $P(bifa) = 0.8$, $P(bifna) = 0.3$, $P(cifb) = 0.4$, and $P(cifnb) = 0.75$. The other variables can be defined in terms of these:

$$b \Leftrightarrow (a \wedge bifa) \vee (\neg a \wedge bifna) \quad (37.1)$$

$$c \Leftrightarrow (b \wedge cifb) \vee (\neg b \wedge cifnb) \quad (37.2)$$

where \wedge means “and,” \vee means “or,” \neg means “not,” and \Leftrightarrow means “if and only if.”

These two formulations are essentially the same; they differ in how the deterministic system is specified, whether it is in Simula or in logic.

Any discrete belief network can be represented as a deterministic system with independent inputs. This was proven by [Poole \[1991, 1993b\]](#) and [Druzdzel and Simon \[1993\]](#). These papers used different languages for the deterministic systems but gave essentially the same construction.

37.2 Possible Worlds Semantics

A probabilistic programming language needs a specification of a deterministic system (given in some programming language) and a way to specify distributions over (independent) probabilistic inputs, or a syntactic variant of this. We will also assume that there are some observations and some query variables for which we want the posterior probability.

In developing the semantics of a probabilistic programming language, we first define the set of possible worlds, and then a probability measure over sets of possible worlds [Halpern 2003]. We first consider the case of discrete variables.

In probabilistic programming, there are (at least) two sets of possible worlds that interact semantically. It is easiest to see these in terms of the above example. In the above belief network, there were three random variables A , B , and C , which had complex inter-dependencies among them. With three binary random variables, there are eight possible worlds. These eight possible worlds give a concise characterization of the probability distribution over these variables.

In the corresponding probabilistic program, there is an augmented space with five inputs, each of which can be considered a random variable (these are A , $Bifa$, $Bifna$, $Cifb$, and $Cifnb$ in the logic representation). With five binary random variables, there are 32 possible worlds. The reason to increase the number of variables, and thus possible worlds, is that in this *augmented* space the random variables can be independent.

Note that the variables in the augmented space do not *have* to be independent. For example, $P(bifna | a)$ can be assigned arbitrarily since, when a is true, no other variable depends on $bifna$. In the augmented space, there is enough freedom to make the variables independent. Thus, we can arbitrarily set $P(bifna|a) = P(bifna|\neg a)$, which will be the same as $P(b|\neg a)$. The independence assumption makes the semantics and some computation simpler, for example, in marginalizing A we can construct a joint that includes $bifa$ and $bifna$.

There are three semantics that could be given to a probabilistic program:

- The rejection sampling semantics: running the program with a random number generator, removing those runs that do not predict the observations; the posterior probability of a proposition is the limit, as the number of runs increases, of the proportion of the non-rejected runs that have the proposition true.
- The independent choice semantics, where a possible world specifies the outcome of all possible draws. Each of these draws is considered to be independent. Given a world, the (deterministic) program would specify what follows. We will call these the augmented worlds. In this semantics, a possible world would select values for all five of the input variables in the example above, and thus gives rise to the augmented space of the above program with 32 worlds. When a program does not have a bounded runtime there can be unaccountably infinite many possible worlds (e.g., if someone keeps playing the lottery until they win, and we want to compute the number of times they play), and we need to define a measure over them.

- The abductive semantics forms a measure over the set of possible worlds where the possible worlds that produce the same value (or have the same proof or derivation for the value) for variables of interest (observed or queried variables) are grouped together. For example, in this semantics, an explanation would specify the values for three of the draws in the program of the previous section, as only three draws are encountered in any run of the program. We will refer to these grouping as concise worlds; they specify the value of a subset of the variables, such that the values of the other variables are irrelevant.

In the logical definition of the belief network (or in the Simula definition if the draws are named), there are 32 worlds in the independent choice semantics:

World	<i>A</i>	<i>Bifa</i>	<i>Bifna</i>	<i>Cifb</i>	<i>Cifnb</i>	Probability
w_0	false	false	false	false	false	$0.9 \times 0.2 \times 0.7 \times 0.6 \times 0.25$
w_1	false	false	false	false	true	$0.9 \times 0.2 \times 0.7 \times 0.6 \times 0.75$
...						
w_{30}	true	true	true	true	false	$0.1 \times 0.8 \times 0.3 \times 0.4 \times 0.75$
w_{31}	true	true	true	true	true	$0.1 \times 0.8 \times 0.3 \times 0.4 \times 0.75$

The probability of each world is the product of the probability of each variable (as each of these variables is assumed to be independent). Note that in worlds w_{30} and w_{31} , the original variables *A*, *B*, and *C* are all true; the value of *Cifnb* is not used when *B* is true. These variables are also all true in the worlds that only differ in the value of *Bifna*, as, again, *Bifna* is not used when *A* is true.

In the abductive semantics, when *C* is observed or queried there are eight concise worlds for this example, because some of the augmented variables don't participate in the program; the program acts the same no matter what the values for these variables are.

World	<i>A</i>	<i>Bifa</i>	<i>Bifna</i>	<i>Cifb</i>	<i>Cifnb</i>	Probability
w_0	false	\perp	false	\perp	false	$0.9 \times 0.7 \times 0.25$
w_1	false	\perp	false	\perp	true	$0.9 \times 0.7 \times 0.75$
...						
w_7	true	True	\perp	false	\perp	$0.1 \times 0.8 \times 0.6$
w_8	true	True	\perp	true	\perp	$0.1 \times 0.8 \times 0.4$

where \perp means the variable is not defined in this concise world. These worlds cover all eight cases of truth values for the original worlds that give values for *A*, *B*, and

C . The values of A , B , and C can be obtained from the program. The idea is that a run of the program is never going to encounter an undefined value; in particular, the Simula program will not actually make these draws. The augmented worlds can be obtained from the concise worlds by splitting the concise worlds on each value of the undefined variables. Thus, each concise world corresponds to a set of augmented worlds, where the distinctions that are ignored do not make a difference in any inference.

If the observation was $C = \text{true}$, and the query was B , a (minimal) explanation is a (minimal) set of assignments of values to the independent choices that gives $C = \text{true} \wedge B = \text{true}$ or $C = \text{true} \wedge B = \text{false}$. There are four such explanations:

- $A = \text{true}, Bifa = \text{true}, Cifb = \text{true}$
- $A = \text{true}, Bifa = \text{false}, Cifnb = \text{true}$
- $A = \text{false}, Bifna = \text{true}, Cifb = \text{true}$
- $A = \text{false}, Bifna = \text{false}, Cifnb = \text{true}$

The probability of each of these explanations is the product of the choices made, as these choices are independent. Now suppose the C is observed to be true. This observation is evidence that is conditioned on. The posterior probability $P(B|C = \text{true})$ can be computed by the weighted sum of the explanations in which B is true. Note that the same explanations would be used even if C has unobserved descendants.

While it may seem that we have not made any progress, after all this is just a simple Bayesian network, we can do the same thing for any program with discrete probabilistic inputs. We just need to define the independent inputs (often these are called *noise inputs*), and a deterministic program that gives the consequences of the choices of values for these inputs. It is reasonably easy to see that any belief network can be represented in this way, where the number of independent inputs is equal to the number of free parameters of the belief network. However, we are not restricted to belief networks. The programs can be arbitrarily complex. We also do not need special “original variables,” but can define the augmented worlds with respect to any variables of interest. Observations can be any proposition, and queries (about which we want the posterior probability) can be any variables.

There can be uncountably many augmented worlds when the language is Turing-equivalent, but only countably many concise worlds for programs that eventually halt. A typical assumption is that the program eventually infers the observations and the query, that is, each run of the program will eventually (with probability 1) assign a value to any given observation and query. Consider again the person who plays the lottery until they win; each augmented world would consist

of an infinite sequence of lottery outcomes, and so there are uncountably infinitely many of these. There are, however, only countable many concise worlds (one for each integer n , where the person wins after n steps).

In an analogous way to how the probability of a real-variable is defined as a limit of discretizations, we can see the abductive characterization as providing a measure over the independent choice worlds. In terms of the Simula program, explanations correspond to execution paths, in particular, the set of outcomes of the draws in one trace of the program. In the logical formulation (e.g., in a probabilistic logic program), an explanation is a minimal set of assignments to the augmented variables that implies the observations and a value for the query variable. The probability of an explanation is the product of the probabilities of the choices in the explanation, and is the measure of the set of extended worlds that are a superset of the explanation.

Let's now consider a larger non-trivial program. Figure 37.1 shows code for a simple probabilistic context-free grammar. The procedure `gen_exp` returns a number or a list, such as 1, [" + ", 1, 2] or [" + ", [" + ", 1, 2], [" + ", 2, 2]]. Suppose that `eval` takes such a list and evaluates it. The second to last statement of Figure 37.1 specifies that that expression v is observed to evaluate to 7. (as, for example, does the list [" + ", [" + ", 1, 2], [" + ", 2, 2]]). The program represents the distribution over the returned value (v) conditioned on the observation ($eval(v) = 7$).

```
define gen_exp()
begin
  Boolean is_dig, digit
  is_dig := draw(0.7);
  if is_dig then
    digit := draw(0.4);
    if digit
      return 1;
    else
      return 2;
  else
    return ["+", gen_exp(), gen_exp()];
end
v = gen_exp();
observe eval(v) = 7;
return v;
```

Figure 37.1 Pseudo-Simula program for a simple probabilistic context-free grammar.

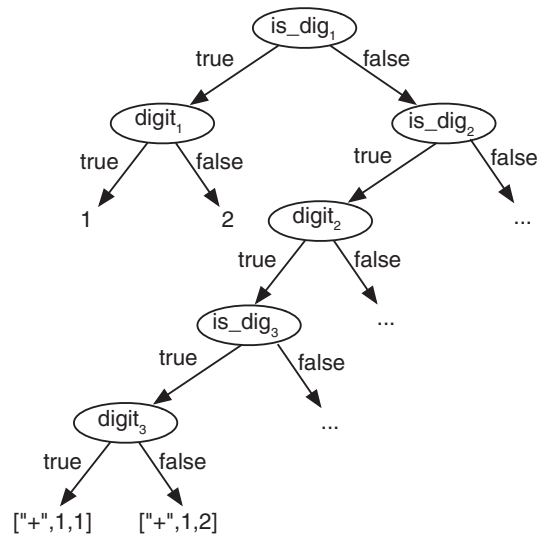


Figure 37.2 Part of the space of choices for *gen_exp()*.

Figure 37.2 shows a tree of choices for this program. Each path in the tree shows one set of choices from the *gen_exp()* program. Given each sequence of choices down a branch, the program is deterministic. Note that the choices are numbered arbitrarily (but in an order they could be encountered by executing the program); the program needs to make various independent choices at different points in the program. The tree is infinite. In the independent choice semantics, a world specifies the values for each of the nodes in this tree. There are countably infinitely many nodes, and so uncountably many worlds. In the abductive semantics, there is a world for each path. Note that, although the existence of a draw in the abductive semantics depends on the previous draws, in the independent choice semantics the draws are still probabilistically independent. Thus, the draws down a path can be multiplied.

The prior probability of any abductive world (and so the measure of the set of augmented worlds compatible with these choices) is the product of the choices made on the path to that world. For example, the probability of generating a 1 is $0.7 * 0.4$ and the probability of producing a 2 is $0.7 * 0.6$ and the probability of producing $["+", 1, 2]$ is $0.3 * 0.7 * 0.4 * 0.7 * 0.6$.

It turns out that for every $\epsilon > 0$ there is a finite subtree such that the probability of paths that lead to abductive worlds (and so have halted) is greater than $1 - \epsilon$. Note that this would not be true if the probability of *is_dig* were smaller. This is not a problem with the probabilistic program per se, but that the grammar

has a positive probability of generating infinite sentences (and a program cannot produce infinite outputs in finite time).

With continuous variables, the semantics is more complicated as the measure theory induced by real variables interacts with the measure theory induced by the control flow of the probabilistic program. In particular, the measure theory of a real-values variable can be defined in terms of binary “<” splits that need to be interleaved with the tree defined above [Kozen 1979, Borgström et al. 2011].

37.3 Inference

The aim of inference is to compute the distribution of query variables conditioned on observations. There are two cases that can be treated separately:

- Discrete variables can be represented as propositions or their characteristic functions (sometimes called the “one-hot encoding”), where a variable X can represent $X = v_i$ with the variable x_i which has value 1 when $X = v_i$ and 0 otherwise. In this case, the expected value of x_i corresponds to the probability of $X = v_i$. We can then use any method for computing expected values to compute probabilities.
- For a continuous variable, the expected value does not characterize the distribution, and there would need to be infinitely many characteristic functions (such as $X < v$ for every value v). While there is a closed form for some distributions (such as the Gaussian or the Dirichlet distributions), in general the best we can do is to draw samples from the distribution. Indeed, it is possible that the probabilistic program is the most concise specification of a distribution.

Another issue that arises with continuous variables is that when a real value is observed, and it is pretended to be of infinite precision, the probability of the observation is zero. In general, we need to be able to specify the precision of observations. The precision is typically more complicated than “to three significant digits,” and we need to model how sensors work.

Earlier algorithms for discrete variables (e.g., Poole 1993a) extract the minimal explanations and compute conditional probabilities from these. Later algorithms, such as those used in Integrated Bayesian Agent Language (IBAL) [Pfeffer 2001], use sophisticated variable elimination to carry out inference in this space. IBAL’s computation graph corresponds to a graphical representation of the choice space of Figure 37.2 (but with shared structure). The original ProbLog [De Raedt et al. 2007] compiled the computation graph into binary decision diagrams (BDDs). More modern approaches [Fierens et al. 2015] compile to first-order weighted model

counting using better data structures for the computation graphs, exploiting more conditional independence, shared structure, and determinism, and allow for lifted inference (automatically determining when individual cases need to be reasoned about separately, and when we can compute a probability analytically by counting).

In algorithms that exploit the conditional independent structure, like variable elimination or recursive conditioning, the order that variables are summed out or split can make a big difference to efficiency. In the independent choice semantics, there are more options available for summing out variables, thus there are more options available for making inference efficient. For example, consider the following fragment of a Simula program:

```
begin
  Boolean x;
  x := draw (0.1);
  if x then
    begin
      Boolean y := draw (0.2);
      ...
    end
  else
    begin
      Boolean z := draw (0.7);
      ...
    end
  ...
end
```

Here y is only defined when x is true and z is only defined when x is false. In the abductive semantics, y and z are never both defined in any world. In the independent choice semantics, y and z are defined in all worlds. Efficiency considerations may mean that we want to sum out X first. In the independent choice semantics, there is no problem, the joint probability on X and Y makes perfect sense. However, in the abductive semantics, it isn't clear what the joint probability of X and Y means. In order to allow for flexible elimination orderings in variable elimination or splitting ordering in recursive conditioning, the independent choice semantics is a natural choice, but we need to avoid unnecessary splitting of (sets of) worlds.

Languages that allow worlds with continuous-valued random variables have inspired the development of approximate inference algorithms that become correct only asymptotically with continued computation. These algorithms are de facto necessary because the partitions of continuous variables' domains driven by

the guard functions to conditional branching (if) statements can become arbitrarily small. This leads to onerous program evaluation requirements like tracking the partitioning of each domain into subregions. Further, in most practical instances, it also leads to an exponential proliferation of discrete random variables that indicate tiny subregions of each real valued variable's domain. This makes variable elimination and other exact methods impractical.

A powerful class of *general purpose* inference algorithms for probabilistic programming [Milch et al. 2005, Goodman et al. 2008, McCallum et al. 2009, Nori et al. 2014, Ritchie et al. 2016] derive from Markov chain Monte Carlo (MCMC) statistical inference algorithms [Neal 1993, Gelman et al. 2013]. MCMC probabilistic programming inference methods remain compatible with the discrete program semantics presented above but also work when the probabilistic programming language is endowed with continuous random variables.

It is possible to do MCMC in either of the spaces of worlds above. The difference arises in the way conditionally present variables are treated. In the augmented space, for the example above, an MCMC state would include values for all of X , Y , and Z . In the abductive semantics, it would contain values for X and Y when $X = \text{true}$, and values for X and Z when $X = \text{false}$, as Y and Z are never simultaneously defined. Suppose in an MCMC step X 's value changes from *true* to *false*. In the augmented space it would just use the remembered values for Z . In the program-trace semantics, Z was not defined when Z was true; thus changing X from *true* to *false* means re-sampling all of the variables defined in that branch, including Z .

Original implementations of different probabilistic programming language evaluators used different MCMC strategies: BLOG [Milch et al. 2005] and Church [Goodman et al. 2008] assigned values to all of the variables in the augmented space and used Gibbs sampling and trace-based MCMC, respectively. FACTORIE [McCallum et al. 2009] worked in what we have called the abductive space. Anglican [Tolpin et al. 2016] and WebPPL [Goodman and Stuhlmüller 2014] use a hybrid “random database” approach [Wingate et al. 2011], reusing the values of random variables when possible while lazily pruning the store of retained sample values. Which of these is more efficient remains an empirical question with a program-dependent answer.

van de Meent et al. [2018] provide a comprehensive overview of general purpose inference algorithms for probabilistic programming languages that also include continuous choices. Scibior et al. [2018] go some ways beyond this by establishing a compositional semantics for inference algorithms themselves, providing the theoretical basis for sound implementations of new and existing inference algorithms.

37.4 Learning

There are many definitions of learning, but one appealing one is that we want the distribution of hypotheses given data. Bayes [1763] showed how to compute the probability that the posterior probability (conditioning on data) of some event is in some range; he used the beta distribution but the idea is more general. In a probabilistic program, we condition on a dataset—a dataset is the observations—and determine the distribution over the parameters of interest or use the program to then make predictions for other cases. In this way learning is just inference, and, as inference is the function of probabilistic programming languages, learning comes for free.

While this reductionist view is accurate, it is often computationally advantageous to define learning in terms of computing a point estimate of the values of some subset of the parameters of a probabilistic program. The algorithmic workhorse of such learning is the Expectation Maximization (EM) algorithm [Dempster et al. 1977], as well as related marginal maximum a posteriori and variational inference algorithms [Wainwright and Jordan 2008].

Learning parameter values using EM in probabilistic programming languages has a long history and was described by Sato [1995] and Koller and Pfeffer [1997] for learning parameters (probabilities in conditional probability tables) in discrete models. EM also forms the basis for learning in Prism [Sato and Kameya 1997, 2001], IBAL [Pfeffer 2001, 2007], and many subsequent languages. Modern deep probabilistic programming systems [Tran et al. 2016, 2017, Uber 2018] can be purposed to use variational methods to learn model parameters in an EM-like way. Referencing the context-free grammar example above, the production probabilities 0.7 and 0.4 could be replaced with variables whose values could be learned from data (for instance, an independent identically distributed [i.i.d.] dataset of single observations 7, 7, 9, 8, and 9). EM, which must integrate out the latent abstract syntax trees for each observation, would adjust the 0.7 and 0.4 values to be those that maximize the probability of producing the observed values.

Structure learning in probabilistic programming languages was historically explored in the context of logic programs, where the techniques of inductive logic programming can be applied. De Raedt et al. [2008] overviews early research in this area. With the advent of universal (or Turing-complete) probabilistic programming languages, this has been complimented by a spectrum of work on program induction [Hwang et al. 2011, Lake et al. 2015, Perov and Wood 2016, Gulwani et al. 2017]. As before, what counts as learning versus what counts as inference is somewhat in the eye of the beholder. Universal languages make it possible and usually straightforward to specify a higher-level prior on the structure of the model itself. This means that the tools of automatic inference can be employed to learn model

structure as well as perform more traditional latent variable inference directly from observational data. In the context-free grammar example from above, one might imagine a structural change to the program that would evaluate whether the `is_dig = false` branch allows for more than one type of binary mathematical operation and in what proportion. Such a construct could use another “structure-level” random choice to choose between a program with one or some other number of binary operators and their meanings. Whether point estimates or distributional outcomes for such structure learning are of interest is a problem and developer-specific consideration.

37.5 Other Issues

In addition to inference and learning, the design and implementation of probabilistic programming languages and systems give rise to a number of other fascinating and sometimes complex issues.

One issue that arises is how to handle probabilistic programs that have not yet halted. Suppose, for example, that there is a Boolean query variable, and we can determine that its value is false in 10% of the runs, true in 20% of the runs, but the algorithm has not halted for the other 70% of the runs. We know the probability will eventually be between 0.2 and 0.9. In general, it is plausible that the samples that can be derived quickly might be different than the samples that require much computation; it is not reasonable to assume that they are a random sample. We also need to consider what a reasonable answer is if we know that some of the probability mass will not halt (e.g., if the rest of the probability mass is just in a never-ending loop). As another example, if the other 70% of the probability mass consists of searching for proofs and counter-examples to a difficult problem (such as $P = NP$), it might not be reasonable to infer any probability, and the halted probability mass might be unrelated to the part that has not halted. [Poole \[1993a\]](#) computed upper and lower bounds of the posterior probability of a query based on parts of the programs that have halted.

As is evident in this chapter and inherent to the tensions between statistics and computer science, early work in computer science on probabilistic programming languages focused on exact probability computation in discrete variable models. In statistics, approximate inference in mixed continuous and discrete variable models was and remains the focus [[Spiegelhalter et al. 1995](#), [Stan Development Team 2014](#)]. One might ask in this light, is it not silly to compute pretending we have real continuous values when digital computers use inherently discrete representations? Furthermore, isn't it silly to pretend as if a measured value has infinite precision rather than being, maximally, a statement that a value lives in some precision

range? Answering these questions forces us to consider issues that range from the practical to deep mathematical semantics. Programming languages have nearly always been endowed with continuous-valued variables (but typically to finite precision). It is *convenient* to use such types. Unfortunately, unlike the possible worlds discrete semantics discussed earlier in this chapter that are only applicable to discrete random-variable languages, the very mathematical meaning of higher-order probabilistic programs with continuous random variables was only very recently established and in fact required new measure-theoretic mathematical foundations [Staton et al. 2016, Heunen et al. 2017, Staton 2017, Scibior et al. 2018].

37.6 Causal Models

It is interesting that the research on causal modeling and probabilistic programming languages have gone on in parallel, with similar foundations, but only recently have researchers started to combine them by adding causal constructs to probabilistic programming languages [Finzi and Lukasiewicz 2003, Baral and Hunsaker 2007, Vennekens et al. 2009].

In some sense, the programming languages can be seen as representations for all of the counterfactual situations. A programming language gives a model when some condition is true, but also defines the “else” part of a condition: what happens when the condition is false.

In the future, we expect that programming languages will be the preferred way to specify causal models, and for interventions and counterfactual reasoning to become part of the repertoire of probabilistic programming languages.

37.7 Some Pivotal References

Probabilistic Horn abduction [Poole 1991, 1993b] was the first language with a probabilistic semantics that allowed for conditioning. The initial parts of this chapter were presented there in the context of logic programs. Probabilistic Horn abduction was refined into the Independent Choice Logic [Poole 1997] that allowed for choices made by multiple agents, where there is a clean integration with negation as failure [Poole 2000]. Prism introduced learning into essentially the same framework [Sato and Kameya 1997, 2001]. More recently, ProbLog [De Raedt et al. 2007] has become a focus to implement many logical languages into a common framework.

In parallel to the work on probabilistic logic programming languages has been work on developing probabilistic functional programming languages starting with Stochastic Lisp [Koller et al. 1997], including IBAL [Pfeffer 2001, 2007], A-Lisp [Andre

and Russell 2002], Church [Goodman et al. 2008], Venture [Mansinghka et al. 2014], and Anglican [Tolpin et al. 2016].

Other probabilistic programming languages are based on more imperative languages such as CES [Thrun 2000] and probabilistic-C [Paige and Wood 2014], both based on C, Turing [Ge et al. 2018] based on Julia, and the languages BLOG [Milch et al. 2005] and FACTORIE [McCallum et al. 2009] based on object-oriented languages. BLOG concentrates on number and identity uncertainty, where the probabilistic inputs include the number of objects and whether two names refer to the same object or not. Modern production languages such as STAN [Stan Development Team 2014] and related research languages [Zhou et al. 2019] also use imperative syntactic constructs.

The most modern probabilistic programming systems use imperative syntax and are implemented as domain-specific embedded languages within host languages that include automatic differentiation capabilities. These languages allow interesting marriages to form between deep learning and probabilistic programming techniques. Besides research languages like PyProb [Le et al. 2017] and ProbTorch [Siddharth et al. 2017], modern industrial giants have invested heavily as well in languages like Pyro [Uber 2018] and Tensorflow Probability/Edward [Tran et al. 2016].

The distinctions discussed here in terms of syntactic variation are more carefully picked apart in terms of language capability by van de Meent et al. [2018]. Unlike this work which touches on delineations of languages that allow discrete and continuous random variables, they posit a sharp transition between languages that allow a finite number of random variables and those that do not. In the finite case, direct compilation correspondences are established to graphical models. In the infinite case, general purpose inference algorithms are explained, up to and including the variational techniques employed in the most modern systems.

37.8 Conclusion

This chapter has concentrated on the foundational similarities, rather than the differences, between the probabilistic programming languages.

There has been considerable research in probabilistic programming over the last few years, with multiple languages developed and a better understanding of the principles underlying any probabilistic programming language [van de Meent et al. 2018]. There have also been considerable advances in both exact and approximate inference algorithms to the point where they are now routinely used for real-world problems. Probabilistic programming languages have an exciting future.

References

- D. Andre and S. Russell. 2002. State abstraction for programmable reinforcement learning agents. In *Proceedings of the 18th National Conference on Artificial Intelligence (IJCAI)*.
- C. Baral and M. Hunsaker. 2007. Using the probabilistic logic programming language P-log for causal and counterfactual reasoning and non-naive conditioning. In *Proceedings of the International Conference of Scalable Uncertainty Management*. 243–249.
- T. Bayes. 1763. An essay towards solving a problem in the doctrine of chances. *Philos. Trans. R. Soc. Lond.* 53, 370–418. DOI: <https://doi.org/10.1098/rstl.1763.0053>.
- J. Borgström, A. D. Gordon, M. Greenberg, J. Margetson, and J. van Gael. 2011. Measure transformer semantics for Bayesian machine learning. In *European Symposium on Programming*, Springer, 77–96. DOI: https://doi.org/10.1007/978-3-642-19718-5_5.
- O.-J. Dahl and K. Nygaard. 1966. Simula: An ALGOL-based simulation language. *Commun. ACM* 9, 9, 671–678. DOI: <https://doi.org/10.1145/365813.365819>.
- L. De Raedt, A. Kimmig, and H. Toivonen. 2007. ProbLog: A probabilistic Prolog and its application in link discovery. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI)*. 2462–2467.
- L. De Raedt, P. Frasconi, K. Kersting, and S. H. Muggleton (Eds.). 2008. *Probabilistic Inductive Logic Programming*. Springer. DOI: https://doi.org/10.1007/978-3-540-78652-8_1.
- A. P. Dempster, N. M. Laird, and D. B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Series B Methodol.* 39, 1, 1–22. DOI: <https://doi.org/10.1111/j.2517-6161.1977.tb01600.x>.
- M. J. Druzdzel and H. A. Simon. 1993. Causality in Bayesian belief networks. In *Proceedings of the 9th Conference on Uncertainty in Artificial Intelligence (UAI-93)*. 3–11. DOI: <https://doi.org/10.1016/B978-1-4832-1451-1.50005-6>.
- D. Fierens, G. van den Broeck, J. Renkens, D. Shterionov, B. Gutmann, I. Thon, G. Janssens, and L. De Raedt. 2015. Inference and learning in probabilistic logic programs using weighted Boolean formulas. *Theor. Prac. Log. Prog.* 15, 4, 358–401. DOI: <https://doi.org/10.1017/S1471068414000076>.
- A. Finzi and T. Lukasiewicz. 2003. Structure-based causes and explanations in the independent choice logic. In *Proceedings of the 19th Conference on Uncertainty in Artificial Intelligence (UAI 2003)*. 225–232.
- H. Ge, K. Xu, and Z. Ghahramani. 2018. Turing: A language for flexible probabilistic inference. In *Proceedings of the 21st International Conference on Artificial Intelligence and Statistics, PMLR 84*, 1682–1690.
- A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. 2013. *Bayesian Data Analysis* (3rd ed.). Chapman Hall, Boca Raton, FL.
- W. R. Gilks, A. Thomas, and D. J. Spiegelhalter. 1994. A language and program for complex Bayesian modelling. *J. R. Stat. Soc. Series D: Stat.* 43, 1, 169–177. DOI: <https://doi.org/10.2307/2348941>.
- N. D. Goodman and A. Stuhlmüller. 2014. The Design and Implementation of Probabilistic Programming Languages. <http://dippl.org>. Accessed:2017-8-22.

- N. Goodman, V. Mansinghka, D. M. Roy, K. Bonawitz, and J. Tenenbaum. 2008. Church: A language for generative models. In *Proceedings of the Uncertainty in Artificial Intelligence (UAI)*. <https://arxiv.org/abs/1206.3255>.
- S. Gulwani, O. Polozov, R. Singh. 2017. Program synthesis. *Found. Trends Prog. Lang.* 4, 1–2, 1–119. DOI: <https://doi.org/10.1561/25000000010>.
- J. Y. Halpern. 2003. *Reasoning about Uncertainty*. MIT Press, Cambridge, MA. DOI: <https://doi.org/10.7551/mitpress/10951.001.0001>.
- C. Heunen, O. Kammar, S. Staton, and H. Yang. 2017. A convenient category for higher-order probability theory. In *32nd Annual ACM/IEEE Symposium on Logic in Computer Science, LICS 2017*. Reykjavik, Iceland, June 20–23, 2017, 1–12. DOI: <https://doi.org/10.1109/LICS.2017.8005137>.
- I. Hwang, A. Stuhlmüller, and N. D. Goodman. 2011. Inducing probabilistic programs by Bayesian program merging. arXiv:1110.5667.
- D. Koller and A. Pfeffer. 1997. Learning probabilities for noisy first-order rules. In *Proceedings of the 15th International Joint Conference on Artificial Intelligence (IJCAI)*. Nagoya, Japan, 1316–1321.
- D. Koller, D. McAllester, and A. Pfeffer. 1997. Effective Bayesian inference for stochastic programs. In *Proceedings of the 14th National Conference on Artificial Intelligence (AAAI)*. 740–747.
- D. Kozen. 1979. Semantics of probabilistic programs. In *20th Annual Symposium on Foundations of Computer Science (sfcs 1979)*. IEEE. 101–114. DOI: <https://doi.org/10.1109/SFCS.1979.38>.
- B. M. Lake, R. Salakhutdinov, and J. B. Tenenbaum. 2015. Human-level concept learning through probabilistic program induction. *Science* 350, 6266, 1332–1338. DOI: <https://doi.org/10.1126/science.aab3050>.
- P. S. Laplace. 1814. *Essai philosophique sur les probabilités*. Courcier. Reprinted (1812) in English, F. W. Truscott and F. L. Emory (Trans.). Wiley, New York.
- T. A. Le, A. G. Baydin, and F. Wood. 2017. Inference compilation and universal probabilistic programming. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, Vol. 54. *Proceedings of Machine Learning Research*. PMLR, Fort Lauderdale, FL, 1338–1348.
- V. Mansinghka, D. Selsam, and Y. Perov. 2014. Venture: A higher-order probabilistic programming platform with programmable inference. *arXiv* 78. <http://arxiv.org/abs/1404.0099>.
- A. McCallum, K. Schultz, and S. Singh. 2009. FACTORIE: Probabilistic programming via imperatively defined factor graphs. In *Neural Information Processing Systems Conference (NIPS)*.
- B. Milch, B. Marthi, S. Russell, D. Sontag, D. L. Ong, and A. Kolobov. 2005. BLOG: Probabilistic models with unknown objects. In *Proceedings of the 19th International Joint Conference Artificial Intelligence (IJCAI-05)*.

- R. M. Neal. 1993. Probabilistic inference using Markov chain Monte Carlo methods. Technical Report CRG-TR-93-1, University of Toronto.
- A. V. Nori, C.-K. Hur, S. K. Rajamani, and S. Samuel. 2014. R2: An efficient MCMC sampler for probabilistic programs. In *Association for the Advancement of Artificial Intelligence*. 2476–2482.
- B. Paige and F. Wood. 2014. A compilation target for probabilistic programming languages. In *Proceedings of the 31st International Conference on Machine Learning*. Vol. 32, *JMLR: W&CP*, 1935–1943.
- J. Pearl. 1988. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann. DOI: <https://doi.org/10.1016/C2009-0-27609-4>.
- J. Pearl. 2000. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, Cambridge. DOI: <https://doi.org/10.1017/S0266466603004109>.
- Y. Perov and F. Wood. 2016. Automatic sampler discovery via probabilistic programming and approximate Bayesian computation. In *Artificial General Intelligence*. 262–273. DOI: https://doi.org/10.1007/978-3-319-41649-6_27.
- A. Pfeffer. 2001. IBAL: A probabilistic rational programming language. In *Proceedings of the 17th International Joint Conference Artificial Intelligence (IJCAI-01)*.
- A. Pfeffer. 2007. The design and implementation of IBAL: A general-purpose probabilistic language. In L. Getoor and B. Taskar (Eds.), *Statistical Relational Learning*. MIT Press, Cambridge, MA.
- D. Poole. 1991. Representing Bayesian networks within probabilistic Horn abduction. In *Proceedings of the 7th Seventh Conference on Uncertainty in Artificial Intelligence (UAI-91)*. 271–278.
- D. Poole. 1993a. Logic programming, abduction and probability: A top-down anytime algorithm for computing prior and posterior probabilities. *New Gener. Comput.* 11, 3–4, 377–400. DOI: <https://doi.org/10.1007/BF03037184>.
- D. Poole. 1993b. Probabilistic Horn abduction and Bayesian networks. *Artif. Intell.* 64, 1, 81–129. DOI: [https://doi.org/10.1016/0004-3702\(93\)90061-F](https://doi.org/10.1016/0004-3702(93)90061-F).
- D. Poole. 1997. The independent choice logic for modelling multiple agents under uncertainty. *Artif. Intell.* 94, 7–56. <http://cs.ubc.ca/~poole/abstracts/icl.html>. Special Issue on Economic Principles of Multi-agent Systems. DOI: [https://doi.org/10.1016/S0004-3702\(97\)00027-1](https://doi.org/10.1016/S0004-3702(97)00027-1).
- D. Poole. 2000. Abducing through negation as failure: Stable models within the independent choice logic. *J. Log. Program.* 44, 1–3, 5–35. <http://cs.ubc.ca/~poole/abstracts/abnaf.html>. DOI: [https://doi.org/10.1016/S0743-1066\(99\)00071-0](https://doi.org/10.1016/S0743-1066(99)00071-0).
- D. Ritchie, A. Stuhlmüller, and N. Goodman. 2016. C3: Lightweight incrementalized MCMC for probabilistic programs using continuations and callsite caching. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*. 28–37.
- T. Sato. 1995. A statistical learning method for logic programs with distribution semantics. In *Proceedings of the 12th International Conference on Logic Programming (ICLP95)*. 715–729.

- <http://sato-www.cs.titech.ac.jp/reference/ICLP95.pdf>. DOI: <https://doi.org/10.7551/mitpress/4298.003.0069>.
- T. Sato and Y. Kameya. 1997. PRISM: A symbolic-statistical modeling language. In *Proceedings of the 15th International Joint Conference on Artificial Intelligence (IJCAI-97)*. 1330–1335.
- T. Sato and Y. Kameya. 2001. Parameter learning of logic programs for symbolic-statistical modeling. *J. Artif. Intell. Res. (JAIR)* 15, 391–454. <http://www.cs.cmu.edu/afs/cs/project/jair/pub/volume15/sato01a.pdf>. DOI: <https://doi.org/10.1613/jair.912>.
- A. Scibior, O. Kammar, M. Vákár, S. Staton, H. Yang, Y. Cai, K. Ostermann, S. K. Moss, C. Heunen, and Z. Ghahramani. 2018. Denotational validation of higher-order Bayesian inference. In *Proceedings of the ACM Program. Lang.* 2(POPL), 60, 1–60, 29. DOI: <https://doi.org/10.1145/3158148>.
- N. Siddharth, B. Paige, J.-W. van de Meent, A. Desmaison, N. D. Goodman, P. Kohli, F. Wood, and P. Torr. 2017. Learning disentangled representations with semi-supervised deep generative models. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), *Advances in Neural Information Processing Systems 30*, 5927–5937. Curran Associates, Inc. <http://papers.nips.cc/paper/7174-learning-disentangled-representations-with-semi-supervised-deep-generative-models.pdf>.
- Stan Development Team. 2014. Stan: A C++ Library for Probability and Sampling, Version 2.4.
- S. Staton. 2017. Commutative semantics for probabilistic programming. In *Programming Languages and Systems—26th European Symposium on Programming, ESOP 2017, Held as Part of the European Joint Conferences on Theory and Practice of Software, ETAPS 2017*. Uppsala, Sweden, April 22–29, 2017, Proceedings, 855–879. DOI: https://doi.org/10.1007/978-3-662-54434-1_32.
- S. Staton, H. Yang, F. Wood, C. Heunen, and O. Kammar. 2016. Semantics for probabilistic programming: Higher-order functions, continuous distributions, and soft constraints. In *Proceedings of the 31st Annual ACM/IEEE Symposium on Logic in Computer Science, LICS '16*. New York, July 5–8, 525–534. DOI: <https://doi.org/10.1145/2933575.2935313>.
- S. Thrun. 2000. Towards programming tools for robots that integrate probabilistic computation and learning. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. DOI: <https://doi.org/10.1109/ROBOT.2000.844075>.
- D. Tolpin, J. W. van de Meent, H. Yang, and F. Wood. 2016. Design and implementation of probabilistic programming language Anglican. *arXiv preprint arXiv:1608.05263*. DOI: <https://doi.org/10.1145/3064899.3064910>.
- D. Tran, A. Kucukelbir, A. B. Dieng, M. Rudolph, D. Liang, and D. M. Blei. 2016. Edward: A library for probabilistic modeling, inference, and criticism. *arXiv preprint arXiv:1610.09787*.
- D. Tran, M. D. Hoffman, R. A. Saurous, E. Brevdo, K. Murphy, and D. M. Blei. 2017. Deep probabilistic programming. *arXiv preprint arXiv:1701.03757*.
- Uber. 2018. Pyro. <http://pyro.ai/>. [Online; accessed 15-Aug-2018].

- J.-W. van de Meent, B. Paige, H. Yang, and F. Wood. 2018. An introduction to probabilistic programming. *arXiv preprint arXiv:1809.10756*.
- J. Vennekens, M. Denecker, and M. Bruynooghe. 2009. CP-logic: A language of causal probabilistic events and its relation to logic programming. *Theor. Pract. Log. Program. (TPLP)* 9, 3, 245–308. DOI: <https://doi.org/10.1017/S1471068409003767>.
- M. J. Wainwright and M. I. Jordan. 2008. Graphical models, exponential families, and variational inference. *Found. Trends Mach. Learn.* 1, 1–2, 1–305. DOI: <https://doi.org/10.1561/2200000001>.
- D. Wingate, A. Stuhlmüller, and N. D. Goodman. 2011. Lightweight implementations of probabilistic programming languages via transformational compilation. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics*, 131.
- Y. Zhou, B. J. Gram-Hansen, T. Kohn, T. Rainforth, H. Yang, and F. Wood. 16–18 Apr 2019. LF-PPL: A low-level first order probabilistic programming language for non-differentiable models. In K. Chaudhuri and M. Sugiyama (Eds.), *Proceedings of Machine Learning Research*, Vol. 89. *Proceedings of Machine Learning Research*, 148–157. PMLR. <http://proceedings.mlr.press/v89/zhou19b.html>.

An Interventionist Approach to Mediation Analysis

James M. Robins (Harvard T. H. Chan School of Public Health),
Thomas S. Richardson (University of Washington),
Ilya Shpitser (Johns Hopkins University)

Judea Pearl’s insight that, when errors are assumed independent, the Pure (aka Natural) Direct Effect (PDE) is non-parametrically identified via the Mediation Formula was “path-breaking” in more than one sense. In the same paper, Pearl described a thought-experiment as a way to motivate the PDE. Analysis of this experiment led Robins and Richardson to a novel way of conceptualizing direct effects in terms of interventions on an expanded graph in which treatment is decomposed into multiple separable components. We further develop this novel theory here, showing that it provides a self-contained framework for discussing mediation without reference to cross-world (nested) counterfactuals or interventions on the mediator. The theory preserves the dictum “no causation without manipulation” and makes questions of mediation empirically testable in future randomized controlled trials. Even so, we prove the interventionist and nested counterfactual approaches remain tightly coupled under a non-parametric structural equation model except in the presence of a “recanting witness.” In fact, our analysis also leads to a simple, sound, and complete algorithm for determining identification in the (non-interventionist) theory of path-specific counterfactuals.

38.1 Introduction

In the companion paper [Chapter 41], we described graphical counterfactual models corresponding to the finest fully randomized causally interpretable structured tree graph (FFRCISTG) models of [Robins \[1986\]](#). Such models correspond to

conditional independencies encoded by the d-separation criterion in single-world intervention graphs (SWIGs). We gave a general identification theory for treatment effects in such models with hidden variables that is a generalization of the theory described in [Tian and Pearl 2002, Shpitser and Pearl 2006a]. We also described the differences between the FFRCISTG/SWIG model and the non-parametric structural equation model with independent errors (NPSEM-IE).

Given that a treatment effect of A on Y is established, it is often desirable to try to understand the contribution to the total effect of A on Y of the other variables M that lie on causal pathways from A to Y . Variables lying on the causal pathways are called mediators of the effect of A on Y . This leads to consideration of direct and indirect effects. Several different types of direct effect have been considered previously. Most have asked whether “the outcome (Y) would have been different had cause (A) been different, but the level of the mediator (M) remained unchanged.” Differences arise regarding what it means to say that M remains unchanged.

In Section 38.2, we will first review these notions, the assumptions under which they are identified and the extent to which identification claims can be verified (in principle) via an experiment. These considerations will lead to a novel way of conceptualizing direct effects introduced in Robins and Richardson [2010] and generalized in Section 38.3 herein. This novel interventionist approach does not require counterfactuals defined in terms of the mediator. We first describe our interventionist approach to causal mediation analysis in the context of direct and indirect effects. This approach (i) need not assume that the mediator M has well-defined causal effects and/or counterfactuals, but (ii) instead hypothesizes that treatment variable A can be decomposed into multiple separable components each contributing to the overall effect of treatment, (iii) preserves the dictum “no causation without manipulation,” (iv) makes questions of mediation empirically testable in future randomized controlled trials, (v) may facilitate communication with subject matter experts, and (vi) when identified from data, the identifying formulae under an interventionist approach and those obtained from the other (mediator-based) approach are identical; however, the causal effects being identified differ since they refer to intervening on different variables. This theory has been extended and applied by others in the context of mediation in survival analysis [Didelez 2019, Aalen et al. 2020] and recently to competing risks [Stensrud et al. 2020a, 2020c] and interference problems [Shpitser et al. 2017]; see also Lok [2016].

Finally, in Section 38.4, we extend this approach to arbitrary (identified) path-specific effects. This allows the (generalized) ID algorithm described in Chapter 41 plus one extra step to be used as a sound and complete algorithm for determining the identification of path-specific distributions; we also provide a version

for handling conditional queries. Finally, we describe the differences between the nested counterfactual approach and the interventionist approach under non-identification.

38.2 Approaches to Mediation Based on Counterfactuals Defined in Terms of the Mediator: The CDE and PDE

One approach to having M remain unchanged would be for M to be fixed via an intervention. On this view both A and M are treatments, and we consider the difference between an intervention setting A to a and M to m versus an intervention setting A to a' and M to m . This leads to the definition of a *controlled direct effect* (CDE):

$$CDE_{a,a'}(m) \equiv \mathbb{E}[Y(a', m) - Y(a, m)], \quad (38.1)$$

where $Y(a, m)$ is the counterfactual response of Y had A and M been set, possibly contrary to fact, to values a and m , respectively. Note that there is a controlled direct effect for every value m of M . Given a causal graph \mathcal{G} , a straightforward application of graphical causal identification theory¹ determines whether the distribution $P(Y(a, m))$, and thus whether the $CDE_{a,a'}(m)$ contrast, is identified.

There are situations in which other notions of direct effect are more natural. In particular, there are contexts in which we wish to know whether A taking the value of a' (rather than the baseline level a) would lead to a change in the value of Y if the effect of A on M were “blocked.” Specifically, if the effect on M of A taking a' was blocked so that the mediator M takes the value $M(a)$ that M would take were A set to the baseline value a .

Along these lines, in many contexts we may ask what “fraction” of the (total) effect of A on Y may be attributed to a particular causal pathway. For example, consider a randomized controlled trial that investigates the effect of an anti-smoking intervention (A) on the outcome myocardial infarction (MI) at 2 years (Y) among non-hypertensive smokers. For simplicity, assume everyone in the treatment arm and no one in the placebo arm quit cigarettes, that all subjects were tested for new-onset hypertension (M) at the end of the first year, and no subject suffered an MI in the first year. Hence A , M , and Y occur in that order. Suppose the trial showed smoking cessation had a beneficial effect on both hypertension and MI. It is natural to ask: “What fraction of the total effect of smoking cessation (A) on MI (Y) is via a pathway that does not involve hypertension (M)?”

1. See the companion paper [Chapter 41] in this volume and references therein.

These ideas lead to the *pure direct effect*² (PDE) defined in [Robins and Greenland \[1992\]](#), which in counterfactual notation can be written as follows:

$$PDE_{a,a'} \equiv \mathbb{E}[Y(a', M(a)) - Y(a, M(a))] = \mathbb{E}[Y(a', M(a)) - Y(a)]. \quad (38.2)$$

This is the difference between two quantities: first, the outcome Y that would result if we set A to a' , while “holding fixed” M at the value $M(a)$ that it would have taken had A been a ; second, the outcome Y that would result from simply setting A to a [and thus having M again take the value $M(a)$]. In [Robins and Greenland \[1992\]](#), the first was alternately described as the result of setting A to a' on Y when the effect of A on M is blocked, thereby leaving M unchanged from its reference value $M(a)$. Thus the PDE interprets had “ M remained unchanged” to mean “had (somehow) M taken the value that it would have taken had we fixed A to a .”

As another application of this idea, [Pearl \[2009, p. 131\]](#) cites the following legal opinion arising in a discrimination case: “*The central question in any employment discrimination case is whether the employer would have taken the same action had the employee been of a different race (age, sex, religion, national origin etc.) and everything else had been the same.*” [Carson vs. Bethlehem Steel Corp., 70 FEP Cases 921, 7th Cir. (1996)].

Here A corresponds to membership in a protected class, while M corresponds to criteria, such as qualifications, that are permitted to be considered in such decisions. If the PDE is non-zero, then discrimination has taken place.

A notion of the indirect effect may be defined similarly, using the nested counterfactual $Y(a', M(a))$. On the additive scale, direct and indirect effects may be used to give a decomposition of the average causal effect [[Robins and Greenland 1992](#)]:

$$\begin{aligned} \underbrace{\mathbb{E}[Y(a')] - \mathbb{E}[Y(a)]}_{\text{average causal effect}} &= \underbrace{(\mathbb{E}[Y(a')] - \mathbb{E}[Y(a', M(a))])}_{\text{total indirect effect}} + \underbrace{(\mathbb{E}[Y(a', M(a))] - \mathbb{E}[Y(a)])}_{\text{pure direct effect}} \\ &= \underbrace{(\mathbb{E}[Y(a')] - \mathbb{E}[Y(a, M(a'))])}_{\text{total direct effect}} + \underbrace{(\mathbb{E}[Y(a, M(a'))] - \mathbb{E}[Y(a)])}_{\text{pure indirect effect}}. \end{aligned} \quad (38.3)$$

Notice that the PDE depends on $Y(a', M(a))$ —a variable in which two different levels of a are nested within the counterfactual for Y . Consequently, in contrast to $CDE_{a,a'}$, this counterfactual does not correspond to any experimental intervention on A and M . This is because in order to know the value $M(a)$ for a unit, it is necessary to set A to a , but this then precludes setting A to a' . This is a manifestation of the fundamental problem of causal inference: it is not possible for a single unit to

2. Also called the natural direct effect in [Pearl \[2001\]](#).

receive two different levels of the same treatment at the same time. For this reason, the counterfactual $Y(a', M(a))$ is referred to as a *cross-world* counterfactual.³

The differences in the assumptions made by the FFRCISTG and NPSEM-IE models lead to quite different identification results for the PDE in the context of the simple DAG shown in Figure 38.3(a). These differences reflect important epistemological distinctions between the two frameworks that are described in Sections 38.2.3 and 38.2.4 below.

38.2.1 Two Hypothetical River Blindness Treatment Studies

We will use as a running example the following pair of hypothetical studies that are represented together in the single causal graph shown in Figure 38.1(a), as described below: A random sample of individuals in an impoverished medically underserved catchment area are selected to participate in a double-blind placebo controlled randomized trial ($A = 1$ selected, $A = 0$ otherwise) of single dose therapy with the drug ivermectin ($M = 1$ received the drug, $M = 0$ otherwise) for the treatment of onchocerciasis (river blindness). The outcome is diminished vision 9 months later ($Y = 1$ if worse than 20/100 and $Y = 0$ otherwise). All subjects in the trial complied with their assigned therapy. The trial was motivated in part by the fact that ivermectin was already being sold by local shop owners as a cure for river blindness without evidence for effectiveness or safety in the local population. After the trial finished, a non-governmental organization carried out a retrospective observational cohort study on a random subset of non-selected subjects ($A = 0$), collecting data on M and Y .

In a subset of patients ivermectin can actually decrease visual acuity. This occurs when the larvae in the eye killed by ivermectin induce an overvigorous immune response. Because of this side effect, a clinic available to all trial participants was established to screen for the above side-effect and treat with immunosuppressive drugs if required ($R = 1$ if treated with immunosuppressants, $R = 0$ otherwise) to prevent further eye damage. The patients not selected for the study had neither access to the clinic nor to immunosuppressive therapy. Thus letting $S = 1$ ($S = 0$) denote access (no access) to a clinic, we have $S = 1$ iff $A = 1$ and $R = 0$ if $A = 0$. Finally we let U denote an unmeasured variable with $U = 1$ and $U = 0$ denoting individuals with greater versus lesser propensity to take medical treatments if offered. In the unselected subjects ($A = 0$), U is thus positively correlated with M . In contrast, owing to random assignment of M , U , and M are independent in selected subjects ($A = 1$). On the other hand, in selected subjects, conditional

3. Formally, we will say that a counterfactual expression is *cross-world* if it involves assigning more than one value to a single index such as a and a' in $Y(a', M(a))$.

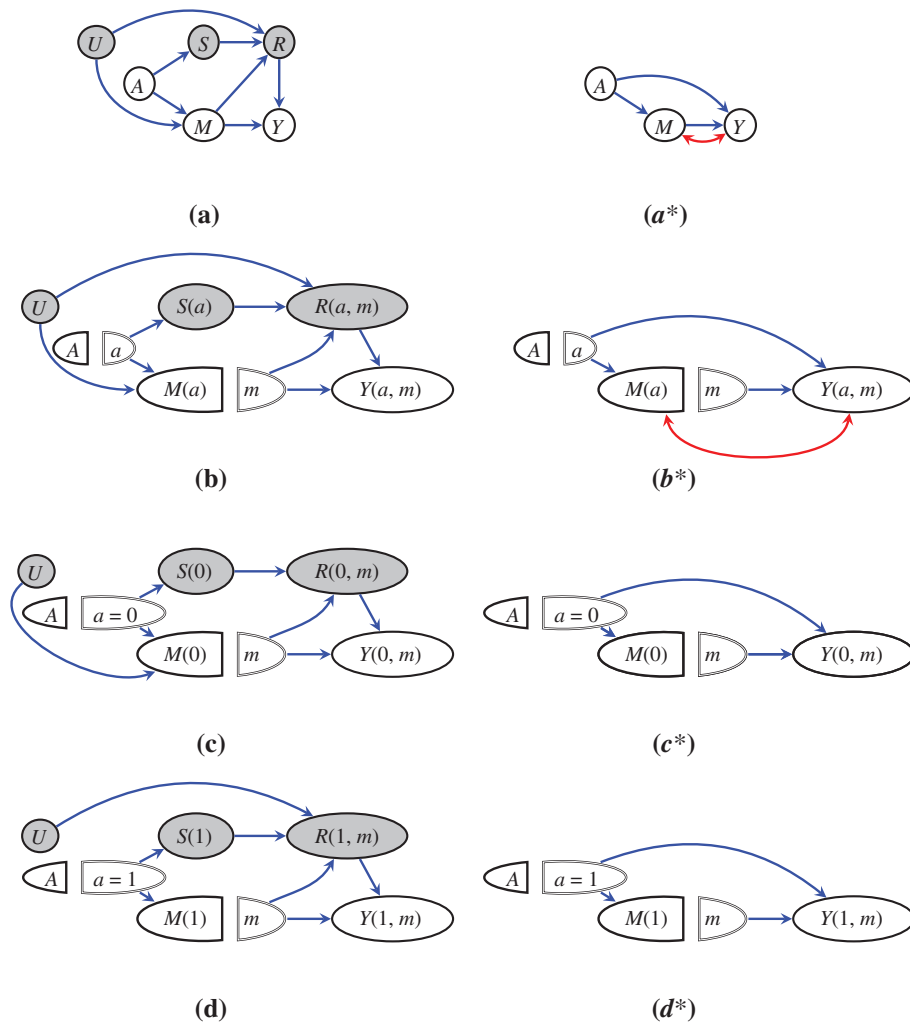


Figure 38.1 (a) A DAG \mathcal{G} representing the underlying structure in the combined observational and randomized trial of treatments for river blindness. Shaded variables are not observed. Here A indicates a randomized trial ($a = 1$) or an observational study ($a = 0$); M is whether the patient received ivermectin; Y is the patient’s vision; S indicates access to a clinic; R is treatment with immunosuppressants; U indicates the inclination of the patient to avail themselves of medical care that is offered. (b) The SWIG $\mathcal{G}(a, m)$ resulting from \mathcal{G} ; (c) and (d) show SWIGs $\mathcal{G}(a = 0, m)$ and $\mathcal{G}(a = 1, m)$ that incorporate additional context-specific causal information. (a*), (b*), (c*), (d*) show the corresponding latent projections.

on M , U is positively correlated with R and thus with Y ; in unselected subjects U and Y are independent given M because treatment with R is not available. (Here

we assume U has an effect on Y only through the treatments actually received.) Finally, records recording which of the selected subjects received immunosuppressive therapy in the trial were destroyed in a fire so the data available for analysis were solely (A, M, Y) , the same variables available in the observational study.

38.2.2 The PDE and CDE in the River Blindness Studies

We next explain the meaning of the CDE and PDE in the context of the ivermectin studies. There we suppose the true data-generating process is as described by the DAG in Figure 38.2(a). Recall that the corresponding latent projection⁴ is given in Figure 38.1(a*) because in Figure 38.1(a) there is a causal path from A to Y , and U is a common cause of M and Y . In this context $M(a = 0)$ is the ivermectin treatment that the patient would select in the observational study where A is 0. Likewise $Y(a = 1, M(a = 0))$ is the patient's outcome if they received the ivermectin treatment as in the observational study (where A is 0) but, as in the randomized study (where A is 1), a clinic ($S = 1$) was made available to them.⁵ To see why, note that $Y(a = 1, M(a = 0))$ corresponds to the effect of setting $a = 1$ on Y through all causal pathways ($S \rightarrow R \rightarrow Y$) not passing through ivermectin (M) when M remains at its self-selected value $M(a = 0)$ in the observational study. Finally, $Y(a = 0) \equiv Y(a = 0, M(a = 0))$ is the patient's outcome if they were assigned to the observational study and did not have access to the clinic. The PDE is the mean of the difference $Y(a = 1, M(a = 0)) - Y(a = 0, M(a = 0))$.

The $CDE(m = 0)$ is the mean effect on Y of having ($s = 1$) versus not having ($s = 0$) a clinic available had no one received ivermectin ($m = 0$).⁶ Likewise,

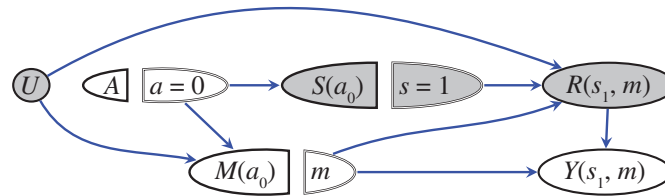


Figure 38.2 The SWIG $\mathcal{G}(a = 0, s = 1, m)$ associated with the DAG shown in Figure 38.1(a); here a_0 and s_1 are short for $a = 0$ and $s = 1$.

4. See Section 41.4.1 in the companion paper [Chapter 41] in this volume for this definition and other standard graphical definitions.

5. If patients know that a clinic is available then this may influence their decision to take ivermectin; thus, in order for the patient's decision to be as in the observational study, they would need to make decisions relating to ivermectin treatment before being told that a clinic is available.

6. This is because $Y(a, m) = Y(S(a), m) = Y(s, m)$ since $S(a) = a$.

$CDE(m = 1)$ is the clinic effect when all subjects received ivermectin ($m = 1$). In contrast, as noted above, the PDE is the mean effect of having versus not having a clinic available had subjects chosen ivermectin treatment (M) as in the observational study.

Identification of the CDE in the River Blindness Studies

We next consider whether the $CDE(m = 0)$ and $CDE(m = 1)$ are identified from the available data on A , M , and Y . From examining the latent projection shown in Figure 38.1(a*), one would expect that the CDE is not identified owing to the bidirected edge $M \leftrightarrow Y$. In particular, identification does not follow from existing methods such as the *do*-calculus, the ID algorithm,⁷ or the back-door criterion [Pearl 2009], although see Tikka et al. [2019].

However, we will now show, perhaps surprisingly, that we do have identification of the CDE. This is a consequence of context specific independencies. Although such independencies cannot be represented using standard causal DAGs (or their latent projections) an extension to context-specific SWIGs due to Dahabreh et al. [2019] and Sarvet et al. [2020] makes this possible.

Consider the SWIG $\mathcal{G}(a, m)$ resulting from a joint intervention setting A to a and M to m as shown in Figure 38.1(b) and its latent projection shown in Figure 38.1(b*). This SWIG $\mathcal{G}(a, m)$, shown in Figure 38.1(b), represents the conditional independence relations that hold for all four of the possible counterfactual distributions ($a, m \in \{0, 1\}$) resulting from jointly intervening on A and M .

Consider the two interventions setting $a = 1$ and $m \in \{0, 1\}$. These interventions correspond to performing the randomized trial in which treatment M is randomly assigned and thus $U \perp\!\!\!\perp M(a = 1)$. Consequently, the distribution of the counterfactuals may be represented by the SWIG $\mathcal{G}(a = 1, m)$, shown in Figure 38.1(d), in which the edge $U \rightarrow M(a = 1)$ is absent.

Now consider the remaining two interventions setting $a = 0$, that is, assigned to the observational study. Recall that the clinic (S) and (hence) treatment with immunosuppressants (R) are only available to those people who were selected for the trial ($a = 1$). Thus, for all subjects in the observational study ($a = 0$), $S(0) = R(0, m) = 0$. Consequently, $U \perp\!\!\!\perp R(a = 0, m) | A, M(a = 0)$, since the inclination (U) of the patient does not affect whether they have access to immunosuppressants. Hence the distribution of the counterfactuals may be represented by the

7. Note that the ID algorithm is complete [Shpitser and Pearl 2006a] with respect to models defined in terms of the independences holding in a (standard) causal DAG; these independences are not context specific.

SWIG $\mathcal{G}(a = 0, m)$, shown in Figure 38.1(c), in which the edge $U \rightarrow R(a = 0, m)$ is absent.⁸

Applying d-separation to the latent projections in Figure 38.1(c*) and (d*), we see that⁹

$$Y(a, m) \perp\!\!\!\perp M(a), A \quad \text{for } a = 0, 1. \quad (38.4)$$

Consequently, since $Y = Y(a, M(a) = m)$ on the event $A = a, M(a) = m$,

$$p(Y | A = a, M = m) = p(Y(a, m)), \quad (38.5)$$

is identified for both $a = 0$ and $a = 1$. Hence the $CDE(m)$ of A and M on Y is identified.

As noted above, the identification (Equation (38.5)) does not follow from the DAG in Figure 38.1(a), with the latent projection in Figure 38.1(a*). However, the independences $U \perp\!\!\!\perp R(a = 0, m) | A, M(a = 0)$ and $U \perp\!\!\!\perp M(a = 1) | A$ can be encoded by and read from the context-specific SWIGs shown in Figure 38.1(c) and (d), respectively.¹⁰

Together with consistency, these independences imply, respectively, $U \perp\!\!\!\perp R | A = 0$ and $U \perp\!\!\!\perp M | A = 1$. Since, in addition to Equation (38.4), we also have $M(a) \perp\!\!\!\perp A$ for $a = 0, 1$, it follows that the distribution of the counterfactuals $\{A, M(a), Y(a, m)$ for all $a, m\}$ obeys the FFRCISTG model associated with the graph shown in Figure 38.3(a) in which there are no bi-directed edges. Interestingly, we show below in Section 38.2.5 that $M(a = 0) \not\perp\!\!\!\perp Y(a = 1, m)$. Consequently, the distribution of the counterfactuals does not obey the NPSEM-IE associated with Figure 38.3(a), and thus as discussed below, the PDE is not identified without additional assumptions beyond those of the NPSEM-IE model.¹¹

8. Since $S(0)$ and $R(0, m)$ are constants we could also leave out the edges $S(0) \rightarrow R(0, m)$ and $m \rightarrow R(0, m)$ on the same basis, but this would not change our conclusions.

9. Recall that when testing d-separation in SWIGs, fixed nodes such as $a = 0$ in Figure 38.1(c*) and $a = 1$ in Figure 38.1(d*) always block paths they occur on (when they are not end points).

10. These independences cannot be read from the SWIG $\mathcal{G}(a, m)$ shown in Figure 38.1(b) that was constructed from \mathcal{G} . Note that the graph $\mathcal{G}(a, m)$ contains the union of edges in the two context-specific SWIGs $\mathcal{G}(a = 0, m)$ and $\mathcal{G}(a = 1, m)$ shown in Figure 38.1(c) and (d), respectively. $\mathcal{G}(a, m)$ thus represents the (non-context-specific) conditional independence relations that are common to all four instantiations $a, m \in \{0, 1\}$.

11. The distribution does obey the NPSEM-IE (hence also the FFRCISTG) associated with Figure 38.1(a*) that includes the $M \leftrightarrow Y$ edge; see also Footnote 23.

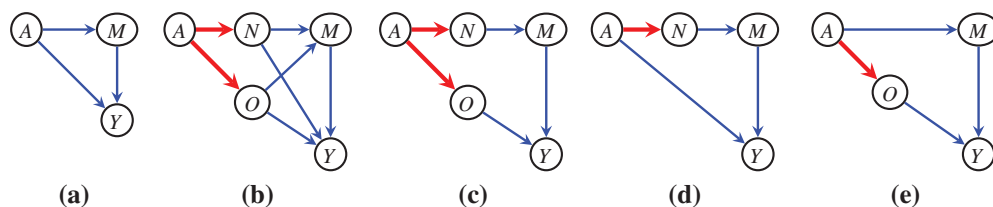


Figure 38.3 (a) A simple DAG \mathcal{G} representing a causal model with a treatment A , a mediator M , and a response Y . (b) An expanded causal model where N and O are a decomposition of A ; N and O are, respectively, the nicotine and non-nicotine components of tobacco. Thicker red edges indicate deterministic relations. (c) An expanded version \mathcal{G}^{ex} of the DAG \mathcal{G} in (a), and an edge subgraph of the DAG in (b), where N does not cause Y directly, and O does not cause M directly. In this graph the direct and indirect effects of A on Y may be defined via interventions on N and O . (d) Special case of (c) in which A plays the role of O ; (e) Special case of (c) in which A plays the role of N .

38.2.3 Identification of the PDE via the Mediation Formula under the NPSEM-IE for Figure 38.3(a)

We now consider the situation in which, unlike the ivermectin example above, the NPSEM-IE associated with the graph in Figure 38.3(a) holds. In this case the $PDE_{a,a'}$ is identified via the following Mediation Formula [Pearl 2001, 2012]:

$$\text{med}_{a,a'} \equiv \sum_m (\mathbb{E}[Y | m, a] - \mathbb{E}[Y | m, a']) p(m | a') \quad (38.6)$$

$$= \left(\sum_m \mathbb{E}[Y | m, a] p(m | a') \right) - \mathbb{E}[Y | a']. \quad (38.7)$$

The proof of this result under the NPSEM-IE is as follows:

$$\begin{aligned} p(Y(a, M(a') = m) = y) & \\ &= \sum_m p(Y(a, M(a') = m) = y | M(a') = m) p(M(a') = m) \\ &= \sum_m p(Y(a, m) = y) p(M(a') = m) \\ &= \sum_m p(Y = y | A = a, M = m) p(M = m | A = a'). \end{aligned}$$

Here the first line follows from elementary probability, the second from the cross-world NPSEM-IE independence:

$$Y(a, m) \perp\!\!\!\perp M(a')$$

which follows from (41.5); the third follows from the FFRCISTG independence (41.6) and thus also from (41.5). Under the NPSEM-IE associated with Figure 41.5 (a*) identification fails because this cross-world independence does not hold.

38.2.4 Partial Identification of the PDE Under the FFRCISTG for Figure 38.3(a)

In contrast, under the less restrictive FFRCISTG model associated with the graph in Figure 38.3(a), the PDE is not, in general, identified. This follows from the fact that the second equality in the previous proof relies on the cross-world independence $Y(a, m) \perp\!\!\!\perp M(a')$; but the FFRCISTG model does not assume any cross-world independencies. However, under the FFRCISTG the observed data implies bounds on the PDE. For example, in the case in which M and Y are binary we have the following sharp bounds [Robins and Richardson 2010]:

$$\begin{aligned} & \max\{0, p(M = 0 | A = a') + p(Y = 1 | A = a, M = 0) - 1\} + \\ & \max\{0, p(M = 1 | A = a') + p(Y = 1 | A = a, M = 1) - 1\} - p(Y = 1 | A = a') \\ & \leq PDE_{a, a'} \leq \\ & \min\{p(M = 0 | A = a'), p(Y = 1 | A = a, M = 0)\} + \\ & \min\{p(M = 1 | A = a'), p(Y = 1 | A = a, M = 1)\} - p(Y = 1 | A = a'). \end{aligned}$$

A proof is given in the [Appendix](#).

38.2.5 An Example in Which an FFRCISTG Model Holds, but an NPSEM-IE Does Not

The differing results on identifiability for the PDE in the previous two sections raise the question as to whether it is most appropriate to adopt the NPSEM-IE (41.5) or FFRCISTG (41.6) assumptions in practice. As shown above, this choice matters since if one assumes the NPSEM-IE associated with the simple graph in Figure 38.3(a) then one will believe the PDE is point identified, while if one assumes the FFRCISTG associated with Figure 38.3(a) only bounds may be obtained.

It has been argued that it is hard to conceive of a realistic data-generating process under which the FFRCISTG model holds, but the NPSEM-IE model does not. However, we now show that the river blindness studies described above provides a counterexample.

Recall that due to records being destroyed in a fire, only the three variables (A, M, Y) are observed. The first (A) is randomly assigned and the analyst assumes (possibly incorrectly) that there is no other variable (either measured or unmeasured) that is a cause of both the mediator M and the final response Y . This hypothesis would imply the causal structure depicted in Figure 38.3(a). The graph in Figure 38.3(a) would be the latent projection of Figure 38.1(a) if it were assumed incorrectly that there is no unmeasured confounder U . Both the FFRCISTG and NPSEM-IE models associated with Figure 38.3(a) imply:

$$Y(a, m) \perp\!\!\!\perp M(a), \quad \text{for } a = 0, 1. \quad (38.8)$$

The NPSEM-IE also implies the cross-world independence:

$$Y(a = 1, m) \perp\!\!\!\perp M(a = 0). \quad (38.9)$$

We have already shown that the underlying data-generating process in the ivermectin example implies Equation (38.8); see Equation (38.4) above. We now show that Equation (38.9) fails to hold for almost all laws corresponding to the NPSEM-IE associated with Figure 38.1(a). This remains true even if we impose, in addition, the context-specific counterfactual independences needed to identify the CDE that are encoded in Figure 38.1(c) and (d).

To see this, first consider the SWIG $\mathcal{G}(a = 0, s = 1, m)$ shown in Figure 38.2 that is constructed from the DAG in Figure 38.1(a). There is a d-connecting path $Y(s = 1, m)$ to $M(a = 0)$, namely $M(a = 0) \leftarrow U \rightarrow R(s = 1, m) \rightarrow Y(s = 1, m)$. Consequently, for almost all distributions in the FFRCISTG model, it holds that $M(a = 0) \not\perp\!\!\!\perp Y(s = 1, m)$.¹² Further, we have:

$$Y(s = 1, m) = Y(S(a = 1), m) = Y(a = 1, m),$$

where in the first equality we use that, due to determinism, $S(a = 1) = 1$ for all individuals; the second follows from recursive substitution.¹³ Hence Equation (38.9) does not hold¹⁴ from which it follows that the PDE is not identified; see Appendix 38.A.2. Consequently, the distribution of the counterfactuals $\{A, M(a), Y(a, m)$ for all $a, m\}$ is not in the NPSEM-IE model associated with Figure 38.3(a). Interestingly we see that by considering a SWIG with interventions on three variables A, S , and M we have shown that the FFRCISTG model plus recursive substitution and some determinism can be used to prove that a cross-world independence fails to hold; see Footnote 23 and Appendix 38.A.3.

Finally, we note that in the ivermectin example the Acyclic Directed Mixed Graph (ADMG)¹⁵ with fewest edges (over A, M, Y) that represents the distribution of the counterfactuals $\{A, M(a), Y(a, m)$ for all $a, m\}$ is Figure 38.3(a) under the FFRCISTG. In contrast, the minimal ADMG that represents this distribution under the NPSEM-IE is the graph with an additional bi-directed confounding arc shown in Figure 38.1(a*). This has the following interesting consequence: typically, people

12. This is also true if we assume the NPSEM-IE associated with the graph in Figure 38.1(a).

13. See Equation (41.1) in Chapter 41 in this volume.

14. Intuitively, this should not be surprising since U is a “common cause” of $M(a = 0)$ and $Y(a = 1, m)$ in that $U \rightarrow M(a = 0)$ in Figure 38.1(c) while there is a path $U \rightarrow R(a = 1, m) \rightarrow Y(a = 1, m)$ in Figure 38.1(d).

15. see Section 41.2.1 in Chapter 41 in this volume.

may make a statement such as “Figure 38.3(a) is *the* true causal graph.” However, we now see that this statement does not have a truth-value without clarifying whether we are referring to the NPSEM-IE or the FFRCISTG as the true underlying counterfactual model.

On the other hand, one might prefer to replace “Figure 38.3(a)” with “Figure 38.1(a^*)” in the statement above, since then the modified statement holds for both counterfactual models;¹⁶ furthermore, Figure 38.1(a^*) accurately indicates that there is confounding for the PDE. However, the disadvantage of this choice is that the inclusion of the bi-directed edge $M \leftrightarrow Y$ does not reveal that the CDE is identified via $E[Y(a, m) | A = a, M = m] = E[Y | A = a, M = m]$.

38.2.6 Testable Versus Untestable Assumptions and Identifiability

Given a graph such as Figure 38.3(a), in principle, there is an empirical test of the FFRCISTG model.¹⁷ However, there is no additional empirical test (on the variables in the graph) for the extra assumptions made by the corresponding NPSEM-IE, which are required to identify the PDE. Consequently, there is a qualitative distinction in the testability of the identification assumptions for these two contrasts.

In more detail, identification of the $CDE(m)$, $m = 0, 1$ is, in principle, subject to direct empirical test: one conducts a four-armed randomized experiment on subjects drawn from the same population, in which both A and M are randomly assigned to their four possible joint values. If for any (a, m) the distribution in the four-arm (A, M) randomized trial $p(Y(a, m) = y)$ ¹⁸ differs from the conditional distribution in the two-armed trial $p(Y(a) = y | M(a) = m) = p(Y = y | M = m, A = a)$ in which only A was randomized, then we may infer that an unmeasured common cause was present between M and Y and hence it was incorrect to postulate the causal DAG in Figure 38.3(a), regardless of whether we are considering the FFRCISTG or NPSEM-IE models associated with this graph.

16. Since the distribution of $\{A, M(a), Y(a, m)\}$ for all a, m obeys the FFRCISTG corresponding to Figure 38.3(a), which is a subgraph of Figure 38.1(a^*), the distribution also obeys the FFRCISTG corresponding to this latter graph.

17. Formally, this requires that one can observe the “natural” value of a variable prior to intervention.

18. Here and throughout this section, we will use $p(A, M, Y)$ to denote the observed distribution in which only A is randomized; we will use $\{p(Y(a, m))\}$ for $a, m \in \{0, 1\}$ to indicate the four distributions that *would* be observed if *both* A and M were to be randomized in a four-arm trial. This is because in such a trial in the arm in which people are assigned to $A = a, M = m$ we would observe the counterfactual $Y(a, m)$.

In contrast, whether the PDE equals the mediation formula (38.7) cannot be empirically tested using data on (A, M, Y) . Even if we can directly manipulate M (in addition to A), there is no experiment involving A and M such that the resulting contrast corresponds to the PDE. This is for the following reason: to observe, for a given subject, the cross-world counterfactual $Y(a = 1, M(a = 0))$ that occurs in the PDE, one would need to “first” assign them to $a = 0$ and record $M(a = 0)$, and “then” perform a “second” experiment (on the same subject) in which they are assigned to $a = 1$ and the recorded value $M(a = 0)$ from the “first” experiment. However, this is usually not possible for the simple reason that having assigned the patient to $a = 0$ in the “first” experiment precludes “subsequently” assigning them to $a = 1$, except in the rare circumstances where a valid cross-over trial is feasible.

These considerations are particularly relevant in a setting such as the ivermectin example, where, as shown above, the distribution over the counterfactual variables $\{A, M(a), Y(a, m)$ for all $a, m\}$ obeys the FFRCISTG but not the NPSEM-IE model associated with the causal DAG in Figure 38.3(a). In particular, an analyst who was unaware of the variables U, R, S in Figure 38.1(a) and posited the model in Figure 38.3(a) would find no evidence of confounding between M and Y even if they were to subsequently perform a four-arm (A, M) randomized trial.

In summary, the PDE identification via the mediation formula (38.7) requires not only that there be no *detectable* single-world confounding between M and Y (as assumed by the FFRCISTG), but, in addition, that undetectable cross-world confounding also be absent.¹⁹ Consequently, as with the ivermectin example, it is possible for the mediation formula to give an inconsistent estimate of the PDE, yet for this to be undetectable given any randomized experiment that could be performed using the variables A, M, Y on the graph in Figure 38.3(a).

38.3 Interventionist Theory of Mediation

The above considerations motivate a theory of mediation based on interventions on sub-components of treatment, rather than on the mediator.

38.3.1 Interventional Interpretation of the PDE Under an Expanded Graph

As described above, the counterfactual $E[Y(a = 1, M(a = 0))]$ and thus the PDE cannot be empirically tested by any intervention on the variables on the graph.

19. If, after carrying out an experiment in which A and M are randomly assigned, it is observed that $p(Y(a, m))$ and $p(Y|A = a, M = m)$ are statistically indistinguishable, then this would almost certainly increase the probability that a Bayesian would assign to cross-world independence holding. However, the ivermectin example shows the importance of the investigator—Bayesian or not—thinking carefully about the underlying data-generating mechanism.

However, curiously, Pearl has often argued that the PDE is a causal contrast of substantive and public-health importance by offering examples along the following lines.

Example: Nicotine-Free Cigarette

Consider the example discussed in Section 38.2 where we have data from a randomized smoking-cessation trial. We have data available on smoking status A , hypertensive status M 6 months after randomization, and myocardial infarction (MI) status Y at 1 year.

Following a similar argument given in Pearl [2001]²⁰ to motivate the PDE, suppose that nicotine-free cigarettes will be newly available starting a year from now. The substantive goal is to use the already collected data from the smoking cessation trial to estimate the difference two years from now in the incidence of MI if all smokers were to change to nicotine-free cigarettes when they become available (in a year) compared to the incidence if all smokers were to stop smoking altogether (in a year).

Further suppose it is believed that the entire effect of nicotine on MI is through its effect on hypertensive status, while the non-nicotine toxins in cigarettes have no effect on hypertension and that there do not exist unmeasured confounders for the effect of hypertension on MI.

In this context, a researcher following the approach that has been advocated by Pearl may postulate that the smoking cessation trial is represented by the NPSEM-IE model associated with Figure 38.3(a). Under these assumptions, the MI incidence in smokers of cigarettes free of nicotine would be $E[Y(a = 1, M(a = 0))]$ since the hypertensive status of smokers of nicotine-free cigarettes will equal their hypertensive status under non-exposure to cigarettes. Thus $E[Y(a = 1, M(a = 0))]$ is precisely the incidence of MI in smokers two years from now were all smokers to change to nicotine-free cigarettes a year from now, and thus the PDE:

$$PDE = E[Y(a = 1, M(a = 0))] - E[Y(a = 0, M(a = 0))] \quad (38.10)$$

is the causal contrast of interest.

Given the assumption of an NPSEM-IE, it follows that $E[Y(a = 1, M(a = 0))]$ equals $\sum_m E[Y | A = 1, M = m]p(m | A = 0)$, and therefore the PDE is identified from the mediation formula applied to the data from the smoking cessation trial.

20. Pearl [2001, section 2] considers a similar example, but where A is a drug, M is aspirin taken to mitigate side-effects, and Y is the final outcome.

What is interesting about Pearl's motivation is that to argue for the substantive importance of the parameter $E[Y(a = 1, M(a = 0))]$, he tells a story about the effect of a manipulation—a manipulation that makes no reference to M at all. Rather, in this context, the manipulation is to intervene to eliminate the nicotine component of cigarettes.

The most direct representation of this story is provided by the *expanded* DAG \mathcal{G}^{ex} in Figure 38.3(c), where N is a binary variable representing nicotine exposure and O is a binary variable representing exposure to the other non-nicotine components of a cigarette. The bolded arrows from A to N and O indicate deterministic relationships: $N(a) = O(a) = a$. This is because in the factual data (with probability one) either one smokes normal cigarettes so $A = N = O = 1$ or one is a non-smoker (i.e., ex-smoker) and $A = N = O = 0$.

This expanded graph now provides a simple interventional interpretation of the PDE. The researcher's assumption of the NPSEM-IE associated with Figure 38.3(a) together with the existence of the variables N and O and associated counterfactuals imply that the nested counterfactual $Y(a = 1, M(a = 0))$ is equal to the simple counterfactual $Y(n = 0, o = 1)$, the outcome had we intervened to expose all subjects to the non-nicotine components, but not to the nicotine components.²¹ It follows that

$$PDE = E[Y(n = 0, o = 1)] - E[Y(n = 0, o = 0)]. \quad (38.11)$$

Furthermore, these assumptions imply that the graph in Figure 38.3(c) is an FFRCISTG. It then follows from proposition 41.3 in Chapter 41 in this volume that

$$E[Y(n = 0, o = 1)] = \sum_m E[Y | O = 1, M = m] p(m | N = 0), \quad (38.12)$$

where the right-hand side (RHS) is the g-formula. Thus $E[Y(n = 0, o = 1)]$ is identified provided that the terms in the g-formula are functions of the distribution of the factials $p(A, N, O, M, Y)$. Since in the factual data now available there is no subject with $N = 0$ and $O = 1$ positivity fails and one might suppose that $E[Y(n = 0, o = 1)]$ is not identified, but in fact it is. To see this, note that the event $\{O = 1, M = m\}$ is equal to the event $\{A = 1, M = m\}$ and similarly the event $\{N = 0, M = m\}$ is equal to the event $\{A = 0, M = m\}$ owing to determinism. Thus,

21. Notice that here we show that the “cross-world” counterfactual $Y(a = 1, M(a = 0))$ defined in the DAG in Figure 38.3(a) is equal (as a random variable) to the “non-cross cross-world” counterfactual $Y(n = 0, o = 1)$ associated with the DAG in Figure 38.3(c); see Section 38.3.6 for further discussion.

by substituting these events we conclude that

$$E[Y(n = 0, o = 1)] = \sum_m E[Y | A = 1, M = m]p(m | A = 0), \quad (38.13)$$

which coincides with (one of the terms in) the mediation formula.²²

For a researcher following Pearl [2001], having at the outset assumed an NPSEM-IE associated with the DAG in Figure 38.3(a), the story involving N and O does not contribute to identification; rather, it served only to show that the PDE encodes a substantively important parameter.

However, from the FFRCISTG point of view, the story not only provides an interventional interpretation of the PDE but in addition makes the PDE identifiable with the mediation formula being the identifying formula. Furthermore, the story makes refutable the claim that the PDE is identified by the mediation formula. Specifically, when nicotine-free cigarettes become available, Pearl's claim can be tested by an intervention that forces a random sample of the population to smoke nicotine-free cigarettes; if the mean of Y under this intervention differs from the RHS of Equation (38.12), Pearl's claim is falsified. As this refers to an actual intervention, the variables (N, O) are not simply formal constructions. Without knowledge of the substantive meaning of N and O , the trial in which N is set to 0 and O is set to 1 is not possible, even in principle. See Robins and Richardson [2010] and Stensrud et al. [2020c] for discussion of substantive considerations regarding

22. Implications of Determinism: The FFRCISTG models associated with Figures 38.3(b) and (c) both imply the factuials (A, N, O, M, Y) factor with respect to the corresponding graph. Now in Figure 38.3(c) note (i) N is d-separated from Y given $\{M, O\}$ and (ii) O is d-separated from M given N which imply $N \perp\!\!\!\perp Y | O, M$ and $O \perp\!\!\!\perp M | N$, respectively. In contrast on Figure 38.3(b) neither of the above d-separations hold. Yet, since by the determinism $A = N = O$ as random variables, both independencies also hold for the DAG in Figure 38.3(b).

Furthermore, $E[Y(n = 0, o = 1)]$ would be identified by the g-formula:

$$\sum_m E[Y | N = 0, O = 1, M = m]p(m | N = 0, O = 1)$$

were the formula a function of the factual distribution. However, it is not; the event $\{N = 0, O = 1\}$ has probability zero due to determinism. Notwithstanding this, a naive application of the above independencies might lead one to conclude that this g-formula is equal to the RHS of (Equation 38.12) and thus is identified by (Equation 38.13). The error in this argument is that $O \perp\!\!\!\perp M | N$ does not imply $p(M = m | N = 0)$ equals $p(M = m | N = 0, O = 1)$ when the event $\{O = 1, N = 0\}$ has probability zero, since the latter is not well-defined.

Note that, in the presence of determinism, we have two DAGs with different adjacencies that represent the same set of factual distributions. The counterfactuals corresponding to the DAGs in Figures 38.3(b) and (c) do not represent the same set of counterfactual distributions; see Footnote 27.

whether variables N and O exist that satisfy the no direct effect assumptions of Figure 38.3(c).²³

Remark 38.1 It should be noted that, in the following sense, it is sufficient to find one of the variables N or O : Specifically, if we have a well-defined intervention N , satisfying:

- (n1) $N(a) = a$ so that $N = A$ in the observed data;
- (n2) N has no direct effect on Y relative to A and M , so $Y(a, n, m) = Y(a, m)$;
- (n3) A has no direct effect on M relative to N so $M(a, n) = M(n)$,

then A will satisfy the conditions for O ; see Figure 38.3 (d). Conversely, if there is a well-defined intervention O such that:

- (o1) $O(a) = a$ so that $O = A$ in the observed data;
- (o2) A has no direct effect on Y relative to O and M , so $Y(a, o, m) = Y(o, m)$;
- (o3) O has no direct effect on M relative to A so $M(a, o) = M(a)$,

then A will satisfy the conditions for N ; see Figure 38.3(e).

We use all three variables (A, N, O) in our subsequent development since this choice is symmetric in N and O , covers both cases and more closely aligns with the original motivating nicotine intervention.

Lastly, note that, as in the ivermectin example, the existence of the interventions N, O satisfying the no direct effect conditions do not imply that the mediation formula identifies the PDE because confounding between M and Y may still be present.²⁴

23. One can trivially construct artificial variables $N^* \equiv A$ and $O^* \equiv A$ such that the (degenerate) joint distribution of the factials $p(A, N^*, O^*, M, Y)$ will factorize according to the DAG in Figure 38.3(c) and thus satisfy $M \perp O^*, A | N^*$ and $Y \perp N^* | A, O^*, M$. However, these latter independencies are tautologies owing to determinism and thus do not establish, for example, $Y(o, m) \perp M(n)$ as required by the FFRCISTG associated with Figure 38.3(c).

24. An attentive reader might wonder how it is that the first expression in the mediation formula fails to identify $P(Y(n, o))$ in the ivermectin example (with A as “ N ” and S as “ O ”; see Remark 38.1 above). As noted earlier, the counterfactual variables $A, M(a), Y(a, m)$ follow the conditional independencies implied by the FFRCISTG model associated with the DAG in Figure 38.3(a) (in which there is no bi-directed arc between M and Y); see the end of Section 38.2.5. Under this FFRCISTG model, the absence of the $M \leftrightarrow Y$ edge implies that there is no confounding between M and Y that is *detectable* via interventions on A and M , since $p(Y(a, m)) = p(Y | A = a, M = m)$.

However, in the ivermectin example, the expanded set of counterfactual variables $A, O, M(a), Y(a, o)$ do *not* follow the FFRCISTG model corresponding to the DAG in Figure 38.3(e). To see this, consider performing, in addition to the randomized trial where $a = 1$ and $s = 1$, an intervention setting A to 0 and $O \equiv S$ to 1, which corresponds to an observational study but with clinics and immunosuppressants. Notice that the confounding variable U in Figure 38.1(a) now becomes detectable in the following sense: If U were not present in Figure 38.13(a) then

38.3.2 Direct and Indirect Effects via the Expanded Graph

In this section we formally introduce the interventionist theory of mediation first introduced in [Robins and Richardson \[2010\]](#) and greatly generalized herein. We do so by continuing with the Nicotine Example. Recall that N and O were substantively meaningful variables, and the goal was to use data from a smoking cessation trial to estimate $E[Y(n = 0, o = 1)]$, the incidence of MI through year 2 if all smokers were to change to nicotine-free cigarettes at one year. As noted above, this policy intervention was used by Pearl to motivate consideration of the PDE. However, given the public health importance of the policy question, one could instead focus directly on estimating the effect of the proposed substantive intervention given data on (A, M, Y) without regard to whether it is equal to the PDE.

In fact, there are many situations where mediation analysis is applied, in which interventions on the putative mediator M are not well-defined. Consequently, substantive researchers may not wish to make reference to the corresponding counterfactuals,²⁵ regardless of whether they may be formally constructed. For such researchers, the PDE parameter may not be substantively meaningful. Fortunately, the interventionist theory described herein, in contrast to Pearl's approach, does not require reference to counterfactuals indexed by m , such as $Y(a, m)$.²⁶

As noted, the interventionist theory only requires that N , O , and interventions on them are substantively meaningful. This is a major advantage since it makes it straightforward to discuss with subject matter experts, for example, physicians, experiments that would shed light on causal pathways; this is a property not shared by the PDE.

Up to this point we have motivated our interventionist theory, based on the expanded graph, as providing an empiricist foundation for the existing mediation theory that is based on cross-world (nested) counterfactuals. However, the interventionist theory can be viewed as autonomous,²⁷ providing a self-contained framework for discussing mediation without reference to cross-world (i.e., nested)

$p(Y(a = 0, s = 1) | M(a = 0)) = p(Y(a = 1, s = 1) | M(a = 1))$. This follows from Rule 3 of the calculus [[Malinsky et al. 2019](#), chapter 41] applied to the SWIG $\mathcal{G}(a, s)$ derived from Figure 38.3(e) after replacing " O " with S . However, we show in Appendix 38.A.3 that in this example if U is present then this equality does not hold in general. Consequently, the latent projection over the variables $A, O \equiv S, M, Y$ includes an $M \leftrightarrow Y$ edge; see also the last paragraph of Section 38.2.6.

25. We regard the existence of interventions as necessary for counterfactuals to be well-defined, but [Pearl \[2018, 2019\]](#) and others may take a different view.

26. The $CDE(m)$ and PDE are defined in terms of such counterfactuals.

27. Following [Wittgenstein \[1922, section 6.54\]](#), we may view the theory based on nested counterfactuals as a "ladder" that we climbed to reach our interventionist theory, and now having done so, we may choose to "kick it away."

counterfactuals. In this section, we adopt this viewpoint. However, we will see that we can prove the two theories are tightly coupled in certain settings; see Section 38.3.6.²⁸

Concretely, consider the DAG associated with Figure 38.3(b). Unlike the model associated with Figure 38.3(a), which involves counterfactuals $M(a)$ and $Y(a)$, the expanded FFRCISTG model associated with Figure 38.3(b) involves counterfactuals $M(n, o)$ and $Y(n, o)$. Taking the interventionist view as primitive, in what follows we will discuss counterfactuals such as $Y(n = 0, o = 1)$ that, without further assumptions, are defined solely within this larger expanded model.

Identification of Four Arms from Two

For the purposes of our development, consider the following three datasets all derived from the same distribution p over the one-step-ahead counterfactuals in the FFRCISTG model associated with the graph in Figure 38.3(b):

- (i) The original observed data from the trial in which A was randomized, namely A, M, Y ;
- (ii) Data from a putative four-arm (N, O) randomized trial; the data in each arm $(n, o) \in \{0, 1\}^2$ corresponds to $M(n, o), Y(n, o)$;
- (iii) A dataset obtained from the four-arm (N, O) trial (ii) by restricting to the two arms in which $n = o$.

Note that in dataset (i) among people with $A = a$ we observe $N = O = a$, owing to determinism; hence, we observe $M(n = a, o = a)$ and $Y(n = a, o = a)$ on this

28. In related work, Lok [2016] has developed an interventional approach to mediation that also does not require interventions on the mediator in order to be well-defined. Lok introduces a notion of an ‘organic intervention’ ($I = 1$) that is required to satisfy certain conditions. Lok then defines notions of direct and indirect effects in term of such organic interventions. Our interventional definitions introduced here are similar in spirit to Lok’s conditions.

However, in order to capture certain aspects of the concept of direct and indirect effects, we, unlike Lok, also require that the variables $(N$ and $O)$ defining our additional interventions (e.g., on N) be equal to A in the observed data. Among other things, this ensures that $N(a) = O(a) = a$ and hence $M(a) = M(n = a, o = a)$ and $Y(a) = Y(n = a, o = a)$.

In contrast, an organic intervention could change the mechanism by which the mediator is produced so that the relevant counterfactual random variables for the mediator under the organic intervention ($M(a = 0, i = 1)$) do not correspond to those in the absence of the organic intervention ($M(a = 1)$), although they have the same distribution. See Robins [2003], section 3 for additional discussion in terms of blocking paths.

event.²⁹ By randomization of A , it follows that for $a \in \{0, 1\}$,

$$\begin{aligned} p(M = m, Y = y | A = a) &= p(M(n = a, o = a) = m, Y(n = a, o = a) = y | A = a) \\ &= p(M(n = a, o = a) = m, Y(n = a, o = a) = y) \end{aligned}$$

which is the distribution of individuals in dataset (iii). We conclude that the distribution of the data in (iii) is identified from the observed data (i). Thus our goal becomes the identification of $E[Y(n, o)]$ for $n \neq o$ from data on $M(n, o)$ and $Y(n, o)$ in the two arms with $n = o$. Motivated by the nomenclature of [Stensrud et al. \[2020c\]](#), when this identification is possible we will say that the effects of N and O on M and Y are *separable*. The following proposition provides sufficient conditions.

Identifying the Results of a Future Four-arm Study from a Current Two-arm Study

Proposition 38.1 If for some $x \in \{0, 1\}$ the following two conditions hold:

$$p(M(n = x, o = 0) = m) = p(M(n = x, o = 1) = m), \quad (38.14)$$

$$\begin{aligned} p(Y(n = 1, o = x^*) = y | M(n = 1, o = x^*) = m) \\ = p(Y(n = 0, o = x^*) = y | M(n = 0, o = x^*) = m), \end{aligned} \quad (38.15)$$

where $x^* = 1 - x$, then:

$$\begin{aligned} p(M(n = x, o = x^*) = m, Y(n = x, o = x^*) = y) \\ = p(Y(n = x^*, o = x^*) = y | M(n = x^*, o = x^*) = m) p(M(n = x, o = x) = m). \end{aligned} \quad (38.16)$$

Note that by consistency and randomization of A in dataset (i), the RHS of Equation (38.16) when summed over m is simply the first expression in the mediation formula (38.7). Hence under Equations (38.15) and (38.14) $E[Y(n, o)]$ is identified from data set (iii) by this expression in the mediation formula.

Proof.

$$\begin{aligned} p(M(n = x, o = x^*), Y(n = x, o = x^*)) \\ = p(Y(n = x, o = x^*) | M(n = x, o = x^*)) p(M(n = x, o = x^*)) \\ = p(Y(n = x^*, o = x^*) | M(n = x^*, o = x^*)) p(M(n = x, o = x)). \end{aligned}$$

29. Note that the assumption that $p(M(a)) = p(M(n = a, o = a))$ and $p(Y(a) | M(a)) = p(Y(n = a, o = a) | M(n = a))$. This assumption is subject to empirical test by examining whether the distribution from (i) and (iii) are the same. This corresponds to the six-arm trial described by [Stensrud et al. \[2020b\]](#). The distribution of (i) and (iii) could differ when, for example, the treatment A contains additional sub-components that are present in neither N nor O .

The constraints in Equations (38.15) and (38.14) are implied by the SWIG given in Figure 38.4(b) with treatments n and o over the random variables N , O , $M(n, o)$ and $Y(n, o)$.³⁰ Note that $Y(n, o)$ is d-separated from the fixed node n given $M(n, o)$, which implies under the SWIG global Markov property that the distribution $p(Y(n, o) | M(n, o))$ does not depend on n , which is equivalent to Equation (38.15).

Similarly, since $M(n, o)$ is d-separated from the fixed node o , $p(M(n, o))$ does not depend on o . Thus, we see that the constraints (38.14) are implied by the FFRCISTG model.³¹ ■

Thus, the identifiability result in Proposition 38.1 follows from the fact that in the SWIG in Figure 38.4(b), there is no variable whose conditional distributional distribution, given its (random) parents, depends on both n and o .

Proposition 38.1 above, and indeed all the results in the remainder of this subsection, holds when counterfactuals indexed by the mediator m are not well-defined. Recall that Robins [1986] and Robins and Richardson [2010] develop an FFRCISTG model in which only interventions on a subset of variables are considered well-defined.

Remarks:

1. The reader may wonder why in this SWIG we have labeled M with (n, o) , rather than simply with (n) . This is to emphasize that in this subsection our results do not require the assumption that missing arrows on a SWIG imply the absence of the associated direct effect for all individuals. Rather in this sub-section we only impose the weaker assumption that any SWIG



Figure 38.4 (a, b) SWIGs derived from the corresponding expanded DAGs G^{ex} shown in Figures 38.3(b) and (c), respectively. In the SWIG shown in (b), the mediator is labeled $M(n, o)$ to indicate that this is a population FFRCISTG for G^{ex} , which does not assume the absence of individual-level direct effects.

30. Thus, although, as noted previously in Footnote 21, under determinism, the DAGs in Figure 38.3(b) and (c) imply the same conditional independence relations on $p(A, M, Y, N, O)$, they lead to different counterfactual models since the constraints (Equation 38.15) and (Equation 38.14) are implied by the SWIG in Figure 38.4(b), but not the SWIG in Figure 38.4(a).

31. As noted in remark 38.1, in some settings A can play the role of N or O .

is a “population causal graph.” A population causal graph [Richardson and Robins 2013, section 7] assumes the distribution of the variables on the graph factor according to the graph, but does not impose the assumption that a missing arrow implies no individual level effects. Thus the variable $M(n, o)$ need not equal the variable $M(n) = M(n, O)$ and thus $M(n, o)$ cannot be labeled as $M(n)$. That is, in the underlying FFRCISTG model associated with the SWIG, the one-step-ahead counterfactuals $M(n, o)$ depend on both n and o . In other words, this population FFRCISTG contains the counterfactual variables present in the SWIG shown in Figure 38.4(a) that results from splitting N and O in the graph shown in Figure 38.3(b). It does not correspond to an NPSEM associated with the graph in Figure 38.3(c) since that NPSEM assumes well-defined counterfactuals intervening on M and also assumes that $M(n, o) = M(n)$; see also Footnotes 17 and 22. In particular, under the SWIG derived from the population graph, the constraints (Equation 38.15) and (Equation 38.14) correspond to the absence of the edges $n \rightarrow Y(n, o)$ and $o \rightarrow M(n, o)$, respectively.

2. Since Equations (38.15) and (38.14) are restrictions on the distribution of the counterfactuals in Figure 38.4(b), there exist consistent tests of the conditions (Equation 38.15) and (Equation 38.14) given the data from (ii). These conditions cannot be tested given only the data (iii) [or equivalently (i)].³²
3. We have seen that, under the FFRCISTG associated with Figure 38.4(b), the distribution of $Y(n, o)$ for all four arms is identified from the two arms in which $n = o$; thus the structure of this SWIG is sufficient for this identification. However, this structure is also “necessary” in that it is the only population SWIG over $M(n, o)$ and $Y(n, o)$ where this identification is possible.³³ To see this, first note that if $Y(n, o)$ depends on both n and o then n and o are both ancestors of Y . By Proposition 38.2 below, if n and o are both parents of Y , then identifiability fails to hold. Given that we only have one other measured variable then this implies that we must have one fixed node that is a parent of M (and M in turn a parent of Y); the other fixed node is then a parent of Y . If any other edges are present between $\{n, o\}$ and $\{M(n, o), Y(n, o)\}$, then again by Proposition 38.2 the conditional distributions will not be identifiable. Likewise, if there is an unmeasured confounder between M and Y , then $Y(n, o)$ will not be d-separated from n given $M(n, o)$.

32. We note that consistent tests of individual level no direct effect conditions such as $Y(n, o) = Y(o)$ do not exist since it is possible that $E[Y(n, o)] = E[Y(n, o')]$ and yet $Y(n, o) \neq Y(n, o')$ (as random variables); but see also footnote 29.

33. Here we are ignoring the structure relating the random variables N and O .

We have the following more general result.

Proposition 38.2 Assume the distribution of the variables on an unexpanded DAG \mathcal{G} is positive. Under an FFRCISTG corresponding to the population SWIG $\mathcal{G}^{ex}(n, o)$ there is no vertex that has both n and o as parents if and only if the joint distributions $p(V(n = x, o = x^*))$ for $x \neq x^*$ are identified from the counterfactual distributions $p(V(n = x, o = x))$. Further, since $P(V(n = x, o = x)) = P(V(a = x))$, also from the distribution of the variables in \mathcal{G} .³⁴

Proof. Consider the g-formula of proposition 3 in Chapter 41 applied to the graph \mathcal{G}^{ex} under an intervention on n and o . This formula is a function of the joint distribution of the observables if and only if none of the terms in the g-formula have both N and O in the conditioning event. Note that each term in the g-formula is the conditional distribution of a variable given its parents on the population SWIG $\mathcal{G}^{ex}(n, o)$. The requirement here that there is no vertex that is a child of both n and o is directly analogous to the “no recanting witness” condition in the theory developed by Avin et al. [2005]. ■

Consider the following examples from Robins and Richardson [2010]. Suppose our original causal graph \mathcal{G} in Figure 38.3(a) for the cigarette cessation trial was incorrect and the correct causal graph is shown in Figure 38.5(a). There exist three possible (N, O) elaborations of this graph, which are shown in Figure 38.6. These represent different causal theories about the causal effect of the treatment variables N, O on L, M , and Y . Figure 38.7 shows the corresponding population SWIGs. Under the SWIGs in Figure 38.7(a) and (b) the distributions of the four arms $Y(n, o)$ are identified given the distributions $Y(n = x, o = x)$:

$$\begin{aligned} p(Y(x, x^*)) &= \sum_{m,l} p(Y(x^*, x^*) | M(x^*, x^*), L(x^*, x^*)) p(M(x, x) | L(x, x)) p(L(x, x)) \\ &= \sum_{m,l} p(Y | m, l, a = x^*) p(m | l, a = x) p(l | a = x) \end{aligned} \quad (38.17)$$

and

$$\begin{aligned} p(Y(x, x^*)) &= \sum_{m,l} p(Y(x^*, x^*) | M(x^*, x^*), L(x^*, x^*)) p(M(x, x) | L(x, x)) p(L(x^*, x^*)) \\ &= \sum_{m,l} p(Y | m, l, a = x^*) p(m | l, a = x) p(l | a = x^*) \end{aligned} \quad (38.18)$$

respectively, where here we are using $Y(i, j)$ to denote $Y(n = i, o = j)$. Note that the identifying formulae are different. See Stensrud et al. [2020a] for generalizations of these results.

34. Note that the result here holds because the DAG \mathcal{G} here does not contain hidden variables.

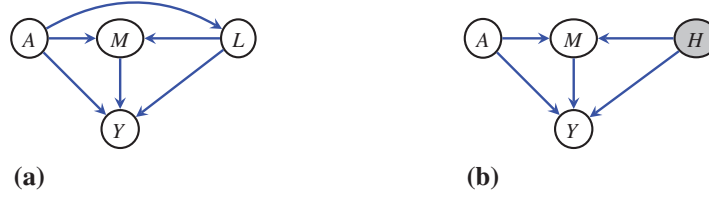


Figure 38.5 (a) DAG containing an observed common cause L of the mediator M and outcome Y that is also caused by A ; (b) DAG containing an unobserved common cause H of the mediator M and outcome Y .

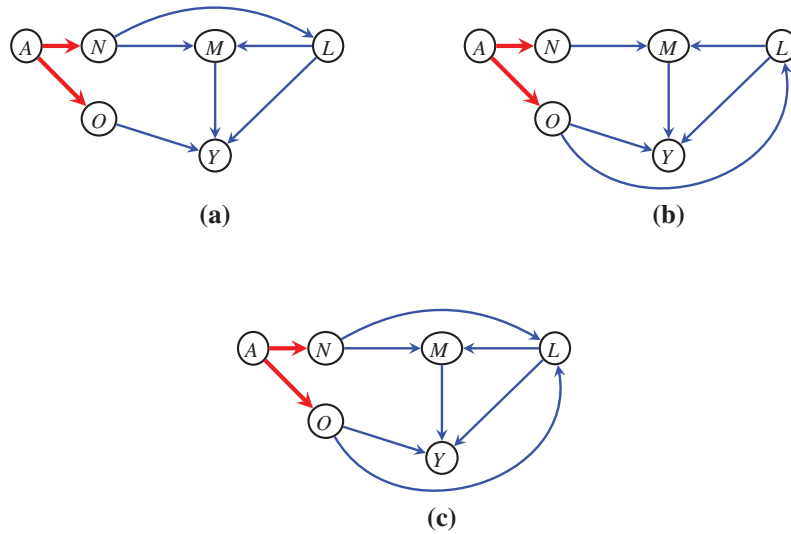


Figure 38.6 Elaborations of the graph in Figure 38.5(a), with additional edges. These represent different causal theories about the causal effect of the treatment variables N, O on L, M, Y . As before, the thicker red edges indicate deterministic relations.

In contrast, under the SWIG in Figure 38.7(c), $p(Y(n, o))$ for $n \neq o$ is not identified from the data on the two arms with $n = o$ (equivalently the observed data) because in $\mathcal{G}^{ex} L$ has both N and O as parents; hence the term $p(l | n = x, o = x^*)$ in the g-formula for $p(Y(x, x^*))$ is not a function of the observed data.

Given that N and O are real interventions, at most one of the expanded causal graphs shown in Figure 38.6 can represent the true causal structure. If, in the future, we obtain data from a four arm (N, O) trial we can test between the three competing theories associated with these expanded graphs.³⁵

35. Note however, that if the results from the four-arm trial do not correspond to the identifying formulae obtained from either Figure 38.7(a) or (b), then although it is possible that the DAG in

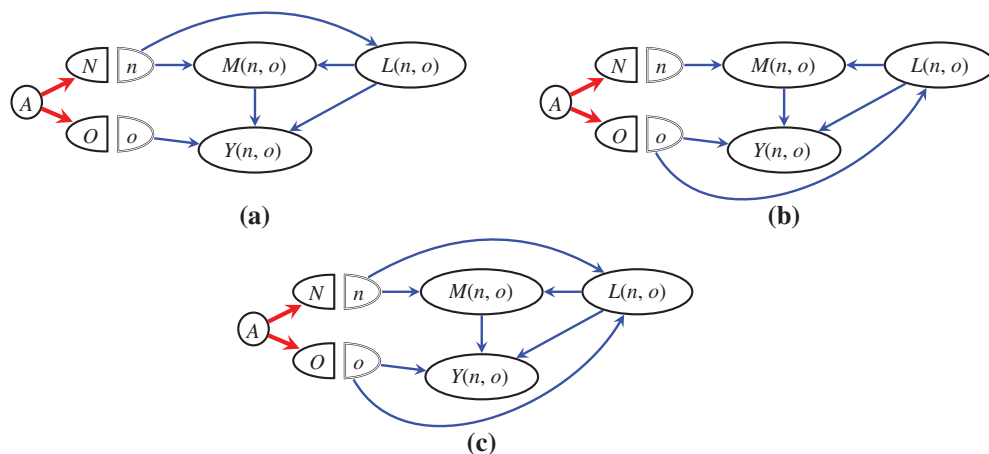


Figure 38.7 Three different population SWIGs associated with the three expanded DAGs shown in Figure 38.6. Under the SWIGs (a), (b) the effects of N and O on Y are separable so that the distribution of $Y(n, o)$ in the four arms are identified given the distribution of $Y(n = x, o = x)$ although the identification formulae differ; (c) A SWIG under which the four arms $Y(n, o)$ are not identified from the two arms $Y(n = x, o = x)$.

38.3.3 Expanded Graphs for a Single Treatment

We formally define expanded graphs as follows:

Given a DAG \mathcal{G} with a single treatment variable A , an *expanded graph* \mathcal{G}^{ex} for A is a DAG constructed by first adding a set of new variables $\{A^{(1)}, \dots, A^{(p)}\}$ corresponding to a decomposition of the treatment A into p separate components (proposed by the investigator); every variable $A^{(i)}$ is a child of A with the same state space and $A^{(i)}(a) = a$, but A has no other children in \mathcal{G}^{ex} ; each child C_j of A in \mathcal{G} has in \mathcal{G}^{ex} a subset of $\{A^{(1)}, \dots, A^{(p)}\}$ as its set of parents.

Lemma 38.1 Assume the distribution of the variables on an unexpanded DAG \mathcal{G} is positive. Under an FFRCISTG corresponding to an expanded (population) graph \mathcal{G}^{ex} for treatment A , the intervention distribution $p(V(a^{(1)} = x^{(1)}, \dots, a^{(p)} = x^{(p)}))$ is identified by the g-formula applied to \mathcal{G}^{ex} from the data on \mathcal{G} if for every child C_j of A in \mathcal{G} the set of parents of C_j in \mathcal{G}^{ex} that are components of A take the same value.³⁶

Figure 38.7(c) holds, it is also possible that the true graph could correspond to Figure 38.7(a) or (b) with an added unmeasured confounder between any pair of the variables L, M and Y .

36. More formally, we require that for all $C_j \in \text{ch}_{\mathcal{G}}(A)$, if $A^{(k)}, A^{(l)} \in \text{pa}_{\mathcal{G}^{ex}}(C_j) \cap \{A^{(1)}, \dots, A^{(p)}\}$ then $x^{(k)} = x^{(l)}$.

Proof. As in the proof of Proposition 38.2, consider the g-formula for $p(V(a^{(1)} = x^{(1)}, \dots, a^{(p)} = x^{(p)}))$ of proposition 3 in Chapter 41 applied to the graph \mathcal{G}^{ex} . The g-formula will be a function of the joint distribution of the observables since each term in the g-formula always conditions on a single value of A . ■

Robins and Richardson [2010, section 6.2] also described the special case in which each child C_j of A in \mathcal{G} has exactly one component $A^{(j)}$ as a parent.³⁷ We will refer to this as the *edge expanded graph for A* , which we often denote as \mathcal{G}^{edge} . In this case $p = |ch_{\mathcal{G}}(A)|$ and in this special case \mathcal{G}^{ex} corresponds to the graph formed from \mathcal{G} by replacing each edge $A \rightarrow C_j$ with $A \rightarrow A^{(j)} \rightarrow C_j$. Note that \mathcal{G}^{edge} is unique.³⁸ See Figure 38.9 for an example.

Corollary 38.1 Under the assumptions of Lemma 38.1, if \mathcal{G}^{ex} is the edge expanded graph \mathcal{G}^{edge} for A then for all assignments $x^{(1)}, \dots, x^{(p)}$ $p(V(a^{(1)} = x^{(1)}, \dots, a^{(p)} = x^{(p)}))$ is identified from the data on \mathcal{G} .

When the conditions of this corollary hold, we will say, following the nomenclature in Stensrud et al. [2020c], that the treatment components $\{A^{(1)}, \dots, A^{(p)}\}$ have *separable effects*.

38.3.4 On the Substantive Relationship between Different \mathcal{G}^{ex} Graphs and \mathcal{G}^{edge}

In the context of the smoking cessation trial recall that the two expanded graphs in Figure 38.6(a) and (b) led to different identifying formulae for $P(Y(n = x, o = x^*))$ given, respectively, by Equations (38.17) and (38.18). The identifying formulae also arise in the context of the graph \mathcal{G}^{edge} shown in Figure 38.9. Specifically, $p(Y(a^{(1)} = x, a^{(2)} = x, a^{(3)} = x^*))$ and $p(Y(a^{(1)} = x^*, a^{(2)} = x, a^{(3)} = x^*))$ are identified by Equations (38.17) and (38.18), respectively. The FFRCISTG models associated with the expanded graphs shown in Figure 38.6(a) and (b) correspond to distinct mutually exclusive causal structures, which lead to different identifying formulae for the distribution of the counterfactual $Y(n = x, o = x^*)$ in an arm of the four arm (N, O) trial in which one intervenes to set $n = x$, and $o = x^*$. However, given the FFRCISTG model corresponding to the graph in Figure 38.9, we are able to interpret the identifying expressions (38.17) and (38.18) as identifying two *different* interventions on $A^{(1)}$, $A^{(2)}$, and $A^{(3)}$ on a single graph \mathcal{G}^{edge} .

37. More formally, we have that for all children C_j of A , $pa_{\mathcal{G}^{ex}}(C_j) \cap \{A^{(1)}, \dots, A^{(p)}\} = \{A^{(j)}\}$.

38. Since the expanded graph \mathcal{G}^{edge} for A postulates a separate component of treatment corresponding to each child of A , such a graph will be unlikely to represent the substantive understanding of an investigator when p is large.

Robins and Richardson [2010] note that the above may seem to, but do not, contradict one another. Recall that N and O represent the substantive variables recording the presence or absence of nicotine and all other cigarette components.

Suppose we further divide the other cigarette components (O) into Tar (T) and cigarette components other than tar and nicotine (O^*). Thus, substantively setting O to a value corresponds to setting both O^* and T to that value. Furthermore, the graph \mathcal{G}^{edge} in Figure 38.9 being an FFRCISTG implies that the graph \mathcal{G}^{ex} in Figure 38.6(b) formed by (re)combining O^* and T is an FFRCISTG. Thus an intervention setting ($O = x, N = x^*$) on Figure 38.6(b) substantively corresponds to the intervention on the graph \mathcal{G}^{edge} in Figure 38.9 with $O^* = A^{(3)} = x, T = A^{(1)} = x, N = A^{(2)} = x^*$. Thus these interventions in this \mathcal{G}^{ex} and in \mathcal{G}^{edge} give the same identifying formula (38.18).

Given we have locked in the substantive interpretation of the $A^{(j)}$ in Figure 38.9, the intervention $O^* \equiv A^{(3)} = x, T \equiv A^{(1)} = x^*, N \equiv A^{(2)} = x^*$ on Figure 38.9 corresponds to the intervention in which N and T are set to the same value and thus does not represent a joint intervention on the substantive variables N and O (since $O = O^* = T$ as random variables in the observed data). However, the identifying formula for this intervention if Figure 38.9 were the causal graph happens to have the same identifying formula (38.17) as the intervention setting $O = x, N = x^*$ if Figure 38.6(a) were the causal graph. This might seem surprising since, under the substantive meanings of the components ($A^{(1)}, A^{(2)}, A^{(3)}$), specifically, $A^{(2)} \equiv N$, Figure 38.9 is compatible with Figure 38.6(b) and *not* Figure 38.6(a). However, it is *not* surprising from a formal perspective, if we notice that by considering a DAG with the same *structure* as Figure 38.6(a), but in which N is replaced by a variable $N^\dagger \equiv N \times T$ indicating the presence of both Tar *and* Nicotine, then we may represent the intervention that sets $O^* = x, N = T = x^*$ via an intervention on N^\dagger and O^* .³⁹

If, instead of dividing O , one divides Nicotine, N , into sub-components corresponding to two different isotopes and re-defines the variables in Figure 38.9 as $A^{(1)} \equiv \text{Nicotine Isotope 1}, A^{(2)} \equiv \text{Nicotine Isotope 2}, A^{(3)} \equiv O$, then the mirror image of the above holds. Specifically, Figure 38.9 is compatible with Figure 38.6(a),⁴⁰ and not Figure 38.6(b).

Note, however, that there is no way to re-define $A^{(1)}, A^{(2)}$, and $A^{(3)}$ such that Figure 38.9 is compatible with Figure 38.6(c). Specifically, an intervention setting $N = x$ and $O = x^* \neq x$ cannot be represented via an intervention on $(A^{(1)}, A^{(2)}, A^{(3)})$, since in Figure 38.6(c) L has two parents N and O while in Figure 38.6(c) L has

39. Though the graph constructed in this way has the same topology as Figure 38.6(a), it represents a different substantive hypothesis since the component T of N^\dagger is a parent of L and not N itself.

40. Figure 38.6(a) with the variables N and O (not N^\dagger and O).

only one. Of course, this must be the case because the intervention on N and O in Figure 38.6(c) is not identified from the observed data. In contrast, by Corollary 38.1 any intervention on $A^{(1)}, A^{(2)}, A^{(3)}$ in the graph shown in Figure 38.9 is identified.

Lastly, note that \mathcal{G}^{edge} assumes that $A^{(1)}, A^{(2)}$, and $A^{(3)}$ each directly affect only L, M , and Y , respectively; this hypothesis could, in principle, be tested if one were to perform an eight-arm $(A^{(1)}, A^{(2)}, A^{(3)})$ trial.

In general, if \mathcal{G}^{ex} is an FFRCISTG with separable (hence, identified) effects and further \mathcal{G}^{edge} is an FFRCISTG, then the counterfactual variables in \mathcal{G}^{ex} may be obtained from those in \mathcal{G}^{edge} by imposing the equality $a^{(i)} = a^{(j)}$ whenever the corresponding children C_i and C_j of A in \mathcal{G} share a common parent in \mathcal{G}^{ex} . Note that this is directly analogous to the way in which the counterfactual variables $V_i(a)$ in $\mathcal{G}(a)$ are equal to the counterfactuals $V_i(n = a, o = a)$ present in $\mathcal{G}(n, o)$.

38.3.5 Generalizations

The foregoing development may be further generalized in several ways:

- (a) Rather than having data from a single treatment variable A , we may consider a study in which there were multiple treatment variables $\{A_1, \dots, A_k\}$. In this setting it may be of interest to attempt to identify the distribution $V(n_1, o_1, \dots, n_k, o_k)$ of a hypothetical future study where N_i and O_i are components of A_i such that in the original study $A_i = O_i = N_i$. See Figure 38.10. More generally each A_i may have p_i components. The natural generalization of Lemma 38.1 holds.⁴¹
- (b) We may consider a setting in which some variables (H) in the underlying causal DAG $\mathcal{G}(V \cup H)$ are not observed; though variables we intervene on are observed, so $A \subseteq V$.

Here, we proceed in two steps. In the first step, we check identification of a standard interventional distribution. Specifically, we construct an ADMG \mathcal{G}^{ex} containing the variables $\{N_1, O_1, \dots, N_k, O_k\}$, such that N_i and O_i have only A_i as a parent; the only edges with an arrowhead at N_i and O_i are of the form $A_i \rightarrow$ while the only edges out are of the form $\rightarrow C$ where C is a child of A in \mathcal{G} . We then apply the extended ID algorithm described in Chapter 41 of this volume to first determine whether $p(V(n_1, o_1, \dots, n_k, o_k))$ would be identified given a positive distribution $p(V \cup \{N_1, O_1, \dots, N_k, O_k\})$ over the observed variables and the treatment components.⁴²

41. That is, we have identification if, for each i and each child C of A_i in \mathcal{G} , the subset of the p_i components of A_i that are parents of C in the expanded graph \mathcal{G}^{ex} take the same value.

42. This would correspond to an observational study where the variables $V \cup \{N_1, O_1, \dots, N_k, O_k\}$ are observed in a population in which N_i and O_i are no longer deterministic functions of A_i , but for which all other one-step-ahead counterfactuals remain the same; variables in H are not observed.

In the second step, we check whether the identification would hold under the weaker conditions where we only have access to a positive distribution on $p(V)$. This corresponds to (i) making sure that identification of every term in the identifying formula given by the ID Algorithm via the inductive application of Proposition 41.5 from Chapter 41 in this volume ensures that the splitting operation is applied to any $A_i \in A$ before any N_i or O_i (this ensures that the positivity requirement for Proposition 41.5 is met), and (ii) confirming that for every (N_i, O_i) , in the identifying formula given by the ID Algorithm⁴³ there is no district D that has both N_i and O_i as parents; Shpitser [2013] terms such districts, which violate this condition, *recanting districts*.⁴⁴

A simple example of such a structure arises when there is an unobserved common cause of M and Y , as shown in Figure 38.5(b): though the four distributions $E[Y(n, o)]$ are identified given a four-arm (N, O) trial, the distributions for which $n \neq o$ are not identified solely given data on (A, M, Y) . This is because $M \leftarrow H \rightarrow Y$ forms a district and both N and O are parents of this district, hence it is recanting.

If the two-step procedure yields identification, the resulting functional structurally resembles the functional obtained from the ID algorithm, except each term in the functional that depends on treatments is evaluated at its own treatment value, corresponding to either n_i or o_i (but never both at once).

See Section 38.4 below for a general method of addressing all complications above simultaneously.

38.3.6 Identification of Cross-world Nested Counterfactuals of DAG \mathcal{G} under an FFRCISTG Model for its Expanded Graph \mathcal{G}^{ex}

From Section 38.3.2 to this point, we only studied the distribution of counterfactuals associated with interventions on (N_i, O_i) or, more generally, $(A^{(1)}, \dots, A^{(p)})$; no other counterfactuals associated with an expanded graph \mathcal{G}^{ex} were mentioned. As argued in Section 38.3.2, for most purposes these counterfactuals constitute an adequate basis for formulating contrasts relating to the mediation of effects. However, since the prior literature on mediation has been formulated in terms of

43. See Equation (41.21) and subsequent discussion in Chapter 41 in this volume.

44. Note that in the special case where $D = \{V_i\}$ is a singleton, then D will be a recanting district in \mathcal{G} if and only if N_i and O_i have a common child in \mathcal{G}^{ex} . In this case V_i is a “recanting witness” as defined by Avin et al. [2005]; see also Section 38.4. Thus, when no hidden variables exist, the first step in (b) always succeeds; hence, as implied by Lemma 38.1, identification fails if and only if there exists an N_i and O_i that have a common child.

cross-world nested counterfactuals associated with the original (unexpanded) DAG \mathcal{G} , we return to our earlier discussion.

The PDE

Recall that in Section 38.1 we related the PDE from the NPSEM associated with Figure 38.3(a) to a four-arm (N, O) trial under the FFRCISTG associated with the SWIG $\mathcal{G}^{ex}(n, o)$ shown in Figure 38.4(b). This may be broken down into two steps:

- (1) Show that $E[Y(n = 0, o = 1)]$ was identified from data on $p(A, M, Y)$ via Equation (38.12). This step only required that \mathcal{G}^{ex} was a population FFRCISTG.⁴⁵ That is, this identification follows from the deterministic relationship $N(a) = O(a) = a$ together with the population level conditions (Equation 38.15) and (Equation 38.14), without requiring that counterfactuals for interventions on M be well-defined.
- (2) Next show that $Y(a = 1, M(a = 0)) = Y(n = 0, o = 1)$ holds under the individual level no direct effect assumptions encoded in the NPSEM associated with \mathcal{G}^{ex} in Figure 38.8. Note that this step does not require that Figure 38.3(c) is an FFRCISTG, only that it be an NPSEM associated with \mathcal{G}^{ex} .

In more detail, the NPSEM associated with \mathcal{G}^{ex} implies the following:

- (i) $M(n) = M(a, n, o)$;
- (ii) $Y(o, m) = Y(a, n, o, m)$.

Under conditions (i) and (ii) we have that:

$$\begin{aligned}
 M(a = 0) &= M(N(a = 0)) = M(n = 0), \\
 Y(a = 1, m) &= Y(O(a = 1), m) = Y(o = 1, m), \\
 Y(a = 1, M(a = 0)) &= Y(o = 1, M(n = 0)) = Y(n = 0, o = 1).
 \end{aligned}
 \tag{38.19}$$

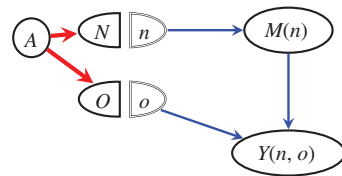


Figure 38.8 A SWIG derived from the expanded DAG \mathcal{G}^{ex} in Figure 38.3(c), under the assumption that the counterfactual variables obey the NPSEM associated with \mathcal{G}^{ex} . Consequently, in contrast to Figure 38.4(b), the graph contains $M(n)$ rather than $M(n, o)$.

45. Thus, this step does not require that the counterfactual variables follow an NPSEM associated with \mathcal{G}^{ex} .

Consequently, under the assumptions (i) and (ii) it follows that:

$$P(Y(a = 1, M(a = 0))) = P(Y(n = 0, o = 1)). \quad (38.20)$$

Remarks:

1. If $N(a) = O(a) = a$ and conditions (i) and (ii) hold, then $PDE = E[Y(n = 0, o = 1)] - E[Y(n = 0, o = 0)]$ even if conditions (38.15) and (38.14) fail, for example due to (M, Y) confounding. This is the situation discussed in Section 38.1 where data from the four-arm (N, O) trial makes it possible to estimate the PDE and hence determine whether it equals the mediation formula.⁴⁶
2. The conditions (38.15) and (38.14) alone, that is, without (i) and (ii) above, are not sufficient to identify the PDE,⁴⁷
- 3.(i) If condition (i) fails, so that O has a (population) direct effect on M (relative to A, N), then the counterfactual $M(n, o)$ is not a function of the original one-step-ahead counterfactuals $Y(a, m)$ and $M(a)$. This can be seen from the fact that whereas M has a single parent A in Figure 38.3(a), under the elaboration that includes N and O it now has two: $\{N, O\}$.
- 3.(ii) Similarly, if condition (ii) fails so that N has a direct effect on Y (relative to A, M, O), then the counterfactual $Y(n, o)$ is not a function of the original one-step-ahead counterfactuals $Y(a, m)$ and $M(a)$. Y has two parents $\{A, M\}$ in Figure 38.3(a), under the elaboration that includes N and O it would have three $\{N, O, M\}$.

Example: The River Blindness Studies

Returning to the river blindness study, note that intervening to set $s = 1$ and $a = 0$ gives $Y(s = 1, a = 0) = Y(a = 1, M(a = 0))$ as random variables. Hence the PDE is identified from data in a four-arm (A, S) trial.⁴⁸ It is not identified from data on

46. Note, however, that if $P(Y(n = 0, o = 1))$ does not equal the first expression in the mediation formula it is possible that this is solely because (N, O) do not satisfy (i) and (ii) but that there are other subcomponents of A , say (N^*, O^*) that do satisfy (i) and (ii).

47. This is because

$$E[Y(a = 1, M(a = 0))] = E[Y(n = 1, o = 1, M(n = 0, o = 0))] \quad (38.21)$$

$$= \sum_m E[Y(n = 1, o = 1, m) | M(n = 0, o = 0) = m]p(M(n = 0, o = 0) = m), \quad (38.22)$$

but this latter conditional expectation term is cross-world in terms of the counterfactuals in \mathcal{G}^{ex} although, as noted, they do identify $E[Y(n = 0, o = 1)] - E[Y(n = 0, o = 0)]$.

48. Recall that in $Y(s = 1, a = 0)$, A is serving as “ N ” and S is “ O ”; see Figure 38.3(e) and Footnote 23.

A, M, Y because $M \leftrightarrow Y$ forms a “recanting district,” as defined in Shpitser [2013]. Note that had identification of the PDE failed due to a recanting witness, that is, A and S having a common child, then additional interventions on the variables in the graph would not have led to identification. Finally, we note that, owing to the context-specific conditional independences in this example, the distribution $p(Y(s = 0, a = 1)) = Y(a = 0, M(a = 1))$, which occurs in the Total Direct Effect is identified given data from the two arms $p(Y(s = x, a = x)), x \in \{0, 1\}$ by the formula given in (38.A), and hence also from $p(A, M, Y)$; see Appendix 38.A.2.

Counterfactuals Related to the DAG in Figure 38.5(a)

Recall that $E [Y(n = 0, o = 1)]$ is identified under the FFRCISTG models associated with the graphs in Figures 38.6(a) and (b), but not (c).

Let $Y(a, l, m), M(a, l)$, and $L(a)$ denote the one-step-ahead counterfactuals associated with the graph in Figure 38.5(a). It follows from the deterministic counterfactual relation $N(a) = O(a) = a$ and the NPSEM associated with Figure 38.5(a), and its associated expanded graph \mathcal{G}^{ex} in Figure 38.6(a) that the random variable

$$Y(n = 0, o = 1) = Y(o = 1, L(n = 0), M(n = 0, L(n = 0)))$$

associated with the NPSEM in Figure 38.6(a) can be written in terms of the counterfactuals associated with the graph in Figure 38.5(a) as the cross-world counterfactual

$$Y(a = 1, L(a = 0), M(a = 0)) = Y(a = 1, L(a = 0), M(a = 0, L(a = 0))). \tag{38.23}$$

Similarly, if we assume that \mathcal{G}^{edge} of Figure 38.9 is an NPSEM, the counterfactual (38.23) also equals, as a random variable, the counterfactual $Y(A^{(1)} = 0, A^{(2)} = 0, A^{(3)} = 1)$. Thus if either Figure 38.6(a) or Figure 38.9 represented the FFRCISTG generating the data, the distribution of Equation (38.23) is identified by the same formula (38.17).

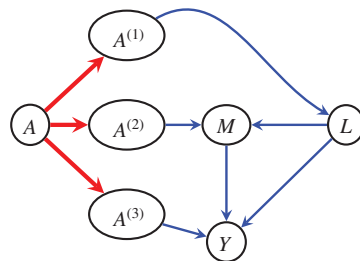


Figure 38.9 The edge expanded graph \mathcal{G}^{edge} associated with the DAG shown in Figure 38.5(a).

Likewise,

$$Y(n = 0, o = 1) = Y(o = 1, L(o = 1), M(n = 0, L(o = 1)))$$

associated with the graph in Figure 38.6(b) equals (as a random variable) the cross-world counterfactual

$$Y(a = 1, L(a = 1), M(a = 0, L(a = 1))). \quad (38.24)$$

associated with the graph in Figure 38.5(a). Again, if we assume that \mathcal{G}^{edge} of Figure 38.9 is an NPSEM, the counterfactual (38.24) also equals $Y(A^{(1)} = 1, A^{(2)} = 0, A^{(3)} = 1)$ (as a random variable). Thus, in this case if either Figure 38.6(b) or Figure 38.9 represented the FFRCISTG generating the data, the distribution of Equation (38.24) is identified by the same formula (38.18).

In contrast, $E[Y(n = 0, o = 1)]$ associated with the graph in Figure 38.6(c) is not the mean of any counterfactual defined from $Y(a, l, m)$, $M(a, l)$, and $L(a)$ under the graph in Figure 38.5(a) since L , after intervening to set $n = 0, o = 1$, is neither $L(a = 1)$ nor $L(a = 0)$ as both imply a counterfactual for L under which $n = o$. Furthermore, the parameter occurring in the PDE

$$E[Y(a = 1, M(a = 0))] = E[Y(a = 1, L(a = 1), M(a = 0, L(a = 0)))]$$

and associated with the graph in Figure 38.5(a) is not identified under the FFRCISTGs associated with any of the three graphs in Figures 38.6(a), (b), and (c). This is because all three of these graphs are compatible with the NPSEM in Figure 38.5(a) under which, by recursive substitution, we have the following equality $Y(a_1, M(a_0)) = Y(a_1, L(a_1), M(a_0, L(a_0)))$, as random variables. But the counterfactual $Y(a_1, L(a_1), M(a_0, L(a_0)))$, since it involves $L(a_0)$ and $L(a_1)$, does *not* correspond to an intervention on n and o under any of the expanded graphs in Figures 38.6(a), (b), or (c). Similarly, $Y(a_0, M(a_1))$ is not identified under any of these graphs. These results follow from the fact that L is a recanting witness for A in Figure 38.5(a).

In the following section, we define path-specific effects associated with the NPSEM model. We show that the two identified nested cross-world counterfactuals are identified path-specific effects under the NPSEM-IE for the graph in Figure 38.5(a). In particular, $E[Y(a = 1, L(a = 0), M(a = 0))]$ is the path-specific effect associated with the path $A \rightarrow Y$ and $E[Y(a = 1, L(a = 1), M(a = 0, L(a = 1)))]$ is the path-specific effect associated with the paths $A \rightarrow Y$, $A \rightarrow L \rightarrow Y$, and $A \rightarrow L \rightarrow M \rightarrow Y$. Thus, for Pearl, identification of these nested cross-world counterfactuals associated with Figure 38.5(a) follows from the assumption that the

distribution of the counterfactuals obeys the NPSEM-IE model associated with the graph in Figure 38.5(a). In contrast, the identification of these cross-world counterfactuals under an FFRCISTG model follows from two different extensions of Pearl's original story: the first is identified under the extension in which N but not O is a cause of L and the second from the extension that O but not N is a cause of L . The respective identifying formulae are the same under both theories.

38.4 Path-Specific Counterfactuals

In Section 38.1 we considered different notions of *direct* effect, which led to the notion of nested counterfactuals and the PDE, which is identified under the NPSEM-IE associated with the graph in Figure 38.3(a) via the associated Mediation Formula. In Section 38.3.2, following Robins and Richardson [2010], we discuss the notion of separability of effects in the sense of Stensrud et al. [2020c] via an expanded (N, O) graph. We showed that the counterfactuals defining the PDE were equal to ordinary (non-cross-world) interventional counterfactuals in the NPSEM given by the expanded graph (with determinism). We also showed that under the corresponding FFRCISTG these effects were identified by the mediation formula. In this section, we now consider path-specific effects which generalize the notion of direct and indirect effects.

In the simplest setting, the intuition behind an indirect effect is to consider all paths from A to Y *other than* the edge $A \rightarrow Y$. This can be generalized to settings where the effect along a particular *set* of causal paths from A to Y is of interest. In what follows we will show that each such path-specific effect will correspond to a cross-world counterfactual contrast associated with the (original) graph \mathcal{G} . We will see that in the absence of recanting witnesses these cross-world counterfactuals are equal (as random variables) to interventional counterfactuals in the NPSEM associated with \mathcal{G}^{edge} , the edge expanded graph associated with the set of treatment variables A .⁴⁹ Consequently, these path-specific counterfactuals will be identified if and only if the corresponding intervention is identified under the FFRCISTG associated with \mathcal{G}^{edge} . Thus all identifying formulae for path-specific cross-world counterfactuals on \mathcal{G} may be derived from \mathcal{G}^{edge} . Further, it follows that all identifying formulae for path-specific cross-world counterfactuals may also be obtained under the assumption that \mathcal{G} is an NPSEM-IE.

49. As noted earlier, the counterfactuals associated with \mathcal{G}^{edge} may not have clear substantive meaning. For \mathcal{G}^{edge} to be substantive it is necessary for each treatment variable A to be decomposable into components each of which could, in principle, be intervened on (separately) and affect one and only one of its children; see Footnote 34, Section 38.3.4 for further discussion. The graph \mathcal{G}^{edge} may still be useful as a formal construction.

The general theory developed by Shpitser [2013] associates a random variable with each subset of causal paths between a treatment A and an outcome Y . The intuition is that this subset of proper⁵⁰ causal paths from A to Y denoted π remain active, while all other causal paths, denoted $\bar{\pi}$, from A to Y are blocked.⁵¹ Next, pick a pair of value sets a and a' for elements in A ; a will be associated with active paths, a' with those that are blocked.

For any $V_i \in V$, define the potential outcome $V_i(\pi, a, a')$ by setting A to a for the purposes of paths in π that end in V_i , and setting A to a' for the purposes of proper causal paths from A to V_i not in π .⁵² Formally, the definition is as follows, for any $V_i \in V$:

$$\begin{aligned} V_i(\pi, a, a') &\equiv a \quad \text{if } V_i \in A, \\ V_i(\pi, a, a') &\equiv V_i(\{V_j(\pi, a, a') \mid V_j \in \text{pa}_i^\pi\}, \{V_j(a') \mid V_j \in \text{pa}_i^{\bar{\pi}}\}). \end{aligned} \quad (38.25)$$

where $V_j(a') \equiv a'$ if $V_j \in A$, and given by recursive substitution otherwise, pa_i^π is the set of parents of V_i along an edge which is a part of a path in π , and $\text{pa}_i^{\bar{\pi}}$ is the set of all other parents of V_i .

A counterfactual $V_i(\pi, a, a')$ is said to be *edge inconsistent* if, for some edge $A_k \rightarrow V_j$ in \mathcal{G} , counterfactuals of the form $V_j(a_k, \dots)$ and $V_j(a'_k, \dots)$ occur in $V_i(\pi, a, a')$, otherwise it is said to be *edge consistent*. In the former case V_j is said to be a *recanting witness (for π)*. It is simple to verify using Equation (38.25) that edge consistent counterfactuals are precisely those where no paths in π and $\bar{\pi}$ share the initial edge. Shpitser [2013] and Shpitser and Tchetgen Tchetgen [2016] have shown that a joint distribution $p(V(\pi, a, a'))$ containing an edge-inconsistent counterfactual $V_i(\pi, a, a')$ is not identified in the NPSEM-IE (nor weaker causal models) in the presence of a recanting witness.

As an example, consider the graph shown in Figure 38.5(a) and the counterfactual given in Equation (38.23) that corresponds to the path $\pi_1 = \{A \rightarrow Y\}$:

$$Y(\pi_1, a = 1, a = 0) \equiv Y(a = 1, L(a = 0), M(a = 0, L(a = 0))).$$

50. A proper causal path intersects the set A only once at the source node.

51. It is important to understand that the predicates “active” and “blocked” are applied to paths. In particular, it is possible for every edge and vertex on a blocked path to also be present on some active path, and vice-versa.

52. Note that it follows from the definition of proper causal path that each path in π is associated with a unique vertex in A ; similarly for each path in $\bar{\pi}$. (Two paths may be associated with the same vertex in A .)

The counterfactual associated with the paths $\pi_2 = \{A \rightarrow Y, A \rightarrow L \rightarrow Y\}$ is given by:

$$Y(\pi_2, a = 1, a = 0) \equiv Y(a = 1, L(a = 1), M(a = 0, L(a = 0))).$$

Note that $Y(\pi_1, a = 1, a = 0)$ is edge consistent while $Y(\pi_2, 1, 0)$ is edge-inconsistent due to the presence of $L(a = 0)$ and $L(a = 1)$.⁵³

This result is proved in [Shpitser and Tchetgen Tchetgen 2016]:

Theorem 38.1 If $V(\pi, a, a')$ is edge consistent, then under the NPSEM-IE for the DAG \mathcal{G} ,

$$p(V(\pi, a, a')) = \prod_{i=1}^K p(V_i | a \cap \text{pa}_i^\pi, a' \cap \text{pa}_i^{\bar{\pi}}, \text{pa}_i^{\mathcal{G} \setminus A}). \quad (38.26)$$

As an example of such an identification consider the distribution $p(Y(\pi, a, a'))$ of the edge consistent counterfactual in Figure 38.5(a). It follows from Theorem 38.1 that

$$\begin{aligned} p(Y(\pi_1, a = 1, a = 0)) &= p(Y(a = 1, L(a = 0), M(a = 0, L(a = 0)))) \\ &= \sum_{m,l} p(Y | m, l, a = 1) p(m | l, a = 0) p(l | a = 0), \end{aligned}$$

a marginal distribution derived from Equation (38.26).

In the following, we exploit an equivalence between edge consistent counterfactuals $V_i(\pi, a, a')$ and standard potential outcomes based on edge expanded graphs \mathcal{G}^{edge} , already defined in the case of a single treatment variable in Section 38.3.3.⁵⁴ In this section, we will abbreviate \mathcal{G}^{edge} as \mathcal{G}^e for conciseness. The edge expanded graph both simplifies complex nested potential outcome expressions and enables us to leverage the prior result in Shpitser and Pearl [2006b] to identify conditional path-specific effects.

We now extend the definition of expanded graph to sets of treatments $|A| > 1$: Given an ADMG $\mathcal{G}(V)$, define for each $A_i \in A \subseteq V$ a new set of variables, $A_i^{\text{Ch}} \equiv \{A_i^j | V_j \in \text{Ch}_i\}$ with state spaces $\mathfrak{X}_{A_i^j} \equiv \mathfrak{X}_{A_i}$; thus for each directed edge $A_i \rightarrow V_j$ in $\mathcal{G}(V)$ from a treatment variable to its child, we have created a new variable A_i^j with the same state space as A_i . Denote the full set of new variables as $A^{\text{Ch}} \equiv \bigcup_{A_i \in A} A_i^{\text{Ch}}$. We define the edge expanded graph of $\mathcal{G}(V)$, written $\mathcal{G}^e(V \cup A^{\text{Ch}})$, as the graph with

53. The above development may be generalized to k different assignments rather than two, by partitioning the set of paths into k . These and other generalizations are termed *path interventions* by Shpitser and Tchetgen Tchetgen [2016].

54. A similar construction was called the “extended graph” in Malinsky et al. [2019].

the vertex set $V \cup A^{Ch}$; the edge expanded graph contains all the edges in \mathcal{G} except for the edges $A_i \rightarrow V_j$ that join a treatment variable A_i to its child V_j , in addition we add the edges $A_i \rightarrow A_i^j \rightarrow V_j$ (which thus “replace” the removed edge $A_i \rightarrow V_j$).

As an example, the edge expanded graph for the DAG in Figure 38.3(a), with $A_1 = V$, is shown in Figure 38.10. For conciseness, we will generally drop explicit references to vertices $V \cup A^{Ch}$, and denote edge expanded graph of $\mathcal{G}(V)$ by \mathcal{G}^e .

More generally, we associate a causal model with \mathcal{G}^e as follows. Let \mathbb{V} be the set of one-step-ahead potential outcomes associated with the original graph \mathcal{G} . Similarly, we let \mathbb{V}^e denote the set of one-step-ahead potential outcomes associated with \mathcal{G}^e , constructed as follows: For every $V_i(\text{pa}_i) \in \mathbb{V}$, we let $V_i(\text{pa}_i^e)$ be in \mathbb{V}^e . Note that this is well-defined, since vertices V_i in \mathcal{G} and \mathcal{G}^e share the number of parents, and the parent sets for every V_i share state spaces. In addition, for every $A_i^j \in A^{Ch}$, we let $A_i^j(a_i)$ for $a_i \in \mathcal{X}_{A_i}$ be in \mathbb{V}^e .

The edges $A_i \rightarrow A_i^j$ in \mathcal{G}^e are understood to represent *deterministic* equality relationships, such that $A_i^j = A_i$. More precisely, every $A_i^j \in A^{Ch}$ has a single parent A_i , and we let $A_i^j(a_i) = a_i$ so that $A_i^j(a_i)$ is a degenerate (constant) random variable corresponding to a point-mass at a_i .

The FFRCISTG model associated with the edge expanded graph \mathcal{G}^e includes these deterministic relationships. Note that it follows that given a distribution $p(\mathbb{V})$ over the counterfactuals given by the NPSEM \mathcal{G} there is a unique distribution $p^e(\mathbb{V}^e)$ over the counterfactuals given by \mathcal{G}^e that satisfies these deterministic relationships.

We now show the following two results. First, we show that an edge-consistent $V(\pi, a, a')$ may be represented without loss of generality by a counterfactual response to an intervention on a subset of A^{Ch} in \mathcal{G}^e with the causal model defined above. Second, we show that this response is identified by the same functional (38.26).

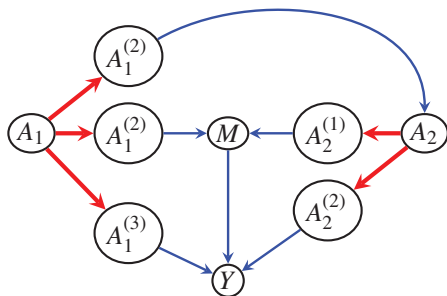


Figure 38.10 An edge expanded graph that considers components of the treatment variables A_1 and A_2 .

Given an edge consistent $V(\pi, a, a')$, define \mathcal{G}^e for $A \subseteq V$. We define a^π that assigns a_i to $A_i^j \in A^{\text{Ch}}$ if $A_i \rightarrow V_j$ in $\mathcal{G}(V)$ is in π , and assigns a'_i to $A_i^j \in A^{\text{Ch}}$ if $A_i \rightarrow V_j$ in $\mathcal{G}(V)$ is not in π . The resulting set of counterfactuals $V(a^\pi)$ is well defined in the model for \mathbb{V}^e , and we have the following result, proved in the [Appendix](#).

Proposition 38.3 Fix a distribution $p(\mathbb{V})$ in the NPSEM-IE for a DAG $\mathcal{G}(V)$, and consider the corresponding distribution $p^e(\mathbb{V}^e)$ in the FFRCISTG model associated with a DAG $\mathcal{G}^e(V \cup A^{\text{Ch}})$. Then for every $V_i \in V$, the random variable V_i in the original model associated with \mathcal{G} is equal to the random variable V_i in the restricted model associated with \mathcal{G}^e that includes the equalities defining the variables in A^{Ch} . Moreover, for any edge consistent π, a, a' , the random variable $V_i(\pi, a, a')$ in the original model associated with \mathcal{G} is equal to the random variable $V_i(a^\pi)$ in the restricted model associated with \mathcal{G}^e .

Theorem 38.2 Under the FFRCISTG model associated with the edge expanded DAG \mathcal{G}^e , for any edge consistent π, a, a' :

$$p^e(V(a^\pi)) = \prod_{i=1}^K p^e(V_i | a^\pi \cap \text{pa}_i^{\mathcal{G}^e}, \text{pa}_i^{\mathcal{G}^e} \setminus \mathcal{A}). \quad (38.27)$$

Note that since, by definition, the distribution $p^e(V \cup A^{\text{Ch}})$ is a deterministic function of $p(V)$, hence by Equation (38.27) $p^e(V(a^\pi))$ is identified by $p(V)$. Recall that if \mathcal{G}^e is an FFRCISTG model then this requires that all of the interventions on the variables in A^{Ch} are well-defined.

Corollary 38.2 Under the NPSEM-IE model associated with the DAG \mathcal{G} , for any edge consistent π, a, a' :

$$p(V(\pi, a, a')) = p^e(V(a^\pi)) = \prod_{i=1}^K p^e(V_i | a^\pi \cap \text{pa}_i^{\mathcal{G}^e}, \text{pa}_i^{\mathcal{G}^e} \setminus \mathcal{A}),$$

where \mathcal{G}^e is the edge expanded graph corresponding to \mathcal{G} .

In fact, Corollary 38.2 provides an alternative proof of Theorem 38.1 since, by definition, for any edge consistent π, a, a' :

$$p(V_i | a \cap \text{pa}_i^\pi, a' \cap \text{pa}_i^{\bar{\pi}}, \text{pa}_i^{\mathcal{G}} \setminus \mathcal{A}) = p^e(V_i | a^\pi \cap \text{pa}_i^{\mathcal{G}^e}, \text{pa}_i^{\mathcal{G}^e} \setminus \mathcal{A}).$$

Though, as discussed previously, if the graph \mathcal{G} is an NPSEM-IE the graph \mathcal{G}^e may be regarded as a purely formal construction that aids in the derivation of the identifying formulae. However, if \mathcal{G}^e is interpreted as a causal model, so that there are well-defined counterfactuals associated with intervening on each of the vertices

in A^{Ch} , then \mathcal{G}^e provides an interventionist interpretation of path-specific counterfactuals; in fact these interventions allow the identification results above to be checked, in principle, in a hypothetical randomized trial.

In the causal models derived from DAGs with unobserved variables (e.g., $\mathcal{G}(V \cup H)$), identification of distributions on potential outcomes such as $p(V(a))$ or $p(V(\pi, a, a'))$ may be stated without loss of generality on the latent projection ADMG $\mathcal{G}(V)$. A complete algorithm for identification of path-specific effects in NPSEM-IEs with hidden variable was given in Shpitser [2013] and presented in a more concise form in Shpitser and Sherman [2018].

We now show that identification theory for $p(V(\pi, a, a'))$ in latent projection ADMGs $\mathcal{G}(V)$ may be restated, without loss of generality, in terms of identification of $p^e(V(a^\pi))$ in $\mathcal{G}^e(V \cup A^{\text{Ch}})$.

Proposition 38.4 Let $\mathcal{G}(V \cup H)$ be a DAG and $Y \subseteq V$ be an ancestral set in $\mathcal{G}(a)$, so $an_{\mathcal{G}(a)}(Y(a)) = Y(a)$. Under the FFRCISTG model associated with the edge expanded DAG $\mathcal{G}^e(V \cup A^{\text{Ch}} \cup H)$, for any edge consistent π, a, a' , it follows that $Y(\pi, a, a') = Y(a^\pi)$ and thus $p(Y(\pi, a, a'))$ is identified given $p(V)$ if and only if $p^e(Y(a^\pi))$ is identified from $p^e(V \cup A^{\text{Ch}})$.

This proposition is a generalization of Theorem 38.2 from DAGs to latent projection ADMGs. To determine whether $p^e(Y(a^\pi))$ is identified we examine the ADMG $\mathcal{G}^e(V, A^{\text{Ch}})$. Since this graph is a standard latent projection ADMG (albeit with deterministic relationships relating A and A^{Ch}), the extended ID algorithm decomposes the distribution $p^e(Y(a^\pi))$ into a set of factors as in (41.21). However, in order for $p^e(Y(a^\pi))$ to be identified given data $p(V)$,⁵⁵ an additional requirement must be placed on the terms of this decomposition. Specifically, for each term $p^e(V_D(a^\pi, v_{pas_D^{\mathcal{G}^e(Y(a^\pi))}} = v_D))$ in (41.21), it must be the case that a^π assigns consistent values to each element of A that is in $pas_D^{\mathcal{G}^e(Y(a^\pi))} \equiv pa_D^{\mathcal{G}^e(Y(a^\pi))} \setminus (D \cup a^\pi)$, the random parents of district D that are not in D . This requirement corresponds to the *recanting district criterion* that was introduced and shown to be complete in Shpitser [2013]. Aside from this requirement, each term must be identified by the extended ID algorithm described in the companion paper [Chapter 41].⁵⁶

55. Note that the vertex set for $\mathcal{G}^e(V, A^{\text{Ch}})$ comprises $V \cup A^{\text{Ch}}$. Further, by construction, the distribution over $p^e(V \cup A^{\text{Ch}})$ is degenerate since the variables in A^{Ch} are deterministic functions of those in A .

56. See Point (b) in Section 38.3.5.

38.4.1 Conditional Path-specific Distributions

Having established that we can identify path-specific effects by working with potential outcomes derived from the \mathcal{G}^e model, we turn to the identification of conditional path-specific effects using the po-calculus. In [Shpitser and Pearl \[2006b\]](#), the authors present the conditional identification (IDC) algorithm for identifying quantities of the form $p(Y(x) | W(x))$ (in our notation), given an ADMG. Since conditional path-specific effects correspond to exactly such quantities defined on the model associated with the edge expanded graph \mathcal{G}^e , we can leverage their scheme for our purposes. The idea is to reduce the conditional problem, identification of $p^e(Y(a^\pi) | W(a^\pi))$, to an unconditional (joint) identification problem for which a complete identification algorithm already exists.

The algorithm has three steps: first, exhaustively apply Rule 2 of the po-calculus to reduce the conditioning set as much as possible; second, identify the relevant joint distribution using Proposition 38.4 and the complete algorithm of [Shpitser and Sherman \[2018\]](#); third, divide that joint by the marginal distribution of the remaining conditioning set to yield the conditional path-specific potential outcome distribution.

Note that we make use of SWIGs defined from edge expanded graphs of the form $\mathcal{G}^e(a^\pi, z)$. Beginning with \mathcal{G}^e the SWIG $\mathcal{G}^e(a^\pi, z)$ is constructed by the usual node-splitting operation: split nodes Z and A_i^j into random and fixed halves, where A_i^j has a fixed copy a if $A_i \rightarrow V_j$ in $\mathcal{G}(V)$ is in π , and a'_i if $A_i \rightarrow V_j$ in $\mathcal{G}(V)$ is not in π . Relabeling of random nodes proceeds as previously described. This procedure is in fact complete, as shown by the following result with a proof found in [Malinsky et al. \[2019\]](#).⁵⁷

Theorem 38.3 Let $p(Y(\pi, a, a') | W(\pi, a, a'))$ be a conditional path-specific distribution in the NPSEM-IE model for \mathcal{G} , and let $p^e(Y(a^\pi) | W(a^\pi))$ be the corresponding distribution under the FFRCISTG associated with the edge expanded graph $\mathcal{G}^e(V \cup A^{\text{Ch}})$. Let Z be the maximal subset of W such that $p^e(Y(a^\pi) | W(a^\pi)) = p^e(Y(a^\pi, z) | W(a^\pi, z)Z(a^\pi, z))$. Then $p^e(Y(a^\pi) | W(a^\pi))$ is identifiable in \mathcal{G}^e if and only if $p^e(Y(a^\pi, z), W(a^\pi, z) \setminus Z(a^\pi, z))$ is identifiable in \mathcal{G}^e .

As an example, $p(Y(a, M(a')))$ is identified from $p(C, A, M, Y)$ in the NPSEM-IE model for the graph in Figure 38.11(a) via

$$p(Y(a, M(a'))) = \sum_m \left(\frac{\sum_c p(Y, m | a, c)p(c)}{\sum_c p(m | a, c)p(c)} \right) \left(\sum_{c^*} p(m | a', c^*)p(c^*) \right).$$

57. [Malinsky et al. \[2019\]](#) assumed an NPSEM-IE but the proof only uses the rules of the po-calculus, hence also applies under the less restrictive FFRCISTG assumptions for \mathcal{G}^e .

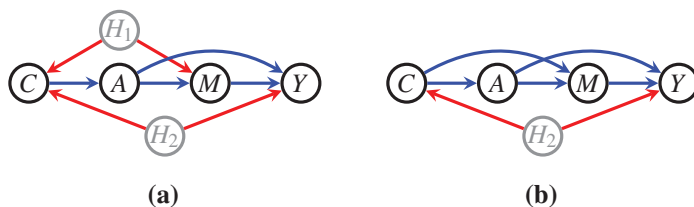


Figure 38.11 (a) A hidden variable causal DAG where $p(Y(a, M(a')))$ is identified but $p(Y(a, M(a'))|C)$ is not identified. (b) A seemingly similar hidden variable causal DAG where both $p(Y(a, M(a')))$ and $p(Y(a, M(a'))|C)$ are identified.

However, $p(Y(a, M(a'))|C)$ is not identified since $p(Y(a, M(a')), C)$ must first be identified, and this joint distribution is not identified via results in Shpitser [2013]. On the other hand, $p(Y(a, M(a'))|C)$ is identified from $p(C, A, M, Y)$ in a seemingly similar graph in Figure 38.11(b), via $\sum_M p(Y|M, a, C)p(M|a', C)$.

38.5 Conclusion

We have shown here that graphical insights derived from Pearl's thought experiment may be used to place mediation analysis on a firm interventionist footing, and yield analyses of direct, indirect, and path-specific effects that are amenable to falsification, and explainable to practitioners.

Acknowledgments

The first author was supported by the grants ONR N000141912446 and National Institutes of Health (NIH) awards R01 AG057869, R01 AI127271, R37 AI102634, U01 CA261277-01, and R01 CA222147-02.

The second author was supported by the grant ONR N00014-19-1-2446.

The third author was supported by the grants ONR N00014-18-1-2760, NSF CAREER 1942239, NSF 1939675, and R01 AI127271-01A1.

The authors would like to thank F. Richard Guo for his insightful comments that improved this manuscript.

38.A Appendix

38.A.1 Proof of PDE Bounds under the FFRCISTG Model

We here prove the bounds on the PDE implied by the FFRCISTG model associated with the graph in Figure 38.3(a).

Proof. It follows from the definition that: $PDE_{a,a'} = P(Y(a, M(a')) = 1) - P(Y = 1 | A = a')$. Note that

$$\begin{aligned} p(Y(a, M(a')) = 1) &= p(Y(a, m = 0) = 1 | M(a') = 0)p(M(a') = 0 | A = a') \\ &\quad + p(Y(a, m = 1) = 1 | M(a') = 1)p(M(a') = 1 | A = a') \\ &= p(Y(a, m = 0) = 1 | M(a') = 0)p(M = 0 | A = a') \\ &\quad + p(Y(a, m = 1) = 1 | M(a') = 1)p(M = 1 | A = a'). \end{aligned}$$

The quantities $p(Y(a, m = 0) = 1 | M(a') = 0)$ and $p(Y(a, m = 1) = 1 | M(a') = 1)$ are constrained by the law for the observed data via:

$$\begin{aligned} p(Y = 1 | A = a, M = 0) &= p(Y(a, m = 0) = 1) \\ &= p(Y(a, m = 0) = 1 | M(a') = 0)p(M(a') = 0) \\ &\quad + p(Y(a, m = 0) = 1 | M(a') = 1)p(M(a') = 1) \\ &= p(Y(a, m = 0) = 1 | M(a') = 0)p(M = 0 | A = a') \\ &\quad + p(Y(a, m = 0) = 1 | M(a') = 1)p(M = 1 | A = a'), \end{aligned}$$

$$\begin{aligned} p(Y = 1 | A = a, M = 1) &= p(Y(a, m = 1) = 1) \\ &= p(Y(a, m = 1) = 1 | M(a') = 0)p(M(a') = 0) \\ &\quad + p(Y(a, m = 1) = 1 | M(a') = 1)p(M(a') = 1) \\ &= p(Y(a, m = 1) = 1 | M(a') = 0)p(M = 0 | A = a') \\ &\quad + p(Y(a, m = 1) = 1 | M(a') = 1)p(M = 1 | A = a'). \end{aligned}$$

It then follows from the analysis in Section 2.2 in [Richardson and Robins \[2010\]](#) that the set of possible values for the pair

$$(\alpha_0, \alpha_1) \equiv (p(Y(a, m = 0) = 1 | M(a') = 0), p(Y(a, m = 1) = 1 | M(a') = 1))$$

compatible with the observed joint distribution $p(m, y | a)$ is given by:

$$(\alpha_0, \alpha_1) \in [l_0, u_0] \times [l_1, u_1]$$

where,

$$\begin{aligned} l_0 &= \max\{0, 1 + (p(Y = 1 | A = a, M = 0) - 1)/p(M = 0 | A = a')\}, \\ u_0 &= \min\{p(Y = 1 | A = a, M = 0)/p(M = 0 | A = a'), 1\}, \end{aligned}$$

$$\begin{aligned} l_1 &= \max\{0, 1 + (p(Y = 1 | A = a, M = 1) - 1)/p(M = 1 | A = a')\}, \\ u_1 &= \min\{p(Y = 1 | A = a, M = 1)/p(M = 1 | A = a'), 1\}. \end{aligned} \quad \blacksquare$$

38.A.2 Proof that the PDE is Not Identified in the River Blindness Study

Consider the NPSEM-IE corresponding to the DAG shown in Figure 38.12, which may be obtained by marginalizing R from the causal DAG representing the river blindness studies shown in Figure 38.1(a).

The one-step-ahead counterfactuals defining the model are: $U, A, S(a), M(a, u), Y(m, s, u)$. We will assume that all variables, including U , are binary. Though not shown explicitly on the graph, we also have the following constraints: (i) $Y(m, s_0, u_0) = Y(m, s_0, u_1) \equiv Y(m, s_0)$; (ii) $Y(m_0, s_1, u_0) = Y(m_0, s_1, u_1) \equiv Y(m_0, s_1)$; (iii) $M(a_1, u_0) = M(a_1, u_1) \equiv M(a_1)$, and (iv) $S(a) = a$, where here we use x_i as a shorthand for $x = i$. These arise from, respectively, (i) the fact that if immunosuppressants are not available ($s = 0$) then the patient’s outcome (Y) is not influenced by their predisposition (U) to use medicine; (ii) the availability of immunosuppressants, and hence the patient’s predisposition (U) to use them, is not relevant to patients who do not receive ivermectin;⁵⁸ (iii) the fact that in a randomized trial ($a = 1$) the treatment the patient receives is not influenced by U ; and (iv) the fact that the clinic is available ($s = 1$) if and only if a patient is in the randomized trial ($a = 1$). The NPSEM-IE is then defined by the following 10 parameters:

$$\begin{aligned}
 U : & \quad \theta_U \equiv p(U = 1); \\
 A : & \quad \theta_A \equiv p(A = 1); \\
 M(a, u) : & \quad \theta_M(a_0, u) \equiv p(M(a_0, u) = 1), \quad \text{for } u \in \{0, 1\}, \\
 & \quad \theta_M(a_1) \equiv p(M(a_1) = 1); \\
 Y(m, s, u) : & \quad \theta_Y(m_1, s_1, u) \equiv p(Y(m_1, s_1, u) = 1), \quad \text{for } u \in \{0, 1\}, \\
 & \quad \theta_Y(m, s) \equiv p(Y(m, s) = 1), \quad \text{for } (m, s) \in \{(0, 0), (0, 1), (1, 0)\}.
 \end{aligned}$$

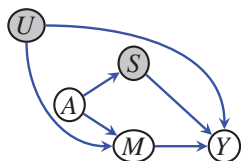


Figure 38.12 The DAG corresponding to the projection of the DAG \mathcal{G} shown in Figure 38.1(a) after marginalizing R .

58. This is an additional restriction not present in the original account but introduced here solely to reduce the number of parameters. This corresponds to removing the $U \rightarrow R$ edge in the SWIG $\mathcal{G}(a_1, m_0, s_1)$ where m is set to 0; see Figure 38.2. (Since non-identifiability under a sub-model implies non-identifiability in the larger model, we can make this assumption without loss of generality.)

Let θ denote a vector containing all of these parameters. The corresponding observed distribution $p_\theta(A, M, Y)$ is given by the following equations:

$$\begin{aligned}
 p_\theta(A = 1) &= \theta_A, \\
 p_\theta(M = 1 | A = a) &= \begin{cases} \theta_M(a_0, u_0)(1 - \theta_U) + \theta_M(a_0, u_1)\theta_U, & \text{if } a = 0, \\ \theta_M(a_1), & \text{if } a = 1, \end{cases} \\
 p_\theta(Y = 1 | A = a, M = m) &= \begin{cases} \theta_Y(m, s_a), & \text{if } (a, m) \in \{(0, 0), (0, 1), (1, 0)\}, \\ \theta_Y(m_1, s_1, u_0)(1 - \theta_U) + \theta_Y(m_1, s_1, u_1)\theta_U, & \text{if } (a, m) = (1, 1). \end{cases}
 \end{aligned}$$

Given θ , consider a perturbed vector $\tilde{\theta}$ defined by taking

$$\begin{aligned}
 \tilde{\theta}_M(a_0, u_0) &\equiv \theta_M(a_0, u_0) + \varepsilon/(1 - \theta_U), \\
 \tilde{\theta}_M(a_0, u_1) &\equiv \theta_M(a_0, u_1) - \varepsilon/\theta_U,
 \end{aligned}$$

for sufficiently small ε and leaving the other 8 parameters unchanged from θ . It is simple to see that the resulting observed distribution is unchanged, so $p_{\tilde{\theta}}(A, M, Y) = p_\theta(A, M, Y)$, since the perturbation only changes the expression for $p(M = 1 | A = 0)$, but the additional terms involving ε cancel.

Turning to the PDE we see that:

$$\begin{aligned}
 &p_\theta(Y(a_1, M(a_0)) = 1) \\
 &= p_\theta(Y(M(a_0), s_1) = 1) \\
 &= \sum_u p_\theta(Y(M(a_0), s_1) = 1 | U = u) p_\theta(U = u) \\
 &= \sum_{u,k} p_\theta(Y(m_k, s_1) = 1 | U = u, M(a_0) = k) p_\theta(M(a_0) = k | U = u) p_\theta(U = u) \\
 \text{consistency} &= \sum_{u,k} p_\theta(Y(m_k, s_1, u) = 1 | U = u, M(a_0) = k) p_\theta(M(a_0, u) = k | U = u) p_\theta(U = u) \\
 \text{independence} &= \sum_{u,k} p_\theta(Y(m_k, s_1, u) = 1) p_\theta(M(a_0, u) = k) p_\theta(U = u) \\
 &= \theta_Y(m_0, s_1)(1 - \theta_M(a_0, u_0))(1 - \theta_U) + \theta_Y(m_0, s_1)(1 - \theta_M(a_0, u_1))\theta_U \\
 &\quad + \theta_Y(m_1, s_1, u_0)\theta_M(a_0, u_0)(1 - \theta_U) + \theta_Y(m_1, s_1, u_1)\theta_M(a_0, u_1)\theta_U.
 \end{aligned}$$

A simple calculation shows that

$$p_{\tilde{\theta}}(Y(M(a_0), s_1) = 1) = p_\theta(Y(M(a_0), s_1) = 1) + \varepsilon(\theta_Y(m_1, s_1, u_0) - \theta_Y(m_1, s_1, u_1))$$

so that the PDE will take a different value under $p_{\tilde{\theta}}$, so long as $\theta_Y(m_1, s_1, u_0) \neq \theta_Y(m_1, s_1, u_1)$.⁵⁹

That the PDE is not identified in the river blindness study example should not be surprising in light of the results in Section 38.4. In particular, we know that $Y(a_1, M(a_0)) = Y(s_1, a_0)$.⁶⁰ In addition, we see that in the SWIG $\mathcal{G}(a = 0, s = 1)$ shown in Figure 38.13(a), which here plays the role of the SWIG for the expanded graph \mathcal{G}^{ex} , $Y(a_0, s_1)$ is in the same district as $M(a_0)$, but the fixed nodes $a = 0$ and $s = 1$ are both parents of this district, but are both associated with A in the original graph. Consequently $M(a_0) \leftrightarrow Y(a_0, s_1)$ forms a recanting district. However, formally Section 38.4 considers models defined solely via ordinary conditional independences, whereas the model in the river blindness study also incorporated context-specific independences. For this reason, since a quantity may be unidentified in a model but identified in a submodel, we provided an explicit construction of an NPSEM-IE for the DAG in Figure 38.1(a).⁶¹

Similarly, consideration of the SWIG $\mathcal{G}(a = 1, s = 0)$ shown in Figure 38.13(b) shows that $p(Y(a_1, s_0))$ is identified from $p(A, M, Y)$ because there is no recanting district.⁶²

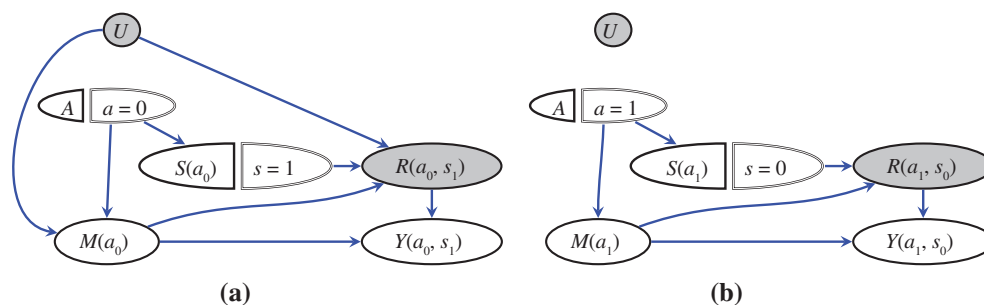


Figure 38.13 (a) The SWIG $\mathcal{G}(a = 0, s = 1)$ associated with the DAG shown in Figure 38.1(a); (b) The SWIG $\mathcal{G}(a = 1, s = 0)$ associated with the DAG shown in Figure 38.1(a). The $U \rightarrow M$ edge is absent because this is a randomized study ($a = 1$). The $U \rightarrow R$ edge is not present because the clinic is not available ($s = 0$), hence patients do not have the option to take immunosuppressants.

59. This inequality corresponds to the presence of the edge $U \rightarrow Y$ in Figure 38.12.

60. This assumes that missing edges correspond to the absence of direct effects at the individual level, so that $M(a_0, s_1) = M(a_0)$.

61. Since the NPSEM-IE is a submodel of the FFRCISTG, this also establishes that the PDE is not identified under the latter interpretation of the DAG in Figure 38.1(a).

62. Note that this identification argument does not use the additional constraint (ii).

$$\begin{aligned}
p(Y(a_1, s_0) = 1) &= \sum_{r,m} p(y = 1 | r, m) p(r | s = 0, m) p(m | a = 1) \quad \text{g-formula} \\
\text{independence} &= \sum_{r,m} p(y = 1 | r, m, s = 0) p(r | s = 0, m) p(m | a = 1) \\
\text{marginalization} &= \sum_m p(y = 1 | m, s = 0) p(m | a = 1) \\
\text{determinism} &= \sum_m p(y = 1 | m, a = 0) p(m | a = 1). \tag{38.A}
\end{aligned}$$

Thus, in this example, $p(Y(a, s)) = p(Y(s, M(a)))$ is identified for $(a, s) \in \{(0, 0), (1, 1), (1, 0)\}$ but is not identified for $(a, s) = (0, 1)$. Since $Y(a_0, M(a_1)) = Y(a_1, s_0)$, it follows that $p(Y(a_0, M(a_1)))$, which forms part of the Total Direct Effect (38.3), is also identified.

This conclusion also follows from Proposition 38.1, here letting A be “ N ” and S be “ O ”, since equality (38.15) holds with $x = 1$ (though not with $x = 0$).

38.A.3 Detecting Confounding via Interventions on A and S

In the river blindness study, it is not possible to detect the confounder U via interventions on A and M since the distribution $\{A, M(a), Y(a, m)\}$ obeys the FFRCISTG model corresponding to Figure 38.3(a). This is due to the context-specific independences that hold in this example; see Section 38.2.5. However, confounding becomes detectable if we are able to intervene on S and A . To see this, note that:

$$\begin{aligned}
p(Y(a_0, s_1) = 1 | M(a_0) = 1) &= \sum_u p(Y(a_0, s_1) = 1 | M(a_0) = 1, U = u) p(u | M(a_0) = 1) \\
&= \sum_u p(Y(m_1, s_1, u) = 1 | M(a_0) = 1, U = u) p(u | M(a_0) = 1) \\
&= \sum_u p(Y(m_1, s_1, u) = 1) p(u | M(a_0) = 1) \\
&= \frac{\theta_Y(m_1, s_1, u_0)(1 - \theta_U)\theta_M(a_0, u_0) + \theta_Y(m_1, s_1, u_1)\theta_U\theta_M(a_0, u_1)}{(1 - \theta_U)\theta_M(a_0, u_0) + \theta_U\theta_M(a_0, u_1)}
\end{aligned}$$

which will depend on $\theta_M(a_0, u_0)$ and $\theta_M(a_0, u_1)$, (if $\theta_Y(m_1, s_1, u_0) \neq \theta_Y(m_1, s_1, u_1)$).

However,

$$\begin{aligned}
p(Y(a_1, s_1) | M(a_1) = 1) &= \sum_u p(Y(a_1, s_1) | M(a_1) = 1, U = u) p(u | M(a_1) = 1) \\
&= \sum_u p(Y(m_1, s_1, u) | M(a_1) = 1, U = u) p(u | M(a_1) = 1)
\end{aligned}$$

$$\begin{aligned}
&= \sum_u p(Y(m_1, s_1, u))p(u | M(a_1) = 1) \\
&= \sum_u p(Y(m_1, s_1, u))p(u) \\
&= \theta_Y(m_1, s_1, u_0)(1 - \theta_U) + \theta_Y(m_1, s_1, u_1)\theta_U.
\end{aligned}$$

Thus $p(Y(a_0, s_1) = 1 | M(a_0) = 1) \neq p(Y(a_1, s_1) = 1 | M(a_1) = 1)$, while if U were absent we would have equality.⁶³ That the equality holds if U is absent may be seen by removing U from the SWIG $\mathcal{G}(a, s)$ constructed from \mathcal{G} in Figure 38.1(a) and then applying Rule 3 of the po-calculus [Shpitser et al. 2021]: the equality follows from the fact that a is d-separated from $Y(a, s)$ given $M(a)$. Conversely, that the equality fails to hold when U is present is not surprising since we have the d-connecting path $a \rightarrow M(a) \leftarrow U \rightarrow R(m, s) \rightarrow Y(m, s)$ in Figure 38.13(a).

38.A.4 Proof of Proposition 38.3

Proof. For any $V_i \in V$, $V_i(a^\pi)$ is defined via (41.1) applied to \mathcal{G}^{edge} ,

$$V_i(a^\pi) \equiv V_i \left(a^\pi_{pa_i^{\mathcal{G}^{edge}} \cap A^{Ch}}, V_{pa_i^{\mathcal{G}^{edge}} \setminus A^{Ch}}(a^\pi) \right).$$

Similarly, $V_i(\pi, a, a')$ is defined via Equation (38.25) applied to \mathcal{G} as

$$\begin{aligned}
V_i(\pi, a, a') &\equiv a \text{ if } V_i \in A, \\
V_i(\pi, a, a') &\equiv V_i(\{V_j(\pi, a, a') | V_j \in pa_i^\pi\}, \{V_j(a') | V_j \in pa_i^{\bar{\pi}}\})
\end{aligned}$$

where $V_j(a') \equiv a'$ if $V_j \in A$, and given by (41.1) otherwise, pa_i^π is the set of parents of V_i along an edge that is a part of a path in π , and $pa_i^{\bar{\pi}}$ is the set of all other parents of V_i .

By definition of \mathcal{G}^{edge} , the induction on the tree structure of Equation (38.25) in \mathcal{G} matches the induction on the tree structure of (41.1) in \mathcal{G}^{edge} .

Finally, since $V_i(\pi, a, a')$ is edge consistent, any edge of the form $A_k \rightarrow V_j$ in \mathcal{G} , for any $A_k \in A$ that starts a proper causal path that ends at V_i is assigned precisely one value (either a_{A_k} or a'_{A_k}). Moreover, this value in the base case of the induction of Equation (38.25), by definition of a^π matches the value $a_{A_k}^\pi$ assigned by the corresponding base case of the induction of (41.1). This establishes our conclusion.

That the corresponding observed data variables V_i match in models represented by the original graph \mathcal{G} , and the edge expanded graph \mathcal{G}^{edge} follows by the above

63. Note that under the additional assumption (iii), we have: $p(Y(a_0, s_1) = 1 | M(a_0) = 0) = p(Y(a_1, s_1) = 1 | M(a_1) = 0)$ since $\theta_Y(m_0, s_1, u_0) = \theta_Y(m_0, s_1, u_1)$. Without (iii) the equality will not hold.

argument, using the fact that all elements in A^{Ch} are deterministically related to the appropriate elements in A .

Malinsky et al. [2019] prove a weaker result establishing equality in distribution.

References

- O. O. Aalen, M. J. Stensrud, V. Didelez, R. Daniel, K. Røysland, and S. Strohmaier. 2020. Time-dependent mediators in survival analysis: Modeling direct and indirect effects with the additive hazards model. *Biom. J.* 62, 3, 532–549. DOI: <https://doi.org/10.1002/bimj.201800263>.
- C. Avin, I. Shpitser, and J. Pearl. 2005. Identifiability of path-specific effects. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence (IJCAI-05)*. Vol. 19, Morgan Kaufmann, San Francisco, 357–363.
- I. J. Dahabreh, J. M. Robins, S. J. Haneuse, and M. A. Hernán. 2019. Generalizing causal inferences from randomized trials: Counterfactual and graphical identification. *arXiv preprint arXiv:1906.10792*.
- V. Didelez. 2019. Defining causal mediation with a longitudinal mediator and a survival outcome. *Lifetime Data Anal.* 25, 4, 593–610. DOI: <https://doi.org/10.1007/s10985-018-9449-0>.
- J. J. Lok. 2016. Defining and estimating causal direct and indirect effects when setting the mediator to specific values is not feasible. *Stat. Med.* 35, 22, 4008–4020. DOI: <https://doi.org/10.1002/sim.6990>.
- D. Malinsky, I. Shpitser, and T. S. Richardson. 2019. A potential outcomes calculus for identifying conditional path-specific effects. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*. arXiv:1903.03662.
- J. Pearl. 2001. Direct and indirect effects. In *Proceedings of the 17th Conference on Uncertainty in Artificial Intelligence (UAI-01)*. Morgan Kaufmann, San Francisco, 411–420.
- J. Pearl. 2009. *Causality: Models, Reasoning, and Inference*. (2nd. ed.). Cambridge University Press. ISBN: 978-0521895606.
- J. Pearl. 2012. The causal mediation formula—A guide to the assessment of pathways and mechanisms. *Prev. Sci.* 13, 4, 426–436. DOI: <https://doi.org/10.1007/s11221-011-0270-1>.
- J. Pearl. 2018. Does obesity shorten life? Or is it the soda? On non-manipulable causes. *J. Causal Inference.* 6. DOI: <https://doi.org/10.1515/jci-2018-2001>.
- J. Pearl. 2019. On the interpretation of $do(x)$. *J. Causal Inference.* 7. DOI: <https://doi.org/10.1515/jci-2019-2002>.
- T. S. Richardson and J. M. Robins. 2010. Analysis of the binary instrumental variable model. *Heuristics, Probability and Causality: A Tribute to Judea Pearl*. 415–444.
- T. S. Richardson and J. M. Robins. 2013. Single world intervention graphs (SWIGs): A unification of the counterfactual and graphical approaches to causality. preprint: <http://www.csss.washington.edu/Papers/wp128.pdf>.

- J. M. Robins. 1986. A new approach to causal inference in mortality studies with sustained exposure periods—Application to control of the healthy worker survivor effect. *Math. Model.* 7, 1393–1512. DOI: [https://doi.org/10.1016/0270-0255\(86\)90088-6](https://doi.org/10.1016/0270-0255(86)90088-6).
- J. M. Robins. 2003. Highly structured stochastic systems. In P. J. Green, N. L. Hjort, and S. Richardson (Eds.), *Semantics of Causal DAG Models and the Identification of Direct and Indirect Effects*. Oxford University Press, Oxford.
- J. M. Robins and S. Greenland. 1992. Identifiability and exchangeability of direct and indirect effects. *Epidemiology* 3, 143–155. DOI: <https://doi.org/10.1097/00001648-199203000-00013>.
- J. M. Robins and T. S. Richardson. 2010. Alternative graphical causal models and the identification of direct effects. In P. E. Shrout, K. M. Keyes, and K. Ornstein (Eds.), *Causality and Psychopathology: Finding the Determinants of Disorders and their Cures*. Oxford University Press, Oxford.
- A. L. Sarvet, K. N. Wanis, M. J. Stensrud, and M. A. Hernán. 2020. A graphical description of partial exchangeability. *Epidemiology* 31, 365–368. DOI: <https://doi.org/10.1097/EDE.0000000000001165>.
- I. Shpitser. 2013. Counterfactual graphical models for longitudinal mediation analysis with unobserved confounding. *Cog. Sci.* 37, 1011–1035. DOI: <https://doi.org/10.1111/cogs.12058>.
- I. Shpitser and J. Pearl. 2006a. Identification of joint interventional distributions in recursive semi-Markovian causal models. In *Proceedings of the 21st National Conference on Artificial Intelligence (AAAI-06)*. AAAI Press, Palo Alto.
- I. Shpitser and J. Pearl. 2006b. Identification of conditional interventional distributions. In *Proceedings of the 22nd Conference on Uncertainty in Artificial Intelligence (UAI-06)*. AUAI Press, Corvallis, OR, 437–444.
- I. Shpitser and E. Sherman. 2018. Identification of personalized effects associated with causal pathways. In *Proceedings of the 34th Annual Conference on Uncertainty in Artificial Intelligence (UAI-18)*.
- I. Shpitser and E. J. Tchetgen Tchetgen. 2016. Causal inference with a graphical hierarchy of interventions. *Ann. Stat.* 44, 6, 2433–2466. DOI: <https://doi.org/10.1214/15-AOS1411>.
- I. Shpitser, E. J. Tchetgen Tchetgen, and R. Andrews. 2017. *Modeling Interference Via Symmetric Treatment Decomposition*. Working paper. <https://arxiv.org/abs/1709.01050>.
- I. Shpitser, T. S. Richardson, and J. M. Robins. 2021. Multivariate counterfactual systems and causal graphical models. This Volume.
- M. Stensrud, M. Hernán, E. Tchetgen Tchetgen, J. M. Robins, V. Didelez, and J. Young. 2020a. Generalized interpretation and identification of separable effects in competing risk settings. *arXiv preprint arXiv:2004.14824*.
- M. J. Stensrud, J. M. Robins, A. Sarvet, E. J. Tchetgen Tchetgen, and J. G. Young. 2020b. Conditional separable effects. *arXiv:2006.15681* [stat.ME].
- M. Stensrud, J. Young, V. Didelez, J. M. Robins, and M. Hernán, 2020c. Separable effects for causal inference in the presence of competing events. *J. Am. Stat. Assoc.* DOI: <https://doi.org/10.1080/01621459.2020.1765783>.

- J. Tian and J. Pearl. 2002. On the testable implications of causal models with hidden variables. In *Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence (UAI-02)*, Vol. 18. AUAI Press, Corvallis, OR, 519–527.
- S. Tikka, A. Hyttinen, and J. Karvanen. 2019. Identifying causal effects via context-specific independence relations. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett (Eds.), *Advances in Neural Information Processing Systems 32*, Curran Associates, Inc., 2800–2810. <http://papers.nips.cc/paper/8547-identifying-causal-effects-via-context-specific-independence-relations.pdf>.
- L. Wittgenstein. 1922. *Tractatus Logico-Philosophicus*. Routledge, London, 1981.

Causality for Machine Learning

Bernhard Schölkopf (Max Planck Institute for Intelligent Systems)

Abstract

Graphical causal inference as pioneered by Judea Pearl arose from research on artificial intelligence (AI), and for a long time had little connection to the field of machine learning. This chapter discusses where links have been and should be established, introducing key concepts along the way. It argues that the hard open problems of machine learning and AI are intrinsically related to causality, and explains how the field is beginning to understand them.

39.1 Introduction

The machine learning community's interest in causality has significantly increased in recent years. My understanding of causality has been shaped by Judea Pearl and a number of collaborators and colleagues, and much of it went into a book written with Dominik Janzing and Jonas Peters [[Peters et al. 2017](#)]. I have spoken about this topic on various occasions,¹ and some of it is in the process of entering the machine learning mainstream, in particular the view that causal modeling can lead to more invariant or robust models. There is excitement about developments at the interface of causality and machine learning, and the present article tries to put my thoughts into writing and draw a bigger picture. I hope it may not only be useful by discussing the importance of causal thinking for AI, but it can also serve as an introduction to some relevant concepts of graphical or structural causal models (SCMs) for a machine learning audience.

1. For example, [Schölkopf \[2017\]](#), talks at the International Conference on Learning Representations, the Asian Conference on Machine Learning, and in machine learning labs that have meanwhile developed an interest in causality (e.g., DeepMind); much of the present paper is essentially a written-out version of these talks.

In spite of all recent successes, if we compare what machine learning can do to what animals accomplish, we observe that the former is rather bad at some crucial feats where animals excel. This includes transfer to new problems, and any form of generalization that is not from one data point to the next one (sampled from the same distribution), but rather from one problem to the next one—both have been termed *generalization*, but the latter is a much harder form thereof. This shortcoming is not too surprising since machine learning often disregards information that animals use heavily: interventions in the world, domain shifts, temporal structure—by and large, we consider these factors a nuisance and try to engineer them away. Finally, machine learning is also bad at *thinking* in the sense of Konrad Lorenz, that is, acting in an imagined space.² I will argue that causality, with its focus on modeling and reasoning about interventions, can make a substantial contribution toward understanding and resolving these issues and thus take the field to the next level. I will do so mostly in non-technical language, for many of the difficulties of this field are of a conceptual nature.

39.2 The Mechanization of Information Processing

The first industrial revolution began in the late 18th century and was triggered by the steam engine and waterpower. The second one started about a century later and was driven by electrification. If we think about it broadly, then both are about how to generate and convert forms of *energy*. Here, the word “generate” is used in a colloquial sense—in physics, energy is a conserved quantity and can thus not be created, but only converted or harvested from other energy forms. Some think we are now in the middle of another revolution, called the digital revolution, the big data revolution, and, more recently, the AI revolution. The transformation, however, really started already in the mid-20th century under the name of cybernetics. It replaced energy by *information*. Like energy, information can be processed by people, but to do it at an industrial scale we needed to invent computers, and to do it intelligently we now use AI. Just like energy, information may actually be a conserved quantity, and we can probably only ever convert and process it, rather than generating it from thin air. When machine learning is applied in industry, we often convert user data into predictions about future user behavior and thus money. Money may ultimately be a form of information—a view not inconsistent with the idea of bitcoins generated by solving cryptographic problems. The first industrial revolutions rendered energy a universal currency [Smil 2017]; the same may be happening to information.

2. “I do not see how thinking should fundamentally differ from such tentative acting in imagined space, taking place only in the brain” [Lorenz 1973].

Like for the energy revolution, one can argue that the present revolution has two components: the first one built on the advent of electronic computers, the development of high-level programming languages, and the birth of the field of computer science, engendered by the vision to create AI by manipulation of symbols. The second one, which we are currently experiencing, relies upon learning. It allows to extract information also from unstructured data, and it automatically infers rules from data rather than relying on humans to conceive of and program these rules. While Judea's approach arose out of classic AI, he was also one of the first to recognize some of the limitations of hard rules programmed by humans, and thus led the way in marrying classic AI with probability theory [Pearl 1988]. This gave birth to graphical models, which were adopted by the machine learning community, yet largely without paying heed to their causal semantics. In recent years, genuine connections between machine learning and causality have emerged, and we will argue that these connections are crucial if we want to make progress on the major open problems of AI.

At the time, the invention of automatic means of processing energy transformed the world. It made human labor redundant in some fields, and it spawned new jobs and markets in others. The first industrial revolution created industries around coal, the second one around electricity. The first part of the information revolution built on this to create electronic computing and the IT industry, and the second part is transforming IT companies into "AI first" as well as creating an industry around data collection and "clickwork." While the latter provides labeled data for the current workhorse of AI, supervised machine learning [Vapnik 1998], one may anticipate that new markets and industries will emerge for causal forms of directed or interventional information, as opposed to just statistical dependences.

The analogy between energy and information is compelling, but our present understanding of information is rather incomplete, as was the concept of energy during the course of the first two industrial revolutions. The profound modern understanding of the concept of energy came with the mathematician Emmy Noether, who understood that energy conservation is due to a symmetry (or covariance) of the fundamental laws of physics: they look the same no matter how we shift the time, in present, past, and future. Einstein, too, was relying on covariance principles when he established the equivalence between energy and mass. Among fundamental physicists, it is widely held that information should also be a conserved quantity, although this brings about certain conundra especially in cosmology.³ One could speculate that the conservation of information might also

3. What happens when information falls into a black hole? According to the *no hair conjecture*, a black hole seen from the outside is fully characterized by its mass, (angular) momentum, and charge.

be a consequence of symmetries—this would be most intriguing, and it would help us understand how different forms of (phenomenological) information relate to each other, and define a unified concept of information. We will below introduce a form of invariance/independence that may be able to play a role in this respect.⁴ The intriguing idea of starting from symmetry transformations and defining objects by their behavior under these transformations has been fruitful not just in physics but also in mathematics [Klein 1872, MacLane 1971].

Clearly, digital goods are different from physical goods in some respects, and the same holds true for information and energy. A purely digital good can be copied at essentially zero cost [Brynjolfsson et al. 2019], unless we move into the quantum world [Wootters and Zurek 1982]. The cost of copying a physical good, on the other hand, can be as high as the cost of the original (e.g., for a piece of gold). In other cases, where the physical good has a non-trivial informational structure (e.g., a complex machine), copying it may be cheaper than the original. In the first phase of the current information revolution, copying was possible for software, and the industry invested significant effort in preventing this. In the second phase, copying extends to datasets, for given the right machine learning algorithm and computational resources others can extract the same information from a dataset. Energy, in contrast, can only be used once.

Just like the first industrial revolutions had a major impact on technology, economy, and society, the same will likely apply for the current changes. It is arguably our information processing ability that is the basis of human dominance on this planet, and thus also of the major impact of humans on our planet. Since it is about information processing, the current revolution is thus potentially even more significant than the first two industrial revolutions. We should strive to use these technologies well to ensure they will contribute toward the solutions of humankind's and our planet's problems. This extends from questions of ethical generation of energy (e.g., environmental concerns) to ethical generation of information (privacy, clickwork) all the way to how we are governed. In the beginnings of the information revolution, cybernetician Stafford Beer worked with Chile's Allende government to build cybernetic governance mechanisms [Medina 2011]. In the

4. Mass seemingly played two fundamentally different roles (inertia and gravitation) until Einstein furnished a deeper connection in general relativity. It is noteworthy that causality introduces a layer of complexity underlying the symmetric notion of statistical mutual information. Discussing source coding and channel coding, Shannon [1959] remarked: "This duality can be pursued further and is related to a duality between past and future and the notions of control and knowledge. Thus we may have knowledge of the past but cannot control it; we may control the future but have no knowledge of it." According to Kierkegaard, "Life can only be understood backwards; but it must be lived forwards."

current data-driven phase of this revolution, China is beginning to use machine learning to observe and incentivize citizens to behave in ways deemed beneficial [Chen and Cheung 2018, Dai 2018]. It is hard to predict where this development takes us—this is science fiction at best, and the best science fiction may provide insightful thoughts on the topic.⁵

39.3 From Statistical to Causal Models

39.3.1 Methods Driven by Independent and Identically Distributed Data

Our community has produced impressive successes with applications of machine learning to big data problems [LeCun et al. 2015]. In these successes there are multiple trends at work: (1) we have massive amounts of data, often from simulations or large-scale human labeling, (2) we use high-capacity machine learning systems (i.e., complex function classes with many adjustable parameters), (3) we employ high-performance computing systems, and finally (often ignored, but crucial when it comes to causality) (4) the problems are independent and identically distributed (IID). The settings are typically either IID to begin with (e.g., image recognition using benchmark datasets), or they are artificially made IID, for example, by carefully collecting the right training set for a given application problem, or by methods such as DeepMind’s “experience replay” [Mnih et al. 2015] where a reinforcement learning agent stores observations in order to later permute them for the purpose of further training. For IID data, strong universal consistency results from statistical learning theory apply, guaranteeing convergence of a learning algorithm to the lowest achievable risk. Such algorithms do exist, for instance, nearest neighbor classifiers and support vector machines [Vapnik 1998, Schölkopf and Smola 2002, Steinwart and Christmann 2008]. Seen in this light, it is not surprising that we can indeed match or surpass human performance if given enough data. Machines often perform poorly, however, when faced with problems that violate the IID assumption yet seem trivial to humans. Vision systems can be grossly misled if an object that is normally recognized with high accuracy is placed in a context that *in the training set* may be negatively correlated with the presence of the object. For instance, such a system may fail to recognize a cow standing on the beach. Even more dramatically, the phenomenon of “adversarial vulnerability” highlights how even tiny but targeted violations of the IID assumption, generated by adding suitably chosen noise to images (imperceptible to humans), can lead to

5. Quoting from Asimov [1951]: “Hari Seldon [...] brought the science of psychohistory to its full development. [...] The individual human being is unpredictable, but the reactions of humans mobs, Seldon found, could be treated statistically.”

dangerous errors such as confusion of traffic signs. Recent years have seen a race between “defense mechanisms” and new attacks that appear shortly after and reaffirm the problem. Overall, it is fair to say that much of the current practice (of solving IID benchmark problems) as well as most theoretical results (about generalization in IID settings) fail to tackle the hard open problem of generalization across problems.

To further understand the way in which the IID assumption is problematic, let us consider a shopping example. Suppose Alice is looking for a laptop rucksack on the Internet (i.e., a rucksack with a padded compartment for a laptop), and the web store’s recommendation system suggests that she should buy a laptop to go along with the rucksack. This seems odd because she probably already has a laptop, otherwise she would not be looking for the rucksack in the first place. In a way, the laptop is the cause and the rucksack is an effect. If I am told whether a customer has bought a laptop, it reduces my uncertainty about whether she also bought a laptop rucksack, and vice versa—and it does so by the same amount (the *mutual information*), so the directionality of cause and effect is lost. It is present, however, in the physical mechanisms generating statistical dependence, for instance the mechanism that makes a customer want to buy a rucksack once she owns a laptop. Recommending an item to buy constitutes an intervention in a system, taking us outside the IID setting. We no longer work with the observational distribution, but a distribution where certain variables or mechanisms have changed. This is the realm of causality.

Reichenbach [1956] clearly articulated the connection between causality and statistical dependence. He postulated the *Common Cause Principle*: if two observables X and Y are statistically dependent, then there exists a variable Z that causally influences both and explains all the dependence in the sense of making them independent when conditioned on Z . As a special case, this variable can coincide with X or Y . Suppose that X is the frequency of storks and Y the human birth rate (in European countries, these have been reported to be correlated). If storks bring the babies, then the correct causal graph is $X \rightarrow Y$. If babies attract storks, it is $X \leftarrow Y$. If there is some other variable that causes both (such as economic development), we have $X \leftarrow Z \rightarrow Y$.

The crucial insight is that without additional assumptions, we cannot distinguish these three cases using observational data. The class of observational distributions over X and Y that can be realized by these models is the same in all three cases. A causal model thus contains genuinely more information than a statistical one.

Given that already the case where we have two observables is hard, one might wonder if the case of more observables is completely hopeless. Surprisingly, this

is not the case: the problem in a certain sense becomes easier, and the reason for this is that in that case there are non-trivial conditional independence properties [Spohn 1978, Dawid 1979, Geiger and Pearl 1990] implied by causal structure. These can be described by using the language of causal graphs or SCMs, merging probabilistic graphical models and the notion of interventions [Spirtes et al. 2000, Pearl 2009a] best described using directed functional parent–child relationships rather than conditionals. While conceptually simple in hindsight, this constituted a major step in the understanding of causality, as later expressed by Pearl [2009a, p. 104]:

We played around with the possibility of replacing the parents–child relationship $P(X_i | \mathbf{PA}_i)$ with its functional counterpart $X_i = f_i(\mathbf{PA}_i, U_i)$ and, suddenly, everything began to fall into place: We finally had a mathematical object to which we could attribute familiar properties of physical mechanisms instead of those slippery epistemic probabilities $P(X_i | \mathbf{PA}_i)$ with which we had been working so long in the study of Bayesian networks.

39.3.2 Structural Causal Models

The SCM viewpoint is intuitive for those machine learning researchers who are more accustomed to thinking in terms of estimating functions rather than probability distributions. In it, we are given a set of *observables* X_1, \dots, X_n (modeled as random variables) associated with the vertices of a directed acyclic graph (DAG). We assume that each observable is the result of an assignment

$$X_i := f_i(\mathbf{PA}_i, U_i), \quad (i = 1, \dots, n), \quad (39.1)$$

using a deterministic function f_i depending on X_i 's parents in the graph (denoted by \mathbf{PA}_i) and on a stochastic *unexplained* variable U_i . Directed edges in the graph represent direct causation, since the parents are connected to X_i by directed edges and through Equation (39.1) directly affect the assignment of X_i . The noise U_i ensures that the overall object (39.1) can represent a general conditional distribution $p(X_i | \mathbf{PA}_i)$, and the set of noises U_1, \dots, U_n are assumed to be *jointly independent*. If they were not, then by the Common Cause Principle there should be another variable that causes their dependence, and thus our model would not be *causally sufficient*.

If we specify the distributions of U_1, \dots, U_n , recursive application of Equation (39.1) allows us to compute the entailed observational joint distribution $p(X_1, \dots, X_n)$. This distribution has structural properties inherited from the graph [Lauritzen 1996, Pearl 2009a]: it satisfies the *causal Markov condition* stating that

conditioned on its parents, each X_j is independent of its non-descendants. Intuitively, we can think of the independent noises as “information probes” that spread through the graph (much like independent elements of gossip can spread through a social network). Their information gets entangled, manifesting itself in a footprint of conditional dependences rendering the possibility to infer aspects of the graph structure from observational data using independence testing. Like in the gossip analogy, the footprint may not be sufficiently characteristic to pin down a unique causal structure. In particular, it certainly is not if there are only two observables since any non-trivial conditional independence statement requires at least three variables.

We have studied the two-variable problem over the last decade. We realized that it can be addressed by making additional assumptions, as not only the graph topology leaves a footprint in the observational distribution but the functions f_i do, too. This point is interesting for machine learning, where much attention is devoted to properties of function classes (e.g., priors or capacity measures), and we shall return to it below. Before doing so, we note two more aspects of Equation (39.1). First, the SCM language makes it straightforward to formalize *interventions* as operations that modify a subset of assignments (39.1), for example, changing U_i , or setting f_i (and thus X_i) to a constant [Spirites et al. 2000, Pearl 2009a]. Second, the graph structure along with the joint independence of the noises implies a canonical factorization of the joint distribution entailed by Equation (39.1) into causal conditionals that we will refer to as the *causal (or disentangled) factorization*,

$$p(X_1, \dots, X_n) = \prod_{i=1}^n p(X_i | \mathbf{PA}_i). \quad (39.2)$$

While many other *entangled factorizations* are possible, for example,

$$p(X_1, \dots, X_n) = \prod_{i=1}^n p(X_i | X_{i+1}, \dots, X_n), \quad (39.3)$$

Equation (39.2) is the only one that decomposes the joint distribution into conditionals corresponding to the structural assignments (39.1). We think of these as the *causal mechanisms* that are responsible for all statistical dependences among the observables. Accordingly, in contrast to Equation (39.3), the disentangled factorization represents the joint distribution as a product of causal mechanisms.

The conceptual basis of statistical learning is a joint distribution $p(X_1, \dots, X_n)$ (where often one of the X_i is a response variable denoted as Y), and we make assumptions about function classes used to approximate, say, a regression $\mathbb{E}(Y | X)$. *Causal learning* considers a richer class of assumptions and seeks to exploit the

fact that the joint distribution possesses a causal factorization (Equation (39.2)). It involves the causal conditionals $p(X_i | \mathbf{PA}_i)$ [i.e., the functions f_i and the distribution of U_i in Equation (39.1)], how these conditionals relate to each other, and interventions or changes that they admit. We shall return to this below.

39.4 Levels of Causal Modeling

Being trained in physics, I like to think of a set of coupled differential equations as the gold standard in modeling physical phenomena. It allows us to predict the future behavior of a system, to reason about the effect of interventions in the system, and—by suitable averaging procedures—to predict *statistical* dependences that are generated by coupled time evolutions.⁶ It also allows us to gain insight in a system, explain its functioning, and in particular read off its causal structure: consider the coupled set of differential equations

$$\frac{d\mathbf{x}}{dt} = f(\mathbf{x}), \quad \mathbf{x} \in \mathbb{R}^d, \quad (39.4)$$

with initial value $\mathbf{x}(t_0) = \mathbf{x}_0$. The Picard–Lindelöf theorem states that at least locally, if f is Lipschitz, there exists a unique solution $\mathbf{x}(t)$. This implies in particular that the immediate future of \mathbf{x} is implied by its past values.

If we formally write this in terms of infinitesimal differentials dt and $d\mathbf{x} = \mathbf{x}(t + dt) - \mathbf{x}(t)$, we get:

$$\mathbf{x}(t + dt) = \mathbf{x}(t) + dt \cdot f(\mathbf{x}(t)). \quad (39.5)$$

From this, we can ascertain which entries of the vector $\mathbf{x}(t)$ determine the future of others $\mathbf{x}(t + dt)$. This tells us that if we have a physical system whose physical mechanisms are correctly described using such an ordinary differential Equation (39.4), solved for $\frac{d\mathbf{x}}{dt}$ (i.e., the derivative only appears on the left-hand side), then its causal structure can be directly read off.

While a differential equation is a rather complete description of a system, a statistical model can be viewed as a much more superficial one. It usually does not talk about time; instead, it tells us how some of the variables allow prediction of others as long as experimental conditions do not change. For example, if we drive a differential equation system with certain types of noise, or we average over time, then it may be the case that statistical dependences between components of \mathbf{x} emerge, and those can then be exploited by machine learning. Such a model does not allow us to predict the effect of interventions; however, its strength is that it can often be learned from data, while a differential equation usually requires an

6. Indeed, one could argue that all statistical dependences in the world are due to such coupling.

Table 39.1 A simple taxonomy of models. The most detailed model (top) is a mechanistic or physical one, usually in terms of differential equations. At the other end of the spectrum (bottom), we have a purely statistical model; this can be learned from data, but it often provides little insight beyond modeling associations between epiphenomena. Causal models can be seen as descriptions that lie in between, abstracting away from physical realism while retaining the power to answer certain interventional or counterfactual questions. See also Mooij et al. [2013] for a formal link between physical models and SCMs

Model	Predict in IID setting	Predict under distr. shift/intervention	Answer counterfactual questions	Obtain physical insight	Learn from data
Mechanistic/physical	Yes	Yes	Yes	Yes	?
Structural causal	Yes	Yes	Yes	?	?
Causal graphical	Yes	Yes	No	?	?
Statistical	Yes	No	No	No	Yes

intelligent human to come up with it. Causal modeling lies in between these two extremes. It aims to provide understanding and predict the effect of interventions. Causal discovery and learning tries to arrive at such models in a data-driven way, using only weak assumptions.⁷ The overall situation is summarized in Table 39.1, adapted from [Peters et al. 2017].

39.5 Independent Causal Mechanisms

We now return to the disentangled factorization [Equation (39.2)] of the joint distribution $p(X_1, \dots, X_n)$. This factorization according to the causal graph is always possible when the U_i are independent, but we will now consider an additional notion of independence relating the factors in Equation (39.2) to one another. We can informally introduce it using an optical illusion known as the Beuchet chair, shown in Figure 39.1.

Whenever we perceive an object, our brain makes the assumption that the object and the mechanism by which the information contained in its light reaches our brain are *independent*. We can violate this by looking at the object from an accidental viewpoint. If we do that, perception may go wrong: in the case of the Beuchet chair, we perceive the three-dimensional (3D) structure of a chair that in reality is not there. The above independence assumption is useful because in

7. It has been pointed out that this task is impossible without assumptions, but this is similar for the (easier) problems of machine learning from finite data. We *always* need assumptions when we perform non-trivial inference from data.

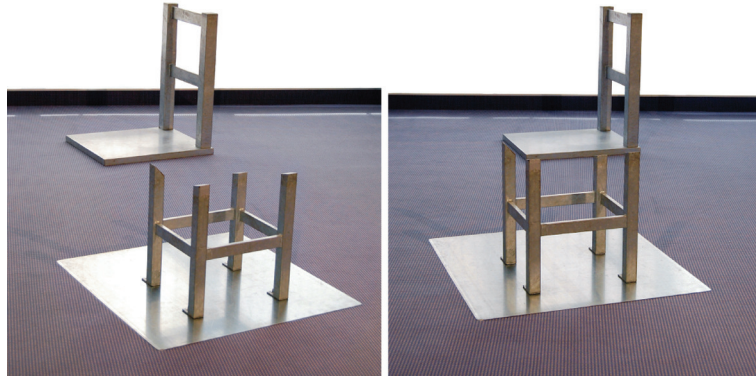


Figure 39.1 Beuchet chair, made up of two separate objects that appear as a chair when viewed from a special vantage point violating the independence between object and perceptual process. (Image courtesy of Markus Elsholz, reprinted from [Peters et al. \[2017\]](#).)

practice it holds most of the time, and our brain thus relies on objects being independent of our vantage point and the illumination. Likewise, there should not be accidental coincidences, 3D structures lining up in two-dimensional (2D), or shadow boundaries coinciding with texture boundaries. In vision research, this is called the generic viewpoint assumption. Likewise, if we move around the object, our vantage point changes, but we assume that the other variables of the overall generative process (e.g., lighting, object position and structure) are unaffected by that. This is an *invariance* implied by the above independence, allowing us to infer 3D information even without stereo vision (“structure from motion”). An example of an extreme violation of this principle would be a head-mounted virtual reality display tracking the position of a perceiver’s head and adjusting the display accordingly. Such a device can create the illusion of visual scenes that do not correspond to reality.

For another example, consider a dataset that consists of altitude A and average annual temperature T of weather stations [[Peters et al. 2017](#)]. A and T are correlated, which we believe is due to the fact that the altitude has a causal effect on the temperature. Suppose we had two such datasets, one for Austria and one for Switzerland. The two joint distributions may be rather different since the marginal distributions $p(A)$ over altitudes will differ. The conditionals $p(T|A)$, however, may be rather similar since they characterize the physical mechanisms that generate temperature from altitude. However, this similarity is lost upon us if we only look at the overall joint distribution, without information about the causal structure

$A \rightarrow T$. The causal factorization $p(A)p(T|A)$ will contain a component $p(T|A)$ that generalizes across countries, while the entangled factorization $p(T)p(A|T)$ will exhibit no such robustness. Cum grano salis, the same applies when we consider interventions in a system. For a model to correctly predict the effect of interventions, it needs to be robust with respect to generalizing from an observational distribution to certain *interventional* distributions.

One can express the above insights as follows [Schölkopf et al. 2012, Peters et al. 2017]:

Independent Causal Mechanisms (ICM) Principle The causal generative process of a system’s variables is composed of autonomous modules that do not inform or influence each other.

In the probabilistic case, this means that the conditional distribution of each variable given its causes (i.e., its mechanism) does not inform or influence the other mechanisms.

This principle subsumes several notions important to causality, including separate intervenability of causal variables, modularity and autonomy of subsystems, and invariance [Pearl 2009a, Peters et al. 2017]. If we have only two variables, it reduces to an independence between the cause distribution and the mechanism producing the effect distribution.

Applied to the causal factorization [Equation (39.2)], the principle tells us that the factors should be independent in the sense that

- (a) changing (or intervening upon) one mechanism $p(X_i | \mathbf{PA}_i)$ does not change the other mechanisms $p(X_j | \mathbf{PA}_j)$ ($i \neq j$), and
- (b) knowing some other mechanisms $p(X_i | \mathbf{PA}_i)$ ($i \neq j$) does not give us information about a mechanism $p(X_j | \mathbf{PA}_j)$.

Our notion of independence thus subsumes two aspects: the former pertaining to influence and the latter to information.

We view any real-world distribution as a product of causal mechanisms. A change in such a distribution (e.g., when moving from one setting/domain to a related one) will always be due to changes in at least one of those mechanisms. Consistent with the independence principle, we hypothesize that smaller *changes tend to manifest themselves in a sparse or local way*, that is, they should usually not affect all factors simultaneously (*sparse mechanism shift*). In contrast, if we consider a non-causal factorization, for example, Equation (39.3), then many terms will be affected simultaneously as we change one of the physical mechanisms responsible for a system’s statistical dependences. Such a factorization may thus be called

entangled, a term that has gained popularity in machine learning [Bengio et al. 2012, Locatello et al. 2018a, Suter et al. 2018].

The notion of invariant, autonomous, and independent mechanisms has appeared in various guises throughout the history of causality research.⁸ Our contribution may be in unifying these notions with the idea of informational independence, and in showing that one can use rather general independence measures [Steudel et al. 2010], a special case of which (algorithmic information) will be described below.

Measures of dependence of mechanisms Note that the dependence of two mechanisms $p(X_i | \mathbf{PA}_i)$ and $p(X_j | \mathbf{PA}_j)$ does not coincide with the statistical dependence of the random variables X_i and X_j . Indeed, in a causal graph, many of the random variables will be dependent even if all the mechanisms are independent.

Intuitively speaking, the independent noise terms U_i provide and parametrize the uncertainty contained in the fact that a mechanism $p(X_i | \mathbf{PA}_i)$ is non-deterministic, and thus ensure that each mechanism adds an independent element of uncertainty. I thus like to think of the ICM Principle as containing the independence of the unexplained noise terms in an SCM [Equation (39.1)] as a special case.⁹ However, it goes beyond this, as the following example illustrates. Consider two variables and structural assignments $X := U$ and $Y := f(X)$. That is, the cause X is a noise variable (with density p_X), and the effect Y is a deterministic function of the cause. Let us moreover assume that the ranges of X and Y are both $[0, 1]$, and f is strictly monotonically increasing. The principle of ICMs then reduces to the independence of p_X and f . Let us consider p_X and the derivative f' as random variables on the probability space $[0, 1]$ with Lebesgue measure, and use their

8. Early work on this was done by Haavelmo [1944], stating the assumption that changing one of the structural assignments leaves the other ones invariant. Hoover [2008] attributes to Herb Simon the *invariance criterion*: the true causal order is the one that is invariant under the right sort of intervention. Aldrich [1989] provides an overview of the historical development of these ideas in economics. He argues that the “most basic question one can ask about a relation should be: How autonomous is it?” [Frisch et al. 1948, Preface]. Pearl [2009a] discusses autonomy in detail, arguing that a causal mechanism remains invariant when other mechanisms are subjected to external influences. He points out that causal discovery methods may best work “in longitudinal studies conducted under slightly varying conditions, where accidental independencies are destroyed and only structural independencies are preserved.” Overviews are provided by Aldrich [1989], Hoover [2008], Pearl [2009a], and Peters et al. [2017, section 2.2].

9. See also Peters et al. [2017]. Note that one can also implement the independence principle by assigning independent priors for the causal mechanisms. We can view ICM as a meta-level independence, akin to assumptions of time-invariance of the laws of physics [Bohm 1957].

correlation as a measure of dependence of mechanisms.¹⁰ It can be shown that for $f \neq id$, independence of p_X and f' implies dependence between p_Y and $(f^{-1})'$ (see Figure 39.2). Other measures are possible and admit information-geometric interpretations. Intuitively, under the ICM assumption, the “irregularity” of the effect distribution becomes a sum of irregularity already present in the input distribution and irregularity introduced by the function, that is, the irregularities of the two mechanisms add up rather than compensating each other, which would not be the case in the anticausal direction (for details, see Janzing et al. [2012]).

Algorithmic independence So far, I have discussed links between causal and statistical structures. The fundamental of the two is the causal structure, since it captures the physical mechanisms that generate statistical dependences in the first place. The statistical structure is an epiphenomenon that follows if we make the unexplained variables random. It is awkward to talk about the (statistical) information contained in a mechanism since deterministic functions in the generic case neither generate nor destroy information. This motivated us to devise an algorithmic model of causal structures in terms of Kolmogorov complexity [Janzing and Schölkopf 2010]. The Kolmogorov complexity (or algorithmic information) of a bit string is essentially the length of its shortest compression on a Turing machine, and thus a measure of its information content. Independence of mechanisms can be defined as vanishing mutual algorithmic information; that is, two conditionals are considered independent if knowing (the shortest compression of) one does not help us achieve a shorter compression of the other one.

Algorithmic information theory provides a natural framework for non-statistical graphical models. Just like the latter are obtained from SCMs by making the unexplained variables U_i random, we obtain algorithmic graphical models by making the U_i bit strings (jointly independent across nodes) and viewing the node X_i as the output of a fixed Turing machine running the program U_i on the input \mathbf{PA}_i . Similar to the statistical case, one can define a local causal Markov condition, a global one in terms of d-separation, and an additive decomposition of the joint Kolmogorov complexity in analogy to Equation (39.2), and prove that they are implied by the SCM [Janzing and Schölkopf 2010]. What is elegant about this approach is that it shows that causality is not intrinsically bound to statistics, and that independence of noises and the independence of mechanisms now coincide since the independent programs play the role of the unexplained noise terms.

10. Other dependence measures have been proposed for high-dimensional linear settings and time series by Janzing et al. [2010], Shajarisales et al. [2015], Besserve et al. [2018a], and Janzing and Schölkopf [2018]; see also Janzing [2019].

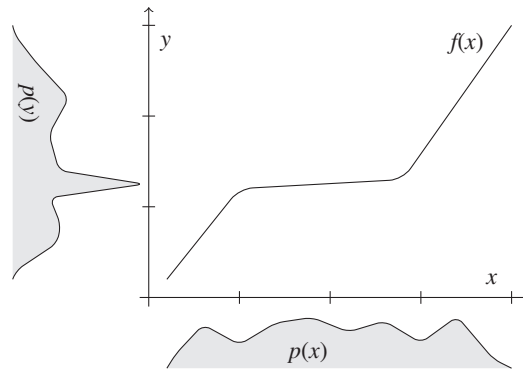


Figure 39.2 If f and p_x are chosen independently, then peaks of p_y tend to occur in regions where f has small slope and f^{-1} has large slope. Thus p_y contains information about f^{-1} . (From Peters et al. [2017].)

The assumption of algorithmically independent mechanisms has intriguing implications for physics, as it turns out to imply the second law of thermodynamics (i.e., the arrow of time) [Janzing et al. 2016]. Consider a process where an incoming ordered beam of particles (the cause) is scattered by an object (the mechanism). Then the outgoing beam (the effect) contains information about the object. That is what makes vision and photography possible: photons contain information about the objects at which they have been scattered. Now we know from physics that, microscopically, time evolution is reversible. Nevertheless, the photons contain information about the object only *after* the scattering. Why is this the case, or in other words, why do photographs show the past rather than the future?

The reason is the independence principle, which we apply to initial state and system dynamics, postulating that the two are algorithmically independent, that is, knowing one does not allow a shorter description of the other one. Then we can prove that the Kolmogorov complexity of the system's state is non-decreasing under the time evolution. If we view Kolmogorov complexity as a measure of entropy, this means that the entropy of the state can only stay constant or increase, amounting to the second law of thermodynamics and providing us with the thermodynamic arrow of time.

Note that this does not contradict microscopic irreversibility of the dynamics; the resulting state after time evolution is clearly *not* independent of the system dynamic: it is precisely the state that when fed to the inverse dynamics would return us to the original state, that is, the ordered particle beam. If we were able to freeze all particles and reverse their momenta, we could thus return to the original configuration without violating our version of the second law.

39.6 Cause–Effect Discovery

Let us return to the problem of causal discovery from observational data. Subject to suitable assumptions such as *faithfulness* [Spirtes et al. 2000], one can sometimes recover aspects of the underlying graph from observations by performing conditional independence tests. However, there are several problems with this approach. One is that in practice our datasets are always finite, and conditional independence testing is a notoriously difficult problem, especially if conditioning sets are continuous and multi-dimensional. So, while in principle the conditional independences implied by the causal Markov condition hold true irrespective of the complexity of the functions appearing in an SCM, for finite datasets conditional independence testing is hard without additional assumptions.¹¹ The other problem is that in the case of only two variables, the ternary concept of conditional independences collapses and the Markov condition thus has no non-trivial implications.

It turns out that both problems can be addressed by making assumptions on function classes. This is typical for machine learning, where it is well-known that finite-sample generalization without assumptions on function classes is impossible. Specifically, although there are learning algorithms that are universally consistent, that is, that approach minimal expected error in the infinite sample limit, for any functional dependence in the data there are cases where this convergence is arbitrarily slow. So, for a given sample size, it will depend on the problem being learned whether we achieve low expected error, and statistical learning theory provides probabilistic guarantees in terms of measures of complexity of function classes [Devroye et al. 1996, Vapnik 1998].

Returning to causality, we provide an intuition why assumptions on the functions in an SCM should be necessary to learn about them from data. Consider a toy SCM with only two observables $X \rightarrow Y$. In this case, Equation (39.1) turns into

$$X = U \tag{39.6}$$

$$Y = f(X, V) \tag{39.7}$$

with $U \perp\!\!\!\perp V$. Now think of V acting as a random selector variable choosing from among a set of functions $\mathcal{F} = \{f_v(x) \equiv f(x, v) \mid v \in \text{supp}(V)\}$. If $f(x, v)$ depends on v in a non-smooth way, it should be hard to glean information about the SCM from

11. We had studied this for some time with Kacper Chwialkowski, Arthur Gretton, Dominik Janzing, Jonas Peters, and Ilya Tolstikhin; a formal result was obtained by Shah and Peters [2018].

a finite dataset, given that V is not observed and it randomly switches between arbitrarily different f_v .¹² This motivates restricting the complexity with which f depends on V . A natural restriction is to assume an *additive noise model*

$$X = U \tag{39.8}$$

$$Y = f(X) + V. \tag{39.9}$$

If f in Equation (39.7) depends smoothly on V , and if V is relatively well concentrated, this can be motivated by a local Taylor expansion argument. It drastically reduces the effective size of the function class—without such assumptions the latter could depend exponentially on the cardinality of the support of V .

Restrictions of function classes not only make it easier to learn functions from data, but it turns out that they can break the symmetry between cause and effect in the two-variable case: one can show that given a distribution over X, Y generated by an additive noise model, one cannot fit an additive noise model in the opposite direction (i.e., with the roles of X and Y interchanged) [Hoyer et al. 2009, Mooij et al. 2009, Kpotufe et al. 2014, Peters et al. 2014, Bauer et al. 2016], cf. also the work of Sun et al. [2006]. This is subject to certain genericity assumptions, and notable exceptions include the case where U, V are Gaussian and f is linear. It generalizes the results of Shimizu et al. [2006] for linear functions, and it can be generalized to include non-linear rescalings [Zhang and Hyvarinen 2009], loops [Mooij et al. 2011], confounders [Janzing et al. 2009], and multi-variable settings [Peters et al. 2011]. We have collected a set of benchmark problems for cause–effect inference, and by now there is a number of methods that can detect causal direction better than chance [Mooij et al. 2016], some of them building on the above Kolmogorov complexity model [Budhathoki and Vreeken 2016], and some directly learning to classify bivariate distributions into causal vs. anticausal [Lopez-Paz et al. 2015]. This development has been championed by Isabelle Guyon whom (along with Andre Elisseeff) I had known from my previous work on kernel methods, and who had moved into causality through her interest in feature selection [Guyon et al. 2007].

Assumptions on function classes have thus helped address the cause–effect inference problem. They can also help address the other weakness of causal discovery methods based on conditional independence testing. Recent progress in

12. Suppose X and Y are binary, and U, V are uniform Bernoulli variables, the latter selecting from $\mathcal{F} = \{id, not\}$ (i.e., identity and negation). In this case, the entailed distribution for Y is uniform, independent of X , even though we have $X \rightarrow Y$. We would be unable to discern $X \rightarrow Y$ from data.

(conditional) independence testing heavily relies on kernel function classes to represent probability distributions in reproducing kernel Hilbert spaces [Gretton et al. 2005a, 2005b, Fukumizu et al. 2008, Zhang et al. 2011, Chalupka et al. 2018, Pfister et al. 2018b].

We have thus gathered some evidence that ideas from machine learning can help tackle causality problems that were previously considered hard. Equally intriguing, however, is the opposite direction: can causality help us improve machine learning? Present-day machine learning (and thus also much of modern AI) is based on statistical modeling, but as these methods becomes pervasive, their limitations are becoming apparent. I will return to this after a short application interlude.

39.7 Half-sibling Regression and Exoplanet Detection

The application described below builds on causal models inspired by additive noise models and the ICM assumption. By a stroke of luck, it enabled a recent breakthrough in astronomy, detailed at the end of the present section.

Launched in 2009, the National Aeronautics and Space Administration (NASA)'s Kepler space telescope initially observed 150,000 stars over four years in search of exoplanet transits. These are events where a planet partially occludes its host star, causing a slight decrease in brightness, often orders of magnitude smaller than the influence of instrument errors. When looking at stellar light curves with our collaborators at New York University, we noticed that not only were these light curves very noisy, but the noise structure was often shared across stars that were light years apart. Since that made direct interaction of the stars impossible, it was clear that the shared information was due to the instrument acting as a confounder. We thus devised a method that (a) predicts a given star of interest from a large set of other stars chosen such that their measurements contain no information about the star's astrophysical signal, and (b) removes that prediction in order to cancel the instrument's influence.¹³ We referred to the method as "half-sibling" regression since target and predictors share a parent, namely the instrument. The method recovers the random variable representing the desired signal almost surely (up to a constant offset), for an additive noise model, and subject to the assumption that the instrument's effect on the star is in principle predictable from the other stars [Schölkopf et al. 2016a].

13. For events that are localized in time (such as exoplanet transits), we further argued that the same applies for suitably chosen past and future values of the star itself, which can thus also be used as predictors.

Meanwhile, the Kepler spacecraft suffered a technical failure, which left it with only two functioning reaction wheels, insufficient for the precise spatial orientation required by the original Kepler mission. NASA decided to use the remaining fuel to make further observations, however the systematic error was significantly larger than before—a godsend for our method designed to remove exactly these errors. We augmented it with models of exoplanet transits and an efficient way to search light curves, leading to the discovery of 36 planet candidates [Foreman-Mackey et al. 2015], of which 21 were subsequently validated as bona fide exoplanets [Montet et al. 2015]. Four years later, astronomers found traces of water in the atmosphere of the exoplanet K2-18b—the first such discovery for an exoplanet in the habitable zone, that is, allowing for liquid water [Benneke et al. 2019, Tsiaras et al. 2019]. The planet turned out to be one that had been first been detected in our work [Foreman-Mackey et al. 2015, exoplanet candidate EPIC 201912552].

39.8 Invariance, Robustness, and Semi-supervised Learning

Around 2009 or 2010, we started getting intrigued by how to use causality for machine learning. In particular, the “neural net tank urban legend”¹⁴ seemed to have something to say about the matter. In this story, a neural net is trained to classify tanks with high accuracy, but subsequently found to have succeeded by focusing on a feature (e.g., time of day or weather) that contained information about the type of tank only due to the data collection process. Such a system would exhibit no robustness when tested on new tanks whose images were taken under different circumstances. My hope was that a classifier incorporating causality could be made invariant with respect to this kind of changes, a topic that I had earlier worked on using non-causal methods [Chapelle and Schölkopf 2002]. We started to think about connections between causality and covariate shift, with the intuition that causal mechanisms should be invariant, and likewise any classifier building on learning these mechanisms. However, many machine learning classifiers were not using causal features as inputs, and indeed, we noticed that they more often seemed to solve anticausal problems, that is, they used effect features to predict a cause.

Our ideas relating to invariance matured during a number of discussions with Dominik, Jonas, Joris Mooij, Kun Zhang, Bob Williamson and others, from a departmental retreat in Ringberg in April 2010 to a Dagstuhl workshop in July 2011. The pressure to bring them to some conclusion was significantly stepped up

14. For a recent account, cf. <https://www.gwern.net/Tanks>.

when I received an invitation to deliver a Posner lecture at the Neural Information Processing Systems conference. At the time, I was involved in founding a new Max Planck Institute, and it was getting hard to carve out enough time to make progress.¹⁵ Dominik and I thus decided to spend a week in a Black Forest holiday house to work on this full time, and during that week in November 2011 we completed a draft manuscript suitably named *invariant.tex*, submitted to the arXiv shortly after [Schölkopf et al. 2011]. The paper argued that causal direction is crucial for certain machine learning problems, that robustness (invariance) to covariate shift is to be expected and transfer is easier for learning problems where we predict effect from cause, and it made a non-trivial prediction for semi-supervised learning (SSL).

39.8.1 Semi-supervised Learning

Suppose our underlying causal graph is $X \rightarrow Y$, and at the same time we are trying to learn a mapping $X \rightarrow Y$. The causal factorization (39.2) for this case is

$$p(X, Y) = p(X)p(Y|X). \quad (39.10)$$

The ICM Principle posits that the modules in a joint distribution's causal decomposition do not inform or influence each other. This means that in particular $p(X)$ should contain no information about $p(Y|X)$, which implies that SSL should be futile in as far as it is using additional information about $p(X)$ (from unlabeled data) to improve our estimate of $p(Y|X = x)$. What about the opposite direction, is there hope that SSL should be possible in that case? It turns out that the answer was yes, due to the work on cause-effect inference using ICMs mentioned in Section 39.5. This work was done by Povilas Daniušis et al. [2010].¹⁶ It introduced a measure of dependence between the input and the conditional of output given input, and showed that if this dependence is zero in the causal direction then it would be strictly positive in the opposite direction. Independence of cause and mechanism in the causal direction would thus imply that in the backward direction (i.e., for anticausal learning) the distribution of the input variable should contain information about the conditional of output given input, that is, the quantity that machine learning is usually concerned with. I had previously worked on SSL [Chapelle et al. 2006], and it was clear that this was exactly the kind of information that SSL required when trying to improve the estimate of output given input by

15. Meanwhile, Google was stepping up their activities in AI, and I even forwent the chance to have a personal meeting with Larry Page to discuss this arranged by Sebastian Thrun.

16. Povilas was an original Erasmus intern visiting from Lithuania. If an experiment was successful, he would sometimes report this with a compact “works.” The project won him the best student paper prize at the Uncertainty in Artificial Intelligence conference.

using unlabeled inputs. We thus predicted that *SSL should be impossible for causal learning problems, but feasible otherwise*, in particular for anticausal ones.

I presented our analysis and the above prediction in the Posner lecture. Although a few activities relating to causality had been present at the conference during the years before, in particular a workshop in 2008 [Guyon et al. 2010], it is probably fair to say that the Posner lecture helped pave the way for causality to enter the machine learning mainstream. Judea, who must have been waiting for this development for some time, sent me a kind e-mail in March 2012, stating “[...] I watched the video of your super-lecture at nips. A miracle.”

A subsequent meta-analysis of published SSL benchmark studies corroborated our prediction, was added to the arXiv report, and the paper was narrowly accepted for the ICML [Schölkopf et al. 2012]. We were intrigued with these results since we felt they provided some structural insight into *physical* properties of learning problems, thus going beyond the applications or methodological advances that machine learning studies usually provided. The line of work provided rather fruitful [Zhang et al. 2013, Weichwald et al. 2014, Zhang et al. 2015, Blöbaum et al. 2016, Gong et al. 2016, Huang et al. 2017, Zhang et al. 2017, Guo et al. 2018, Li et al. 2018a, 2018b, Lipton et al. 2018, Magliacane et al. 2018, Rabanser et al. 2018, Rojas-Carulla et al. 2018, Subbaswamy et al. 2018, Wang et al. 2019] and nicely complementary to studies of Elias Bareinboim and Judea [Bareinboim and Pearl 2014, Pearl and Bareinboim 2015]. When Jonas moved to Zürich to complete and defend his PhD in Statistics at ETH, he carried on with the invariance idea, leading to a thread of work in the statistics community exploiting invariance for causal discovery and other tasks [Peters et al. 2016, Heinze-Deml and Meinshausen 2017, Heinze-Deml et al. 2017, Pfister et al. 2018a].¹⁷

On the SSL side, subsequent developments include further theoretical analyses [Janzing and Schölkopf 2015, Peters et al. 2017, section 5.1.2] and a form of conditional SSL [von Kügelgen et al. 2019]. The view of SSL as exploiting dependencies between a marginal $p(x)$ and a non-causal conditional $p(y|x)$ is consistent with the common assumptions employed to justify SSL [Chapelle et al. 2006]. The *cluster assumption* asserts that the labeling function (which is a property of $p(y|x)$) should not change within clusters of $p(x)$. The *low-density separation assumption* posits that the area where $p(y|x)$ takes the value of 0.5 should have small $p(x)$; and

17. Jonas also played a central role in spawning a thread of causality research in industry. In March 2011, Leon Bottou, working for Microsoft at the time, asked me if I could send him a strong causality student for an internship. Jonas was happy to take up the challenge, contributing to the work of Bottou et al. [2013], an early use of causality to learn large scale interacting systems. Leon, one of the original leaders of the field of deep learning, has since taken a strong interest in causality [Lopez-Paz et al. 2017].

the *semi-supervised smoothness assumption*, applicable also to continuous outputs, states that if two points in a high-density region are close, then so should be the corresponding output values. Note, moreover, that some of the theoretical results in the field use assumptions well-known from causal graphs (even if they do not mention causality): the *co-training theorem* [Blum and Mitchell 1998] makes a statement about learnability from unlabeled data, and relies on an assumption of predictors being conditionally independent given the label, which we would normally expect if the predictors are (only) caused by the label, that is, an anticausal setting. This is nicely consistent with the above findings.

39.8.2 Adversarial Vulnerability

One can hypothesize that causal direction should also have an influence on whether classifiers are vulnerable to *adversarial attacks*. These attacks have recently become popular, and consist of minute changes to inputs, invisible to a human observer yet changing a classifier’s output [Szegedy et al. 2013].

This is related to causality in several ways. First, these attacks clearly constitute violations of the IID assumption that underlies predictive machine learning. If all we want to do is prediction in an IID setting, then statistical learning is fine. In the adversarial setting, however, the modified test examples are not drawn from the same distribution as the training examples—they constitute interventions optimized to reveal the non-robustness of the (anticausal) $p(y|x)$.

The adversarial phenomenon also shows that the kind of robustness current classifiers exhibit is rather different from the one a human exhibits. If we knew both robustness measures, we could try to maximize one while minimizing the other. Current methods can be viewed as crude approximations to this, effectively modeling the human’s robustness as a mathematically simple set, say, an l_p ball of radius $\varepsilon > 0$: they often try to find examples which lead to maximal changes in the classifier’s output, subject to the constraint that they lie in an l_p ball in the pixel metric. This also leads to procedures for adversarial training, which are similar in spirit to old methods for making classifiers invariant by training on “virtual” examples (e.g., Schölkopf and Smola [2002]).

Now consider a factorization of our model into components [cf. Equation (39.3)]. If the components correspond to causal mechanisms, then we expect a certain degree of robustness since these mechanisms are properties of nature. In particular, if we learn a classifier in the causal direction, this should be the case. One may thus hypothesize that for causal learning problems (predicting effect from cause) adversarial examples should be impossible, or at least harder to find [Schölkopf 2017, Kilbertus et al. 2018]. Recent work supports this view: it was shown that a possible defense against adversarial attacks is to solve the anticausal classification

problem by modeling the causal generative direction, a method which in vision is referred to as *analysis by synthesis* [Schott et al. 2019].

More generally, also for graphs with more than two vertices, we can speculate that structures composed of autonomous modules, such as given by a causal factorization [Equation (39.2)], should be relatively robust with respect to swapping out or modifying individual components. We shall return to this shortly.

Robustness should also play a role when studying *strategic behavior*, that is, decisions or actions that take into account the actions of other agents (including AI agents). Consider a system that tries to predict the probability of successfully paying back a credit, based on a set of features. The set could include, for instance, the current debt of a person as well as their address. To get a higher credit score, people could thus change their current debt (by paying it off), or they could change their address by moving to a more affluent neighborhood. The former probably has a positive causal impact on the probability of paying back; for the latter, this is less likely. We could thus build a scoring system that is more robust with respect to such strategic behavior by only using causal features as inputs [Khajehnejad et al. 2019].

39.8.3 Multi-task Learning

Suppose we want to build a system that can solve multiple tasks in multiple environments. Such a model could employ the view of learning as compression. Learning a function f mapping x to y based on a training set $(x_1, y_1), \dots, (x_n, y_n)$ can be viewed as conditional compression of y given x . The idea is that we would like to find the most compact system that can recover y_1, \dots, y_n given x_1, \dots, x_n . Suppose Alice wants to communicate the labels to Bob, given that both know the inputs. First, they agree on a finite set of functions \mathcal{F} that they will use. Then Alice picks the best function from the set, and tells Bob which one it is (the number of bits required will depend on the size of the set, and possibly on prior probabilities agreed between Alice and Bob). In addition, she might have to tell him the indices i of those inputs for which the function does not correctly classify X_i , that is, for which $f(x_i) \neq y_i$. There is a trade-off between choosing a huge function class (in which case it will cost many bits to encode the index of the function) and allowing too many training errors (which need to be encoded separately). It turns out that this trade-off beautifully maps to standard VC bounds from statistical learning theory [Vapnik 1995]. One could imagine generalizing this to a multi-task setting: suppose we have multiple datasets, sampled from similar but not identical SCMs. If the SCMs share most of the components, then we could compress multiple datasets (sampled from multiple SCMs) by encoding the functions in the SCMs, and it is plausible that the correct structure (in the two-variables case, this would

amount to the correct causal direction) should be the most compact one since it would be one where many functions are shared across datasets, and thus need only be encoded once.

39.8.4 Reinforcement Learning

The program to move statistical learning toward causal learning has links to reinforcement learning (RL), a sub-field of machine learning. RL used to be (and still often is) considered a field that has trouble with real-world high-dimensional data, one reason being that feedback in the form of a reinforcement signal is relatively sparse when compared to label information in supervised learning. The DeepQ agent [Mnih et al. 2015] yielded results that the community would not have considered possible at the time, yet it still has major weaknesses when compared to animate intelligence. Two major issues can be stated in terms of questions [Schölkopf 2017]; cf. also Schölkopf [2015]:

Question 1: why is RL on the original high-dimensional ATARI games harder than on downsampled versions? For humans, reducing the resolution of a game screen would make the problem harder, yet this is exactly what was done to make the DeepQ system work. Animals likely have methods to identify objects (in computer game lingo, “sprites”) by grouping pixels according to “common fate” (known from Gestalt psychology) or common response to intervention. This question thus is related to the question of what constitutes an object, which concerns not only perception but also concerns how we interact with the world. We can pick up one object, but not half an object. Objects thus also correspond to modular structures that can be separately intervened upon or manipulated. The idea that objects are defined by their behavior under transformation is a profound one not only in psychology but also in mathematics, cf. Klein [1872] and MacLane [1971].

Question 2: why is RL easier if we permute the replayed data? As an agent moves about in the world, it influences the kind of data it gets to see, and thus the statistics change over time. This violates the IID assumption, and as mentioned earlier, the DeepQ agent stores and re-trains on past data (a process the authors liken to dreaming) in order to be able to employ standard IID function learning techniques. However, temporal order contains information that animate intelligence uses. Information is not only contained in temporal order but also in the fact that slow changes of the statistics effectively create a multi-domain setting. Multi-domain data have been shown to help identify causal (and thus robust) features, and more generally in the search for causal structure by looking for invariances [Peters et al. 2017]. This could enable RL agents to find robust components in their models that are likely to generalize to other parts of the state space. One way to do this is to employ model-based RL using SCMs, an approach that can help

address a problem of confounding in RL where time-varying and time-invariant unobserved confounders influence both actions and rewards [Lu et al. 2018]. In such an approach, non-stationarities would be a feature rather than a bug, and agents would actively seek out regions that are different from the known ones in order to challenge their existing model and understand which components are robust. This search can be viewed and potentially analyzed as a form of *intrinsic motivation*, a concept related to latent learning in Ethology that has been gaining traction in RL [Chentanez et al. 2005].

Finally, a large open area in causal learning is the connection to dynamics. While we may naively think that causality is always about time, most existing causal models do not (and need not) talk about time. For instance, returning to our example of altitude and temperature, there is an underlying temporal physical process that ensures that higher places tend to be colder. On the level of microscopic equations of motion for the involved particles, there is a clear causal structure (as described above, a differential equation specifies exactly which past values affect the current value of a variable). However, when we talk about the dependence or causality between altitude and temperature, we need not worry about the details of this temporal structure—we are given a dataset where time does not appear, and we can reason about how that dataset would look if we were to intervene on temperature or altitude. It is intriguing to think about how to build bridges between these different levels of description. Some progress has been made in deriving SCMs that describe the interventional behavior of a coupled system that is in an equilibrium state and perturbed in an “adiabatic” way [Mooij et al. 2013], with generalizations to oscillatory systems [Rubenstein et al. 2018]. There is no fundamental reason why simple SCMs should be derivable in general. Rather, an SCM is a high-level abstraction of an underlying system of differential equations, and such an equation can only be derived if suitable high-level variables can be defined [Rubenstein et al. 2017], which is probably the exception rather than the rule.

RL is closer to causality research than the machine learning mainstream in that it sometimes effectively directly estimates do-probabilities. For example, on-policy learning estimates do-probabilities for the interventions specified by the policy (note that these may not be hard interventions if the policy depends on other variables). However, as soon as off-policy learning is considered, in particular in the batch (or observational) setting [Lange et al. 2012], issues of causality become subtle [Gottesman et al. 2018, Lu et al. 2018]. Recent work devoted to the field between RL and causality includes Bareinboim et al. [2015], Bengio et al. [2017], Buesing et al. [2018], Lu et al. [2018], Dasgupta et al. [2019], and Zhang and Bareinboim [2019].

39.9 Causal Representation Learning

Traditional causal discovery and reasoning assumes that the units are random variables connected by a causal graph. Real-world observations, however, are usually not structured into those units to begin with, for example, objects in images [Lopez-Paz et al. 2017]. The emerging field of causal representation learning hence strives to learn these variables from data, much like machine learning went beyond symbolic AI in not requiring that the symbols that algorithms manipulate be given a priori (cf. Bonet and Geffner [2019]). Defining objects or variables that are related by causal models can amount to coarse-graining of more detailed models of the world. Subject to appropriate conditions, structural models can arise from coarse-graining of microscopic models, including microscopic structural equation models [Rubenstein et al. 2017], ordinary differential equations [Rubenstein et al. 2018], and temporally aggregated time series [Gong et al. 2017]. Although every causal models in economics, medicine, or psychology uses variables that are abstractions of more elementary concepts, it is challenging to state general conditions under which coarse-grained variables admit causal models with well-defined interventions [Chalupka et al. 2015, Rubenstein et al. 2017].

The task of identifying suitable units that admit causal models is challenging for both human and machine intelligence, but it aligns with the general goal of modern machine learning to learn meaningful representations for data, where meaningful can mean *robust, transferable, interpretable, explainable, or fair* [Kilbertus et al. 2017, Kusner et al. 2017, Zhang and Bareinboim 2018]. To combine structural causal modeling [Equation (39.1)] and representation learning, we should strive to embed an SCM into larger machine learning models whose inputs and outputs may be high-dimensional and unstructured, but whose inner workings are at least partly governed by an SCM. A way to do so is to realize the unexplained variables as (latent) noise variables in a generative model. Note, moreover, that there is a natural connection between SCMs and the modern generative models: they both use what has been called the *reparametrization trick* [Kingma and Welling 2013], consisting of making desired randomness an (exogenous) input to the model (in an SCM, these are the unexplained variables) rather than an intrinsic component.

39.9.1 Learning Transferable Mechanisms

An artificial or natural agent in a complex world is faced with limited resources. This concerns training data, that is, we only have limited data for each individual task/domain, and thus need to find ways of pooling/re-using data, in stark contrast to the current industry practice of large-scale labeling work done by humans. It also concerns computational resources: animals have constraints on the size of their brains, and evolutionary neuroscience knows many examples where brain

regions get re-purposed. Similar constraints on size and energy apply as ML methods get embedded in (small) physical devices that may be battery powered. Future AI models that robustly solve a range of problems in the real world will thus likely need to re-use components, which requires that the components are robust across tasks and environments [Schölkopf et al. 2016b]. An elegant way to do this is to employ a modular structure that mirrors a corresponding modularity in the world. In other words, if the world is indeed modular, in the sense that different components of the world play roles across a range of environments, tasks, and settings, then it would be prudent for a model to employ corresponding modules [Goyal et al. 2019]. For instance, if variations of natural lighting (the position of the sun, clouds, etc.) imply that the visual environment can appear in brightness conditions spanning several orders of magnitude, then visual processing algorithms in our nervous system should employ methods that can factor out these variations, rather than building separate sets of face recognizers, say, for every lighting condition. If our brain were to compensate for the lighting changes by a gain control mechanism, say, then this mechanism in itself need not have anything to do with the physical mechanisms bringing about brightness differences. It would, however, play a role in a modular structure that corresponds to the role the physical mechanisms play in the world's modular structure. This could produce a bias toward models that exhibit certain forms of structural isomorphism to a world that we cannot directly recognize, which would be rather intriguing, given that ultimately our brains do nothing but turn neuronal signals into other neuronal signals.

A sensible inductive bias to learn such models is to look for ICMs [Locatello et al. 2018b], and competitive training can play a role in this: for a pattern recognition task, Parascandolo et al. [2018] show that learning causal models that contain independent mechanisms helps in transferring modules across substantially different domains. In this work, handwritten characters are distorted by a set of unknown mechanisms including translations, noise, and contrast inversion. A neural network attempts to undo these transformations by means of a set of modules that over time specialize on one mechanism each. For any input, each module attempts to produce a corrected output, and a discriminator is used to tell which one performs best. The winning module gets trained by gradient descent to further improve its performance on that input. It is shown that the final system has learned mechanisms such as translation, inversion, or denoising, and that these mechanisms transfer also to data from other distributions, such as Sanskrit characters. This has recently been taken to the next step, embedding a set of dynamic modules into a recurrent neural network, coordinated by a so-called attention mechanism [Goyal et al. 2019]. This allows learning modules whose dynamics operate independently much of the time but occasionally interact with each other.

39.9.2 Learning Disentangled Representations

We have earlier discussed the ICM Principle implying both the independence of the SCM noise terms in Equation (39.1) and thus the feasibility of the disentangled representation

$$p(S_1, \dots, S_n) = \prod_{i=1}^n p(S_i | \mathbf{PA}_i) \quad (39.11)$$

as well as the property that the conditionals $p(S_i | \mathbf{PA}_i)$ be independently manipulable and largely invariant across related problems. Suppose we seek to reconstruct such a *disentangled representation using independent mechanisms* [Equation (39.11)] from data, but the causal variables S_i are not provided to us a priori. Rather, we are given (possibly high-dimensional) $X = (X_1, \dots, X_d)$ (below, we think of X as an image with pixels X_1, \dots, X_d), from which we should construct causal variables S_1, \dots, S_n ($n \ll d$) as well as mechanisms, cf. Equation (39.1),

$$S_i := f_i(\mathbf{PA}_i, U_i), (i = 1, \dots, n), \quad (39.12)$$

modeling the causal relationships among the S_i . To this end, as a first step, we can use an *encoder* $q : \mathbb{R}^d \rightarrow \mathbb{R}^n$ taking X to a latent “bottleneck” representation comprising the unexplained noise variables $U = (U_1, \dots, U_n)$. The next step is the mapping $f(U)$ determined by the structural assignments f_1, \dots, f_n .¹⁸ Finally, we apply a *decoder* $p : \mathbb{R}^n \rightarrow \mathbb{R}^d$. If n is sufficiently large, the system can be trained using reconstruction error to satisfy $p \circ f \circ q \approx id$ on the observed images.¹⁹ To make it causal, we use the ICM Principle, that is, we should make the U_i statistically independent, and we should make the mechanisms independent. This can be done by ensuring that they be largely invariant across problems (one could call this *sparse causal shift training*), or that they can be independently intervened upon: if we manipulate some of them, they should thus still produce valid images (one could refer to this as *counterfactual training*). The latter could be trained using the discriminator of a generative adversarial network [Goodfellow et al. 2014].

While we ideally manipulate causal variables or mechanisms, we discuss the special case of intervening upon the latent noise variables.²⁰ One way to intervene

18. Note that for a DAG, recursive substitution of structural assignments reduces them to functions of the noise variables only. Using recurrent networks, cyclic systems may be dealt with.

19. If the causal graph is known, the topology of a neural network implementing f can be fixed accordingly; if not, the neural network decoder learns the composition $\tilde{p} = p \circ f$. In practice, one may not know f , and thus only learn an autoencoder $\tilde{p} \circ q$, where the causal graph effectively becomes an unspecified part of \tilde{p} . By choosing the network topology, one can ensure that each noise should only feed into one subsequent unit (using connections skipping layers), and that all DAGs can be learnt.

20. Interventions on the S_i can be done accordingly, including the case of decoders without encoder (e.g., GANs).

is to replace noise variables with the corresponding values computed from other input images, a procedure that has been referred to as hybridization by [Besserve et al. \[2018b\]](#). In the extreme case, we can hybridize latent vectors where *each* component is computed from another training example. For an IID training set, these latent vectors have statistically independent components by construction.

In such an architecture, the encoder is an anticausal mapping that recognizes or reconstructs causal drivers in the world. These should be such that in terms of them mechanisms can be formulated that are transferable (e.g., across tasks). The decoder establishes the connection between the low-dimensional latent representation (of the noises driving the causal model) and the high-dimensional world; this part constitutes a causal generative image model. The ICM assumption implies that if the latent representation reconstructs the (noises driving the) true causal variables, then interventions on those noises (and the mechanisms driven by them) are permissible and lead to valid generation of image data.

39.9.3 Learning Interventional World Models and Reasoning

Modern representation learning excels at learning representations of data that preserve relevant statistical properties [[Bengio et al. 2012](#), [LeCun et al. 2015](#)]. It does so, however, without taking into account causal properties of the variables, that is, it does not care about the interventional properties of the variables it analyzes or reconstructs. I expect that going forward causality will play a major role in taking representation learning to the next level, moving beyond the representation of statistical dependence structures toward models that support intervention, planning, and reasoning, realizing Konrad Lorenz's notion of *thinking as acting in an imagined space*. This ultimately requires the ability to reflect back on one's actions and envision alternative scenarios, possibly necessitating (the illusion of) free will [[Pearl 2009b](#)]. The biological function of self-consciousness may be related to the need for a variable representing oneself in one's Lorenzian *imagined space*, and free will may then be a means to communicate about actions taken by that variable, crucial for social and cultural learning, a topic that has not yet entered the stage of machine learning research although it is at the core of human intelligence [[Henrich 2016](#)].

39.10 Personal Notes and Conclusion

My first conscious encounter with Judea Pearl was in 2001, at a symposium on the *Interface of Computing Sciences and Statistics*.²¹ We both spoke at this symposium, and I recall his talk, formalizing an area of scientific inquiry that I had previously considered solidly part of the realm of philosophy. It stirred the same

21. <https://www.ics.uci.edu/~interfac/>.

fascination that had attracted me to the research that I was doing at that time, in statistical learning theory and kernel methods. I had a background in mathematics and physics, had dabbled in neural networks, and was impressed when in 1994 I met Vladimir Vapnik, who taught me a statistical theory underlying the philosophical problems of induction and generalization. Judea Pearl, another giant of our still young field of AI, seemed to be doing the same on a rather different but equally fascinating problem. Like Vladimir, Judea left me with a lasting impression as someone who has mastered not just technicalities but has gained access to profound philosophical understanding. With kernel methods and learning theory taking off, I did not manage to go into depth on causality at the time. I did follow some of the work in graphical models which became a staple in machine learning, and I knew that although most researchers shied away from presenting these models as causal, this interpretation existed and formed a conceptual motivation for that field.

I was brought in touch with causality research for the second time in 2004 by my study friend Dominik Janzing. He was at the time working on quantum information, and spoke about causality in a course he taught in Karlsruhe. The student Xiaohai Sun followed that lecture and convinced Dominik to start working with him on a project. Eventually, the question of a PhD project came up, and Dominik (who felt his own field was too far from that) decided to ask me whether a joint supervision would make sense. At the time, Vladimir Vapnik was visiting my lab, and after a long conversation, he agreed this could be interesting (“you should decide if you want to play this game”—by his standards, a fairly enthusiastic endorsement). I decided to take the risk, Xiaohai became a student in my lab in Tübingen, and in 2007, Dominik joined us. We also recruited the student Jonas Peters, who had taken part in a summer course I had taught in 2006, as well as the postdocs Joris Mooij and Kun Zhang, both independently driven toward the problem of causality. With Andre Elisseeff and Steffen Lauritzen, Dominik and I wrote a proposal to organize a causality workshop in Dagstuhl. This workshop took place in 2009, and helped us become members of the causality community; it was where I first personally met Peter Spirtes.

I feel fortunate to have had such a strong team of people to do this work (including many whom I did not mention by name above), and I believe we have made a contribution to modern causality research and especially its links to machine learning: both by using learning methods to develop data-driven causal methods, and by using inspiration from causality to better understand machine learning and develop new learning methods. In that respect, representation learning and disentanglement are intriguing fields. I recall a number of discussions with Yoshua Bengio when I was a member of the review panel and advisory board of the CIFAR

program. He was driven by the goal to disentangle the underlying factors of variation in data using deep learning, and I was arguing that this is a causal question. Our opinions have since then converged, and research has started to appear that combines both fields [Goudet et al. 2017, Locatello et al. 2018a, Suter et al. 2018, Bengio et al. 2019, Goyal et al. 2019].

All this is still in its infancy, and the above account is personal and thus biased—I apologize for any omissions. With the current hype around machine learning, there is much to say in favor of some humility toward what machine learning can do, and thus toward the current state of AI—the hard problems have not been solved yet, making basic research in this field all the more exciting.

Acknowledgments

Many thanks to all past and present members of the Tübingen causality team, without whose work and insights this article would not exist, in particular to Dominik Janzing and Chaochao Lu who read a version of the manuscript. The text has also benefitted from discussions with Elias Bareinboim, Yoshua Bengio, Christoph Bohle, Leon Bottou, Anirudh Goyal, Isabelle Guyon, Judea Pearl, and Vladimir Vapnik. Wouter van Amsterdam, and Julius von Kügelgen have pointed out typos that have been corrected in this second version.

References

- J. Aldrich. 1989. Autonomy. *Oxf. Econ. Pap.* 41, 15–34. DOI: <https://doi.org/10.1093/oxfordjournals.oep.a041889>.
- I. Asimov. 1951. *Foundation*. Gnome Press, New York.
- E. Bareinboim and J. Pearl. 2014. Transportability from multiple environments with limited experiments: Completeness results. In *Advances in Neural Information Processing Systems 27*, 280–288.
- E. Bareinboim, A. Forney, and J. Pearl. 2015. Bandits with unobserved confounders: A causal approach. In *Advances in Neural Information Processing Systems 28*, 1342–1350.
- S. Bauer, B. Schölkopf, and J. Peters. 2016. The arrow of time in multivariate time series. In *Proceedings of the 33rd International Conference on Machine Learning*, Vol. 48. *JMLR Workshop and Conference Proceedings*, 2043–2051.
- Y. Bengio, A. Courville, and P. Vincent. 2012. Representation learning: A review and new perspectives. *IEEE Trans. Softw. Eng.* 35, 8, 1798–1828. DOI: <https://doi.org/10.1109/TPAMI.2013.50>.
- E. Bengio, V. Thomas, J. Pineau, D. Precup, and Y. Bengio. 2017. Independently controllable features. arXiv:1703.07718.
- Y. Bengio, T. Deleu, N. Rahaman, R. Ke, S. Lachapelle, O. Bilaniuk, A. Goyal, and C. Pal. 2019. A meta-transfer objective for learning to disentangle causal mechanisms. arXiv:1901.10912.

- B. Benneke, I. Wong, C. Piaulet, H. A. Knutson, I. J. M. Crossfield, J. Lothringer, C. V. Morley, P. Gao, T. P. Greene, C. Dressing, D. Dragomir, A. W. Howard, P. R. McCullough, E. M. R. K. J. Fortney, and J. Fraine. 2019. Water vapor on the habitable-zone exoplanet K2-18b. arXiv:1909.04642.
- M. Besserve, N. Shajarisales, B. Schölkopf, and D. Janzing. 2018a. Group invariance principles for causal generative models. In *Proceedings of the 21st International Conference on Artificial Intelligence and Statistics (AISTATS)*. 557–565.
- M. Besserve, R. Sun, and B. Schölkopf. 2018b. Counterfactuals uncover the modular structure of deep generative models. arXiv:1812.03253.
- P. Blöbaum, T. Washio, and S. Shimizu. 2016. Error asymmetry in causal and anticausal regression. *Behaviormetrika* 2017. arXiv:1610.03263. DOI: <https://doi.org/10.1007/s41237-017-0022-z>.
- A. Blum and T. Mitchell. 1998. Combining labeled and unlabeled data with co-training. In *Proceedings of the 11th Annual Conference on Computational Learning Theory*. ACM, New York, 92–100. DOI: <https://doi.org/10.1145/279943.279962>.
- D. Bohm. 1957. *Causality and Chance in Modern Physics*. Routledge & Kegan Paul, London.
- B. Bonet and H. Geffner. 2019. Learning first-order symbolic representations for planning from the structure of the state space. arXiv:1909.05546.
- L. Bottou, J. Peters, J. Quiñonero-Candela, D. X. Charles, D. M. Chickering, E. Portugaly, D. Ray, P. Simard, and E. Snelson. 2013. Counterfactual reasoning and learning systems: The example of computational advertising. *J. Mach. Learn. Res.* 14, 3207–3260.
- E. Brynjolfsson, A. Collis, W. E. Diewert, F. Eggers, and K. J. Fox. 2019. GDP-B: Accounting for the value of new and free goods in the digital economy. Working Paper 25695, National Bureau of Economic Research.
- K. Budhathoki and J. Vreeken. 2016. Causal inference by compression. In *IEEE 16th International Conference on Data Mining*. DOI: <https://doi.org/10.1109/ICDM.2016.0015>.
- L. Buesing, T. Weber, Y. Zwols, S. Racaniere, A. Guez, J.-B. Lespiau, and N. Heess. 2018. Woulda, coulda, shoulda: Counterfactually-guided policy search. arXiv:1811.06272.
- K. Chalupka, P. Perona, and F. Eberhardt. 2015. Multi-level cause–effect systems. arXiv:1512.07942.
- K. Chalupka, P. Perona, and F. Eberhardt. 2018. Fast conditional independence test for vector variables with large sample sizes. arXiv:1804.02747.
- O. Chapelle and B. Schölkopf. 2002. Incorporating invariances in nonlinear SVMs. In T. G. Dietterich, S. Becker, and Z. Ghahramani (Eds.), *Advances in Neural Information Processing Systems 14*. MIT Press, Cambridge, MA, 609–616. DOI: <https://doi.org/10.7551/mitpress/1120.003.0083>.
- O. Chapelle, B. Schölkopf, and A. Zien (Eds.). 2006. *Semi-Supervised Learning*. MIT Press, Cambridge, MA. <http://www.kyb.tuebingen.mpg.de/ssl-book/>. DOI: <https://doi.org/10.7551/mitpress/9780262033589.001.0001>.

- Y. Chen and A. Cheung. 2018. The transparent self under big data profiling: Privacy and Chinese legislation on the social credit system. *J. Comp. Law* 12, 2, 356–378. DOI: <http://dx.doi.org/10.2139/ssrn.2992537>.
- N. Chentanez, A. G. Barto, and S. P. Singh. 2005. Intrinsically motivated reinforcement learning. In *Advances in Neural Information Processing Systems 17*. MIT Press, 1281–1288.
- X. Dai. 2018. Toward a reputation state: The social credit system project of China. <https://ssrn.com/abstract=3193577>. DOI: <http://dx.doi.org/10.2139/ssrn.3193577>.
- P. Daniušis, D. Janzing, J. M. Mooij, J. Zscheischler, B. Steudel, K. Zhang, and B. Schölkopf. 2010. Inferring deterministic causal relations. In *Proceedings of the 26th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*. 143–150.
- I. Dasgupta, J. Wang, S. Chiappa, J. Mitrovic, P. Ortega, D. Raposo, E. Hughes, P. Battaglia, M. Botvinick, and Z. Kurth-Nelson. 2019. Causal reasoning from meta-reinforcement learning. arXiv:1901.08162.
- A. P. Dawid. 1979. Conditional independence in statistical theory. *J. R. Stat. Soc. B* 41, 1, 1–31.
- L. Devroye, L. Györfi, and G. Lugosi. 1996. *A Probabilistic Theory of Pattern Recognition*, Vol. 31: Applications of Mathematics. Springer, New York. DOI: <http://dx.doi.org/10.1007/978-1-4612-0711-5>.
- D. Foreman-Mackey, B. T. Montet, D. W. Hogg, T. D. Morton, D. Wang, and B. Schölkopf. 2015. A systematic search for transiting planets in the K2 data. *Astrophys. J.* 806, 2. <http://stacks.iop.org/0004-637X/806/i=2/a=215>. DOI: <http://dx.doi.org/10.1088/0004-637X/806/2/215>.
- R. Frisch, T. Haavelmo, T. Koopmans, and J. Tinbergen. 1948. *Autonomy of Economic Relations*. Universitets Sosialøkonomiske Institutt, Oslo, Norway.
- K. Fukumizu, A. Gretton, X. Sun, and B. Schölkopf. 2008. Kernel measures of conditional dependence. In *Advances in Neural Information Processing Systems 20*. 489–496.
- D. Geiger and J. Pearl. 1990. Logical and algorithmic properties of independence and their application to Bayesian networks. *Ann. Math. Artif. Intell.* 2, 165–178. DOI: <https://doi.org/10.1007/BF01531004>.
- M. Gong, K. Zhang, T. Liu, D. Tao, C. Glymour, and B. Schölkopf. 2016. Domain adaptation with conditional transferable components. In *Proceedings of the 33rd International Conference on Machine Learning*. 2839–2848.
- M. Gong, K. Zhang, B. Schölkopf, C. Glymour, and D. Tao. 2017. Causal discovery from temporally aggregated time series. In *Proceedings of the 33rd Conference on Uncertainty in Artificial Intelligence (UAI)*. ID 269.
- I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. 2014. Generative adversarial nets. In *Advances in Neural Information Processing Systems 27*. Curran Associates, Inc., 2672–2680.
- O. Gottesman, F. Johansson, J. Meier, J. Dent, D. Lee, S. Srinivasan, L. Zhang, Y. Ding, D. Wihl, X. Peng, J. Yao, I. Lage, C. Mosch, L. wei H. Lehman, M. Komorowski,

- M. Komorowski, A. Faisal, L. A. Celi, D. Sontag, and F. Doshi-Velez. 2018. Evaluating reinforcement learning algorithms in observational health settings. arXiv:1805.12298.
- O. Goudet, D. Kalainathan, P. Caillou, I. Guyon, D. Lopez-Paz, and M. Sebag. 2017. Causal generative neural networks. arXiv:1711.08936.
- A. Goyal, A. Lamb, J. Hoffmann, S. Sodhani, S. Levine, Y. Bengio, and B. Schölkopf. 2019. Recurrent independent mechanisms. arXiv:1909.10893.
- A. Gretton, O. Bousquet, A. Smola, and B. Schölkopf. 2005a. Measuring statistical dependence with Hilbert–Schmidt norms. In *Algorithmic Learning Theory*. Springer-Verlag, 63–78. DOI: https://doi.org/10.1007/11564089_7.
- A. Gretton, R. Herbrich, A. Smola, O. Bousquet, and B. Schölkopf. 2005b. Kernel methods for measuring independence. *J. Mach. Learn. Res.* 6, 2075–2129.
- R. Guo, L. Cheng, J. Li, P. R. Hahn, and H. Liu. 2018. A survey of learning causality with data: Problems and methods. arXiv:1809.09337. DOI: <https://doi.org/10.1145/3397269>.
- I. Guyon, C. Aliferis, and A. Elisseeff. 2007. Causal feature selection. In *Computational Methods of Feature Selection*. Chapman and Hall/CRC, Boca Raton, FL, 75–97.
- I. Guyon, D. Janzing, and B. Schölkopf. 2010. Causality: Objectives and assessment. In I. Guyon, D. Janzing, and B. Schölkopf (Eds.), *JMLR Workshop and Conference Proceedings*. Vol. 6. MIT Press, Cambridge, MA, 1–42.
- T. Haavelmo. 1944. The probability approach in econometrics. *Econometrica* 12, (supplement), S1–S115.
- C. Heinze-Deml and N. Meinshausen. 2017. Conditional variance penalties and domain shift robustness. arXiv:1710.11469.
- C. Heinze-Deml, J. Peters, and N. Meinshausen. 2017. Invariant causal prediction for nonlinear models. arXiv:1706.08576.
- J. Henrich. 2016. *The Secret of our Success*. Princeton University Press, Princeton, NJ.
- K. D. Hoover. 2008. Causality in economics and econometrics. In S. N. Durlauf and L. E. Blume (Eds.), *The New Palgrave Dictionary of Economics* (2nd. ed.). Palgrave Macmillan, Basingstoke, UK.
- P. O. Hoyer, D. Janzing, J. M. Mooij, J. Peters, and B. Schölkopf. 2009. Nonlinear causal discovery with additive noise models. In *Advances in Neural Information Processing Systems 21 (NIPS)*. 689–696.
- B. Huang, K. Zhang, J. Zhang, R. Sanchez-Romero, C. Glymour, and B. Schölkopf. 2017. Behind distribution shift: Mining driving forces of changes and causal arrows. In *IEEE 17th International Conference on Data Mining (ICDM 2017)*. 913–918. DOI: <https://doi.org/10.1109/ICDM.2017.114>.
- D. Janzing. 2019. Causal regularization. In *Advances in Neural Information Processing Systems 33*.
- D. Janzing and B. Schölkopf. 2010. Causal inference using the algorithmic Markov condition. *IEEE Trans. Inf. Theory* 56, 10, 5168–5194. DOI: <https://doi.org/10.1109/TIT.2010.2060095>.

- D. Janzing and B. Schölkopf. 2015. Semi-supervised interpolation in an anticausal learning scenario. *J. Mach. Learn. Res.* 16, 1923–1948.
- D. Janzing and B. Schölkopf. 2018. Detecting non-causal artifacts in multivariate linear regression models. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*. 2250–2258.
- D. Janzing, J. Peters, J. M. Mooij, and B. Schölkopf. 2009. Identifying confounders using additive noise models. In *Proceedings of the 25th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*. 249–257.
- D. Janzing, P. Hoyer, and B. Schölkopf. 2010. Telling cause from effect based on high-dimensional observations. In J. Fürnkranz and T. Joachims (Eds.), In *Proceedings of the 27th International Conference on Machine Learning*. 479–486.
- D. Janzing, J. M. Mooij, K. Zhang, J. Lemeire, J. Zscheischler, P. Daniušis, B. Steudel, and B. Schölkopf. 2012. Information-geometric approach to inferring causal directions. *Artif. Intell.* 182–183, 1–31. DOI: <https://doi.org/10.1016/j.artint.2012.01.002>.
- D. Janzing, R. Chaves, and B. Schölkopf. 2016. Algorithmic independence of initial condition and dynamical law in thermodynamics and causal inference. *New J. Phys.* 18, 093052, 1–13. DOI: <https://doi.org/10.1088/1367-2630/18/9/093052>.
- M. Khajehnejad, B. Tabibian, B. Schölkopf, A. Singla, and M. Gomez-Rodriguez. 2019. Optimal decision making under strategic behavior. arXiv:1905.09239.
- N. Kilbertus, M. Rojas Carulla, G. Parascandolo, M. Hardt, D. Janzing, and B. Schölkopf. 2017. Avoiding discrimination through causal reasoning. In *Advances in Neural Information Processing Systems 30*. 656–666.
- N. Kilbertus, G. Parascandolo, and B. Schölkopf. 2018. Generalization in anti-causal learning. arXiv:1812.00524.
- D. P. Kingma and M. Welling. 2013. Auto-encoding variational Bayes. arXiv:1312.6114.
- F. Klein. 1872. *Vergleichende Betrachtungen über neuere geometrische Forschungen*. Verlag von Andreas Deichert, Erlangen.
- S. Kpotufe, E. Sgouritsa, D. Janzing, and B. Schölkopf. 2014. Consistency of causal inference under the additive noise model. In *Proceedings of the 31st International Conference on Machine Learning*. 478–486.
- M. J. Kusner, J. Loftus, C. Russell, and R. Silva. 2017. Counterfactual fairness. In *Advances in Neural Information Processing Systems 30*. Curran Associates, Inc., 4066–4076.
- S. Lange, T. Gabel, and M. Riedmiller. 2012. Batch reinforcement learning. In M. Wiering and M. van Otterlo (Eds.), *Reinforcement Learning: State-of-the-Art*. Springer, Berlin, 45–73.
- S. L. Lauritzen. 1996. *Graphical Models*. Oxford University Press, New York.
- Y. LeCun, Y. Bengio, and G. Hinton. 2015. Deep learning. *Nature* 521, 7553, 436–444. DOI: <https://doi.org/10.1038/nature14539>.
- Y. Li, M. Gong, X. Tian, T. Liu, and D. Tao. 2018a. Domain generalization via conditional invariant representation. arXiv:1807.08479.

- Y. Li, X. Tian, M. Gong, Y. Liu, T. Liu, K. Zhang, and D. Tao. 2018b. Deep domain generalization via conditional invariant adversarial networks. In *The European Conference on Computer Vision (ECCV)*.
- Z. C. Lipton, Y.-X. Wang, and A. Smola. 2018. Detecting and correcting for label shift with black box predictors. arXiv:1802.03916.
- F. Locatello, S. Bauer, M. Lucic, G. Rätsch, S. Gelly, B. Schölkopf, and O. Bachem. 2018a. Challenging common assumptions in the unsupervised learning of disentangled representations. In *Proceedings of the 36th International Conference on Machine Learning*.
- F. Locatello, D. Vincent, I. Tolstikhin, G. Rätsch, S. Gelly, and B. Schölkopf. 2018b. Competitive training of mixtures of independent deep generative models. arXiv:1804.11130.
- D. Lopez-Paz, K. Muandet, B. Schölkopf, and I. Tolstikhin. 2015. Towards a learning theory of cause–effect inference. In *Proceedings of the 32nd International Conference on Machine Learning*. 1452–1461.
- D. Lopez-Paz, R. Nishihara, S. Chintala, B. Schölkopf, and L. Bottou. 2017. Discovering causal signals in images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 58–66.
- K. Lorenz. 1973. *Die Rückseite des Spiegels*. R. Piper & Co. Verlag, Munich.
- C. Lu, B. Schölkopf, and J. M. Hernández-Lobato. 2018. Deconfounding reinforcement learning in observational settings. arXiv:1812.10576.
- S. MacLane. 1971. *Categories for the Working Mathematician*. Vol. 5. Graduate Texts in Mathematics. Springer-Verlag, New York.
- S. Magliacane, T. van Ommen, T. Claassen, S. Bongers, P. Versteeg, and J. M. Mooij. 2018. Domain adaptation by using causal inference to predict invariant conditional distributions. In *Proceedings of the NeurIPS*. arXiv:1707.06422.
- E. Medina. 2011. *Cybernetic Revolutionaries: Technology and Politics in Allende's Chile*. The MIT Press, Cambridge, MA.
- V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis. 2015. Human-level control through deep reinforcement learning. *Nature* 518, 7540, 529–533. DOI: <https://doi.org/10.1038/nature14236>.
- B. T. Montet, T. D. Morton, D. Foreman-Mackey, J. A. Johnson, D. W. Hogg, B. P. Bowler, D. W. Latham, A. Bieryla, and A. W. Mann. 2015. Stellar and planetary properties of K2 Campaign 1 candidates and validation of 17 planets, including a planet receiving Earth-like insolation. *Astrophys. J.* 809, 1, 25.
- J. M. Mooij, D. Janzing, J. Peters, and B. Schölkopf. 2009. Regression by dependence minimization and its application to causal inference. In *Proceedings of the 26th International Conference on Machine Learning (ICML)*. 745–752. DOI: <https://doi.org/10.1145/1553374.1553470>.

- J. M. Mooij, D. Janzing, T. Heskes, and B. Schölkopf. 2011. On causal discovery with cyclic additive noise models. In *Advances in Neural Information Processing Systems 24 (NIPS)*.
- J. M. Mooij, D. Janzing, and B. Schölkopf. 2013. From ordinary differential equations to structural causal models: The deterministic case. In *Proceedings of the 29th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*. 440–448.
- J. M. Mooij, J. Peters, D. Janzing, J. Zscheischler, and B. Schölkopf. 2016. Distinguishing cause from effect using observational data: Methods and benchmarks. *J. Mach. Learn. Res.* 17, 32, 1–102.
- G. Parascandolo, N. Kilbertus, M. Rojas-Carulla, and B. Schölkopf. 2018. Learning independent causal mechanisms. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*. 4033–4041.
- J. Pearl. 1988. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers Inc., San Francisco, CA.
- J. Pearl. 2009a. *Causality: Models, Reasoning, and Inference*. (2nd. ed.). Cambridge University Press, New York.
- J. Pearl. 2009b. Giving computers free will. *Forbes*.
- J. Pearl and E. Bareinboim. 2015. External validity: From do-calculus to transportability across populations. *Stat. Sci.* 2014, 29, 4, 579–595. arXiv:1503.01603. DOI: <https://doi.org/10.1214/14-STS486>.
- J. Peters, J. M. Mooij, D. Janzing, and B. Schölkopf. 2011. Identifiability of causal graphs using functional models. In *Proceedings of the 27th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*. 589–598.
- J. Peters, J. M. Mooij, D. Janzing, and B. Schölkopf. 2014. Causal discovery with continuous additive noise models. *J. Mach. Learn. Res.* 15, 2009–2053.
- J. Peters, P. Bühlmann, and N. Meinshausen. 2016. Causal inference using invariant prediction: Identification and confidence intervals. *J. R. Stat. Soc. Series B Stat. Methodol.* 78, 5, 947–1012. DOI: <https://doi.org/10.1111/rssb.12167>.
- J. Peters, D. Janzing, and B. Schölkopf. 2017. *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press, Cambridge, MA.
- N. Pfister, S. Bauer, and J. Peters. 2018a. Identifying causal structure in large-scale kinetic systems. arXiv:1810.11776.
- N. Pfister, P. Bühlmann, B. Schölkopf, and J. Peters. 2018b. Kernel-based tests for joint independence. *J. R. Stat. Soc. Series B Stat. Methodol.* 80, 1, 5–31. DOI: <https://doi.org/10.1111/rssb.12235>.
- S. Rabanser, S. Günnemann, and Z. C. Lipton. 2018. Failing loudly: An empirical study of methods for detecting dataset shift. arXiv:1810.11953.
- H. Reichenbach. 1956. *The Direction of Time*. University of California Press, Berkeley, CA. DOI: <https://doi.org/10.2307/2216858>.
- M. Rojas-Carulla, B. Schölkopf, R. Turner, and J. Peters. 2018. Invariant models for causal transfer learning. *J. Mach. Learn. Res.* 19, 36, 1–34. DOI: <https://dl.acm.org/doi/10.5555/3291125.3291161>.

- P. K. Rubenstein, S. Weichwald, S. Bongers, J. M. Mooij, D. Janzing, M. Grosse-Wentrup, and B. Schölkopf. 2017. Causal consistency of structural equation models. In *Proceedings of the 33rd Conference on Uncertainty in Artificial Intelligence*.
- P. K. Rubenstein, S. Bongers, B. Schölkopf, and J. M. Mooij. 2018. From deterministic ODEs to dynamic structural causal models. In *Proceedings of the 34th Conference on Uncertainty in Artificial Intelligence (UAI)*.
- B. Schölkopf. 2015. Artificial intelligence: Learning to see and act. *Nature* 518, 7540, 486–487. DOI: <https://doi.org/10.1038/518486a>.
- B. Schölkopf. 2017. Causal learning. In *Invited Talk, 34th International Conference on Machine Learning (ICML)*. <https://vimeo.com/238274659>.
- B. Schölkopf and A. J. Smola. 2002. *Learning with Kernels*. MIT Press, Cambridge, MA.
- B. Schölkopf, D. Janzing, J. Peters, and K. Zhang. 2011. Robust learning via cause–effect models. <https://arxiv.org/abs/1112.2738>.
- B. Schölkopf, D. Janzing, J. Peters, E. Sgouritsa, K. Zhang, and J. Mooij. 2012. On causal and anticausal learning. In J. Langford and J. Pineau (Eds.), *Proceedings of the 29th International Conference on Machine Learning (ICML)*. Omnipress, New York, 1255–1262. <http://icml.cc/2012/papers/625.pdf>.
- B. Schölkopf, D. Hogg, D. Wang, D. Foreman-Mackey, D. Janzing, C.-J. Simon-Gabriel, and J. Peters. 2016a. Modeling confounding by half-sibling regression. *Proc. Natl. Acad. Sci. U. S. A.* 113, 27, 7391–7398. DOI: <https://doi.org/10.1073/pnas.1511656113>.
- B. Schölkopf, D. Janzing, and D. Lopez-Paz. 2016b. Causal and statistical learning. *Oberwolfach Rep.* 13, 3, 1896–1899. DOI: <https://doi.org/10.4171/OWR/2016/33>.
- L. Schott, J. Rauber, M. Bethge, and W. Brendel. 2019. Towards the first adversarially robust neural network model on MNIST. In *International Conference on Learning Representations*. <https://openreview.net/forum?id=S1EHOsC9tX>.
- R. D. Shah and J. Peters. 2018. The hardness of conditional independence testing and the generalised covariance measure. *Ann. Statist.* 48, 3, 1514–1538. arXiv:1804.07203. DOI: <https://doi.org/10.1214/19-AOS1857>.
- N. Shajarisales, D. Janzing, B. Schölkopf, and M. Besserve. 2015. Telling cause from effect in deterministic linear dynamical systems. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*. 285–294.
- C. E. Shannon. 1959. Coding theorems for a discrete source with a fidelity criterion. In *IRE International Convention Records*. Vol. 7, 142–163.
- S. Shimizu, P. O. Hoyer, A. Hyvärinen, and A. J. Kerminen. 2006. A linear non-Gaussian acyclic model for causal discovery. *J. Mach. Learn. Res.* 7, 2003–2030.
- V. Smil. 2017. *Energy and Civilization: A History*. MIT Press, Cambridge, MA. DOI: <https://doi.org/10.7551/mitpress/10752.001.0001>.
- P. Spirtes, C. Glymour, and R. Scheines. 2000. *Causation, Prediction, and Search* (2nd. ed.). MIT Press, Cambridge, MA. DOI: <https://doi.org/10.1002/sim.1415>.
- W. Spohn. 1978. *Grundlagen der Entscheidungstheorie*. Scriptor-Verlag.
- I. Steinwart and A. Christmann. 2008. *Support Vector Machines*. Springer, New York.

- B. Steudel, D. Janzing, and B. Schölkopf. 2010. Causal Markov condition for submodular information measures. In *Proceedings of the 23rd Annual Conference on Learning Theory (COLT)*. 464–476.
- A. Subbaswamy, P. Schulam, and S. Saria, 2018. Preventing failures due to dataset shift: Learning predictive models that transport. arXiv:1812.04597.
- X. Sun, D. Janzing, and B. Schölkopf. 2006. Causal inference by choosing graphs with most plausible Markov kernels. In *Proceedings of the 9th International Symposium on Artificial Intelligence and Mathematics*.
- R. Suter, Đ. Miladinović, B. Schölkopf, and S. Bauer. 2018. Robustly disentangled causal mechanisms: Validating deep representations for interventional robustness. arXiv:1811.00007. Proceedings ICML.
- C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. 2013. Intriguing properties of neural networks. arXiv:1312.6199.
- A. Tsiaras, I. Waldmann, G. Tinetti, J. Tennyson, and S. N. Yurchenko. 2019. Water vapour in the atmosphere of the habitable-zone eight-earth-mass planet K2-18b. *Nat. Astron.* 3, 1–6. DOI: <https://doi.org/10.1038/s41550-019-0878-9>.
- V. Vapnik. 1995. *The Nature of Statistical Learning Theory*. Springer, New York. DOI: <https://doi.org/10.1007/978-1-4757-2440-0>.
- V. N. Vapnik. 1998. *Statistical Learning Theory*. Wiley, New York.
- J. von Kügelgen, A. Mey, M. Loog, and B. Schölkopf. 2019. Semi-supervised learning, causality and the conditional cluster assumption. <https://arxiv.org/abs/1905.12081>.
- H. Wang, Z. He, Z. C. Lipton, and E. P. Xing. 2019. Learning robust representations by projecting superficial statistics out. arXiv:1903.06256.
- S. Weichwald, B. Schölkopf, T. Ball, and M. Grosse-Wentrup. 2014. Causal and anti-causal learning in pattern recognition for neuroimaging. In *4th International Workshop on Pattern Recognition in Neuroimaging (PRNI)*. IEEE. DOI: <https://doi.org/10.1109/PRNI.2014.6858551>.
- W. K. Wootters and W. H. Zurek. 1982. A single quantum cannot be cloned. *Nature* 299, 5886, 802–803. DOI: <https://doi.org/10.1038/299802a0>.
- J. Zhang and E. Bareinboim. 2018. Fairness in decision-making—The causal explanation formula. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*. New Orleans, LA, 2037–2045.
- J. Zhang and E. Bareinboim. 2019. Near-optimal reinforcement learning in dynamic treatment regimes. In *Advances in Neural Information Processing Systems 33*.
- K. Zhang and A. Hyvärinen. 2009. On the identifiability of the post-nonlinear causal model. In *Proceedings of the 25th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*. 647–655.
- K. Zhang, J. Peters, D. Janzing, and B. Schölkopf. 2011. Kernel-based conditional independence test and application in causal discovery. In *Proceedings of the 27th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*. 804–813.

- K. Zhang, B. Schölkopf, K. Muandet, and Z. Wang. 2013. Domain adaptation under target and conditional shift. In *Proceedings of the 30th International Conference on Machine Learning (ICML)*. 819–827.
- K. Zhang, M. Gong, and B. Schölkopf. 2015. Multi-source domain adaptation: A causal view. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence*. 3150–3157.
- K. Zhang, B. Huang, J. Zhang, C. Glymour, and B. Schölkopf. 2017. Causal discovery from nonstationary/heterogeneous data: Skeleton estimation and orientation determination. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI 2017)*. 1347–1353. DOI: <https://doi.org/10.24963/ijcai.2017/187>.

Why Did They Do That?

Ross Shachter (Stanford University),
David Heckerman (Amazon, Seattle)

Abstract

Judea Pearl argues that people as well as machines with artificial intelligence must have the ability to apply causal reasoning to make decisions and to explain or justify those decisions. We wholeheartedly agree with Judea on this point, but show that the ability to identify available alternatives and the ability to express preferences are also necessary for making and explaining decisions. We briefly review the basic principles of decision theory, showing how these three abilities come together in decision-making. We illustrate these principles with examples including Judea's incisive depiction of the story of Adam and Eve.

40.1 Introduction

At the start of *The Book of Why*, Judea Pearl places us in the Garden of Eden [Pearl and Mackenzie 2018]. When God asked Adam, “Have you been eating of the tree I forbade you to eat?” Adam replied, “It was the woman you put with me; she gave me the fruit and I ate it.” Eve added, “The serpent tempted me, and I ate” [Jones 1971]. Judea notes that instead of responding “Yes” or “No” to God’s questions, they gave excuses, explanations of *why* they ate the fruit, attempting to shift the blame from themselves to others. They knew that they had disobeyed God and they hoped they could avoid the consequences that God declared, “of the tree of the knowledge of good and evil you are not to eat, for on the day you eat of it you shall most surely die.”

Judea argues that asking “Why did they do that?” is a natural question and argues that causal reasoning, the second level of his ladder of causation, is a necessary tool to answer this question. We wholeheartedly agree with these points [Heckerman and Shachter 1995]. In particular, in order to know why Adam and Eve ate the fruit, we need to know how they thought about the causal consequences of

their possible actions, and the probabilities of those consequences are obtained through causal reasoning using methods such as those pioneered by Judea. Judea also argues that machinery for causal reasoning is a necessary component of true artificial intelligence. We agree with this point as well. That said, we use this article to argue that there are two additional machineries needed to answer the question of “Why did I do that?” and further argue that these machineries are equally important components of a true artificial intelligence. The first additional component is the ability to identify available alternatives—that is, what an entity can do in a given situation. The second component is the ability of the entity to express preferences over the possible outcomes. Together, these three components allow us to choose the best available alternative.

These three components for answering the question “Why did I do that?” are precisely the key aspects of decision theory [Ramsey 1926, von Neumann and Morgenstern 1947, Blackwell and Girshick 1954, Savage 1954, Raiffa 1968, Howard 1970]. It is from this perspective that we explore the story of the Garden of Eden to illustrate the approach. (We do not intend to convey any theological insight.) Before we return to their decision and a description of how they could use decision theory to give a more complete explanation of why they ate the fruit, let’s consider a simple decision problem and how it is represented by decision theory.

40.2 Some Examples

Consider whether to pay for parking if we expect to be staying for only a short time to run an errand. This situation is represented by the decision tree shown in Figure 40.1.

We can choose to pay for the parking or we can choose to not pay for it. These are our alternatives. If our violation is detected, then we will receive a ticket and have to pay a fine, but if it goes undetected, then we will have parked for free. For many people, parking for free is the best outcome and paying a fine is the worst outcome,

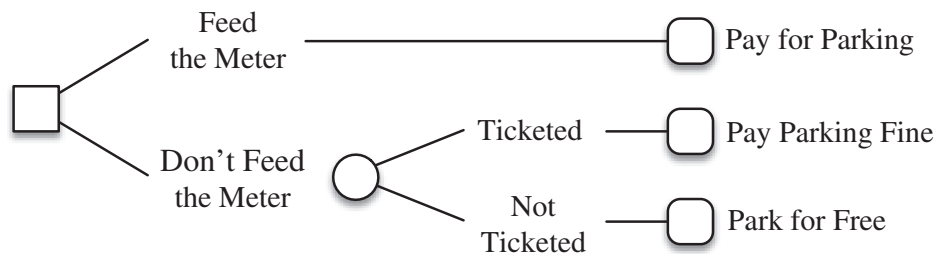


Figure 40.1 We can choose to feed the meter and pay for parking, or choose not to feed it and then either pay a fine or park for free, depending on whether our car is ticketed.

so there is some tradeoff involving the cost of parking, the cost of the fine, and the probability that our violation will be detected. This probability is determined through causal reasoning.

While we don't want to get much into the details of decision theory, we note that, once the ingredients of alternatives, probabilities over possible outcomes (from causal reasoning), and preferences are specified, the theory offers a prescription to act. In this simple example, we should pay for parking if

$$P\{detection\}(\text{Cost of Fine}) > (\text{Cost of Parking}).$$

Another important feature of decision theory is that different decision makers can have different alternatives, probabilities, and preferences. For example, a decision maker may know there is free parking nearby and add that as an alternative, may be uncertain about the cost of the fine if they fear their car might be towed, or prefer not to park without paying. These differences among decision makers can lead them to make different choices, even if they agree on the causal reasoning.

The determination of probabilities through causal reasoning, the identification of alternatives, and the assessment of preferences are all important ingredients to making a decision, and answering the question “Why did I do that?” In a recent *CACM* article, Judea talks about how causal reasoning can help someone who asks “I am about to quit my job, but should I?” [Pearl 2019]. We agree that causal reasoning is an important component to answering their question. But it is not enough. Are they considering another job or going back to school? Is the issue salary, work situation, family, health, or commute? Causal reasoning lets us consider the effects of each alternative on these different concerns. However, to answer their question, we also need to know their preferences and the alternatives available to them.

40.3 Back to the Garden of Eden

Now let us return to Adam and Eve's decision, viewing it from their perspective *before* they ate the fruit. (In this story, we are in a very unusual situation where eating the fruit actually changes who they are, potentially changing their ability to identify alternatives and to reason causally, and potentially changing their preferences. So, we must be careful to look at their decision before they ate the fruit.) Their situation is represented by the decision tree shown in Figure 40.2. Their alternatives are to eat the fruit or to not eat it. Thinking about the consequences about their possible actions—that is, thinking causally—they reason that, if they don't eat the fruit, everything will remain as it has been. If they eat the fruit and God finds out, God has said that they shall most surely die, presumably by God's hand.

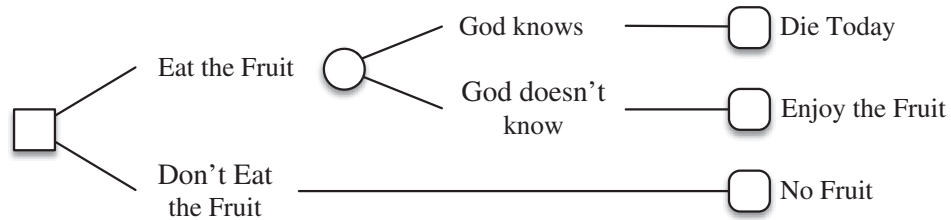


Figure 40.2 Adam and Eve can choose to eat the fruit and then either die or live having enjoyed the fruit, depending on whether God finds out, or they can choose not to eat the fruit and never get to enjoy it.

In contrast, the serpent has said “No! You will not die! God knows in fact that on the day you eat it your eyes will be opened and you will be like gods, knowing good and evil.” So they reason that there are two possibilities: either God will find out and they will die, or God won’t find out and they will get to enjoy the fruit and any knowledge that comes with it. “The woman saw that the tree was good to eat and pleasing to the eye, and that it was desirable for the knowledge that it could give.” Therefore, enjoying the fruit is the best outcome and dying is the worst outcome, and they need to think about how likely it is that God will find out and how desirable the fruit must be to take that risk.

We know that they ate the fruit. Perhaps they had considered other alternatives—for example, pleading with God if God were to find out. Or perhaps they considered other outcomes—for example, that they might not die if God finds out. However, the Bible is silent on these other possibilities, so we leave the representation of their decision—“Why did they do that?”—as shown in the decision tree in Figure 40.2.

40.4 Decision Theory and Decision Analysis

There are other interesting aspects surrounding the story of Adam and Eve, but before we return to these, let us consider decision theory in more detail. Decision theory is a *normative* theory for decision-making. That is, its principles follow from a small set of axioms that a decision maker should follow. For example, one of the axioms is that a decision maker’s preferences over the possible outcomes are totally ordered. If not, for example if A is preferred to B, B is preferred to C, and C is preferred to A, then a third party could extract money from the decision maker by getting him to pay for a preferred outcome over another, until he is back with the outcome he started with, but with less money. Decision theory is not a descriptive theory of decision-making—that is, it does not accurately describe *how* people actually make decisions.

Prospect theory is such a theory, attempting to explain and characterize the processes people use in decision-making, incorporating their biases and heuristics [Kahneman and Tversky 2006]. Prospect theory would be an appropriate theory for artificial intelligence if our goal were merely to simulate human decision makers. However, it seems more compelling to consider the normative decision theory developed to explain and characterize the processes humans might want to use for important decisions.

Decision analysis, developed by Ron Howard and Howard Raiffa, is a discipline centered on the application of decision theory to real decisions in practice. One of the fundamental distinctions in decision analysis is between the quality of a decision made and the quality of the resulting outcome. Although it is commonplace in our society to judge a decision by the outcome, they are quite different, as good decisions can have bad outcomes and bad decisions can have good outcomes. We use decision analysis to help us make a decision *before* we act, improving the quality of our decision. Using decision analysis only to explain our decision *later* leaves most of its benefits on the table.

Unfortunately, it is when a decision maker experiences a bad outcome that they are most likely to ask themselves (again) “Why did I do that?” That leads to the question “What would have happened had I made another choice instead?” This is often how we learn to make better decisions, and it is the type of counterfactual reasoning on the third level of Judea’s ladder of causation.

Yet another important notion from decision analysis is that, in any given situation, a pure $\text{Do}(X = x)$ alternative might not be available, and the decision maker must choose from the limited set of realistic alternatives. For example, consider the decision made by a patient and their physician about which treatment they should receive for a serious disease. Ideally, they would like the alternative $\text{Do}(\text{disease} = \text{false})$ with no side effects, but this is rarely available as an alternative. When we consider the treatment options actually available, we need to integrate those side effects into our causal reasoning.

An extreme example of side effects can be found in the classic horror story of “The Monkey’s Paw” [Jacobs 1902]. In that story, each owner of a monkey’s paw was granted three wishes. Although their wishes were fulfilled, it was not done so as the wisher had intended, such as when his wish for money was fulfilled by the death of his son leading to an insurance payout.

Indeed, identifying realistic alternatives can be difficult. Nonetheless, this identification is a necessary component for answering the question “Why did I do that?” and a necessary component for true artificial intelligence. In humans, the process of identifying alternatives seems to arise naturally from the experience of free will. Interestingly, recent experiments in neurobiology [Soon et al. 2008]

and physics [Proietti et al. 2019] cast doubts on whether humans actually have free will or merely perceive that we do. Nonetheless, many researchers, including Judea, believe that at least the perception of free will offers advantages to artificial intelligence.

We agree and offer the following observations. With regard to taking action, a very different alternative to decision theory is the use of situation–action rules. Such rules are evolutionarily built in to almost all forms of life, even reflexive actions in humans. A key advantage of situation–action rules is computational simplicity, and hence ease of implementation within a biological or human-made system. In contrast, a key advantage of decision theory, where many complex alternatives and their consequences are imagined (afforded by the perception of free will) and then selected, is improved quality of action, leading to increased chances of survival. Furthermore, decision-theoretic thinking has an important property that makes implementation feasible.

That property is modularity or decomposability of its parts. In particular, as illustrated in our examples, decision-theoretic thinking flows naturally in the sequence: (1) identify available alternatives, (2) consider the causal effects of those alternatives on the possible outcomes, and (3) consider preferences over those possible outcomes. The causal effects flowing from each alternative and the resulting preferences on those effects can be considered separately for each alternative. Finally, causal reasoning itself is modular in that cause–effect relationships can be pieced together to form other relationships. As a simple example, if we know X causes Y, and Y causes Z, it is usually safe to conclude X causes Z.

In summary, decision-theoretic thinking offers strong survival advantages, and its modularity permits feasible implementation. Furthermore, the perception of free will permits such thinking. A key unanswered question is whether the perception of free will is necessary for decision-theoretic thinking.

40.5 Back Again in the Garden of Eden

Now consider when God asks Adam and Eve “Have you been eating of the tree I forbade you to eat?” Adam blames Eve, who looks back at her decision-making and uses causal reasoning to recognize that the serpent’s words led her to both lower her probability that God would find out and increase her preference for the fruit, together causing her to choose to eat the fruit. Eve summarizes this to God, saying “The serpent tempted me, and I ate.” Of course, Eve leaves out the important details of the temptation, which she must have surmised—with her new knowledge of good and evil—that God would not appreciate. Interestingly, the serpent’s statement turned out to be true, if misleading: Adam and Eve did not die that day, as God had threatened, and they indeed gained the knowledge of good and evil.

40.6 Conclusion: God's Decision

We conclude by thinking about the decision that God made after confronting Adam and Eve from the perspective of decision theory. When God created them, we learn “Now both of them were naked, the man and his wife, but they felt no shame in front of each other.” However, when God found them in the Garden of Eden after they ate the fruit, Adam says “I was afraid because I was naked, so I hid.” God’s causal model kicks in: “Who told you that you were naked? ... Have you been eating of the tree I forbade you to eat?”

It might not have taken any inference, as some believe that God is all knowing and can predict the future as well. Hence they believe that God knew all along that Adam and Eve would eat the fruit. However, others believe that when God created Adam and Eve, he endowed them with curiosity and free will. God could only control their behaviors imperfectly and indirectly through punishment and reward. And to do so perfectly, God would have needed to understand their alternatives, their causal reasoning, and their preferences!

Once God confirmed with Adam and Eve that they ate the fruit, there was a decision to be made about their punishment. If they died that day, then God’s promise would have been fulfilled. If nothing happened, then God had fears: “See, the man has become like one of us, with his knowledge of good and evil. He must not be allowed to stretch his hand out next and pick from the tree of life also, and eat some and live forever.” It would seem that God had decided that they should die.

However, God introduced a new alternative of banishing them to a life of suffering as mortal humans, as shown in Figure 40.3. But why? God makes it clear that it was neither forgiveness nor mercy, for he cursed them and all of their descendants with lives of pain and suffering. Did God merely threaten death to dissuade them from eating the fruit knowing that otherwise they could not resist the temptation? If only we knew God’s available alternatives, causal reasoning, and preferences well enough to understand “Why did God do that?”

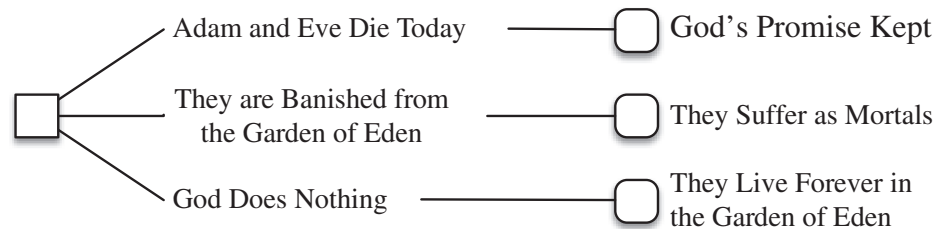


Figure 40.3 Knowing they have eaten the fruit, God has three choices: either Adam and Eve die as God promised, they are banished from the Garden of Eden to a life of suffering as mortals, or they will live forever in the Garden of Eden.

References

- D. Blackwell and M. A. Girshick. 1954. *Theory of Games and Statistical Decisions*. John Wiley & Sons, Inc., New York. DOI: <https://doi.org/10.1002/nav.3800010313>.
- D. Heckerman and R. Shachter. 1995. Decision-theoretic foundations for causal reasoning. *J. Artificial Intel. Res.* 3, 405–430. DOI: <https://doi.org/10.1613/jair.202>.
- R. A. Howard. 1970. Decision analysis: Perspectives on inference, decision, and experimentation. *Proc. IEEE* 58, 5, 632–643. DOI: <https://doi.org/10.1109/PROC.1970.7719>.
- W. W. Jacobs. 1902. “The Monkey’s Paw.” *Harper’s Monthly*, 105, 634–639.
- A. Jones. 1971. *The Jerusalem Bible*. Doubleday, Garden City, NY.
- D. Kahneman and A. Tversky. 2006. Prospect theory: An analysis of decision under risk. *Econometrica* 74, 2, 263–292. DOI: <https://doi.org/10.2307/1914185>.
- J. Pearl. 2019. The seven tools of causal inference, with reflections on machine learning. *Commun. ACM* 62, 3, 54–60. DOI: <https://doi.org/10.1145/3241036>.
- J. Pearl and D. Mackenzie. 2018. *The Book of Why*. Basic Books, New York.
- M. Proietti, A. Pickston, F. Graffitti, P. Barrow, D. Kundys, C. Branciard, M. Ringbauer, and A. Fedrizzi. 2019. Experimental rejection of observer-independence in the quantum world. <http://arxiv.org/abs/1902.05080>. DOI: <https://doi.org/10.1126/sciadv.aaw9832>.
- H. Raiffa. 1968. *Decision Analysis: Introductory Lectures on Choices under Uncertainty*. Addison-Wesley. DOI: <https://doi.org/10.2307/2987280>.
- F. P. Ramsey. 1926. Truth and probability. *Philosophical Papers*, 52–94.
- L. J. Savage. 1954. *The Foundations of Statistics*. John Wiley & Sons, Inc., New York. DOI: <https://doi.org/10.1002/nav.3800010316>.
- C. S. Soon, M. Brass, H. J. Heinze, and J. D. Haynes. 2008. Unconscious determinants of free decisions in the human brain. *Nat. Neurosci.* 11, 5, 543–545. DOI: <https://doi.org/10.1038/nn.2112>.
- J. von Neumann and O. Morgenstern. 1947. *Theory of Games and Economic Behavior* (2nd ed). Princeton University Press, Princeton, NJ.

Multivariate Counterfactual Systems and Causal Graphical Models

Ilya Shpitser (Johns Hopkins University),
Thomas S. Richardson (University of Washington),
James M. Robins (Harvard T. H. Chan School of Public Health)

Among Judea Pearl's many contributions to causality and statistics, the graphical d-separation criterion, and the do-calculus stand out. In this chapter we show that d-separation provides direct insight into an earlier causal model originally described in terms of potential outcomes and event trees. In turn, the resulting synthesis leads to a simplification of the do-calculus that clarifies and separates the underlying concepts, and a simple counterfactual formulation of a complete identification algorithm in causal models with hidden variables.

41.1 Introduction

For the last three decades, Judea Pearl has been a leading advocate for the adoption of causal models throughout the sciences. Pearl [1995] introduced causal models based on non-parametric structural equation models (NPSEMs).¹ NPSEMs encode direct causal relations between variables. More precisely, each variable V is modeled as a function of its direct causes and an error term ϵ_V ; this is the “structural equation” for V ; see Table 41.2. These causal relationships can be represented

1. See also Pearl [2009, p. 69]. More recently Pearl has used the term (Structural) Causal Model (SCM) to refer to NPSEMs; see Pearl [2009, p. 203, Definition 7.1.1]. However, (S)CM is sometimes also used to denote NPSEMs in which, in addition, the error terms are assumed to be independent either explicitly (see Pearl [2009, p. 44, Definition 2.2.2] and Forré and Mooij [2019]) or implicitly (see Lee et al. [2020]). For this reason, we prefer to use Pearl's earlier terminology.

naturally by the directed arrows on a directed acyclic graph (DAG) in which there is an edge $X \rightarrow V$ if X is present in the structural equation for V . The resulting graph is often called a causal DAG or diagram. However, further probabilistic assumptions are required to link the NPSEM to the distribution of the data.

Pearl has often considered a submodel of an NPSEM, hereafter referred to as the NPSEM-IE, which assumes the Independence of Error terms. NPSEM-IEs typically include both observed and hidden variables.² Thus, although these models assume that errors are independent, they still allow a modeler to postulate non-causal dependence between observed variables X and Y by including a hidden variable $X \leftarrow H \rightarrow Y$ (instead of allowing errors ε_X and ε_Y to be dependent).

Under the NPSEM-IE the distribution over the factual (i.e., hidden plus observed) variables factorizes according to the causal DAG. This allows one to reason about conditional independence in the distribution for the factual variables via d-separation relations on the causal graph. Based on this insight, Pearl developed an influential reasoning system called the *do*-calculus that allows complex derivations to be made linking causal and observed quantities by appealing to d-separation in graphs derived from the causal DAG.

Causal graphs plus d-separation turn a difficult mathematical problem into a simple one of graph topology. The use of causal DAGs, as championed by Pearl, has revolutionized causal reasoning in many fields, including fields such as epidemiology and sociology, precisely because causal reasoning based on DAGs and d-separation is so “user-friendly.” That is, individuals lacking the necessary mathematical background to understand probabilistic inference based solely on an NPSEM-IE have been given a tool with which they can solve subtle problems in causal inference. In fact, even the mathematically sophisticated find causal reasoning with graphs to be much easier than algebraically manipulating the underlying structural equations. As Pearl emphasizes, this is largely because causal DAGs faithfully represent the way humans, including scientists and mathematicians, encode causal relations.

The use of DAGs to encode causal relationships dates back to the work of the geneticist Sewall Wright [Wright 1921] in the 1920s, who used a special case of the NPSEM associated with linear structural equations, and Gaussian errors for pedigree analysis among other applications in biology. These ideas were further developed and applied by Wright, Haavelmo, the Cowles Commission, Strotz & Wold, and Fisher [Wright 1921, Haavelmo 1943, Simon 1953, Strotz and Wold 1960, Fisher 1969, 1970].

2. Causal DAG models with unobserved variables are also referred to as “semi-Markovian” by Pearl [2009, pp. 69 and 76].

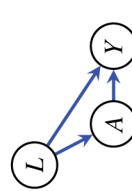
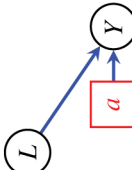
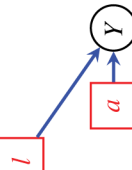
Table 41.1.1 Graphical causal models based on the SWIG/FFRCISTG counterfactual framework
 FFRCISTG/SWIG Potential Outcome Model [Richardson and Robins 2013, Robins 1986]

	One-Step Ahead Counterfactuals	Passive Observation	Experimental Intervention on A
Graph:			
Variables:	L $A(l)$ $Y(a, l)$	L $A \equiv A(L)$ $Y \equiv Y(A, L)$	L $A \equiv A(L)$ $Y(a) \equiv Y(a, L)$
Interpretation:	Counterfactuals when A and L are intervened on;	Observed System;	Prior to intervention: L and A; after intervention: Y(a).
Meaning of A:	(A does not appear)	Natural value of A;	Natural value of A (observed prior to intervention on A).

Single-World No Confounding Assumption: for each pair a, l : $L \perp\!\!\!\perp A(l) \perp\!\!\!\perp Y(a, l)$

Note that the meaning of a variable such as A or Y does not depend on the graph in which it appears; compare to Table 41.2. The SWIG no confounding assumption is less restrictive than the assumption of independent errors.

Table 41.2 Structural equation models and their relationship to counterfactuals
Non-Parametric Structural Equation Model with Independent Errors (NPSEM-IE) [Pearl 2009, Strotz and Wold 1960]

	Passive Observation	Experimental Intervention on A	Experimental Intervention on A and L
Graph:			
Variables:	$L = f_L(\epsilon_L)$ $A = f_A(L, \epsilon_A)$ $Y = f_Y(A, L, \epsilon_Y)$	$L = f_L(\epsilon_L)$ $A = a$ $Y = f_Y(A, L, \epsilon_Y)$	$L = l$ $A = a$ $Y = f_Y(A, L, \epsilon_Y)$
Interpretation:	Observed system;	Variables in system in which A is set to a;	L and A after each is intervened on; Y after both interventions.
Meaning of A:	Natural value of A;	Value of A after intervention on A;	Value of A after intervention on A (and L).

Independent Errors No Confounding Assumption: $\epsilon_L \perp \epsilon_A \perp \epsilon_Y$

Relationships to Counterfactuals:

Error terms: $\epsilon_L = L; \quad \epsilon_A = \{A(l) \text{ for all } l\}; \quad \epsilon_Y = \{Y(a, l) \text{ for all } l, a\}$

Structural equations: $L = f_L(\epsilon_L) \quad A(l) = f_A(l, \epsilon_A) \quad Y(a, l) = f_Y(a, l, \epsilon_Y)$

Note that the meaning of a variable such as A or Y is context specific, it depends on the graph in which it appears. This NPSEM is a strict submodel of the SWIG given in Table 41.1. This is (solely) because the SWIG no confounding assumption is less restrictive than the assumption of independent errors.

In statistics, (non-graphical) causal inference models have a long history also dating back to the 1920s [Neyman 1923, Rubin 1974, Robins 1986]. These models are based on counterfactual variables (potential outcomes) that encode the value the variable would have if, possibly contrary to fact, a particular treatment had been given. Causal contrasts in these models compare the distributions of potential outcomes under two or more treatments.

In general, these counterfactual models considered treatments or exposures at a single point in time. Extending the framework introduced by Neyman to allow for treatment at multiple time-points, Robins introduced *causally interpretable structured tree graph* (CISTG) models. These counterfactuals models, which were represented using event tree graphs, extended the point treatment model of Neyman [1923] to longitudinal studies with time-varying treatments, direct and indirect effects, and feedback of one cause on another.

Pearl has noted that an NPSEM (even without assumptions on the distribution of the errors) implies the existence of potential outcomes and thus an NPSEM model also allows reasoning about counterfactuals; see Halpern and Pearl [2001, 2005]. Indeed, Robins and Richardson have shown that in fact a particular finest CISTG model (“as detailed as the data”) is mathematically isomorphic to an NPSEM model in the sense that any such CISTG model can be written as an (acyclic) NPSEM model and vice versa. A finest CISTG “as detailed as the data” is a counterfactual causal model in which all the underlying variables can be intervened on—an assumption that Pearl has sometimes also adopted.³ Other versions of CISTG models, unlike the NPSEM, assume that only a subset of the variables can be thought of as treatments with associated counterfactuals; thus, interventions and causal effects are only defined for this subset. Henceforth, unless stated otherwise, the term “CISTG model” will be used to denote a “finest CISTG model as detailed as the data.”

Since counterfactual variables are not directly observed, assumptions are needed to link counterfactuals and their distributions to those of the factual data. A necessary assumption is consistency, which states that for a unit their observed outcome (Y) and their potential outcome ($Y(a)$) had a particular treatment a been

3. See Galles and Pearl [1998], Definitions 2 and 3 and Footnote 2, also Pearl [2009] Definitions 7.1.2 and 7.1.3. However, in more recent work, Pearl [2018, 2019] has made a further distinction between hypothetical interventions and a concept of causation based on variables that “listen to others.” Pearl continues to assume that for every variable there are counterfactuals associated with applying the *do* operator to that variable. However, the model resulting from applying the *do* operator and removing structural equations need no longer correspond to an actual intervention. This leaves open the question as to whether there are predictions made by these removals and, if so, how they can be validated.

assigned, will coincide, if in reality the treatment they received (A) is a . However, since both counterfactuals are not directly observed for any individual—the fundamental problem of causal inference—distributions of causal effects are not identified without additional assumptions, beyond consistency.

These assumptions typically take the form of Markov (conditional independence) assumptions that link the distribution of the factual data to that of the counterfactuals, as further discussed below. The simplest example is a randomized clinical trial that assigns treatment via the flip of a coin, and thus treatment is independent of the potential outcomes, so for all a , $A \perp\!\!\!\perp Y(a)$. Together, consistency and Markov assumptions allow population-level causal contrasts to be identified from observed data. In contrast, individual-level effects are not typically identifiable.

Under the NPSEM-IE model, the additional Markov assumptions follow from the assumption that the errors in the structural equation for each variable (hidden or observed) are independent of the errors in the structural equations for the other variables.

Robins [1986] similarly added independence assumptions to the CISTG model. Robins referred to the version of this model in which all variables can be intervened on as the “finest fully randomized CISTG model as detailed as the data,” which we henceforth refer to as the “FFRCISTG model,” unless stated otherwise. Interestingly, the NPSEM-IE implies many more counterfactual independence assumptions than does the corresponding FFRCISTG model. In fact, if we consider complete graphs on p binary variables, then the difference between the number of assumptions implied by the NPSEM-IE and the FFRCISTG model grows at a doubly exponential rate.⁴

The NPSEM-IE allows the identification of certain causal effects—the pure and total direct and indirect effects and more generally path-specific effects⁵—by making use of additional independence assumptions that cannot be confirmed, even in principle, by any experiment conducted using the variables represented on the graph. In contrast, under the less-restrictive FFRCISTG model all counterfactual independence assumptions are in principle experimentally testable,⁶ and the pure and total direct effects are not identifiable (from the variables on the graph). However, ordinary intervention distributions of the type that arise in Pearl’s *do*-calculus are identifiable under the FFRCISTG model.

4. With three binary variables, the difference in the dimension of the two models is 94, with four it is 32,423 [Richardson and Robins 2013].

5. See Chapter 38 of this volume for more detail on these effects.

6. This assumes that it is possible to observe the natural value of a variable and then intervene on it an instant later; see discussion in Section 41.2.

Many statisticians and econometricians exclusively use counterfactuals (without graphs) when carrying out causal data analyses. Pearl has developed purely graphical criteria to reason about confounding and many other causal questions. Since graphical criteria, such as Pearl’s *do*-calculus, make no reference to counterfactuals, they can appear confusing to those unused to causal graphs. Indeed, only factual variables typically appear on Pearl’s causal diagrams so any connection between Pearl’s graphical criteria and the statistician’s counterfactual criteria appear at first glance to be obscure. This is true even though Pearl and others have shown mathematically that the two approaches to evaluation of confounding are effectively logically equivalent.

In this chapter, we will describe an approach that unifies the graphical and counterfactual approaches to causality, via a graph known as a Single-World Intervention Graph (SWIG).⁷ The SWIG is defined by the counterfactual independencies implied by the FFRCISTG model. The nodes on a SWIG correspond to the counterfactual random variables present in these independencies. Furthermore, Pearl’s d-separation criterion can be applied to the SWIG to read off counterfactual independencies implied by the FFRCISTG model. In fact, we will show that SWIGs lead directly to a simpler reformulation of the *do*-calculus in terms of potential outcomes that allows a considerable simplification of Rule 3. This reformulated calculus, which we term the potential outcome calculus or *po*-calculus, is also strictly stronger than Pearl’s in that it may be used to infer equalities that are not expressible in terms of the *do*(·) operator. We use the *po*-calculus to derive a new simple formulation of an extended version of the ID algorithm for identification of causal queries in the presence of hidden variables. The extended algorithm identifies joint distributions over sets of counterfactual outcomes, where some outcomes are the “natural” values that treatment variables would take were they not intervened on.

7. The approach taken here is inspired by, but distinct from, earlier approaches to combining graphs and counterfactuals such as Pearl’s twin network approach [Balke and Pearl 1994, Shpitser and Pearl 2008, Pearl 2009, Section 7.1.4]. However, the d-separation criterion on twin networks is not complete as there are deterministic relationships that are present—but not represented graphically—among the variables in a twin network. Consequently, it is possible for there to be a d-connecting path and yet the corresponding conditional independence holds for all distributions in the model; see the Appendix for a simple example. In contrast, d-separation is complete for a SWIG. However, it should be noted that twin networks are addressing a harder problem than SWIGs since their goal is to determine all independencies implied by an NPSEM-IE model, including “cross-world” independencies.

Finally, we note that twin-network graphs have not typically used (minimal) labelings, which turn out to be important in some applications; see Section 41.2.3.

41.2 Graphs, Non-parametric Structural Equation Models, and the g -do Operator

Fix a set of indices $V \equiv \{1, \dots, K\}$ under a total ordering \prec , define the sets $\text{pre}_i \equiv \{1, \dots, i - 1\}$. For each index $i \in V$, associate a random variable X_i with state space \mathfrak{X}_i ; the ordering here could be given by temporal ordering but need not be.⁸ Given $A \subseteq V$, we will denote subsets of random variables indexed by A as $X_A \in \mathfrak{X}_A \equiv \times_{i \in A} \mathfrak{X}_i$. For notational conciseness we will sometimes use index sets to denote random variables themselves, using V and A to denote X_V and X_A , respectively, and similarly using lower case a to denote $x_A \in \mathfrak{X}_A$. Similarly, by extension, we will also use V_A to denote X_A and V_i to denote X_i .

We assume the existence of all one-step-ahead *potential outcome* (also called counterfactual) random variables of the form $V_i(x_{\text{pa}_i})$, where pa_i is a fixed subset of pre_i , and x_{pa_i} is any element in $\mathfrak{X}_{\text{pa}_i}$.⁹ The variable $V_i(x_{\text{pa}_i})$ denotes the value of V_i had the set V_{pa_i} of *direct causes of* V_i been set, possibly contrary to fact, to values pa_i . The existence of a total ordering \prec on indices and the fact that $\text{pa}_i \subseteq \text{pre}_i$ precludes the existence of cyclic causation. That is, we consider causal models that are *recursive*. $V_i(x_{\text{pa}_i})$ may be conceptualized as the output of a *structural equation* $f_i : (x_{\text{pa}_i}, \varepsilon_i) \mapsto x_i$, a function representing a causal mechanism that maps values of x_{pa_i} , as well as the value of a variable ε_i , to values of V_i . Specifically, we may define the error term ε_i to be the vector comprising the set of random variables $\{V_i(x_{\text{pa}_i}) | x_{\text{pa}_i} \in \mathfrak{X}_{\text{pa}_i}\}$ and f_i to be such that $f_i(x_{\text{pa}_i}, \varepsilon_i) \equiv (\varepsilon_i)_{x_{\text{pa}_i}} = V_i(x_{\text{pa}_i})$.

We define NPSEMs as sets of densities over the set of random variables

$$\mathbb{V} \equiv \{V_i(x_{\text{pa}_i}) | i \in V, x_{\text{pa}_i} \in \mathfrak{X}_{\text{pa}_i}\}.$$

Note that \mathbb{V} includes variables V_i which have no parents, and which are thus factual. For simplicity of presentation, we assume \mathfrak{X}_i is always finite, and thus ignore the measure-theoretic complications that arise with defining densities over sets of random variables in the case where some state spaces $\mathfrak{X}_{\text{pa}_i}$ are infinite.

Given a set of one-step-ahead potential outcomes \mathbb{V} , for any $A \subseteq V$ and $i \in V$, the potential outcome $V_i(a)$, the response of V_i had variables in V_A been set to $a \in \mathfrak{X}_A$, is the one step ahead counterfactual $V_i(\text{pa}_i) \in \mathbb{V}$ if $V_A = V_{\text{pa}_i}$, and is otherwise defined via *recursive substitution*:

$$V_i(a) \equiv V_i(a_{\text{pa}_i}, V_{\text{pa}_i \setminus A}(a)). \quad (41.1)$$

8. If some variables do not affect variables later in time, then many non-temporal orders may be used; see [Robins \[1986, Chapter 11\]](#) and later.

9. pa here is short for “parent,” which will be motivated subsequently when we later build a connection to directed graphs.

In other words, this states that $V_i(a)$ is the potential outcome where variables in both pa_i and A are set to their corresponding values in a , and all elements of pa_i not in A are set to whatever values their recursively defined counterfactual versions would have had had V_A been set to a . This is well defined because of the requirement that $\text{pa}_i \subseteq \text{pre}_i$.

Equivalently, $V_i(a)$ is the random variable induced by a modified set of structural equations: specifically the set of functions f_j for V_j such that $A \cap \text{pa}_j \neq \emptyset$ are replaced by modified functions $f_j^a : (x_{\text{pa}_j \setminus A}, \varepsilon_j) \mapsto x_j$ that are obtained from $f_j : (x_{\text{pa}_j}, \varepsilon_j) \mapsto x_j$ by always evaluating $\text{pa}_j \cap A$ at the corresponding values in a .

We will extend our notational shorthand by using index sets to denote sets of potential outcomes themselves. Thus, for $B \subset V$, we let $B(a)$ denote the set of potential outcomes $V_B(a)$. We denote by \mathbb{V}^* the set of all variables derived by Equation (41.1) from \mathbb{V} , for all possible choices of the set A (together with the set \mathbb{V} itself).¹⁰

While the potential outcome and the structural equation formalisms both yield the same causal model, there are some differences regarding the way in which the frameworks are typically presented. Specifically, regarding which “objects” are taken as primitive and which are derived.

The counterfactual formalism here starts with one-step-ahead counterfactuals that intervene on every parent (direct cause) of every variable, and constructs all other counterfactuals by means of recursive substitution. Recursive substitution implies, in particular, that $A(a) \equiv A$. This accords with the substantive claim that it is possible to *first* learn the “natural” value a variable A would take on, and *then* an instant *later* intervene setting it to a specific value a , resulting in all subsequent variables V_i behaving as counterfactual variables $V_i(a)$.

On the other hand, the structural equation formalism typically starts with a set of unaltered structural equations that yields the observed data distribution (via substitution). Counterfactual distributions representing an intervention that sets elements in A to a are generated by replacing structural equations corresponding to elements in A by degenerate functions that yield constants in a [Strotz and Wold 1960, Pearl 2009]. The resulting modified equation system thus represents the set of variables (including A) *after* the action of setting A to the value a . Consequently, there are two subtle but important notational (not conceptual) distinctions:

- Under the standard presentation of structural equation models, used by Pearl, the meaning of a variable such as A , Y , or L , is dependent on the set

10. The set \mathbb{V}^* corresponds to Robins’ *Finest Causally Interpreted Structured Tree Graph as Detailed as the Data*. See Appendix C in Richardson and Robins [2013].

of equations in which it appears. For example in Table 41.2, Y in the left (unmodified) display corresponds to the natural value; in the middle display Y corresponds to the value after intervening on a or $Y(a)$ in counterfactual notation; in the right display Y denotes the value after intervening on A and L , or $Y(a, l)$. In contrast, in the potential outcome framework, variables that are affected by an intervention take on a new name.

- Second, in the standard presentation, the variable “ A ” in the modified set of equations represents the variable after it has been intervened on. Thus, for Pearl, $\text{do}(A = a)$ implies that $A = a$, a property he terms “effectiveness.”¹¹

In this chapter, we will follow the notation conventions that are used in the potential outcome framework, but we stress that formally, NPSEMs and one-step-ahead counterfactuals are equivalent conceptually.¹² See the *Variables* rows in Tables 41.1 and 41.2 to see the correspondence between sets of one-step-ahead counterfactuals and systems of structural equations; see Pearl [1995] and Imbens [2014] for further discussion of the representation of structural equations via potential outcomes.

Given a set $A \subseteq V$, the distribution on $V \setminus A$ resulting from setting A to a by interventions has been denoted in Robins [1986] by $p(V \setminus A | g = a)$, and subsequently as $p(V \setminus A | \text{do}(a))$ [Pearl 2009]. The potential outcome view also allows us to consider distributions $p(V(a))$ for any $A \subseteq V$. In such a distribution, variables in A occur both as random and intervened on versions. We later consider identification theory for distributions of this sort, where the set of treatment variables and outcome variables may intersect.

Recursive substitution in NPSEMs provides a link between observed variables and potential outcomes. In particular, it implies the *consistency property*: for any disjoint $A, B \subseteq V$, $i \in V \setminus (A \cup B)$, $a \in \mathfrak{X}_A$, $b \in \mathfrak{X}_B$,

$$V_B(a) = b \text{ implies } V_i(a, b) = V_i(a). \quad (41.2)$$

See Robins [1986, 1987] and, for a proof using notation similar to this chapter, Richardson and Robins [2013] and Malinsky et al. [2019]. Consistency is

11. If we were to use $A(a)$ to designate the value taken by a variable A after an intervention on A , then we could express this as $A(a) = a$. However, as noted, we use $A(a)$ to designate the value taken by a variable A immediately prior to the intervention.

12. However, as described below, structural equation models are often used under an additional (strong) assumption of independent errors. Since this is a stronger assumption than typically used in the potential outcome framework, we use the acronyms NPSEM and NPSEM-IE to distinguish whether this additional assumption is being made.

often phrased in a simpler form where $A = \emptyset$, yielding the identity $V_i(b) = V_i$ if $V_B = b$.

Equation (41.1) also implies the *causal irrelevance property*, namely that every $V_i(a)$ can be written as a function of a unique minimally causally relevant subset of a , as follows. (See Robins [1986] and Richardson and Robins [2013] and, for a formulation similar to that used here, Malinsky et al. [2019].) Given \mathbb{V}^* derived from \mathbb{V} via Equation (41.1), let $V_i(a) \in \mathbb{V}^*$, and let A^* be the maximal subset of A such that for every $j \in A^*$ there exists a sequence w_1, \dots, w_m that does not intersect A , where $j \in \text{pa}_{w_1}$, $w_i \in \text{pa}_{w_{i+1}}$, for $i = 1, \dots, m - 1$, and $w_m \in \text{pa}_i$. Then, $V_i(a) = V_i(a^*)$.

As an example, given the indices $\{1, 2, 3\}$, under the ordering $1 \prec 2 \prec 3$, if $\text{pa}_2 = \{1\}$ and $\text{pa}_3 = \{2\}$, we have one-step-ahead potential outcomes $V_1, V_2(v_1), V_3(v_2)$, for any values v_1, v_2 . We can define other counterfactuals via Equation (41.1), for example $V_3(v_1) \equiv V_3(V_2(v_1))$. Consistency implies statements of the form $V_1 = v_1 \Rightarrow V_2 = V_2(v_1)$, while causal irrelevance implies $V_3(v_2, v_1) = V_3(v_2)$.

Both consistency and causal irrelevance hold in any NPSEM in the sense that these properties are implied by the existence of a total order on variables we wish to consider, the existence of one-step-ahead counterfactuals, and Equation (41.1). While useful, these properties on their own fail to capture many of the hypotheses that arise in causal inference problems (either by design or assumption). These additional constraints correspond to conditional independence restrictions concerning the error terms in non-parametric structural equations. Although causal models are well defined without reference to graphs, much conceptual clarity may be gained by viewing them graphically. Thus, before describing causal models in detail, we introduce graphs and graphical models.

41.2.1 Graphical Models

Statistical and causal models can be associated with graphs, where vertices represent variables and edges represent (potential) statistical or causal relationships. Formally, random variables are indexed by vertices. However, when we depict graphs we will display them with the random variables as vertices.

We will consider graphs with either directed edges only (\rightarrow), or mixed graphs with both directed and bidirected (\leftrightarrow) edges. Bidirected edges naturally arise as a way to represent (classes of) DAGs with latent variables; see Section 41.4.1. In all cases we will require the absence of directed cycles, meaning that whenever the graph contains a path of the form $V_i \rightarrow \dots \rightarrow V_j$, the edge $V_j \rightarrow V_i$ cannot exist. Directed graphs with this property are called DAGs, and mixed graphs with this property are called acyclic directed mixed graphs (ADMGs). We will refer to graphs by $\mathcal{G}(V)$, where V is the set of random variables indexed by $\{1, \dots, K\}$. We will write \mathcal{G} in place of $\mathcal{G}(V)$ when the vertex set is clear. We will use the following standard

definitions for sets of vertices in a graph:

$$\begin{aligned}
\text{pa}_i^{\mathcal{G}} &\equiv \{j \mid V_j \rightarrow V_i \text{ in } \mathcal{G}\} && \text{(parents of } V_i) \\
\text{an}_i^{\mathcal{G}} &\equiv \{j \mid V_j \rightarrow \dots \rightarrow V_i \text{ in } \mathcal{G}, \text{ or } V_j = V_i\} && \text{(ancestors of } V_i) \\
\text{de}_i^{\mathcal{G}} &\equiv \{j \mid V_j \leftarrow \dots \leftarrow V_i \text{ in } \mathcal{G}, \text{ or } V_j = V_i\} && \text{(descendants of } V_i) \\
\text{dis}_i^{\mathcal{G}} &\equiv \{j \mid V_j \leftrightarrow \dots \leftrightarrow V_i \text{ in } \mathcal{G}, \text{ or } V_j = V_i\} && \text{(the district of } V_i) \\
\text{mb}_i^{\mathcal{G}} &\equiv \{j \mid V_j \leftrightarrow \dots \leftrightarrow V_i \text{ in } \mathcal{G}\} \cup \\
&\quad \{j \mid V_j \rightarrow \circ \leftrightarrow \dots \leftrightarrow V_i \text{ in } \mathcal{G}\} && \text{(the Markov blanket of } V_i).^{13} \quad (41.3)
\end{aligned}$$

We will generally drop the superscript \mathcal{G} if the relevant graph is obvious. By definition, $\text{an}_i^{\mathcal{G}} \cap \text{de}_i^{\mathcal{G}} \cap \text{dis}_i^{\mathcal{G}} = \{V_i\}$. We define these relations on sets disjunctively. For example, $\text{an}_A^{\mathcal{G}} \equiv \bigcup_{V_i \in A} \text{an}_i^{\mathcal{G}}$.

Given a DAG $\mathcal{G}(V)$, a statistical DAG model (also called a Bayesian network) associated with $\mathcal{G}(V)$ is a set of distributions that factorize (equivalently are Markov) with respect to $\mathcal{G}(V)$:

$$p(V) = \prod_{i=1}^K p(V_i \mid V_{\text{pa}_i^{\mathcal{G}}}). \quad (41.4)$$

Given a distribution $p(V)$ that factorizes relative to a DAG $\mathcal{G}(V)$, conditional independence relations that are implied in $p(V)$ by Equation (41.4) can be derived using the well-known d-separation criterion [Pearl 1988]. More precisely, if $p(V)$ is Markov relative to $\mathcal{G}(V)$, then the following *global Markov property* holds: for any disjoint X, Y, Z (where Z may be empty)

$$(X \perp\!\!\!\perp_d Y \mid Z)_{\mathcal{G}(V)} \Rightarrow (X \perp\!\!\!\perp Y \mid Z)_{p(V)}.$$

Here $(X \perp\!\!\!\perp_d Y \mid Z)_{\mathcal{G}(V)}$ denotes that X is d-separated from Y given Z in $\mathcal{G}(V)$; $(X \perp\!\!\!\perp Y \mid Z)_{p(V)}$ indicates that X is independent of Y given Z in $p(V)$.

The global Markov property given by d-separation allows reasoning about conditional independence restrictions implied by the statistical DAG model using qualitative, visual reasoning on paths in the graph.

41.2.2 Causal Models Associated with DAGs

NPSEMs may be associated with directed graphs as well, by associating vertices with indices, and edges with relations given by pa_i , $i \in \{1, \dots, k\}$. Specifically, given

13. Other authors often define the Markov blanket for a variable to be the minimal set M that makes V_i m-separated from $V \setminus (\{V_i\} \cup M)$. Our definition corresponds to the minimal set M such that V_i is m-separated from its non-descendants.

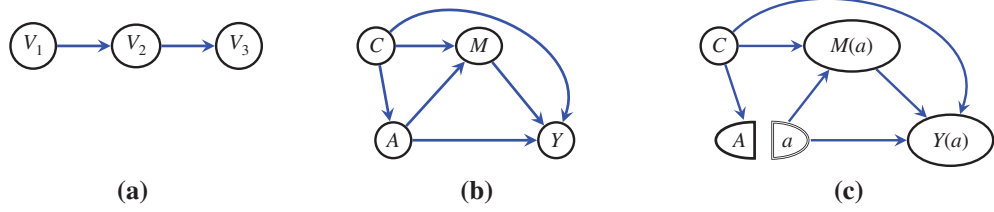


Figure 41.1 (a) A DAG representing a simple NPSEM. (b) A simple causal DAG \mathcal{G} , with a treatment A , an outcome Y , a vector C of baseline variables, and a mediator M . (c) A SWIG $\mathcal{G}(a)$ derived from (a) corresponding to the world where A is intervened on setting it to value a .

a (recursive) NPSEM defined on \mathbb{V} given the sets $\{\text{pa}_i | i \in \{1, \dots, k\}\}$, we construct a *causal diagram*, a DAG $\mathcal{G}(V)$ with a vertex for every V_i , $i \in \{1, \dots, k\}$, and a directed edge from V_j to V_i if $j \in \text{pa}_i$. In other words, $\mathcal{G}(V)$ is defined by the NPSEM by letting $\text{pa}_i^{\mathcal{G}} \equiv \text{pa}_i$ for every i . As an example, the NPSEM defined on the indices $\{1, 2, 3\}$ described in the previous section corresponds to the DAG in Figure 41.1(a). See the *Graph* rows of Tables 41.1 and 41.2 for graphs corresponding to one-step-ahead counterfactuals and structural equations.

Substantive knowledge may motivate additional independence assumptions relating to the set of one-step-ahead counterfactuals \mathbb{V} . As we will show below, such assumptions may also allow causal effects to be identified even when hidden variables are present. Below we introduce two sets of such assumptions.

41.2.2.1 Non-parametric Structural Equations with Independent Errors

A *non-parametric structural equation model with independent errors*, or NPSEM-IE, is the set of distributions such that the K different sets of one-step-ahead variables satisfy:

$$\{V_1\} \perp\!\!\!\perp \{V_2(x_{\text{pa}_2}) | x_{\text{pa}_2} \in \mathfrak{X}_{\text{pa}_2}\} \perp\!\!\!\perp \dots \perp\!\!\!\perp \{V_K(x_{\text{pa}_K}) | x_{\text{pa}_K} \in \mathfrak{X}_{\text{pa}_K}\} \quad (41.5)$$

so that they are mutually independent of one another. Phrased in terms of structural equations $f_i : (x_{\text{pa}_i}, \varepsilon_i) \mapsto x_i$ for each V_i , the NPSEM-IE states that the joint distribution of the disturbance terms factorizes into a product of marginals: $p(\varepsilon_1, \dots, \varepsilon_K) = \prod_{i=1}^K p(\varepsilon_i)$.

NPSEMs with independent errors arise naturally as putative data-generating processes for a closed system. For example, if we are simulating every variable in a model, then it is natural to do this in a stepwise process by specifying a set of structural equations. The equations provide recipes for generating a value for each

variable in turn, given the previous values that have already been simulated plus an independently simulated error term.¹⁴ See Table 41.2 for an example.

However, from an empiricist point of view the assumption of independent errors may be regarded as stronger than necessary. In particular, this assumption permits the identification of causal contrasts that are not subject to experimental verification even in principle;¹⁵ see the discussion in section 38.1 in Chapter 38 in this volume. At the same time, many causal contrasts of interest, including all intervention distributions, may be identified under a much smaller set of assumptions.

41.2.2.2 A Less Restrictive Model: Non-parametric Structural Equations with Single-World (FFRCISTG) Independences

The above observations motivate an alternative approach based on the *finest fully randomized causally interpretable structured tree graph (as detailed as the data)*, or FFRCISTG model of Robins [1986].

The FFRCISTG model is ontologically liberal but epistemologically conservative. Specifically, all the counterfactual queries that may be formulated within the scope of an NPSEM are still well defined under this alternative, but, in contrast to the NPSEM-IE, only those contrasts that could in principle be experimentally verified by experiments on the variables in the system are identified.

An NPSEM with FFRCISTG independences is the set of counterfactual distributions satisfying

$$\text{For each } x_V \in \mathfrak{X}_V, \text{ we have } V_1 \perp\!\!\!\perp V_2(x_{pa_2}) \perp\!\!\!\perp \dots \perp\!\!\!\perp V_K(x_{pa_K}); \quad (41.6)$$

see Robins and Richardson [2010]. Thus, for each $x_V \in \mathfrak{X}_V$ there is a set of K random variables (the K one-step-ahead counterfactuals associated with X_V) and the variables *within* each such set are assumed to be mutually independent. As V_1 is first in the ordering, it has no parents.

The FFRCISTG assumptions could be empirically verified in a set of randomized experiments, one for each X_V , under which we intervene on every variable in turn, setting V_i to the value x_i , but just before doing so, we are able to observe the random variable $V_i(x_{pa_i})$, resulting from our earlier interventions. (Here it is assumed that because we intervene to set V_i to x_i an instant after it is measured, the value $V_i(x_{pa_i})$

14. Note that an NPSEM-IE may also contain unobserved variables, so that they include models described by Pearl as semi-Markovian [Pearl 2009].

15. Specifically, even if it were possible to carry out a randomized experiment manipulating any subset of the variables in the system, we could not directly observe certain counterfactual contrasts that are identified via an NPSEM-IE.

does not causally influence any subsequent variable.) Note the counterfactual random variables in Equation (41.6) all refer to a specific set of values X_V , which thus correspond to a single counterfactual “world.” Note that Equation (41.5) imposes all restrictions in Equation (41.6), and in general exponentially many more.¹⁶ Thus the FFRCISTG is less restrictive than the NPSEM-IE model; in other words, the NPSEM-IE is a strict submodel of the FFRCISTG.

As an example, the NPSEM associated with Figure 41.1(b) is defined using one-step-ahead counterfactuals $C, A(c), M(c, a)$, and $Y(c, a, m)$, for every value set c, a, m . Then the FFRCISTG model restrictions for this NPSEM imply that

$$\text{For each set of values } c, a, m, C \perp\!\!\!\perp A(c) \perp\!\!\!\perp M(a, c) \perp\!\!\!\perp Y(c, a, m), \quad (41.7)$$

while the NPSEM-IE restrictions for the NPSEM state that

$$\text{For each set of values } c, c', c'', a, a', C \perp\!\!\!\perp A(c) \perp\!\!\!\perp M(a, c') \perp\!\!\!\perp Y(c'', a', m). \quad (41.8)$$

The restrictions in Equation (41.7) are a strict subset of the restrictions in Equation (41.8), which are themselves a subset of the restrictions defining the NPSEM-IE.

Interventional distributions of the form $p(V(a))$, for $A \subseteq V$ in both of the above models, may be represented in graphical form by a simple splitting operation on DAGs. The graphs resulting from this operation will be called SWIGs [Richardson and Robins 2013] for reasons that will be described below.

41.2.3 Single-World Intervention Graphs

SWIGs were introduced in Richardson and Robins [2013] as graphical representations of potential outcome distributions that help unify the graphical and potential outcome formalisms. Given a set A of variables and an assignment a to those variables, a SWIG $\mathcal{G}(V(a))$ may be constructed from $\mathcal{G}(V)$ by splitting all vertices in A into a random half and a fixed half, with the random half inheriting all edges with an incoming arrowhead and the fixed half inheriting all outgoing directed edges. Then, all random vertices V_i are re-labeled as $V_i(a)$ or equivalently (due to causal irrelevance) as $V_i(a_{\text{an}_i^*})$, where an_i^* consists of the fixed vertices that are ancestors of V_i in the split graph; the latter labeling is referred to as the *minimal labeling* of the SWIG. By using minimal labeling the SWIG encodes the property of causal irrelevance, so that, for example, if $Y(x)$ appears in $\mathcal{G}(x, z)$ then $Y(x, z) = Y(x)$.

16. In fact, in the case where all variables are binary, the fraction of experimentally untestable constraints implied by the NPSEM-IE rises at a doubly exponential rate in the number of variables. See Richardson and Robins [2013, Section 41.5.], and footnote 4.

Fixed nodes are enclosed by a double line. For an example of a SWIG representing the joint density $p(Y(a), M(a), C(a), A(a)) = p(Y(a), M(a), C, A)$, under the FFRCISTG model (and thus under an NPSEM-IE) associated with the DAG of Figure 41.1(b), see Figure 41.1(c). If the vertex set V is assumed or obvious, we will denote $\mathcal{G}(V(a))$ by $\mathcal{G}(a)$, just as $\mathcal{G}(V)$ is denoted by \mathcal{G} .

Thus a SWIG $\mathcal{G}(V(a))$ is a DAG with vertex set $V(a) \cup a$; the vertices in $V(a)$ correspond to random variables while vertices in a are fixed, taking a specific value.

In a SWIG, every treatment variable has two versions: a fixed version representing the intervention on that treatment, and a random version (which corresponds to measuring the treatment variable just before the intervention took place). This feature of SWIGs allows them to directly express, using d-separation, independence restrictions linking observed versions of treatments, and counterfactual variables representing responses after treatments have been set.

Restrictions of this type, which generalize the well-known conditional ignorability restriction,¹⁷ will be used later to reformulate the second rule of the *do*-calculus, using the language of SWIGs and potential outcomes.

Pearl's "mutilated graphs," which are an alternative graphical representation of interventional distributions, only contain the fixed versions of treatments. This makes it difficult to express restrictions such as conditional ignorability. Instead, the *do*-calculus uses a variant of the mutilated graph where certain outgoing edges are also removed. An additional difficulty with this variant, though it is formally correct, is that vertices on it are not labeled as counterfactual random variables.

Tables 41.1 and 41.2 illustrate, via simple examples, how SWIGs and mutilated graphs differ.

The edges among random variables on the SWIG encode the factorization of the joint distribution $p(V(a))$. More precisely, the FFRCISTG model (and thus the NPSEM-IE) imply that for any $A \subseteq V$, and $a \in \mathfrak{X}_A$, the distribution $p(V(a))$ factorizes with respect to $\mathcal{G}(V(a))$. In other words,

$$p(V(a)) = \prod_{i=1}^K p(V_i(a) | V_{\text{pa}_i \setminus A}(a)). \quad (41.9)$$

Fixed nodes do not occur in the conditioning sets for the terms in Equation (41.9) and thus the presence or absence of edges ($a_i \rightarrow V_i(a_j)$) from fixed nodes to random nodes in $\mathcal{G}(V(a))$ are not reflected in this expression (41.9). However, the fact that a random node is not a descendant of a fixed node does encode information about causal irrelevance. Specifically, if there is no directed path from the

17. Specifically, $Y(a) \perp\!\!\!\perp A | C$, for some set of baseline covariates C .

fixed node a_j to $V_i(a)$ then $V_i(a) = V_i(a_{A \setminus \{j\}})$, hence under minimal labeling a_j will not appear in the label for the vertex $V_i(a_{\text{an}_i^*})$ in $\mathcal{G}(V(a))$.¹⁸ Thus, as noted earlier, by causal irrelevance, $V_i(a) = V_i(a_{\text{an}_i^*})$, where an_i^* consists of the fixed vertices that are ancestors of $V_i(a)$ in $\mathcal{G}(V(a))$. Thus Equation (41.9) may be expressed as:

$$p(V(a)) = \prod_{i=1}^K p \left(V_i(a_{\text{an}_i^*}) \left| \left\{ V_j(a_{\text{an}_j^*}), \text{ for } j \in \text{pa}_i \setminus A \right\} \right. \right).$$

More generally, paths commencing with a fixed node but on which every other node is random also encode information about functional dependence. A path π in a SWIG $\mathcal{G}(a)$ is said to be *Markov relevant* if at most one endpoint is a fixed vertex, and every non-endpoint is random. A Markov relevant path π in $\mathcal{G}(a)$ is *d-connecting given $V_Z(a)$* if every collider on π is an ancestor of a vertex in $V_Z(a)$ and every non-collider on π is not in $V_Z(a)$.

It follows directly from Equation (41.9) that if $V_X(a)$ is d-separated from $V_Y(a)$ given $V_Z(a)$ in $\mathcal{G}(a)$ then $V_X(a) \perp\!\!\!\perp V_Y(a) \mid V_Z(a)$ in $p(V(a))$, so that d-separation relations among random variables encode conditional independence. In addition, the absence of any d-connecting path in $\mathcal{G}(V(a))$ between a fixed node a_j and a set of random vertices $V_Y(a)$, given a (possibly empty) set of random variables $V_Z(a)$, encodes that $p(V_Y(a) \mid V_Z(a))$ does not depend on the value of a_j . Thus we allow d-separation statements of the form $(a_j \perp\!\!\!\perp_a V_Y(a) \mid V_Z(a))_{\mathcal{G}(V(a))}$.¹⁹ More generally, given three disjoint subsets $Y, X, Z \subseteq V$, where Z may be empty, and a set $A' \subseteq A$, then

$$(V_Y(a) \perp\!\!\!\perp_a V_X(a), a_{A'} \mid V_Z(a))_{\mathcal{G}(V(a))} \quad (41.10)$$

if in the SWIG $\mathcal{G}(V(a))$ there is no path d-connecting a random vertex $V_i(a)$ with $i \in X$ or a fixed vertex a_j with $j \in A'$ to a random vertex in $V_j(a)$ with $j \in Y$ given $V_Z(a)$. Note that, by definition, fixed vertices may only arise on one side of a d-separation statement [Equation (41.10)]. Conversely, a possibly d-connecting path may only contain at most one fixed node in which case it is an endpoint vertex (thus, as in

18. In the [Appendix](#) we briefly consider using the SWIG to make inferences about weaker causal models, including the agnostic causal model, and models in which the absence of a directed edge corresponds to the absence of a population-level direct effect. In the latter models, the equality $V_i(a) = V_i(a \cap \text{an}_i^*)$ would no longer hold, and minimal labelings are constructed using a (possibly strict) edge super-graph of the graph used for the factorization (Equation (41.9)); see Section 7 in [\[Richardson and Robins 2013\]](#).

19. This represents an extension of the notion of d-separation in [Richardson and Robins \[2013\]](#). Our extension here consists only in allowing fixed vertices to appear in, at most, one side of a d-separation statement (not the conditioning set). The semantics for these extended d-separation statements are given in Equation (41.11).

Richardson and Robins [2013], fixed nodes never arise as non-endpoint vertices on d-connecting paths).

Results for DAG models with fixed nodes [Richardson et al. 2017] imply the following:

Proposition 41.1 SWIG Global Markov property

Under the FFRCISTG for \mathcal{G} , for every set A , disjoint sets of random vertices $V_X(a)$, $V_Y(a)$, $V_Z(a)$ and a set of fixed nodes $a_{A'}$, where $A' \subseteq A$,

$$\begin{aligned} \text{if } (V_Y(a) \perp\!\!\!\perp_d V_X(a), a_{A'} \mid V_Z(a))_{\mathcal{G}(V(a))} \text{ then, for some } f(\cdot), \\ p(V_Y(a) = v_Y \mid V_Z(a) = v_Z, V_X(a) = v_X) = p(V_Y(a) = v_Y \mid V_Z(a) = v_Z) \\ = f(v_Y, v_Z, a_{A'}). \end{aligned} \tag{41.11}$$

Example 41.1 Consider the global Markov property associated with the SWIG $\mathcal{G}(a)$ in Figure 41.2(b), corresponding to the FFRCISTG model shown in Figure 41.2(a). Since a is d-separated from $Y(a)$ given $M(a)$ in $\mathcal{G}(a)$,

$$p(Y(a) = y \mid M(a) = m) = f(y, m). \tag{41.12}$$

Hence this distribution is not a function of a , even though $M(a)$ and $Y(a)$ are minimally labeled, so $M(a) \neq M(a')$ and $Y(a) \neq Y(a')$ for $a \neq a'$. In addition, it is well known that in the FFRCISTG model corresponding to Figure 41.2(a),

$$p(Y(a) = y \mid M(a) = m) = \sum_{a'} p(y \mid m, a') p(a'),$$

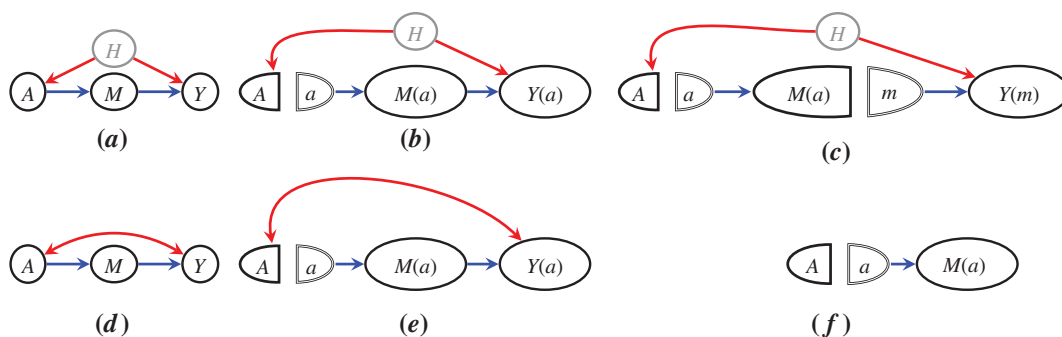


Figure 41.2 (a) A hidden variable causal model. (b) A SWIG corresponding to an intervention that sets A to a in the causal model represented by (a). (c) A SWIG corresponding to an intervention that sets A to a and M to m . (d) A latent projection of the DAG in (a). (e) A latent projection of the SWIG in (b). (f) A latent projection of the SWIG in (b) onto an ancestral set of vertices $A, M(a)$ and a .

which is not equal to $p(y|m)$ under the given model. Hence, the function $f(v_Y, v_Z, a_{A \setminus A'})$ is not necessarily equal to the conditional distribution $p(V_Y(a_{A \setminus A'}) | V_Z(a_{A \setminus A'}))$.

Remark 41.1 Since, by construction, all edges in $\mathcal{G}(V(a))$ are directed out of a_j , in the case where Z is the empty set, there is a d-connecting path between a_j and $V_i(a)$ if and only if a_j is an ancestor of $V_i(a)$ in $\mathcal{G}(V(a))$; as noted earlier, this is automatically reflected with the minimal labeling of the vertices.

In Equation (41.12) we see an example where $p(Y(a) | M(a))$ does not depend on a , even though $Y(a)$ and $M(a)$ are minimally labeled. One might wonder whether it is possible to have the converse situation whereby a conditional distribution *does* depend on a fixed vertex that is not present in any minimal label. The Proposition 41.2 shows that this cannot arise:

Proposition 41.2 In a minimally labeled SWIG $\mathcal{G}(a)$, if a fixed vertex a_i is d-connected to $V_j(a_{\text{an}_j^*})$ given $\{V_{k_1}(a_{\text{an}_{k_1}^*}), \dots, V_{k_p}(a_{\text{an}_{k_p}^*})\}$ then either $i \in \text{an}_j^*$ or $i \in \text{an}_{k_s}^*$ for some s .

In other words, if a fixed vertex is d-connected by a path to a random vertex given some conditioning set, then the fixed vertex either appears in the minimal label for the other endpoint, or a vertex in the conditioning set. This follows since if there is a d-connecting path on which a_i is an endpoint then, since a_i only has children in $\mathcal{G}(a)$, the path is directed out of a_i . The conclusion then follows since if the path contains no colliders then $V_j(a_{\text{an}_j^*})$ is a descendant of a_i ; if the path contains a collider then a_i is an ancestor of that collider, which, by definition of d-connection is itself an ancestor of a vertex in $V_{k_s}(a_{\text{an}_{k_s}^*})$.

41.2.4 Factorization Associated with the SWIG Global Markov Property

As noted earlier, the factorization [Equation (41.9)] corresponds solely to the induced subgraph of $\mathcal{G}(a)$ on the random vertices. We now derive the factorization corresponding to the SWIG global Markov property. Consider a single term in Equation (41.9):

$$\begin{aligned} p(V_i(a) = v_i | V_{\text{pa}_i \setminus A}(a) = v_{\text{pa}_i \setminus A}) \\ &= p(V_i(a, v_{\text{pa}_i \setminus A}) = v_i | V_{\text{pa}_i \setminus A}(a, v_{\text{pa}_i \setminus A}) = v_{\text{pa}_i \setminus A}) \\ &= p(V_i(a, v_{\text{pa}_i \setminus A}) = v_i) \\ &= p(V_i(a_{\text{pa}_i \cap A}, v_{\text{pa}_i \setminus A}) = v_i) \end{aligned} \tag{41.13}$$

$$\begin{aligned} &= p(V_i(a_{\text{pa}_i \cap A}, v_{\text{pa}_i \setminus A}) = v_i | V_{\text{pa}_i \cap A}(a_{\text{pa}_i \cap A}, v_{\text{pa}_i \setminus A}) = a_{\text{pa}_i \cap A}, \\ &\quad V_{\text{pa}_i \setminus A}(a_{\text{pa}_i \cap A}, v_{\text{pa}_i \setminus A}) = v_{\text{pa}_i \setminus A}) \\ &= p(V_i = v_i | V_{\text{pa}_i \cap A} = a_{\text{pa}_i \cap A}, V_{\text{pa}_i \setminus A} = v_{\text{pa}_i \setminus A}). \end{aligned} \tag{41.14}$$

Here the first equality follows from consistency; the second follows from Equation (41.9) for $\mathcal{G}(a, v_{\text{pa}_i \setminus A})$. The third equality follows from causal irrelevance since if we intervene on all the parents of V_i then no other variables have a causal effect on V_i . The fourth line follows from Equation (41.9) for $\mathcal{G}(a_{\text{pa}_i \cap A}, v_{\text{pa}_i \setminus A})$. The fifth line again follows from consistency. Thus, we have:

Proposition 41.3 Under the FFRCISTG models associated with a graph \mathcal{G} , we have the following identification formula:

$$p(V(a) = v) = \prod_{i=1}^K p(V_i(a) = v_i \mid V_{\text{pa}_i \setminus A}(a) = v_{\text{pa}_i \setminus A}) \quad (41.15)$$

$$= \prod_{i=1}^K p(v_i \mid a_{\text{pa}_i \cap A}, v_{\text{pa}_i \setminus A}). \quad (41.16)$$

Thus $p(V(a))$ is identified if all of the conditional distributions in Equation (41.16) are identified.²⁰

Now consider a DAG $\mathcal{G}^*(V \cup A^*)$ containing disjoint sets of vertices V and A^* , with the same set of edges as in $\mathcal{G}(a)$ under the natural correspondence: $V_i \Leftrightarrow V_i(a)$ and $A_i^* \Leftrightarrow a_i$. Then Equation (41.16) corresponds syntactically to the (subset of) terms in the DAG factorization for \mathcal{G}^* associated with the variables in V . This then establishes the SWIG global Markov property via results on conditional graphs [Richardson et al. 2017].²¹

The modified factorization [Equation (41.16)] is known as the *extended g-formula* [Robins et al. 2004, Richardson and Robins 2013]. Like the original factorization (41.4), Equation (41.16) has a term for every $V_i \in V$ not merely for every $V_i \in V \setminus A$.²² An alternative proof of the extended g-formula is given in Richardson and Robins [2013].

Proposition 41.4 follows directly from Equation (41.16) and is included here because a generalization of this result, Proposition 41.5 below, plays an important role in the identification of causal effects in DAGs with hidden variables.

20. This may not hold in the absence of positivity; see Section 41.2.1 for further discussion.

21. For the sole purpose of establishing the SWIG global Markov property, it is sufficient to show that $p(V_i(a) = v_i \mid V_{\text{pa}_i \setminus A}(a) = v_{\text{pa}_i \setminus A})$ is not a function of the fixed nodes that are not in pa_i , that is $a_{A \setminus \text{pa}_i}$. This is established by Equation (41.13). Under the FFRCISTG, $p(V_i(a_{\text{pa}_i \cap A}, v_{\text{pa}_i \setminus A}))$ exists even if $p(a_{\text{pa}_i \cap A}, v_{\text{pa}_i \setminus A}) = 0$.

22. This is because the extended g-formula includes the value a variable takes on just before it is intervened upon and set to a constant a_i .

Proposition 41.4 If \mathcal{G} is a DAG with SWIG $\mathcal{G}(a)$ then for all $c_k \in \mathfrak{X}_k$

$$\begin{aligned} p(V(a, c_k) = v) \\ = p(V_{-k}(a) = v_{-k}, V_k(a) = c_k) \times \frac{p\left(V_k(a) = v_k \mid V_{\text{pa}_k^{\mathcal{G}(a)}}(a) = v_{\text{pa}_k^{\mathcal{G}(a)}}\right)}{p\left(V_k(a) = c_k \mid V_{\text{pa}_k^{\mathcal{G}(a)}}(a) = v_{\text{pa}_k^{\mathcal{G}(a)}}\right)}, \end{aligned}$$

where $V_{-k} \equiv \{V_1, \dots, V_{k-1}, V_{k+1}, \dots, V_K\}$, provided the conditional probability in the denominator is positive.

41.2.5 SWIG Representation of the Defining FFRCISTG Assumptions

Consider the special case in which $A = V$; in the resulting graph $\mathcal{G}(V(v))$ every variable (in V) has been split and thus no pair of random variables are joined by an edge. The factorization [Equation (41.9)] then becomes:

$$p(V(v^*) = v) = \prod_{i=1}^K p(V_i(v^*) = v_i) = \prod_{i=1}^K p\left(V_i(v_{\text{pa}_i}^*) = v_i\right), \quad (41.17)$$

and thus for a fixed $v^* \in \mathfrak{X}_V$ the one-step-ahead counterfactuals $V_1(v_{\text{pa}_1}^*), \dots, V_K(v_{\text{pa}_K}^*)$ are independent. Note that Equation (41.17) holding for all $v^* \in \mathfrak{X}_V$ is equivalent to Equation (41.6) and thus defines the FFRCISTG model.

41.3 The Potential Outcomes Calculus and Identification

Pearl presented the three rules of *do*-calculus as an inference system for deriving identification results for causal inference problems. The *do*-calculus is stated as three identities involving (conditional) interventional distributions, with preconditions given by d-separation (or m-separation) statements on graphs derived from the causal diagram $\mathcal{G}(V)$.

Here we reformulate and extend these three rules as a “potential outcomes calculus” or “po-calculus” for short. The rules are as follows:

- 1: $p(Y(x)|Z(x), W(x)) = p(Y(x)|W(x))$ if $(Y(x) \perp\!\!\!\perp Z(x)|W(x))_{\mathcal{G}(x)}$,
- 2: $p(Y(x, z)|W(x, z)) = p(Y(x)|W(x), Z(x) = z)$ if $(Y(x, z) \perp\!\!\!\perp Z(x, z)|W(x, z))_{\mathcal{G}(x, z)}$,
- 3: $p(Y(x, z)) = p(Y(x))$ if $(Y(x, z) \perp\!\!\!\perp z)_{\mathcal{G}(x, z)}$,

where $\mathcal{G}(x)$ and $\mathcal{G}(x, z)$ are SWIGs describing interventions on X and $X \cup Z$. The sets Z , Y , and W are assumed to be disjoint; X may overlap with the other sets, but if $Z \cap X \neq \emptyset$ then we require $x_{X \cap Z} = z_{X \cap Z}$, so that the assignments are consistent.

Rule 1 can be viewed as the part of the SWIG global Markov property pertaining to random (rather than fixed) variables.

Rule 2 can be viewed as a kind of generalized conditional ignorability rule which follows from Rule 1 and recursive substitution. Specifically, by recursive substitution or minimal labeling $Z(x, z) = Z(x)$; further,

$$\begin{aligned} p(Y(x, z) | W(x, z)) &= p(Y(x, z) | W(x, z), Z(x) = z) \\ &= p(Y(x) | W(x), Z(x) = z) \end{aligned}$$

here the first equality follows by the given d-separation and Rule 1, while the second follows from consistency (or recursive substitution) since $Y(x, z) = Y(x)$ and $W(x, z) = W(x)$ given that $Z(x) = z$.

Rule 3 expresses the property of causal irrelevance that interventions only affect descendants: Note that the Rule 3 condition $(Y(x, z) \perp\!\!\!\perp z)_{\mathcal{G}(x, z)}$ is, by definition, equivalent to the fixed vertex (or vertices) z not being an ancestor of any vertex in $Y(x, z)$ in the SWIG $\mathcal{G}(x, z)$ where the vertices in X and Z have been split.

Further, if a variable $Y(x)$ appears in the SWIG $\mathcal{G}(x, z)$ (with minimal labeling), then there is no directed path from any fixed vertex in z to $Y(x)$, and $Y(x) = Y(x, z)$.²³ Thus the minimal labeling of the SWIG implicitly encodes all applications of Rule 3, in the sense that if the (minimally labeled) vertex $Y(x)$ is present in the SWIG then for any set Z , disjoint from X , $p(Y(x)) = p(Y(x, z))$.

As we have shown here, the po-calculus directly follows from the SWIG global Markov property, which is implied by both the FFRCISTG model (and thus the NPSEM-IE), consistency, and causal irrelevance, where the latter two hold for any NPSEM.

Rule 3, as stated here,²⁴ simply states that interventions only affect descendants and thus is simpler than Rule 3 in the original formulation of the *do*-calculus. It is proved in Malinsky et al. [2019] that this reformulated Rule 3, in conjunction with the other two rules, is equivalent to Pearl's *do*-calculus in the sense that the three rules stated here imply the original three rules. The rules stated here are more general in that we allow X to overlap with Y , Z , and W , which is not possible within the framework and notation of the original *do*-calculus. As a consequence, as we will show below, there are additional identification results that follow from the po-calculus, but not the *do*-calculus. However, if we restrict the po-calculus

23. Thus under the counterfactual model, as defined by one-step-ahead counterfactuals, $V_i(a_{pa_i})$ and recursive substitution (41.1), we have a stronger implication than Rule 3: if $(Y(x, z) \perp\!\!\!\perp z)_{\mathcal{G}(x, z)}$ then $Y(x, z) = Y(x)$. Notwithstanding this, we formulate Rule 3 in terms of the equality of distributions because we wish these rules to be logically equivalent to the original *do*-calculus and also apply to weaker causal models; see Footnote 17 and Section 41.A.2.

24. Rule 3 in this chapter is called Rule 3* by Malinsky et al. [2019].

rules to the case where X does not overlap with Y, Z then they are equivalent to the *do*-calculus.

It may also be noted that the po-calculus is formulated using a uniform type of graph, the SWIG, for displaying the preconditions for each rule.²⁵

Remark 41.2 We note that there are other types of equality between distributions that do not correspond to a single application of the po-calculus rules. For instance, in Example 41.1 it follows that

$$p(Y(a) | M(a)) = p(Y(a') | M(a')) \quad \text{for } a, a' \in \mathcal{X}_A. \quad (41.18)$$

This holds even though $Y(a)$ and $M(a)$ depend on a and $p(Y(a) | M(a)) \neq p(Y | M)$. This is a form of independence.²⁶ Such constraints are captured by the full global Markov property for SWIGs: notice that a is d-separated from $Y(a)$ given $M(a)$ in the SWIG shown in Figure 41.2(b). However, the equality in (41.18) may be derived from three applications of the po-calculus (or the do-calculus) rules.

41.4 Identification in Hidden Variable Causal Models

If some variables in an NPSEM are unobserved, identification becomes more complicated, and some interventional distributions become non-identified. Identification theory in NPSEMs associated with $\mathcal{G}(V \cup H)$, where H are hidden variables, is often described in terms of a special ADMG $\mathcal{G}(V)$ obtained from $\mathcal{G}(V \cup H)$ via the *latent projection* operation [Verma and Pearl 1990]. Any two distinct hidden variable DAGs $\mathcal{G}_1(V \cup H_1), \mathcal{G}_2(V \cup H_2)$ that share the latent projection $\mathcal{G}(V) = \mathcal{G}_1(V) = \mathcal{G}_2(V)$ also share all equality constraints on the observed marginal distribution [Evans, 2018], as well as non-parametric identification theory, in the sense that effects are identified in \mathcal{G}_1 if and only if they are identified in \mathcal{G}_2 , and by the same functional [Richardson et al. 2017].

In cases where $p(V(a))$ is identified, the functional is a kind of modified factorization associated with nested Markov models of ADMGs [Richardson et al. 2017].

41.4.1 Latent Projection ADMGs

Given a DAG $\mathcal{G}(V \cup H)$, where V are observed and H are hidden variables, a latent projection $\mathcal{G}(V)$ is the following ADMG with a vertex set V . An edge $A \rightarrow B$ exists in $\mathcal{G}(V)$ if there exists a directed path from A to B in $\mathcal{G}(V \cup H)$ with all intermediate

25. Whereas the original *do*-calculus involves three different constructions: $G_{\bar{X}}$, $G_{\bar{X}Z}$, and $G_{\bar{X}(Z(\bar{W}))}$.

26. Formally we may think of $P(Y(a) | M(a))$ as forming a kernel $q(y | m, a)$, which is a set of conditional distributions indexed by a . The constraint is then an independence in this kernel [Richardson et al. 2017].

vertices in H . Similarly, an edge $A \leftrightarrow B$ exists in $\mathcal{G}(V)$ if there exists a path without consecutive edges $\rightarrow \circ \leftarrow$ from A to B with the first edge on the path of the form $A \leftarrow$, the last edge on the path of the form $\rightarrow B$, and all intermediate vertices on the path in H . Latent projections of hidden variable DAGs may be viewed as graphical versions of marginal distributions, in the following sense. Just as conditional independences may be read off a DAG using d-separation, they may be read from an ADMG via the natural extension of d-separation to ADMGs, which is called m-separation [Richardson 2003].

If $p(V \cup H)$ factorizes with respect to $\mathcal{G}(V \cup H)$, then for any disjoint subsets A, B, C of V , if A is m-separated from B given C , then A is independent of B conditionally on C in the marginal distribution $p(V)$. Since latent projections define an infinite class of hidden variable DAGs that share identification theory, identification algorithms are typically defined directly on latent projections for simplicity.

Given $A \subseteq V$ in a hidden variable DAG $\mathcal{G}(V \cup H)$, we can construct the latent projection of the SWIG $\mathcal{G}(V(a) \cup H(a))$ directly from the ADMG $\mathcal{G}(V)$, we denote the resulting ADMG (with fixed nodes) by $\mathcal{G}(V(a))$. We can extend d-separation on SWIGs constructed from DAGs to m-separation on SWIGs constructed from ADMGs, and define the SWIG global Markov property on SWIG ADMGs analogously to the SWIG global Markov property on SWIG DAGs. Similarly, we can restate po-calculus rules using m-separation on SWIG ADMGs.

As an example, the latent projection of the hidden variable DAG in Figure 41.2(a) is shown in Figure 41.2(d), while the latent projection of the SWIG in Figure 41.2(b) is shown in Figure 41.2(e).

All vertex relations defined in (41.3) translate without change to any SWIG $\mathcal{G}(V(a))$, except by convention $\text{dis}_i^{\mathcal{G}(V(a))}$, $\text{mb}_i^{\mathcal{G}(V(a))}$, and $\text{pre}_i^{\mathcal{G}(V(a))}$ may only contain random vertices, in other words, they are subsets of $V(a)$.

We will describe a complete identification algorithm in hidden variable DAG models for all distributions of the form $p(Y(a))$, where Y may potentially intersect A . The original formulation of the problem in Tian and Pearl [2002], Shpitser and Pearl [2006a], and Richardson et al. [2017] assumed $Y \cap A = \emptyset$, and yielded the *ID algorithm*.

We call our version of the algorithm the *extended ID algorithm*, by analogy with the *extended g-formula* [Equation (41.16)]. The extended ID algorithm will be formulated using SWIGs defined on latent projection ADMGs of the underlying hidden variable DAG. The algorithm will take advantage of the fact that under certain assumptions given by the causal model, a single splitting operation that defines a counterfactual distribution in a SWIG can be phrased in terms of the observed data distribution. This insight can be applied inductively to obtain results of multiple splitting operations as functionals of the observed data distribution.

The extended ID algorithm expresses the functional for $p(Y(a))$ as a counterfactual factorization in a certain SWIG ADMG, where terms of the factorization correspond to districts in the SWIG. It then aims to identify each term by finding a sequence of splitting operations, possibly interleaved with marginalization operations. Perhaps surprisingly, this always suffices to obtain identification whenever identification is possible.

41.4.2 The Identified Splitting Operation in a SWIG

A general identification algorithm for interventional distributions in hidden variable DAG models involves, as an essential step, expressing the counterfactual distribution $p(V(a, c_k))$ as a function of another counterfactual distribution $p(V(a))$, where one fewer variable (V_k) has been intervened on, using restrictions in $\mathcal{G}(V(a))$. Specifically, we have the following generalization of Proposition 41.4:

Proposition 41.5 Given an ADMG $\mathcal{G}(V)$ with SWIG $\mathcal{G}(V(a))$, if $V_k(a)$ is not split, so $k \notin A$, and $V_k(a)$ is such that there is no other random vertex that is both a descendant of $V_k(a)$ and in the same district as $V_k(a)$ then for all $c_k \in \mathfrak{X}_k$:

$$\begin{aligned} p(V(a, c_k) = v) \\ = p(V_{-k}(a) = v_{-k}, V_k(a) = c_k) \times \frac{p\left(V_k(a) = v_k \mid V_{\text{mb}_k^{\mathcal{G}(V(a))}}(a) = v_{\text{mb}_k^{\mathcal{G}(V(a))}}\right)}{p\left(V_k(a) = c_k \mid V_{\text{mb}_k^{\mathcal{G}(V(a))}}(a) = v_{\text{mb}_k^{\mathcal{G}(V(a))}}\right)}, \end{aligned}$$

where $V_{-k} \equiv \{V_1, \dots, V_{k-1}, V_{k+1}, \dots, V_K\}$, provided the conditional probability in the denominator is positive.

In other words, this proposition states that if $V_k(a)$ satisfies the graphical condition in $\mathcal{G}(V(a))$ then $p(V(a, c_k))$, the joint distribution over all variables (including A and V_k) resulting from intervening to set A to a and V_k to c_k , may be obtained from $p(V(a))$ by evaluating at $V_k(a) = c_k$ and multiplying by a ratio of conditional densities for $V_k(a)$.

The graphical condition may be interpreted as requiring that in the world where we have already intervened on A , there is no sequence of variables between V_k and any of its causal descendants such that there is an unmeasured confounder between each pair.

There exist counterfactual distributions which are identified, but where the above proposition does not directly apply to the observed data distribution. For example, in the graph in Figure 41.3(a), $p(Y_1(a)) = p(Y_1|a)$, and $p(Y_2(a)) = p(Y_2)$. Nevertheless, the preconditions to applying Proposition 41.5 do not apply to the original graph, meaning that the distribution represented by the SWIG in

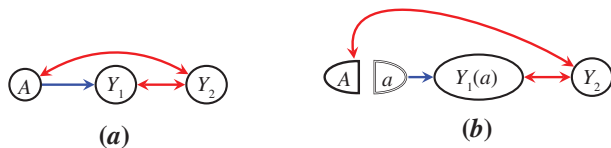


Figure 41.3 (a) A graph where $p(Y_1(a))$ and $p(Y_2(a))$ are identified, but Proposition 41.5 may not be applied. (b) A SWIG showing a splitting operation that is not identified according to Proposition 41.5.

Figure 41.3(b) is not equal to the functional of the observed data distribution described in the proposition. In fact, the joint distribution associated with this SWIG is not identified at all, as was shown in Tian and Pearl [2002]. Nevertheless, identification of $p(Y_2(a))$ and $p(Y_1(a))$ may be obtained by the identification algorithm we describe below.

In the next section, we will apply the proposition iteratively, in conjunction with marginalization steps, in order to obtain a complete algorithm for identifying a margin $p(Y(a))$.

Proof. Fix an ordering \prec' on vertex indices such that \prec' is topological in $\mathcal{G}(V(a))$ and such that no element in the district of $V_k(a)$ occurs later in the ordering than $V_k(a)$. Such a topological ordering exists because, by hypothesis, no vertex in the district of $V_k(a)$ is a descendant of $V_k(a)$. For any index j , define pre'_j to be the set of predecessor indices according to \prec' .

$$\begin{aligned}
 p(V(a, c_k) = v) &= \prod_{i \in \text{pre}'_k} p(V_i(a, c_k) = v_i \mid V_{\text{pre}'_i}(a, c_k) = v_{\text{pre}'_i}) \times p(V_k(a, c_k) = v_k \mid V_{\text{pre}'_k}(a, c_k) = v_{\text{pre}'_k}) \\
 &\quad \times \prod_{j \in \{k\} \cup \text{pre}'_k} p(V_j(a, c_k) = v_j \mid V_{\text{pre}'_j}(a, c_k) = v_{\text{pre}'_j}) \\
 &= \prod_{i \in \text{pre}'_k} p(V_i(a) = v_i \mid V_{\text{pre}'_i}(a) = v_{\text{pre}'_i}) \times p(V_k(a) = v_k \mid V_{\text{pre}'_k}(a) = v_{\text{pre}'_k}) \\
 &\quad \times \prod_{j \in \{k\} \cup \text{pre}'_k} p(V_j(a, c_k) = v_j \mid V_{\text{pre}'_j}(a, c_k) = v_{\text{pre}'_j}) \\
 &= \prod_{i \in \text{pre}'_k} p(V_i(a) = v_i \mid V_{\text{pre}'_i}(a) = v_{\text{pre}'_i}) \times p(V_k(a) = v_k \mid V_{\text{pre}'_k}(a) = v_{\text{pre}'_k}) \\
 &\quad \times \prod_{j \in \{k\} \cup \text{pre}'_k} p(V_j(a) = v_j \mid V_{\text{pre}'_j \setminus \{k\}}(a) = v_{\text{pre}'_j}, V_k(a) = c_k) \\
 &= \prod_{i \in \text{pre}'_k} p(V_i(a) = v_i \mid V_{\text{pre}'_i}(a) = v_{\text{pre}'_i}) \times p(V_k(a) = c_k \mid V_{\text{pre}'_k}(a) = v_{\text{pre}'_k}) \\
 &\quad \times \prod_{j \in \{k\} \cup \text{pre}'_k} p(V_j(a) = v_j \mid V_{\text{pre}'_j \setminus \{k\}}(a) = v_{\text{pre}'_j}, V_k(a) = c_k) \\
 &\quad \times p(V_k(a) = v_k \mid V_{\text{pre}'_k}(a) = v_{\text{pre}'_k}) / p(V_k(a) = c_k \mid V_{\text{pre}'_k}(a) = v_{\text{pre}'_k})
 \end{aligned}$$

$$\begin{aligned}
&= p(V_{-k}(a) = v_{-k}, V_k(a) = c_k) \\
&\quad \times p(V_k(a) = v_k \mid V_{\text{pre}'_k}(a) = v_{\text{pre}'_k}) / p(V_k(a) = c_k \mid V_{\text{pre}'_k}(a) = v_{\text{pre}'_k}) \\
&= p(V_{-k}(a) = v_{-k}, V_k(a) = c_k) \frac{p\left(V_k(a) = v_k \mid V_{\text{mb}_k^{\mathcal{G}(V(a))}}(a) = v_{\text{mb}_k^{\mathcal{G}(V(a))}}\right)}{p\left(V_k(a) = c_k \mid V_{\text{mb}_k^{\mathcal{G}(V(a))}}(a) = v_{\text{mb}_k^{\mathcal{G}(V(a))}}\right)}.
\end{aligned}$$

Here the first identity is via the chain rule of probability applied to $p(V(a), c_k)$ using the ordering \prec' , the second by Rule 3 (causal irrelevance) applied to elements indexed by $\{k\} \cup \text{pre}'_k$ in $\mathcal{G}(V(a), c_k)$, the third by Rule 2 (generalized ignorability) applied to every term in the second product in $\mathcal{G}(V(a), c_k)$, and the assumption on \prec' that all elements of $\text{dis}_k^{\mathcal{G}(V(a))}$ are in $V_{\text{pre}'_k}(a)$, the fourth by multiplying and dividing by $p(V_k(a) = c_k \mid V_{\text{pre}'_k}(a) = v_{\text{pre}'_k})$, the fifth by the chain rule, the sixth by Rule 1 (m-separation) applied to $\mathcal{G}(V(a))$ and the definition of $\text{mb}_k^{\mathcal{G}(V(a))}$. ■

41.4.3 The Extended ID Algorithm

There are SWIGs $\mathcal{G}(V(a))$ for which, for some variable $V_k(a)$ we are not able to apply Proposition 41.5, but where it may be applied to a SWIG $\mathcal{G}(Y(a))$, where $Y(a)$ is an ancestral subset of $V(a)$ in $\mathcal{G}(V(a))$. Here a set Y of vertices in a (SWIG) ADMG \mathcal{G}^* is said to be *ancestral* if $V_i \in Y$ implies $\text{an}_i^{\mathcal{G}^*} \subseteq Y$.

Marginal distributions $p(Y(a))$ obtained from $p(V(a))$ that correspond to ancestral sets in $\mathcal{G}(V(a))$ have the nice property that a latent projection $\mathcal{G}(Y(a))$ is always equal to an induced subgraph $(\mathcal{G}(V(a)))_{Y(a)}$ of a SWIG $\mathcal{G}(V(a))$, with $\mathcal{G}(Y(a))$ having strictly fewer vertices and edges than $\mathcal{G}(V(a))$ if $Y(a) \subset V(a)$. For example, given the SWIG in Figure 41.2(e), the latent projection onto the ancestral subset A , $M(a)$ and a yields the SWIG shown in Figure 41.2(f). We describe the precise way in which splitting and ancestral margin operations are used to obtain identification below.

Specifically, complete non-parametric identification for intervention distributions associated with the FFRCISTG model may be obtained from: (i) the district factorization in the appropriate SWIG, (ii) the identified splitting operation described in the previous section, and (iii) marginalization steps that lead to marginal distributions corresponding to ancestral sets of vertices in SWIGs. All of these steps may be justified via the po-calculus.

For any (possibly intersecting) subsets Y, A of V in a latent projection $\mathcal{G}(V)$ representing a causal DAG $\mathcal{G}(V \cup H)$, define $Y^*(a)$ to be the random ancestors of $Y(a)$ in $\mathcal{G}(V(a))$. Clearly, if $p(Y^*(a))$ is identified, then we may recover $p(Y(a))$ since:

$$P(Y(a) = y) = \sum_{u \in \mathfrak{X}_{Y^* \setminus Y}} p(V_Y(a) = y, V_{Y^* \setminus Y}(a) = u). \quad (41.19)$$

Though less obvious, extensions of results in Shpitser and Pearl [2006a] imply that the converse also holds, so that if $p(Y(a))$ is identified (for all parameter values) then $p(Y^*(a))$ is identified. Consequently, in the foregoing we will assume that $Y(a)$ is an ancestral set of (random) vertices in $\mathcal{G}(V(a))$.

If $p(Y(a))$ is identified, then this may be obtained by breaking this joint distribution into districts in $\mathcal{G}(Y(a))$. For each such district $D(a)$, define the set of *strict (random) parents* as $\text{pas}_{D(a)}^{\mathcal{G}(Y(a))} \equiv \text{pa}_{D(a)}^{\mathcal{G}(Y(a)) \setminus (D(a) \cup a)}$.

First, we show that $p(Y(a) = y)$ can be factorized into a set of terms of the form $p(D(a), \mathbf{v}_{\text{pas}_{D(a)}^{\mathcal{G}(Y(a))}})$, as follows.

$$\begin{aligned}
p(Y(a) = \mathbf{v}_Y) &= \prod_{i \in Y} p(V_i(a) = v_i \mid V_{Y \cap \text{pre}_i}(a) = \mathbf{v}_{Y \cap \text{pre}_i}) \quad (41.20) \\
&= \prod_{D \in \mathcal{D}(\mathcal{G}(Y(a)))} \prod_{i \in D} p(V_i(a) = v_i \mid V_{Y \cap \text{pre}_i}(a) = \mathbf{v}_{Y \cap \text{pre}_i}) \\
&= \prod_{D \in \mathcal{D}(\mathcal{G}(Y(a)))} \prod_{i \in D} p(V_i(a, \mathbf{v}_{\text{pas}_D^{\mathcal{G}(Y(a))}}) = v_i \mid V_{D \cap \text{pre}_i}(a, \mathbf{v}_{\text{pas}_D^{\mathcal{G}(Y(a))}}) = \mathbf{v}_{D \cap \text{pre}_i}) \quad (41.21) \\
&= \prod_{D \in \mathcal{D}(\mathcal{G}(Y(a)))} p(V_D(a, \mathbf{v}_{\text{pas}_D^{\mathcal{G}(Y(a))}}) = \mathbf{v}_D).
\end{aligned}$$

Here the first two lines follow by the chain rule of probability, term grouping, and the fact that in any ADMG, including a SWIG ADMG, the set of districts partitions the set of random vertices. The third equality follows because of the following:

$$\begin{aligned}
p(V_i(a) = v_i \mid V_{Y \cap \text{pre}_i}(a) = \mathbf{v}_{Y \cap \text{pre}_i}) &= p(V_i(a) = v_i \mid V_{\text{mb}_i^{\mathcal{G}(Y(a)) \cap \text{pre}_i}(a)} = \mathbf{v}_{\text{mb}_i^{\mathcal{G}(Y(a)) \cap \text{pre}_i}}). \quad (41.22) \\
&= p(V_i(a) = v_i \mid V_{D \cap \text{pre}_i}(a) = \mathbf{v}_{D \cap \text{pre}_i}, \\
&\quad V_{\text{pas}_D^{\mathcal{G}(Y(a)) \cap \text{pre}_i}(a)} = \mathbf{v}_{\text{pas}_D^{\mathcal{G}(Y(a)) \cap \text{pre}_i}}) \\
&= p(V_i(a, b_i) \mid V_{D \cap \text{pre}_i}(a, b_i) = \mathbf{v}_{D \cap \text{pre}_i}) \\
&= p(V_i(a, \mathbf{v}_{\text{pas}_D^{\mathcal{G}(Y(a))}}) \mid V_{D \cap \text{pre}_i}(a, \mathbf{v}_{\text{pas}_D^{\mathcal{G}(Y(a))}}) = \mathbf{v}_{D \cap \text{pre}_i}) \quad (41.23)
\end{aligned}$$

where $b_i = \mathbf{v}_{\text{pas}_D^{\mathcal{G}(Y(a)) \cap \text{pre}_i}}$, and $D = \text{dis}_i^{\mathcal{G}(Y(a))}$. Here the first equality follows from Rule 1;²⁷ the second follows from the definition of the Markov blanket of $V_i(a)$ in $\mathcal{G}(Y(a))$; the third follows from Rule 2 since $V_i(a, b_i)$ is m-separated from $B_i(a, b_i) \equiv$

27. Note that the Markov blanket of i in the subgraph of $\mathcal{G}(Y(a))$ restricted to predecessors of i is, in general, a strict subset of the predecessors of i in the Markov blanket of i in $\mathcal{G}(Y(a))$. Consequently, the conditioning set in the terms of (41.22) may not be minimal.

$V_{\text{pas}_D^{\mathcal{G}(Y(a))} \cap \text{pre}_i}(a, b_i)$ in $\mathcal{G}(a, b_i)$; the fourth is an application of Rule 3 since vertices in $V_{\text{pas}_D^{\mathcal{G}(Y(a))} \setminus \text{pre}_i}$ are ordered after V_i and hence are not ancestors of V_i in \mathcal{G} , and thus also in $\mathcal{G}(a, b_i)$.

Next, we consider whether each term of the form $p(V_D(a, v_{\text{pas}_D^{\mathcal{G}(Y(a))}}))$ is identified from $p(V)$ by inductively applying the identified splitting operation in Proposition 41.5 to every element V_j in $A \cup (V \setminus D)$ in a sequence such that the precondition of Proposition 41.5 is satisfied at every step, and marginalizing $V_j(a)$ at every step whenever $V_j \notin D$. (Hence V_j will be split unless $V_j \in D \setminus A$.) $p(Y(a))$ is identified if for every district $D \in \mathcal{D}(\mathcal{G}(Y(a)))$, the corresponding term $p(V_D(a, v_{\text{pas}_D^{\mathcal{G}(Y(a))}}))$ is identified in this way. In fact, the above method of identification is sufficient and necessary for identification of $p(Y(a))$. See the Appendix for details.

For the special case where $Y \cap A = \emptyset$, the resulting identified functionals were first described as an algorithm in Tian and Pearl [2002], and proven to be complete in Huang and Valtorta [2006] and Shpitser and Pearl [2006a]. In both versions of the algorithm, the identifiable terms corresponding to districts $D(a)$ in $\mathcal{G}(Y(a))$ form parts of the *nested Markov factorization* of an ADMG, and the algorithm may thus be viewed as giving a modified nested factorization of an ADMG, just as the extended g-formula is a modified DAG factorization. For more details, see Richardson et al. [2017].

41.4.4 Identification of Conditional Interventional Distributions

Targets of inference in causal inference are often functions of *conditional* counterfactual distributions $p(Y(a) | Z(a))$ rather than marginal distributions $p(Y(a))$. Such targets arise, for instance, when effects within certain subgroups are of interest, or when investigating relationships between primary and secondary outcomes. A straightforward modification of the above algorithm yields identification in such settings.

Fix Y, Z, A where Y, Z are disjoint, but may both intersect A . Fix the largest subset $W \subseteq Z$, with $Z' = Z \setminus W$, such that $Z'(a, z')$ is m-separated from $Y(a, z')$ given $W(a, z')$ in $\mathcal{G}(V(a, z'))$. Then, by Rule 2, $p(Y(a) | W(a), Z'(a) = z') = p(Y(a, z') | W(a, z'))$. Next, let A' be a maximal subset of $Z \cap A$ such that $A'(a, z') \perp\!\!\!\perp Y(a, z') | \{W(a, z') : W \in Z \setminus (Z' \cup A')\}$. Then $p(Y(a) | Z(a))$ is identified if

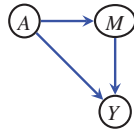


Figure 41.4 A simple DAG containing a treatment A , an intermediate M , and a response Y .

$p(Y(a, z'), \{W(a, z'): W \in Z \setminus (Z' \cup A')\})$ is identified. In fact, we have:

$$p(Y(a) | Z(a)) = \frac{p(Y(a, z'), \{W(a, z'): W \in Z \setminus (Z' \cup A')\})}{p(\{W(a, z'): W \in Z \setminus (Z' \cup A')\})} \Big|_{Z_{A'} = z_{A'}}.$$

As we show in the [Appendix](#), this condition is also necessary since if $p(Y(a, z'), \{W(a, z'): W \in Z \setminus (Z' \cup A')\})$ is not identified, $p(Y(a)|Z(a))$ is also not identified.

41.4.5 Representing Context-specific Independence using SWIGs

We now discuss an extension of SWIGs due to [Dahabreh et al. \[2019\]](#) and [Sarvet et al. \[2020\]](#) who demonstrate that SWIGs have greater expressive power than standard causal DAGs because of their ability to represent interventional context-specific conditional independence.

Consider the causal DAG shown in Figure 41.5(a) where A , M , and Y are observed and U , R , and S are unobserved. The latent projection is given in Figure 41.5(a*). The SWIG resulting from a joint intervention setting A to a and M to m is shown in Figure 41.5(b); the latent projection of this SWIG is shown in Figure 41.5(b*). The distribution of $Y(a, m)$ is not identified owing to the presence of the edges $M \rightarrow Y \leftrightarrow M$ (also called a bow arc).

However, suppose that additional context-specific subject matter knowledge²⁸ implies that the following counterfactual independences hold:

$$U \perp\!\!\!\perp R(a = 0, m); \quad U \perp\!\!\!\perp M(a = 1).$$

As a consequence, the edges $U \rightarrow R(0, m)$ in $\mathcal{G}(0, m)$ and $U \rightarrow M(1)$ in $\mathcal{G}(1, m)$ may be removed, leading to the SWIGs shown in Figure 41.5(c) and (d), with corresponding latent projections shown in Figure 41.5(c*) and (d*).

Applying d-separation to the latent projections in Figure 41.5(c*) and (d*), we see that²⁹

$$Y(a, m) \perp\!\!\!\perp M(a), A \quad \text{for } a = 0, 1. \quad (41.24)$$

Consequently,

$$P(Y | A = a, M = m) = P(Y(a, m)), \quad (41.25)$$

so that the joint effect of A and M on Y is identified for both $a = 0$ and $a = 1$.

28. See the ivermectin study described in the companion paper [Robins et al. \[2021\]](#), Chapter 38 in this volume.

29. Recall that when testing d-separation in SWIGs, fixed nodes such as $a = 0$ in Figure 41.5(c*) and $a = 1$ in Figure 41.5(d*) always block paths on which they occur as non-endpoint vertices.

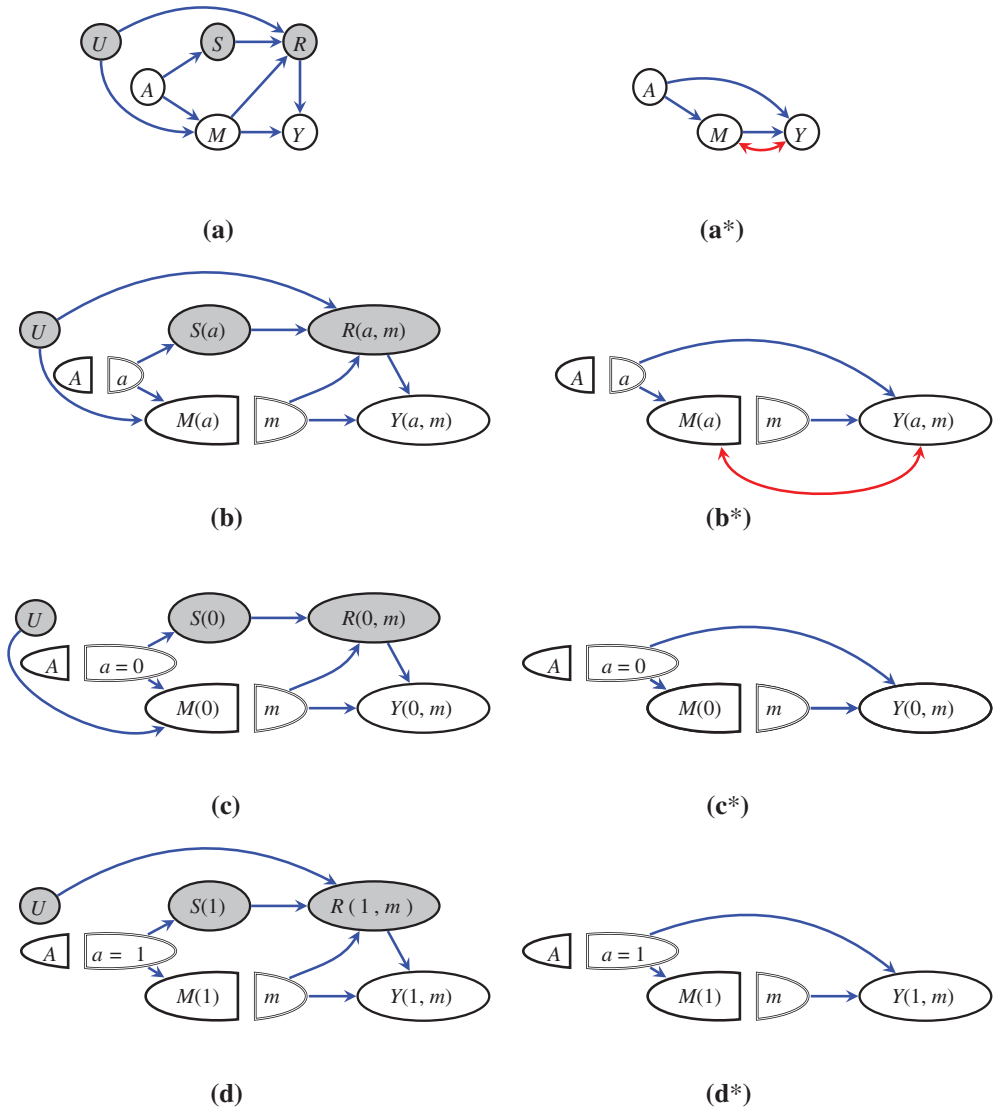


Figure 41.5 (a) A DAG \mathcal{G} representing two studies of river blindness, described in Section 38.2.1 in Chapter 38 in this volume. (b) The SWIG $\mathcal{G}(a, m)$ resulting from \mathcal{G} ; (c) and (d) show SWIGs $\mathcal{G}(a = 0, m)$ and $\mathcal{G}(a = 1, m)$ that incorporate additional context specific causal information. (a*), (b*), (c*), (d*) show the corresponding latent projections.

Given solely the DAG in Figure 41.5(a), with the latent projection in Figure 41.5(a*), the equality (41.25) would not be expected since it does not follow from existing methods such as the *do*-calculus, the ID algorithm, or the back-door criterion [Pearl 2009], though see the recent work by Tikka et al. [2019]. However,

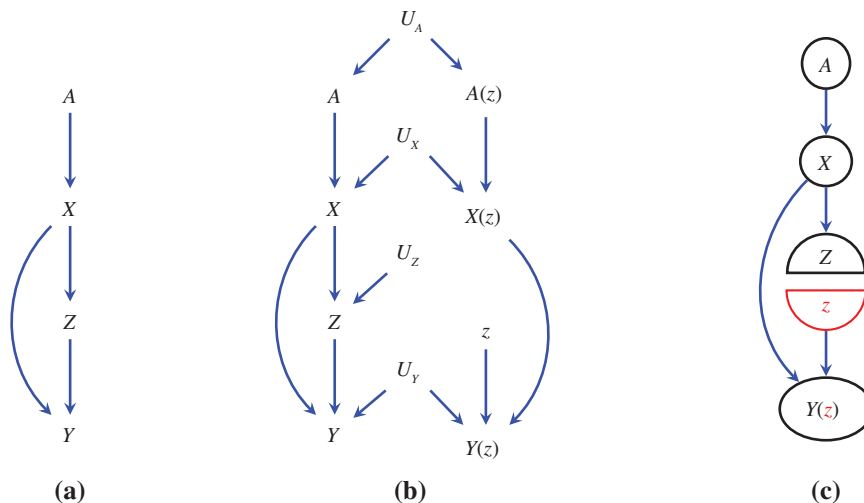


Figure 41.6 (a) A DAG \mathcal{G} . (b) The twin network arising from intervening to set Z to z . (c) The SWIG $\mathcal{G}(z)$.

Equation (41.25) has a structural explanation in terms of the SWIGs corresponding to different treatment values of A .

In particular, the context-specific SWIG independences $U \perp\!\!\!\perp R(0, m) | A, M(0)$ and $U \perp\!\!\!\perp M(1) | A$, coupled with consistency, imply, respectively, the context-specific independences $U \perp\!\!\!\perp R | A = 0$ and $U \perp\!\!\!\perp M | A = 1$ on the factual distribution. These independences cannot be read off from the standard causal DAG shown in Figure 41.5(a). This is because the absence of the $U \rightarrow M$ edge when A is set to 1 and the $U \rightarrow Y$ path when A is set to 0 are not represented in this DAG.

Since, in addition to Equation (41.24), we also have $M(a) \perp\!\!\!\perp A$ for $a = 0, 1$, it follows that the distribution of the counterfactuals $\{A, M(a), Y(a, m)\}$ for all a, m obeys the FFRCISTG model associated with the graph shown in Figure 41.4 in which there are no bidirected edges. However, interestingly, the distribution of these counterfactuals does not obey the NPSEM-IE associated with Figure 41.4, though the distribution does obey the NPSEM-IE (hence also the FFRCISTG) associated with Figure 41.5(a^{*}).³⁰

41.5 Conclusion

We wholeheartedly applaud Judea Pearl for his development and advocacy of graphical approaches to causal modeling. His approach represents a fundamental advance leading to many new insights and methods, including complete

30. See Section 38.2.6 in chapter 38 of this volume.

identification theory for causal queries of all types, and extensions of d-separation to complex questions in causal modeling and missing data.

Acknowledgments

The first author was supported by the grants ONR N00014-18-1-2760, NSF CAREER 1942239, NSF 1939675, and R01 AI127271-01A1. The second author was supported by the grant ONR N00014-19-1-2446. The third author was supported by the grants ONR N00014-19-1-2446 and National Institutes of Health (NIH) awards R01 AG057869, R01 AI127271, R37 AI102634, U01 CA261277-01, and R01 CA222147-02. The authors would like to thank F. Richard Guo for his insightful comments that improved this manuscript.

41.A

41.A.1

Appendix

Incompleteness of d-Separation in Twin Networks due to Deterministic Relations

Twin networks [Balke and Pearl 1994] are an alternative way to combine graphs and counterfactuals that allow some of the counterfactual independence relations implied by the NPSEM-IE to be read off via d-separation; see also Shpitser and Pearl [2008] and Pearl [2009, Section 7.1.4]. However, d-separation is not complete for twin networks [Richardson and Robins 2013] since, as a consequence of consistency, certain variables in a twin network may be deterministically related. Consequently, it is possible for there to be a d-connecting path in a twin network and yet the corresponding conditional independence holds for all distributions in the model.

To see a simple example, consider the DAG shown in Figure 41.6(a), with the twin network and SWIG associated with intervening to set Z to z , shown in Figure 41.6(b) and (c), respectively. Note that A and $Y(z)$ are d-connected given X in the twin network by two different d-connecting paths.³¹ However, in spite of this $A \perp\!\!\!\perp Y(z) | X$ under the associated NPSEM-IE because $X(z) = X$, and A and $Y(z)$ are d-separated given $X(z)$ in the twin network. The SWIG $\mathcal{G}(z)$ shown in Figure 41.6(c) makes manifest that A is d-separated from $Y(z)$ given X , hence $A \perp\!\!\!\perp Y(z) | X$ under the FFRCISTG, hence also under the NPSEM-IE.

In addition, it may also be inferred from the SWIG that $A(z) \perp\!\!\!\perp Y(z) | X$, $A \perp\!\!\!\perp Y(z) | X(z)$, and $A(z) \perp\!\!\!\perp Y(z) | X(z)$ hold under the FFRCISTG (and hence also the NPSEM-IE). This is because it follows from causal irrelevance that, given a SWIG

31. Precisely: $A \leftarrow U_A \rightarrow A(z) \rightarrow X(z) \rightarrow Y(z)$ and $A \rightarrow X \leftarrow U_X \rightarrow X(z) \rightarrow Y(z)$.

$\mathcal{G}(a)$, if a label a_i is present on some random node (equivalently if the SWIG contains a fixed node a_i), then a_i may always be added to the label of any random node on which it is not already present. Consequently, we are free to add z to the label for X and A in $\mathcal{G}(z)$, from which these independences follow. Note that in the twin network, although A and $A(z)$ are d-separated from $Y(z)$ given $X(z)$, the path $A(z) \rightarrow X(z) \rightarrow Y(z)$ d-connects $A(z)$ and $Y(z)$ given X , hence we cannot read off $A(z) \perp\!\!\!\perp Y(z)|X$ from the twin network.

Shpitser and Pearl [2008] provide an algorithm for merging nodes in a twin network, under a particular instantiation of the variables. This algorithm is conjectured to be complete for checking equality of the probability of counterfactual events. A conditional independence statement corresponds to a (potentially exponential) set of equalities between probabilities of events. Thus, if the conjecture holds, then the algorithm of Shpitser and Pearl [2008] provides a way to check counterfactual conditional independence implied by an NPSEM-IE. Though this approach is more involved, as noted earlier in Footnote 6, it addresses a harder problem than SWIGs since it is determining all independencies implied by an NPSEM-IE model that also includes “cross-world” independencies.

41.A.2 Weaker Causal Models to Which the po-Calculus Also Applies

In Section 41.3, we chose to express Rule 3 of the po-calculus on the distribution level: $p(Y(x, z)) = p(Y(x))$, although the equality holds on the individual level: $Y(x, z) = Y(x)$ under the FFRCISTG, see also Footnotes 18 and 23. We chose to do so for several reasons. First, this form is closer in spirit to Pearl’s original formulation of the *do*-calculus.

Second, the weaker equality is expressible in the language of interventions, say via the *do* operator: $p(Y|\text{do}(x, z)) = p(Y|\text{do}(x))$. This allows us to apply this rule and other rules of po-calculus to causal models that are not counterfactual, but which allow discussion of interventional distributions, such as the *agnostic causal model* of Spirtes et al. 2001, which is *defined* by the relationship between the observed data distribution and interventional distributions given by the extended g-formula [Equation (41.16)] re-expressed via the *do* operator. Indeed, the FFRCISTG and the NPSEM-IE imply all distribution-level interventional statements that hold under the agnostic causal model, and these are the only statements that are relevant for the purposes of identification of interventional quantities expressible by the *do* operator. Note that the distribution-level equality has a graphical representation via *population SWIGs* in which missing edges correspond to the absence of population-level direct effects, whereas the individual-level counterfactuals are not necessarily the same. See also Section 7 of Richardson and Robins [2013].

41.A.3 Completeness Proofs

Here we describe a number of completeness results referred to in the main body of the chapter. Before doing so, we state necessary preliminaries. Given an acyclic directed mixed graph (ADMG) $\mathcal{G}(V)$ and a set $S \subseteq V$, an induced subgraph $\mathcal{G}(V)_S$ is defined to be a graph containing vertices S , and all edges in $\mathcal{G}(V)$ between elements in S .

Given an ADMG $\mathcal{G}(V)$, we define a set $W \subseteq V$ to be fixable if $W = \emptyset$, or $W = \{W_1, W_2, \dots\}$ and there exists a set of ADMGs $\mathcal{G}_0(V)$, $\mathcal{G}_1(V \setminus \{W_1\})$, $\mathcal{G}_2(V \setminus \{W_1, W_2\})$, \dots , $\mathcal{G}_k(V \setminus W)$, such that

- $\mathcal{G}_0(V) = \mathcal{G}(V)$.
- For every $i = 0, \dots, k-1$, W_{i+1} has no element $V_j \in V \setminus \{W_1, \dots, W_i, W_{i+1}\}$ with a directed path from W_{i+1} to V_j and a path consisting exclusively of bidirected edges from W_{i+1} to V_j in \mathcal{G}_i .
- For every $i = 1, \dots, k$, $\mathcal{G}_i(V \setminus \{W_1, \dots, W_i\})$ is obtained from $\mathcal{G}_{i-1}(V \setminus \{W_1, \dots, W_{i-1}\})$ by removing W_i and all edges adjacent to W_i .

If $W \subseteq V$ is fixable, the set $S \equiv V \setminus W$ is said to be *reachable*. A set S reachable in $\mathcal{G}(V)$ is said to be *intrinsic* if the vertices in $\mathcal{G}(V)_S$ form a bidirected connected set. Note the relationship between reachable sets and the precondition for Proposition 41.5. We have the following result.

Theorem 41.1 Fix possibly intersecting sets Y, A such that $Y(a)$ is ancestral in the SWIG $\mathcal{G}(V(a))$. Then

$$p(Y(a) = v_Y) = \prod_{D \in \mathcal{D}(\mathcal{G}(Y(a)))} p\left(V_D(a, v_{\text{pas}_D^{\mathcal{G}(Y(a))}}) = v_D\right),$$

and $p(Y(a))$ is not identified if there exists $D \in \mathcal{D}(\mathcal{G}(Y(a)))$ such that no inductive sequence of applications of Proposition 41.5 exists where every element $V_j \in A \cup (V \setminus D)$ is split such that the precondition of Proposition 41.5 is satisfied at every step, and $V_j(a)$ is marginalized from the resulting SWIG whenever $V_j \notin D$.

Proof. Assume such a set D exists. Assume D is not a reachable set in $\mathcal{G}(V)$. Then the results in Richardson et al. [2017] imply that there exists a hedge for $p(Y(a))$ and that $p(Y(a))$ is not identified [Shpitser and Pearl 2006a].

Assume D is a reachable set, but some element $A_i \in D$ cannot be split by applying Proposition 41.5. This implies there exists a set of vertices W_1, \dots, W_k in D that are bidirected connected, and W_k is a child of A_i in $\mathcal{G}(V)$. Since W_1, \dots, W_k , being elements of D , are in the set of ancestors of Y in $\mathcal{G}(V(a))$, the sets $\{A_i\}$, and $\{A_i, W_1, \dots, W_k\}$ form a hedge for $p(Y(a))$, so $p(Y(a))$ is not identifiable. ■

Theorem 41.2 Fix subsets Y, Z, A of V , in some ADMG $\mathcal{G}(V)$, where Y, Z are disjoint, but may both intersect A . Fix the largest subset $W \subseteq Z$, with $Z' = Z \setminus W$, such that $Z'(a, z')$ is m-separated from $Y(a, z')$ given $W(a, z')$ in $\mathcal{G}(V(a, z'))$, and let A' be a maximal subset of $Z \cap A$ such that $A'(a, z')$ is m-separated from $Y(a, z')$ given $\{W(a, z'): W \in Z \setminus (Z' \cup A')\}$. Then $p(Y(a)|Z(a))$ is identified if $p(Y(a, z'), \{W(a, z'): W \in Z \setminus (Z' \cup A')\})$ is identified. If identification holds, we have:

$$p(Y(a)|Z(a)) = \frac{p(Y(a, z'), \{W(a, z'): W \in Z \setminus (Z' \cup A')\})}{p(\{W(a, z'): W \in Z \setminus (Z' \cup A')\})} \Big|_{Z_{A'}=z_{A'}}.$$

Proof. If the stated assumptions hold, and $p(Y(a, z'), \{W(a, z'): W \in Z \setminus (Z' \cup A')\})$ is identified, the conclusion follows by definition of conditioning.

Assume $p(Y(a, z'), \{W(a, z'): W \in Z \setminus (Z' \cup A')\})$ is not identified. It suffices to consider the case where $p(\{W(a, z'): W \in Z \setminus (Z' \cup A')\})$ is not identified. The proof then follows the proof structure for analogous results in [Shpitser and Pearl \[2006b\]](#) and [Malinsky et al. \[2019\]](#), with the fact that $A \cap Y$ is potentially not an empty set not influencing the structure of the proof.

Non-identification of $p(\{W(a, z'): W \in Z \setminus (Z' \cup A')\})$ implies the existence of a hedge, and the preconditions of the theorem imply the existence of an m-connecting (given W) path from an element in W in the hedge to some element in Y . Non-identification is established by induction on the structure of this path. Specifically, fix an element L on the path such that the inductive hypothesis that $p(L(a, z')|W'(a, z'))$ is not identified holds, where W' is the subset of W involved in the hedge, or in the m-connecting path from the hedge to L . Thus, there exist two elements of the causal model that disagree on this distribution, but agree on the observed data distribution. The induction then establishes that the distribution $p(L'(a, z')|W''(a, z'))$, where L' is the next element on the m-connecting path, and W'' are all elements of W that are either in the hedge, or witness m-connection of the path from the hedge to L' , is also not identified. This is established by extending the existing two elements with an appropriate distribution that yields a one-to-one mapping from distributions $p(L(a, z')|W'(a, z'))$ to distributions $p(L'(a, z')|W''(a, z'))$. ■

References

- A. Balke and J. Pearl. 1994. Probabilistic evaluation of counterfactual queries. In *Proceedings of the Twelfth Conference on Artificial Intelligence (AAAI-94)*. Morgan Kaufmann, San Francisco. 230–237.
- I. J. Dahabreh, J. M. Robins, S. J. Haneuse, and M. A. Hernán. 2019. Generalizing causal inferences from randomized trials: Counterfactual and graphical identification. *arXiv preprint arXiv:1906.10792*.

- R. J. Evans. 2018. Margins of discrete Bayesian networks. In *Annals of Statistics*, 46(6A), 2623–2656.
- F. Fisher. 1969. Causation and specification in economic theory and econometrics. *Synthese* 20, 489–500. In *Econometrics: Essays in Theory and Applications: Collected Papers*. MIT Press, 1992.
- F. M. Fisher. 1970. A correspondence principle for simultaneous equation models. *Econometrica* 38, 73–92. DOI: <https://doi.org/10.2307/1909242>.
- P. Forré and J. M. Mooij. 2019. Causal calculus in the presence of cycles, latent confounders and selection bias. In *Proceedings of the 35th Annual Conference on Uncertainty in Artificial Intelligence (UAI-19)*.
- D. Galles and J. Pearl. 1998. An axiomatic characterization of causal counterfactuals. *Found. Sci.* 3, 151–182. DOI: <https://doi.org/10.1023/A:1009602825894>.
- T. Haavelmo. 1943. The statistical implications of a system of simultaneous equations. *Econometrica* 11, 1–12. DOI: <https://doi.org/10.2307/1905714>.
- J. Halpern and J. Pearl. 2001. Causes and explanations: A structural-model approach. Part I: Causes. *Proceedings of UAI-01*. 411–420.
- J. Halpern and J. Pearl. 2005. Causes and explanations: A structural-model approach. Part II: Explanations. *B. J. Philos. Sci.* 56, 889–911.
- Y. Huang and M. Valtorta. 2006. Pearl’s calculus of interventions is complete. In *Twenty Second Conference on Uncertainty in Artificial Intelligence (UAI-06)*.
- G. W. Imbens. 2014. Instrumental variables: An econometrician’s perspective. *Stat. Sci.* 29, 3, 323–358. DOI: <https://doi.org/10.1214/14-STS480>.
- S. Lee, J. D. Correa, and E. Bareinboim. 2020. Generalized transportability: Synthesis of experiments from heterogeneous domains. In *Proceedings of the Thirty Fourth AAAI Conference on Association for the Advancement of Artificial Intelligence*.
- D. Malinsky, I. Shpitser, and T. S. Richardson. 2019. A potential outcomes calculus for identifying conditional path-specific effects. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*.
- J. Neyman. 1923. Sur les applications de la thar des probabilités aux expériences agraires: Essay des principe. excerpts reprinted (1990) in English. *Stat. Sci.* 5, 463–472.
- J. Pearl. 1988. *Probabilistic Reasoning in Intelligent Systems*. Morgan and Kaufmann, San Mateo.
- J. Pearl. 1995. Causal diagrams for empirical research. *Biometrika* 82, 4, 669–709. DOI: <https://doi.org/10.2307/2337329>.
- J. Pearl. 2009. *Causality: Models, Reasoning, and Inference* (2nd. ed.). Cambridge University Press. ISBN: 978-0521895606.
- J. Pearl. 2018. Does obesity shorten life? Or is it the soda? On non-manipulable causes. *J. Causal Inference* 6, 1–7. DOI: <https://doi.org/10.1515/jci-2018-2001>.
- J. Pearl. 2019. On the interpretation of do(x). *J. Causal Inference* 7. DOI: <https://doi.org/10.1515/jci-2019-2002>.

- T. S. Richardson. 2003. Markov properties for acyclic directed mixed graphs. *Scand. J. Stat.* 30, 1, 145–157. DOI: <https://doi.org/10.1111/1467-9469.00323>.
- T. S. Richardson and J. M. Robins. 2013. Single world intervention graphs (SWIGs): A unification of the counterfactual and graphical approaches to causality. *preprint*: <http://www.csss.washington.edu/Papers/wp128.pdf>.
- T. S. Richardson, R. J. Evans, J. M. Robins, and I. Shpitser. 2017. Nested Markov properties for acyclic directed mixed graphs. <https://arxiv.org/abs/1701.06686>.
- J. M. Robins. 1986. A new approach to causal inference in mortality studies with sustained exposure periods—Application to control of the healthy worker survivor effect. *Math. Model.* 7, 1393–1512. DOI: [https://doi.org/10.1016/0270-0255\(86\)90088-6](https://doi.org/10.1016/0270-0255(86)90088-6).
- J. M. Robins. 1987. Errata to “A new approach to causal inference in mortality studies with sustained exposure periods—Application to control of the healthy worker survivor effect.” *Comput. Math. App.* 14, 917–921.
- J. M. Robins and T. S. Richardson. 2010. Alternative graphical causal models and the identification of direct effects. In P. Shrouf, K. Katherine, and K. Ornstein (Eds.), *Causality and Psychopathology: Finding the Determinants of Disorders and their Cures*. Oxford University Press.
- J. M. Robins, M. A. Hernan, and U. Siebert. 2004. Effects of multiple interventions. *Comparative Quantification of Health Risks: Global and Regional Burden of Disease Attributable to Selected Major Risk Factors*. 2, 28, 2191–2230.
- J. M. Robins, T. S. Richardson, and I. Shpitser. 2021. An interventionist approach to mediation analysis. Chapter 38 in this Volume.
- D. B. Rubin. 1974. Estimating causal effects of treatments in randomized and non-randomized studies. *J. Educ. Psychol.* 66, 688–701. DOI: <https://doi.org/10.1037/h0037350>.
- A. L. Sarvet, K. N. Wanis, M. J. Stensrud, and M. A. Hernán. 2020. A graphical description of partial exchangeability. *Epidemiology* 31, 3, 365–368.
- I. Shpitser and J. Pearl. 2006a. Identification of joint interventional distributions in recursive semi-Markovian causal models. In *Proceedings of the Twenty-First National Conference on Artificial Intelligence (AAAI-06)*. AAAI Press, Palo Alto.
- I. Shpitser and J. Pearl. 2006b. Identification of conditional interventional distributions. In *Proceedings of the Twenty Second Conference on Uncertainty in Artificial Intelligence (UAI-06)*. AUAI Press, Corvallis, Oregon 437–444.
- I. Shpitser and J. Pearl. 2008. Complete identification methods for the causal hierarchy. *J. Mach. Learn. Res.* 9, Sep, 1941–1979.
- H. A. Simon. 1953. Causal ordering and identifiability. In W. C. Hood and T. C. Koopmans (Eds.), *Studies in Econometric Method*. Wiley.
- P. Spirtes, C. Glymour and R. Scheines. 2001. *Causation, Prediction, and Search*. (2nd. ed.). Springer Verlag, New York. ISBN: 978-0262194402.
- R. H. Strotz and H. O. A. Wold. 1960. Recursive versus non-recursive systems: An attempt at synthesis. *Econometrica* 28, 2, 417–427. DOI: <https://doi.org/10.2307/1907731>.

- J. Tian and J. Pearl. 2002. On the testable implications of causal models with hidden variables. In *Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence (UAI-02)*, Vol. 18. AUAI Press, Corvallis, Oregon., 519–527.
- S. Tikka, A. Hyttinen, and J. Karvanen. 2019. Identifying causal effects via context-specific independence relations. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett (Eds.), *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc, 2800–2810. <http://papers.nips.cc/paper/8547-identifying-causal-effects-via-context-specific-independence-relations.pdf>.
- T. S. Verma and J. Pearl. 1990. *Equivalence and Synthesis of Causal Models*. Technical Report R-150, Department of Computer Science, University of California, Los Angeles.
- S. Wright. 1921. Correlation and causation. *J. Agric. Res.* 20, 557–585.

Causal Bayes Nets as Psychological Theory

Steven A. Sloman (Brown University)

Abstract

Causal Bayes nets (CBNs) have offered deep and lasting insights about how to make causal and counterfactual inferences. This chapter summarizes some of what is known about whether people obey these prescriptions when making inferences. CBNs do model some aspects of human reasoning well, but fall short in certain respects: Physical causal reasoning takes advantage of mental simulations not captured by CBNs. There is evidence that people are sensitive to a causal Markov condition, but it requires assuming that people make up their own causal structure to some extent. People do tend to explain away as the CBN formalism prescribes, but often insufficiently. Much of human reasoning appears qualitatively but not quantitatively similar to CBN reasoning. CBNs capture how people reason about action, but counterfactual inference presents a much bigger problem. Further developments are also necessary to capture the fact that human knowledge is distributed; it resides across a community of knowledge. This raises deep and difficult questions about mental representation.

One of Judea Pearl's great achievements was to win cognitive science's most prestigious award, the Rumelhart Prize, in 2010, for computational contributions to the study of the mind. Pearl has been making a convincing case that people think in terms of causal relations from the time that he began his work on causal models (and before, see [Pearl \[1988\]](#)). His argument is essentially that core assumptions of CBNs¹ are inference schema that people find highly intuitive. The Markov property,

1. I use the term CBN to describe a framework for reasoning about both interventions and counterfactuals. [Pearl \[2000\]](#) distinguishes between the two, reserving the term "CBN" for interventions and "Causal Diagrams" to represent counterfactuals.

for instance, describes an inference that people make naturally, at least when they are reasoning about causal chains. It seems obvious that A and C are rendered independent if the only variables that connect them are held constant. And who could doubt explaining away? Clearly, learning that one explanation for an effect is true makes other explanations less likely for they are unnecessary. Moreover, everybody recognizes the difference between intervention and observation. Observing that the ground is wet suggests different precursors than making the ground wet yourself. And the fact that the basic elements of the causal modeling framework seem so intuitive make their product—causal models—seem equally good descriptors of human causal reasoning.

So what do CBNs really have to do with the mind? Is human knowledge—at least, human causal knowledge—structured like a CBN? I will address this question in three parts. First, are people as sensitive to causality as Pearl suggests? Do people structure events around causal relations? Second, I will offer a brief history of tests of CBNs as psychological theory. Do the core assumptions of CBNs match human intuition? Finally, I will address the question from a broader perspective. Is the psychological evidence consistent with the view that people walk around with CBNs that represent the causal structure of the world in their brains?

42.1 The Human Conception of Causality

Philosophers have offered a variety of interpretations of what it means for A to cause B. One, celebrated by philosophers like [Woodward \[2003\]](#) is that A causes B if and only if a sufficiently strong intervention on A would affect B. From a psychological perspective, it is self-evident that this hypothesis offers a sufficient definition of causality. However, it is not necessary. There are other conceptions of causality that do not depend on intervention. For instance, young infants perceive launching events (like one billiard ball hitting another) causally even though no agent is in any obvious way intervening on the cause [[Bechlivanidis et al. 2019](#)]. More generally, events that involve forces in either a literal or metaphorical sense (like a road sign forcing a driver to turn) are perceived causally in the absence of an intervening agent [[Wolff 2007](#)]. The perception of appropriate temporal relations among events induce a causal interpretation. If A happens and then B does followed by C, people will perceive a causal chain $A \rightarrow B \rightarrow C$ as long as the temporal delays are consistent with whatever causal mechanism is likely to relate the variables [[Buehner and May 2002](#), [Lagnado and Sloman 2006](#)]. So, intervention is a strong cue for causality, but there are others.

If intervention is not necessary for people to conceive of a causal link, is correlation sufficient? On the one hand, there is no question that people can misinterpret correlations as causal. It is true that coffee consumption is correlated with reduced

mortality [Loftfield et al. 2018], but nobody has shown a causal relation. Nevertheless, the causal interpretation is compelling because of the ease of imagining a mechanism relating cause and effect. People are willing to assume causation when they think they can generate a story about why the cause would lead to the effect [Heider and Simmel 1944]. On the other hand, people seem to be largely unable to infer causal structure from merely correlational data (without helpful temporal or spatial information or prior knowledge). Simply showing people mounds of data displaying correlations and conditional correlations does not induce an impression of causality. The converse is sufficient: If the impression of causality is already there, then people will see correlations that do not exist [Chapman and Chapman 1969].

Human conceptions of causality are guided by beliefs about operating mechanisms. Causality is generally understood, not in terms of correlation, but in terms of process. The perception of physical causality requires some quantity (like energy) that travels continuously from cause to effect through space and time [Dowe 2000]. People are not satisfied by explanations that appeal to correlations, they want knowledge about specific mechanisms [Ahn and Bailenson 1996]. People will only describe an event as causal if they can imagine a quantity being transferred from cause to effect; correlations between cause and effect are not enough [Wolff 2007]. When attributing cause, people insist that some entity travel along a spatiotemporal pathway. The fact that the entity makes a difference to the outcome—that the outcome would have been different if not for the putative cause—is not good enough [Walsh and Sloman 2011]. Causal inferences about the physical world appeal to mechanisms, not to knowledge that entities are associated.

This is not true when putative causes are intentional [Lombrozo 2010]. When an agent achieves an outcome because they desire the outcome and are able to make it happen, then the specific mechanism is less important. If an outcome is desired by an intelligent agent, the agent can generally achieve it using an alternative mechanism if necessary. If the army can't get the territory through negotiation, they can resort to force. Because they are intentional agents, they will create an alternative causal pathway if necessary, and thus, the event seems causal even if one doesn't know the specific process that led to the outcome. Therefore, intentional causation does not require a mechanistic process in the same way that physical causation does; an appeal to the counterfactual "if the agent had not acted, then the outcome would have been different" is generally sufficient.

In sum, causal Bayes nets (CBNs) are rich ways to represent causal structure. Pearl and many others have demonstrated their value in machine learning, suggesting that it is often useful to reduce causality to probability relations with an

intervention operator. But that is apparently not how people think about causality. In the physical domain, people want to know the process that unfolds continuously over space and time that leads from cause to effect. In the intentional domain, CBNs are closer to being right: People want to know that the appropriate counterfactual is supported, that the effect would not have occurred but for the cause. CBNs might capture how people represent causal structure, but there's more to how people think about causality itself than the CBN framework captures.

42.2 Core Properties

I now briefly summarize some of the key evidence regarding people's sensitivity to three of the core properties of CBNs.

Core property 1: The Markov property. A number of studies have examined whether people's judgments satisfy the Markov property associated with causal graphical models for three-variable common cause structures. This condition requires that a variable be statistically independent of its non-descendants conditional on the state of its immediate parents. For common cause situations, this means that one effect should be independent of the other effect given that the (common) cause has been conditioned on, whatever its value. The common finding is that people violate this condition, instead treating the two effects as conditionally correlated such that movement in one variable leads to corresponding movement in judgments of the other variable (see [Rottman and Hastie \[2014\]](#) and [Hagmayer \[2016\]](#) for reviews).

In the many experiments reviewed, participants are presented with or infer a causal structure from supporting data. [Park and Sloman \[2013, 2014\]](#) argue that people don't rest with the causal structures they are given, rather they import more structure based on prior knowledge. The structure they import depends on the nature of the mechanisms represented by the causal model. For instance, when relations are probabilistic, people will explain a cause's intermittent failure to produce an effect by appealing to a disabling condition. Thus, they import disabling conditions that they were not told about. In the case of a common cause structure, there are two relevant mechanisms, one for each cause. If those mechanisms are different, then participants will import two disablers, one to explain the probabilistic functioning of each mechanism. Consider a model with smoking as a common cause of both impairment of lung function and of an additional financial burden on the family budget. Smoking can have both of these effects, but it produces them in very different ways via different mechanisms. Disablers of each mechanism are likely to be different (e.g., smoking a lower tar brand vs. buying a cheaper brand). But if the mechanisms are the same, then one disabler will do, for it could disable both mechanisms. If the two effects in the model were impairment of lung function

and damage to blood vessels, then the lower tar brand might prevent both. But if only one disabler is introduced, then that also introduces another common cause of both effects, a backdoor path from one to the other. This will introduce a new dependence explaining the violation of the Markov condition. Accordingly, Park and Sloman show that Markov violations only occur when people treat the mechanisms represented by a common cause model as the same, and not when they are different.

The implication is that people do seem to be responsive to the logic of CBNs. However, it's not trivial to determine which CBN an individual is using to make a judgment. People have a habit of elaborating what they are told by importing causal knowledge into their representations.

Core property 2: Explaining away. Another central idea of CBNs is explaining away: When two independent causes have a common effect that they independently contribute to, the causes become dependent conditional on the effect. Specifically, conditional on the values of one cause and the effect, the probability of the other cause should be discounted. It is no longer helpful in explaining the effect; it has been explained away.

People were shown to be responsive to this inference schema many years ago in social psychology (e.g., Kelley [1973]). The focus in the social psychological literature was on how people explain other people's behavior. Do they appeal to the other person's personality or the environment of the behavior? The data show that telling them that one of these causes occurred reduces judgments of the probability of the other. Using a large variety of other scenarios including non-social ones, research on causal reasoning has largely validated this type of inference (see Rottman and Hastie [2016] and Liefgreen et al. [2018] for reviews).

However, although people do engage in explaining away, they generally discount insufficiently and, in a few cases, not at all (reviewed in Rottman and Hastie [2016]). A potential reason for insufficient discounting is that people may sometimes answer the wrong question. In a standard experiment, people are told the probability that each cause produces the target effect in general and then are asked to judge the probability that the cause produced the effect in a particular (token) case, one where the other cause and the effect are known to have occurred. In some cases, people may be judging the cause's general propensity, a value that doesn't change over the course of the experiment, rather than the updated probability for the target event. For instance, if one cause is that the outcome of a flip of a fair coin determines whether one wins or loses, then instead of judging the probability that the coin came up heads on the last trial given that one won, people might judge the propensity of the coin to deliver heads (0.5 as the coin is fair; Liefgreen et al. [2018]).

People may be answering a question about type, rather than the question that was asked, about a token event. People are known to have a tendency to substitute easy questions for hard ones [Kahneman and Frederick 2002].

Core property 3: Seeing versus doing. The hallmark of CBNs, what distinguishes them from run-of-the-mill probabilistic models, is the “do” operator, the means to represent intervention. Intervening on a variable in a CBN both changes the value of that variable and breaks the edges pointing to it; that is, the intervention is determining the variable’s value, so its normal causes are not. According to Pearl [2000], such an operation is a means to represent both action and counterfactual thoughts, interventions on the actual world and on other possible worlds. As Halpern [2016] shows, such interventions are critical for explaining how to attribute cause to outcomes.

Psychological data are not required to conclude that people represent intervention correctly when making inferences about their own actions. Anybody who chokes somebody to death and then concludes the person died from lack of oxygen is understood to be either a liar, a tyrant, or psychotic. Even young children are sensitive to the logic of intervention [Schulz et al. 2007].

The situation is much less clear regarding counterfactual inference. Pearl [2000] proposes a three-step procedure to model counterfactual inference of $Y = y$ given a counterfactual assumption $X = x$ and new evidence e :

Step 1 (abduction): Update one’s causal model to accommodate e .

Step 2 (action): Apply the $\text{do}(X = x)$ operation to construct a causal model that represents the counterfactual world.

Step 3 (prediction): Use the modified model to compute the probability of $Y = y$.

A number of experiments have been run to evaluate the psychological reality of Pearl’s procedure (e.g., Sloman and Lagnado [2005], Rips [2010], Han et al. [2014]). The results are decidedly mixed. Most of the work focuses on counterfactual backtracking, people’s willingness to make a diagnostic inference from a counterfactually assumed effect to its cause. Pearl’s theory clearly disallows such an inference because Step 2 (action) involves breaking the edges to the counterfactual effect, yet people make it on occasion. There are cases when such an inference would be sensible. For instance, a physician might say to medical students “if the symptoms were different, then the disease would have been different” as a way of teaching the relations between diseases and symptoms. Or a counterfactual claim might be best interpreted as a request to treat the statement as diagnostic. If someone says, “if only the trees were green and flowers blooming,” a reasonable interpretation is

that the person wishes winter were over and spring had arrived. In other words, they are intending for the listener to change the values of the causes of their counterfactual statement. The do operation is just not always what counterfactual inference calls for.

Other theories have emerged as competitors to Pearl's [2000]. Hiddleston's [2005] minimal network theory claims that when reasoning counterfactually, the changes introduced to one's representation of the actual world in order to represent the counterfactual world should be minimal in violating as few causal laws that govern the system as possible; the counterfactual model should be kept as similar as possible to the actual model, with the minimal number of edge breaks and the maximum number of intact variables. Rips [2010] offers evidence in support of this theory and against Pearl's. Another theory has been offered by Lucas and Kemp [2012] who propose a double modifiable structural model. They propose that reasoners hold essentially two representations, one with and one without edges into the counterfactually assumed variables. In other words, they define an augmented twin network that includes the original observational or world causal model, plus a copy with intervention implemented. This allows inference from both intervention and observation. A free parameter represents the degree of mutability of the counterfactual model, and helps produce reasonable fits to the published data on counterfactual backtracking.

In sum, CBNs do an excellent job of modeling human reasoning about action. But they are incomplete representations of the subtle and complex world of counterfactual reasoning. In favor of CBNs, not only do competitor models also face challenges, but humans themselves often disagree about which counterfactual world is at issue and about the correct response to a counterfactual question.

42.3 The Broader Perspective: The Community of Knowledge

Large amounts of data from the study of human judgment, reasoning, and decision-making show clearly that causal inference is central to human thought [Sloman 2005, Sloman and Lagnado 2015]. However, data also show that causal inference tends to be based on very limited and coarse knowledge; people's causal representations are remarkably superficial [Rozenblit and Keil 2002, Fernbach et al. 2013]. People are unable to explain how very basic artifacts operate and they are remarkably ignorant about the consequences of social policy [Zaller 1992] and about the causal models underlying common events [Zemla et al. 2017].

So, most causal models of most things are surprisingly superficial. Yet humans have developed science, arts, and technology that are rich, deep, and mind-numbingly complicated. How can relatively ignorant individuals create and survive in such a sophisticated environment? It is because most of the knowledge people

use resides in other people's heads. We live and operate in a community of knowledge [Clark and Chalmers 1998, Sloman and Fernbach 2017]. Causal inference, like all cognition, should be conceived as a collective enterprise, not an individual endeavor.

This collective view of causal inference has ramifications for how we should understand representations of causal knowledge. Pearl may well be right that knowledge should be represented using causal models, but those models need a certain kind of hierarchical structure, not only because causal systems in the world have a hierarchical structure, but also because the only way to get a community of knowledge off the ground is to distribute knowledge in a way that conforms to a hierarchical principle.

What does it mean to “know gardening?” There are aspects of gardening that every independent, functioning member of society knows: It involves soil, water, sunshine, and plants. Everybody even has a causal structure to relate these entities: soil, water, and sunshine together produce plants. To actually engage in gardening requires more causal knowledge though: You have to know about seeds, that they get planted, that they need water to grow, that they become specific plants, and that plants require sunshine. A better gardener will have detailed knowledge about some plants and what they require. A scientific gardener might know exactly how much sun and water each plant requires and what the soil should be composed of, or even that soil isn't strictly necessary (as in the case of hydroponics). In sum, pretty much everyone shares a basic superficial causal model of gardening. This is common ground, and allows broad conversation and humor about gardening. Then there are different levels of expertise that involve having unpacked versions of common ground, more detailed causal models that unpack common ground using more variables and their causal relations to one another.

Experts themselves differ in what they know. Some expert gardeners know more about flowers and others know more about edible crops. And within each group, expertise varies. The expertise of someone who grows potatoes in the Andes is not the same as someone who grows cannabis hydroponically. Then there are experts on the components of gardening: those who know about irrigation, experts on tractors and harvest equipment, masters of pollination, and so on.

A representation of human knowledge requires a theory of how this all fits together. How are causal models nested within one another? The topmost—common ground—is the least articulated. That superficial knowledge is what most people know, other than the fact that there is more to know. People know that others can unpack their superficial knowledge into complete mechanisms. So, each component of common ground is really a placeholder, a gesture to something richer. Those placeholders get cashed out in the minds of experts. And each expert

can only cash out so much, perhaps one or two mechanisms. The experts themselves need placeholders that indicate that their knowledge can also be unpacked (by other experts). Those placeholders represent both the aspects of common ground that they are not expert in, as well as the expertise they have that could be further cashed out by someone else. For instance, a soil expert might know the constituents of a good soil, but they might need a biochemist to explain how those constituents interact to produce growth. Expert knowledge can almost always be unpacked into the more detailed knowledge of someone with deeper (but narrower) expertise.

42.4 Collective Causal Models

This hierarchy of causal models suggests some constraints on how to represent knowledge. The claim is that higher-level knowledge is (i) sparse and (ii) a representation of more detailed knowledge that sits elsewhere. (ii) has at least two interpretations. The first is that the detailed knowledge is a set of lower-order elements that constitute the higher-level object in the sense that individual water molecules taken together comprise a set that constitutes the “water” that, at a higher-level, has causal force (by serving to wash a dirty dog or to slake thirst). Chalupka et al. [2017] propose an algorithm that partitions lower-level data to find causally relevant partitions of both putative causes and effects and then uses the intervention operator to bin resulting partitions to create a higher-level representation of causal structure and test its viability. Chalupka et al. note that, in some cases, the linkage between higher and lower levels is causal, not constitutive. For instance, psychometric models assume latent variables like intelligence that are measured through a set of questions given to test takers. The assumption is that the higher-level entity, intelligence, generates performance on the test, not that the questions constitute intelligence. Chalupka et al.’s algorithm does not directly apply to such cases.

For some applications, identifying a causal structure through the constituents of its variables is useful. Conceptually, however, constituents have causal structures of their own. Water molecules are not static but rather causal entities themselves. In other words, objects at the lower level are not a mere set, but should be represented as causal structures themselves. To model such constituent structure, Casini et al. [2011] propose a recursive Bayesian network (RBN) formalism. In RBNs, variables at higher levels represent Bayes’ nets that reside at a lower level. In other words, higher-level causal structures can be unpacked such that each variable is itself a causal structure at a lower level. Gerharter [2014] proposes a different model, multi-level causal models. These are similar except that, instead of

unpacking high-level variables into causal structures, arrows representing causal dependencies are unpacked. [Casini \[2016\]](#) argues in favor of unpacking variables rather than arrows.

These kinds of models offer a first pass for representing a community of knowledge. The idea would be that common ground is represented at the highest level, and that knowledge would serve as a pointer to progressively more detailed unpackings at levels underneath. Those more detailed unpackings would sit in the heads of experts. Each expert might be responsible for a single or a small number of such lower-level representations. Experts might have different models of the causal structure of the situation. In that case, how one makes use of expert opinion depends on whether the experts' causal models are compatible or not [[Alrajeh et al. 2018](#)].

To use RBNs to model a community of knowledge, a number of independence assumptions would be required in the lower-level knowledge. If high-level causal knowledge carries any value, then the sets corresponding to lower-level knowledge must be (relatively) independent of one another conditional on the high-level variables. For instance, if I believe that water slakes thirst, then the set of water molecules must be independent of the set of biological entities corresponding to thirst conditional on the higher-level representation “water” and “thirst.” If they are not, then the higher-level structure cannot offer predictions and explanations on its own and becomes redundant because lower-level considerations would always be necessary.

These independence assumptions are especially important if we are using RBNs to represent a community of knowledge. In a community of knowledge, different causal structures representing the lower-level knowledge of different higher-level variables sit in the heads of different experts. Such structures are thus useless if they are not in some sense independent of one another; too much dependence would prevent the individual using them to make inferences. Experts presumably carry useful knowledge that does not depend on the knowledge carried by a different expert.

Such a representation leaves many open questions. [Glymour \[2007\]](#) offers an insightful discussion of some of the philosophical and scientific issues that arise when attempting to aggregate variables into a causal structure. One concerns the nature of the links between levels. [Casini \[2016\]](#) argues that the links are constitutive and not causal. This is important as it allows interventions on variables at the lower level to affect variables at the higher level. If the links were causal and descending from higher to lower levels, then [Pearl \[2000\]](#) would require that an intervention at the lower level would render lower-level variables independent of everything at the higher level. But if we assume that intervention does not break

constitutive edges—only causal ones—then a lower-level intervention could still influence effects at the higher level.

But what about Chalupka et al.'s [2017] observation that sometimes links are causal, not constitutive, as in the case of intelligence? Then a lower-level intervention should break edges and render the variable independent of higher-level effects. Actually, this does not appear to be a problem, at least for the example. Intervening on an intelligence test by having an outside agent answer a question does not increase intelligence and would not affect whatever effects intelligence is supposed to have. Indeed, the example suggests that constitutive relations between higher and lower levels, and not causal ones, are what is needed to distinguish the knowledge of different individuals within a community of knowledge. An expert on intelligence would not be distinguished from a non-expert merely by their ability to enumerate the questions on an intelligence test. They would presumably know something about the constituents of intelligence that govern the choice of questions. And an intervention on those constituents would indeed affect intelligence.

42.5 Conclusion

CBNs have been invaluable in the study of how people make causal inferences. However, CBNs are not the whole story. People reason by mentally simulating how a mechanism unfolds over time and space. CBNs don't capture this dynamic process. Nevertheless, many of CBNs' inference schema do seem available to mere mortals. Once we take people's tendency to make up their own causal structure, there is evidence that people are sensitive to a Markov condition. People do tend to explain away as the CBN formalism prescribes, but often insufficiently. In both cases, much of human reasoning appears qualitatively similar to CBN prescription, but not quantitatively similar, and certainly not probabilistically coherent. While the notion of intervention is critical for capturing reasoning about action, we are a long way from understanding counterfactual inference. How people agree on which world they're talking about—when they do—remains an open question.

Other open questions concern how to represent a community of knowledge. Recursive structures with constituent relations provide a starting point, but serious modeling is yet to be done. Another big open question concerns the justification for belief. If people know so little as individuals, why are they so sure about so many things? It is not merely hubris. In many cases, people should have confidence even without knowledge. If people don't take a firm stand on issues like war and climate change, catastrophe can result. Our knowledge representations are responsible for maintaining accurate knowledge, even if the details are distributed throughout a community.

Acknowledgments

This work was funded by a grant from the program on Humility and Conviction in Public Life at the University of Connecticut and The John Templeton Foundation. I thank Sabina Sloman and Semir Tatlidil for comments.

References

- W. Ahn and J. Bailenson. 1996. Mechanism-based explanations of causal attribution: An explanation of conjunction and discounting effects. *Cogn. Psychol.* 31, 82–123. DOI: <https://doi.org/10.1006/cogp.1996.0013>.
- D. Alrajeh, H. Chockler, and J. Y. Halpern. 2018, April. Combining experts' causal judgments. In *Thirty-Second AAAI Conference on Artificial Intelligence*.
- C. Bechlivanidis, A. Schlottmann, and D. A. Lagnado. 2019. Causation without realism. *J. Exp. Psychol. Gen.* 148, 5, 785–804. DOI: <https://doi.org/10.1037/xge0000602>.
- M. J. Buehner and J. May. 2002. Knowledge mediates the timeframe of covariation assessment in human causal induction. *Think. Reason.* 8, 4, 269–295. DOI: <https://doi.org/10.1080/13546780244000060>.
- L. Casini. 2016. How to model mechanistic hierarchies. *Philos. Sci.* 83, 5, 946–958.
- L. Casini, P. M. Illari, F. Russo, and J. Williamson. 2011. Models for prediction, explanation and control: Recursive Bayesian networks. *THEORIA. Revista de Teoría, Historia y Fundamentos de la Ciencia* 26, 1, 5–33.
- K. Chalupka, F. Eberhardt, and P. Perona. 2017. Causal feature learning: An overview. *Behaviormetrika* 44, 1, 137–164. DOI: <https://doi.org/10.1007/s41237-016-0008-2>.
- L. J. Chapman and J. P. Chapman. 1969. Illusory correlation as an obstacle to the use of valid psychodiagnostic signs. *J. Abnorm. Psychol.* 74, 271–280. DOI: <https://doi.org/10.1037/h0027592>.
- A. Clark and D. Chalmers. 1998. The extended mind. *Analysis* 58, 1, 7–19. DOI: <https://www.jstor.org/stable/3328150>.
- P. Dowe. 2000. *Physical Causation*. Cambridge University Press. <https://www.jstor.org/stable/3489203>.
- P. M. Fernbach, T. Rogers, C. R. Fox, and S. A. Sloman. 2013. Political extremism is supported by an illusion of understanding. *Psychol. Sci.* 24, 6, 939–946. DOI: <https://doi.org/10.1177/0956797612464058>.
- C. Glymour. 2007. When is a brain like the planet? *Philos. Sci.* 74, 3, 330–347. DOI: <https://doi.org/10.1086/521968>.
- Y. Hagmayer. 2016. Causal Bayes nets as psychological theories of causal reasoning: Evidence from psychological research. *Synthese* 193, 4, 1107–1126. DOI: <https://doi.org/10.1007/s11229-015-0734-0>.
- J. Y. Halpern. 2016. *Actual Causality*. MIT Press. <https://www.jstor.org/stable/j.ctt1f5g5p9>.

- J. H. Han, W. Jimenez-Leal, and S. Sloman, S. 2014. Conditions for backtracking with counterfactual conditionals. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, Vol. 36, No. 36.
- F. Heider and M. Simmel. 1944. An experimental study of apparent behavior. *Am. J. Psychol.* 57, 2, 243–259. DOI: <https://doi.org/10.2307/1416950>.
- E. Hiddleston. 2005. A causal theory of counterfactuals. *Nous* 39, 4, 632–657. <https://www.jstor.org/stable/3506114>.
- D. Kahneman and S. Frederick. 2002. Representativeness revisited: Attribute substitution in intuitive judgment. *Heuristics and Biases: The Psychology of Intuitive Judgment* 49, 81. DOI: <https://doi.org/10.1017/CBO9780511808098.004>.
- H. H. Kelley. 1973. The processes of causal attribution. *Am. Psychol.* 28, 2, 107. DOI: <https://doi.org/10.1037/h0034225>.
- D. Lagnado and S. A. Sloman. 2006. Time as a guide to cause. *J. Exp. Psychol. Learn. Mem. Cogn.* 32, 451–460. DOI: <https://doi.org/10.1037/0278-7393.32.3.451>.
- A. Liefgreen, M. Tesic, and D. Lagnado. 2018. Explaining away: Significance of priors, diagnostic reasoning, and structural complexity. In *Proceedings of the 40th Annual Meeting of the Cognitive Science Society*.
- E. Loftfield, M. C. Cornelis, N. Caporaso, K. Yu, R. Sinha, and N. Freedman. 2018. Association of coffee drinking with mortality by genetic variation in caffeine metabolism: Findings from the UK Biobank. *JAMA Internal Med.* 178, 8, 1086–1097. DOI: <https://doi.org/10.1001/jamainternmed.2018.2425>.
- T. Lombrozo. 2010. Causal-explanatory pluralism: How intentions, functions, and mechanisms influence causal ascriptions. *Cogn. Psychol.* 61, 4, 303–332. DOI: <https://doi.org/10.1016/j.cogpsych.2010.05.002>.
- A. Lucas and C. Kemp. 2012. A unified theory of counterfactual reasoning. In N. Miyake, D. Peebles, and R. P. Cooper (Eds.), *Proceedings of the 34th Annual Meeting of the Cognitive Science Society*, Japan. 707–712. <https://escholarship.org/uc/item/9qr693dn>.
- J. Park and S. A. Sloman. 2013. Mechanistic beliefs determine adherence to the Markov property in causal reasoning. *Cogn. Psychol.* 67, 186–216. DOI: <https://doi.org/10.1016/j.cogpsych.2013.09.002>.
- J. Park and S. A. Sloman. 2014. Causal explanation in the face of contradiction. *Mem. Cognit.* 1, 1–15. DOI: <https://doi.org/10.3758/s13421-013-0389-3>.
- J. Pearl. 1988. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, Palo Alto, CA.
- J. Pearl. 2000. *Causality: Models, Reasoning, and Inference*. Cambridge University Press. <https://www.jstor.org/stable/3533601>.
- L. J. Rips. 2010. Two causal theories of counterfactual conditionals. *Cogn. Sci.* 34, 2, 175–221. DOI: <https://doi.org/10.1111/j.1551-6709.2009.01080.x>.
- B. M. Rottman and R. Hastie. 2014. Reasoning about causal relationships: Inferences on causal networks. *Psychol. Bull.* 140, 1, 109. DOI: <https://doi.org/10.1037/a0031903>.

- B. M. Rottman and R. Hastie. 2016. Do people reason rationally about causally related events? Markov violations, weak inferences, and failures of explaining away. *Cogn. Psychol.* 87, 88–134. DOI: <https://doi.org/10.1016/j.cogpsych.2016.05.002>.
- L. Rozenblit and F. Keil. 2002. The misunderstood limits of folk science: An illusion of explanatory depth. *Cogn. Sci.* 26, 5, 521–562. DOI: https://doi.org/10.1207/s15516709cog2605_1.
- L. E. Schulz, A. Gopnik, and C. Glymour. 2007. Preschool children learn about causal structure from conditional interventions. *Dev. Sci.* 10, 322–332. DOI: <https://doi.org/10.1111/j.1467-7687.2007.00587.x>.
- S. A. Sloman. 2005. *Causal Models: How People Think About the World and its Alternatives*. Oxford University Press. DOI: <https://doi.org/10.1093/acprof:oso/9780195183115.001.0001>.
- S. A. Sloman and P. Fernbach. 2017. *The Knowledge Illusion: Why We Never Think Alone*. Riverhead Press, NY.
- S. A. Sloman and D. A. Lagnado. 2005. Do we “do”? *Cogn. Sci.* 29, 1, 5–39. DOI: https://doi.org/10.1207/s15516709cog2901_2.
- S. A. Sloman and D. Lagnado. 2015. Causality in thought. *Annu. Rev. Psychol.* 66, 223–247. DOI: <https://doi.org/10.1146/annurev-psych-010814-015135>.
- C. R. Walsh and S. A. Sloman. 2011. The meaning of cause and prevent: The role of causal mechanism. *Mind Lang.* 26, 1, 21–52. DOI: <https://doi.org/10.1111/j.1468-0017.2010.01409.x>.
- P. Wolff. 2007. Representing causation. *J. Exp. Psychol. Gen.* 136, 1, 82–111. DOI: <https://doi.org/10.1037/0096-3445.136.1.82>.
- J. Woodward. 2003. *Making Things Happen: A Theory of Causal Explanation*. Oxford University Press. DOI: <https://doi.org/10.1111/j.1933-1592.2007.00012.x>.
- J. R. Zaller. 1992. *The Nature and Origins of Mass Opinion*. Cambridge University Press, Cambridge.
- J. C. Zemla, S. A. Sloman, C. Bechlivanidis, and D. A. Lagnado. 2017. Evaluating everyday explanations. *Psychon. Bull. Rev.* 24, 5, 1488–1500. DOI: <https://doi.org/10.3758/s13423-017-1258-z>.

Causation: Objective or Subjective?

Wolfgang Spohn (University of Konstanz)

Abstract

We, and scientific practice, tend to conceive of causation as an objective relation characterizing the external world. Philosophy has been more ambiguous. This chapter intends to renew the doubts. If causation is only a model-relative notion and if causation is tightly entangled with notions that are best understood in a subject-relative way, then the objectivity of causation is at least undermined. The paper discusses these doubts and concludes that the objectivity of causation must not be presupposed, but must be constructively earned.

43.1 Causation: A Bunch of Attitudes

I am glad that philosophy's voice is to be represented in this volume as well—after all, Judea Pearl not only won the Turing Award but also the Lakatos Prize, a highly, if not the most highly renowned award in the philosophy of science—and I am honored that I am invited to contribute as a philosopher. However, philosophy is different; with its more distant view it is prone to have a more critical perspective. Indeed, I feel that this perspective is wanting nowadays.

After the eventual breakdown of positivism, behaviorism, and similar doctrines around 1960, great methodological uncertainty spread, and philosophy, or at least the philosophy of science, seemed much needed. This has thoroughly changed over the past 20 to 30 years. Not that the problems have been solved in a generally accepted way. Philosophy certainly has not solved them; to expect so would be a misunderstanding of the nature of philosophy. Rather, the natural and social sciences have consolidated. They are just no longer irritated. Solving foundational problems has little impact on scientific practice. And methodological problems

have shifted. Data is all-important, but it is also overwhelming, and so data mining, data analysis, machine learning, and computer science in general are the new methodological aids. Philosophical aid seems outmoded.

I can understand this development to some extent, but it is detrimental. I would like to exemplify this with causation, the “cement of the universe.”¹ There is hardly any other notion that is of such universal scientific importance than that of causation. Sciences struggle with it every day. It is thus useful to take again the more distant, 2,500-year-old perspective of philosophy. This is not philosophy’s private perspective. Rather, almost every cognitive enterprise used to run under the label of philosophy. It’s the common heritage of all sciences. Today, though, it’s only philosophy that cultivates this heritage. And it is worth doing so. (Of course, if philosophy would do only this, it would be doomed.)

Aristotle, the first and still most embracing universal scientist, distinguished four notions of causes, which would be better called grounds nowadays. A thousand years later, one of them, the notion of efficacious cause—that’s our modern notion of causation—took center stage. However, it remained under almost complete theological control for another thousand years. Allah or God is the sole or the ultimate cause of everything. And who would dare question Allah’s or God’s ways?

The modern discussion starts with David Hume’s *Treatise of Human Nature*, Vol. I, in 1739, which he wrote at the age of 28. Well, he did not merely start it; he prepared the entire playing ground on which we still move today. Of course, this is a forbiddingly rough summary.² The all-importance of Hume, though, cannot be understated. He confusingly offered two definitions of causation. One, the *regularity view*, is highlighted as an advent of science, although it is recognized as insufficient. The other one is not the counterfactual view, as Lewis [1973b] and his readers, including Pearl [2000, p. 238, 2018, p. 20] state. It is rather what I like to call the *associationist view*. It seems repressed nowadays. According to it, causation is in the eye of the beholder, a habit of thought. This is a gross oddity, it is natural to discard it as a misunderstanding. Hume himself says about it: “I am sensible that of all paradoxes, which I have had, or shall hereafter have occasion to advance in the course of this treatise, the present one is the most violent” (1739, p. 166). In effect, he is so ambiguous about it that interpreters have puzzled over the relation of his two definitions till the present day (see, e.g., Beebee [2006]).

1. This is the phrase of Hume [1740]. More precisely, he says that the “principles of association ... resemblance ... contiguity ... causation... are really to us the cement of the universe.” So, actually, and interestingly, he is talking about “epistemic cement.”

2. The epilogue of Pearl [2000] gives a much longer, but still brief and very entertaining overview.

“A habit of thought,” this sounds so understated. In the more elevated German way, Kant [1781] turned this into a *pure category of thought*. What a label. The idea, though, is basically the same as Hume’s. Causation is a relation we impose on the world. It is not a notion we acquire from experience, not an idea of sensation, but rather *an idea of reflection*, in Hume’s words—although, of course, experience is required to learn how the relation realizes. I think Kant is right. However, I will not use this chapter to positively defend this claim. The only aim I am pursuing in this chapter is to create some awareness of the fact that contemporary theories of causation are not safe at all from being infected by these old and important ideas about causation.

For, what is the contemporary attitude toward causation? In the positivistic times mentioned above, causation was a shunned notion, bad metaphysics, not imposed by our mind—this would be preposterous—but also not to be found in the world. This changed with Hempel and Oppenheim’s [1948] theory of deductive-nomological explanation, which was, *in nuce*, Hume’s regularity theory of causation. The irony was: there was no causation in that theory, as became clear about 15 years later. But the ban was broken, and causation is continuously among the hottest topic in the philosophy of science up to the present day. It was not so different in the various sciences, although each has its own speed. Pearl [2018] tells impressive stories about how obstinate the community of statisticians was and still is.

So, the importance of causation is acknowledged almost everywhere now. The natural and social sciences came up with a really surprising variety of ideas and conceptions. If you study them, it’s hard to believe that they all talk about the same thing. The goal was to have specific and useful accounts, not just sublime philosophy. However, the matter turned out very difficult, and the ideas were quite idiosyncratic and tentative.

The field is still scattered. However, a certain paradigm emerged around thirty years ago, which by now seems to be the dominating one, sharing wide agreement and applicability. I am referring to the *interventionist theory of causal Bayes nets*, the cornerstones of which are Pearl [1988, 2000], Spirtes et al. [1993], and Woodward [2003]; it was substantially adumbrated, though, in Spohn [1978, 1980]. The three are by no means identical; there is quite a lot of divergence in detail. Still, it is legitimate to subsume them under one broadly conceived heading. And there is no doubt that no one did more than Judea Pearl to familiarize other disciplines with this doctrine and to convince them of its wide applicability—perhaps because as an AI researcher he is closer to the needs of the sciences.

This is tremendous progress and unprecedented success. However, when it comes to the nature of this doctrine, its contenders are surprisingly silent, Judea

Pearl included. Unlike many predecessors, starting with Hume, they don't try to define causation. This may be plausible. There must be some basic concepts, and then causation is likely to be one. Glymour [2004] emphasizes the liberating effect of this move. Similarly, Pearl [2018, p. 27]: His approach, which he attributes to Alan Turing, "is exceptionally fruitful when we are talking about causality because it bypasses long and unproductive discussions of what exactly causality is and focuses instead on the concrete and answerable question 'What can a causal reasoner do?'" However, this strategy does not avoid conceptual issues. If not definable, causation is at least closely related to other basic notions and thus at least infected by their character, as we will see below.

Instead, the main interest was to build causal models, to study their behavior, and to say how they can be tested. This was explored in great constructive detail; only thereby could wide applicability be acquired. What does this procedure leave to be desired? The background ideology certainly is that there are sort of objectively true causal models. This much seems to be tacitly understood, even if one is modest in claiming truth for the models one entertains. And the account of Pearl and others is the best way to get on to the track of the true models. Pearl [2018] does not explicitly speak of true causal models, but he explains the many inferences causal models allow, provided—that's a recurring phrase—"your causal model accurately reflects the real world" (p. 335).

Whenever I talk to scientists, this seems to be their common attitude as well. Of course, causation is an objective feature of the world, and science is there to uncover it. Anything else would undermine the self-conception of science as truth-seeking. And now we finally have a grip on how to do it.

Really? Are Hume and Kant thereby refuted? And the positivists defeated? I would like to cast doubt on this attitude. The objectivity of the notion of causation is not guaranteed at all. A crucial quote from Pearl [2018, p. 21] is: "If I could sum up the message of this book in one pithy phrase, it would be that you are smarter than your data. Data do not understand causes and effects; humans do." But what is it that makes us smarter than the data? The answer Pearl [2018] unfolds is that it is the second and the third rung of his so-called ladder of causation, acting/intervening and imagining the counterfactual. Maybe, though, we are smarter not because of being able to represent more objective truth than the data, but because we are able to add something to the data?

As said, I do not want to defend an answer to this question. But I want to suggest that objectivity must not be simply assumed and is not so easily earned. Subjectivity creeps in from at least two directions, which I want to briefly discuss in this paper. One point is the model relativity of the notion of causation, and the other is the potential subject-relativity of the notions with which causation is at least

intrinsically related. This does not yet confirm Hume or Kant; but it shows that matters are less clear than scientists wish.

43.2 The Model Relativity of Causation

Even if there is no general agreement, we have a fairly good conception of what causal models are and how they behave. Thus, we know what causation is, what the causal relations are *within* those models: they are either directly given by the arrows between the nodes or the variables of a causal graph, or they consist in certain probabilistic conditional dependencies among those variables, or they lie in the structural equations relating those variables, and so on. This is our grasp of causation. It is, however, only a model-relative grasp. Is causation hence a *model-relative* notion?

I observe a profound ambiguity concerning this question. On the one hand, I sense an implicit inclination toward model relativity, although it's hard to find it explicitly endorsed. Perhaps I get this sense because people are only dealing with causal models; this is the only frame within which they talk of causation. On the other hand, this attitude clearly won't do. Causation can't be *only* a model-relative notion.

Compare this with the notion of truth, another notion of utter fundamentality. There, Tarski has provided us with the model-theoretic notion of truth, of what it means that a sentence is true in a model.³ Thereby we have gained a rigorous grip on truth theory, for the first time in history. However, this can't be the full truth about truth. We also have a notion of absolute truth or truth *simpliciter*. "The sun is shining." That's true—full stop (when I am writing this sentence). Relatively speaking, it's only true in one model and false in another (and doesn't get any truth value in a third). So, what's absolute truth? One is tempted to say: absolute truth is truth relative to the true model. But that's blatantly circular.

I won't try to resolve this predicament of truth theory; it's a serious problem. However, the analogy is illuminating. Clearly, acquiescing in the model relativity of causation would introduce an intolerable amount of subjectivity. Causal relations cannot be this way or that way, depending on the causal model we choose; we cannot have it both ways. This would undermine scientific objectivity. Thus, at least implicitly scientists presuppose an *absolute* notion of causation, and this is what their modeling activities try to capture. What is it?

3. Sometimes, people speak only of truth in an interpretation. In any case, this notion of a model differs from, and is much more general than, the notion of a model used in the theory of causation.

The literature, not only in the applied sciences but also in philosophy, hardly comments on this question. To be honest, I find this shocking. Apparently, the question is not really relevant. Somehow, the causal models fit better or worse; so, we know in which direction to improve our causal models; thereby we approach the true causal model; and that's the one that grasps causation in the absolute sense. Thus, there is really no more than the model-relative notion of causation, amended only by the notion of fit or truth of a model. This seems to be the general attitude, and it is certainly the one displayed in [Pearl \[2018\]](#).

I have offered two terms here, "fit" and "truth." "Fit" sounds more cautious, perhaps this is all scientists expect of causal models. However, they cannot waive truth. They may be modest in not claiming to possess the truth. Still, truth must be their guiding aim. Hence, the present discussion is really about the truth of causal models. The general attitude parallels the blatantly circular answer in the case of truth theory. For causation, however, it does not sound circular. But is it any better?

I don't think so. One needs to understand that the truth of a causal model is never a relation between the model as such and reality, as it were. A causal model may fit the data very well, and then there is no reason for suspicion. It may even overfit the data. Often, though, the fit is not so good. One may then adapt the parameters of the model or take similar moves, without essentially changing the model. Usually this won't do, however. Criticism of the model takes the form of an enlarged model accounting for more variables. Almost all discussions are about neglected variables which disturb the picture in one or the other way and because of which all those partial regression and correlation coefficients are misleading. There are common causes, there are confounding variables and selection variables, there are unmeasured and latent variables, Simpson's paradox lurks more often than expected, and so on. By making those neglected variables explicit in the enlarged model one may reach a better fit. This is *always* a possibility, even if the original model fits very well and does not raise suspicion. Surprises can never be excluded.

Of course, practicing scientists rarely aim for perfect models. There is always some slack between the model and the data. Scientists are content when they can be confident to have identified the main causes. They would admit that there always are a lot of further causes blurring the picture. But if they blur it only a little bit, we need not worry. One must always think where to spend one's efforts, and to explore those residual causes may not be worth the efforts. Again, though, surprises can never be excluded. So, this attitude of the practicing scientist only confirms the fact that the truth rather lies in an enlarged model.

Note, however, where this takes us. Isn't this to say that a causal model is true if it is part of an enlarged model, not any enlarged model, of course, but a true enlarged model? And now we are caught again, not in a circularity, but in an infinite

regress. No model is large enough to decide about the truth. We are deferred in the end to what may be called the universal model containing all variables whatsoever, so that no variable is neglected, no further confounding or otherwise disturbing variable can turn up. Surely, though, that's completely ill-defined speech. The universal model is at best a fictitious ideal of which we have no more than the faintest grasp.

For this reason, I am claiming since [Spohn \[2001\]](#) that causation in the intended absolute sense is a model-transcendent notion. Limited causal models do intend it, but whether they grasp it cannot be decided by any other limited model, however enlarged. I want to briefly indicate that this model-transcendence transpires through all of the current theories of causation.

[Spirtes et al. \[1993, pp. 44f.\]](#), for example, make very clear that their basic causal axioms, the causal Markov, minimality, and faithfulness condition apply only to causally sufficient causal models—where, roughly, a model is causally sufficient if it contains all common causes of any two variables in the model. Clearly, if this is taken literally, this means that any causal model must contain the Big Bang, which surely is a remote common cause of any earthly matters. Fairness requires to say that Spirtes et al. have done a lot to weaken this presupposition by exploring how much we can still infer about causal relations in its absence; see, in particular, their second edition.

[Woodward \[2003\]](#) perfectly displays the interventionist agenda on causation. However, if one looks closely, his notion of intervention is model-transcendent, too. If we intervene on the variable X in order to find out whether it is a cause of the variable Y , he requires the intervention on X to be statistically independent of any variable that causes Y along some causal path that does not go through X ; this is condition I4 in [Woodward \[2003, p. 98\]](#). Here, “any variable” must be taken as quantifying not only about the variables in the causal model but also about all variables outside the model. This is his way of model-transcendence.

The same remarks apply to Pearl's *do*-operator. The model-immanent function of $do(X)$ is to causally separate the variable X from all its causal predecessors in the model. However, this separation is to hold for any enlarged model as well. That is, although $do(X)$ is explained by Pearl as just another variable with a special behavior within causal models, it really has a model-transcendent function. Or in more general words: The truth claim of any causal model always carries the implication—“and there are no further neglected variables, confounding or otherwise, which change the causal picture.” This is clearly a model-transcendent implication.⁴

4. This is not to say that by using the relative notion of causation we are bound to make the closed-world assumption (see, e.g., [Pearl \[2000, pp. 252f.\]](#)). As such, relative causation is just causation

Where does this leave us? We are not forced to acquiesce in the model relativity of causation. We can get rid of it, but not in the way commonly assumed. The truth of a causal model is not a local affair that could be locally settled. Rather, we are referred to ever larger causal models, but nothing is ever settled due to the model-transcendence of absolute causation. So, in a way, we indeed deal only with model-relative causation; it's always more of the same in ever larger models. But it is important to be clear on what we are up to with causal models, to be clear about what truth could mean for them.

43.3 Laws

We may thus have banned the subjectivity entering through the model relativity of causation, though in a somewhat unexpected way. Let me turn, hence, to the other potential source of subjectivity, the nature of the concepts with which causation is closely connected, even if one should have given up defining causation by them. When one surveys theories of causation, the connection always refers to one of two kinds of concepts, either to something like regularities, laws, structural equations etc., or to probability, which is the central notion in all statistical contexts.

Of course, causation is essentially connected to still further notions: action (this relation is perhaps sufficiently reflected in interventionist theories of causation), order (if this is explicated as entropy, one may subsume this under the probabilistic connection), and most importantly, space and time. For physicists this relation is absolutely central. In the social sciences it is often marginal. Surely, if we are to model climate change or the proliferation of a pandemic, space and time are indispensable categories. Often, though, these categories do not even play an implicit role. The reason is clear. There is often no temporal order in the data and hence none in the causal model representing the data.⁵ Still, I am wondering how one can ever do causal theorizing while neglecting its first axiom, namely that causes temporally precede their effects.

In the present context, however, we may neglect these other connections because they do not endanger the objectivity of causation. Let me therefore focus on the two connections initially mentioned and first on laws and its ilk. The law connection is the one originally claimed by Hume's regularity of causation. If it would be appropriate, it would bar subjectivity. However, it is not appropriate, for various reasons.

relative to the model. Only when we claim that model-relative causation amounts to absolute causation do we claim the closed-world assumption to be true.

5. See also [Pearl \[2000, section 7.5.1\]](#) for a discussion of this point.

First, Hume was not so sophisticated to distinguish between accidental regularities and genuine laws. Certainly, only the latter create causal relations. Although the natural sciences take it to be clear what they are after when they are after laws, I can only warn the reader to enter the philosophical discussion about what laws are; it's a quagmire.⁶ One certainly finds various opinions giving up on the objectivity of laws and thus of causation.

However, that's presently only a side issue. There are two more important concerns. One concern is that laws by themselves cannot tell about causal relations. As has often been observed, the counterfactuals describing causal relations really are *counternomologicals*; they refer to which laws still hold when some laws are broken. The laws can never tell this by themselves. Let's consider a simple example and assume that the co-occurrence of falling air pressure, the falling of the barometer, and a thunderstorm were sort of a strict law. Now we break the law between air pressure and barometer by manipulating the barometer. The question determining the causal relations then is: which law still holds, that between barometer and thunderstorm, or that between air pressure and thunderstorm? Not both can still hold. The answer is obvious to us. The point is only that the question and the answer are counternomological ones. The example makes clear that interventions are also invoking counternomologicals; they introduce small miracles, in the terminology of Lewis [1973a, pp. 75ff.].

The next question is: what governs those counternomologicals? The answer is not clear at all. Perhaps a similarity ordering à la Lewis [1973a] does the trick, perhaps some epistemic entrenchment order is working in the background (see, e.g., [Gärdenfors 1988, chapter 4]). Something of this sort is required. However, the objectivity of all these auxiliary notions is at least doubtful. I don't want to say that they are hopelessly subjective. *Prima facie*, though, they do look subjective. We might reach intersubjective agreement concerning similarity, entrenchment, or whatever, though we would have to study on which grounds we can do so. Possibly we can even claim some kind of objectivity for our agreement in the end; but again the question would be on which grounds we are able to do so.

What I want to emphasize: The issues I am raising here are not issues of the ordinary scientists. They usually proceed from a tacit understanding and agreement. However, if they reach agreement, it's not due to collectively grasping what is objectively there. It's *not* like: "Why does (almost) everybody say that $2 + 3$ is 5? Because $2 + 3$ is 5." Rather, agreement comes about in some other way. And if it can claim objectivity, it is not the objectivity of ordinary facts. Our dealings with similarity or entrenchment orders and the like are not a scientific but an epistemological issue

6. If you want to disregard my warning, you may start with van Fraassen [1989].

that requires a different sort of study. The point then is: if causation is essentially entangled not just with laws but with all this additional machinery of doubtful objectivity, then causation is deeply infected by this machinery as well. The objectivity of causation cannot be presupposed, but must be earned and constructed in the way required for this machinery.

The other important concern is that we all have this noble ideal of a natural law, allegedly explored in basic physics. But of course, the laws investigated in the social sciences, economics, geology, biology, even in most parts of physics, and so on, are *not* of that ideal kind at all. The ideal is very misleading. Rather, they somehow are soft, non-strict laws; they are, as we say today, *ceteris paribus* (*cp*) laws. A simple physical example is Hooke's law about the proportionality of the extension of a spring and the weight attached to it, which, of course, has countless exceptions. To be sure, all structural equations are of the same kind, wherever they are formulated. This fact is certainly clear to the working scientists, even though they may not have fathomed its epistemological implications. For philosophers of science the insight came quite late; too long were they attached to the ideal. But once they started thinking about them, *cp* laws turned out to be an utter mystery (see, e.g., [Reutlinger et al. \[2019\]](#)).

Look at "*cp*, (all) *Fs* are *G*" (e.g., "*cp*, birds fly," or "*cp*, prices go up, when demand rises"). What does this claim? How must the world look like for this to be true? It's very unclear. Polemically, one might say that it doesn't claim anything at all; it simply says: "all *Fs* are *G*, unless they aren't." This is unfair; scientists don't claim platitudes. But it is very hard to avoid this unintended answer. Another reply is that *cp* laws are statistical laws. Judea Pearl seems to tend to this answer.⁷ Most people, however, would reject it. Hooke's law is not a tacitly statistical law about the manufacture and use of springs. Our schematic law doesn't say "most or 99% *Fs* are *G*." It rather says "normally or typically, *Fs* are *G*." And normality or typicality is not just a matter of proportions. But what is it?

The core problem is that we slip into a similar open-ended situation as we did with causal models in the previous section. We might start with saying: "*cp*, *Fs* are *G*" means "under normal conditions, *Fs* are *G*," leaving the task to specify the normal (and the exceptional) conditions. Maybe we can confirm good hypotheses: "whenever normal conditions *N* hold, *Fs* are *G*," and "whenever *E* (= not-*N*), *Fs* are not *G*." But of course, these hypotheses are not literally true. They are *cp* laws in turn, and we will find further exceptional conditions *E'* and *E''* such that:

7. In [Pearl \[2000\]](#), he explains right on pp. 1f why he turns to probabilities. One reason he gives is that "causal expressions in natural language are subject to exceptions," as are *cp* laws, and that "probability theory" is "especially equipped to tolerate unexplicated exceptions."

“whenever N and E' , F s aren't G , either” and “whenever E and E'' , F s are G , after all.” And so on. With a little phantasy, you can easily take three or four rounds of this game with my sample laws. This process is non-monotonic, as logicians say; strengthening the conditions may always reverse the law. And, in analogy to causal models, the process is open-ended; you are never in a position to say: “Now I have exhausted all conditions under which F s are, or are not, G .”⁸

The upshot is that by claiming a cp law we do not make a claim with a plain truth condition to be ascertained or confirmed in familiar ways. The dialectics of normal and exceptional conditions is a different epistemic game. Of course, it is legitimate to play this game; it's the way of science. However, its rules are quite unclear. It's not an ordinary search for truth. How could it be, when the claims made are qualified by cp clauses and thereby lose a plain truth condition? Again, it seems to be the task of the epistemologist to clarify the matter and to find out about the underlying methodology of this cp science.

It is not so clear what the epistemologist will find. To be sure, nothing is objectively normal or exceptional. Normality is, to put it vaguely, an anthropocentric notion. So, subjectivity lurks again. In scientific contexts we can perhaps restrict the notion of normality to its epistemic uses.⁹ But even in this case it is basically subjective, and we must again find a different explanation for reaching consensus than that the consensus agrees on objective truth.

Thus, my point is the same as above: If causation is closely entangled with soft cp laws, then it is also entangled with this non-objectivity, with this absence of truth conditions just observed. If so, the objectivity of causation can again not be presupposed. Rather we have to study, by studying the epistemology of cp science, how causation may, perhaps only partially, acquire objectivity.

Where do we stand? If we should have hoped to somehow anchor causation in objective lawhood, this has ended in disappointment; cp laws are not the kind of laws to satisfy our idea of objectivity. And even if they were, laws only would not do; they would have to be amended by some machinery answering counter-nomological questions. Maybe, though, we can avoid this muddle by taking the turn that most sciences have taken, anyway. Maybe we can avoid all reference to laws and the like and instead look at the connection between causation and probability.¹⁰

8. As indicated in the previous footnote, this analogy is one motive for Pearl to resort to probabilities. For Woodward [2002] it is a reason to try to analyze cp claims as causal claims. Either way, the problems I am about to display persist.

9. In Spohn [2014] I have tried to explicate this epistemic use in terms of ranking theory, which, I argue, is ideally suited for this job.

10. Van Fraassen [1989] is not about causation. However, it is precisely the probabilistic turn that he propagates there in order to escape the muddle of laws.

This may look promising. However, I would like to indicate in the rest of this paper that we thereby move out of the frying pan into the fire.

43.4 Probability

When one looks at contemporary causal theorizing, one is overwhelmed by its probabilistic character. A hundred years ago, this was unthinkable. Causality was firmly tied to deterministic theorizing. Causes were mostly conceived as necessary and sufficient causes. Things changed when physics turned out to be irreducibly probabilistic. At first, it seemed that we had entirely lost causation in the physical realm. But then it became ever clearer that probabilistic causation makes good sense as well. Nowadays we find this attitude also in all of the social sciences. Our data is probabilistic, and when we hope to find causal relations in it, it can only be in the form of probabilistic tendencies. So, it's not surprising that probability now is the key notion with which causality is wedded.¹¹

However, what do we mean by probability? In philosophy we discuss several different interpretations—five? or more?¹² It should make a big difference for our understanding of causation with which of these interpretations it is connected. Again, I am surprised how little this is discussed in the relevant philosophical and scientific literature. Is it not important? Is it clear, anyway?

Well, whatever the other interpretations may be, the social sciences (medicine, etc., always included) obviously speak of *statistical probabilities*. This appears to be taken as the only relevant interpretation. However, do we know what statistical probabilities are? Did we check whether they are suitable for connecting up with causation? Again, the literature appears to take this as settled. Let me approach these questions by first briefly explaining how rich and unclear the concept of probability is despite the fact that its mathematical structure is unequivocally fixed.¹³

The clearest interpretation is the *subjective* or *Bayesian* one. According to it, probabilities are rational degrees of belief. There are a lot of arguments why rational degrees of belief must take the form of probabilities. We may leave it open how cogent these arguments are and whether there might be other reasonable

11. What *is* surprising is the far-reaching marginalization of deterministic causation. I find it very unlikely that we searched for a chimera for 2,000 years.

12. Galavotti [2005] and Gillies [2000] are two very commendable presentations of this confusing field.

13. Well, almost. There is some uncertainty concerning σ -additivity and concerning the representation of conditional probabilities via Renyi and Popper measures. This need not worry us here.

conceptions of degrees of belief.¹⁴ There is no doubt, though, that probabilities are by far the most familiar conception of these degrees. Subjectivists, Bruno de Finetti ahead, claim that this is indeed the only intelligible interpretation of probability. However, we need not go so far; it suffices to say that it is at least one good and reasonable interpretation.

I was surprised to read in Pearl [2000], right on p. 2: “We will adhere to the Bayesian interpretation of probability, according to which probabilities encode degrees of belief about events in the world and data are used to strengthen, update, or weaken those degrees of belief.” This is a resolute Bayesian avowal extending over the entire book. However, I suspect that he is not consistent in that avowal. The book’s later parts are about statistics, and statistics don’t refer to single events in the world, as I will point out below.¹⁵

And the avowal betrays the quest for objectivity. There are no true subjective probabilities. They can and should be well-informed by the data; but then they can change to being even better informed. They might be called true if they conform to objective probabilities. However, this idea is highly problematic. One reason is that objective probabilities themselves are highly problematic, as we shall see. Another reason is that there is more to know about a fact than its objective probability (if it has one), for instance, the fact itself. Thus, perhaps, only a probability of 1 for the fact can be called true?¹⁶ I conclude: we better abstain from calling subjective probabilities true or false or taking them as representing reality.

For causation this entails that there are nothing but causal beliefs, which may be more or less well-informed, which, however, cannot be called true. They do *not* represent any causal reality. This runs counter to the general attitude we meet in the sciences. And it seems to run counter to Pearl’s own attitude that I have quoted in Section 43.1. Time and again, he slips into realistic talk, from causal beliefs to belief in causal facts. However, within the Bayesian interpretation this is an illegitimate

14. For decades I have been propagating ranking theory as another model of degrees of belief. Not the least of my reasons is that ranking theory allows to state a theory of deterministic causation (which speaks of causes making their effects possible or necessary) in close parallel to probabilistic theories (which speak of causes making their effects more probable). See Spohn [2012], in particular chapter 14.

15. For instance, Pearl [2000, section 7.5.4] discusses singular versus general causes. But his discussion refers to statistical probabilities concerning populations and not to subjective probabilities about single events.

16. Neither does it help to say that the proper probability is true only before the fact in question realizes, and probability 1 is true only after the fact. This would make truth time-dependent in unwanted ways.

move.¹⁷ So, it seems we should attempt to avoid the Bayesian interpretation in our context.

But beware. Whenever we get into trouble with other interpretations, the subjective interpretation is the only one that always works, that makes sense in every application. We can always resort to making assertions only about our and other people's (causal) beliefs. So, whenever a fallback position is needed, we may well be forced back into the subjective interpretation.

Let me add a few remarks about *Bayesian statistics*. Bayesian statisticians, or Bayesians for short, are not subjectivists; they certainly grant objective probabilities in some sense. They only claim that in doing statistics we need to consider our prior assumptions as well, represented by our subjective probabilities. The traditional Neyman–Pearson school hopes to do without these subjective elements. What sounds like a principled disagreement—it indeed is—apparently turns into a fair cooperation in practice.

However, I think Bayesians have a delicate standing in our present context.¹⁸ It won't do for Bayesians just to use subjective as well as objective probabilities. For, each probability measure must have a uniform interpretation; one cannot mix different interpretations within one measure. So, the Bayesian needs to assume bridge principles translating between objective and subjective probabilities. Such principles are not hard to come by. For instance, if all I know about a given event is that its objective probability is x , my subjective probability for that event should obviously also be x . By introducing his so-called Principal Principle, Lewis [1980] has initiated a big philosophical discussion about such bridge principles; they may need generalization and modification.

This is, however, not a satisfactory rescue for the Bayesian. One problem is that not all kinds of objective probability are equally suited for such bridge principles. We shall see below that the so-called statistical probabilities are indeed ill-suited. Another problem is that we cannot turn all probabilities the Bayesian refers to into objective ones. The required uniform interpretation of probabilities can only be a

17. More generally, the tendency to slip from conditional belief to belief in conditional propositions is ubiquitous. This step is so easy. However, the move hides all ambiguities between epistemic and realistic world conceptions. It is not innocuous at all. Stalnaker [1984, chapters 6 and 7] is a paradigmatic, but in my view unsuccessful, struggle with what is going on in this move.

18. Pearl [2018, p. 90] complains that “Bayesian subjectivity in mainstream statistics did nothing to help the acceptance of causal subjectivity.” The latter means for Pearl that each causal inquiry must start with positing a subjective causal model and must grant the possibility that data may not decide between two different causal models subjectively posited. This sharply differs from the subjectivity I am discussing here.

subjective one. This point then extends to his account of causation, which must be similarly subjective. Thus, we are back at the position above which Pearl avowed, but which we might want to avoid.

Let's turn, hence, to *objective* probabilities. Here, interpretational variance starts. Still, there is a common anchor. Everybody agrees that probabilities somehow ground in relative frequencies; that's their connection to reality. However, this even holds for the Bayesian interpretation; of course, well-informed subjective probabilities listen, and are usually close, to observed frequencies. In particular, though, it holds for all objective probabilities. This grounding is spelled out in the fundamental law of large numbers, proved by Jacob Bernoulli already in 1689 and called the "Golden Theorem" by him. It guarantees that the relative frequencies in infinite independent repetitions converge to the single-case probabilities—though only in a probabilistic sense. This means that some notion of probability is already presupposed by the law of large numbers, and it says then how those probabilities probably manifest in frequencies.

Frequentism, which has been very popular among working probability theorists, wants to turn around the relation. It is the doctrine that probability is *defined* as the limit of relative frequency in random sequences, where random sequences are subject to further qualifications, most notably complexity-theoretic ones. Frequentism's crucial problem, with no good answer to date, is that it applies only to infinite sequences, strictly speaking. It cannot be employed for the single case, which is the one we are interested in. We want to know the probabilities governing the next throw of the coin, and this is about the next throw, not about an infinity of throws. Thus, frequentism is *not* supported by the law of large numbers, which already presupposes those single-case probabilities.

Or to address our present concern: Suppose we could isolate a causal system modeled with its probabilities, what does it teach about the causal relations, if we run the system very often and speculate about the limiting frequencies? These relations are in the system, and they are somehow connected with the probabilities in the system and not with the frequencies in the repetitions. With respect to causation, too, we need a notion of objective probability that applies to the single case.

There is such a notion that serves our purpose; philosophers call it the *propensity interpretation*. According to it, the objective probability of, for example, a die showing a 6, is something attributable to the die as such, an intrinsic feature of the die and its set-up, the throwing device, a disposition that can only be described probabilistically and not deterministically, viz. a propensity. The single-case propensity is basic, but it entails, of course, a long-run propensity, which

converges as described by the law of large numbers. A die or a roulette wheel are already good examples, although one may argue about whether they “really” are deterministic devices.

The ultimate examples can be found in quantum and nuclear physics. Radium atoms, for example, have a propensity to decay. We could say that there are many different kinds of radium atoms, each with a different deterministic decay time. Determinism saved. However, this would make no sense at all, since there is no way to tell the kinds apart; it would be a distinction without a difference. Hence, it is much more reasonable to say that all radium atoms have the *same* irreducibly probabilistic propensity to decay governed by an exponential distribution. This is a genuine statistical law: all objects of a certain kind show the same stochastic behavior.

The decay propensity of radium atoms is not immutable. It can change. For instance, we can excite the nucleus by various kinds of radiation and thereby accelerate its decay in various ways. We may set up causal models representing these propensities and their potential changes. Such a model would describe genuine probabilistic causal laws applying to each single case in the same way. Understanding probabilities objectively in this way would thus allow us understanding of causality in the same objective way.

This is what we were looking for. However, the crucial point for the rest of the paper is: Success does not extend; propensity is *not* the kind of probability referred to in most applications of causal models discussed in the literature. These applications belong to the social sciences, medicine, epidemiology, for example, and the statistical probabilities they refer to are not propensities as just described. Let me explain.

Many of those models are a matter of life and death. I certainly have a deterministic propensity to die sometime. But it does not make sense to speak of any propensity of mine to die before 80 or after 80. There are millions of potential causes of my death, most of which are not within my reach. The chance set-up in which my future death is located spreads more or less over the entire surface of the earth and further. The hope that a universal wave function could decide about this propensity would be nothing but a silly reductionist phantasy.

However, aren't there mortality tables? Sure. They don't tell anything, though, about my propensity. They tell how likely men in my age are to reach 80. But I am not an average man, nobody is. It is entirely unclear which specific subgroup would consist of the men relevantly similar to me, and if it were clear, there would definitely not exist any statistics for that subgroup. This is the well-known *problem of*

the reference class, which has no good answer.¹⁹ It prevents transferring statistics to the single case; we can't statistically infer single-case propensities.

Certainly, though, we are inclined to reason as follows: If 60% of the men in my age group reach 80, and if you have no information about me that makes me in any way special, then your subjective probability for my reaching 80 should be 60%, too. This reasoning applies a kind of bridge principle relating objective and subjective probability, or perhaps relating only frequency and subjective probability. Presumably, we use this kind of reasoning, at least roughly, whenever we read a statistic.²⁰ But note that we thereby return to subjective probabilities about the single case, which always make sense. And note that the premise of the argument, the absence of special knowledge, is, strictly speaking, almost never satisfied. Usually, we do have special knowledge about a given single case, which we reasonably conjecture to be statistically relevant, a bit at least, even though we do not have a relevant statistic.

Note how different this is from my physical example. There we could legitimately assume single-case probabilities that entail the statistical behavior of large samples. Here we only have the statistical behavior of large samples without any underpinning by objective single-case probabilities; only shaky subjective inferences about the single case are feasible. This is a *world* of difference.

Of course, I have chosen a graphic example, the probability of death, something potentially caused in more ways than anything else. My point seems obvious in this example. However, the radium atom was an equally clear example for objective propensities. Where on the scale from the one to the other example do the propensities get lost? I do not know; it seems very hard to say. They do get lost through the multitude and the externality of the causes of the object's states to be probabilistically assessed. In view of this multitude what can we still attribute to the object itself? Already a person's propensity to recover from a certain disease after taking a certain drug is a very unclear case, I think. Moreover, the onus is not on me to say where propensities are lost. The onus is on the friend of objective probabilities to show that he is still legitimately speaking of them.

Perhaps, though, he is not speaking of propensities at all. The scientists fitting their causal models to statistical data refer to statistical probabilities; that's what they would say. Let's finally ask, then: what are *statistical probabilities*? Primarily,

19. Probably, the reference class relevant for me consists only of myself—not good for doing any statistics.

20. For an affirmative discussion of this statistical bridge principle, see Schurz [2019, pp. 57–68].

they are just relative frequencies in a given population, which behave like mathematical probabilities. But frequencies are not probabilities. Genuine probabilities enter only through the random mechanism by which individuals are selected from the population. If each individual has an equal chance (= objective probability = propensity) to be selected, then the chance that an individual with a certain feature is selected is the same as the relative frequency of that feature in the population. However, speaking of probabilities in this sense is only a roundabout way of speaking of the frequencies.

Usually, the procedure is the other way around. We cannot register the entire population; we can only observe a representative sample, which is selected by such a random mechanism. Then inferential statistics is needed to probabilistically infer the distribution of the features in the population from that in the sample. But note that these inferred probabilities are not objective probabilities for the shape of the population; they are subjective probabilities expressing our expectations about this shape. Of course, this does not mean that they are arbitrary. They proceed from an objective base in the representative selection mechanism by statistical inference. However, making a random selection from the population does not make the population itself in any way chancy.

Where does this leave us with respect to a causal model? It contains a set of variables that characterize the shape of the population, it contains causal arrows between those variables, and it contains many quantities that look like absolute and conditional probabilities. But these quantities are either observed frequencies in the sample or estimated frequencies in the population. And they confirm, or do not confirm, the causal arrows via the methods of causal inference. However, it must be clear that causes and effects in the model are nothing but relative frequencies in the population. By changing the relative frequencies for the cause variable one can change the relative frequencies of the effect variable. This is most useful information, for sure. But it is *this* kind of information and nothing else. And, I find, it makes the causal model appear quite mysterious because the causal story it delivers is a brute story about the population level without any underpinning from causal stories on the individual level.

This may appear as a very unfair presentation of what is going on in causal models. In particular, my claim about the missing underpinning from the individual level rests on my claim that it rarely makes sense in the applications in the social sciences to speak of individual propensities. I suspect that the general attitude rather is to simply *postulate* those individual propensities. We may not know much about them, and they may have considerable variance. But we do know that they generate the frequencies we observe in the samples or estimates for the population. Hence, what we observe and estimate is an (statistically qualified) average individual propensity.

For instance, if 60% of the men of my age reach 80, then the average propensity of men of my age to reach 80 is 60%. The individual propensities diverge in unknown ways, but they must (roughly) have this average. If a certain drug raises the recovery rate for a certain disease from 20% to 50% in a sample (or in the test group as compared with the control group in a randomized controlled trial) or probably in the population, then the average recovery propensity of those having the disease is raised by the drug from 20% to 50%. Again, the individual propensities will diverge, but they must (roughly) have this average. As stated above, however, any inference to those individual propensities almost inevitably results in subjective probabilities about the individual cases. I have more than 60% confidence to reach 80, and if a person recovers after taking the drug, this is perhaps not just because a 50% propensity has played out well.

This is in no way to question the great value of knowledge about average propensities (= observed or estimated frequencies) and about how to change these average propensities. However, what is the conceptual gain of this move? We now have a hypothetical individual underpinning of the population frequencies. This is indeed a causal underpinning by hypothetical causal stories about hypothetical individual propensities adding up to average propensities. However, we do not know much about that underpinning beyond the frequencies to which it leads. We do not have any statistical laws for the individual cases. And as explained, this underpinning is at best hypothetical and at worst meaningless.

Let me make clear once more what my dialectics on the previous pages was supposed to be. I started out saying that a subjective interpretation of probabilities can be applied everywhere. This would be fine—except that it does not satisfy our objectivistic intuitions concerning causation. This motivated the search for suitable, more objective interpretations that could save the objectivity of causation. This search was perhaps not entirely negative, but the objectivists can hardly be pleased by its weak and problematic results.

My general moral hence is: We must not presuppose the objectivity of causation and of the notions with which it is related. The safe fallback position is always the subjectivistic one; perhaps we should indeed start with Pearl's avowal of Bayesianism. And starting from there, we must work hard to earn and establish objectivity, without guarantee of success. I don't claim that Hume and Kant are thereby confirmed. However, I hope I have succeeded in pleading for more openness toward their doctrines.

This need not undermine the self-conception of scientists as truth seekers. It only suggests a more complicated picture of truth-seeking. Truth seeking is not just somehow adequately representing reality. It has much more to do with subjective belief, with intersubjective agreement, with rational belief and belief change, guided by principles of epistemic rationality, which must be agreed upon in turn.

All of this must be made explicit. When we do so, we may (have to) take recourse to another kind of objectivity, the objectivity of rationality. This is of a normative kind and as such delicate, contested, and not secured at all. It must be earned as well. This picture of science is more complicated, also more difficult to explain to the public than the simple picture of just objectively representing reality. In the end, though, it is a more honest picture.

Acknowledgments

This research was supported by the DFG Cluster of Excellence “Machine Learning—New Perspectives for Science,” EXC 2064/1, project number 390727645. I am most grateful to Judea Pearl for 30 years of partnership. We never did joint work, but we extensively worked at common themes, to my great benefit. Concerning this paper, I am indebted to Eric Raidl for the support and many helpful comments.

References

- H. Beebe. 2006. *Hume on Causation*. Routledge, London.
- P. Gärdenfors. 1988. *Knowledge in Flux: Modeling the Dynamics of Epistemic States*. MIT Press, Cambridge, MA. DOI: <https://dx.doi.org/10.2307/2275379>.
- M. C. Galavotti. 2005. *Philosophical Introduction to Probability*. CSLI Publications, Stanford. DOI: <https://doi.org/10.1007/s10670-007-9083-9>.
- D. Gillies. 2000. *Philosophical Theories of Probability*. Routledge, London. DOI: <https://doi.org/10.4324/9780203132241>.
- C. Glymour. 2004. Critical notice on: James Woodward, making things happen. *Br. J. Phil. Sci.* 55, 779–790. DOI: <https://doi.org/10.1093/bjps/55.4.779>.
- C. G. Hempel and P. Oppenheim. 1948. Studies in the logic of explanation. *Phil. Sci.* 15, 135–175. DOI: <https://doi.org/10.1086/286983>.
- D. Hume. 1739. *A Treatise of Human Nature*, Vol. I: Of the Understanding. (Page numbers refer to the edition of L.A. Selby-Bigge, Oxford: Clarendon Press 1896.)
- D. Hume. 1740. *An Abstract of 'A Treatise of Human Nature.'* DOI: <https://doi.org/10.1093/oseo/instance.00046221>.
- I. Kant. 1781. *Kritik der reinen Vernunft*.
- D. K. Lewis. 1973a. *Counterfactuals*. Oxford, Blackwell. DOI: <https://doi.org/10.2307/2273738>.
- D. K. Lewis. 1973b. Causation. *J. Phil.* 70, 556–567. DOI: <https://doi.org/10.2307/2025310>.
- D. K. Lewis. 1980. A subjectivist's guide to objective chance. In R. C. Jeffrey (Ed.), *Studies in Inductive Logic and Probability*, Vol. II. University of California Press, Berkeley, 263–293.
- J. Pearl. 1988. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, San Mateo, CA.
- J. Pearl. 2000. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, Cambridge.

- J. Pearl and D. MacKenzie. 2018. *The Book of Why: The New Science of Cause and Effect*. Basic Books, New York.
- A. Reutlinger, G. Schurz, and A. Hüttemann. 2019. *Ceteris paribus* laws. In E. Zalta (Ed.), *Stanford Encyclopedia of Philosophy*. <http://plato.stanford.edu/entries/ceteris-paribus/>.
- G. Schurz. 2019. *Hume's Problem Solved: The Optimality of Meta-Induction*. MIT Press, Cambridge, MA.
- P. Spirtes, C. Glymour, and R. Scheines. 1993. *Causation, Prediction, and Search*. (2nd. ed.). Springer, Berlin. 2000, MIT Press, Cambridge, MA.
- W. Spohn. 1978. *Grundlagen der Entscheidungstheorie*. Kronberg/Ts.: Scriptor, out of print, pdf-version: <http://www.uni-konstanz.de/FuF/Philo/Philosophie/Mitarbeiter/spohn.shtml>.
- W. Spohn. 1980. Stochastic independence, causal independence, and shieldability. *J. Phil. Log.* 9, 73–99. DOI: <https://doi.org/10.1007/BF00258078>.
- W. Spohn. 2001. Bayesian nets are all there is to causal dependence. In M. C. Galavotti, P. Suppes, and D. Costantini (Eds.), *Stochastic Dependence and Causality*. CSLI Publications, Stanford, 157–172.
- W. Spohn. 2012. *The Laws of Belief: Ranking Theory and Its Philosophical Applications*. Oxford University Press, Oxford.
- W. Spohn. 2014. The epistemic account of ceteris paribus conditions. *Eur. J. Philos. Sci.* 4, 385–408. DOI: <https://doi.org/10.1007/s13194-014-0093-6>.
- R. C. Stalnaker. 1984. *Inquiry*. MIT Press, Cambridge, MA.
- B. C. van Fraassen. 1989. *Laws and Symmetry*. Oxford University Press, Oxford. DOI: <https://doi.org/10.1093/0198248601.001.0001>.
- J. Woodward. 2002. There is no such thing as a *ceteris paribus* law. *Erkenntnis* 57, 303–328. DOI: <https://doi.org/10.1023/A:1021578127039>.
- J. Woodward. 2003. *Making Things Happen: A Theory of Causal Explanation*. Oxford University Press, Oxford. DOI: <https://doi.org/10.1093/0195155270.001.0001>.

Editors' Biographies

Hector Geffner



Hector Geffner was born and grew up in Buenos Aires, Argentina. He received a BSc in Electrical Engineering from the Universidad Simón Bolívar, Caracas, in 1983, and a PhD in Computer Science from UCLA in 1989. He then worked at the IBM T.J. Watson Research Center in New York for 2 years, and at the Universidad Simón Bolívar, in Caracas, Venezuela. Since 2001, Hector has been a researcher at the Catalan Institute of Advanced Research (ICREA) and a professor at the Universitat Pompeu Fabra, Barcelona. Currently, he is also a Guest Wallenberg Professor at the University of Linköping, Sweden.

Hector's major research interests are in computational models of reasoning, action, learning, and planning. His research group pioneered a number of ideas in these areas including the formulation of planning as heuristic search, the formulation of goal recognition as planning, and formulations and methods for learning generalized plans. Currently, he is leading a project on representation learning for planning (RLeap) which is funded by an advanced grant from the European Research Council and that is aimed at bringing together model-free and model-based methods for acting and planning.

Hector is a Fellow of AAAI and EurAI, and was an associate editor of the *Artificial Intelligence* journal and of the *Journal of Artificial Intelligence Research*.

Among the awards he has received, he won best paper awards for papers published at JAIR and at the planning conference (ICAPS), three ICAPS Influential Paper Awards, and the 1990 ACM Dissertation Award for a thesis supervised by Judea Pearl. He is the author of the book *Default Reasoning: Causal and Conditional*

Theories (MIT Press, 1992), and with Blai Bonet of *A Concise Introduction to Models and Methods for Automated Planning* (Morgan and Claypool, 2013). Together with Rina and Joe, Hector edited another book for Judea: *Heuristics, Probability, and Causality: a Tribute to Judea Pearl* (College Publications, 2010). Hector teaches courses on AI, logic, and social and technological change.

Rina Dechter

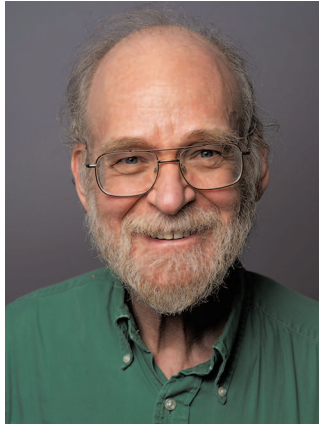


Rina Dechter is a Distinguished Professor of Computer Science at University of California, Irvine. She holds a PhD in Computer Science from UCLA (1985), an MS in Applied Mathematics from the Weizmann Institute, Rehovot, Israel (1975), and a BS in Mathematics and Statistics from the Hebrew University in Jerusalem (1973).

Dechter's research centers on computational aspects of automated reasoning and knowledge representation including search, constraint processing and probabilistic reasoning. She is the author of *Constraint Processing* published by Morgan Kaufmann (2003), and of *Reasoning with Probabilistic and Deterministic Graphical Models: Exact Algorithms* published by Morgan and Claypool Publishers (2013, second ed. 2019). She has authored and coauthored almost 200 research papers.

Dechter is a Fellow of the Association for the Advancement of Artificial Intelligence (AAAI) (1994), of the Association for Computing Machinery (ACM) (2013), and the American Association for the Advancement of Science (AAAS) (2021), and was a Radcliffe Fellow from 2005 to 2006. She received the Association of Constraint Programming (ACP) Research Excellence Award (2007), and a Classic AI Paper Award for a paper co-authored by Itay Meiri and Judea Pearl. She served as a co-editor-in-chief of *Artificial Intelligence* from 2011 to 2018. She also served on the editorial boards of *Artificial Intelligence*, *Constraints Journal*, the *Journal of Artificial Intelligence Research*, and the *Journal of Machine Learning Research*. Dechter served as program chair or conference chair of several conferences including Constraint Programming in 2000, AAAI in 2002, and Uncertainty in AI (UAI) in 2006 and 2007. She is the conference chair-elect of International Joint Conference on Artificial Intelligence (IJCAI)-2022.

Joseph Halpern



Joseph Halpern received a BSc in Mathematics from the University of Toronto in 1975 and a PhD in mathematics from Harvard in 1981. In between, he spent 2 years as the head of the Mathematics Department at Bawku Secondary School, in Ghana. After a year as a visiting scientist at the Massachusetts Institute of Technology (MIT), he joined the IBM Almaden Research Center in 1982, where he remained until 1996, also serving as a consulting professor at Stanford. In 1996, he joined the Computer Science Department at Cornell University, where he is currently the Joseph C. Ford Professor and was department chair from 2010 to 2014.

Halpern's major research interests are in reasoning about knowledge and uncertainty, security, distributed computation, decision theory, and game theory. Together with his former student, Yoram Moses, he pioneered the approach of applying reasoning about knowledge to analyzing distributed protocols and multi-agent systems. He has coauthored five patents, three books (*Reasoning About Knowledge*, MIT Press 2003; *Reasoning about Uncertainty*, MIT Press 2003; and *Actual Causality*, MIT Press 2016), and over 360 technical publications.

Halpern is a Fellow of the AAAI, the American Association for the Advancement of Science (AAAS), the American Academy of Arts and Sciences, ACM, IEEE, the Game Theory Society, the National Academy of Engineering, and the Society for the Advancement of Economic Theory (SAET). Among other awards, he received the Kampe de Frier Award in 2016, the ACM SIGART Autonomous Agents Research Award in 2011, the Dijkstra Prize in 2009, the ACM/AAAI Newell Award in 2008, and the Godel Prize in 1997, and was a Guggenheim Fellow from 2001 to 2002, and a Fulbright Fellow from 2001 to 2002 and 2009 to 2010. Two of his papers have won best-paper prizes at the IJCAI (1985 and 1991), and another two received best-paper awards at the Knowledge Representation and Reasoning Conference (2006 and 2012). He was editor-in-chief of the *Journal of the ACM* (1997–2003) and has been program chair of a number of conferences, including the Symposium on Theory in Computing (STOC), Logic in Computer Science (LICS), UAI, Principles of Distributed Computing (PODC), and Theoretical Aspects of Rationality and Knowledge (TARK). He started and continues to be the administrator of CoRR, the computer science section of arxiv.org.

Index

- A* algorithm, 108
- AAAI. *See* Association for the Advancement of Artificial Intelligence (AAAI)
- Abduction, 327, 692
- Absolute notion of causation, 871
- Actions, 321–325, 327
- Active learning, 600
- Actual causation, 625–626
 - causal models and but-for causation, 626–631
 - intransitivity and overdetermination, 634–636
 - Pearl’s achievement, 642–643
 - Pearl’s definitions, 637–642
 - pre-emption and Lewis, 631–634
- Actual causes
 - desert traveler, 359–360
 - from necessity and sufficiency to, 354
 - singular sufficient causes, 356–359
 - structural information, 354–356
 - sufficiency and necessity given forensic reports, 361–364
- Acyclic directed mixed graphs (ADMGs), 724, 823, 847
 - latent projection, 835–837
 - nested Markov factorization of, 841
- Acyclic model, 516
- Add-list, 114
- Additive noise model, 781
- ADJ(y, z) condition, 115–116
- Adjacency, 225
- Adjustment, detecting heterogeneity through, 492–494
- Admissibility, 108
 - heuristic, 108
- Admissible heuristics, mechanical generation of, 114–117
- Admissible sequence, recoverability in absence of, 418–419
- Admissible sets, 486–488
- Adversarial attacks, 786
- Adversarial vulnerability, 769, 786–787
- Agnostic causal model, 846
- AI. *See* Artificial intelligence (AI)
- Algebraic reduction method, 307
- Algorithmic independence, 778
- Algorithmic information theory, 778
- Algorithms, 611–612
- Alpha–Beta procedure (α - β procedure), 85
 - branching factor of, 85–88
- α - β pruning algorithm, 91
 - analysis, 94–101

- analytical results, 93–94
- binary game tree, 93
- evaluation of α - β , 97–101
- informal description, 92–93
- integral formula for $N_{n,d}$, 94–96
- Analogical models, 114
- Analysis by synthesis, 787
- Ancestral graph, 559
- Anticipatory node, 135, 160
- Arbitrary equation, 351
- Arbitrary graph, 194
- Arithmetic peculiarity, 403
- Armistead’s critique, 408–409
- Artificial intelligence (AI), 4, 59, 125, 318, 767
 - community, 217
 - renaissance, 600
 - researchers, 130
 - revolution, 766
- Association, 191
- Association for the Advancement of Artificial Intelligence (AAAI), 11, 396
 - proceedings, 12
- Associationist view, 868
- “Assumption-based” reasoning, 169
- Assumptions, 310, 452–454
- Asymptopia, 599
- Asymptotic behavior, 63
- Atomic intervention, 261
- Attention mechanism, 791
- Attrition, 422–423
- Augmented DAGs, 566–567
- Automatic problem-solvers, 117
- Autonomous propagation as
 - computational paradigm, 148–151
- Autonomy, 672
- Average controlled direct effect, 381
- Average indirect effect, 386–387
- Average natural direct effect, 382
- Back-door criterion, 258, 262–264, 282, 302–303
- Balke, Alexander, 12
- Baseline bias, 489
- Bayes inference, 131
- Bayes network, 145–146, 240
 - structure, 147
 - topology, 145
- Bayes’ rule, 34, 164, 598
- Bayesian networks (BNs), 3, 6–7, 33–34, 50–51, 125–126, 144, 217–218, 249, 534, 692, 824
- Bayesian probabilities, 878
- Bayesian statistics, 880
- Bayesian tree, 132
- Bayesian updating, 34
- “Bayesian” approach, 598
- Bayesianism, 616–617
- BDDs. *See* Binary decision diagrams (BDDs)
- Belief propagation, 6
 - data fusion, 154–156
 - flow of belief, 162–163
 - propagation mechanism, 157–162
 - properties of updating scheme, 163
 - in trees, 151
- Beliefs, 130
 - maintenance architecture, 130
 - networks, 125, 141–145
- Bell Labs, 30
- Berkson’s paradox, 407, 460*n*9
- Bi-directional link, 227
- Big Bang, 873
- Big data revolution, 766
- Binary decision diagrams (BDDs), 700

- Blocking, 407
- BNs. *See* Bayesian networks (BNs)
- Boltzman machines, 168, 173
- Bonawitz, Elizabeth, 600
- Book of Why, The* (Pearl), 36–37, 43–47, 805
- Bottom-up inferences, 151
- Bottom-up propagation, 135, 159
- Bounding function, 106
- Brak, Bnei, 29
- Branching factor, 70, 92, 94
 - of alpha–beta (α - β) procedure, 85–88
 - of SOLVE algorithm, 73, 75
- Brooklyn Polytechnic Institute, 30
- But-for causation, 626–631

- Calculus of intervention, 265
 - causal inference by surrogate experiments, 269
 - inference rules, 265–267
 - preliminary notation, 265
 - symbolic derivation of causal effects, 267–268
- CAR. *See* Coarsened at random (CAR)
- Carey, Susan, 594
- Causal analysis, 614–616
- Causal Bayes nets, 595, 855
 - collective causal models, 861–863
 - community of knowledge, 859–861
 - core properties, 856–859
 - human conception of causality, 854–856
- Causal Bayesian Network (CBN), 514, 537–539
 - CBN-Semi-Markovian, 546–547
 - cross-layer inferences through CBNs with latent variables, 547–551
 - with latent variables, 545–547
- Causal blocking, 607
- Causal calculus, 458–460
- Causal DAG, 814
- Causal diagram, 256–257, 537, 542–543, 825
 - calculus of intervention, 265–269
 - confounding bias, 262–265
 - formal semantics of, 258
 - graphical models and manipulative account of causation, 258–262
 - graphical tests of identifiability, 269–275
- Causal Diagrams for Empirical Research’ (Pearl), 218, 282–313
- “Causal discovery” methods, 565, 682
- Causal effects, 261–262, 267, 457–458
 - recoverability of, 444–447
 - symbolic derivation, 267–268
 - transportability of, 471–475
- Causal factorization, 772–773, 776
- Causal graphs, 577–579, 655, 774, 814
 - missingness graphs, 656–657
 - recoverability, 658–664
 - testability, 664–666
- Causal hierarchy, 514–524
- Causal Hierarchy Theorem (CHT), 528–533
- Causal inference, 255, 258, 287, 511, 533, 860
 - with interference between units, 650–651
 - via \mathcal{L}_2 -constraints, 535–551
 - logical foundations of, 454–461

- notation, 514
- for ODE-based systems, 681–682
- with panel data, 648–649
- Pearl hierarchy, 524–551
- roadmap, 512–514
- structural causal models and
 - causal hierarchy, 514–524
- successful and unsuccessful, 296–297
- by surrogate experiments, 269
- Causal information, 510
- Causal irrelevance property, 823
- Causal kinetic models, 677
 - causal kinetic models with driving noise, 678–679
 - causal kinetic models with measurement noise, 677–678
 - causal models for dynamical systems and related work, 681
 - interventions, 679–680
- Causal learning, 772, 789
- Causal Markov condition (CMC), 538n30, 771
- Causal mechanisms, 772
- Causal modeling, 765
- Causal models, 222, 293–295, 321–325, 626–631, 705, *See also* Structural causal models (SCMs)
 - applications to synthesis of, 231–234
 - associated with DAGs, 824
 - embedded, 227–231
 - as inference engines, 454–456
 - less restrictive model, 826–827
 - levels of, 773–774
 - methods driven by independent and identically distributed data, 769–771
 - non-parametric structural equations with independent errors, 825–826
 - patterns of, 224–226
 - SCM, 771–773
 - from statistical to, 769
- Causal network, 144
- Causal power, 320, 338
- Causal queries, recovering, 420–422
- Causal reasoning, 15–16, 805–807
- Causal representation learning, 790
 - learning disentangled representations, 792–793
 - learning interventional world models and reasoning, 793
 - learning transferable mechanisms, 790–791
- Causal Revolution, 46, 652
- Causal sufficiency, 355
- Causal theory, 222, 246
- Causality, 6, 15, 51–53, 145, 169–173, 600, 608–609, 671–672, 765, 878
 - human conception of, 854–856
 - from invariance to causality and generalizability, 682–683
 - relating causality to traditional statistical philosophies and “objective” statistics, 616–618
 - theory, 617
- Causality* (Pearl), 13, 36, 219, 400, 593, 597
- Causality into statistics, 612–613
- Causally interpretable structured tree graph models (CISTG models), 817

- Causation, 35, 191, 217, 310–311, 318, 339, 607, 805, 867–868, *See also* Actual causation
 in 20th-century statistics, 613–614
 contemporary theories of, 869
 laws, 874–878
 model relativity of, 871–874
 notion of, 870–871
 probability, 878–886
- Cause-effect, 310
 discovery, 780–782
 puzzle, 39
 relationships, 35, 40, 255
- CBN. *See* Causal Bayesian Network (CBN)
- CDE. *See* Controlled direct effect (CDE)
- Centrally organized architecture, 170
- Ceteris paribus laws (cp laws), 876
- Chain-rule formula, 142
- Chemical reaction networks and ODEs, 675–677
- Child machine, 13
- CHT. *See* Causal Hierarchy Theorem (CHT)
- Cigarettes, 39
- CISTG models. *See* Causally interpretable structured tree graph models (CISTG models)
- City-block distance, 105
- CLEAR (predicate), 115
- CLEAR(z) condition, 115–116
- CLOSED node, 108
- Closed path, 578
- “Closest world” approach, 238
- Cluster assumption, 785
- CMC. *See* Causal Markov condition (CMC)
- Co-training theorem, 786
- Coarsened at random (CAR), 423
- Cognitive development, 593, 600
- Cognitive enterprise, 868
- Cognitive Systems Laboratory, 4
- Collapse, 528
- Collective causal models, 861–863
- Collider, 578
- Common Cause Principle*, 770
- Community of knowledge, 859–861
- Commutativity, 119
- Compact ranking, 212
- Compatibility, 538n29
- Competitive training, 791
- Complete case analysis, 659
- Completed pattern, 226
- Completeness, 141–142
 proofs, 847–848
- Computation, 510
- Conceptual analysis. *See also* Formal analysis
 descriptive interpretation of indirect effects, 378–380
 descriptive vs. prescriptive interpretation, 376–377
 direct vs. total effects, 375–376
 policy implications of descriptive interpretation, 377–378
- Concomitants, 262, 289–290
- Conditional dependence, 141
- Conditional entailment, 210–211
- Conditional ignorability, 47, 550n39
- Conditional independence, 5, 132, 145–148, 169–173, 258–259, 534
 in probability theory, 191
- Conditional interventional distributions, identification of, 841–842

- Conditional path-specific
 - distributions, 753–754
- Conditional probability, 140–141, 143, 150
- Conditioning, 168
- Confounded component, 543–545
- Confounding, 319, 434, 488–489
- Confounding bias, 262
 - back-door criterion, 262–264
 - front-door criteria, 264–265
- Connectedness, 191
- Consequence relations, 204–206, 213
- Consistency, 108–109, 141–142, 202, 823
 - property, 822
- Constraint-propagation mechanisms, 150
- Constructivism, 594
- Context-specific independence using SWIGs, 842–844
- Controlled direct effect (CDE), 380–381, 715
 - in river blindness studies, 719–722
- Controlled distribution, 458
- Controlled effect, 376
- Coronary Drug Project, 297
- Correlational graphoids, 198
- Counterfactual queries, 240
 - evaluating, 245–248
 - linear-normal models, 250–252
 - notation, 240–241
 - party example, 241–242, 248–250
 - probabilistic vs. functional specification, 242–245
- Counterfactuals, 19, 23–24, 304–306, 321–325, 457–458
 - analysis, 304
 - empirical content of, 365–368
 - form, 365
 - formalism, 820
 - interpretation of counterfactual antecedents, 240
 - model, 859
 - sentence, 237
 - structural origin of, 503–505
 - training, 792
 - triumph, 25
- Counternomologicals, 875
- Covariate adjustment, 581
- Covariate-induced heterogeneity, 485
 - assessing, 485–486
 - special cases, 486–488
- Covariate-specific methods, 484
- cp laws. *See* Ceteris paribus laws (cp laws)
- Cross-layer inferences through CBNs with latent variables, 547–551
- Cybernetic governance mechanisms, 768
- D-map. *See* Dependency map (D-map)
- d-separation, 259, 278, 458–460, 559
 - criterion, 223, 714, 819
 - incompleteness in twin networks due to deterministic relations, 845–846
- d* separation, 421
- DAGs. *See* Directed acyclic graphs (DAGs)
- Dangerous graphs, myth of, 312–313
- Daniel Pearl Foundation, 8
- DARPA machine common sense program, 600
- Darwiche, Adnan, 12
- Data, 611–612, 868
 - data-driven causal methods, 794
 - data-driven sciences, 672

- data-generating model, 435
- data-node, 135
- data-sharing philosophy, 476
- fusion, 154–156
- node, 160
- Data Fusion theory, 396
- Dawid, Phil, 7
- Decision analysis, 606, 808–810
- Decision makers, 807
- Decision theory, 807–810
- “Decision-theoretic” approach, 565
- Decomposability, 117
- Deduction, 327
- Deep learning, 34, 38, 600
- “Deep reinforcement” learning, 600
- DeepMind, 39
- DeepQ agent, 788
- Defense mechanisms, 770
- Degree, 205–206
- Delete-list, 114
- Dependence, 141, 310–311
- Dependency equivalence, 224
- Dependency map (D-map), 193–194
- Descriptive interpretations, 374–377
 - of indirect effects, 378–380
 - policy implications, 377–378
- Desert traveler, 359–360
- Design of Experiments* (Fisher), 614
- Deterministic algorithm, 69–70
- Deterministic causal kinetic model, 677
- Deterministic necessity, 357
- Deterministic SCM, 672
- Deterministic sufficiency, 357–358
- Deterministic systems, 693–694
- Developmental psychology, 593–594
- Diagnostic rules, 5
- Diagram, 814
- Differential equation, 773
- Differential treatment effect bias, 489
- Digital goods, 768
- Digital revolution, 766
- Dimensionality of augmented space, 692
- Direct effects, 25, 374–376
- Direct transportability, 470
- Directed acyclic graphs (DAGs), 142, 222–223, 246, 404, 557, 649, 656, 814
 - Dag-isomorphic distribution, 231
 - doing, 564–569
 - imagining, 569–571
 - ladder of causation, 558–559
 - model, 824
 - properties, 226
 - seeing, 560–564
 - syntax, 559–560
 - use of, 222, 259
- Directed graph models, 285
- Directional algorithm, 70, 73
- Disentangled factorization, 774
- Distributed hierarchical approach
 - combining top and bottom evidences, 132–133
 - definitions and nomenclature, 131
 - propagation of information through network, 134–135
 - properties of updating scheme, 136
 - structural assumptions, 131–132
 - summary of proofs, 136–137
 - token game, 135–136
- Divide-and-conquer’ principle, 118
- Do calculus, 7, 12, 405–406, 514, 548–551, 596, 814, 819, 833
 - evolution, 19
 - rules of, 460–461

- Do*-formalism, 672
- Doing, 519–522, 564
 - augmented DAGs, 566–567
 - downsizing and upsizing, 568
 - empirical assessment, 567–568
 - functional intervention DAGs, 568–569
 - intervention DAGs, 565–566
- Double modifiable structural model, 859
- Downsizing, 563–564, 568
- Driving noise
 - causal kinetic models with, 678–679
 - SCMs with, 673–674
- Dummy node, 160

- Eberhardt, Frederick, 600
- Edge expanded graph, 739
- Effect identifiability, 547–548
- Effectiveness, 822
- 8-puzzle problem, 109–110, 114
- 8-queens problem, 108
- Electronic Memories, 31
- EM algorithm. *See* Expectation Maximization algorithm (EM algorithm)
- Embedded causal models, 223–224, 227–231
- Embedded pattern, 228
- Emotion Machine, The* (Minsky), 40–41
- Empirical assessment, 562
- Encoder, 793
- “End-means” strategy, 118
- Energy, 766, 767
 - revolution, 767
- Entangled factorizations, 772
- Epidemiology, 318
- Epistemic rationality, 885

- Equivalence, 231
 - of counterfactual and structural analyses, 306–308
- “Equivalence and synthesis of causal models” paper (Verma), 218
- Error variables, 557–558
- Evidence, 140, 153
- Evidential reasoning, 151
- Excess-risk-ratio, 337
- Exclusion, 579
 - restriction, 291, 306, 454
- Exogeneity
 - bounds and basic relationships under, 334–336
 - identifiability under, 336–339
- Exoplanet detection, 782–783
- Expanded graph
 - direct and indirect effects via, 731–738
 - identification of cross-world nested counterfactuals of DAG \mathcal{G} under FFRCISTG model for, 742–747
 - interventional interpretation of PDE under, 726–730
 - for single treatment, 738–739
- Expectation Maximization algorithm (EM algorithm), 703
- “Experience replay”, 769
- Experimental identification, 382–383
- “Experimental” distribution, 458
- Expert system, 5*n*1
- “Explaining-away” phenomenon, 5
- Explanation, 24
- Extended *g*-formula, 836
- Extended ID algorithm, 836, 839–842
- External data
 - recoverability with, 440–444
 - recoverability without, 437–440

- External Validity*, 395
- External validity, 452
 - formalizing transportability, 465–470
 - inference across populations, 461–465
 - preliminaries, 454–461
 - threats vs. assumptions, 452–454
 - transportability of causal effects, 471–475
- Factorization
 - associated with SWIG global Markov Property, 831–833
 - implied by semi-Markov condition, 545
- Fair coin toss, 342–344
- Faithfulness condition, 285–286, 780
- Faraday’s law, 30
- FFRCISTG model. *See* Finest fully randomized CISTG model (FFRCISTG model)
- FFRCISTG models. *See* Finest fully randomized causally interpretable structured tree graph models (FFRCISTG models)
- Finest fully randomized causally interpretable structured tree graph models (FFRCISTG models), 713–714, 725
 - identification of cross-world nested counterfactuals of DAG \mathcal{G} under, 742–747
 - proof of PDE bounds under, 754–755
- Finest fully randomized CISTG model (FFRCISTG model), 818, 819
 - NPSEM with, 826–827
 - SWIG representation of defining FFRCISTG assumptions, 833
- Firing squad, 344–346
- Fixed conditional-probability matrix, 165
- Ford, Martin, 29
 - interview by, 29–42
- Formal analysis
 - controlled direct effects, 380–381
 - natural direct effects, 381–385
 - natural indirect effects, 386–388
 - notation, 380
 - path-specific effects, 388–390
- Formal semantics of causal diagrams, 258
- 4-tuple topology, 181
- Frequentism, 616, 881
- Front-door criteria, 264–265
- Functional DAGs, 562
- Functional intervention DAGs, 568–569
- Functional specification, 242–245
- Fundamental laws of physics, 767
- Fusion*, 126
- Fusion, 148
 - data fusion, 154–156
 - equations, 164–166
- g*-do operator, 820
- G*-computation algorithm, 286
- Galles, David, 23
- Game searching methods, 93
- Game trees
 - with arbitrary distribution of terminal values, 65–69
 - solving, testing, and evaluating, 75–78
- Garden of Eden, 805–808
 - back again in, 810

- Gaussian errors for pedigree analysis, 814
- Geffner, Hector, 7
- Gelman, Susan, 594
- General ordered factorization, 417
- Generalizability, 682–683
- Generalization, 452, 741–742, 766
- Generalized conditional ignorability rule, 834
- GeNIe, 561
- Genuine cause, 233–234
- Geographical information systems (GIS), 652
- Gestalt psychology, 788
- GIS. *See* Geographical information systems (GIS)
- Global Markov property, 824
- Glymour, Clark, 22, 595, 597
- Golden Theorem, 881
- Goodman, Noah, 597–598
- Gotlieb, Kelly, 11
- Graph separability, 145–148
- Graph-induced graphoids, 196–197
- Graphical approach, 47
- Graphical causal models, 647–648, 815
 - applications, 648–651
- Graphical counterfactual models, 713
- Graphical formalism, 285–286
- Graphical identification criterion, 384–385
- Graphical inference rules, 293
- Graphical models, 767, 823
 - and manipulative account of causation, 258–262
 - theory of, 404–405
- Graphical tests of identifiability, 269
 - identifying models, 271–273
 - nonidentifying models, 273–275
- Graphoids, 195–196
 - and open problems, 196–198
 - probabilistic dependencies and graphical representation, 192–195
 - theory of, 126, 191
- Graphs, 190, 258–259, 304–306, 820
 - as models of interventions, 259–262
- Greedy algorithm, 116, 119
- Griffiths, Tom, 597–599
- Half-sibling regression, 782–783
- Halpern, Joe, 23
- HEARSAY system, 130
- Hernan, Miguel, 46
- Heterogeneity, 483–484
 - assessing heterogeneity in structural equation models, 503–506
 - covariate-induced heterogeneity, 485–488
 - latent heterogeneity between treated and untreated, 488–490
 - in recruitment, 495–497
 - three ways of detecting, 490–495
- Heuristics*, 31
- Heuristics, 49–50, 109–114
 - goal tree in 8-puzzle, 104
 - mechanical generation of admissible heuristics, 114–117
 - program, 117–119
 - properties, 107–109
 - search, 59
 - uses, 103–111
- Heuristics: Intelligent Search Strategies for Computer Problem Solving*, 4

- Hidden causes, 148, 180–181
- Hidden variable causal models, 830
 - extended ID algorithm, 839–842
 - identification in, 835
 - identification of conditional interventional distributions, 841–842
 - identified splitting operation in SWIG, 837–839
 - latent projection ADMGs, 835–837
 - representing context-specific independence using SWIGs, 842–844
- Hierarchical inference systems, 130
- Hooke’s law, 16–17
- Human conception of causality, 854–856
- Humans’ inferential reasoning, 140
- Hume, David, 868
- Hybrid graph, 227
- Hybridization, 793
- Hypotheses, 140
- Hypothetical interventions, 308–309
- “Hypothetical” reasoning, 169

- I-map. *See* Independency map (I-map)
- i-mapness, 538*n*29
- IBAL. *See* Integrated Bayesian Agent Language (IBAL)
- IBM, 30
- IC-algorithm. *See* Inductive causation algorithm (IC-algorithm)
- ICM. *See* Independent causal mechanisms (ICM)
- Identifiability, 263, 334, 459, 725–726
 - under monotonicity and exogeneity, 336–339
 - under monotonicity and non-exogeneity, 339–342
- Identification, 458–460, 809
 - bounds and basic relationships under exogeneity, 334–336
 - of conditional interventional distributions, 841–842
 - conditions, 331
 - definitions, notations, and basic relationships, 331–334
 - in hidden variable causal models, 835
 - theory, 714
- Identifying models, 271–273
- Ignorability, 302–303
- IID problems. *See* Independent and identically distributed problems (IID problems)
- Imagined space, 793
- Imagining, 569–571
- Imputation based methods, 423
- In-depth understanding, 172
- Incompleteness of d-separation in twin networks due to deterministic relations, 845–846
- Independence restrictions, 306
- Independency map (I-map), 193–194, 196
- Independent and identically distributed problems (IID problems), 769
- Independent causal mechanisms (ICM), 774–779
- Indirect effects, 25, 374
 - descriptive interpretation of, 378–380
- Inducing path, 228–229, 421
- Induction problem, 594
- Inductive causation algorithm (IC-algorithm), 232–234

- Industrial revolutions, 766, 768
- Inference, 700–702
 - net, 131
 - across populations, 461–465
 - rules, 265–267
- Infinitesimal analysis, 209
- Influence networks, 144
- Information, 766
 - revolution, 767
- Instrumental variable analysis (IV analysis), 575–576, 579–580
 - causal graphs, 577–579
 - qualitative analysis, 580–581
 - quantitative analysis, 581–588
- Integrated Bayesian Agent Language (IBAL), 700
- Interpretation process, 148–149
- Interval truncation, 580
- Interventional interpretation of PDE
 - under expanded graph, 726–730
- Interventional SCM, 519
- Interventionally equivalent
 - distribution, 674
- Interventionist theory of causal Bayes nets, 869
- Interventionist theory of mediation, 726
 - direct and indirect effects via expanded graph, 731–738
 - expanded graphs for single treatment, 738–739
 - generalizations, 741–742
 - identification of cross-world nested counterfactuals of DAG \mathcal{G} under FFRCISTG model, 742–747
 - interventional interpretation of PDE under expanded graph, 726–730
 - on substantive relationship between different \mathcal{G}^{ex} graphs and \mathcal{G}^{edge} , 739–741
- Interventions, 457–458, 628, 674
 - as conditionalisation, 309–310
 - DAGs, 565–566
- Intransitivity, 634–636
- Intrinsic motivation, 789
- Intuition, 5
- Invariance, 672, 783–789
 - criterion, 777n8
- Inverse Probability Weighted Methods, 423
- Israel Institute of Technology, 30
- IV analysis. *See* Instrumental variable analysis (IV analysis)
- Jeffrey’s rule, 162
- Joint distribution, 175, 264, 458
- Joint probability distribution, 19, 143, 517
- Kepler space telescope, 782
- Knowledge. *See also* Learning community of, 859–861
 - knowledge-based systems, 5
- Kolmogorov complexity, 778
- \mathcal{L}_2 connection, 539
- Ladder of causation, 558–559
- “Language of thought” probabilistic logics, 599
- Languages, 701
- Latent heterogeneity
 - extreme case, 501–503
 - between treated and untreated, 488–490

- Latent variables, 415
- Laws, 874–878
- Learning, 703–704
 - disentangled representations, 792–793
 - methods, 794
 - transferable mechanisms, 790–791
- Legal reasoning, 318
- Legal responsibility from
 - experimental and nonexperimental data, 349–351
- Lewis, David, 25, 631–634
 - chain criterion, 355
 - relation to Lewis' counterfactuals, 327
- Likelihood, 141
 - ratio, 133, 154
- Linear difference equation, 82
- Linear-normal models, 250–252
- Linearity, 288
- Listwise deletion, 659
- Local compactness, 212
- Local invertability, 353–354
- Loop-cut conditioning, 126
- Lotka–Volterra model, 676, 677
- Low-density separation assumption, 785
- Machine learning, 34, 593–594, 765–766, 782
 - causal representation learning, 790–793
 - cause–effect discovery, 780–782
 - half-sibling regression and exoplanet detection, 782–783
 - independent causal mechanisms, 774–779
 - invariance, robustness, and semi-supervised learning, 783–789
 - levels of causal modeling, 773–774
 - mechanization of information processing, 766–769
 - from statistical to causal models, 769–773
- Magic Shield of David, 668–669
- Magic square, 668
- Manhattan distance, 105
- Manifest distribution, 416
- Manipulated graph, 266
- MAR. *See* Missing At Random (MAR)
- Marginal probability, 143
- Markov assumption, 328–329
- Markov boundary, 195
- Markov chain, 147
- Markov chain Monte Carlo statistical inference algorithms (MCMC statistical inference algorithms), 51, 702
- Markov equivalent, 560
- Markov fields approach, 144
- Markov Fields theory, 194
- Markov network, 232
- Markov property, 171, 853–854, 856–857
- Markov relative, 545
- Markov relevant path, 829
- MARKOV-NET, 194
- Markovian Causal Bayesian Networks, 535–540
- Markovian models, 341–342, 385, 516
- MAX nodes, 92, 95
- Maximum Likelihood method, 423
- Maximum-entropy approach (ME approach), 208–210

- MCAR. *See* Missing Completely At Random (MCAR)
- McDonnell Foundation, 597
- MCMC statistical inference
algorithms. *See* Markov chain Monte Carlo statistical inference algorithms (MCMC statistical inference algorithms)
- ME approach. *See* Maximum-entropy approach (ME approach)
- Mean complexity of solving
(h, d, P_0)-game, 69–75
- Measurement noise
causal kinetic models with, 677–678
SCMs with, 672–673
- Mediating instruments, detecting heterogeneity through, 494–495
- Mediation, 24
formula, 713
- Mediation analysis, 25, *See also* Conceptual analysis
approaches to mediation based on counterfactuals defined in mediator, 715–726
FFRCISTG models, 713–714
interventionist theory of mediation, 726–747
path-specific counterfactuals, 747–754
path-specific distributions, 714–715
- Mediator (M), 714
approaches to mediation based on counterfactuals defined in, 715
FFRCISTG model, 723–725
PDE, 715–716
PDE and CDE in river blindness studies, 719–722
PDE identification via mediation formula under NPSEM-IE, 722
testable *vs.* untestable assumptions and identifiability, 725–726
two hypothetical river blindness treatment studies, 717–719
- Message passing scheme, 136
- Meta-analysis, 26
- Meta-transportability, 476
- Metaphorical models, 114
- Michaelis–Menten kinetics, 677
- Microscopic models, 790
- Microscopic structural equation models, 790
- MIN nodes, 92, 95
- Mind Magazine*, 12
- Mini Turing Test, 13
- Minimal labeling, 827, 834
- Minimal network theory, 859
- Minimal ranking function, uniqueness of, 211–213
- Minimum spanning tree (MST), 107
heuristic, 111
- Minimum-spanning-tree problem, 118
- Minsky, Marvin, 40
- Miraculous Analysis* (Lewis), 240
- Missing At Random (MAR), 420, 423, 655
recoverability in, 658–660
- Missing Completely At Random (MCAR), 420, 423, 655
recoverability in, 658–660
- Missing data, 423–424

- Missing Not At Random (MNAR), 424, 656
- Missingness Graphs (m-graphs), 414, 424, 656–657, 660–664
graphical representation, 657
and recoverability, 414–416
- Missingness process, 426–427
- MNAR. *See* Missing Not At Random (MNAR)
- Model, 202
model-immanent function, 873
relativity of causation, 871–874
- Modelling errors, 311–312
- Modern probabilistic programming systems, 706
- Modern production languages, 706
- Modularity, 672
advantage of, 33
feature of, 34
- Money, 766
- Monotonicity
assumption, 291
identifiability under, 336–342
- Moralization, 559–560
- MOVE(X_1, Y_1, Z_1) operator, 115
- MST. *See* Minimum spanning tree (MST)
- Multi-level causal models, 861
- Multi-task learning, 787–788
- Multidirectional propagation process, 149
- Multiply connected networks,
summary and extensions
for, 167–169
- Multivariate Gaussian, 672
- Mutilated graphs, 828
- Mutual information, 770
- n -cycle tree, 79
- NASA. *See* National Aeronautics and Space Administration (NASA)
- National Aeronautics and Space Administration (NASA), 782
- Natural direct-effect (NDE), 381–385
- Natural effect, 376
- Natural indirect effects, 386–388
- Nature, 17–18, 20–21
- NDE. *See* Natural direct-effect (NDE)
- Necessary causes, 331–342
- Necessary-and-sufficient cause, 318
- Negative causal assertions, 257
- Neighbor system, 196
- Neighborhood, 191
- Nested Markov factorization, 841
- Neural networks, 40
- Neuron diagrams, 321 n 5
- Neyman–Rubin model, relation to, 330–331
- Neyman–Rubin–Holland model, 304
- Nobel Prize of Computing, 11
- Nodes, 131, 415
processors, 134
- Noise inputs, 697
- Noisy AND gates, 145, 353
- Noisy OR gates, 145, 353
- Non-directional algorithm, 88
- Non-exogeneity, identifiability under, 339–342
- (Non-graphical) causal inference models, 817
- Non-monotonic models,
identification in, 351–354
- Non-monotonic reasoning, 126
- Non-parametric identification, 551
- Non-parametric structural equation models (NPSEMs), 293, 295, 713, 813–814, 820

- Non-parametric structural equations
 - with independent errors (NPSEM-IE), 814, 824–826
 - PDE identification via mediation formula under, 722
- Non-recoverability, 664
 - criteria for joint and conditional distributions, 419
- Nonexperimental identification, 383–384
- Nonidentifying models, 273–275
- Nonparametric analysis, 484
- Nonparametric models, assumptions in, 456–457
- Nonparametric structural equations, 259
 - with single-world independences, 826–827
- Normality, 641
- Normative decision theory, 809
- Normative theory, 808
- NPSEM-IE. *See* Non-parametric structural equations with independent errors (NPSEM-IE)
- NPSEMs. *See* Non-parametric structural equation models (NPSEMs)
- Objective probabilities, 881
- Observational distribution, 672
- Observational studies, 46
- Observationally equivalent distribution, 674
- Observed data distribution, 657
- ODEs
 - challenges in causal inference for ODE-based systems, 681–682
 - chemical reaction networks and, 675–677
- Ohm's law, 365
- Olendorf, Franz, 4
- OLS. *See* Ordinary least squares (OLS)
- 1-entailment, 205–208
- “One-hot encoding”, 700
- OPEN node, 108
- Open path, 578
- Optimal assignment problem, 107
- Optimal branching factor, 74, 77
- Ordinary differential equations, 790
- Ordinary least squares (OLS), 575
- Overdetermination, 634–636
- Parent–child distributions, 561
- Partial correlation coefficient, 198
- Partial interference assumption, 650
- Partially-directed graphs, 226–227
- Passive observational studies, 616
- Path, 578
 - models, 300
 - parameter, 577
- Path-specific counterfactuals, 747–754
 - conditional path-specific distributions, 753–754
 - edge consistent, 749–751
 - NPSEM-IE model associated with DAG, 751–752
 - potential outcome, 748
- Path-specific effects, 375, 388–390
- Patterns of causal models, 224–226
- Paz, Azaria, 126
- PCH. *See* Pearl Causal Hierarchy (PCH)
- PD. *See* Probability of disablement (PD)
- PDE. *See* Pure direct effect (PDE)

- PE. *See* Probability of enablement (PE)
- Pearl, Judea, 3, 7–8, 11–12, 29–42, 59–60, 125–126, 217–219, 395–396, 510, 512, 625
 annotated bibliography by, 49
 Bayesian networks, 6–7, 50–51
 causal, casual, and curious, 53–55
 causality, 51–53
 search and heuristics, 49–50
 teaching courses, 5
 vortex, 4, 30
- Pearl, Ruth, 7
- Pearl Causal Hierarchy (PCH), 512, 515, 526–528
 doing, 519–522
 imagining counterfactual worlds, 522–524
 seeing, 517–519
- Pearl hierarchy, 524
 graphical perspective, 533–551
 logical perspective, 524–533
- Pearl-y Cognitive Science, 598
- Pearl’s theory, 858
- Pedigree analysis, 814
- Philosophy, 867
 of science, 593–594
- Physical good, 768
- Piaget, Jean, 594
- Pilot randomized experiment, 495–496
- PN. *See* Probability of Necessity (PN)
- PNS. *See* Probability of necessity and sufficiency (PNS)
- po*-calculus, 819
- Pohl, Ira, 5
- Poincaré Equation, 98
- Point truncation, 580
 proof of adjustment as, 590–591
- Polytree, 33
- Population SWIGs, 846
- Postintervention distribution, 458
- Potential cause, 233–234
- Potential outcomes (po), 820, 833
 calculus, 833
 and identification, 833–835
 weaker causal models to, 846
- Potential response, 519–520
 variables, 244
- Practical interventions, 308–309
- Pre-emption, 631–634
- Precondition-list, 114
- Predictive form, 365
- Preferential selection, 434–435
- Preliminaries, 454–461
- Preprocessing approach, 169
- Prequential model, 284
- Prescriptive interpretations, 374–377
- Primitive predicates, 114–115
- Primitive processors, 167
- Principal Principle, 880
- Probabilistic causal model, 323–325
- Probabilistic causality, relation to, 328–330
- Probabilistic DAG, 561
- Probabilistic dependencies and graphical representation, 192–195
- “Probabilistic evaluation of counterfactual queries” paper (Balke), 218
- Probabilistic graphoids, 197–198
- Probabilistic Horn abduction, 705
- Probabilistic independence demands
 physical independence, strength of, 609–610
- Probabilistic methods, 5
- Probabilistic models, 113

- and deterministic systems, 693–694
- Probabilistic programming
 - languages, 691–692
 - causal models, 705
 - inference, 700–702
 - learning, 703–704
 - other issues, 704–705
 - possible worlds semantics, 694–700
 - probabilistic models and deterministic systems, 693–694
- Probabilistic Reasoning* (Pearl), 13
- Probabilistic Reasoning in Intelligent Systems*, 6, 217
- Probabilistic semantics, 561
- Probabilistic specification, 242–245
- Probability, 607, 612, 874, 878–886
 - distribution, 165
 - theory, 33, 140, 876n7
 - tree, 299–300
 - of winning standard h -level game tree with random WIN positions, 62–65
- Probability of causation, 319
 - examples and applications, 342–351
 - identification in non-monotonic models, 351–354
 - necessary and sufficient causes, 331–342
 - from necessity and sufficiency to “actual cause”, 354–364
 - structural model semantics, 321–331
- Probability of disablement (PD), 332
- Probability of enablement (PE), 332–333
- Probability of Necessity (PN), 319–320, 331
- Probability of necessity and sufficiency (PNS), 332, 524
- Probability of Sufficiency (PS), 319–320, 331–332
- ProbLog, 700, 705
- ProbTorch, 706
- Processor, 134
- Propagation, 148
 - autonomous propagation as computational paradigm, 148–151
 - belief propagation in trees, 151–163
 - equations, 166–167
 - of information through network, 134–135
 - in singly connected networks, 163–167
- Propensity interpretation, 881
- Proposition, proof of, 760–761
- Prospect theory, 809
- PROSPECTOR, 5
- PS. *See* Probability of Sufficiency (PS)
- Pseudo-Simula program, 698
- Psychology, 318
- Purcell, Ed, 7
- Pure direct effect (PDE), 713, 715–716, 743–747
 - detecting confounding via interventions on A and S , 759–760
 - interventional interpretation of PDE under expanded graph, 726–730
 - PDE identification via mediation formula under NPSEM-IE, 722

- proof of PDE bounds under
 - FFRCISTG model, 754–755
 - proof of PDE not identified in
 - river blindness study, 756–759
 - proof of proposition, 760–761
 - in river blindness studies, 719–722
- Purple expressions, 24
- PyProb, 706
- Qualitative analysis, 580–581
- Qualitative controlled unit-level
 - direct-effect, 380
- Qualitative unit-level indirect effect, 386
- Qualitative unit-level natural direct
 - effect, 381–382
- Quantifier, 150
- Quantitative analysis, 581
 - selection as function of mediator, 584–586
 - selection as function of
 - treatment, 581–584
 - selection on treatment and
 - unobserved confounder, 587–588
- Quantitative controlled unit-level
 - direct-effect, 380–381
- Query, 427
- Radiation effect on leukemia, 346–349
- Rajchman, Jan, 30–31
- Random variables, 288, 820
- Randomized clinical trials (RCT), 46
- Randomized trials, detecting
 - heterogeneity in, 491–492
- Ranking
 - function, 211–212
 - theory, 879n14
- Rational closure, 205
- Rational constructivism, 598
- Rational monotony, 206
- Rational monotony of admissible
 - rankings, 213
- RBN. *See* Recursive Bayesian network (RBN)
- RCA Laboratories, 30
- RCT. *See* Randomized clinical trials (RCT)
- Realistic statistical analysis, 611
- Recanting district criterion, 752
- Recoverability, 414–416, 658
 - in absence of admissible
 - sequence, 418–419
 - of causal effects, 444–447
 - without external data, 437–440
 - with external data, 440–444
 - as guide for estimation, 659–660
 - of joint distribution in MCAR
 - and MAR models, 658–659
 - in MAR and MCAR problems, 658–660
 - in MNAR problems, 660–664
- Recruitment, heterogeneity in, 495–497
- Recursive Bayesian network (RBN), 861–862
- Recursive model. *See* Acyclic model
- Recursive substitution, 820
- Reference class, 883
- Reflection, 869
- Regularity theory of causation, 869
- Regularity view, 868
- Reinforcement learning, 39–40
- Reinforcement learning (RL), 788–789
- Relaxed models, 111–112
- Relaxed $N \times N$ -puzzle, 118

- Relevance, 190–191, 579
 - boundary, 196
 - sphere, 196
- Reliable independence, 234
- Reparametrization trick, 790
- Response function, 243
- Response-function variable, 244
- Risk-difference, 337
- River blindness studies, 744–745
 - PDE and CDE in river blindness studies, 719–722
 - proof of PDE not identified in, 756–759
 - two hypothetical river blindness treatment studies, 717–719
- RL. *See* Reinforcement learning (RL)
- Road map, shortest path in, 105
- Robot planning, 107–108
- Robustness, 783–789
- Ron, Amiram, 4
- Root node, 160
- Rosenbaum, Paul, 219
- Rubin’s model, 260
 - for causal inference, 301
- Rudimentary pattern, 226
- Rule-based systems, 5
- Rule-priority relation, 210
- Run-of-the-mill probabilistic models, 858

- S*-admissibility, 471–475
- s*-Recoverability, 437–444
- Scenarios invite reversals, 404–405
- Scheines, Richard, 22, 595, 597
- Schulz, Laura, 600
- Scientific inference, 606–608
- SCMs. *See* Structural causal models (SCMs)
- SCOUT, 94
 - algorithm, 78–79, 85, 88
 - analysis of SCOUT’s expected performance, 79–85
 - asymptotic performance, 86
 - flow-chart, 80
- Screening neighborhood, 147
- Search, 49–50
- Seeing, 517–519, 560
 - downsizing and upsizing, 563–564
 - empirical assessment, 562, 564
 - functional DAGs, 562
 - qualitative structure, 560–561
 - quantitative structure, 561
- Selection bias, 396, 433, 576
 - recoverability of causal effects, 444–447
 - recoverability with external data, 440–444
 - recoverability without external data, 437–440
 - related work and contributions, 435–437
- Selection diagrams, 396, 465–468
- Selection variables, 465–468
- Selection-backdoor adjustment, 446–447
- Selection-backdoor criterion, 445–446
- Self-consciousness, 793
- Semantics, 694–700
- Semi-decomposable models, 118
- Semi-Markov relative, 545
- Semi-Markovian causal Bayes networks, 540–551
- Semi-Markovian models, 516, 814*n*2
 - revisiting locality in, 543–545
- Semi-supervised learning (SSL), 783–789

- Semi-supervised smoothness
 - assumption, 786
- SEMs. *See* Structural equation models (SEMs)
- Sensible interpretation, 400
- Sensitivity to generative process, 319
- Separable effects, 739
- Separation, 147, 560
- Sequential factorization, 417–418
 - recovering probabilistic queries by, 416–418
- Short-term success, 33
- Shortest path in road map, 105
- Simple Attrition, 422
 - recovering causal effects under, 423
 - recovering joint distributions under, 422
- Simpliciter, 871
- Simpson, Edward H., 400
- Simpson’s paradox, 46, 395, 399
 - Armistead’s critique, 408–409
 - correct decision, 405–408
 - history, 399–402
 - resolution, 402–408
 - scenarios invite reversals, 404–405
 - surprise, 403–404
- Simula program, 691, 701
- Single-world intervention graphs (SWIGs), 714, 819, 827–831, *See also* Directed acyclic graphs (DAGs)
 - context-specific independence using, 842–844
 - factorization associated with SWIG global Markov Property, 831–833
 - global Markov Property, 830
 - identified splitting operation in, 837–839
 - representation of defining FFRCISTG assumptions, 833
- Singly connected networks
 - derivation of updating rules for, 181–183
 - propagation in, 163–167
- Singular sufficient causes, 356–359
- Singular-event sufficiency, 358–359
- Social networks, 652
- Social psychology, 857
- Social sciences
 - future of causal research in, 652
 - research, 647–648
- Solution tree, 69
- SOLVE, 70–73
 - structural identity, 75
- Sparse causal shift training, 792
- Sparse mechanism shift, 776
- Specific evidence, 153
- Speech recognition, 107–108
- Spirtes, Peter, 22, 595, 597
- Spurious correlation, 171, 434
- SRI. *See* Stanford Research Institute (SRI)
- SSL. *See* Semi-supervised learning (SSL)
- STAN, 706
- Stanford Research Institute (SRI), 5*n*1
- Star distribution, 173
- Star-decomposability, conditions for, 183–185
- Star-decomposable
 - distribution, 173
 - triplets, 176–178
- State, 171
- State-approach, 107
- Statistical analysis, 288

- Statistical learning theory, 769
- Statistical probabilities, 878, 883
- Statistical structure, 778
- Statistical tasks, 395
- Statistics, 35–36, 606–607
- Stochastic causal kinetic model, 679
- Stochastic relaxation, 168
- Stochastic SCM, 673
- Strategic behavior, 787
- Stratified protocol, 223
- Strict exogeneity, 649
- STRIPS (robot-planning program), 114, 118
- Structural assumptions, 131–132
- Structural causal models (SCMs), 454*n*3, 511, 514–524, 672, 765, 771–773, 789, 813
 - with driving noise, 673–674
 - interventions, 674
 - with measurement noise, 672–673
 - time-dependent data, 675
- Structural constraints, 534
- Structural equation formalism, 820
- Structural equation models (SEMs), 300–301, 373–374, 453–454, 456–457, 626, 648, 816, 822
 - assessing heterogeneity in, 503–506
- Structural equations, 263, 304–306, 820
- Structural error, 577
- Structural information, 354–356
- Structural models, 321
 - causal models, actions and counterfactuals, 321–325
 - examples, 325–327
 - relation to Lewis' counterfactuals, 327
 - relation to Neyman–Rubin model, 330–331
 - relation to probabilistic causality, 328–330
 - semantics, 321
- Structure learning, 682, 703
- Structuring causal trees, 169
 - causality, conditional independence, and tree architecture, 169–173
 - open questions, 180–181
 - problem definition and nomenclature, 173–175
 - star-decomposable triplets, 176–178
 - tree-reconstruction procedure, 178–180
- Subjective probabilities, 878
- Submodels, 322–323
- Successors, 108
- Sufficient causes, 318, 331–342
- Superconducting supercollider of selection, 610–611
- Superconductivity, 30–31
- Surrogate endpoint, 464
- Survival of the fittest method, 13
- Susceptibility, 338
- SWIGs. *See* Single-world intervention graphs (SWIGs)
- Symbolic languages, 525–526
- Symmetric overdetermination, 635
- Syntax, 559–560
- System-Z, 126
 - conditional entailment, 210–211
 - consequence relations, 204–206
 - illustrations, 206–208
 - ME approach, 208–210
- Systematic relaxations, 112

- Technion Magazine*, 3
- Tel Aviv, 3
- Temporally aggregated time series, 790
- Tenenbaum, Josh, 597–598
- TEST, 75
 - flow-chart, 76
 - optimality, 78
 - structural identity, 75
 - superiority, 79
- Testability, 664–666
- Testable assumptions, 725–726
- Testing, 310
 - compatibility between underlying and manifest distributions, 427
- Theorem proving, 107–108
- Theory of mind, 598
- “Theory theory”, 594
- Thinking, 793
- Threats, 452–454
- Time-dependent data, 675
- Token game, 135–136
- Toleration, 202
- Top-down inferences, 151
- Top-down propagation, 135, 159
- Total effect, 374–376
- Traditional statistical analysis, 614–616
- Transparency, 33, 142
- “Transparent” revision process, 149
- Transportability, 24–25
 - of causal effects, 471–475
 - definitions and examples, 468–470
 - formalizing, 465–470
 - selection diagrams and selection variables, 465–468
- Traveling Salesman Problem (TSP), 106–107
- Treatise of Human Nature, Vol. I* (Hume), 868
- Treatment effects, 256
- Treatment on the treated, 488
- Treatment-dependent sample, 435
- Treatment-induced IV selection bias, 576
- Tree architecture, 169–173
- Tree dependence, 172
- Tree-decomposable distribution, 174
- Tree-dependent random variables, 173
- Tree-reconstruction procedure, 178–180
- Tree-structured influence networks, 151
- Trivial transportability, 470
- Trivially recoverable query, 420
- True distribution, 657
- Truncated factorization product, 539–540
- Truncation, 580
- Truncation bias expressions, 589–590
- TSP. *See* Traveling Salesman Problem (TSP)
- Turing Award Lecture (Pearl), 11
 - Bayesian nets to causality and counterfactuals, 19
 - causal calculus, 24
 - causal hierarchy and physics, 17
 - causes, counterfactuals, and sense of justice, 16
 - child machine, 13
 - counterfactuals, 23
 - logic and experiment for science of cause and effect, 26

- structured causal models and truncated factorization, 20, 22
- Turing Test and plurality of mini-Turing Tests, 15
- Turing machine, 692
- Twin network approach, 819n7
 - incompleteness of d-separation in twin networks due to deterministic relations, 845–846
- Two hypothetical river blindness treatment studies, 717–719
- Tyranny of Euphonious Monosyllable, 598
- UCLA, 126, 667
- Underlying distribution, 416
- Undirected graph, 193
- Untestable assumptions, 725–726
- Updating scheme, properties of, 136, 163
- Upsizing, 563–564, 568
- Variable-effect bias, 489
 - separating fixed-effect from, 489–490
- Verma, Thomas, 218
- Virtual evidence, 153
- Visual processing algorithms, 791
- Vortex Theory of Superconductive Memories*, 4
- Warranted inferences, 297–298
- Weaker causal models to po-calculus, 846
- Wellman, Henry, 594
- WIN-LOSS
 - assignment, 63
 - status, 59
- Winship, Chris, 7
- Woodward, James, 596
- Wright, Sewall, 36
- Xu, Fei, 598
- Z-ordering, 203
- Z-ranking, 203, 205
- 0-entailment, 204–205

Probabilistic and Causal Inference

The Works of Judea Pearl

Hector Geffner, Rina Dechter, Joseph Y. Halpern (Editors)

Professor Judea Pearl won the 2011 Turing Award “for fundamental contributions to artificial intelligence through the development of a calculus for probabilistic and causal reasoning.” This book contains the original articles that led to the award, as well as other seminal works, divided into four parts: heuristic search, probabilistic reasoning, causality, first period (1988–2001), and causality, recent period (2002–2020). Each of these parts starts with an introduction written by Judea Pearl. The volume also contains original, contributed articles by leading researchers that analyze, extend, or assess the influence of Pearl’s work in different fields: from AI, Machine Learning, and Statistics to Cognitive Science, Philosophy, and the Social Sciences. The first part of the volume includes a biography, a transcript of his Turing Award Lecture, two interviews, and a selected bibliography annotated by him.

ABOUT ACM BOOKS



ACM Books is a series of high-quality books published by ACM for the computer science community. ACM Books publications are widely distributed in print and digital formats by major booksellers and are available to libraries and

library consortia. Individual ACM members may access ACM Books publications via separate annual subscription.

BOOKS.ACM.ORG · **WWW.MORGANCLAYPOOLPUBLISHERS.COM**

ISBN 978-1-4503-9587-8



9 781450 395878