TECHNICAL REPORT
R-528
July 2023

UNIVERSITY OF CALIFORNIA
Los Angeles

Causal Analysis for Generalized Interference Problems

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Computer Science

by

Chi Zhang

2023

ABSTRACT OF THE DISSERTATION

Causal Analysis for Generalized Interference Problems

by

Chi Zhang
Doctor of Philosophy in Computer Science
University of California, Los Angeles, 2023
Professor Judea Pearl, Chair

Causal inference studies the causal relationships between factors by modeling the underlying data generating process. A common goal in causal inference research is to answer what the effects are of the treatments on the outcomes. Traditional causal inference techniques assume data are independent and identically distributed (IID) and thus ignore interactions among single units. However, a unit's treatment may affect another unit's outcome (interference), a unit's treatment may be correlated with another unit's outcome, or a unit's treatment and outcome may be spuriously correlated through another unit. Those unit-level interactions are referred to as *generalized interference*. To capture such nuances, this work proposes a graphical model, "interaction models," which can model the data generating process of data with generalized interference using causal graphs. In this work, I focus on the estimation of causal effects given data with generalized interference, and use interaction models to conduct a systematic analysis of the bias caused by different types of interactions among units. I start with assuming linearity and present the graphical framework, interaction models. The framework applies to a more general setting where interactions can occur between any units. I derive theorems to detect, quantify, and remove the interaction bias. Those results rely

on knowing the exact interaction patterns between units. Next, I show how this assumption can be relaxed and present results for when the exact interaction pattern is unknown, where bounding or unbiasedly estimating the causal effects might be possible. I then show how the interaction model framework and the bias analysis results can be generalized for non-parametric models. Finally, I will discuss a special setting where interactions only occur between separated "blocks," so non-IID data can be reduced to block-IID data.

The dissertation of Chi Zhang is approved.

Karthika Mohan

Yizhou Sun

Adnan Darwiche

Onyebuchi Arah

Judea Pearl, Committee Chair

University of California, Los Angeles

2023

To Dad and Mom

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I am forever grateful to my Ph.D. advisor, Judea Pearl, who has helped me through this Ph.D. journey. I had the pleasure to have so many inspiring research discussions with you, where you taught me not answers but questions.

I want to thank my committee members, Onyebuchi Arah, Adnan Darwiche, Yizhou Sun, and Karthika Mohan. Thank you for your support and thoughtful feedback.

Parts of this work have benefited from discussions and collaborations with Ang Li, Bryant Chen, Carlos Cinelli, Karthika Mohan, Rumen Iliev, Scott Mueller, Totte Harinen, and Yujia Shen. Thank you, Karthika, for all the guidance and patience. Thank you, Yujia, for being my "rubber duck".

Thank you Kaoru Mulvihill for assisting with the administrative matters.

Finally, I thank my family and friends for their love and support. Thank you to my friends from the UCLA CS department, who helped make UCLA my home. Thank you to the Cognitive Systems Laboratory members, who I have learned so much from. Thank you to my family for always being there for me.

VITA

2021–2023 Graduate Student Researcher, Computer Science Department, UCLA.

2022–2022 Research Intern, Toyota Research Institute, Los Altos.

2017–2020 Teaching Assistant, Computer Science Department, UCLA. Taught Intro to CS I and II, Intro to AI.

2017 B.S. (Electrical and Computer Engineering), Shanghai Jiao Tong University.

2017 B.S. (Computer Science), University of Michigan, Ann Arbor.

2014–2014 Teaching Assistant, UM-JI Joint Institute, Shanghai Jiao Tong University. Taught Honors Mathematics.

PUBLICATIONS

Zhang, Chi, Karthika Mohan, and Judea Pearl. "Causal Inference with Non-IID Data under Model Uncertainty." Proceedings of Machine Learning Research vol TBD 1 (2023): 14.

Zhang, Chi, Karthika Mohan, and Judea Pearl. "Causal Inference with Non-IID Data using Linear Graphical Models." Advances in Neural Information Processing Systems. 2022.

Zhang, Chi, et al. "Exploiting equality constraints in causal inference." International Conference on Artificial Intelligence and Statistics. PMLR, 2021.

Zhang, Chi, Bryant Chen, and Judea Pearl. "A simultaneous discover-identify approach to causal inference in linear models." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 34. No. 06. 2020.

CHAPTER 1

Introduction

1.1 Causality

Understanding cause-effect relationships is an important task in many scientific disciplines, including engineering, epidemiology, economics, social science and medicine. Engineers want to find out the reason of a defective product. Medical researchers want to find out the effect of a drug. Politicians want to find out the effect of a policy. Such questions are important in our daily lives, too. For instance, I want to know what might have caused my migraine today. Some causal questions can be answered through controlled experiments. However, if controlled experiments are not possible due to cost or ethical concerns, some causal knowledge can be obtained from past experience. For instance, if among most of the occurrences of my migraine, I stayed up late the night before, I may conjecture that staying up late was the cause of my migraine.

Humans subconsciously use causal assumptions to answer causal questions. With the aid of internet and modern computing resources, we are able to learn not only from personal experience, but also from the experience of other people across the world. However, intelligent machines usually cannot infer causal directions or derive causal conclusions by only looking at the data. This is because the data only convey correlation between migraine and staying up late, but do not imply causation without further assumptions. This limitation cannot be overcome by using more data. A recent intelligent chatbot, ChatGPT, learns from massive

text datasets including billions of words and characters[1], and can answer questions from various domains. Such technology, although much more powerful than humans in finding correlation patterns, still stumbles when it comes to causal reasoning.

Numerous causal reasoning frameworks and approaches have been developed and improved over the past three decades. Those works have provided solutions to different aspects of causal reasoning including identification (reducing causal effects to observed quantities), discovery (learning causal structures), counterfactual (answering "what if things had been different"), missing data, external validity, assessment of direct and indirect effects, etc. [Pea19]. However, one important aspect that has not been systematically studied is the violation of IID assumptions (introduced next), and is the main focus of this work.

1.2 Independent and Identically Distributed (IID)

Many of the algorithms and techniques used in empirical sciences, including causal reasoning and machine learning, rely on the Independent and Identically Distributed (IID) assumption [Sch22, Pea09, IR15, Rub78]. Data are IID when each sample is generated through the same data-generating process, and the manner in which each sample is generated is independent of other samples. The graphical model framework uses graphs to represent causal relationships among variables, and assumes IID to explain the data observed. The potential outcomes framework assumes SUTVA (Stable Unit Treatment Value Assumption), which disallows the outcome of any unit to be affected by the treatment of other units. The IID assumption is convenient because it simplifies both the the modeling process (information about interaction patterns need not be collected, stored or modeled) as well as the underlying mathematics (by facilitating tractable solutions to hard problems [Cao22]). However, IID does not hold true in many real-world datasets. Typical examples include interactions among users in social media and spreading patterns of infectious diseases such as covid or habits such as smoking

[1]The number came from the answer by ChatGPT itself.

[FSY22, CMM22, Cao14].

1.2.1 Real-World Examples of IID Violation

Coupon Effectiveness Problem: A company plans to send coupons at different discount rate for their product to customers. Suppose they want to analyze the effect of a discount rate on sales volume. Specifically, they want to know if sending a coupon of certain discount rate to an average potential customer would increase the customer's chance of purchase. Assuming that observational data are available in the form of (DiscountRate, Purchase) pairs from 50 customers. The conventional way would be to construct a causal model, $DiscountRate \rightarrow Purchase$ and estimate the causal effect of interest by regressing Purchase on DiscountRate, using the 50 data pairs.

However in reality, there might be interactions between customers. Suppose customer i received a coupon, purchased the product, and advertised it on the social network. Customer j, seeing i's post, purchased the product without receiving a coupon. In another scenario, customers k and l live together, and k received a coupon. k and l both used k's coupon for purchase. Such scenarios might make us wrongly conclude that the discount rate of the coupon is not so attractive, since many customers still purchased the product without a coupon. As a result, we would misestimate the causal effect.

Vaccination Problem: Suppose we are interested in studying the effectiveness of Covid-19 vaccines. Specifically, we are interested in the causal effects of vaccine doses, V, on the severity of sickness S. A naive method would involve building a causal model on V, S, and other related factors, and estimating the causal effect of V on S using available data. However, this method may result in inaccurate estimation primarily because individuals in the sample are not isolated from each other in the pandemic setting. Below are a few instances of IID violation ([ETP22]).

Case 1: The vaccination V of a unit i, (V_i), decreases their viral load, L_i, which in turn

decreases the transmission rate of the virus, and hence decreases the severity of sickness S of another unit j, (S_j), who comes into contact with i. *V_i causally affects S_j.*

$$V_i \to S_j$$

Case 2: V_i is affected by the area A that i lives in, and a contact j who lives in vaccine deprived areas and areas with a higher incidence of Covid-19 infection is more likely to get sick. *V_i and S_j are confounded.*

$$V_i \leftrightarrow S_j$$

Case 3: S_i, determines whether or not i is quarantined and thus affects whether i transmits the disease to another unit j. *S_i causally affects S_j.*

$$S_i \to S_j$$

Such interactions between units plague both observational and experimental studies. If the latter is performed in a controlled environment where subjects are isolated from each other, the results would not be valid for the target environment, where subjects affect each other, and vice versa. Note that this problem is also not resolved by increasing sample size.

1.2.2 Generalized Interference

One line of existing work that analyzes interactions between units is interference [Cox58]. Interference holds when treatment of unit i (discount rate of i's coupon, vaccination dose of i, ...) causally affects the outcome of another unit j (j's purchase, sickness of j, ...). This is modeled by the existence of a causal pathway from i's treatment to j's outcome. However, interference is not the only type of interaction between units that can cause biased estimates. In Case-2 above, V_i and S_j are confounded and V_i is not a cause of S_j. Another example is an instance where unit i's treatment affects their own outcome through an attribute of unit j i.e., $V_i \to W_j \to S_i$, for some $W_j \notin \{S_j, V_j\}$. In both these cases units interact

with each other in a way that might bias the estimation of causal effects although they may not typically be classified as interference. In spite of the prevalence of such interactions in applications related to health care, infectious diseases, social networks, and ad placements, they have not been systematically studied. It is this deficiency that this work attempts to overcome. In particular, I aim to model and develop methods for handling interactions not limited to interference, which I call "generalized interference."

1.3 Problem Setting and Preliminaries

The scenarios exemplified above raises several questions regarding the computation of causal effects given data with generalized interference. How can we model different types of interactions among units in the population? Under what conditions can we safely ignore unit interactions with the guarantee that assuming IID (and applying existing estimation techniques) will result in negligible bias? If assuming IID would yield a biased estimate, then how can we get rid of this bias?

1.3.1 Average Causal Effect

In this work, I am primarily interested in the estimation of ACE for data with generalized interference. Average causal effect (ACE), also named as Average Treatment Effect (ATE)[2] [Rub77, Hol88] is often used to represent the size of causal effects. The ACE of a treatment X on an outcome Y represents how much Y is expected to change if X is intervened to change from one value to another value. Given a causal model M, the average causal effect (ACE) of $X = t$ vs $X = c$ (t and c are constants) on Y for k units is defined as $ACE_{XY} = \frac{1}{k}\sum_i(Y_{iX_i=t} - Y_{iX_i=c})$. ACE is defined under the assumption that Y_i depends only on factors of unit i (including X_i) [Hol88].

[2]For consistency, I use ACE to refer to both ACE and ATE.

1.3.2 Linear and Non-Parametric Models

I will start the discussion by developing the framework for acyclic linear structural causal models (SCMs) [Wri21, Pea09], since the generalized interference problem is simpler in the linear case. Formally, linear SCMs are represented by a system of linear equations $X = \Lambda^T X + \epsilon$ where X is a vector of observed variables, ϵ is a vector of latent variables, and Λ is an upper triangular matrix of direct effects, whose ijth element, $\lambda_{v_i v_j}$ gives the magnitude of the direct causal effect of v_i on v_j. In linear models, the error term ϵ is commonly assumed to be normally distributed with covariance matrix $\mathcal{E}$. This means that the covariance matrix of the observed data $\Sigma := XX^T$ fully characterizes the observational distribution. This matrix can be linked to the underlying structural parameters through the system of polynomial equations $\Sigma = XX^T = (I - \Lambda)^{-T}\mathcal{E}(I - \Lambda)^{-1}$, and the problem of causal effect estimation reduces then to finding the elements of Λ that are uniquely determined by the above system.

For the linear case, without loss of generality, I assume $t = c + 1$[3]. In linear models, ACE of X on Y can be identified as β_{YX}, the linear regression slope of Y on X, if there is no backdoor (non-directed open paths) between X and Y [PGJ16, Pea17].

Later in the discussion, I will generalize the framework to non-parametric models. In real-world applications, relationships between variables might not be perfectly linear. For example, certain drug can interact with a vaccine to boost or weaken its performance. In a non-parametric model, the value of a variable Y is determined by the values of its parents by a function f, i.e., $Y = f(Pa(Y))$. Contrary to linear models, non-parametric models do not assume any function f needs to follow any parametric forms.

1.3.3 Partial Interference

Recent years have witnessed a rise in papers on interference that employ graphical models ([OV14], [SS18], [NPB20], etc.). These works rely on *partial interference* which divides units

[3]If $t \neq c + 1$, the ACE is multiplied by the constant $(t - c)$.

into equal-sized blocks under the assumption that interactions occur only within a block but not across different blocks. Partial interference is useful in simplifying non-IID data to "block-IID" data. Such methods do not generalize to cases where units are allowed to interact with anyone, or when each block does not have the same structure. Nonetheless, assuming partial interference strengthens causal identification power that goes beyond current state of the art. I will present a causal identification method under the partial interference setting in Chapter 6.

1.4 Thesis Outline

This thesis is organized as follows. In Chapter 2, I define generalized interference, and introduce a new graphical framework, the interaction models, for modeling generalized interference. In Chapter 3, I define the non-IID version of average causal effects, named *True Average Causal Effects* (TACE), and present theorems for generalized interference bias detection, quantification, and removal for TACE. In Chapter 4, I discuss the case where specific interaction patterns are unavailable, and present results for bounding and estimating TACE. In Chapter 5, I derive an extension of the interaction models and generalize bias analysis results to the non-parametric setting. In Chapter 6, I propose a partial-interference-based causal identification method to utilize equality information between units that can be applied to solve many diverse problems in addition to non-IID data. Finally, in Chapter 7, I recapitulate major contributions and conclude the thesis.

CHAPTER 2

Linear Interaction Models

2.1 Introduction

In this chapter, I first formally define generalized interference, which is the type of IID violation that this work focuses on. Existing causal models usually make the IID assumption, where each unit is assumed to be the "same." Such models are incapable of modeling interactions between units, which is required in order to approach the non-IID problem. I will next present a new graphical model, named *interaction models*, for this purpose. Interaction models are first developed assuming linearity, where the total effects are simple summations of different components, so that the effects can be naturally separated. In Chapter 5, I will discuss how the linearity assumption can be relaxed.

2.2 Generalized Interference

In a traditional causal model $M(G, S)$ ([Pea09], Definition 7.1.1), G is the causal graph (e.g., Figure 2.1) and S is the set of structural equations of variables (e.g., Equations 2.1-2.3)[1]. I refer to the variables in a traditional causal model as *generic variables*. X, C, Y in Figure 2.1 are generic variables. The structural equations 2.1-2.3 represent causal relationships among the variables. An *explicit variable* is similar to a generic variable except that it represents an attribute/event of one specific unit (or sample or individual). For example, "treatment

[1]In the remaining text, unless specified, the independent random error variables such as U_X, U_Y will be omitted for simplicity.

(X)" is a generic variable, and "the treatment of unit i (X_i)" is an explicit variable.

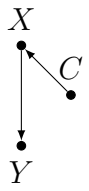


Figure 2.1: Traditional causal DAG.

$$C = U_C \tag{2.1}$$

$$X = f(C, U_X) \tag{2.2}$$

$$Y = g(X, U_Y) \tag{2.3}$$

When IID is violated, the corresponding variable (treatment/outcome/etc.) of different units in the sample may have different data generating processes. The type of non-IID violation that I focus on in this work is named as *generalized interference*, which I formally define as follows.

Definition 1 (Generalized Interference)**.** *Generalized interference between two units i, j in a sample is defined as an explicit variable of i being caused by an explicit variable of j or vice versa.*

A traditional interference is defined as the treatment of a variable causes the outcome of another variable. In generalized interference, the interaction is generalized to be between any variables of two different units, not limited to the treatment and the outcome. Such interactions usually make units non-IID. For example, if the sickness of i is affected by the viral load of j (L_j), then the sickness of i (S_i) and the sickness of j (S_j) are correlated through the common factor L_j. As a result, S_i and S_j are not independent. They are also likely not identical, unless we assume S_i is also affected by some L_k in the same way.

Note that generalized interference is not the only way non-IID can be violated. The data generating processes of the corresponding variable of two units (e.g., S_i and S_j) can be different even without interactions. Two units might have different characteristics (e.g., health conditions) that cause their corresponding variables to have different probability dis-

tributions. However, this is beyond the scope of this work, while in this work I limit the attention to non-IID caused by generalized interference.

2.3 Graphical Modeling of Generalized Interference

In this section, I define a graphical model derived from traditional causal models for modeling generalized interference.

Definition 2 (Interaction model $M^*(G^*, S^*)$)**.** *An interaction model, $M^*(G^*, S^*)$, is a causal model where G^* is the interaction network and S^* is the set of structural equations defining the data generating process of the observed explicit variables. An interaction network, G^*, is a directed acyclic graph with each node representing an explicit variable and each directed edge $A_i \to B_j$ representing A_i causes B_j.*

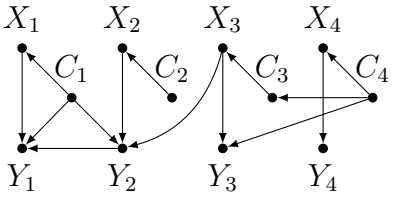


Figure 2.2: Interaction network with 4 units and 12 explicit variables (X_i, Y_i, C_i for $i = 1, 2, 3, 4$).

$$X_1 = U_{X_1} \tag{2.4}$$

$$Y_1 = X_1 + Y_2 + 3C_1 + U_{Y_1} \tag{2.5}$$

$$Y_2 = 2X_2 - C_1 + X_3 + U_{Y_2} \tag{2.6}$$

$$\dots$$

An example of interaction model $M^*(G^*, S^*)$, is the interaction network, G^*, portrayed in Figure 2.2 and the structural equations S^* (part of) specified beside it; U_{V_i} denotes the unobserved exogenous error of an explicit variable V_i. Observe that interaction networks allow edges between explicit variables of the same unit (e.g., $X_1 \to Y_1$), as well as two distinct units (e.g., $C_1 \to Y_2$).

We are now ready to define an *isolated interaction model* for an interaction model M^*. It is the "ideal" model constructed from M^* by eliminating all interactions between units.

Definition 3 (Isolated interaction model $IM^*(IG^*, IS^*)$)**.** *$IM^*(IG^*, IS^*)$ is the* Isolated interaction model *of an interaction model $M^*(G^*, S^*)$ if IM satisfies the following conditions:*

1. *$IG^* = G'$ where G' is the graph obtained by removing from G^* all edges $A_i \to B_j$, $i \neq j$,*

2. *$IS^* = S'$ where S' is the set of equations obtained by removing from each equation $X_i = f(Pa(X_i))$[2] in S^* all terms containing any Y_j, $\forall j \neq i$.*

For example, the interaction model $M^*(G^*, S^*)$ has Figure 2.2 as G^*, and Equations (2.4-2.6) as part of S^*. The isolated model for M^* is denoted $IM^*(IG^*, IS^*)$. IG^* is given in Figure 2.3 below. And IS^* for Equations (2.4-2.6) are given by Equations (2.7-2.9).

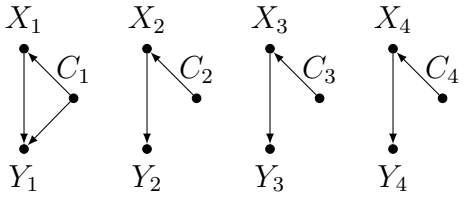


Figure 2.3: Interaction network with 4 units.

$$X_1 = U_{X_1} \tag{2.7}$$

$$Y_1 = X_1 + 3C_1 + U_{Y_1} \tag{2.8}$$

$$Y_2 = 2X_2 + U_{Y_2} \tag{2.9}$$

2.4 Symmetry Assumptions

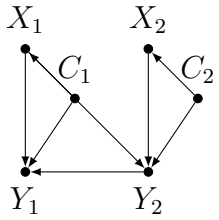
In real-world applications, we will have at our disposal limited (usually just one) observations corresponding to a unit which in turn will make it hard to draw useful conclusions if the model is completely arbitrary. In traditional causal inference techniques this is not a problem since they assume IID, which is assuming for each variable, the distribution is the same and independent for all units. In general, it is nearly impossible to obtain meaningful results given data on units that behave completely differently. So it might be reasonable to assume that the unit model for each unit would behave the same way if the units were isolated from

[2] *$Pa(X_i)$ denotes the parents of X_i in G^*.*

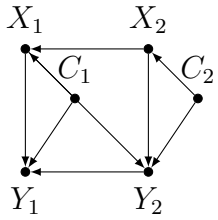
each other. While we do not make strong assumptions such as IID, we need to make weaker symmetry restrictions (definitions 4, 5), in order to quantify bias and identify ACE. We only require some of the variables are IID instead of all.

Definition 4 (Balanced interaction model $M^*(G^*, S^*)$)**.** *Let $M^*(G^*, S^*)$ be an interaction model with isolated model IM^*. M^* is a* balanced *interaction model if IM^* has the same unit-model $(IM_i^*(IG_i^*, IS_i^*))$ for every unit i.*

Let G^* be the graph in Figure 2.2 and S^* be the set of equations (2.4-2.6) corresponding to $M^*(G^*, S^*)$. IG^* in Figure 2.3 is the graph and IS^* are the equations (2.7), (2.8) and (2.9) that correspond to IM^*, which is the isolated model of M^*. The unit-graph for unit 1 is different from unit 2. Also, the structural equations for Y_1 and Y_2 of the isolated interaction model (Equations (2.8) and (2.9)) are different. Hence, M^* is not a balanced interaction model.



(a) X satisfies ASDC.



(b) X does not satisfy ASDC.

$$C_i = U_{C_i}, i \in \{1, 2\}$$
$$X_i = C_i + U_{X_i}, i \in \{1, 2\}$$
$$Y_1 = 2X_1 - C_1 + Y_2 + U_{Y_1}$$
$$Y_2 = 2X_2 - C_2 + U_{Y_2}$$

(c) Structural equations for (a)

Figure 2.4: Two balanced interaction networks and the structural equations for (a)

For another example, the interaction model M^* is balanced where G^* is the graph in Figure 2.4(a), and S^* is the set of equations given in Figure 2.4(c).

Remark 1. *Note that a balanced interaction model M^* does not imply that data generated by it are IID. Being balanced only requires all units share the same causal relationships within each unit itself, but permits interactions and effects from other units. For example, the*

parents of explicit variables Y_i and Y_j, $i \neq j$ can be different in G^* i.e., Y_i can be caused by a set of variables S_k corresponding to unit, k, and Y_j can be caused by a distinct set of variables T_k. However, for M^* to be balanced it is required that for all distinct units i and j, all Y_i have the same relationship with i's explicit variables as Y_j with j's variables.

We further note that if M^{**} is balanced then all the unit-models $IM_i^*(IG_i^*, IS_i^*)$ in Definition 4 are identical (with no edges between IG_i^* and IG_j^*), and can be succinctly represented by a (single) causal model $M(G, S)$ where G and S can be constructed from any IG_i^* and IS_i^* by replacing explicit variables with generic variables.

In addition to the assumption that the isolated components being the same, it would be helpful if we also have *symmetrical* assumptions on the underlying distributions of specific sets of variables. For example, it is reasonable to assume all units' treatments have the same distribution, i.e., for any treatment $X = x$, all units have an equal chance of getting the treatment $X = x$.

Definition 5 (Ancestral same-distribution condition (ASDC))**.** *In the interaction network G^* a balanced interaction model, generic variable W to satisfies the ancestral same-distribution condition (ASDC) if for all unit i, 1) $Pa(W_i)$ satisfies ASDC, and 2) $Pa(W_i) \subseteq \mathcal{V}_{(i)}$, and 3) for any different unit $j \neq i$, $Pa(W_i)$ and $Pa(W_j)$ have the same set of generic variables, and their exogenous errors U_{W_i} and U_{W_j} have the same distribution. (When i=j, the condition is automatically satisfied.)*

For example, in Figures 2.4(a) and 2.4(b), X satisfies ASDC in the former (assuming the condition on exogenous errors is satisfied) but not in the latter, since in the latter $Pa(X_1) \neq Pa(X_2)$. ASDC implies IID as stated in the following lemma.

Lemma 1. *If W satisfies ASDC, then any two explicit variables W_i and W_j are IID (Independent and Identically Distributed.)*

Remark 2. *The descendants of an ASDC variable need not be IID. For example, in Figure 2.4(a), X satisfies ASDC, and Y_i and Y_j are descendants of X_i and X_j. Y_i and Y_j have different sets of parents, making their distributions different, so Y is non-IID.*

Finally, I define the notion of a *generic network*, which is the expected causal DAG if all interactions are removed.

Definition 6 (Generic Network)**.** *The* generic network *for a balanced interaction model $M^*(G^*, S^*)$ is defined as the the shared unit-graph for the isolated model IM^* of M, with the nodes relabeled as the corresponding generic variables.*

For example, given the balanced interaction model M^* with interaction network as Figure 2.5 left, the generic network is defined as the causal DAG in Figure 2.5 right.

Figure 2.5: A balanced interaction network (left) and its corresponding generic network (right).

2.5 Discussion and Summary

In this chapter, I defined the notion generalized interference, which is the problem focus of this work. I defined a new graphical model, named interaction models, for handling generalized interference in data.

One of the most studied concepts related to interactions among units is interference [Cox58]. Majority of literature in empirical fields assume no-interference. In fact, SUTVA is a common assumption in causal inference [Rub78]. Recent years have witnessed a rise in

papers on interference that employ graphical models. These include [OV14] that was the first to model interference using DAGs, [SS18] that modeled interference using chain graphs which permits modeling unknown interactions between units and [BMS20] that proposed structure learning methods for chain graphs. These works rely on partial interference which divides units into equal-sized blocks under the assumption that interactions occur only within a block but not across different blocks. In addition, partial interference requires corresponding units in different blocks to satisfy the *'identical'* condition in IID. Thus partial interference methods assume "block IID," which is weaker than "unit IID" assumed by traditional causal methods. However, in many domains such as infectious diseases, it is unrealistic to assume that the samples in the dataset can be divided into blocks that satisfy the requirements for partial interference. For instance, if blocks pertain to families then all families may not have the same number of members and individuals in the family are likely to interact with people outside the family. The framework presented in this chapter is not limited to partial interference.

Some related works demonstrating application values include [NPB20], which developed methods for identification and estimation of multiple queries under conditions of interference and homophily, and applied the results to the problem of ad-placements. [Sob06] was the first to notice the effect of interference in the housing mobility problem, and proposes causal estimands for this application.

[AS17], [SA17] modeled *general interference* (without assuming partial inference) by constructing a function to define a unit's exposure level on the number of treated neighbors they have. The methods are less restricted than partial interference methods, and allow units to be affected by any number of neighbors. However, they are limited to interference and do not handle other forms of interactions.

CHAPTER 3

Interaction Bias of the Causal Effect

3.1 Introduction

In this chapter, I conduct a systematic analysis on the interaction bias resulting from using IID methods on generalized interfering data. I will start by formally define the query of interest and the interaction bias. Next, I will present theoretical results on quantifying, detecting and removing the interaction bias. Finally, I test the performance and coverage of the proposed methods through simulations and a case study.

3.2 Quantity of Interest: True Average Causal Effect (TACE)

I generalize traditional ACE to the non-IID setting. Examine the interactions depicted in Figure 3.1.

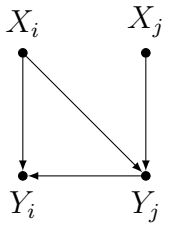


Figure 3.1: An example interaction: X_i causes Y_i through Y_j.

Unit i's treatment X_i affects its outcome through unit j's outcome Y_j. This effect is not part of the effect that is the interest of this work, since it results from interactions between

units. Hence, $X_i \to Y_j \to Y_i$ is considered a "spurious" causal path. In other words, I am interested in computing the ACE of a unit's treatment on their outcome, excluding the effects transmitted via spurious paths from its neighbors/contacts. In an experimental setting, interactions might be eliminated by isolating all subjects. In an observational setting or an experimental setting where subjects are not isolated, the data are non-IID. I am interested in computing the average causal effect of treatment on outcome *as if all units were isolated*. The formal definition of this quantity of interest is presented below.

Definition 7 (True Average Causal Effect ($TACE_{XY}$))**.** *Let M^* be an interaction model. True average causal effect of X on Y, denoted as $TACE_{XY}$, is defined as the ACE of X on Y in the isolated interaction model IM^* corresponding to M^*.*

$TACE$ is the non-IID version of ACE and is the same as ACE in a traditional causal model where all samples are isolated. Again, without loss of generality, I assume the difference between the treatment value and the outcome value is 1, i.e. treatment is $X = c + 1$ and the outcome is $X = c$.

3.3 Defining Interaction Bias for TACE

Many machine learning algorithms including those that employ causal techniques assume that data are IID ([Sch22], section 3). In other words, the theoretical and performance guarantees of these algorithms are based on data being IID. As such it would be useful to determine conditions under which an algorithm meant for IID data can be applied on non-IID data with the certainty that the resulting *bias* would be negligible.

Interaction bias is the bias induced by falsely assuming IID on datasets with interactions. We now formally define the interaction bias for a balanced interaction model.

Definition 8 (Interaction Bias)**.** *Let balanced interaction model M^* be the true model that generated the (available) non-IID dataset D. Let Q denote the query of interest and let Q^**

be its true value. Let Γ *denote an algorithm that outputs an unbiased estimate of* Q *given data that are IID and the causal graph that generated the IID data. Let* $G^{\dagger}$ *denote the unit default interaction graph for any unit in* M^**, with the explicit variables are relabeled as the corresponding generic variables. Let* $\hat{Q}$ *be the estimate computed by* Γ *using* $G^{\dagger}$ *and* D *as input. Interaction bias is given by* $||Q^* - E[\hat{Q}]||$.

For example, given the data D on two variables X and Y and assuming IID, with $G^{\dagger}$ as $X \to Y$, the IID method would output $\hat{Q}$ as the regression of Y on X. However, if D is generated from a non-IID process, then $G^{\dagger}$ cannot perfectly characterize the data generating process, and the estimation using $G^{\dagger}$ will be biased. The resulting bias is defined as the interaction bias.

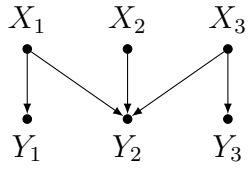


Figure 3.2: Interaction bias might cancel out.

$$Y_1 = 2X_1 \tag{3.1}$$

$$Y_2 = 2X_2 - X_1 + X_3 \tag{3.2}$$

$$Y_3 = 2X_3 \tag{3.3}$$

With certain parametrizations of the structural equations, the bias might appear to be 0, while indeed it is an "accidental cancel out." We do not want to count this case as unbiased because it is unlikely to occur. The "unbiased" property discussed in this work will always refer to *unbiased almost everywhere*, which we define below.

Definition 9 (Unbiased almost everywhere)**.** *θ defines the parametrization (structural equation functions) of the interaction model. An estimator $\hat{A}$ is unbiased almost everywhere if $E[\hat{A}]$ converges to the true value of A when sampled infinite times, except when θ resides on a set of Lebesgue measure zero.*

For example, given an interaction model with network Figure 3.3 and structural equations (3.4)-(3.6), the interaction bias is calculated to be 0 by the definition (see appendix for the

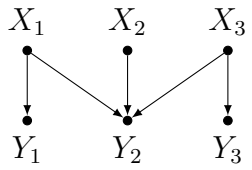


Figure 3.3: Interaction bias might cancel out.

$$Y_1 = 2X_1 \tag{3.4}$$

$$Y_2 = 2X_2 - X_1 + X_3 \tag{3.5}$$

$$Y_3 = 2X_3 \tag{3.6}$$

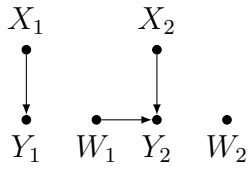


Figure 3.4: Interaction bias is 0 not due to accidental cancellations.

$$Y_1 = 2X_1 \tag{3.7}$$

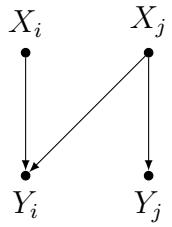
$$Y_2 = 2X_2 + W_1 \tag{3.8}$$

full derivation). However, if the parametrization (function) changes to $Y_2 = 2X_2 - 2X_1 + X_3$, then the interaction bias is not 0. This implies that the bias being 0 is an "accidental cancel out," since it occurs only for parametrizations satisfying certain constraints. Hence, this interaction model is not considered unbiased almost everywhere. On the contrary, given an interaction model with network Figure 3.4 and structural equations (3.7)-(3.8), the interaction bias is always 0 regardless of the parametrization. For example, changing the structural equation to $Y_2 = 2X_2 + 2W_1$ still results in 0 interaction bias. So this model is unbiased almost everywhere.

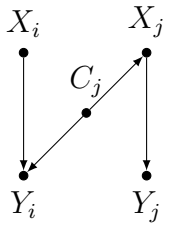
3.4 Quantifying, Detecting and Removing Interaction Bias for TACE

3.4.1 Quantifying Bias

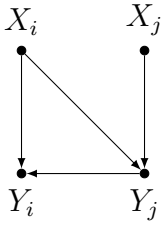
I define the two main types of problematic graphical structures in a linear interaction network that introduces bias in the estimation of $TACE$.



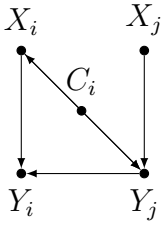
(a) X_j causes Y_i through a directed path (deflecting bias).



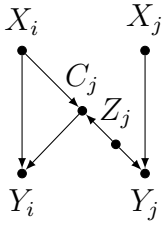
(b) X_j and Y_i are confounded (deflecting bias).



(c) X_i causes Y_i through Y_j (reflecting bias).



(d) X_i and Y_i have a confounding path (reflecting bias).



(e) X_i causes Y_i through C_j (reflecting bias).

Figure 3.5: Deflecting and reflecting interaction types.

Definition 10 (Deflecting bias structure)**.** *A deflecting bias structure for $TACE_{XY}$ in an interaction network G^* is an open path between X_j and Y_i for $i \neq j$.*

Deflecting bias structures are open paths from one unit to another unit. For example, Figures 3.5(a) and 3.5(b) contain deflecting bias structures. The interaction network in Figure 3.5(a) has a directed open path between X_j and Y_i, and the interaction network in Figure 3.5(b) has a confounded open path between X_j and Y_i.

Definition 11 (Reflecting bias structure)**.** *A reflecting bias structure for $TACE_{XY}$ in an interaction network G^* is an open path between X_i and Y_i through some explicit variable W_j with $i \neq j$.*

Reflecting bias structures are open paths that go from a unit through another unit and back to the same unit. For example, Figures 3.5(c) and 3.5(d) contain a reflecting bias structure. In each of them, there is an open path from X_i to Y_i through Y_j. In some cases, there can be a deflecting bias structure embedded in a reflecting bias structure, as in Figures 3.5(c) and 3.5(d). However, this is not necessary. Figure 3.5(e) contains only a reflecting bias structure ($X_i \rightarrow C_j \rightarrow Y_i$) but no deflecting bias structure.

Theorem 1. *Let $M^*(G^*, S^*)$ be a balanced interaction model in which treatment variable X_i and outcome variable Y_i are not confounded by any variable in $\mathcal{V}_i$, $\forall i$. Let D be the available data generated by M^* and let $G^\dagger$ be the generic network. Let $TACE_{XY}$ be identifiable in $G^\dagger$ and be given by β_{YX}, the regression coefficient of Y on X. Let α denote the true value of $TACE_{X,Y}$ in M^*. If X satisfies ASDC then the interaction bias is given by,*

$$\left| E[\hat{\beta_{YX}}] - \alpha \right| = \left| \frac{1}{n} \sum_{1 \leq i \leq n} \sum_{p \in P[iji]} Val(p) \frac{\sigma^2_{R_p}}{\sigma^2_X} - \frac{1}{n(n-1)} \sum_{1 \leq i \leq n} \sum_{p \in P[ji]} Val(p) \frac{\sigma^2_{R_p}}{\sigma^2_X} \right|,$$

where $P[iji]$ is the set of reflecting bias structures between X_i and Y_i through any explicit variable W_j of unit j with $i \neq j$, $P[ji]$ is the set of deflecting bias structures between X_j and Y_i with $i \neq j$, and R_p is the root of path p.

It follows from Theorem 1 that in a balanced interaction model in which no X_i and Y_i are confounded by any variable in $\mathcal{V}_i$, the reflecting and deflecting structures are the only two structures that will bias the identification of $TACE$. Note that although definition of interaction bias (Definition 8) on TACE is for any unbiased estimator for ACE, I focus only on the ordinary least squares estimator in this paper. This is because among the class of unbiased linear estimators, the OLS estimator has the minimum variance [JW14].

Next I exemplify theorem 1.

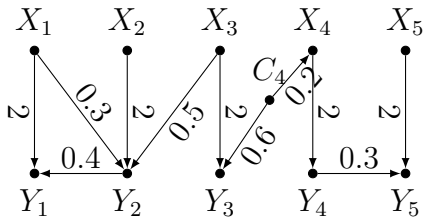


Figure 3.6: Interaction network with 4 units. The numbers represent edge coefficients. (C_1, C_2, C_3, C_5 are omitted)

Example 1. *Figure 3.6 shows an example of an interaction model with 4 units where $X_1, \ldots, X_5$ are the treatments, and $Y_1, \ldots, Y_5$ the outcomes. The numbers on the edges are the edge coefficients. C satisfies ASDC, and C_i for $i = 1, 2, 3, 5$ are omitted from the graph for simplicity.*

Suppose we want to estimate the ACE of X on Y as if the units were isolated: ***Input:*** *the interaction network G^* as shown in Figure 3.6 (no parameter i.e., S^* is not an input),* ***Output:*** *the $TACE_{XY}$ (should equal to 2). If we estimate ACE_{XY} ignoring the connections between units, our estimator will be $\hat{\beta_{YX}}$, with $Y = \{Y_1, \ldots, Y_5\}$ and $X = \{X_1, \ldots, X_5\}$. This is because ignoring the connections, the graph becomes $X_i \to Y_i$ separated for $i = 1, \ldots, 5$, so is essentially $X \to Y$ [Pea09]. However, by Theorem 1,*

$$|\beta_{YX} - 2| = |\frac{0.3 \cdot 0.4}{5} - \frac{1}{20} \cdot 0.5 - \frac{1}{20} \cdot 2 \cdot 0.4 - \frac{1}{20} \cdot 0.5 \cdot 0.4 - \frac{1}{20} \frac{0.6 \cdot 0.2\sigma_C^2}{\sigma_X^2} - \frac{1}{20} \cdot 2 \cdot 0.3|$$
$$\neq 0.$$

Hence, the result is biased, and does not give us what we want. I show later in Theorem 2 how to compute an unbiased estimate of $TACE$.

3.4.2 Detecting Bias

In this section, I provide a graphical criterion resulting from Theorem 1, to detect interaction bias.

Corollary 1. *Let $M^{**}(G^{**}, S)$ be a balanced interaction model in which X satisfies ASDC and TACE is identified as $\beta_{YX} = \alpha$ in the generic network, then interaction bias exists iff G^{**} contains a reflecting or deflecting bias structure.*

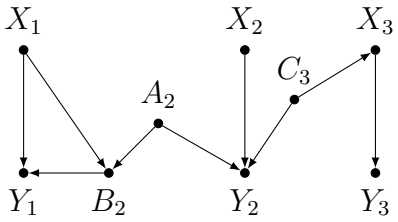


Figure 3.7: Interaction network with 3 units. (Other A, B, C variables including $A_1, B_1, \ldots$ are omitted)

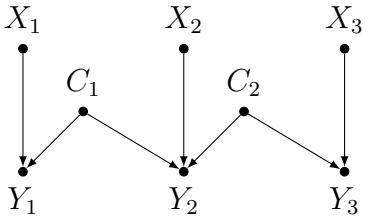


Figure 3.8: Interaction network with 3 units. (C_3 is omitted)

For example, Figure 3.7 contains both reflecting and deflecting bias structures. Figure 3.8 does not contain any bias structure. So Figure 3.7 has interaction bias and Figure 3.8 does not. Note that the interactions in Figure 3.8 do not qualify as bias structures by Definitions 10 and 11.

3.4.3 Removing Bias

Theorem 2 presents a technique for computing an unbiased estimate of TACE in cases where theorem 1 predicts significant bias. It proceeds by applying linear regression on a set of samples B that satisfy the condition that no bias inducing structures exist between any two distinct units i and j. In particular, a subset of samples/units B is termed as a **bias-free subset** for $TACE_{XY}$ if no reflecting bias structures exist for any $i \in S$ and no deflecting bias structures exist in G^*_S where G^*_S is the latent projection of G^* on B (Definition 2.6.1, [Pea09]). For example in figure 3.6, B comprises of units 2 & 5 and G^*_S is $X_2 \to Y_2$ $X_5 \to Y_5$. However, B *is not unique* for a given interaction network. Another candidate for B is units 2 & 4 and the associated G^*_S is $X_2 \to Y_2$ $X_4 \to Y_4$. An algorithm for constructing B is

presented in Algorithm 1, with an example and discussion in the appendix. This algorithm starts by randomly initializing B with a sample. Then it goes through the rest of the samples and adds a sample to B if its inclusion does not create bias structures in the resultant graph, G^*_S.

Algorithm 1 Select a bias-free subset B from an interaction network G^* and return the largest subset from t iterations

Input: an interaction network G^*, iterations t

Output: the largest bias-free subset B selected from t iterations

function FINDSUB(G^*, t)
  $\mathbf{B} = \emptyset$
  for $i = 1, \ldots, t$ **do**
    $Units$ = randomly sorted list $1, \ldots, n$
    $B = \{Units[1]\}$ (The indices for $Units$ start from 1)
    for $i = 2, \ldots, n$ **do**
      if $Units[i]$ has no reflecting bias structure in G^* **then**
        if $Units[i]$ has no deflecting bias structure in G^* with an element in B **then**
          $B = B \cup \{Unit[i]\}$
    $\mathbf{B} = \mathbf{B} \cup \{B\}$
  return Largest B in $\mathbf{B}$

Theorem 2. *Let G^* be an interaction network. Given the conditions in Theorem 1 and B a bias-free subset for G^*, $TACE_{XY} = E[\hat{\beta_{YX}}]$ where the regression coefficient is calculated using only samples in set B.*

Note that bias-free subset of samples B *used in Theorem 2 is not always IID*. While I insist that no reflecting or deflecting bias structures exist in G^*_S, I do not restrict other forms of interactions among these samples. For example, in Figure 3.8, Units {1, 2, 3} constitute

a bias-free subset. In this case, Y is not IID (Y_1 and Y_2 are dependent, Y_2 and Y_3 are dependent) and hence the bias-free subset is non-IID.

Also note that to compute an unbiased estimate using Theorem 2, we have at our disposal a smaller set of samples; so the variance of estimation will be larger. There is a trade off between ignoring interaction (large bias, small variance), and using theorem 2 (no bias, large variance). It remains future work to quantify the variance of the estimator in Theorem 2 for different interaction models, but in Section 3.5, I provide simulation results and case analysis study to empirically show its performance.

Applicability of theorems 1 & 2 to real world problems: A natural question that arises at this juncture is whether we need an entire interaction network to apply these results to real world problems. Theorem 1 quantifies bias and in doing so reveals to us if and how various factors such as sample size and strength of connections (value of path coefficients) influence bias. This in turn allows us to use available information about the problem from prior experience, domain knowledge or external sources to determine if bias would be negligible or not. Specifically, bias becomes smaller as the number of bias-structure-free samples increases. In fact, if the numbers of deflecting and reflecting structures are fixed, the bias terms diminishes as n increases, indicated by the $1/n$ for the reflecting bias term and $1/n(n-1)$ for the deflecting bias term. It is also evident that if the values of path coefficients are high, $Val(p)$ would be high and this will result in increased bias. Finally, if the interaction connections are sparse (fewer edges between units), the reduction in the total number of paths could potentially lower bias but more importantly the number of samples in the bias-free set B used in theorem 2 will tend to be larger, which in turn will help in computing better quality estimates.

3.5 Experiments

3.5.1 Simulations

Simulated Model I randomly generate balanced interaction network with n units (i.e., the sample size is n), with $C_i \to X_i \to Y_i$ and $X_i \to M_i$ for all $i = 1, \ldots, n$. For all ordered pairs of distinct units i, j, I randomly add deflecting bias structures in the form of $X_i \leftarrow C_i \to Y_j$ with probability $dRate$. For all units i, I randomly add reflecting bias structures in the form of $X_i \to M_k \to Y_i$ with probability $rRate$ for a random $k \neq i$.

Experiment: Bias of REG It follows from theorem 1 that larger sample sizes and smaller path values on the bias structures result in smaller bias. I perform two simulations to show how bias varies as a function of sample size and path values.I simulate data such that for each variable, the exogenous error term follows a Gaussian distribution with mean 0 and standard deviation 1. For each set of parameters, I randomly generate an interaction network, and simulate the data 10000 times. Each time, I record the result from a naive regression of Y on X (REG). As a comparison, I also record the result from Theorem 2 (THM-2). I run the algorithm (provided in the appendix) to randomly select bias-free subsets for 10 times and select the largest subset.

Simulation 1: $X_i \to Y_i$'s edge coefficient is 100, the edge coefficients of $C_i \to X_i, X_i \to M_j, M_j \to Y_i$ are all set to 10, the numbers of deflecting bias structures and additional reflecting bias structures are both 100.

Simulation 2: Number of units $n = 1000$, $X_i \to Y_i$'s edge coefficient is 100, the numbers of deflecting bias structures and additional reflecting bias structures are both 100. The results are plotted in Figure 3.9. As seen in the plots, as n increases or the path values on the bias structures decreases (both with all other parameters fixed), β_{YX} from a naive regression approaches $TACE$. Such results coincide with Theorem 1. The β_{YX} computed by THM-2 is very close to TACE and the two lines almost overlap.

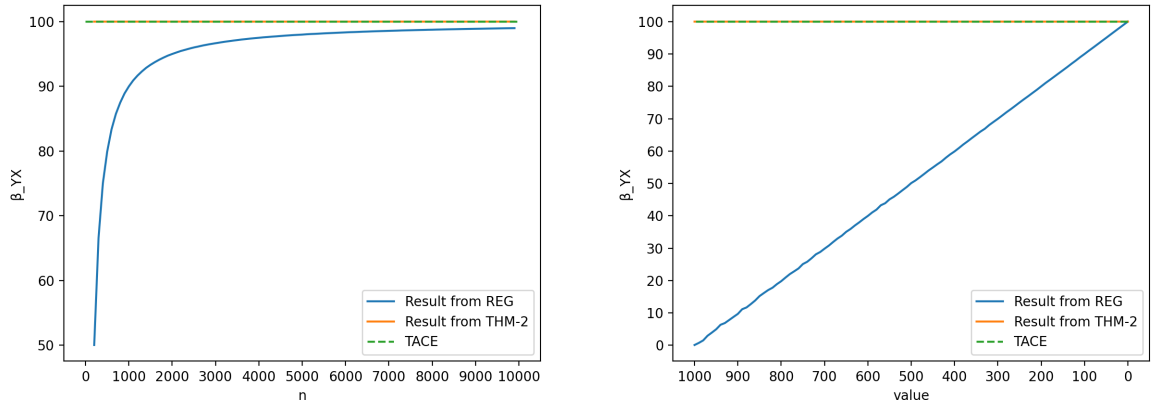


Figure 3.9: Left: β_{YX} vs. number of units n. Right: β_{YX} vs. path value on the bias structures. $TACE = 100$.

3.5.2 Case Study

Settings I am interested in analyzing the effect of tutoring time on students' grades. In particular, I wish to compute the effect provided through the tutoring program only, but not through "side effects" from other units, such as learning from classmates, although such interactions are encouraged in this scenario. For instance, unit i might help unit j understand the course materials better which in turn might improve j's grade. If unit i helped unit j improve their understanding and unit j states this in the peer review, then it would boost i's grade. To construct an interaction network and apply the proposed results, we ask the students to fill out a survey including 1) their tutoring time, 2) their grade, 3) whom they helped, 4) who helped them, 5) peer review score.

Construction of the Interaction Network Three generic variables are T (tutoring time in hours), U (understanding of course materials), and R (grade). For each unit i, $T_i \to U_i \to R_i$. In addition, if i helped j, add $U_i \to U_j$ (deflecting bias structure). If i first helped j and j mentioned this in the peer review and thus boosted i's grade, add $U_i \to U_j \to R_i$ (reflecting bias structure). I assume no additional back-and-forth help happens.

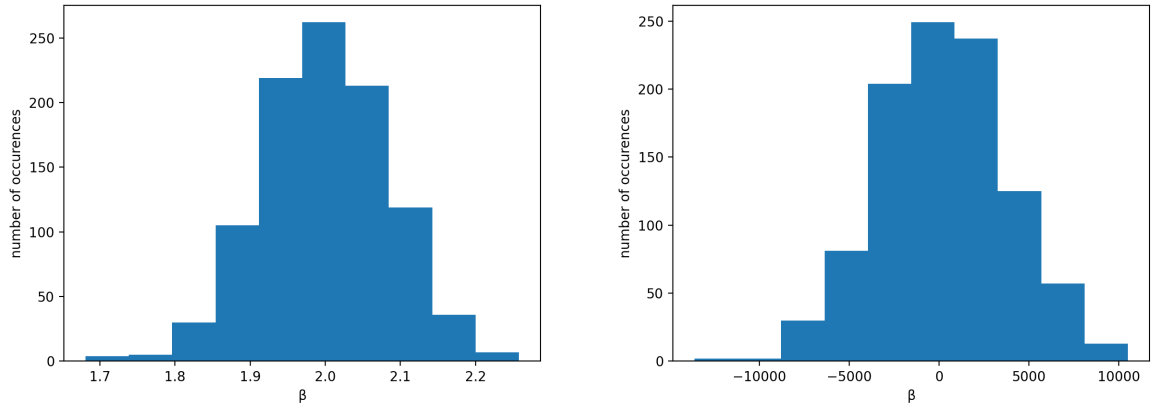


Figure 3.10: Left: estimated $TACE$ distribution from THM-2. Right: estimated $TACE$ distribution from REG.

Simulation Let there be 500 students, assume each student on average help 5 other students, and the other student has a 0.5 chance of helping back. Let $TACE = 2$, and the $U_i \to U_j$ and $U_i \to R_j$ edges both have the value 2. I randomly generate an interaction network and simulate data based on these parameters.

Results I apply THM-2 to select a bias-free subset, and compute β_{GT} using data from that subset. I get the result 1.963, with the size of the subset 72. As a result, the effect of tutoring time on students' grades not through other units is estimated to be 1.963, which is close to the ground truth $TACE$ (2). I further repeat the experiment 1000 times to show the distribution of the results. Each time a random structure is generated and random data are simulated. THM-2 is on average able to select a bias-free subset of size 76, and the average recovered $TACE = 2.0002$. The result from REG had a *significantly high bias with* $TACE$ *averaging at* 194.11. Also since every time the data are regenerated, the model is different, and REG uses all the data, it has a larger variance. The two plots in Figure 3.10 show the distribution of results from THM-2 and REG. The histograms of the results of β_{YX} computed by THM-2 and REG are shown in Figure 3.10.

3.6 Discussion and Summary

In this chapter, I derived theorems to quantify the interaction bias for average treatment effects in linear models, when generalized interference are present. I provided sufficient and necessary graphical conditions to detect interaction bias. Additionally, I developed a method to compute an unbiased estimate of causal effect in cases where blindly assuming IID is expected to yield a significant bias. Finally, I tested the performance of the proposed method through simulation studies.

[JPV20] proposed a quasi-coloring method to estimate direct effect under interference using experimental data. However, it does not easily generalize to observational studies. Other papers along a similar direction include [FZ20], which proposed experiment design to minimize interaction bias and selection bias at the same time, and [LH14], which proposed a two-stage randomization design to minimize interference bias. [TFS21] proposed a g-computation method, which is the first to model general interference using graphical models (chain-graphs), but requires the interference effects to be symmetrical between units. [SAH21] and [HH08] defined queries similar to TACE, named EATE and PADE, respectively. These queries generalize traditional ACE to allow a unit's outcome to be affected by treatments of other units. However, they do not allow outcomes to be affected by other units' variables other than treatments.

[HH08] defined six types of queries in the problems involving interference. Work in interference that focuses on different queries/problems include a few as follows. [VTH12] is the first to decompose the spillover effect (the effect of a unit's treatment on another's outcome ([Qua12])) to contagion and infectiousness effects using counterfactual mediation analysis. [STA17] presented decomposition for units with unknown and symmetrical interaction patterns and analyzed different interference paths. In linear models, the contagion and infectiousness effects reduce to the directed paths from X_j to Y_i. Moreover, their work does not handle reflection bias. [HLW21] was the first to define and provide estimands

for the average indirect effect. [VTH14] developed methods for sensitivity analysis under interference.

Other types of interactions include the contagion effects, which are defined as a unit's outcome affecting another unit's outcome [VA13]. Work on this line usually used longitudinal data, including [Bur87, Lyo11, VOT12]. Homophily effects are defined as the behavior of connected units are similar [JPV20]. Work in this line include [MSC01, JPV20]. The existing work above does not model interactions using graphical models.

CHAPTER 4

Uncertain Interaction Models

4.1 Introduction

In Chapters 2 and 3, I presented results that models generalized interference using interaction models that represent general interaction patterns between units, and is not limited to interference. However, one limitation with interaction models and many other approaches (such as those in [AS17, JPV20]) is that the interaction patterns need to be known in advance. This level of detail is not easily available in real-world datasets. For example, in a drug trial, it may not be feasible to track down each participant; in an online study, it is difficult to know if participants communicated with others in the study. The question of interest is, how can we perform causal analysis given non-IID data when there is uncertainty in the interaction pattern?

The main results of this chapter are as follows. I derive theorems to quantify interaction bias when some interaction paths exist with uncertainty (Thm. 3). I reduce or remove bias when some interaction paths exist with uncertainty (Thm. 4). I present a polynomial algorithm for the bias-reduction/removal method. (Algo. 2). Finally, I derive results for bounding ACE when some interaction paths exist with uncertainty. (Thm. 5 & Cor. 2).

4.2 Bias Reduction for Graph with Uncertain Interactions

While the interaction modeling has the benefit of modeling general arbitrary interactions, they rely on knowing the full interaction graph structure, which is often unavailable. In this section I will generalize those results to handle uncertainty in the interaction patterns among units.

Definition 12 (Uncertain Paths)**.** *An uncertain path between two distinct nodes A and B in a DAG is an open path between A and B that exists with probability θ, $0 < \theta < 1$.*

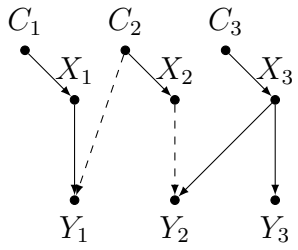


Figure 4.1: An uncertain interaction graph

A *definite* path on the other hand is one that exists with probability 1.

Definition 13 (Uncertain Interaction Graphs)**.** *An uncertain interaction graph is an interaction graph with uncertain paths.*

Figure 4.1 shows an uncertain interaction graph, where uncertain paths are represented as dashed arrows, and definite paths are represented as solid arrows.

There are multiple ways in which units can interact, such as two units' outcomes are confounded, a unit's treatment affects its outcome through another unit's variables, etc. In this chapter I focus on interference, since it is one of the most common and most studied type of interactions. Interference is defined as the phenomenon that one unit's treatment affects another unit's outcome. I assume that the only form of interaction in the interaction model is via interference paths, defined below.

Definition 14 (Interference Paths)**.** *Given an interaction graph, an interference path is a directed path from X_i to Y_j, $i \neq j$.*

I next impose a few additional restrictions on the graph so it is not too arbitrary to draw useful conclusions.

Definition 15 (Balanced Graph for Uncertain Interference (b-G^U, for short))**.** *An interaction graph, G^U, is termed as a* balanced graph for uncertain interference *if*

1. *it is the interaction graph of a balanced interaction model M^*,*
2. *the only type of bias structures in M^* are directed paths from X_i to Y_j where all intermediate nodes belong to either unit i or j.*
3. *only definite edges exist between any two nodes A_i and B_i of unit i, for any i. Uncertain edges may exist only between nodes of distinct units i, j, for any i and j.*
4. *the sum of the values of interference paths from X_i to Y_j (if such exists) is the same as that from X_k to Y_l (if such exists), for all $i \neq j$ and $k \neq l$.*

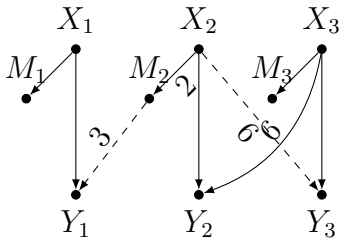


Figure 4.2: A balanced graph for uncertain interference.

Note that each interaction graph with or without uncertainty corresponds to an underlying interaction model that encodes the data generating process. Figure 4.2 is a b-G^U, if the interaction model it corresponds to is balanced. Condition 1 is satisfied. Condition 2 is satisfied since the only such path with an intermediate node is from X_2 to Y_1, with M_2 being an intermediate node, and it belongs to unit 2. Condition 3 is satisfied since the only

uncertain edges $M_2 \to Y_1$ and $X_2 \to Y_3$ are both between distinct units. As for Condition 4, we can calculate the sum of the values on the three interference paths. The edge coefficients are labeled in Figure 4.2, and they all equal to 6. Thus, Condition 4 is also satisfied.

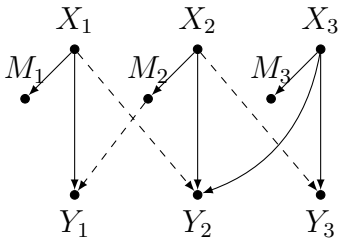


Figure 4.3: A balanced graph for uncertain interference.

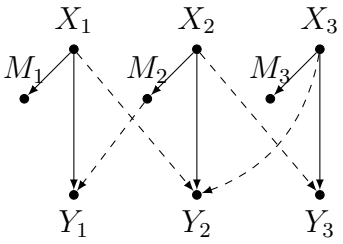


Figure 4.4: A balanced graph for uncertain interference with $N_d = 0$.

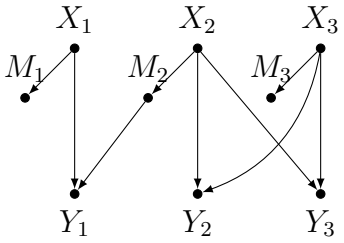


Figure 4.5: The underlying true interaction graph (unavailable).

$$S^* \text{ (unavailable)} \begin{cases} M_1 = 2X_1 + U_{M_1} \\ M_2 = 2X_2 + U_{M_2} \\ M_3 = 2X_3 + U_{M_3} \\ Y_1 = 5X_1 + M_2 + U_{Y_1} \\ Y_2 = 5X_2 + 2X_3 + U_{Y_2} \\ Y_3 = 5X_3 + 2X_2 + U_{Y_3} \end{cases}$$

As is mentioned in the preliminaries, the interaction bias (Definition 6) is the bias resulted from incorrectly assuming IID to estimate the unit "true" ACE ($TACE_{XY}$). Theorem 3 below quantifies the interaction bias in an uncertain interaction graph.

Theorem 3. *Suppose $M^*, D, G^\dagger$ refer to the true model, available data and generic network as specified in definition 6 such that $Q = TACE_{XY}$ and $\hat{Q} = \beta_{YX}$. X_i and Y_i are not confounded by any variable of i, for all i. Let G^U be the b-G^U corresponding to M^*. For all $i \neq j$ pairs, let N_d be the number of pairs of units that have definite interference paths from i to j and let N_θ be the number of pairs of units that have uncertain interference paths from*

i to j with probability θ. Let the sum of the values of the interference paths from X_i to Y_j be p,[1] *for all $i \neq j$. The expected interaction bias is given by*

$$E[\left|E[\hat{\beta_{YX}}] - Q\right|] = \frac{1}{n(n-1)}|p|(N_d + \theta N_\theta).$$

Figure 4.3 is a b-G^U with $N_d = 1$ ($X_3 \to Y_2$) and $N_\theta = 3$ ($X_2 \to M_2 \to Y_1$, $X_1 \to Y_2$, and $X_2 \to Y_3$). $n = 3$ since there are 3 units. The underlying true interaction model (unavailable) is shown in Figure 4.5, with the structural equations on the right. The interference effect $|p|$ is equal to 2, calculated from the structural equations. $TACE_{XY}$ is 5. We can also see that the true θ is 2/3, i.e., out of the 3 uncertain paths, there are 2 that really exist. Although in real-world applications, θ is usually unavailable, so we need an estimate from expert knowledge about the frequency of interference in this sample. Figure 4.4 shows another b-G^U that corresponds to Figure 4.5. In Figure 4.4, there is no definite interference paths. This is in fact an interesting special case, which I will elaborate more in the next section.

The debias method in [ZMP22] selects a bias-free subset of units and uses it to unbiasedly compute TACE given the full interaction graph. When there is uncertainty, if we treat all uncertain interference paths as definite existence, we might end up selecting too small a subset, especially when there are many uncertain interference paths. One solution is to select a larger subset to maybe include some interactions, while still bound the interaction bias at a reasonable level. Theorem 4 below shows such a method.

Theorem 4. *Consider the setting in Theorem 3. Suppose we are additionally given a bias threshold τ, and the interference effect is bounded by a constant Γ times the TACE (i.e., $|p| \leq \Gamma|Q|$). If a subset $\mathcal{B}$ of units satisfies*

$$\frac{1}{|\mathcal{B}|(|\mathcal{B}|-1)}(N'_d + \theta N'_\theta)\Gamma \leq \tau,$$

[1] *I.e., p is equal to the causal effect of X_i on Y_j.*

then using the samples in $\mathcal{B}$, the expected interaction bias will be at most $\tau|Q|$. For all $i \neq j$ pairs with $i, j \in \mathcal{B}$, N'_d denotes the number of pairs with definite interference paths from i to j in G^*, and N'_θ denotes the number of pairs with interference paths from i to j in G^* with probability θ.

If such a subset is found, then the bias is bounded. For example, if the threshold $\tau = 0.1$, the bias will be as large as 10% of the true ACE, computed using the data from the selected subset. This theorem becomes a debias method if $\tau = 0$, since that simply implies that the bias has to be 0. Algorithm 2 is a polynomial greedy algorithm that selects such a subset given threshold τ.

Algorithm 2 Select a subset $\mathcal{B}$ from an uncertain interaction graph G^U that makes the interaction bias $\leq \tau$

Input: an interaction graph G^U, probability of uncertain paths θ, interference/TACE ratio bound constant Γ, bias threshold τ

Output: a subset $\mathcal{B}$ resulting in $\leq \tau$ bias

function FINDSUB(G^U, τ)
  $Units =$ randomly sorted list $1, \ldots, n$
  $\mathcal{B} = Units[1]$
  for $i = 2, \ldots, n$ **do**
    if $\mathcal{B} \cup \{Units[i]\}$ satisfies $1/((|\mathcal{B}+1|)|\mathcal{B}|)(N'_d + \theta N'_\theta)\Gamma \leq \tau$ **then**
      $\mathcal{B} = \mathcal{B} \cup \{Units[i]\}$
  return $\mathcal{B}$

Algorithm 2 goes through all the units, and select units one at a time, until the condition is no longer satisfied, and the selected subset is returned.

4.3 Causal Effect Estimation with Unknown Interference Structures

Next, I present a theorem for unbiased estimation of TACE. Unbiased estimation is possible if the relationship between the interference path strength and the TACE is given, and where the interference paths occur need not be known.

Theorem 5. *Consider the setting described in Theorem 3. Suppose we know the relationship between p (the interference path strength) and Q (TACE) is $p = \gamma Q$, where γ is a constant, then Q is unbiasedly estimated as*

$$Q = \frac{E[\hat{\beta_{YX}}]}{1 - \frac{1}{n(n-1)}\gamma(N_d + \theta N_\theta)}.$$

Applying Theorem 5 to generate bounds In this chapter there are two types of effects under consideration. First, the effect of X_i on Y_i (unit specific effect) and second, the effect of treatment applied to other units such as X_j, $j \neq i$ on Y_i (interference). In many situations such as when treatment is vaccination and outcome is disease, (i) the magnitude of unit-level treatment effects (TACE) can be safely assumed to be higher than those due to interference (p); mathematically, this translates to $|Q| > |p|$ and $0 < \gamma < 1$.

Corollary 2. *Consider the setting described in Theorem 5, if we further assume $0 < \gamma < 1$ and $|Q| > |p|$ and $0 < \gamma < 1$ then Q can be bounded as*

$$\frac{E[\hat{\beta_{YX}}]}{1 - \frac{(N_d + \theta N_\theta)}{n(n-1)}} < Q < E[\hat{\beta_{YX}}].$$

Note that from Corollary 2, Q is always less than $E[\hat{\beta_{YX}}]$. This implies that when the unit specific effect and the interference effect have the same sign, then assuming IID ($E[\hat{\beta_{YX}}]$) always "overestimates" the true unit specific effect (Q).

Remark 3. *Note that there are several interesting special cases with the the results presented in Theorems 3, 4, and 5.*

1. $N_d = N_\theta = 0$. *In this case, there is no interference path (definite or uncertain) in the model, which results in a model without interaction structures. In Theorem 3, the interaction bias is 0. In Theorem 4, the inequality always hold since the l.h.s. is 0, while τ is positive, so we can select any subset $\mathcal{B}$ where $|\mathcal{B}| > 1$. In Theorem 5, $Q = E[\hat{\beta_{YX}}]$, which is consistent with an interference-free setting.*

2. $N_d = 0$. *In this case, there is no definite interference path. This special case is useful when we do not have any information about which units interact with which units in some real-world applications. All those theorems still apply.*

3. $N_\theta = 0$. *In this case, there is no uncertain interference path. This means we have all the information regarding which units interact with which units. The theorems reduce to the results in [ZMP22], where there is no uncertainty in the interaction network.*

4.4 Discussion and Summary

This chapter focused on the problem of interference (non-generalized) when there is uncertainty regarding the interaction patterns. I showed that bias due to interference can be quantified using the interference strength and expected number of interactions. I developed an algorithm that computes true average causal effect such that bias is guaranteed to be less than a given quantity τ. Finally, I bound the average causal effect when it is guaranteed that unit level causal effect is higher than interference.

To my knowledge, there is no existing work that systematically discusses uncertainty in interaction patterns. There exists work that exploits model uncertainty for traditional causal diagrams under the IID assumption. Some uncertain DAGs include *patterns* in [VP91] and *partial ancestral graphs (PAGs)* in [Ric96]. Both graphical frameworks have uncertain edges

in the graph, representing unknown edge orientations. In addition, PAGs are used to represent equivalence classes of maximal ancestral graphs (MAGs) [RS02]. MAGs are abstractions of DAGs that keep only the conditional independence and ancestral relationships. Formally, MAGs are *maximal* and *ancestral*. There is an edge between two nodes A and B in the MAG if and only if there exists no set that can separate A and B in the DAG (maximal), and $A \to B$ is in the MAG if and only if A is an ancestor of B in the DAG (ancestral). [JZB18] introduced a causal identification method for PAGs. Causal discovery methods including the PC algorithm [SGS00] and the IC algorithm [Pea09] learn patterns and the FCI algorithm [SGS00] learns a PAG.

CHAPTER 5

Non-Parametric Interaction Models

5.1 Introduction

In the previous chapters, I presented linear interaction models and discussed bias analysis under the linear setting. The linearity assumption simplifies the generalized interference problem, since the interaction effects are always added to the total effects and can be naturally separated. In this chapter, I relax this assumption and extend interaction models for the non-parametric case. I perform bias analysis from applying estimation techniques meant for IID data such as Horvitz-Thompson estimator [HT52] on non-IID data. I also present an assumption weaker than linearity that can help mitigate bias. Finally, I present results on unbiased estimation design and empirically evaluate the debiasing procedure for different setups.

5.2 Defining Non-Parametric TACE

The main query of interest is the causal effect through units themselves, where hypothetically there is no influence from other units. For example, I am interested in the effect of a vaccine on a disease through a person's own immune system, but not through obtaining immunity from people around them. I first define unit default interaction model, which is the "default" causal model that we expect a unit to have if no interaction with other units exists. I disconnect interacting units by replacing the variables in the interacting terms in the structural equation with their default value. In linear models, removing interacting

effects from a variable means removing the terms that cause interactions, so default values are always 0. In non-parametric models, default values can be any real number that the variable can take on, determined by specific settings. For instance, given that the sickness of a patient is equal to their own immune strength multiplied by their close contact's sickness. If we assume the default model for the sickness of a patient is equal to its immune strength, under no interaction, then the default value for their close contact's sickness is set to a non-zero value 1.

Note that the default value for each generic variable need not be the same across all units, although I assume it is the case to keep the notations simple. Let $Pa(V_i)$ denote the parents of V_i in the interaction network.

Definition 16 (Unit Default Interaction Model)**.** $DM_i^*(DG_i^*, DS_i^*)$ *is the* unit default interaction model *for unit i with respect to an interaction model $M^*(G^*, S^*)$ with a set of default generic variable values dv if it is constructed from M^* in the following way:*

1. *$DG_i^* = G_i'$ where G_i' is the subgraph of G^* containing only variables of i,*
2. *$DS^* = S'$ where S' is the set of equations for variables of i, obtained by substituting each equation $V_i = f(Pa(V_i))$ in S^* any W_j, $\forall j \neq i$, with the corresponding constant in dv.*

Given the equations (5.1)-(5.4), the interaction network in Figure 5.1, and default value $X = 1$, the unit default models are as follows. For unit 1, the unit default model remains unchanged since it is not affected by another unit. However, for unit 2, the unit default model is given below and is obtained by replacing X_1 with its default value 1.

$$X_2 = U_{X_2}$$
$$Y_2 = 2X_2 - 1 + U_{Y_2},$$

In order to utilize data from non-IID units, I will limit our attention to a type of model with some symmetry information shared among the units. I define *balanced interaction model*

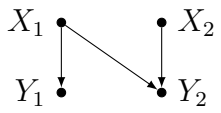


Figure 5.1: A simple interaction network.

$$X_1 = U_{X_1} \tag{5.1}$$
$$X_2 = U_{X_2} \tag{5.2}$$
$$Y_1 = 2X_1 + U_{Y_1} \tag{5.3}$$
$$Y_2 = 2X_2 - X_1 + U_{Y_2} \tag{5.4}$$

$$X_1 = U_{X_1} \tag{5.5}$$
$$X_2 = U_{X_2} \tag{5.6}$$
$$Y_1 = 2X_1 + U_{Y_1} \tag{5.7}$$
$$Y_2 = 2X_2X_1 + U_{Y_2} \tag{5.8}$$

$$X_1 = U_{X_1} \tag{5.9}$$
$$X_2 = U_{X_2} \tag{5.10}$$
$$Y_1 = 3X_1 + U_{Y_1} \tag{5.11}$$
$$Y_2 = 2X_2 - X_1 + U_{Y_2} \tag{5.12}$$

for the non-parametric case. The intuition behind this definition is that if hypothetically all interactions were removed, then the units would behave in the same way (have the same data generating process). Balanced model does not necessarily mean that interactions are non-existent. All it means is that the underlying unit default models are identical.

Definition 17 (Balanced interaction model M^*)**.** *An interaction model M^* is* balanced *with default values dv if the unit default interaction model with dv for each unit is identical.*

Since unit default interaction models are identical in balanced models, I will denote the equations using superscript D. For example, I use $Y^D = 2X^D$ instead of $\forall i \sim Y_i = 2X_i$.

Note that the choice of default value affects whether an interaction model is balanced. The model with interaction network Figure 5.1 and structural equations (5.1)-(5.4) is not balanced when $dv : X = 1$, since the unit default models for Y_1 and Y_2 are different. However, given $dv : X = 0$, the model is balanced. The structural equations of the unit default model is given by

$$X^D = U_{X^D}$$
$$Y^D = 2X^D + U_{Y^D}.$$

The interaction network of the unit default model is $X^D \to Y^D$.

Given the same interaction network, and the structural equations (5.5)-(5.8), the model is balanced when $dv : X = 1$, but is not balanced otherwise. Given the same interaction network, and the structural equations (5.9)-(5.12), the model is not balanced given any dv.

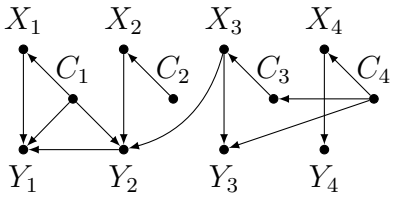


Figure 5.2: Interaction network with 4 units and 12 explicit variables (X_i, Y_i, C_i for $i = 1, 2, 3, 4$).

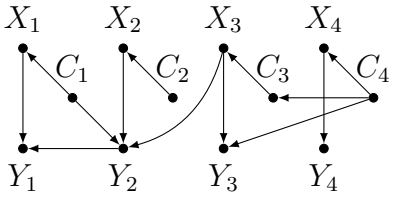


Figure 5.3: Interaction network with 4 units and 12 explicit variables (X_i, Y_i, C_i for $i = 1, 2, 3, 4$).

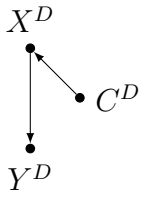


Figure 5.4: The shared unit default model.

$$Y_1 = X_1^2 Y_2 - 4 \qquad (5.13)$$

$$Y_2 = X_2^2 + X_3 - 4 \qquad (5.14)$$

$$Y_3 = X_3^2 - 2C_4 \qquad (5.15)$$

$$Y_4 = X_4^2 - 4 \qquad (5.16)$$

$$Y^D = (X^D)^2 - 4 \qquad (5.17)$$

Certain interaction networks might belong to a balanced model regardless of the default value. For example, Figure 5.2 cannot be the interaction network of a balanced model, since the unit default models are not the same. There is an edge $C_1 \to Y_1$, while there is no such edge for units 2, 3, or 4. Figure 5.3 can be balanced, depending on the structural equations and default values. If the structural equations are Eqs. (5.13)-(5.16) (the equations for C

and X are omitted and assumed to be identical in the unit models), and the default values are $dv : C = 2, X = 0, Y = 1$, then the model is balanced. The unit default interaction network is Figure 5.4, and the structural equation for Y^D is Eq. (5.17).

Assuming IID, an unbiased estimator for average causal effects under no confoundedness is the HT estimator [HT52, AM13]. The average causal effect of $X = 1$ vs. 0 on Y is estimated by

$$\frac{\sum_n X_i Y_i}{m_1} - \frac{\sum_n (1 - X_i) Y_i}{m_0},$$

where m_1 and m_0 are the numbers of $X_i = 1$ and $X_i = 0$ in this sample (I will keep using this notion throughout the manuscript). $m_1 + m_0 = n$. This estimand is essentially taking the difference between the average value of Y_i where $X_i = 1$ and the average value of Y_i where $X_i = 0$. I will use this estimator as the default for estimating ACE assuming the data are IID, assuming no confounder exists between X_i and Y_i for all i. I define the query of interest, which is the average causal effect as if the units were isolated, as follows.

Definition 18 (True Average Causal Effect ($TACE_{XY}$))**.** *Let M^* be a balanced interaction model with default values dv. True average causal effect of $X = 1$ vs 0 on Y, denoted as $TACE_{XY}$, is defined as the ACE of X on Y in the identical unit default model with dv corresponding to M^*.*

For example, given Figure 5.3 as the interaction network, Eqs. (5.13)-(5.16) as the structural equations, and the default values $dv : C = 2, X = 0, Y = 1$, the true average causal effect of $X = 1$ vs $X = 0$ on Y is given by

$$TACE_{XY} = \frac{\sum_i \mathbb{1}_{\{X_i=1\}} Y_i}{m_1} - \frac{\sum_i \mathbb{1}_{\{X_i=0\}} Y_i}{m_0} = (1^2 - 4) - (0^2 - 4) = 1,$$

where $\mathbb{1}_A$ is the indicator function that is equal to 1 if A is true and 0 if A is false.

5.3 Quantifying and Detecting Bias: the General Case

Theorem 6 (Interaction Bias in Non-Param. Models)**.** *Let M^* be an interaction model balanced with default values dv. X satisfies ASDC. The average causal effect of X on Y is estimated as $\hat{Q}_{Y|do(X)}$ in $G^\dagger$. The interaction bias is given by*

$$
\begin{aligned}
&|E[\hat{Q}_{Y|do(X)}] - TACE_{XY}| \\
=&\Bigg| E_{m_1,m_0}\left[\frac{1}{n}\sum_{1\le j\le n} E\left[Y_j|\, m_1, m_0, X_j = 1\right] - \frac{1}{n}\sum_{1\le j\le n} E\left[Y_j | m_1, m_0, X_j = 0\right]\right] \\
&- (E[Y^D|X^D = 1] - E[Y^D|X^D = 0])\Bigg|,
\end{aligned}
$$

where n is the total number of units ($n = m_1 + m_0$).

If the model is non-parametric, i.e., no assumption is made of the function forms (linearity, etc.) of M^*, the graphical criterion for detecting bias is as follows.

Theorem 7. *Let M^* be an interaction model balanced with default values dv. X satisfies ASDC. The average causal effect of X on Y is estimated as $\hat{Q}_{Y|do(X)}$ in $G^\dagger$. There is no interaction bias iff Y is ASDC.*

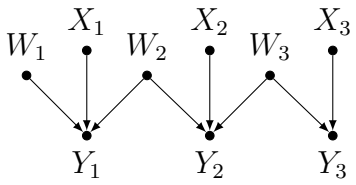


Figure 5.5: Interaction network with 3 units.

$$Y_1 = 3X_1W_1 + W_1W_2 \qquad (5.18)$$

$$Y_2 = 3X_2W_2W_3 + 1 \qquad (5.19)$$

$$Y_3 = 3X_3W_3 + 1 \qquad (5.20)$$

$$Y_1 = 3X_1W_1 + W_2^2 \qquad (5.21)$$

$$Y_2 = 3X_2W_2 + W_3 \qquad (5.22)$$

$$Y_3 = 3X_3W_3 + 1 \qquad (5.23)$$

Example 2. *I use an example to illustrate this theorem. An interaction model has interaction network Figure 5.5, and structural equations (5.18)-(5.20), dv* : $W = 1$. *This model is balanced with shared unit default structural equation* $Y^D = 3X^D W^D + 1$. *The interaction bias given a distribution of* m_1 *and* m_0 *where* $m_1 = 2$ *and* $m_0 = 1$ *is calculated as follows.*

$$
\begin{aligned}
&|E[\hat{Q}_{Y|do(X)}] - TACE_{XY}| \\
=&\Bigg|\frac{1}{3}\sum_{1\leq j\leq 3} E\left[Y_j|\, m_1 = 2, m_0 = 1, X_j = 1\right] - \frac{1}{3}\sum_{1\leq j\leq 3} E\left[Y_j|m_1 = 2, m_0 = 1, X_j = 0\right] \\
&- (E[Y^D|X^D = 1] - E[Y^D|X^D = 0])\Bigg| \\
=&\Bigg|\frac{1}{3}(E[3W_1 + W_1W_2] + E[3W_2W_3 + 1] + E[3W_3 + 1]) \\
&- \frac{1}{3}(E[W_1W_2] + E[1] + E[1]) - (3E[W^D])\Bigg| \\
=&\Big|E[W^D W^{D'}] - E[W^D]\Big| \neq 0.
\end{aligned}
$$

The notion $W^{D'}$ *is used to distinguish from* W^D *so they are not (mistakenly) assumed to be the same variable. However* W^D *and* $W^{D'}$ *are IID. Since* Y *is not ASDC, interaction bias exists and this is also confirmed by Theorem 7.*

5.4 Restricted Additivity

In this section, I define a parametric assumption called restricted additivity, which will provide stronger results by aiding in mitigating bias. Note that this assumption is weaker than both linearity and IID, since it still allows interactions between units and non-linear terms.

Definition 19 (Restricted Additivity)**.** *Given an interaction model* M^*, *let* $\mathcal{T}_{V_i}$ *denote the set of terms in the structural equation of* V_i. M^* *satisfies* restricted additivity *if* $\forall V_i$, $\forall t \in \mathcal{T}_{V_i}$, *one of the following holds true:*

1. t *contains only explicit variables of unit* i *that correspond to ASDC generic variables*

2. t contains only one explicit variable E_i. E_i corresponds to a Non-ASDC generic variable E.

3. t contains only one explicit variable E_j with $j \neq i$ and no other variable.

For example, given the structural equation of an explicit variable V_i as $V_i = A_iB_i + C_j + D_iE_k$. Then, $\mathcal{T}_{V_i} = \{A_iB_i, C_j, D_iE_k\}$. Note that the term $t = D_iE_k$ where $t \in \mathcal{T}_{V_i}$ contains a variable E_k that is not in i, and there is another variable D_i in t. So the interaction model with this V_i does not satisfy restricted additivity. Note that any linear model satisfies restricted additivity. In addition, variables that undergo log transformation might make the model satisfy restricted additivity, although the model might not satisfy it before variable transformation. As long as the model with the given variables satisfies restricted additivity, the theories presented in this section apply.

Restricted additivity is essentially requiring that all non-ASDC variables or variables belonging to a different unit linearly affect their children. As a result, the components that affect each variable can be separated into the sum of two parts, the ASDC part (which is the same for each unit) and the non-ASDC part. For example, given Figure 5.5, and structural equations (5.18)-(5.20), the interaction model does not satisfy restricted additivity. When the structural equations are (5.21)-(5.23), this interaction model satisfies restricted additivity. In this case, similarly, the interaction bias is calculated as follows.

$$
\begin{aligned}
&|E[\hat{Q}_{Y|do(X)}] - TACE_{XY}| \\
=&\left|\frac{1}{3}\sum_{1\le j\le 3} E\left[Y_j|\, m_1=2, m_0=1, X_j=1\right] - \frac{1}{3}\sum_{1\le j\le 3} E\left[Y_j|m_1=2, m_0=1, X_j=0\right]\right. \\
&\left. - (E[Y^D|X^D=1] - E[Y^D|X^D=0])\right| \\
=&\left|\frac{1}{3}(E[3W_1+W_2^2] + E[3W_2+W_3] + E[3W_3+1])\right. \\
&\left. - \frac{1}{3}(E[W_2^2] + E[W_3] + E[1]) - (3E[W^D])\right| \\
=&\left|E[W^D] - E[W^D]\right| = 0.
\end{aligned}
$$

There is no interaction bias, which is consistent with Theorem 8.

Theorem 8. *Let $M(G,S)$ be an interaction model balanced with default values dv. X satisfies ASDC. The conditional average causal effect of X on Y is estimated as $\hat{Q}_{Y|do(X)}$ in $G^\dagger$. Assume restricted additivity. There is interaction bias iff there are deflecting or reflecting bias structures between X and Y in G.*

Compared to Theorem 7, which disallows any form of interaction for the model to be bias-free, this only forbids the two bias structures. Often persimmible structures include $Y_1 \leftarrow W_2 \rightarrow Y_2$, etc. For example, Figure 5.5 is unbiased under restricted additivity (if the conditions for Theorem 8 is met), but is biased without assuming restricted additivity.

5.5 Debiasing

Theorem 9. *A subset of units $\mathcal{B}$ is a bias-free subset for the causal effect of X on Y iff the latent projection [Pea09] on $\mathcal{B}$ does not have interaction bias by Theorem 7 (by Theorem 8 if restricted additivity is satisfied).*

For example, no such subset can be selected for the model in Figure 5.5 not assuming restricted additivity, since no two Y_i is ASDC. If restricted additivity is assumed, then {1,

2, 3} constitutes such a subset. I provide a polynomial algorithm for selecting a bias-free subset in the appendix.

5.6 Experiments

In this section, I present several simulated experiments to illustrate and support the theoretical results. In the interest of space, I have only retained crucial information in this paper. Fine-grained information is available in the appendix.

5.6.1 The Size of Interaction Bias

In the following experiments, I will demonstrate results related to the theorems for bias quantification and bias detection (Theorems 6, 7, 8). I show that interaction bias exists under the proposed graphical criteria, and illustrate how the bias size changes with varies factors. For the analysis of bias, I will run a "naive" method that estimates TACE as if the data were generated by an IID model, i.e., compute $\hat{Q}_{Y|do(X)} = E[Y|X = 1] - E[Y|X = 0]$ from the sample.

5.6.1.1 Bias due to Blindly Assuming IID

The goal of this experiment is to test how well the naive method would perform in the presence of bias inducing structures. In Theorem 8, consider the case where restricted additivity holds, deflecting bias and reflecting bias structures cause interaction bias. I compare the bias of blindly assuming IID with three types of interactions including deflecting bias structures, reflecting bias structures, and non-bias interactions. I simulate data based on interaction models with interactions shown in Figure 5.6, where from left to right are deflecting bias only, reflecting bias only, and non-bias interaction only. I generated different structures with those interactions randomly added to different places of the structures.

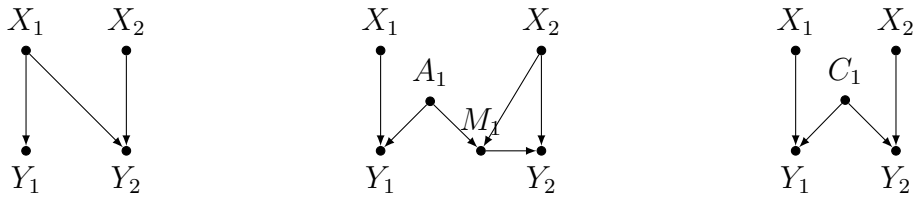


Figure 5.6: Three types of interactions.

Setup: The same default model is used for all three cases ($TACE = 6$), and the same number of corresponding interactions are added. I run the experiment for 1000 iterations to reduce sampling variance. In each iteration, I resample data for each of the three interaction models, and estimate the TACE of X on Y using the naive method.

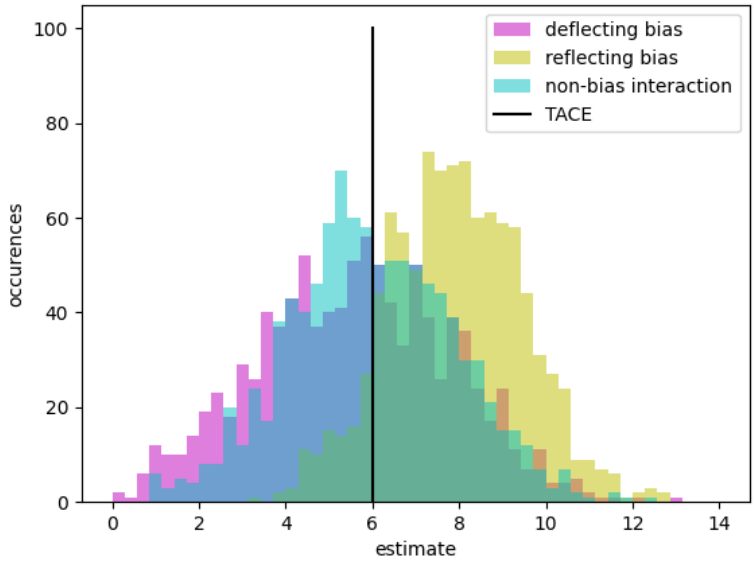


Figure 5.7: Comparison of the estimation assuming IID for three interaction types.

Results: I plot a histogram of the results from the 1000 iterations and is shown in Figure 5.7. The means and variances of the estimates of the three models are listed in the following table.

	def. model	ref. model	non-bias model
Mean	5.57	7.92	5.97
Var.	5.10	2.42	3.68

The mean of the non-bias interaction case is close to $TACE$. The means of the deflecting case and the reflecting case are both biased, where the reflecting case has a larger bias.

5.6.1.2 Size of Interaction Bias

The goal of this experiment is to test how interaction bias size changes with other factors. I first assume a restricted additivity setting, and show how the bias varies with bias (deflecting or reflecting) structure strength and density.

Deflecting bias case: I simulate data from a balanced interaction model, where deflecting bias structures exist in the form of $X_i \to Y_j$ for $i \neq j$. α denotes the bias strength and is equal to 1, -3, or 5 in our experiment. The experiment is run for 1000 iterations and averaged to reduce uncertainty. The plot below shows how bias (Y-axis) varies with sample size (X-axis), with the number of interactions fixed at 50.

As seen in the graph, the bias is larger when the bias structure has a stronger bias. Also, as the sample size increases, the interaction network has less interaction density (since the total number of interactions is fixed by our setting), and the bias becomes smaller.

Reflecting bias case: similarly, I simulate data from a balanced interaction model, where reflecting bias structures exist in the form $X_i \to M_j \to Y_i$ with $i \neq j$. The plot is as follows.

The effect of reflecting biases is in general larger than deflecting biases, while they both become smaller as the bias strengths and the interaction densities decrease. Next, I show the results for a general non-parametric setting without assuming restricted additivity.

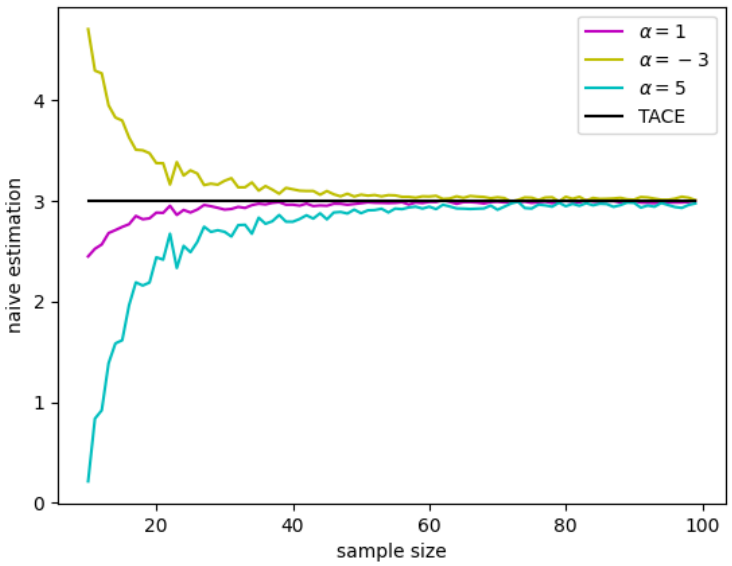


Figure 5.8: Deflecting bias size vs. sample size with total interaction number fixed.

General case: without the restricted additivity assumption, any interaction that causes Y to be non-ASDC will result in interaction bias. So in this experiment, I do not distinguish between deflecting and reflecting bias structures, and instead I use interaction structures that would not cause bias if they were in the restricted additivity setting. The interactions are of the form $Y_i \leftarrow C_i \rightarrow Y_j$ with $i \neq j$. The bias strength α is equal to 0.1, -0.3, or 0.5 in our experiment. The plot below shows how the interaction bias varies with different bias strengths and sample sizes (with the total number of interactions fixed).

The general non-parametric case is different from the restricted additivity cases we have seen above. Although note that the bias still becomes smaller as the interactions become less dense, the value of the parameter α is not necessarily positively correlated with the bias size. The reason for this is that the interactions can affect the estimate in arbitrary ways, so it is likely non-monotonic. Note that an extreme case happens when $\alpha = 0.5$, where there is no bias. This is due to an "accidental" but not "structural" cancel out, since it happens only for certain parameter choices. Such cases are ruled out when assuming unbiased everywhere.

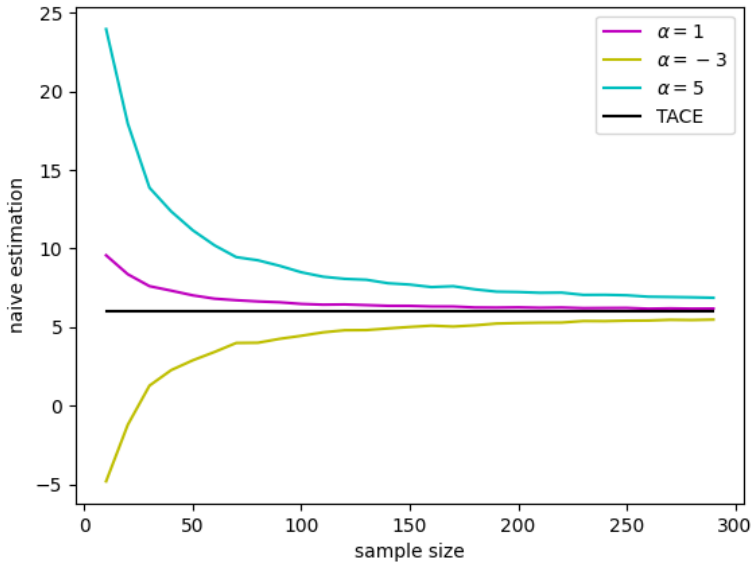


Figure 5.9: Reflecting bias size vs. sample size with total interaction number fixed.

5.6.2 Debias

The goal of the following experiments is to test the performance of the proposed debias method (Theorem 9).

5.6.2.1 Assuming Restricted Additivity

Setup: I simulate data from a balanced interaction model ($TACE = 3$) with both deflecting and reflecting bias structures. I run both a naive method that estimates as if the data were IID using the original dataset (denoted ORI), and our proposed method selecting a bias-free subset (denoted SUBS). I repeat the experiment for 1000 iterations and record all the results. The experiment is done for two different settings: 1. all units have equal chance of being involved in an interaction, and 2. some units have higher chance and some units have lower chance. Setting 1 is a simple setting that is likely to happen in designed experiments, while setting 2 is a more real-world setting (e.g., some people are more isolated

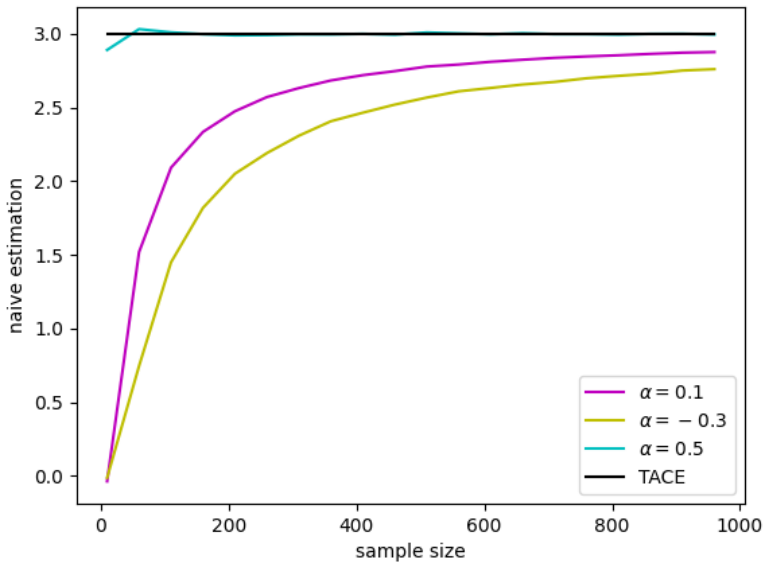


Figure 5.10: General case: interaction bias size vs. sample size with total interaction number fixed.

and some people are more social).

Results: The results are in the plots and tables below.

Table 5.1: Restricted additivity & setting 1.

	ORI	SUBS
mean	3.75	3.00
var.	5.15	12.46

Table 5.2: Restricted additivity & setting 2.

	ORI	SUBS
mean	3.74	2.90
var.	9.59	5.14

In both settings, SUBS has better mean than ORI, which is consistent with the results in this paper, since SUBS is expected to be bias-free while ORI is not. However, for setting 1 where all units have the same interaction chance, SUBS has a larger variance than ORI. The large variance is mainly due to sample size issues. In setting 1, SUBS on average selected 22.363 units for each iteration, while ORI had many more available units: it used all 100

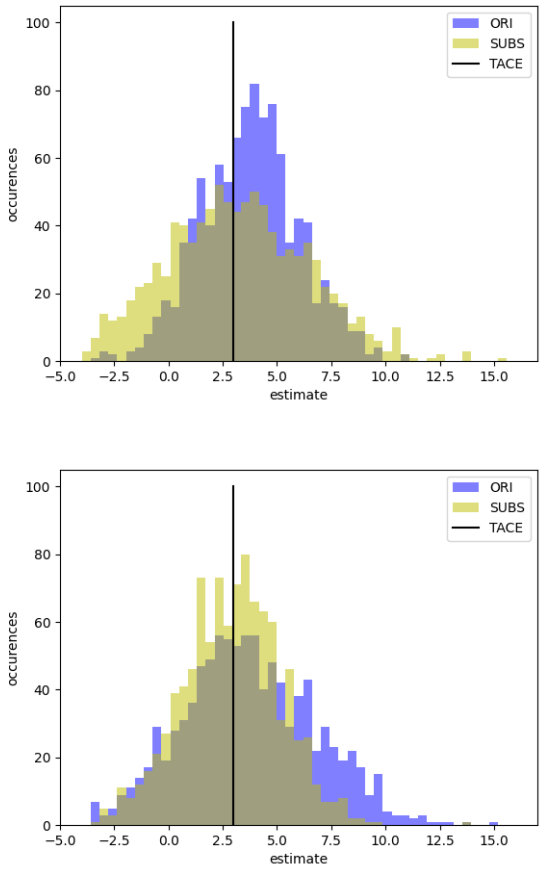


Figure 5.11: Comparison of ORI and SUBS, without restricting sample sizes.

units. So I also tested the results estimated if they had the same sample size. The results are in the following plots and tables.

From the results comparing ORI and SUBS with the same sample size, we see that SUBS subsumes ORI in both settings, and for both mean and variance. Note that the current version of SUBS is a basic one that greedily selects units, which is unlikely to select the largest bias-free subset. It is straightforward to develop efficient algorithms that can select subsets such that more number of samples are used. This would make it comparable to the sample size and substantially improve the performance. One candidate algorithm is provided in the appendix. In addition, as I have shown, if the units vary a lot in the

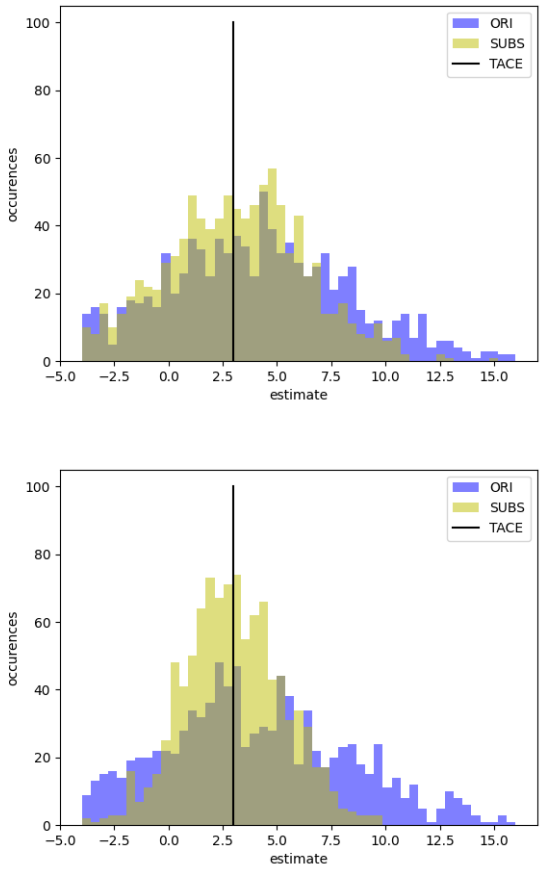


Figure 5.12: Comparison of ORI and SUBS, with same sample size.

possibility of involved in an interaction, SUBS is also able to select more units ($\sim$ 44 vs. $\sim$ 22 in the experiments).

5.6.2.2 General Non-parametric with Interactions

For the non-parametric case, the bias terms are multiplied instead of added, and the TACE is set to 6. The results are in Tables 5.5 and 5.6. The histograms are in the appendix.

I omit the comparison of the results obtained if the sample size were the same since SUBS already has better variance in this case even with smaller sample sizes.

Table 5.3: Restricted additivity & setting 1, with same sample size.

	ORI	SUBS
mean	3.64	2.97
var.	23.12	11.51

Table 5.4: Restricted additivity & setting 2, with same sample size.

	ORI	SUBS
mean	3.78	3.02
var.	22.79	5.34

Table 5.5: Non-parametric & setting 1.

	ORI	SUBS
mean	1602.86	5.98
variance	318153430.56	5.82

Table 5.6: Non-parametric & setting 2.

	ORI	SUBS
mean	5587.52	5.98
variance	2632262981.40	4.81

5.7 Summary

The main points of this chapter are as follows. I first developed non-parametric interaction framework for analyzing bias induced by non-IID data in estimating causal effects. I showed that in non-parametric models bias is inevitable given non-IID data. Next, I derived for restricted additivity models, the graphical condition where blindly applying IID methods would result in bias, which is existence of deflecting or reflecting bias structures. Between these, reflecting bias *e.g.*, $X_i \to Y_j \to Y_i$ is the more harmful one. I then developed debiasing procedures for both the non-parametric setting and the restricted additivity setting. Finally, I ran simulated experiments to test the proposed debiasing procedures for various setups.

CHAPTER 6

Causal Identification under Partial Interference

6.1 Introduction

While interactions among subjects can complicate causal identification, sometimes it may be defensible to assume that certain causal effects apply equally to each subject. For example, a person getting vaccinated reduces the chance of them getting a contagious disease, and in turn reduces the chance of people around them getting the disease. However, the effect of vaccination on one's own chance of getting the disease might be assumed to be same. A person who smoke can affect the health condition of people who live together with them, such effect might be the same as how this person is affected by other smokers. I refer to such equality of effects as *equality constraints* in the following text.

Currently there is no known efficient algorithm that is able to systematically exploit such equality constraints for identification.[1] While in the past few decades significant progress has been made in developing efficient identification algorithms for linear causal models [BP12, FDD12, CKB17, WRD18, KCB19, KCB20], such techniques can only systematically handle two types of assumptions encoded in a causal diagram: (i) the absence of a direct effect between certain variables; and (ii) the absence of association between error terms.

As a result, the current literature handling equality constraints has mostly worked with *ad-hoc* structures on a case-by-case basis. For example, [KS19] discuss the gain-score method

[1] One could use methods from computer algebra [GSS10], but these are often computationally intractable, making it practically infeasible for models larger than 4 or 5 nodes.

for solving certain models; [Cha13, Cha19] provides a more general method in which difference-in-differences is a special case, but still restricted to few cases; and while [CPC18] demonstrate that benchmarking in sensitivity analysis can be reduced to an identification problem with equality constraints, they only do so for specific model structures.

In this chapter, I show how to utilize equality in causal effects to handle causal identification with non-IID (interacting) subjects. The proposed identification method can be applied to real-world applications other than the non-IID settings.

6.2 Interference

A common assumption made when dealing with non-IID data is *partial interference*. Partial interference is a specific type of interference that splits the population into "blocks" (usually with the same number of units per block) such as, for instance, a household [Sob06, Ros07, HH08, TV12]. [OV14] demonstrate how interference in such cases can be represented and solved graphically. Partial interference assumes that interactions only occur between two units that belong to the same block. In addition, partial interference requires corresponding units in different blocks to satisfy the "identical" condition in IID. Thus partial interference methods assume "block IID," which is weaker than "unit IID" assumed by traditional causal

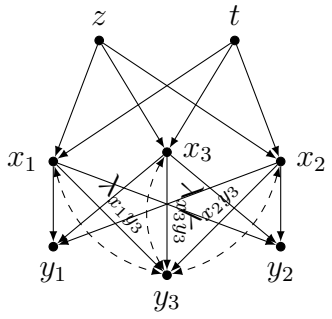


Figure 6.1: The assumption that x_1 and x_2 have equal effects on y_3 allows the identification of $\lambda_{x_1y_3}$, $\lambda_{x_2y_3}$, and $\lambda_{x_3y_3}$. Bidirected edges between other x_i and y_j omitted for clarity.

methods.

Existing approaches handling interference usually do not handle unobserved confounders, which complicates identification and, in some cases, makes it impossible. Luckily, if equality constraints can be defended, they can help identification even under the presence of confounding. For instance, perhaps one could argue that the effect of the treatment on the outcome should be equivalent for subjects within the block. Alternatively, one could also surmise that effects of one subject on another subject (known as *spillover effects*) are similar within the block.

Figure 6.1 graphically depicts the interference structure within a block [OV14], where three subjects are interfering with one another. In this case, x_1, x_2, and x_3 represent treatments for different subjects, and y_1, y_2, and y_3 represent their outcomes; z and t are two instrumental variables (e.g, randomized incentive for taking the treatment) applied to the whole block (e.g, the household). Here, if one posits that the effects of x_1 and x_2 on y_3 are the same, this enables the identification of $\lambda_{x_1y_3}$, $\lambda_{x_2y_3}$, and $\lambda_{x_3y_3}$.

Of course, such strict equality may not always be assumed. In these cases, one could relax the degree of equality, and obtain bounds on the causal effects instead of point identification.

6.3 Problem Setup

For the example presented in the previous section, the identification of causal effects of interest only becomes possible when equality amongst certain structural parameters is known a priori. In this section, I formally define the problem of identification using equality constraints.

I first define C-identifiability, denoting identifiability of model parameters of a linear SCM, M, given a set of external constraints C, beyond those already induced by the causal graph G.

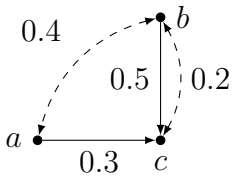


Figure 6.2: Numerical example of C-identifiability. In this model, λ_{ac} and λ_{bc} are not identified just with the constraints provided by the DAG. However, they become identified if the constraint $c\lambda_{ac} + \lambda_{bc} = 0$ is added.

Definition 20 (C-identifiability)**.** *Let M be a linear SCM (as specified by G) and let C be a set of additional constraints on the parameters of M. A causal quantity θ is said to be C-identifiable if θ is uniquely computable from C and the covariance matrix of M.*

In this paper, I consider the problem of C-identifiability specifically when C is composed of equality constraints on two structural parameters where one parameter is a multiple of the other. I restrict attention to two edges because this is the type of equality constraint of interest in the applications cited, and also the main focus of our results in Section 6.5.

Formally, we have the following definition.

Definition 21 (External Equality Constraint)**.** *An* external equality constraint *for a model M is a constraint of the form*

$$c\theta_1 + \theta_2 = 0, \tag{6.1}$$

where c is a constant, and θ_1 and θ_2 are structural parameters of M.

Here, the two structural parameters θ_1 and θ_2 can be two directed edges, two bidirected edges, or one directed edge and another bidirected edge. In fact, benchmarking in sensitivity analysis involves constraints where directed edges are equal to bidirected edges. I discuss that in detail in Section 6.6.

I use an example to illustrate the idea of C-identification. Suppose we are given the SCM of Figure 6.2. If we do not know the value of any of the edges, and we are given only the

graph as well as the correlations among the three variables, then neither λ_{ac} nor λ_{bc} can be identified. To demonstrate, λ_{ac}, λ_{bc}, ε_{ab}, ε_{bc} could be $0.4, 0.25, 0.4, 0.41$ respectively, and this model implies the same correlations as those of the SCM of Figure 6.2 (this can be easily checked using Wright's rules). However, if we know the equality constraint between λ_{ac}, λ_{bc}, i.e., $-5/3\lambda_{ac} + \lambda_{bc} = 0$, then we can uniquely solve for λ_{ac} and λ_{bc}. Thus, λ_{ac} and λ_{bc} are not identifiable but are both C-identifiable with C being $-5/3\lambda_{ac} + \lambda_{bc} = 0$.

As we see, the goal is to find cases where the provided external equality constraints can supplement the limited information we have from the graph G alone, and thus help with the identification of more structural parameters of the model. As I discuss next, I tackle this problem by finding the linear constraints induced by the graph G and combining them with the external equality constraints C. This allows the construction of a system of linear equations that can solve for the parameters of interest.

6.4 Searching for Graph-Induced Linear Constraints

Given a DAG G and the covariance matrix of the modeled variables, some relationships between structural parameters can be deduced. Here I am interested in finding linear equations among the structural parameters, since these equations can be used to solve for the structural parameters using linear algebra. In this section, I provide graphical conditions to find such linear constraints on the graph.

I first formally define this type of linear relationship, which I name *graph-induced linear constraint*.

Definition 22. *Let $\theta_1, \ldots, \theta_p$ be structural parameters of a linear model M. If the graph $G = (V, D, B)$ induces a linear equation of the type,*

$$l_{\theta_1,\ldots,\theta_p} := a_1\theta_1 + a_2\theta_2 + \cdots + a_p\theta_p = c$$

where $a_1 \ldots a_p$ and c are functions of Σ, then we say $l_{\theta_1,\ldots,\theta_p}$ is a graph-induced linear con-

straint on $\theta_1 \ldots, \theta_p$ from G.

One way to search for graph-induced linear constraints is through searching for generalized instrumental sets (IV sets) of [BP12]. If an IV set can be found in the graph, one can then use them to construct a full-rank system of linear equations on certain structural parameters. Instead of aiming for a full-rank system that guarantees point identification, the basic idea of our method is simply to search for such linear relationships among edges, even if we cannot have as many equations as there are unknowns (here including directed and bidirected edges).

For example, in Figure 6.1, if we search for a generalized instrumental set on the edges $\lambda_{x_1y_3}, \lambda_{x_2y_3}, \lambda_{x_3y_3}$, I will not be able to find one, since there are only two possible instruments, z and t, while all other variables violate the requirements for a generalized instrumental set. However, although it is not possible to identify any of the three edges, we can still construct two linear constraints on these three edges:

$$\rho_{zy_3} = \rho_{zx_1}\lambda_{x_1y_3} + \rho_{zx_3}\lambda_{x_3y_3} + \rho_{zx_2}\lambda_{x_2y_3} \tag{6.2}$$

$$\rho_{ty_3} = \rho_{tx_1}\lambda_{x_1y_3} + \rho_{tx_3}\lambda_{x_3y_3} + \rho_{tx_2}\lambda_{x_2y_3} \tag{6.3}$$

Now note that those linear constraints can still be used to identify the three edges, provided we have a third external equality constraint to supplement the missing information.

Below I define *partial-instrumental sets*, which relaxes the traditional definition of generalized instrumental sets of [BP12], by allowing the inclusion of a larger set of directed and bidirected edges.

Definition 23 (Partial-Instrumental Set)**.** *In a graph $G = (V, D, B)$, let y be a variable in V and let E be a set of n edges where $E \subseteq Inc(y)$. Given a set of n' edges, $E' = \{e_1, e_2, \ldots, e_{n'}\}$ where $E' \subseteq E$, and a set of n' variables, $Z = \{z_1, z_2, \ldots, z_{n'}\}$, Z is a* partial-instrumental set *for E on E' if there exists triples $(z_1, W_1, p_1), \ldots, (z_{n'}, W_{n'}, p_{n'})$ such that:*

1. *For $i = 1, \ldots, n'$, the elements of W_i are non-descendants of y, and either:*

(a) $(z_i \perp\!\!\!\perp y | W_i)_{G_{(E \cap D)-}}$, *or*

(b) *if there exists a bidirected edge between* z_i *and* y*:* ε_i*, and* $\varepsilon_i \in E$*,* W_i *are non-descendants of* z_i*, and* $(z_i \perp\!\!\!\perp y | W_i)_{G_{(E \cap D) \cup \{\varepsilon_i\}-}}$.

2. *for* $i = 1, \ldots, n'$*,* p_i *is a path between* z_i *and* y *that is not blocked by* W_i *and passes through* e_i*, and*

3. *for* $1 \leq i < j \leq n'$*, variable* z_j *does not appear in path* p_i*, and if paths* p_i *and* p_j *have a common variable* v*, then both* $p_i[v \sim y]$ *and* $p_j[z_j \sim v]$ *point to* v*.*

In this definition, the set of edges, E, contains the edges we are interested in solving for and might not be able to be removed from consideration by conditioning. Note $|E'|$ number of linear constraints on E can be generated if such a partial-instrumental set exists. The set of edges, E', is considered a "critical set" for the constraints generated, where each constraint matches to an edge in E'. For each i in $1, \ldots, n'$, I say that the constraint l_i generated from z_i "matches to" the edge e_i. The constraint l_i matching to e_i indicates that l_i has additional information about e_i, which cannot be deduced from other constraints.

I explain the matching between constraints and edges using the example in Figure 6.3. Starting with Figure 6.3(a), using Wright's rules, we can find two linear equations on the edges, $\lambda_{x_1 y}$ and $\lambda_{x_2 y}$. They are

$$l_1 : \rho_{z_1 x_1} \lambda_{x_1 y} + \rho_{z_1 x_2} \lambda_{x_2 y} = \rho_{z_1 y} \tag{6.4}$$

$$l_2 : \rho_{z_2 x_1} \lambda_{x_1 y} + \rho_{z_2 x_2} \lambda_{x_2 y} = \rho_{z_2 y} \tag{6.5}$$

When we have two graph-induced linear constraints on two parameters, we have a system of linear equations to solve for both parameters. However, now moving to Figure 6.3(b), note that here the two equations are in fact "equivalent," since the coefficients ($\rho_{z_1 x_1}, \rho_{z_1 x_2}$, and $\rho_{z_1 y}$) in Eq. (6.4) multiplied with $\lambda_{z_2 z_1}$ are equal to the corresponding coefficients ($\rho_{z_2 x_1}, \rho_{z_2 x_2}$, and $\rho_{z_2 y}$) in Eq. (6.5). The reason behind this is, given z_1, there is no additional information z_2 can provide on $\lambda_{x_1 y}$ or $\lambda_{x_2 y}$, because z_2 is connected to $\lambda_{x_1 y}$ or $\lambda_{x_2 y}$ only through z_1. Hence,

(a) Distinct constraints

(b) Redundant constraints

Figure 6.3: Different numbers of independent linear constraints can be constructed in different graphs.

l_2 cannot be "matched to" λ_{x_1y} or λ_{x_2y}, which makes z_2 an invalid candidate instrument when z_1 is present. Condition 3 in Definition 23 is used to guarantee that each constraint generated will have unique information on one edge in E', since it disallows the path for one instrument to subsume the path for another instrument.

Nevertheless, Figure 6.3(b) is still an example of partial-instrumental set. One possible choice of Z, E, E' is $Z = \{z_2\}$, $E = \{\lambda_{x_1y}, \lambda_{x_2y}\}$, $E' = \{\lambda_{x_1y}\}$, so that Z is a partial-instrumental set for E on E'. In this case, $W_1 = \emptyset$ and p_1 is $z_2 \to z_1 \to x_1 \to y$. In other words, although we cannot solve the system, we can still extract one non-redundant linear equation on the two parameters. This may still be useful, as such equation may be combined with an external equality constraint on those parameters to build a full-rank system of equations.

Another example is given in Figure 6.4. If we define $Z = \{z_1, z_2, z_3, z_4\}$, $E = \{\lambda_{x_1y}, \lambda_{x_2y}, \lambda_{x_3y}, \varepsilon_{z_2y}, \varepsilon_{z_4y}\}$, and $E' = \{\lambda_{x_1y}, \varepsilon_{z_2y}, \lambda_{x_2y}, \lambda_{x_3y}\}$, then Z is a partial-instrumental set for E on E'. The constraints generated from z_1, z_2, z_3, z_4 are matched to $\lambda_{x_1y}, \varepsilon_{z_2y}, \lambda_{x_2y}, \lambda_{x_3y}$, respectively, with conditioning sets $W = \{\{a, b\}, \{b\}, \emptyset, \emptyset\}$, and the paths $P = \{z_1 \to x_1 \to y, z_2 \leftrightarrow y, z_3 \to z_2 \to x_2 \to y, z_4 \to x_3 \to y\}$.

Note that when $E' = E$ and E contains only directed edges, Definition 23 degenerates to the traditional generalized instrumental set. Lemma 2 below states that we can construct

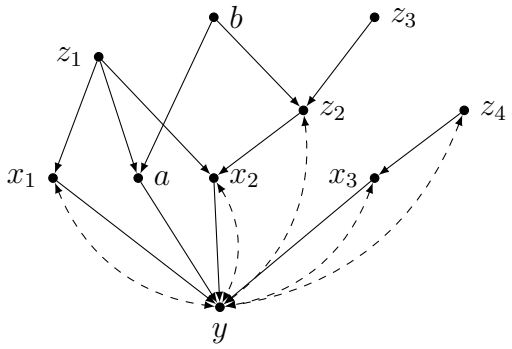


Figure 6.4: It is possible to construct 4 linear equations on 5 edges, $E = \{\lambda_{x_1y}, \lambda_{x_2y}, \lambda_{x_3y}, \varepsilon_{z_2y}, \varepsilon_{z_4y}\}$.

graph-induced linear constraints on edges in E, which might contain both directed and bidirected edges. The number of constraints constructed, $|E'|$, might be fewer than the number of edges involved in the equations, when E' is a strict subset of E. For example, for the DAG in Figure 6.4, we can construct 4 linear equations on 5 edges.

Lemma 2. *For an SCM M with graph $G = (V, D, B)$, if there exists a partial-instrumental set $Z = \{z_1, \ldots, z_{n'}\}$ for $E = \{\theta_1, \ldots, \theta_n\}$ on E' where $|E| = n$ and $|Z| = |E'| = n'$, then there exists a set of n' graph-induced linear constraints on E. Specifically, given the triples in Definition 23 as $(z_1, W_1, p_1), \ldots, (z_{n'}, W_{n'}, p_{n'})$, for each $i = 1, \ldots, n'$, we have a constraint,*

$$l_i : \rho_{z_iy \cdot W_i} = c_{i1}\theta_1 + \cdots + c_{in}\theta_n, \tag{6.6}$$

where c_{ij} is a function on the correlations of variables in M for all $j = 1, \ldots, n$.

See [BP12] for how to compute the coefficients $c_{i1}, \ldots, c_{in}$.

6.5 Incorporating External Equality Constraints

Given a set of linear constraints, it is important to check for the uniqueness of such constraints given the model M—is a newly found constraint equivalent to a previously found one, or can it be deduced from several previously found ones? In other words, what are the criteria

for a set of constraints to be "full-rank?" This question becomes harder when external equality constraints and known edges are provided, since it is not trivial to decide whether one constraint can be a linear combination of several other constraints of any type. As discussed, each constraint of an instrumental set can be "matched to" an edge. The same idea applies to partial-instrumental sets, where more edges of both types are involved. I now show that we can also apply this simple strategy when combining graph-induced linear constraints with external equality constraints and known edges.

To begin with, we have the following lemma.

Lemma 3. *Given only n' constraints constructed in Lemma 2 from the partial-instrumental set Z for E on E', no edge in $E \setminus E'$ can be solved.*

The correctness of this lemma is evident for the reason that, if an edge is not "matched to" by any constraint constructed from a partial-instrumental set, then it cannot be solved given those constraints. In other words, the value of any variable in $E \setminus E'$ cannot be deduced from $\mathcal{L}$. Hence, we can combine external information on the edges $E \setminus E'$ with the constraints of $\mathcal{L}$, without worrying about such external constraints being redundant.

Building on top of this, we have the main theorem of this paper. Theorem 10 provides a sufficient condition that, when satisfied, guarantees a full-rank set system of linear equations can be constructed by combining a set of graph-induced linear constraints, external equality constraints, and the values of known edges.

Theorem 10. *For an SCM M with graph $G = (V, D, B)$, let y be a variable in V and let E be a set of n edges where $E \subseteq Inc(y)$. Suppose there exists a partial-instrumental set, Z, for E on E' where $|Z| = |E'| = n'$, and we are given the following external information:*

1. *a set of n_k edges, $E_k \subseteq E$, whose coefficients are known, and*
2. *a set of n_e linearly independent external equality constraints, L_e, on edges E_e, where $E_e \subseteq E$.*

If $E_k \cap E' = \emptyset$, and there exists a way to simultaneously select one edge from each constraint $l \in L_e$ such that the selected edges 1) are not repetitive, 2) do not contain any edge in $E' \cup E_k$, then there exists a full-rank set of $n' + n_k + n_e$ linear constraints on E.

The intuition behind Theorem 10 is that a full-rank set of constraints can be constructed if we can find an edge for each constraint, where that constraint contains some unique information on that edge. Specifically, the constraints are given by: n' constraints constructed from the partial-instrumental set as in Lemma 2, n_k constraints in the form of $\theta_i = c_i$ where $\theta_i \in E_k$ and c_i is the known value of θ_i for $i = 1, \ldots, n_k$, and n_e external equality constraints. A special case of Theorem 10 is when there exists no partial-instrumental set—we can still construct a full-rank constraint set from known edges and equality constraints only. For example, given $\theta_1 = \theta_2$ and $\theta_1 = k$, they form a full-rank set and we immediately have $\theta_2 = k$.

An immediate result from Theorem 10 is that when $n' + n_k + n_e = |E|$, i.e., the number of linear constraints we can find is equal to the number of structural parameters E that those constraints are on, then we can solve for all the structural parameters in E. I use Figure 6.5 to show how to apply Theorem 10. The set of variables $Z = \{z_1, z_2, z_3\}$ is a partial-instrumental set for $E = \{\lambda_{x_1y}, \lambda_{x_2y}, \varepsilon_{x_2y}, \varepsilon_{z_3y}\}$ on $E' = \{\lambda_{x_1y}, \varepsilon_{x_2y}, \varepsilon_{z_3y}\}$. We can construct three constraints using the instruments z_1, z_2, z_3, and those constraints are matched to $\lambda_{x_1y}, \varepsilon_{x_2y}, \varepsilon_{z_3y}$, with the paths $z_1 \to z_2 \to x_1 \to y, z_2 \leftrightarrow y, z_3 \leftrightarrow y$, respectively.

Now I analyze different possible types of external information given. Let k, l, m denote constants:

1. the constraint $\varepsilon_{x_2y} = k\varepsilon_{x_3y}$ cannot be combined with our graph-induced linear constraints, since both ε_{x_2y} and ε_{z_3y} are in E', and there is no way to select an edge from this equality constraint that is not in $E' \cup E_k$;

2. the constraint $\lambda_{x_1y} = l\lambda_{x_2y}$ can be combined with our graph-induced linear constraints, since λ_{x_2y} is not in E', so we can select the edge λ_{x_2y} from this equality constraint that

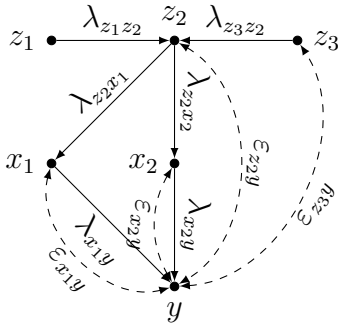


Figure 6.5: Variables z_1, z_2, z_3 form a partial instrument set for $E = \{\lambda_{x_1y}, \lambda_{x_2y}, \varepsilon_{x_2y}, \varepsilon_{z_3y}\}$ on $E' = \{\lambda_{x_1y}, \varepsilon_{x_2y}, \varepsilon_{z_3y}\}$.

is not in $E' \cup E_k$;

3. the constraint $\lambda_{x_2y} = m$ (either from previous identification or prior knowledge) can be combined with our graph-induced linear constraints, since λ_{x_2y} is not in E', so $E_k \cap E' = \emptyset$.

The three graph-induced linear constraints are:

$$\rho_{z_1y} = \rho_{z_1x_1}\lambda_{x_1y} + \rho_{z_1x_2}\lambda_{x_2y} \tag{6.7}$$

$$\begin{aligned}\rho_{z_2y\cdot\{z_3\}} = &\frac{\rho_{z_2x_1} + \rho_{z_3x_1}}{(1-\rho^2_{z_2z_3})^{1/2}(1-\rho^2_{z_3y})^{1/2}}\lambda_{x_1y} \\ &+ \frac{\rho_{z_2x_2} + \rho_{z_3x_2}}{(1-\rho^2_{z_2z_3})^{1/2}(1-\rho^2_{z_3y})^{1/2}}\lambda_{x_2y} \\ &+ \frac{1}{(1-\rho^2_{z_2z_3})^{1/2}(1-\rho^2_{z_3y})^{1/2}}\varepsilon_{x_2y}\end{aligned} \tag{6.8}$$

$$\rho_{z_3y} = \rho_{z_3x_1}\lambda_{x_1y} + \rho_{z_3x_2}\lambda_{x_2y} + \varepsilon_{z_3y} \tag{6.9}$$

By Wright's rules, all three equations above have the equal ratio of the coefficient for λ_{x_1y} to the coefficient for λ_{x_2y}. Hence, λ_{x_1y} and λ_{x_2y} can be eliminated together, and ε_{x_2y} and ε_{z_3y} can thus both be solved. This again explains why we cannot combine the external information $\varepsilon_{x_2y} = k\varepsilon_{z_3y}$ with the three graph-induced linear constraints: this external constraint can be deduced from the three graph-induced linear constraints. On the other hand, if the external

information is $\lambda_{x_1y} = k\lambda_{x_2y}$, since neither edge can be solved from the graph-induced linear constraints, the equality constraint cannot be deduced from the system, and is therefore not redundant.

Though in this paper I present our method based on generalized instrumental sets, I conjecture that this approach can be generalized to combine with most of existing linear causal identification methods. This is due to the nature of identification methods for linear models, most of which construct a system of linear equations to solve for a set of structural parameters, E. I hence believe that we can match each equation to one parameter in E as required by Theorem 10. Proving this conjecture is beyond the scope of this paper and I leave it for future work.

6.6 Case Studies

In this section, I revisit the example in Section 6.2 and show how the proposed method can be used to solve it. I also show how it is useful in other important real-world applications.

6.6.1 Interference

For the interference example in Figure 6.1, there is an equality constraint $\lambda_{x_1y_3} = \lambda_{x_2y_3}$. $Z = \{z, t\}$ is a partial-instrumental set for $E = \{\lambda_{x_1y_3}, \lambda_{x_2y_3}, \lambda_{x_3y_3}\}$ on $E' = \{\lambda_{x_1y_3}, \lambda_{x_3y_3}\}$. Note that here we can either choose E' to be $\{\lambda_{x_1y_3}, \lambda_{x_3y_3}\}$ or $\{\lambda_{x_2y_3}, \lambda_{x_3y_3}\}$ but not $\{\lambda_{x_1y_3}, \lambda_{x_2y_3}\}$. Otherwise, the equality constraint has no edge to select from for it to match to, so the condition in Theorem 10 will fail. If we choose $E' = \{\lambda_{x_1y_3}, \lambda_{x_3y_3}\}$, we have a full-rank set of three equations on $\lambda_{x_1y_3}, \lambda_{x_2y_3}, \lambda_{x_3y_3}$, with two graph-induced linear constraints generated from z and t, and one external equality constraint.

6.6.2 Equiconfounding

A number of identification techniques use the assumption of equiconfounding, where observed variables are equally affected by an unobserved confounder. [Cha13] discusses some special types of equiconfounding where point-identification is possible, including when two joint responses are equiconfounded, and when two causes and one response are equiconfounded.

The most widely applied special case of equiconfounding is "difference-in-differences" [AP09, KS19], which assumes two joint responses are equally affected by unobserved confounding. A commonly cited example involves estimating the effect of raising the minimum wage on unemployment. In this case, the change in employment after minimum wage was increased in New Jersey (NJ) was compared to the change in employment in Pennsylvania (PA) over the same time period, where minimum wage was not changed [CK94]. The usual structure can be depicted as in Figure 6.6, where x represents minimum wage, y represents unemployment after the change in minimum wage, w represents unemployment before the change in minimum wage, and u represents the unobserved confounder. The equality constraint of this model is that $\lambda_{uw} = \lambda_{uy}$ (DAG on the left), without which the causal effect is not identifiable. In the latent projection [Pea09] (DAG on the right) the equality constraint becomes $\varepsilon_{xw} = \varepsilon_{xy}$.

Figure 6.6: When two joint responses are equiconfounded ($\lambda_{uy} = \lambda_{uw}$) this can aid in identification. Left: Latent variable DAG. Right: Latent projection.

As discussed in [Cha13, Chapter 4], another common case of equiconfounding happens when two joint causes and one response are affected by the unobserved confounder by the same or proportional magnitude. For example, in Figure 6.7 (left), we have an equality

constraint on three edges, $\lambda_{ux_1} = \lambda_{ux_2} = \lambda_{uy}$ (in the latent projection (right), this translates to the equality constraint $\varepsilon_{x_1x_2} = \varepsilon_{yx_1} = \varepsilon_{x_2y}$). The causal effect λ_{x_1y} is not identifiable without this constraint.

Figure 6.7: If two joint causes and one response are equiconfounded ($\lambda_{ux_1} = \lambda_{ux_2} = \lambda_{uy}$) this enables identification. Left: Latent variable DAG. Right: Latent projection.

The first example I showed is Figure 6.6, the well-known "difference-in-differences" graph, or the case when two joint responses are equiconfounded [Cha13]. First, we can see that the bidirected edge, ε_{xw} is identified in this latent projection DAG and is equal to ρ_{xw}, since $x \leftrightarrow w$ is the only unblocked path between x and w. So we can plug it in to the equality constraint, $\varepsilon_{xw} = \varepsilon_{xy}$ and get $\varepsilon_{xy} = \rho_{xw}$. Next, we see that $Z = \{x\}$ is a partial-instrumental set for $E = \{\lambda_{xy}, \varepsilon_{xy}\}$ on $E' = \{\lambda_{xy}\}$, so we have the graph-induced linear constraint $\lambda_{xy} + \varepsilon_{xy} = \rho_{yx}$. Together with the known edge constraint $\varepsilon_{xy} = \rho_{xw}$, we have a full-rank set of two constraints on two variables, and λ_{xy} can be solved, which gives $\lambda_{xy} = \rho_{yx} - \rho_{xw}$.

The second example is when two joint causes and one response are equiconfounded, as in Figure 6.7. This case is similar to the previous one. $\varepsilon_{x_1x_2}$ can be identified ($\varepsilon_{x_1x_2} = \rho_{x_1x_2}$), and plugging into the equality constraint identifies ε_{x_2y} and ε_{yx_1}. Next, we observe that $Z = \{x_1, x_2\}$ is a partial-instrumental set for $E = \{\varepsilon_{x_2y}, \lambda_{x_1y}, \lambda_{x_2y}, \varepsilon_{yx_1}\}$ on $E' = \{\lambda_{x_1y}, \lambda_{x_2y}\}$. As a result, we have a full-rank set of four constraints, including two graph-induced constraints and two constant (known edge) constraints, and we can thus solve for all the four edges in E.

Another more complex example, [Cha13, Graph 2], can be solved similarly using our method, and I skip the discussion of that. Our method can solve all the cases where point

identification is possible in [Cha13]. I can also solve other simple generic cases of equiconfounding which have not been discussed in [Cha13]. For instance, by replacing the equality constraint with $\lambda_{ux} = \lambda_{uw}$ or $\lambda_{ux} = \lambda_{uy}$ in Figure 6.6 left, I have two different examples that I can both solve. I leave the discussion to the appendix.

6.6.3 Benchmarking in Sensitivity Analysis

Causal inference requires knowledge or assumptions about the data generating process, and sensitivity analysis aims to understand the extent of bias when these assumptions are violated [Ros10, Ros17]. Often, these violations render the causal effect of interest unidentifiable, and, therefore, additional constraints are needed to identify the causal effect and derive the bias [CKC19].

A common practice is to "benchmark" the extent to which the assumption is violated [CH20]. For example, if we want to assess the sensitivity of our estimate to omitted variable bias, we might ask what the bias would be if the missing confounder were as strong as an observable confounder. One could then argue that, as long as the strongest confounders have been accounted for, this value represents an upper bound on the potential bias due to a missing variable.

Solving this problem again reduces to identification in the presence of an equality constraint. For example, suppose that we wanted to determine the bias if an unobserved confounder, depicted by the bidirected edge in Figure 6.8 right, were k times as strong as the observed confounder, z, for some known constant k. In this case, I posit that $\varepsilon_{xy} = k\lambda_{zx}\lambda_{zy}$. This equality constraint permits the identification of λ_{xy}, enabling us to compute the bias under this hypothesized relative strength of confounding.

For Figure 6.8, we have the external information that $\varepsilon_{xy} = k\lambda_{zx}\lambda_{zy}$. First notice that the edge, λ_{zx} can be identified using z as an instrument to itself, and we get $\lambda_{zx} = \rho_{zx}$. Hence, the equality constraint reduces to $\varepsilon_{xy} = k\rho_{zx}\lambda_{zy}$, which is now in the form of

Figure 6.8: Left: Original DAG. Right: Potential violation with unobserved confounders ε_{xy}. The assumption that $\varepsilon_{xy} = k\lambda_{zx}\lambda_{zy}$ allows identifying λ_{xy}.

$\theta_1 = k'\theta_2$ that our method can handle. We next examine the DAG and see the set $Z = \{x, z\}$ is a partial-instrumental set for $E = \{\varepsilon_{xy}, \lambda_{xy}, \lambda_{zy}\}$ on $E' = \{\lambda_{xy}, \lambda_{zy}\}$. We can thus construct two graph-induced linear constraints, as follows.

$$\varepsilon_{xy} + \lambda_{xy} + \rho_{zx}\lambda_{zy} = \rho_{xy} \tag{6.10}$$

$$\rho_{zx}\lambda_{xy} + \lambda_{zy} = \rho_{zy} \tag{6.11}$$

Together with the equality constraint $\varepsilon_{xy} = k\rho_{zx}\lambda_{zy}$, we have a full-rank set of linear constraints from Theorem 10, where the equality constraint is matched to the edge ε_{xy}. Note that this is just one possible choice of E', and we can also choose $E' = \{\varepsilon_{xy}, \lambda_{xy}\}$, where the equality constraint will be matched to λ_{zy}. Either way, we have three equations on three unknowns, and all of them are solved. Specifically,

$$\lambda_{xy} = \frac{\rho_{zx}\rho_{zy}(k+1) - \rho_{xy}}{(k+1)\rho_{zx}^2 - 1} \tag{6.12}$$

$$\lambda_{zy} = \frac{\rho_{xy}\rho_{zx} - \rho_{zy}}{(k+1)\rho_{zx}^2 - 1} \tag{6.13}$$

$$\varepsilon_{xy} = k\rho_{zx}\frac{\rho_{xy}\rho_{zx} - \rho_{zy}}{(k+1)\rho_{zx}^2 - 1}. \tag{6.14}$$

As we see, those parameters are point-identified if we know the value of k, which is how strong the unobserved confounder is compared to an observed confounder, z. If one does not know the exact value of k, but only its plausible range (for instance, $k \leq 2$), it is still possible to use this result to bound the target parameters.

6.7 Discussion and Summary

I developed a novel graphical criterion that allows researchers to leverage equality constraints for identification in linear systems. I showed how several apparently diverse problems including interference, difference-in-differences, and sensitivity analysis can be reduced to identification with equality constraints, consisting of special cases handled by our method. I hope the results of this paper can be used towards the construction of a systematic, algorithmic approach to exploit equality constraints in causal inference. Extensions to more general forms of equality constraints, and incorporating such results into state-of-the art linear identification algorithms are promising directions for future work.

CHAPTER 7

Concluding Remarks

Many existing machine learning and causal inference methods have relied on the data being IID. Blindly assuming that data are IID when in fact they are not, can potentially bias the outcome of a research study. Such bias can occur for the query: causal effect of treatment on outcome, when there exist bias structures in the interaction pattern. In this work, I focus on causal inference for the generalized interference problem. Under the linear setting or restricted additivity, bias structures include an open (not necessarily directed) path from the treatment of unit i to the outcome of unit j and/or to the outcome of unit i itself such that an intermediate node on the path belongs to unit j. Furthermore, only those two types of interaction structures can induce bias. Under the general non-parametric setting, any effect from a unit i to the outcome of another unit j induces bias. I also presented the bias-quantification formulas, which show what factors affect the size of the interaction bias. In the presence of interaction bias, it is still possible to compute an unbiased estimate by selecting a subset of samples $\mathcal{B}$ such that no biasing paths exist in the interaction graph corresponding to samples in $\mathcal{B}$. More importantly, such a debiasing procedure does not require the selection of IID samples and may contain interactions among them. Such a debiasing procedure can also be done in polynomial time. In the empirical analysis, I randomly generated interaction models and show that the bias can be huge if IID is wrongly assumed on data with generalized interference. The debiasing methods in this work yield unbiased estimates. Finally, with the partial interference assumption, I developed a causal identification method utilizing equality constraints that works even with unobserved confounders. The method advances solutions to other well-known problems including difference-in-differences and sensitivity analysis.

REFERENCES

[AM13] Peter M. Aronow and Joel A. Middleton. "A Class of Unbiased Estimators of the Average Treatment Effect in Randomized Experiments." *Journal of Causal Inference*, **1**(1):135–154, 2013.

[AP09] Joshua Angrist and Jörn-Steffen Pischke. *Mostly harmless econometrics: an empiricists guide*. Princeton: Princeton University Press, 2009.

[AS17] Peter M Aronow and Cyrus Samii. "Estimating average causal effects under general interference, with application to a social network experiment." *The Annals of Applied Statistics*, **11**(4):1912–1947, 2017.

[BMS20] Rohit Bhattacharya, Daniel Malinsky, and Ilya Shpitser. "Causal inference under interference and network uncertainty." In *Uncertainty in Artificial Intelligence*, pp. 1028–1038. PMLR, 2020.

[BP12] Carlos Brito and Judea Pearl. "Generalized instrumental variables." *arXiv preprint arXiv:1301.0560*, 2012.

[Bur87] Ronald S Burt. "Social contagion and innovation: Cohesion versus structural equivalence." *American journal of Sociology*, **92**(6):1287–1335, 1987.

[Cao14] Longbing Cao. "Non-IIDness Learning in Behavioral and Social Data." *The Computer Journal*, **57**(9):1358–1370, 2014.

[Cao22] Longbing Cao. "Beyond i.i.d.: Non-IID Thinking, Informatics, and Learning." *IEEE Intelligent Systems*, **37**(4):5–17, 2022.

[CH20] Carlos Cinelli and Chad Hazlett. "Making sense of sensitivity: extending omitted variable bias." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **82**(1):39–67, 2020.

[Cha13] Karim Chalak. *Identification Without Exogeneity Under Equiconfounding in Linear Recursive Structural Systems*, pp. 27–55. Springer New York, New York, NY, 2013.

[Cha19] Karim Chalak. "Identification of average effects under magnitude and sign restrictions on confounding." *Quantitative Economics*, **10**(4):1619–1657, 2019.

[CK94] David Card and Alan B Krueger. "Minimum wages and employment: A case study of the fast-food industry in New Jersey and Pennsylvania." *The American Economic Review*, **84**(4):772, 1994.

[CKB17] Bryant Chen, Daniel Kumor, and Elias Bareinboim. "Identification and model testing in linear structural equation models using auxiliary variables." In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 757–766. JMLR. org, 2017.

[CKC19] Carlos Cinelli, Daniel Kumor, Bryant Chen, Judea Pearl, and Elias Bareinboim. "Sensitivity Analysis of Linear Structural Causal Models." *International Conference on Machine Learning*, 2019.

[CMM22] Diego F Cuadros, Claudia M Moreno, Godfrey Musuka, F DeWolfe Miller, Phillip Coule, and Neil J MacKinnon. "Association between vaccination coverage disparity and the dynamics of the COVID-19 Delta and Omicron waves in the US." *Frontiers in Medicine*, **9**, 2022.

[Cox58] D.R. Cox. *Planning of Experiments*. Wiley Series in Probability and Statistics - Applied Probability and Statistics Section. Wiley, 1958.

[CPC18] Carlos Cinelli, Judea Pearl, and Bryant Chen. "When Confounders Are Confounded: Naive Benchmarking in Sensitivity Analysis.", Aug 2018.

[ETP22] David W Eyre, Donald Taylor, Mark Purver, David Chapman, Tom Fowler, Koen B Pouwels, A Sarah Walker, and Tim EA Peto. "Effect of Covid-19 Vaccination on Transmission of Alpha and Delta Variants." *New England Journal of Medicine*, 2022.

[FDD12] Rina Foygel, Jan Draisma, Mathias Drton, et al. "Half-trek criterion for generic identifiability of linear structural equation models." *The Annals of Statistics*, **40**(3):1682–1713, 2012.

[FSY22] Guihong Fan, Haitao Song, Stan Yip, Tonghua Zhang, and Daihai He. "Impact of low vaccine coverage on the resurgence of COVID-19 in Central and Eastern Europe." *One Health*, p. 100402, 2022.

[FZ20] Zahra Fatemi and Elena Zheleva. "Minimizing Interference and Selection Bias in Network Experiment Design." *Proceedings of the International AAAI Conference on Web and Social Media*, **14**(1):176–186, May 2020.

[GSS10] Luis D García-Puente, Sarah Spielvogel, and Seth Sullivant. "Identifying causal effects with computer algebra." In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence (UAI). AUAI Press*, 2010.

[HH08] Michael G Hudgens and M Elizabeth Halloran. "Toward causal inference with interference." *Journal of the American Statistical Association*, **103**(482):832–842, 2008.

[HLW21] Yuchen Hu, Shuangning Li, and Stefan Wager. "Average Direct and Indirect Causal Effects under Interference." *Biometrika*, 2021.

[Hol88] Paul W Holland. "Causal inference, path analysis and recursive structural equations models." *ETS Research Report Series*, **1988**(1):i–50, 1988.

[HT52] D. G. Horvitz and D. J. Thompson. "A Generalization of Sampling Without Replacement From a Finite Universe." *Journal of the American Statistical Association*, **47**(260):663–685, 1952.

[IR15] Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

[JPV20] Ravi Jagadeesan, Natesh S Pillai, and Alexander Volfovsky. "Designs for estimating the treatment effect in networks with interference." *The Annals of Statistics*, **48**(2):679–712, 2020.

[JW14] Richard Arnold Johnson, Dean W Wichern, et al. *Applied multivariate statistical analysis*, volume 6. Pearson London, UK:, 2014.

[JZB18] Amin Jaber, Jiji Zhang, and Elias Bareinboim. "Causal identification under Markov equivalence." *arXiv preprint arXiv:1812.06209*, 2018.

[KCB19] Daniel Kumor, Bryant Chen, and Elias Bareinboim. "Efficient Identification in Linear Structural Causal Models with Instrumental Cutsets." In *Advances in Neural Information Processing Systems*, pp. 12477–12486, 2019.

[KCB20] D. Kumor, C. Cinelli, and E. Bareinboim. "Efficient Identification in Linear Structural Causal Models with Auxiliary Cutsets." In *Proceedings of the 37th International Conference on Machine Learning*, volume 119. PMLR, 2020.

[KS19] Yongnam Kim and Peter M Steiner. "Gain scores revisited: A graphical models perspective." *Sociological Methods & Research*, p. 0049124119826155, 2019.

[LH14] Lan Liu and Michael G Hudgens. "Large sample randomization inference of causal effects in the presence of interference." *Journal of the american statistical association*, **109**(505):288–301, 2014.

[Lyo11] Russell Lyons. "The spread of evidence-poor medicine via flawed social-network analysis." *Statistics, Politics, and Policy*, **2**(1), 2011.

[MSC01] Miller McPherson, Lynn Smith-Lovin, and James M Cook. "Birds of a feather: Homophily in social networks." *Annual review of sociology*, **27**(1):415–444, 2001.

[NPB20] Razieh Nabi, Joel Pfeiffer, Murat Ali Bayir, Denis Charles, and Emre Kıcıman. "Causal inference in the presence of interference in sponsored search advertising." *arXiv preprint arXiv:2010.07458*, 2020.

[OV14] Elizabeth L Ogburn and Tyler J VanderWeele. "Causal diagrams for interference." *Statistical science*, **29**(4):559–578, 2014.

[Pea09] Judea Pearl. *Causality*. Cambridge university press, 2009.

[Pea17] Judea Pearl. "A linear "microscope" for interventions and counterfactuals." *Journal of causal inference*, **5**(1), 2017.

[Pea19] Judea Pearl. "The Seven Tools of Causal Inference, with Reflections on Machine Learning." *Commun. ACM*, **62**(3):54–60, 2019.

[PGJ16] Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. *Causal inference in statistics: A primer*. John Wiley & Sons, 2016.

[Qua12] David Quammen. *Spillover: animal infections and the next human pandemic*. WW Norton & Company, 2012.

[Ric96] Thomas Richardson. "A discovery algorithm for directed cyclic graphs." In *Proceedings of the Twelfth international conference on Uncertainty in artificial intelligence*, pp. 454–461. Morgan Kaufmann Publishers Inc., 1996.

[Ros07] Paul R Rosenbaum. "Interference between units in randomized experiments." *Journal of the American Statistical Association*, **102**(477):191–200, 2007.

[Ros10] Paul R Rosenbaum. *Design of observational studies*. Springer Series in Statistics, 2010.

[Ros17] Paul R Rosenbaum. *Observation and experiment: an introduction to causal inference*. Harvard University Press, 2017.

[RS02] Thomas Richardson and Peter Spirtes. "Ancestral graph Markov models." *Ann. Statist.*, **30**(4):962–1030, 08 2002.

[Rub77] Donald B Rubin. "Assignment to treatment group on the basis of a covariate." *Journal of educational Statistics*, **2**(1):1–26, 1977.

[Rub78] Donald B Rubin. "Bayesian inference for causal effects: The role of randomization." *The Annals of statistics*, pp. 34–58, 1978.

[SA17] Daniel L Sussman and Edoardo M Airoldi. "Elements of estimation theory for causal effects in the presence of network interference." *arXiv preprint arXiv:1702.03578*, 2017.

[SAH21] Fredrik Sävje, Peter M Aronow, and Michael G Hudgens. "Average treatment effects in the presence of unknown interference." *The Annals of Statistics*, **49**(2):673–701, 2021.

[Sch22] Bernhard Schölkopf. "Causality for machine learning." In *Probabilistic and Causal Inference: The Works of Judea Pearl*, pp. 765–804. 2022.

[SGS00] Peter Spirtes, Clark N Glymour, Richard Scheines, David Heckerman, Christopher Meek, Gregory Cooper, and Thomas Richardson. *Causation, prediction, and search*. MIT press, 2000.

[Sob06] Michael E Sobel. "What do randomized studies of housing mobility demonstrate? Causal inference in the face of interference." *Journal of the American Statistical Association*, **101**(476):1398–1407, 2006.

[SS18] Eli Sherman and Ilya Shpitser. "Identification and estimation of causal effects from dependent data." *Advances in neural information processing systems*, **31**, 2018.

[STA17] Ilya Shpitser, Eric Tchetgen Tchetgen, and Ryan Andrews. "Modeling interference via symmetric treatment decomposition." *arXiv preprint arXiv:1709.01050*, 2017.

[TFS21] Eric J Tchetgen Tchetgen, Isabel R Fulcher, and Ilya Shpitser. "Auto-g-computation of causal effects on a network." *Journal of the American Statistical Association*, **116**(534):833–844, 2021.

[TV12] Eric J Tchetgen Tchetgen and Tyler J VanderWeele. "On causal inference in the presence of interference." *Statistical methods in medical research*, **21**(1):55–75, 2012.

[VA13] Tyler J VanderWeele and Weihua An. "Social networks and causal inference." *Handbook of causal analysis for social research*, pp. 353–374, 2013.

[VOT12] Tyler J VanderWeele, Elizabeth L Ogburn, and Eric J Tchetgen Tchetgen. "Why and when" flawed" social network analyses still yield valid tests of no contagion." *Statistics, Politics, and Policy*, **3**(1), 2012.

[VP91] Thomas Verma and Judea Pearl. *Equivalence and synthesis of causal models*. UCLA, Computer Science Department, 1991.

[VTH12] Tyler J VanderWeele, Eric J Tchetgen Tchetgen, and M Elizabeth Halloran. "Components of the indirect effect in vaccine trials: identification of contagion and infectiousness effects." *Epidemiology (Cambridge, Mass.)*, **23**(5):751, 2012.

[VTH14] Tyler J VanderWeele, Eric J Tchetgen Tchetgen, and M Elizabeth Halloran. "Interference and sensitivity analysis." *Statistical science: a review journal of the Institute of Mathematical Statistics*, **29**(4):687, 2014.

[WRD18] Luca Weihs, Bill Robinson, Emilie Dufresne, Jennifer Kenkel, Kaie Kubjas Reginald McGee II, McGee II Reginald, Nhan Nguyen, Elina Robeva, and Mathias Drton. "Determinantal generalizations of instrumental variables." *Journal of Causal Inference*, **6**(1), 2018.

[Wri21] Sewall Wright. "Correlation and causation." *Journal of agricultural research*, **20**(7):557–585, 1921.

[ZMP22] Chi Zhang, Karthika Mohan, and Judea Pearl. "Causal Inference with Non-IID Data using Linear Graphical Models." In *Proceedings of Thirty-Sixth Conference on Neural Information Processing Systems*, 2022.

CHAPTER 8

Appendix

8.1 Supplemental Materials for Chapter 3

8.1.1 Example and Analysis for Algorithm 1

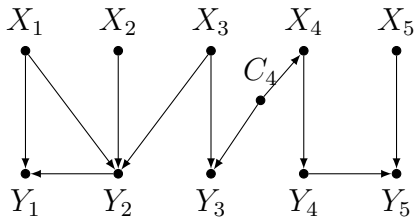


Figure 8.1: An interaction network of 5 individuals.

Example 3. *Input:* $t = 3$*, interaction network in Figure 8.1.*

Iteration 1: Units = [1, 5, 4, 2, 3]. $B = \{5, 2\}$.

Iteration 2: Units = [5, 1, 4, 3, 2]. $B = \{5, 3\}$.

Iteration 3: Units = [5, 2, 3, 1, 4]. $B = \{5, 2\}$.

The three choices of B *all have the same size 2. So the output is any of the three choices of* B.

Note that the subnetwork formed by 5 and 3 contains a bidirected path between Y_3 *and* Y_5 *(due to the path* $Y_3 \leftarrow C_4 \rightarrow X_4 \rightarrow Y_4 \rightarrow Y_5$*), and this does not constitute a bias structure.*

Complexity Analysis The time complexity is $O(tn^2d^p)$. d is the maximum degree of each node (how many other nodes a node is directly connected to), and p is the length (number of edges) of the longest simple path. This is polynomial if the degree is bounded.

Lemma 4. *The following two statements are equivalent. The first statement is used in this algorithm for simpler computation, and the second statement is used in the main text for easier understanding.*

1. *For each individual i in B, i has no deflecting bias structure in G^* with another individual j in B.*

2. *For each individual i in B, i has no deflecting bias structure in the latent projection of G^* on B.*

The definition of latent projection is by [Pea09], as follows.

Definition 24 (Projection[Pea09])**.** *A latent structure $L_{[O]} =< D_{[O]}, O >$ is a projection of another latent structure L if and only if:*

1. *every unobservable variable of $D_{[O]}$ is a parentless common cause of exactly two non-adjacent observable variables; and*

2. *for every stable distribution P generated by L, there exists a stable distribution P' generated by $L_{[O]}$ such that $I(P_{[O]}) = I(P'_{[O]})$.*

Proof of Lemma 4.

Proof. If statement 1 is false, then there exists an open path between X_i and Y_j in G^*, where $i, j \in B$. The latent projection contains both i and j so the open path still exists, which imply a deflecting bias structure in the latent projection.

If statement 2 is false, then there exists an open path between X_i and Y_j in the latent projection. This implies a deflecting bias structure in G^*. □

8.1.2 An Additional Simulation

Experiment: Subset Size of THM-2 We use same parameter settings as the previous experiment, except that we let $dRate$ and $rRate$ vary in $0.01, 0.1, 0.3, 0.5$. The subset sizes selected by THM-2 are in Table 8.1. Observe that as the graph gets denser (larger $dRate$ and $rRate$), THM-2 is unable to use most of the input samples. However, for the tests with samples ≥ 3, THM-2 yields very accurate estimates. Given that the ground truth is 100, **the estimates of THM-2 range between** 99.96 **and** 100.06**.**

		$dRate$			
		0.01	0.1	0.3	0.5
$rRate$	0.01	155	147	131	115
	0.1	26	24	23	23
	0.3	9	8	8	8
	0.5	5	4	3	0

Table 8.1: Each cell denotes the subset size selected using THM-2.

8.1.3 Proof of the Theorems

All lemmas and proofs are attached in Section 8.1.4 of the appendix.

Theorem 1. *Let $M^*(G^*, S^*)$ be a balanced interaction model in which treatment variable X_i and outcome variable Y_i are not confounded by any variable in $\mathcal{V}_i$, $\forall i$. Let D be the available data generated by M^* and let $G^\dagger$ be the generic network. Let $TACE_{XY}$ be identifiable in $G^\dagger$ and be given by β_{YX}, the regression coefficient of Y on X. Let α denote the true value of $TACE_{X,Y}$ in M^*. If X satisfies ASDC then the interaction bias is given by,*

$$\left|E[\hat{\beta_{YX}}] - \alpha\right| = \left|\frac{1}{n}\sum_{1\leq i\leq n}\sum_{p\in P[iji]} Val(p)\frac{\sigma^2_{R_p}}{\sigma^2_X} - \frac{1}{n(n-1)}\sum_{1\leq i\leq n}\sum_{p\in P[ji]} Val(p)\frac{\sigma^2_{R_p}}{\sigma^2_X}\right|,$$

where $P[iji]$ is the set of reflecting bias structures between X_i and Y_i through any explicit variable W_j of unit j with $i \neq j$, $P[ji]$ is the set of deflecting bias structures between X_j and Y_i with $i \neq j$, and R_p is the root of path p.

Proof. By Lemma 11,

$$\begin{aligned} & E[\hat{\beta_{YX}}] \\ =&\alpha \\ &+ \frac{1}{n}(\sum_{p\in\mathcal{P}} Val(p) + \sum_{1\leq i\leq n}\sum_{R\in(\mathcal{R}[iji]\setminus\{X_i\})} c_R\beta_{RX}) \\ &- \frac{1}{n(n-1)}\sum_{1\leq i\leq n}\sum_{R\in\mathcal{R}[ji]} c_R\beta_{RX}, \end{aligned}$$

where $\mathcal{P}$ is the set of directed paths from X_i to Y_i for any i passing through an intermediate node $W_j \in \mathcal{V}_{(j)}$, $i \neq j$, $\mathcal{R}[iji]$ is the set of roots of the open paths between X_i and Y_i through some W_j with $j \neq i$, $\mathcal{R}[ji]$ is the set of roots of the open paths between X_j and Y_i for $j \neq i$, and c_R is the sum of values of the directed paths from a variable R ($\in (\mathcal{R}[iji] \setminus \{X_i\})$ or $\in \mathcal{R}[ji]$) to Y_i not passing through any variable in $\mathcal{R}[iji] \cup \mathcal{R}[ji]$ for any $j \neq i$.

We prove this is equivalent to

$$\begin{aligned} & E[\hat{\beta_{YX}}] \\ =&\alpha \\ &+ \frac{1}{n}\sum_{1\leq i\leq n}\sum_{p\in P[iji]} Val(p)\frac{\sigma^2_{R_p}}{\sigma^2_X} \\ &- \frac{1}{n(n-1)}\sum_{1\leq i\leq n}\sum_{p\in P[ji]} Val(p)\frac{\sigma^2_{R_p}}{\sigma^2_X}, \end{aligned}$$

where $P[iji]$ is the set of open paths between X_i and Y_i through any $W_j \in \mathcal{V}_{(j)}$ with $i \neq j$, $P[ji]$ is the set of open paths between X_j and Y_i through any $W_j \in \mathcal{V}_{(j)}$ with $i \neq j$, and R_p is the root of path p.

We first check the term $\sum_{R\in\mathcal{R}[ji]} c_R\beta_{RX}$. For an R that is the root of a path between X_j and Y_i, since X satisfies ASDC, we must have $R \in \mathcal{V}_{(j)}$. Rename it as R_j. We also have

$\beta_{RX} = \sigma_{RX}/\sigma_X^2$. By Wright's Rules, σ_{RX} is equal to the sum of open path values between R and X times the variance of the root of that path. Recall that $R \in Anc(X)$, X satisfies ASDC, so R satisfies ASDC. So σ_{RX} is equal to the sum of open path values between R_j and X_j times the variance of the root of that path. We prove that each term that appears in $A = \sum_{p \in P[ji]} Val(p)\frac{\sigma_{R_p}^2}{\sigma_X^2}$ also appears in $B = \sum_{R \in \mathcal{R}[ji]} c_R \beta_{RX}$, and there is no extra term.

Each R_p in A is a root between X_j and Y_i for some $j \neq i$, and must be included if it is a root. So we just have to check all the roots between X_j and Y_i for some $j \neq i$. For each root R_p, we check where in B will $\sigma_{R_p}^2/\sigma_X^2$ exist. When R in B is R_p, the term containing $\sigma_{R_p}^2/\sigma_X^2$ in β_{RX} is the sum of paths from R_p to X_j where R_p is the root, so is the sum of directed paths from R_p to X_j. So the term containing $\sigma_{R_p}^2/\sigma_X^2$ in $c_R\beta_{RX}$ is the sum of paths between Y_i and X_j through R_p with 1) R_p being the root and 2) the sub-path from R_p to Y_i does not go through any variable in $\mathcal{R}[iki] \cup \mathcal{R}[ki]$ for any $k \neq i$.

The terms that are left in $Val(p)\frac{\sigma_{R_p}^2}{\sigma_X^2}$ to cover in B are the $X_j - R_p - Y_i$ paths whose sub-path from R_p to Y_i go through some variable in $\mathcal{R}[iki] \cup \mathcal{R}[ki]$ for any $k \neq i$. We just have to go over all types of R in B, and see which ones contain $\sigma_{R_p}^2/\sigma_X^2$.

Case 1: $R \in Anc(R_p)$**.** There is no such a path in c_R or β_{RX}. c_R does not go through R since $R \in \mathcal{R}[ji]c_R\beta_{RX}$. β_{RX} also does not contain $\sigma_{R_p}^2$ since $R \in Anc(R_p)$, so R_p is never a root on any paths between R and X_j. Hence $c_R\beta_{RX}$ does not contain such a path.

Case 2: $R \in Desc(R_p)$**.** Again, c_R does not contain R_p. However β_{RX} contains $\sigma_{R_p}^2$. R_p can be a root on some paths between R and X_j. Those paths are from R_p to R and R_p to X_j. Recall that c_R denotes directed paths from R to Y_i. The term that contains $\sigma_{R_p}^2$ in $c_R\beta_{RX}$ are the paths between X_j and Y_i, that pass through some variable in $\mathcal{R}[iki] \cup \mathcal{R}[ki]$ (R), with R_p being the root. As a result, this case completely cover the missing term.

Case 3: $R \perp\!\!\!\perp R_p$**.** It is easy to derive that in this case, $c_R\beta_{RX}$ does not contain a path that goes through R_p. Otherwise R and R_p would be dependent.

Case 4: R **and** R_p **are only connected through common ancestors.** In this case,

in any path that contains both R and R_p, R_p will not be the root. Their common ancestors will be the roots. So this case also does not provide any term containing $\sigma^2_{R_p}/\sigma^2_X$.

We have proved that for every R_p in A, the coefficient of $\sigma^2_{R_p}/\sigma^2_X$ (equal to a sum of those paths in $P[ji]$ with R_p being the root) is equal to the the coefficient of $\sigma^2_{R_p}/\sigma^2_X$ in B. As stated before, A and B have the same set of roots, so they have the same $\sigma^2_{R_p}/\sigma^2_X$ terms. So the sum of those terms are equal.

Next, we prove the reflecting bias terms are also equal. Observe that $\bigcup_{1\leq i\leq n} P[iji] = \mathcal{P}$, so we just have to prove that $\sum_{p\in P[iji]} Val(p) + \sum_{R\in(\mathcal{R}[iji]\setminus\{X_i\})} c_R\beta_{RX}$ is equivalent to $\sum_{p\in P[iji]} Val(p)\frac{\sigma^2_{R_p}}{\sigma^2_X}$. This can be proven using the exact same reasoning above, so we omit the proof.

Thus, the two expressions for $E[\hat{\beta_{YX}}]$ are equivalent. □

Corollary 1. *Let $M^{**}(G^{**}, S)$ be a balanced interaction model in which X satisfies ASDC and TACE is identified as $\beta_{YX} = \alpha$ in the generic network, then interaction bias exists iff G^{**} contains a reflecting or deflecting bias structure.*

Proof. (if part) Follows from theorem 1.There are two terms that cause bias in theorem 1 and they can be attributed to the two bias structures.
(only if part) Had there been additional structures that caused bias, then theorem 1 would have had additional terms to account for it. Since theorem 1 has only two bias terms fully accounted for by the two structures, there exist no other structure that creates bias. □

Theorem 2. *Let G^* be an interaction network. Given the conditions in Theorem 1 and B a bias-free subset for G^*, $TACE_{XY} = E[\hat{\beta_{YX}}]$ where the regression coefficient is calculated using only samples in set B.*

Proof. We check the interaction network G^*_S formed by B, by treating any variable from $\mathcal{V}_{(j)}$ where $j \notin S$ as unobserved. Next, we calculate $E[\hat{\beta_{YX}}]$ for G^*_S.

By Theorem 1,

$$E[\hat{\beta_{YX}}] = \alpha + \frac{1}{n} Term_2 - \frac{1}{n(n-1)} Term_3.$$

The second term is obtained by summing over paths of the form: $X_i - \cdots - W_j - \cdots - Y_i$, and the third term is obtained by summing over paths of the form: $X_i - \cdots - Y_j$. These paths do not exist in G^*_S. Hence, the two bias terms are 0, and $E[\hat{\beta_{YX}}] = \alpha$. □

8.1.4 Lemmas

Lemma 1. *If W satisfies ASDC, then any two explicit variables W_i and W_j are IID (Independent and Identically Distributed.)*

Proof. If W satisfies ASDC, and W_i is the root for some i, then from the third property of ASDC, W_i must be the root for all i. The roots are only caused by their error terms, the error terms are IID (identically distributed and independent), so W is IID.

If W_i is not the root for any i, W satisfies ASDC, and all its parents are IID, then we have for any i

$$W_i = \sum_{V_i \in Pa(W)} c_{V_i} V_i + U_{W_i},$$

where c_{V_i} is the coefficient of the variable V_i on the edge $V_i \to W_i$. Each term is IID for any $i \neq j$. So W_i and W_j are IID.

If W_i is not the root for any i, W satisfies ASDC, and there exists a parent of W, V such that V_i and V_j are not IID. Then from our previous derivation, there exists a parent of V, V', such that V'_i and V'_j are not IID. Keep tracing up until a root variable R, such that R_i and R_j are not IID. However, this violates our derivation in the beginning, that if a variable is the root and satisfies ASDC, it must be IID. We reach a contradiction. Hence, if W_i is not the root for any i, W satisfies ASDC, then all its parents are IID, and W is thus IID. □

Lemma 5. *Let $\mathcal{X} = \{X_1, \ldots, X_n\}$ be n IID random variables where the $\sigma_X^2 > 0$, and a random variable W_i. Among $\mathcal{X}$, W_i is dependent of X_i only, and $W_i = aX_i + b$ where a and b are constants. Then the following expectation exists.*

$$E\left[\frac{(X_i - \bar{X})W_i}{\sum\limits_{1\leq i\leq n}(X_i - \bar{X})^2}\right].$$

Proof. We have to prove that the function $f(X_1, \ldots, X_n, W_i)$ inside of the expectation is bounded. For convenience, rewrite it by plugging in $W_i = aX_i + b$.

$$E\left[\frac{(X_i - \bar{X})(aX_i + b)}{\sum\limits_{1\leq i\leq n}(X_i - \bar{X})^2}\right].$$

For any X_j with $j \neq i$, the denominator is a quadratic function on X_j, and the numerator is a linear function of X_j from the term $\bar{X}$. For X_i, the denominator is a quadratic function on X_i, and the numerator is a quadratic function on X_i. Since $\sigma_X^2 \neq 0$, $X_1, \ldots, X_n$ cannot take on the same value, so the denominator is always positive. When considering X_i as the variable, f might only go to infinity when X_i goes to infinity or negative infinity, and same with X_j.

When considering X_i as the variable, and X_j for all other j as constants, the denominator can be written in the form of $AX_i^2 + BX_i + C$, with A, B, C being constants. Hence, the order (of the polynomial) of the denominator is 2, and the order of the numerator is 2. So the limit of f when X_i goes to ∞ or $-\infty$ is a finite value equal to the ratio of the coefficient of X_i^2 in the numerator divided by the coefficient of X_i^2 in the denominator.

When considering X_j as the variable, the order of the denominator is 2, and the order of the numerator is 1. So the limit of f when X_i goes to ∞ or $-\infty$ is 0. Hence, f is bounded. □

Lemma 6. *Given a balanced interaction model $M^{**}(G^{**}, S^{**})$, if generic variables V and X*

both satisfy ASDC, and $dSep(V_i, X_i|\emptyset)$ *for all* i *in* G^{**}*, then*

$$E\left[\frac{\sum\limits_{1\leq i\leq n}(X_i-\bar{X})V_i}{\sum\limits_{1\leq i\leq n}(X_i-\bar{X})^2}\right]=0.$$

Proof. The d-separation condition implies $X_i \perp\!\!\!\perp V_i$. V and X are IID implies that we can treat all X_i's as the same variable X, and treat all V_i's as the same variable V. Hence, $X \perp\!\!\!\perp V$ and $\sigma_{XV}=0$, which gives $\beta_{VX} = \sigma_{XV}\sigma_X^2 = 0$. Also note that

$$\hat{\beta_{VX}}=\frac{\sum\limits_{1\leq i\leq n}(X_i-\bar{X})V_i}{\sum\limits_{1\leq i\leq n}(X_i-\bar{X})^2}.$$

Since the ordinary least squares estimator is unbiased, we have $E[\hat{\beta_{VX}}]=\beta_{VX}=0$. □

Lemma 7. *Given a balanced interaction model, with the following conditions: 1)* X_i *and* Y_i *are not confounded by a path containing only variables in* $\mathcal{V}_i$*,* $\forall i$*, and 2)* X_i *satisfies ASDC. Then there exists a set* $\mathcal{S}$ *consisting of the following three subsets of explicit variables:*

1. $\mathcal{S}_1$*:* X_i*,*
2. $\mathcal{S}_2$*: the root variables (excluding* X_i*) of each open path between* X_j *and* Y_i *(*j *can be the same as* i*),*
3. $\mathcal{S}_3$*: the root variables of this interaction network that are in* $Anc(Y_i)$ *and d-separated (by an empty set) from* X_j *for all* j*,*

such that Y_i *can be expressed as a linear function of the variables in* S *i.e.,*

$$Y_i=\sum_{W_t\in\mathcal{S}}c_{W_t}W_t,$$

where c_{W_t} *is equal to the sum of the values of the directed paths from* W_t *to* Y_i *that do not go through any variable in* $\mathcal{S}$*.*

Proof. Consider the following protocol.

- Start from the initial structural equation of Y_i, $Y_i = f(Pa(Y_i))$, denoted $SE(Y_i)$.
- For each variable A_q in the r.h.s. of $SE(Y_i)$,
  - if $A_q \in \mathcal{S}$, keep it.
  - if $A_q \notin \mathcal{S}$ and not a root of the network, replace it with its structural equation, $A_q = g(Pa(A_q))$ and plug it into $SE(Y_i)$.
  - if $A_q \notin \mathcal{S}$ and is a root of the network, keep it.
- Keep replacing until no more replacement can be done in the r.h.s. of $SE(Y_i)$.
- Denote the final $SE(Y_i)$ as $SE_f(Y_i)$.

We prove $SE_f(Y_i)$ is

$$Y_i = \sum_{W_t \in \mathcal{S}} c_{W_t} W_t,$$

where c_{W_t} is equal to the sum of the product of path coefficients of the directed paths from W_t to Y_i that do not go through any variable in $\mathcal{S}$.

First, we prove that the r.h.s. of $SE_f(Y_i)$ contains only variables in $\mathcal{S}$. If it contains a variable, $A_r \notin \mathcal{S}$, then A_r must be a root variable of the network. Otherwise it would have been replaced by its parents according to the protocol. $A_r \notin \mathcal{S}$, so $A_r \notin \mathcal{S}_3$, hence A_r must be d-connected (given an empty set) to at least one X_j for some j. Since A_r is a root of the network, A_r must be the ancestor of X_j. We next discuss if it is X_j for $j = i$ or $j \neq i$.

- $j = i$, i.e., A_r is an ancestor of X_i. Since X is ASDC, X_i cannot be caused by a variable belonging to another unit. Hence, we have $r = i$. If all directed paths from A_r to Y_i pass through variables in $\mathcal{S}$, then A_r cannot be replaced into the r.h.s. of $SE_f(Y_i)$. Hence, there exists at least one directed path from A_r to Y_i that does not pass through any variable in $\mathcal{S}$, which we denote as p_d. Since A_r is an ancestor of X_i and A_r to Y_i is a directed path not through $\mathcal{S}$(including X_i), there exists a confounding path between

X_i and Y_i through A_r. Since X_i and Y_i are not confounded by only variables of i, p_d must go through a variable of a different unit, and is the root of that confounding path. However, then $A_r \in \mathcal{S}_2$ by definition, which contradicts the assumption that $A_r \notin \mathcal{S}$.

- $j \neq i$, i.e., A_r is an ancestor of X_j for some $j \neq i$. Again, there exists at least one directed path from A_r to Y_i that does not pass through any variable in $\mathcal{S}$, which we denote as p_d. Since A_r is ancestor to both X_j and Y_i, there is a confounding path between X_j and Y_i through A_r. A_r is the root on this path, which implies $A_r \in \mathcal{S}_3$, and contradicts the assumption that $A_r \notin \mathcal{S}$.

Thus, our counterproof assumption is wrong, which means the r.h.s. of $SE_f(Y_i)$ generated by the above protocol contains only variables in $\mathcal{S}$. Next we prove that the coefficients C_{W_t} for each $W_t\mathcal{S}$ in the linear combination is equal to the sum of the values of the directed paths from W_t to Y_i that do not go through any variable in $\mathcal{S}$. In the protocol above, every time a variable is replaced by its parents, there is a multiplier equal to the directed edge between each parent and the variable. For example, in SE_{Y_i}, a term is γC_i. If C_i is replaced by its parents, D_j and E_k, where $C_i = \delta D_j + \theta E_k$, then the term in SE_{Y_i} becomes $\gamma(\delta D_j + \theta E_k)$. So the coefficient of D_j is C_i's coefficient γ multiplied by δ, the edge $D_j \to C_i$. Since replacements of a variable stops if it is in $\mathcal{S}$, we have that the final coefficient of a variable is equal to the sum of all directed paths from that variable to Y_i, which do not pass through any other variable in $\mathcal{S}$. □

Lemma 8. *Given n IID random variables $X_1, \ldots, X_n$, and n IID random variables $R_1, \ldots, R_n$. For each i, R_i is not independent of X_i only. Then we have*

$$E\left[\frac{(X_i - \bar{X})R_i}{\sum\limits_{1\leq k\leq n}(X_k - \bar{X})^2}\right] = \frac{\beta_{RX}}{n},$$

and β_{RX} is the OLS regression coefficient of R on X, treating $X_1, \ldots, X_n$ as a single variable X, and $R_1, \ldots, R_n$ as a single variable R.

Proof. The above expression only depends on i, and from the property of IID, it is the same for any i. We sum over i for that expression, and get

$$\begin{aligned}
& nE\left[\frac{(X_i-\bar{X})R_i}{\sum\limits_{1\le k\le n}(X_k-\bar{X})^2}\right] \\
=&\sum_{1\le i\le n} E\left[\frac{(X_i-\bar{X})R_i}{\sum\limits_{1\le k\le n}(X_k-\bar{X})^2}\right] \\
=&E\left[\frac{\sum\limits_{1\le i\le n}(X_i-\bar{X})R_i}{\sum\limits_{1\le k\le n}(X_k-\bar{X})^2}\right] \\
=&E\left[\hat{\beta_{RX}}\right] \\
=&\beta_{RX}.
\end{aligned}$$

Divided by n on both sides, we have the equation in the lemma. □

Lemma 9. *Given n IID random variables $X_1,\ldots,X_n$, and n IID random variables $R_1,\ldots,R_n$. For each i, R_i is not independent of X_i only. Then we have*

$$E\left[\frac{(X_i-\bar{X})R_j}{\sum\limits_{1\le k\le n}(X_k-\bar{X})^2}\right]=-\frac{\beta_{RX}}{n(n-1)},$$

for $i\neq j$, and β_{RX} is the OLS regression coefficient of R on X, treating $X_1,\ldots,X_n$ as a single variable X, and $R_1,\ldots,R_n$ as a single variable R.

Proof. Denote the expectation of interest as E_{ij}. X and R are both IID regarding different units, and X_i and R_j are independent for $i\neq j$. Thus, $E_{ij}=E_{i'j}$, for any $i'\neq j$. Below

when the sum is over $i \neq j$, it means summing over $i \in \{1, \ldots, n\} \setminus \{j\}$. We have

$$\begin{aligned}
(n-1)E_{ij} =& \sum_{i \neq j} E_{ij} \\
=& E\left[\frac{\sum_{i \neq j}(X_i - \bar{X})R_j}{\sum_{1 \leq k \leq n}(X_k - \bar{X})^2}\right] \\
=& E\left[\frac{\sum_{1 \leq i \leq n}(X_i - \bar{X})R_j - (X_j - \bar{X})R_j}{\sum_{1 \leq k \leq n}(X_k - \bar{X})^2}\right] \\
=& E\left[\frac{(\sum_{1 \leq i \leq n}(X_i - \bar{X}) - (X_j - \bar{X}))R_j}{\sum_{1 \leq k \leq n}(X_k - \bar{X})^2}\right] \\
=& E\left[\frac{(0 - (X_j - \bar{X}))R_j}{\sum_{1 \leq k \leq n}(X_k - \bar{X})^2}\right] \\
=& -E\left[\frac{(X_j - \bar{X})R_j}{\sum_{1 \leq k \leq n}(X_k - \bar{X})^2}\right].
\end{aligned}$$

By Lemma 8, we have

$$(n-1)E_{ij} = -\frac{\beta_{RX}}{n}.$$

Divided by $(n-1)$ on both sides, we get the equation we wanted to prove. □

Lemma 10. *Given n IID random variables $X_1, \ldots, X_n$, and a variable L_t independent of $X_1, \ldots, X_n$. Then we have*

$$E\left[\frac{(X_i - \bar{X})L_t}{\sum_{1 \leq k \leq n}(X_k - \bar{X})^2}\right] = 0.$$

Proof. Denote the expectation of interest as E_i, then $E_i = E_j$ for any i, j, since X_i and X_j

are IID. So we have

$$\begin{aligned} nE_i &= \sum_{1\leq i\leq n} E_i \\ &= E\left[\frac{\sum\limits_{1\leq i\leq n}(X_i-\bar{X})L_t}{\sum\limits_{1\leq k\leq n}(X_k-\bar{X})^2}\right] \\ &= 0. \end{aligned}$$

□

To prove Theorem 1, we first prove a slightly different version of it, Lemma 11.

Lemma 11. *Given the interaction network G^* of a balanced linear interaction model, with X_i and Y_i not confounded by any variable in $\mathcal{V}_i$, $\forall i$. Given that X satisfies ASDC, then the expected value of the OLS estimator $\hat{\beta_{YX}}$ is given by*

$$\begin{aligned} & E[\hat{\beta_{YX}}] \\ =&\alpha \\ &+\frac{1}{n}(\sum_{p\in\mathcal{P}} Val(p) + \sum_{1\leq i\leq n}\sum_{R\in(\mathcal{R}[iji]\setminus\{X_i\})} c_R\beta_{RX}) \\ &-\frac{1}{n(n-1)}\sum_{1\leq i\leq n}\sum_{R\in\mathcal{R}[ji]} c_R\beta_{RX}, \end{aligned}$$

where $\mathcal{P}$ is the set of directed paths from X_i to Y_i for all i through any $W_j \in \mathcal{V}_{(j)}$ with $i \neq j$, $\mathcal{R}[iji]$ is the set of roots of the open paths between X_i and Y_i through some W_j with $j \neq i$, $\mathcal{R}[ji]$ is the set of roots of the open paths between X_j and Y_i for $j \neq i$, and c_R is the sum of values of the directed paths from a variable R ($\in (\mathcal{R}[iji]\setminus\{X_i\})$ or $\in \mathcal{R}[ji]$) to Y_i not passing through any variable in $\mathcal{R}[iji]\cup\mathcal{R}[ji]$ for any $j \neq i$.

Proof.

$$
\begin{aligned}
E[\hat{\beta_{YX}}] =& E\left[\frac{\sum_{1\le i\le n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sum_{1\le i\le n}(X_i-\bar{X})^2}\right] \\
=& E\left[\frac{\sum_{1\le i\le n}(X_i-\bar{X})Y_i}{\sum_{1\le i\le n}(X_i-\bar{X})^2}\right] - E\left[\frac{\sum_{1\le i\le n}(X_i-\bar{X})\bar{Y}}{\sum_{1\le i\le n}(X_i-\bar{X})^2}\right] \\
=& E\left[\frac{\sum_{1\le i\le n}(X_i-\bar{X})Y_i}{\sum_{1\le i\le n}(X_i-\bar{X})^2}\right] - E\left[\frac{(\sum_{1\le i\le n}X_i-n\bar{X})\bar{Y}}{\sum_{1\le i\le n}(X_i-\bar{X})^2}\right] \\
=& E\left[\frac{\sum_{1\le i\le n}(X_i-\bar{X})Y_i}{\sum_{1\le i\le n}(X_i-\bar{X})^2}\right] - E\left[\frac{(n\bar{X}-n\bar{X})\bar{Y}}{\sum_{1\le i\le n}(X_i-\bar{X})^2}\right] \\
=& E\left[\frac{\sum_{1\le i\le n}(X_i-\bar{X})Y_i}{\sum_{1\le i\le n}(X_i-\bar{X})^2}\right]
\end{aligned}
$$

Y_i is can be written as a linear combination of the set in Lemma 7, $\mathcal{S}$. By Lemma 7, $\mathcal{S}$ is composed of

1. X_i,
2. the root variables (excluding X_i) of each open path between X_j and Y_i, and
3. the root variables of this interaction network that are in $Anc(Y_i)$ and d-separated (by an empty set) from X_j for all j, denoted by $\mathcal{L}_i$.

The second component can be further divided into two sub-components as follows.

1. $\mathcal{R}[iji] \setminus \{X_i\}$, the set of roots of the open paths between X_i and Y_i through some W_j with $j \neq i$, with X_i excluded, and
2. $\mathcal{R}[ji]$, the set of roots of the open paths between X_j and Y_i for $i \neq j$.

We have

$$Y_i = c_i X_i + \sum_{R \in (\mathcal{R}[iji] \setminus \{X_i\})} c_R R + \sum_{R \in \mathcal{R}[ji]} c_R R + \sum_{L \in \mathcal{L}_i} c_L L,$$

where c_i, c_R, and c_L denote coefficients for the linear combination. The variables in the above expression are $\mathcal{S}$, i.e., $\mathcal{S} = \mathcal{R}[iji] \cup \mathcal{R}[ji] \cup \mathcal{L}_i$. Next, we compute the coefficients c_i, c_R, c_L.

c_i is the sum of the directed path values from X_i to Y_i not passing through any variable in $\mathcal{S}$. There are three types of directed paths from X_i to Y_i:

1. the directed edge $X_i \to Y_i$,
2. directed paths $X_i \to \cdots \to V_i \to \cdots \to Y_i$, and
3. directed paths $X_i \to \cdots \to V_j \to \cdots \to Y_i$ for $j \neq i$.

The first two types belong to TACE by definition. So $c_i = \alpha + c_{i3}$, where c_{i3} is the coefficient contributed by the third type of directed paths. Note that V_j cannot be a root of another path between X_k and Y_l for some $k \neq l$. This is because V_j is caused by X_i, so V cannot be ASDC, so X cannot be ASDC since X_k is caused by V_j, which violates the assumption that X is ASDC. Hence, c_{i3} is equal to the sum of all directed paths from X_i to Y_i through some variable V_j for any j, which is equal to $\sum_{p \in \mathcal{P}}$ in the lemma statement.

For the second and third components in Y_i, each c_R is the sum of the directed paths (multiplications of edge coefficients) from R to Y_i not through variables in $\mathcal{S}$. This follows from Lemma 7.

We have

$$
\begin{aligned}
&E[\hat{\beta_{YX}}] \\
=&E\left[\frac{\sum\limits_{1\le i\le n}(X_i-\bar{X})Y_i}{\sum\limits_{1\le i\le n}(X_i-\bar{X})^2}\right] \\
=&E\left[\frac{\sum\limits_{1\le i\le n}(X_i-\bar{X})(c_iX_i+\sum_{R\in(\mathcal{R}[iji]\setminus\{X_i\})}c_RR+\sum_{R\in\mathcal{R}[ji]}c_RR+\sum_{L\in\mathcal{L}_i}c_LL)}{\sum\limits_{1\le i\le n}(X_i-\bar{X})^2}\right] \\
=&\alpha E\left[\frac{\sum\limits_{1\le i\le n}(X_i-\bar{X})X_i}{\sum\limits_{1\le i\le n}(X_i-\bar{X})^2}\right]+E\left[\frac{\sum\limits_{1\le i\le n}(X_i-\bar{X})c_{i3}X_i}{\sum\limits_{1\le i\le n}(X_i-\bar{X})^2}\right] \\
&+\sum_{1\le i\le n}\sum_{R\in(\mathcal{R}[iji]\setminus\{X_i\})}c_RE\left[\frac{(X_i-\bar{X})R}{\sum\limits_{1\le i\le n}(X_i-\bar{X})^2}\right] \\
&+\sum_{1\le i\le n}\sum_{R\in\mathcal{R}[ji]}c_RE\left[\frac{(X_i-\bar{X})R}{\sum\limits_{1\le i\le n}(X_i-\bar{X})^2}\right]+\sum_{1\le i\le n}\sum_{L\in\mathcal{L}_i}c_LE\left[\frac{(X_i-\bar{X})L}{\sum\limits_{1\le i\le n}(X_i-\bar{X})^2}\right].
\end{aligned}
$$

For the first term: similar to the way $\bar{Y}$ is removed before, in the first term, we can change X_i to $X_i-\bar{X}$. The numerator and the denominator are the same in the expectation. So the first term is α.

The second term is equal to

$$\sum_{1\le i\le n}c_{i3}E\left[\frac{(X_i-\bar{X})X_i}{\sum\limits_{1\le i\le n}(X_i-\bar{X})^2}\right].$$

By Lemma 8, it becomes

$$\sum_{1\le i\le n}c_{i3}\frac{\beta_{XX}}{n},$$

where c_{i3} is the sum of directed paths from X_i to Y_i through V_j for any $j\neq i$ and any V.

For the third term: we look at one single R first. R is the root variable of an open path between X_i and Y_i through some W_j with $j\neq i$, so R causes X_i. Then R must belong to

unit i since X satisfies ASDC. Since R is the root, $R \in Anc(X)$, so R satisfies ASDC, and is IID for different units. So we relabel this R as R_i, and we have IID $R_1, \ldots, R_n$. Applying Lemma 8, we have the expectation term is equal to β_{RX}/n. c_R is the sum of the directed paths from R_i to Y_i, not through variables in $\mathcal{S}$. So the third term is equal to

$$\frac{1}{n} \sum_{1 \leq i \leq n} \sum_{R \in (\mathcal{R}[iji] \setminus \{X_i\})} c_R \beta_{RX}.$$

For the fourth term: we look at one single R first. R is the root variable of an open path between X_j and Y_i, for some $j \neq i$, so either R causes X_j or $R = X_j$. If R causes X_j, then R must belong to unit j, because X satisfies ASDC. So either case R belongs to unit j. Since R is the root, $R \in Anc(X)$, so R satisfies ASDC, and is IID for different units. So we relabel this R as R_j, and we have IID $R_1, \ldots, R_n$. Applying Lemma 9, we have the expectation term is equal to $-\beta_{RX}/(n(n-1))$. c_R is the sum of the directed paths from R_j to Y_i, not through variables in $\mathcal{S}$. So the fourth term is equal to

$$-\frac{1}{n(n-1)} \sum_{1 \leq i \leq n} \sum_{R \in \mathcal{R}[ji]} c_R \beta_{RX}.$$

The fifth term is 0 by Lemma 6.

Finally, recall that $Val(p)$ denotes the value of an open path p. Plugging the above values back into the expression for $E[\hat{\beta_{YX}}]$, we have the results as in Lemma 11. □

8.2 Supplemental Materials for Chapter 4

8.2.1 Proof

Theorem 3. *Suppose $M^*, D, G^\dagger$ refer to the true model, available data and generic network as specified in definition 6 such that $Q = TACE_{XY}$ and $\hat{Q} = \beta_{YX}$. X_i and Y_i are not confounded by any variable of i, for all i. Let G^U be the b-G^U corresponding to M^*. For all $i \neq j$ pairs, let N_d be the number of pairs of units that have definite interference paths from*

i to j and let N_θ be the number of pairs of units that have uncertain interference paths from i to j with probability θ. Let the sum of the values of the interference paths from X_i to Y_j be p,[1] for all $i \neq j$. The expected interaction bias is given by

$$E[\Big|E[\hat{\beta_{YX}}] - Q\Big|] = \frac{1}{n(n-1)}|p|(N_d + \theta N_\theta).$$

Proof. Let G^* be the true interaction graph corresponding to M^* (no uncertainty). By Theorem 1 in [ZMP22],

$$\Big|E[\hat{\beta_{YX}}] - Q\Big| = \Big|\tfrac{1}{n}\sum_{1\leq i\leq n}\sum_{p\in P[iji]} Val(p)\tfrac{\sigma^2_{R_p}}{\sigma^2_X} - \tfrac{1}{n(n-1)}\sum_{1\leq i\leq n}\sum_{p\in P[ji]} Val(p)\tfrac{\sigma^2_{R_p}}{\sigma^2_X}\Big|.$$

Under our settings, there is no reflecting bias structure, but only deflecting bias structures. So the first term on the r.h.s. is 0. Theorem 1 becomes

$$\Big|E[\hat{\beta_{YX}}] - Q\Big| = \Big| - \tfrac{1}{n(n-1)}\sum_{1\leq i\leq n}\sum_{p\in P[ji]} Val(p)\tfrac{\sigma^2_{R_p}}{\sigma^2_X}\Big|.$$

The sole term on the r.h.s. is the sum of all deflecting bias paths' strengths multiplied by the variance factors, divided by $1/(n(n_1))$. Under our settings, interference paths are the only deflecting bias structures, and the roots of those paths are all X_i for some i. So we have $\sigma^2_{R_p}/\sigma^2_X = 1$. In addition, the summation is over all interference paths. It can be rearranged as summing over all pairs of $i \neq j$, and for each pair, sum over all the interaction paths. For each pair, the summation of all the interaction paths is the same and equal to p, from our assumptions. Hence, Theorem 1 can be further simplified as

$$\Big|E[\hat{\beta_{YX}}] - Q\Big| = \Big| - \tfrac{1}{n(n-1)}Np\Big|,$$

where N is the total number of ordered pairs of $i \neq j$ where there are interference paths from X_i to Y_j. The term on the r.h.s. inside of the absolute symbols except p is less than 0. So Theorem 1 becomes

$$\Big|E[\hat{\beta_{YX}}] - Q\Big| = \tfrac{1}{n(n-1)}N|p|.$$

[1] *I.e., p is equal to the causal effect of X_i on Y_j.*

Hence, we have

$$E[\left|E[\hat{\beta_{YX}}] - Q\right|] = E[\frac{1}{n(n-1)}N|p|]$$
$$= \frac{1}{n(n-1)}|p|E[N]. \tag{8.1}$$

Next we evaluate $E[N]$. N is the sum of the pairs whose interference paths appear in G^U as definite paths, and the pairs whose interference paths appear in G^U as uncertain paths. The first part is simply N_d. The second part's expectation is θN_θ by the definition of θ. So we have

$$E[N] = N_d + \theta N_\theta.$$

Plugging this into Equation 8.1, we have the equation in the statement of Theorem 3. □

Theorem 4. *Consider the setting in Theorem 3. Suppose we are additionally given a bias threshold τ, and the interference effect is bounded by a constant Γ times the TACE (i.e., $|p| \leq \Gamma|Q|$). If a subset $\mathcal{B}$ of units satisfies*

$$\frac{1}{|\mathcal{B}|(|\mathcal{B}|-1)}(N'_d + \theta N'_\theta)\Gamma \leq \tau,$$

then using the samples in $\mathcal{B}$, the expected interaction bias will be at most $\tau|Q|$. For all $i \neq j$ pairs with $i, j \in \mathcal{B}$, N'_d denotes the number of pairs with definite interference paths from i to j in G^*, and N'_θ denotes the number of pairs with interference paths from i to j in G^* with probability θ.

Proof. Consider the sub-graph G_{sub} formed by projecting the true interaction graph G^* on $\mathcal{B}$. variables of units in $\mathcal{B}$. Consider an interference path from X_i to Y_j, where $i \neq j$ and $i, j \in \mathcal{B}$. It does not go through a third unit by our assumptions, and since $i, j \in \mathcal{B}$, the path remains unchanged in G_{sub}. Let G_{usub} be the uncertain sub-graph formed by projecting G^U on $\mathcal{B}$. For each pair $i \neq j$ in the original uncertain sub-graph G^U such that the interference paths from X_i to Y_j are uncertain, consider the following scenarios.

1. If $i, j \in \mathcal{B}$, then from the previous discussion, the interference paths from X_i to Y_j remains unchanged after the projection. So the probability of those paths existing is still θ.

2. If $i \in \mathcal{B}, j \notin \mathcal{B}$, then the interference paths are removed in G_{usub}.

3. If $i \notin \mathcal{B}, j \in \mathcal{B}$, if we also have interference paths from X_i to Y_k with $k \neq j$, this will result in a bidirected path between Y_j and Y_k. However, this can be ignored since it is not a bias structure, by Definitions 7 and 8 in [ZMP22].

As a result, we can apply Theorem 3 on G_{usub}, and obtain the interaction bias as

$$\frac{1}{|\mathcal{B}|(|\mathcal{B}|-1)}(N'_d + \theta N'_\theta)|p|.$$

Plugging in $|p| \leq \Gamma|Q|$, we have the interaction bias is at most $\tau|Q|$ by Theorem 4. □

Theorem 5. *Consider the setting described in Theorem 3. Suppose we know the relationship between p (the interference path strength) and Q (TACE) is $p = \gamma Q$, where γ is a constant, then Q is unbiasedly estimated as*

$$Q = \frac{E[\hat{\beta_{YX}}]}{1 - \frac{1}{n(n-1)}\gamma(N_d + \theta N_\theta)}.$$

Proof. From the proof of Theorem 1 in [ZMP22], we have a slightly stronger result than Theorem 3, which is Theorem 3 without the absolute signs. We have

$$E[\hat{\beta_{YX}}] - Q = -\frac{1}{n(n-1)}p(N_d + \theta N_\theta).$$

Plugging in $p = \gamma Q$, we have the expression in the theorem statement. □

Corollary 2. *Consider the setting described in Theorem 5, if we further assume $0 < \gamma < 1$ and $|Q| > |p|$ and $0 < \gamma < 1$ then Q can be bounded as*

$$\frac{E[\hat{\beta_{YX}}]}{1 - \frac{(N_d + \theta N_\theta)}{n(n-1)}} < Q < E[\hat{\beta_{YX}}].$$

Proof. Q is monotonic with respect to γ. Hence, Corollary 2 results from Theorem 5 by plugging in $\gamma = 0$ and $\gamma = 1$. □

8.3 Supplemental Materials for Chapter 5

8.3.1 Derivation of Examples for Definition 9

Given an interaction model with network Figure 3.3 and structural equations (3.4)-(3.6), the interaction bias is calculated as follows.

$$
\begin{aligned}
&|E[\hat{Q}_{Y|do(X)}] - TACE_{XY}| \\
=&\left| E_{m_1,m_0}\left[\frac{1}{n}\sum_{1\le j\le n} E\left[Y_j|\, m_1, m_0, X_j = 1\right] - \frac{1}{n}\sum_{1\le j\le n} E\left[Y_j|m_1, m_0, X_j = 0\right]\right] - (E[Y^D|X^D = 1] - E[Y^D|X^D \right. \\
=&\left| E_{m_1,m_0}\left[\frac{1}{3}(2 + E[2 - X_1 + X_3|X_2 = 1, m_1, m_0] + 2) - \frac{1}{3}(0 + E\left[0 - X_1 + X_3|X_2 = 0, m_1, m_0\right] + 0]\right] - 2\right| \\
=&\left| E_{m_1,m_0}\left[2 + \frac{1}{3}E\left[-X_1 + X_3|X_2 = 1, m_1, m_0\right] - \frac{1}{3}E\left[-X_1 + X_3|X_2 = 0, m_1, m_0\right]\right] - 2\right| \\
=&\frac{1}{3}\left| E_{m_1,m_0}\left[(-E\left[X_1|X_2 = 1, m_1, m_0\right] + E\left[X_3|X_2 = 1, m_1, m_0\right]) + (E\left[X_1|X_2 = 0, m_1, m_0\right] - E\left[X_3|X_2 = 0, \right.\right.\right. \\
=&\frac{1}{3}\left| E_{m_1,m_0}\left[0 + 0\right]\right|
\end{aligned}
$$

Both terms are 0 because X_1 and X_3 are identically distributed given X_2, m_1, m_0. As a result the bias is 0. If the structural equation of Y_2 is changed to $Y_2 = 2X_2 - 2X_1 - X_3$, then the interaction bias becomes

$$
\begin{aligned}
&|E[\hat{Q}_{Y|do(X)}] - TACE_{XY}| \\
=&\left| E_{m_1,m_0}\left[\frac{1}{3}(2 + E[2 - 2X_1 + X_3|X_2 = 1, m_1, m_0] + 2) - \frac{1}{3}(0 + E\left[0 - 2X_1 + X_3|X_2 = 0, m_1, m_0\right] + 0]\right] - 2\right| \\
=&\left| E_{m_1,m_0}\left[2 + \frac{1}{3}E\left[-2X_1 + X_3|X_2 = 1, m_1, m_0\right] - \frac{1}{3}E\left[-2X_1 + X_3|X_2 = 0, m_1, m_0\right]\right] - 2\right| \\
=&\frac{1}{3}\left| E_{m_1,m_0}\left[(-E\left[2X_1|X_2 = 1, m_1, m_0\right] + E\left[X_3|X_2 = 1, m_1, m_0\right]) + (E\left[2X_1|X_2 = 0, m_1, m_0\right] - E\left[X_3|X_2 = \right.\right.\right. \\
=&\frac{1}{3}\left| E_{m_1,m_0}\left[- E\left[X_1|X_2 = 1, m_1, m_0\right] + E\left[X_1|X_2 = 0, m_1, m_0\right]\right]\right|
\end{aligned}
$$

This is non-zero because X_1 has different distributions given X_2, m_1, m_0. Hence, whether the bias is 0 depends on the parameter choice, and the model is not unbiased almost everywhere.

For the other example with network Figure 3.4, the interaction bias is computed as follows.

$$
\begin{aligned}
&|E[\hat{Q}_{Y|do(X)}] - TACE_{XY}| \\
=&\left| E_{m_1,m_0}\left[\frac{1}{2}(E[Y_1|X_1=1,m_1,m_0]+E[Y_2|X_2=1,m_1,m_0]) - \frac{1}{2}(E[Y_1|X_1=0,m_1,m_0]+E[Y_2|X_2=0,m_1,\right.\right. \\
=&\left| E_{m_1,m_0}\left[\frac{1}{2}(E[Y_1|X_1=1]+E[Y_2|X_2=1,m_1,m_0]) - \frac{1}{2}(E[Y_1|X_1=0]+E[Y_2|X_2=0,m_1,m_0])\right] - \alpha\right| \\
=&\left| E_{m_1,m_0}\left[\frac{1}{2}\alpha + \frac{1}{2}(E[Y_2|X_2=1]-E[Y_2|X_2=0])\right] - \alpha\right| \\
=&\left| E_{m_1,m_0}\left[\frac{1}{2}\alpha + \frac{1}{2}\alpha\right] - \alpha\right|
\end{aligned}
$$

This is 0 regardless of parameter choice.

8.3.2 Proof

Proof of Theorem 6.

Proof.

$$
\begin{aligned}
&E[\beta(X,Y)] \\
=&E_{m_1,m_0}\left[E\left[\frac{\sum\limits_{1\le i\le n} X_iY_i}{m_1} - \frac{\sum\limits_{1\le i\le n}(1-X_i)Y_i}{m_0}\middle| m_1,m_0\right]\right] \\
=&E_{m_1,m_0}\left[\sum_{1\le i\le n} E\left[\frac{X_iY_i}{m_1} - \frac{(1-X_i)Y_i}{m_0}\middle| m_1,m_0\right]\right] \\
=&E_{m_1,m_0}\left[\sum_{1\le i\le n} \frac{m_1}{n}E\left[\frac{Y_i}{m_1}\middle| m_1,m_0,X_i=1\right] - \frac{m_0}{n}E\left[\frac{Y_i}{m_0}\middle| m_1,m_0,X_i=0\right]\right] \\
=&E_{m_1,m_0}\left[\sum_{1\le i\le n} \frac{1}{n}\left(E\left[Y_i\middle| m_1,m_0,X_i=1\right] - E\left[Y_i\middle| m_1,m_0,X_i=0\right]\right)\right].
\end{aligned}
$$

TACE is defined as the ACE in the default model, which is

$$E[Y^D|X^D=1] - E[Y^D|X^D=0].$$

The interaction bias is the absolute value of the difference between $E[\beta(X,Y)]$ and TACE, which is the expression in the theorem. □

Proof of Theorem 7.

Proof. TACE is defined as the ACE in the default model, which is

$$E[Y_i^D|X_i = 1] - E[Y_i^D|X_i = 0].$$

Since we are considering the general non-parametric smooth functions, the function forms of the structural equations can be arbitrary. We just need to prove

1. if Y is not ASDC, then there exists a function form for the structural equation such that there is interaction bias, and
2. if Y is ASDC, then there is no interaction bias.

Part 1: If Y is not ASDC, then there exists a unit i and a non-ASDC parent for Y_i. Denote the non-ASDC parent as V_j (j can be equal to i). Let the structural equation of Y_i be $Y_i = X_i V_j$. Then, the difference between Y_i given $X_i = 1, m_1, m_0$ and Y_i given $X_i = 0, m_1, m_0$ are calculated as follows.

$$\begin{aligned} & E\left[Y_i \middle| m_1, m_0, X_i = 1\right] - E\left[Y_i \middle| m_1, m_0, X_i = 0\right] \\ =& E[V_j]. \end{aligned}$$

Another unit l might have different results, since l might not be affected by V_j but another variable. The interaction bias is the average difference above for all units, so the interaction bias is not 0.

Part 2: If Y is ASDC, then $\forall i$, Y is affected by X_i but not $X_j \neq i$, and there is no ancestor of Y_i that belongs to another variable. We do not have to consider descendants of Y_i. This is because in this case, if a variable is a descendant of Y_i, since Y is ASDC, then it must be a descendant of all Y_j's too. So we can safely ignore the descendants of Y_i for any i.

If we try converting the unit default model, we will get the same model, since no replacement to default value needs to be made. Hence, $E[\beta(X, Y)]$ is the same as TACE, and there is no interaction bias. □

Proof of Theorem 8.

Proof. We first write the structural equation of Y_i in a different way using the construction process below.

The structural equation of Y_i might contain non-ASDC variables or variables of another unit. In the structural equation of Y_i, for each variable that violates ASDC but is still unit i's variable, replace it with its structural equation. Recursively replace variables until the new structural equation of Y_i does not contain a non-ASDC variable of i. Denote the variables from another unit that are in the equation as V_{ik} where k is the index for those variables. The term with V_{ik} is a function of V_{ik}, and is denoted as $f_k(V_{ik})$, where f_k is the function. The new structural equation of Y_i can be written as

$$Y_i = Y_i^{ASDC} + \sum_k f_k(V_{ik}),$$

where Y_i^{ASDC} is the ASDC component (a function of the ASDC variables). The non-ASDC term is a summation due to restricted additivity. Since the component is ASDC, and the model is balanced, the component is identical for Y_j for all j. Note that the definition of the unit default model is the model of a unit where all variables that are from another unit are replaced with their default values. So Y_i^D is Y_i with other units' variables replaced with their default values. So we have

$$Y_i^D = Y_i^{ASDC} + \sum_k f_k(v_{ik}),$$

where v_{ik} is the default value of V_{ik}. The second term becomes a constant. Hence, the TACE is given by

$$E[Y_i^{ASDC}|X_i = 1, m_1, m_0] - E[Y_i^{ASDC}|X_i = 0, m_1, m_0].$$

This gives the same result for all i since the model is balanced. Denote the term

$$\sum_k f_k(v_{ik})$$

as Y_i^{NASDC}, then the interaction bias is given by

$$\left| E_{m_1,m_0}\left[\frac{1}{n}\sum_{1\leq j\leq n} E\left[Y_i^{NASDC}\middle|\, m_1, m_0, X_i = 1\right] - \frac{1}{n}\sum_{1\leq j\leq n} E\left[Y_i^{NASDC}|m_1, m_0, X_i = 0\right]\right]\right|.$$

Under the assumption of "unbiased almost everywhere," the cancellation of bias between different units is considered "accidental," since changing the numbers in the structural equation (without changing the interaction graph) will change the values of the estimate. Hence, the interaction bias is 0 iff $\forall i$,

$$E\left[Y_i^{NASDC}\middle|m_1, m_0, X_i = 1\right] = E\left[Y_i^{NASDC}\middle|m_1, m_0, X_i = 0\right]. \tag{8.2}$$

Given $X_i = 1$, the other X_j's for $j \neq i$ need to satisfy the number of $X_j = 1$ is $m_1 - 1$ (minus one because $X_i = 1$), and the number of $X_j = 0$ is m_0. Hence, Equation 8.2 holds iff

$$\forall j, Y_i^{NASDC} \perp\!\!\!\perp X_j.$$

Additionally, $V_{ik} \not\perp\!\!\!\perp X_i$ iff there is an open path between V_{ik} and X_i. Since $V_{ik} \in Anc(Y_i)$, there is an open path between X_i and Y_i through V_{ik}, which is a reflecting bias structure. Hence, $V_{ik} \not\perp\!\!\!\perp X_i$ iff there is a reflecting bias structure.

Similarly, $V_{ik} \not\perp\!\!\!\perp X_j$ for $j \neq i$ iff there is an open path between V_{ik} and X_j. Since $V_{ik} \in Anc(Y_i)$, there is an open path between X_j and Y_i, which is a deflecting bias structure. Hence, $V_{ik} \not\perp\!\!\!\perp X_i$ iff there is a deflecting bias structure.

Combining both cases $V_{ik} \not\perp\!\!\!\perp X_i$ and $V_{ik} \not\perp\!\!\!\perp X_j$ for $j \neq i$, we have that V_{ik} is dependent of some V_j for any j iff there is a reflecting or deflecting bias structure. There exists such V_{ik} iff Equation 8.2 does not hold iff interaction bias is not 0. Thus, there is interaction bias iff there are deflecting or reflecting bias structures between X and Y.

□

Proof of Theorem 9.

Proof. We check the interaction network G_B^* formed by $\mathcal{B}$, by treating any variable from unit $j \notin \mathcal{B}$ as unobserved. G_B^* is a new interaction network from a new interaction model M_B^*, where all the properties remain. Applying Theorems 7 or Theorems 8, we have the corresponding conclusions in this theorem. □

8.3.3 Experiment Details

8.3.4 Section 5.6.1.1 Experiment Setup

Interaction model is balanced on the default value of $X = 0$ and $M = 0$. Variable settings:

- X is random binary with $P(X_i = 1) = 0.5$ for all i
- A_i $N(2, 1)$
- $M_i = A_i + N(0, 1)$
- Basic $Y_i = (X_i + 1)^2 A_i + N(0, 1)$
  - For the model with deflecting bias, $Y_i += 8X_j$ for an interaction between X_j and Y_i
  - For the model with reflecting bias, $M_j += X_i$ and $Y_i += M_j$ for an interaction between X_i and Y_i through M_j
  - For the model with non-bias interactions, $Y_i += 2A_j$ for an interaction between Y_i and Y_j through A_j.

Sample size: 50. Number of interactions for each model: 100. We chose the specific strengths of the interactions because different Y_i receiving different types of interactions would still have similar variance due to the interaction effect.

8.3.5 Section 5.6.1.2 Experiment Setup

8.3.5.1 Deflecting Bias (Restricted Additivity)

Interaction model is balanced on the default value of $X = 0$. Variable settings:

- X is random binary with $P(X_i = 1) = 0.4$ for all i
- Basic $Y_i = (X + 1)^2 + N(0, 1)$
- For the Y_i affected by deflecting bias, $Y_i += \alpha X_j$ for an interaction between X_j and Y_i
- $\alpha = 1, -3, 5$ for three experiments

Sample size ranges from 10 to 100. Number of interactions: 50.

8.3.5.2 Reflecting Bias (Restricted Additivity)

Interaction model is balanced on the default value of $X = 0$ and $M = 0$. Variable settings:

- X is random binary with $P(X_i = 1) = 0.4$ for all i
- C_i $N(2, 1)$
- $M_i = C_i + N(0, 1)$
- Basic $Y_i = (X_i + 1)^2 A_i + N(0, 1)$
- For the Y_i affected by deflecting bias, $M_j += X_i$ and $Y_i += \alpha M_j$ for an interaction between X_i and Y_i through M_j
- $\alpha = 1, -3, 5$ for three experiments

Sample size ranges from 10 to 300. Number of interactions: 50.

8.3.5.3 Interactions (Non-Parametric General)

Interaction model is balanced on the default value of $C = 10, 10/3$, or 2, depending on which experiment (with different interaction strengths) it is. Variable settings:

- X is random binary with $P(X_i = 1) = 0.4$ for all i
- C_i $N(2, 1)$
- Basic $Y_i = (X_i + 1)^2 - C + N(0, 1)$
- For the Y_i affected by deflecting bias, the first term is multiplied by αC_j.
- $\alpha = 0.1, -0.3, 0.5$ for three experiments

Sample size ranges from 10 to 1000. Number of interactions: 50.

8.3.6 Sections 5.6.2.1 and 5.6.2.2 Experiment Setup

Interaction model is balanced on the default value of $X = 0$. Variable settings:

- X is random binary with $P(X_i = 1) = 0.5$ for all i
- C_i $N(2, 1)$
- W_i $N(1, 1)$
- Basic $Y_i = (X_i + 1)^2 C_i W_i + N(0, 1)$
- For the Y_i affected by deflecting bias, $Y_i += 8X_j$
- For the Y_i affected by reflecting bias, $M_j += X_j$ and $Y_i += M_i$

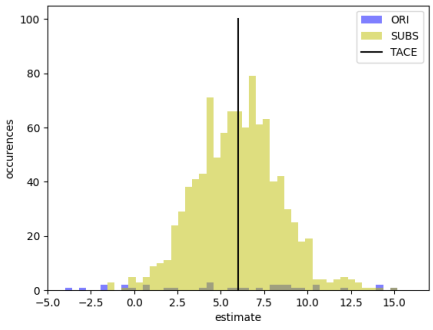
Sample size: 100. Number of deflecting bias interactions: 300. Number of reflecting bias interactions: 100.

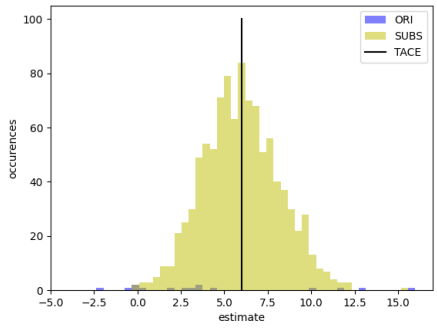
For the setting with different unit interacting chance, 1/3 units are "non-interacting", 1/3 units are "possibly-interacting", and 1/3 units are "likely-interacting". Randomly sample interaction pairs until enough number of interactions are added to the model. For each sampled interaction pair,

- If it contains at least one non-interacting unit, it is not added to the model.
- If it contains two possibly-interacting units, it has 0.25 chance of being added to the model.
- If it contains one possibly-interacting unit and one likely-interacting unit, it has 0.5 chance of being added to the model.
- If it contains two likely-interacting units, it will be added to the model.

The non-parametric experiment uses similar setups except that the interactions are multiplied instead of added.

8.3.7 Histograms for Section 5.6.2.2





8.3.8 Debias Algorithm

8.3.9 Restricted Additivity

The linear debias algorithm in [ZMP22] can be generalized and used for the restricted additivity case.

Algorithm 3 Select a bias-free subset B from an interaction network G^* and return the largest subset from t iterations

Input: an interaction network G^*, iterations t

Output: the largest bias-free subset B selected from t iterations

function FindSub(G^*, t)
  $\mathbf{B} = \emptyset$
  for $i = 1, \ldots, t$ **do**
    $Units$ = randomly sorted list $1, \ldots, n$
    $B = \{Units[1]\}$ (The indices for $Units$ start from 1)
    for $i = 2, \ldots, n$ **do**
      if $Units[i]$ has no reflecting bias structure in G^* **then**
        if $Units[i]$ has no deflecting bias structure in G^* with an element in B **then**
          $B = B \cup \{Unit[i]\}$
    $\mathbf{B} = \mathbf{B} \cup \{B\}$
  return Largest B in $\mathbf{B}$

8.3.10 General Non-Parametric

8.4 Supplemental Materials for Chapter 6

We first define the following lemma, which we will be using in the later proofs.

Algorithm 4 Select a bias-free subset B from an interaction network G^*

Input: an interaction network G^*

Output: a bias-free subset B

```
function FindSub(G*, t)
    B = ∅
    for i = 1, ..., n do
        if Y_i is ASDC then
            B = B ∪ {i}
    return B
```

Lemma 3. *Each constraint l_i in Lemma 1 can be rewritten in the form of*

$$\rho_{z_i y \cdot W_i} \Psi = q_{i1}\theta_1 + \cdots + q_{in}\theta_n, \tag{8.3}$$

such that Ψ is a function on correlations among variables in M, and each q_{il} for all $i = 1, \ldots, n'$ and $l = 1, \ldots, n$ satisfies the following conditions.

1. *If θ_l is a directed edge, then $q_{il} = \sum_{j=0}^{n} b_{i_j} a_{i_j l}$, where $a_{i_j l}$ and b_{i_j} are defined the same way as [BP12] Equation (10).*

2. *If θ_l is a bidirected edge, then $q_{il} = b_{i_0}$.*

The proof of Lemma 3 is given in Section 8.4.4.

8.4.1 Proof of Lemma 1

Proof. Given Lemma 3, and [BP12] Section 7.4, we have all those coefficients are functions on the correlations of variables in M. □

Note that the functions are not necessarily polynomials, since from the proof of Lemma 3, ϕ_i is a polynomial on correlations, while $\rho_{z_i y \cdot W_i}$ is ϕ_i divided by some functions on the correlations, which results in an arbitrary function.

8.4.2 Proof of Lemma 2

We prove Lemma 2 together with Theorem 1.

8.4.3 Proof of Theorem 1

Proof. To prove there exists a full-rank set of $N = n' + n_k + n_e$ linear constraints on E, we first construct a set of constraints, $\mathcal{L}$, such that $|\mathcal{L}| = N$. Then we prove each of the N constraints is linear, and finally we show that the set is full-rank.

Constructing the N constraints: We first construct the first n' constraints. Given a partial-instrumental set Z for E on E', w.l.o.g, denote $Z = \{z_1, \ldots, z_{n'}\}, E = \{e_1, \ldots, e_n\}, E' = \{e_1, \ldots, e_{n'}\}$. Also denote the triples in the definition of a basic-partial-instrumental set as $(z_1, W_1, p_1), \ldots, (z_{n'}, W_{n'}, p_{n'})$. Since each p_i is a path from z_i to $Ta(e_i)$, we can say each z_i matches to an edge $e_i \in E'$. From Lemma 1, we can create l_i, which is matched to z_i and e_i. See Lemma 3.

The left-hand side expression from Equation (8.3) and $q_{i1}, \ldots, q_{in}$ can all be calculated from the data. Hence, the first n' linear equations we construct for $\mathcal{L}$ are Equation (8.3) for $i = 1, \cdots, n'$.

Next, we construct the next group of n_e constraints in $\mathcal{L}$. For each $j = 1, \ldots, n_e$, we write the j-th constraint in E_e as

$$0 = d_j e^e_{j1} + e^e_{j2}, \tag{8.4}$$

where d_j is a constant, and e^e_{j1} and e^e_{j2} are the two edges involved in this equality constraint. W.l.o.g, we assume for each j, in the j-th constraint, the first edge, e^e_{j1}, is selected for the selection defined in the theorem.

Finally, we construct the remaining n_k of the constraints in $\mathcal{L}$. For each $h = 1, \ldots, n_k$,

the h-th edge in E_k is e_h^k, and we have a constraint

$$\lambda_h = e_h^k, \tag{8.5}$$

where λ_h is the known value of e_h^k.

Constructing a matrix of the constraints Now that we have a set of $N = n' + n_e + n_k$ constraints, in order to prove that they are linearly independent, we want to construct a matrix, and prove the matrix is full-row-rank. We first construct an ordering of the edges involved in those constraints.

The first n' edges are the edges in E', in the order of $e_1, \ldots, e_{n'}$. Since there exists a way to non-repetitively select one edge from each equality constraint that certain conditions are satisfied, let the selected edges, E_s, be the next n_e edges, with the ordering the same as the ordering of the equality constraints in $\mathcal{L}$. Denote those edges as $\{e_{11}^e, \ldots, e_{n_e 1}^e\}$, and the edges that are paired with those edges as $\{e_{12}^e, \ldots, e_{n_e 2}^e\}$. Next, the last n_k edges are those in E_k, with the ordering the same as the constraints in $\mathcal{L}$. Finally, any edges in $(E \cup E_k) \setminus (E' \cup E_s \cup E_k)$ can be of any order in the end. We can construct this order because as specified in the theorem, E', E_s, and E_k do not share any element.

Given the ordering of the edges, we can construct a matrix, where each term in the matrix is the coefficient in front of an edge in a constraint. Each row is one constraint in $\mathcal{L}$, in order, and each column is one edge, in the order we just specified. So we have an $N \times |E|$ matrix. To prove this matrix is full-row-rank, it suffices to prove the $N \times N$ sub-matrix containing the first N columns of the original matrix is full-rank. Below we give what the submatrix looks like (the first row in parentheses is used to indicate the edges for the matrix, and is

not part of the matrix.)

$$
\begin{array}{c}
\begin{array}{cccccccccc}
(e_1 & e_2 & \dots & e_{n'} & e^e_{11} & \dots & e^e_{n_e 1} & e^k_1 & \dots & e^k_{n_k})
\end{array} \\
\begin{bmatrix}
q_{11} & q_{12} & \dots & q_{1n'} & U & \dots & U & U & \dots & U \\
\vdots & \ddots & & & & & & & & \vdots \\
\vdots & & \ddots & & & & & & & \vdots \\
q_{n'1} & q_{n'2} & \dots & q_{n'n'} & U & \dots & U & U & \dots & U \\
0 & \dots & \dots & 0 & d_1 & \dots & \dots & 1 & \dots & 0 \\
\vdots & & & & & \ddots & & & & \vdots \\
0 & \dots & 1 & \dots & 0 & \dots & d_{n_e} & \dots & \dots & 0 \\
0 & \dots & \dots & \dots & \dots & \dots & 0 & 1 & \dots & 0 \\
\vdots & & & & & & & & \ddots & \vdots \\
0 & \dots & \dots & \dots & \dots & \dots & \dots & \dots & 0 & 1
\end{bmatrix}
\end{array}
$$

U denotes "unknown", which might be zero (if the edge corresponding to that column is not in E, or is in E but not in the constraint corresponding to that row,) or non-zero (if the edge corresponding to that column is in E and is in the constraint corresponding to that row.)

Proof that the matrix is full-rank To prove this matrix is full-rank, we simply have to prove that the determinant does not vanish. The determinant of an $N \times N$ matrix can be calculated using the Leibniz formula, which is summing up the product of N entries corresponding to all possible permutations of the set $\{1, 2, \dots, N\}$. Hence, we only have to prove that the product we get by selecting the first permutation, i.e., $\{1, 2, \dots, N\}$, cannot be canceled by any other products. In other words, we only have to prove that the product of the diagonal of the matrix has a term that cannot be canceled out by any other term from the expression of the determinant.

We define a term, T^* to be

$$T^* = \prod_{j=1}^{n'} T(p_j) \prod_{i=1}^{n_e} d_i, \tag{8.6}$$

where $T(p_j)$ is the product of the edge coefficients along the path p_j. T^* must exist in the product of the diagonal, since $\prod_{j=1}^{n'} T(p_j)$ exists in the product of the first n' entries from the diagonal (Lemma 3), and $\prod_{i=1}^{n_e} d_i$ is the product of the rest of the diagonal entries.

Suppose that T^* appears at least twice in the expression of the determinant. We first prove that T^* must come from selecting the diagonal terms of the matrix.

Note that each selection must select one entry from each row and each column, from the Leibniz formula. We must select the diagonal for the last n_k entries, since if a non-diagonal entry was selected, that entry must be 0, and the whole product would be 0.

Next, we must also select the diagonal for the middle n_e entries, $d_1, \ldots, d_{n_e}$. We prove this argument by proving that if we do not select the diagonal, then we cannot reproduce the product of the n_e diagonal entries no matter what edges we select, which means any term in our selection cannot cancel out T^*. Suppose this is not true, i.e., even if we do not select the diagonal entries for the middle n_e rows, we can still get the product somewhere else.

Recall that for each $j = 1, \ldots, n_e$, $d_j e^e_{j1} + e^e_{j2} = 0$. Since this equality constraint should comply with the actual values of the edges e^e_{j1}, e^e_{j2} in the model M, we have for each j,

$$d_j = -\frac{e^e_{j2}}{e^e_{j1}}. \tag{8.7}$$

Denote the product of the diagonal entries for the middle n_e rows as T_m, then

$$T_m = (-1)^{n_e} \prod_{j=1}^{n_e} \frac{e^e_{j2}}{e^e_{j1}}. \tag{8.8}$$

Terms cancel out if we have the same edge with one occurrence on the numerator and one occurrence on the denominator. So we might end up having a simplified expression,

$$T_m = (-1)^{n_e} \prod_{i} \frac{e^e_{n_i}}{e^e_{d_i}}. \tag{8.9}$$

Note that T_m cannot be $(-1)^{n_e}$, where all edges cancel out. We next examine where those edges might appear in the matrix. First note that the terms in the first n' rows do not contain any edge in $Inc(y)$ (Lemma 3).

b_{i_j} are polynomials on the correlations among $z_i, W_{i_1}, \ldots, W_{i_k}$ and $a_{ijl} = \rho_{W_{i_j} x_l}$. All edges in $Inc(y)$ have a head y, which means no edge in $Inc(y)$ can appear in the correlations among the non-descendants of y (this can be seen from Wrights' rules.) So q_{il}, which is made up of correlations among W_{i_j}, x_i, z_i (all non-descendants of y,) does not contain any edge in $Inc(y)$.

Hence, the edges in T_m cannot be canceled out by anything in the first n' rows, which means T^* will contain T_m as it is.

Suppose we select only a subset of n'_e the diagonal entries for the middle n_e rows. For each row where the diagonal is not selected, 1 must be selected (otherwise we will have to select 0, and the product will be 0.) So we end up having the product of the selected entries from the middle n_e rows, T'_m as

$$T'_m = (-1)^{n'_e} \prod_i \frac{e^e_{n'_i}}{e^e_{d'_i}}. \tag{8.10}$$

T_m and T'_m cannot be equal to each other. Otherwise, we produce a constraint on those edges by equating T_m and T'_m. However, given that the equality constraints are linearly independent, the values of those edges should vary independently and should not comply to any constraint. In other words, they are equal only when the constraint is satisfied, which has Lebesgue measure 0, so we assume that is not the case. Thus, to cancel out T^*, we must select the diagonal of the middle n_e rows.

We have proved that for the last $n_e + n_k$ rows, we must select the entries on the diagonal. For the first n' rows, we can only select from the first n' columns, since we can only select one entry from each column, and the last $n_e + n_k$ columns already have entries been selected. Therefore, we only need to analyze the top left $n' \times n'$ submatrix. The problem reduces to proving the term

$$t^* = \prod_{j=1}^{n'} T(p_j) \tag{8.11}$$

exists only once in the determinant of this submatrix. We first prove that t^* appears only

once in the product of the diagonal entries. We use the same proof strategy as in [BP12] Proof of Lemma 8. To get t^*, we need to select one term from each diagonal entry such that the product of those terms gives t^*. From [BP12] Proof of Lemma 8, for q_{jj} where column j is a directed edge, if we select the second or the third term of q_{jj} in [BP12] Equation (11), then it must bring in a term that is not in t^*, or causes the product to contain a term in t^* twice. Hence, for those q_{jj} entries, we can only select from the first term in Equation in [BP12] Equation (11). After eliminating those terms from consideration, the remaining terms in the product of the n' diagonal terms are given by

$$t^* \prod_{i \text{ for directed}} (1 + \hat{b}_{i_0}) \prod_{k \text{ for bidirected}} (1 + \hat{b}_{k_0}) \tag{8.12}$$

$$= t^* \prod_{j}^{n'} (1 + \hat{b}_{j_0}), \tag{8.13}$$

From [BP12], $\hat{b}_{j_0}$ are polynomials on correlations among W_i, and they do not have any constant terms. As a result, t^* appears only once in Equation (8.13), and thus appears only once in the product of the diagonal entries.

What remains to prove is that t^* does not appear in the product of another selection of entries, which is different from selecting all the diagonals. For the columns that correspond to bidirected edges, we have to select the diagonal terms, since those are the only terms in those column that are non-zero. We generate a submatrix by removing those columns corresponding to bidirected edges and those rows with the same row numbers as those column numbers. This submatrix is a square matrix, and all columns correspond to directed edges. This reduces to the proof of Theorem 1 from [BP12], where they proved that no matter which selection we have, the term $\prod_{j \text{ for directed}} T(p_j)$ can never be canceled.

To sum up, we showed that one can never find another term in the determinant that can cancel out a term, T^*, which is also in the determinant. Hence, the $N \times N$ sub-matrix is full-rank, and the $N \times E$ matrix is full-row-rank.

Finally, when $N = |E|$, we have a full-rank set of N linear equations on N edges, so we

can solve for all of the edges. □

8.4.4 Proof of Lemma 3

Proof. From Lemma 1 in [BP12], denoting $W_i = \{W_{i_1}, \ldots, W_{i_k}\}$ (we assume W_i contains k single variables), we have

$$\rho_{z_i y \cdot W_i} = \frac{\phi_i(z_i, y, W_{i_1}, \ldots, W_{i_k})}{\psi_i(z_i, W_{i_1}, \ldots, W_{i_k})\psi_i(y, W_{i_1}, \ldots, W_{i_k})}, \tag{8.14}$$

where ϕ is linear on the correlations $\rho_{z_i y}, \rho_{W_{i_1} y}, \ldots, \rho_{W_{i_k} y}$, and the square of each of the ψ functions is a polynomial on correlations among the variables it takes. We can write

$$\phi_i = b_{i_0}\rho_{z_i y} + b_{i_1}\rho_{W_{i_1} y} + \cdots + b_{i_1}\rho_{W_{i_k} y}. \tag{8.15}$$

We only need to prove that ϕ_i is linear on the edges $e_1, \ldots, e_n$ and does not contain any constant term. Since $\rho_{z_i y \cdot W_i}$ vanishes in $G_{E \cap D \cup \{\varepsilon_i\}-}$ from the definition of a partial-instrumental set, $\phi(z_i, y, W_{i_1}, \ldots, W_{i_k})$ must also vanish in $G_{E \cap D \cup \{\varepsilon_i\}-}$. For all bidirected edges in $Inc(y)$, we can treat them as two directed sub-edges connected at the tails. Hence, [BP12]'s Lemmas 6 and 7 apply. Let e'_j be the same as e_j if e_j is directed, and the sub-edge pointing to y if e_j is bidirected, and we immediately have that ϕ_i is linear on the edges $e'_1, \ldots, e'_n$ and does not contain any constant term. If e_j is bidirected, ϕ_i being linear on e'_j is equivalent to that ϕ_i is linear on e_j. From Lemma 7, we have that all edges not in $E \cap D \cup \{\varepsilon_i\}$ have coefficient 0. Hence, either $\varepsilon_{z_i y}$ is the only bidirected edge in the constraint l_i, or there exists no bidirected edge in l_i.

$\rho_{z_i y \cdot W_i}$ can be written in the form of $\rho_{z_i y \cdot W_i} = c_{i1}e_1 + \cdots + c_{in}e_n$. We then apply the results from Section 7.4 in [BP12] and we have for each j where e_j is a directed edge, c_{ij} is a function of the correlations of variables in M.

If there does not exist a bidirected edge among $\theta_1, \ldots, \theta_n$, then the lemma is evident from the result from [BP12]. If there exists a bidirected edge, w.l.o.g, assume θ_n is the bidirected edge. Now we examine every q_{ij}.

First we can decompose θ into two directed edges, one pointing to y and one does not include y. Let the decomposition be $\theta_n = \alpha\beta$, where α is the edge pointing to y. We can thus write $\rho_{z_i y \cdot W_i}$ in the form of

$$\rho_{z_i y \cdot W_i} = q_{i1}\theta_1 + \cdots + q_{i(n-1)}\theta_{n-1} + q_{in}\beta\alpha. \tag{8.16}$$

Now we have a linear equation on directed edges $\theta_1, \ldots, \theta_{n-1}, \alpha$. Hence, the results from [BP12] applies, and we know for j where θ_j is a directed edge, q_{ij} is the same as the way defined in [BP12].

The coefficient of α can also be regarded as $\sum_{j=0}^{n} b_{i_j} a_{i_j n}$. Recall the definition in [BP12], each $a_{i_j n}$ is the sum of paths from z_i or W_i to y passing through θ_n, but not including θ_n. From Definition 4, each W_{i_j} is non-descendant of z_i, so any unblocked path from W_{i_j} to z_i must have an arrowhead at z_i, which makes z_i a collider (also named as "sink" or "convergent") between W_{i_j} and y, and blocks the path between z_i and y. Since no paths from other z_i or W_i can pass through the bidirected edge, the only non-zero $a_{i_j n}$ is $a_{i_0 n}$, which is the sum of paths from z_i to y through θ_n but not including θ_n, which is equal to 1. The corresponding multiplier is b_{i_0}. Since the index of the bidirected edge among $\theta_1, \ldots, \theta_n$ does not matter, we assumed the bidirected edge is of index n for the convenience of discussion. Now we can replace n with l and we have the coefficient $q_{il} = b_{i_0} \cdot 1 = b_{i_0}$.

□

8.4.5 Discussion on the Example in Section 7.1

In Figure 3 left, if the equality constraint is instead $\lambda_{ux} = \lambda_{uw}$, then the equality constraint in the latent projection DAG is $\varepsilon_{xy} = \varepsilon_{wy}$. $\{w, x\}$ form a partial-instrmental set for $\{\varepsilon_{wy}, \varepsilon_{xy}, \lambda_{xy}\}$ on $\{\varepsilon_{wy}, \lambda_{xy}\}$. Together with the equality constraint, we can solve for all edges.

If the equality constraint is instead $\lambda_{ux} = \lambda_{uy}$, then the equality constraint in the latent projection DAG is $\varepsilon_{xw} = \varepsilon_{wy}$. ε_{xw} is identified ($\varepsilon_{xw} = \rho_{xw}$). Then with the equality con-

straint, ε_{wy} is identified. $\varepsilon_{xy} = \varepsilon_{wy}$. $\{w, x\}$ form a partial-instrmental set for $\{\varepsilon_{wy}, \varepsilon_{xy}, \lambda_{xy}\}$ on $\{\varepsilon_{xy}, \lambda_{xy}\}$. Together with the value of ε_{wy}, we can solve for all edges.