In Proceedings of International Joint Conference on Artificial Intelligence, Edinburgh, Schotland, 357--363, August 2005.

TECHNICAL REPORT
R-321
June 2005

Identifiability of Path-Specific Effects*

Chen Avin, Ilya Shpitser, Judea Pearl
Cognitive Systems Laboratory
Department of Computer Science
University of California, Los Angeles
Los Angeles, CA. 90095
{avin, ilyas, judea}@cs.ucla.edu

Abstract

Counterfactual quantities representing path-specific effects arise in cases where we are interested in computing the effect of one variable on another only along certain causal paths in the graph (in other words by excluding a set of edges from consideration). A recent paper [Pearl, 2001] details a method by which such an exclusion can be specified formally by fixing the value of the parent node of each excluded edge. In this paper we derive simple, graphical conditions for experimental identifiability of path-specific effects, namely, conditions under which path-specific effects can be estimated consistently from data obtained from controlled experiments.

1 Introduction

Total, direct and indirect effects are important quantities in practical causal reasoning about legal, medical, and public policy domains, among others. The task of explicating, and computing these quantities has been successfully addressed in the framework of linear structural equation models (SEM), but encountered difficulties in non-linear as well as non-parametric models. See for instance [Robins and Greenland, 1992], [Galles and Pearl, 1995], [Pearl, 2001],

In the linear SEM framework, the *total effect* of Z on Y is the response of Y to a unit change in the setting of Z. On the other hand, the *direct effect* is the effect of Z on Y not mediated by any other variable in the model while the *indirect effect* is the effect of Z on Y excluding the direct effect.

In non-parametric models, we can define the *controlled* direct effect as the change in the measured response of Y to a change in Z, while all other variables in the model, henceforth called *context variables*, are held constant. Unfortunately, there is no way to construct an equivalent notion of controlled indirect effects, since it is not clear to what values other variables in the model need to be fixed in order to measure such an effect.

Recently, a novel formulation of *natural* [Pearl, 2001] or *pure* [Robins and Greenland, 1992] effects was proposed which defined effects in a more refined way by holding variables constant not to predetermined values, but to values they would have attained in some situation. For example, the natural direct effect of Z on Y is the sensitivity of Y to changes in Z, while the context variables are held fixed to the values they would have attained had no change in Z taken place. Similarly, the natural indirect effect is the sensitivity of Y to changes the context variables would have undergone had Z been changed, while Z is actually being fixed.

Being complex counterfactual quantities, natural effects tend to have intricate verbal descriptions. It is often easier to explain such effects using the visual intuitions provided by graphical causal models. Graphical causal models represent causal assumptions as graphs, with vertices representing variables, and edges representing direct causal paths. In such models, natural direct effect can be interpreted as the effect along the edge $Z \rightarrow Y$, with the effect along all other edges 'turned off.' Similarly, the natural indirect effect can be interpreted as the effect along all edges except the one between Z and Y. Using this interpretation, the suggestive next step in the study of natural effects is to consider effects along a select subset of edges between Z and Y which are called *path-specific* effects.

1.1 A Motivating Example

Consider the following example, inspired by [Robins, 1997], A study is performed on the effects of the AZT drug on AIDS patients. AZT is a harsh drug known to cause a variety of complications. For the purposes of the model, we restrict our attention to two – pneumonia and severe headaches. In turn, pneumonia can be treated with antibiotics, and severe headache sufferers can take painkillers. Ultimately, all the above variables, except headache, are assumed to have a direct effect on the survival chances of the patient. The graphical causal model for this situation is shown in Fig. 1.

The original question considered in this model was the total effect of AZT and antibiotics treatment on survival. However, a variety of other questions of interest can be phrased in terms of natural effects. For instance, what is the direct effect of AZT on survival, if AZT produced no side effects in the patient, which is just the natural direct effect of AZT on survival. See Fig. 2 (a). Similarly, we might be interested in how just the side effects of AZT affect survival, independent of the effect of AZT itself. This corresponds to the natural

*This research was partially supported by AFOSR grant #F49620-01-1-0055, NSF grant #IIS-0097082, and ONR (MURI) grant #N00014-00-1-0617.

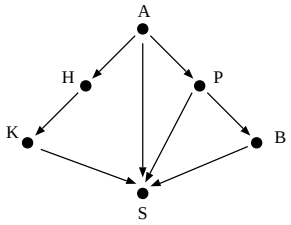

Figure 1: The AZT example. A: AZT, P: pneumonia, H: headaches, B: antibiotics, K: painkillers, S: survival

indirect effect of AZT on survival. See Fig. 2 (b).

Furthermore, certain interesting questions cannot be phrased in terms of either direct or indirect natural effects. For example we might be interested in the interactions between antibiotics and AZT that negatively affect survival. To study such interactions, we might consider the effect of administering AZT on survival in the idealized situation where the antibiotics variable behaved as if AZT was not administered, and compare this to the total effect of AZT on survival. Graphically, this amounts to 'blocking' the direct edge between antibiotics and survival or more precisely, keeping the edge functioning at the level it would have had no AZT been given, while letting the rest of the edges function as usual. This is shown graphically in Fig. 3 (a). The edges which we wish to block will be crossed out in the graph.

1.2 Outline and Discussion of Our Approach

Our goal is to study and characterize situations where path-specific effects like the one from the previous section can be computed uniquely from the data available to the investigator. Our main result is a simple, necessary, graphical condition for the identifiability of path-specific effects from experimental data. Furthermore, our condition becomes sufficient for models with no spurious correlations between observables, also known as Markovian models.

The condition can be easily described in terms of blocked and unblocked paths as follows. Let X, Y be variables in a causal model M inducing a graph G. Then given a set of blocked edges g, the corresponding path-specific effect of X on Y cannot be identified if and only if there exists a node W with an unblocked directed path from X to W, an unblocked directed path from W to Y, and a blocked directed path from W to Y. For instance, the effects of A on S are identifiable in Fig. 2 (a), (b), and Fig. 3 (b), but not in Fig. 3 (a). Therefore, in general we cannot study the interractions of AZT and antibiotics in the way described above, but we can study the interractions of AZT and painkillers. The latter case is made tractable by an absense of blocked and unblocked paths sharing edges.

Our condition also shows that all identifiable path-specific effects are 'equivalent', in a sense made precise later, to effects where only root-emanating edges are blocked. Thus identifiable path-specific effects are a generalization of both natural direct effects, where a single root-emanating edge is unblocked, and of natural indirect effects, where a single root-emanating edge is blocked.

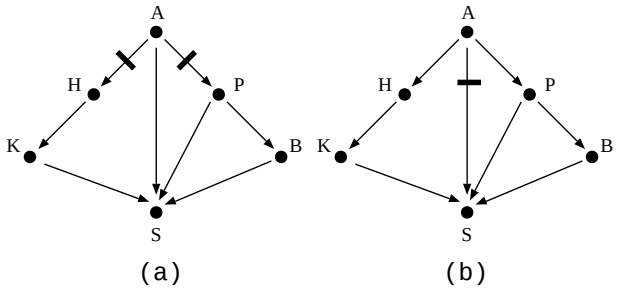

Figure 2: (a) Natural direct effect (b) Natural indirect effect

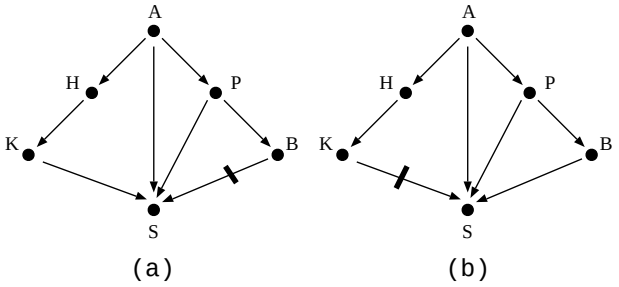

Figure 3: Path specific effects

To obtain this result formally, we treat effects as probabilities of statements in a certain counterfactual logic. However, rather than manipulating these probabilities directly, we convert them to subgraphs of the original causal model, and reason about and perform manipulations on the subgraphs. We then introduce simple counterfactual formulas whose probabilities are not identifiable, and prove that certain simple graphical conditions must be described by such formulas, and lack of such conditions leads to subgraphs corresponding to identifiable effects.

Due to space considerations, the proofs of some lemmas have been omitted, while the proofs included generally are missing some technical details. Our technical report contains the complete proofs.

2 Preliminaries

This paper deals extensively with causal models and counterfactuals. We reproduce their definitions here for completeness. A full discussion can be found in [Pearl, 2000]. For the remainder of the paper, variables will be denoted by capital letters, and their values by small letters. Similarly, sets of variables will be denoted by bold capital letters, sets of values by bold small letters. We will also make use of some graph theoretic abbreviations. We will write $Pa(A)_G$, $De(A)_G$, and $An(A)_G$, to mean the set of parents, descendants (inclusive), and ancestors (inclusive) of node A in graph G. G will be omitted from the subscript when assumed or obvious. If a variable is indexed, i.e. V^i, we will sometimes denote the above sets as Pa^i, De^i, and An^i, respectively.

2.1 Causal Models and Counterfactual Logic

Definition 1 *A probabilistic causal model (PCM) is a tuple* $M = \langle \boldsymbol{U}, \boldsymbol{V}, \boldsymbol{F}, P(\boldsymbol{u}) \rangle$*, where*

(i) $\boldsymbol{U}$ *is a set of background or exogenous variables, which cannot be observed or experimented on, but which can influence the rest of the mode*

(ii) $\boldsymbol{V}$ *is a set* $\{V^1, ..., V^n\}$ *of observable or endogenous variables. These variables are considered to be functionally dependent on some subset of* $\boldsymbol{U} \cup \boldsymbol{V}$.

(iii) $\boldsymbol{F}$ *is a set of functions* $\{f^1, ..., f^n\}$ *such that each* f^i *is a mapping from a subset of* $\boldsymbol{U} \cup \boldsymbol{V} \setminus \{V^i\}$ *to* V^i*, and such that* $\bigcup \boldsymbol{F}$ *is a function from* $\boldsymbol{U}$ *to* $\boldsymbol{V}$.

(iv) $P(\boldsymbol{u})$ *is a joint probability distribution over the variables in* $\boldsymbol{U}$.

A causal model M induces a directed graph G, where each variable corresponds to a vertex in G and the directed edges are from the variables in the domain of f^i (i.e Pa^i) to V^i for all the functions. For the remainder of this paper, we consider causal models which induce directed acyclic graphs.

A Markovian causal model M has the property that each exogenous variable U is in the domain of at most one function f. A causal model which does not obey this property is called semi-Markovian. By convention, nodes corresponding to variables in **U** are not shown in graphs corresponding to Markovian models.

For the purposes of this paper, we will represent counterfactual statements in a kind of propositional modal logic, similar to the one used in [Halpern, 2000]. Furthermore, the distribution $P(\boldsymbol{u})$ will induce an additional probabilistic interpretation on the statements in the logic.

Definition 2 (atomic counterfactual formula) *Let* M *be a causal model, let* X *be a variable and* $\boldsymbol{Z}$ *be a (possibly empty) set of variables. Then for any value* x *of* X*, and values* $\boldsymbol{z}$ *of* $\boldsymbol{Z}$*,* x *is a term, and* $X_{\boldsymbol{z}}(\boldsymbol{u})$ *is a term, taken to mean 'the value* X *attains when* $\boldsymbol{Z}$ *is forced to take on values* $\boldsymbol{z}$*, and* $\boldsymbol{U}$ *attain values* $\boldsymbol{u}$*.'*

For two terms t_1 *and* t_2*, an atomic counterfactual formula has the form* $t_1 = t_2$*. We will abbreviate formulas of the form* $X_{\boldsymbol{z}}(\boldsymbol{u}) = x$ *as* $x_{\boldsymbol{z}}(\boldsymbol{u})$.

The 'forcing' of the variables to $\boldsymbol{z}$ is called an intervention, and is denoted by do($\boldsymbol{z}$) in [Pearl, 2000]. Counterfactual formulas are constructed from atomic formulas using conjunction and negation.

Definition 3 (counterfactual formula)

(i) An atomic formula $\alpha(\boldsymbol{u})$ *is a counterfactual formula.*

(ii) If $\alpha(\boldsymbol{u})$ *is a counterfactual formula, then so is* $(\neg\alpha)(\boldsymbol{u})$.

(iii) If $\alpha(\boldsymbol{u})$ *and* $\beta(\boldsymbol{u})$ *are counterfactual formulas, then so is* $(\alpha \wedge \beta)(\boldsymbol{u})$.

The satisfaction of counterfactual formulas by causal models is defined in the standard way, which we reproduce from [Halpern, 2000].

Definition 4 (entailment) *A causal model* M *satisfies a counterfactual formula* $\alpha(\boldsymbol{u})$*, written* $M \models \alpha(\boldsymbol{u})$*, if all variables appearing in* α *are in* M *and one of the following is true*

(i) $\alpha(\boldsymbol{u}) \equiv t_1 = t_2$ *and for the given setting of* $\boldsymbol{u}$*, the terms* t_1 *and* t_2 *are equal in* M.

(ii) $\alpha(\boldsymbol{u}) \equiv (\neg\beta)(\boldsymbol{u})$ *and* $M \not\models \beta(\boldsymbol{u})$.

(iii) $\alpha(\boldsymbol{u}) \equiv (\beta \wedge \gamma)(\boldsymbol{u})$ *and* $M \models \beta(\boldsymbol{u})$ *and* $M \models \gamma(\boldsymbol{u})$

Thus a formula $\alpha(\boldsymbol{u})$ has a definite truth value in M. If the values $\boldsymbol{u}$ are unknown, we cannot in general determine the truth of α. However, we can easily define a natural notion of probability of α in M as follows:

$$P(\alpha|M) = \sum_{\{\boldsymbol{u}|M \models \alpha(\boldsymbol{u})\}} P(\boldsymbol{u}) \tag{1}$$

We will omit the conditioning on M if the model in question is assumed or obvious.

If we consider each value assignment $\boldsymbol{u}$ as a $possible\ world$, then we can view $P(\boldsymbol{u})$ as describing our degree of belief that a particular world is true, and $P(\alpha)$ as our belief that a particular statement is true in our causal model if viewed as a $type\ 2\ probability\ structure$ [Halpern, 1990].

2.2 Submodels and Identifiability

Definition 5 (submodel) *For a causal model* $M = \langle \boldsymbol{U}, \boldsymbol{V}, \boldsymbol{F}, P(\boldsymbol{u}) \rangle$*, an intervention* $do(\boldsymbol{z})$ *produces a new causal model* $M_{\boldsymbol{z}} = \langle \boldsymbol{U}, \boldsymbol{V}_{\boldsymbol{z}}, \boldsymbol{F}_{\boldsymbol{z}}, P(\boldsymbol{u}) \rangle$*, where* $\boldsymbol{V}_{\boldsymbol{z}}$ *is a set of distinct copies of variables in* $\boldsymbol{V}$*, and* $\boldsymbol{F}_{\boldsymbol{z}}$ *is obtained by taking distinct copies of functions in* $\boldsymbol{F}$*, but replacing all copies of functions which determine the variables in* $\boldsymbol{Z}$ *by constant functions setting the variables to values* $\boldsymbol{z}$.

The joint distribution $P(\boldsymbol{V}_{\boldsymbol{z}})$ over the endogenous variables in $M_{\boldsymbol{z}}$ is called an interventional distribution, and is sometimes denoted as $P_{\boldsymbol{z}}$. For a given causal model M, define P_* as $\{P_{\boldsymbol{z}} | \boldsymbol{Z} \subseteq \boldsymbol{V}, \boldsymbol{z}$ a value assignment of $\boldsymbol{Z}\}$. In other words, P_* is the set of all possible interventional (or experimental) distributions of M.

Intuitively, the submodel is the original causal model, minimally altered to render $\boldsymbol{Z}$ equal to $\boldsymbol{z}$, while preserving the rest of its probabilistic structure.

Because there is no requirement that interventions in atomic counterfactuals in a formula α be consistent with each other, it is in general impossible to alter the original model using only interventions in such a way as to make the entire formula true. Thus, we introduce a causal model which encompasses the 'parallel worlds' described by the counterfactual formula.

Before doing so, we give a simple notion of union of submodels, as follows:

Definition 6 (causal model union) *Let* $M_{\boldsymbol{x}}$*, and* $M_{\boldsymbol{z}}$ *be submodels derived from* M*. Then* $M_{\boldsymbol{x}} \cup M_{\boldsymbol{z}}$ *is defined to be* $M_{\boldsymbol{x}}$ *if* $\boldsymbol{z} = \boldsymbol{x}$*, and* $\langle \boldsymbol{U}, \boldsymbol{V}_{\boldsymbol{x}} \cup \boldsymbol{V}_{\boldsymbol{z}}, \boldsymbol{F}_{\boldsymbol{x}} \cup \boldsymbol{F}_{\boldsymbol{z}}, P(\boldsymbol{u}) \rangle$*, otherwise.*

Definition 7 (parallel worlds model) *Let* M *be a causal model,* α *a counterfactual formula. Then the parallel worlds model* M_α *is the causal model union of the submodels corresponding to atomic counterfactuals of* α.

We call the joint distribution $P(\boldsymbol{V}_\alpha)$ over the endogenous variables in M_α a counterfactual distribution, and will sometimes denote it as P_α. In the language of the potential outcomes framework [Rubin, 1974], we can view P_α as the joint distribution over the unit-response variables mentioned in α.

The parallel worlds model is a generalization of the twin network model, first appearing in [Balke and Pearl, 1994], to more than two possible worlds. It displays independence assumptions between counterfactual quantities in the same way a regular causal model displays independence assumptions between observable quantities – by positing counterfactuals are independent of their non-descendants given their parents.

Given a causal model M and a formula α, we are interested in whether the corresponding counterfactual joint distribution P_α (or its marginal distributions) can be computed uniquely from the set of joint distributions available to the investigator. The formal statement of this question is as follows:

Definition 8 (identifiability) *Let M be a causal model from a set of models $\boldsymbol{M}$ inducing the same graph G, M_α a parallel worlds model, and Q be a marginal distribution of the counterfactual joint distribution P_α. Let K be a set of known probability distributions derived from M. Then Q is K-identifiable in $\boldsymbol{M}$ if it is unique and computable from K in any $M \in \boldsymbol{M}$.*

It follows from the definition that if we can construct two models in $\boldsymbol{M}$ with the same K but different Q, then Q is not identifiable. An important, well-studied special case of this problem – which we call evidential identifiability of interventions – assumes α is an atomic counterfactual, and K is the joint distribution over the endogenous variables in M, or $P(V)$. Being able to identify an interventional marginal in this way is being able to compute the effects of an intervention without having to actually perform the intervention, and instead relying on passive, observational data.

In this paper we are concerned with identifying probabilities of counterfactuals formulas using the set P_* of all interventional distributions of M as a given. In other words, we are interested in computing probabilities of counterfactuals from experimental and observational probabilities.

3 Path-Specific Effects

Our aim is to provide simple, graphical conditions for the P_*-identifiability of path-specific effects. To do so, we must formalize such effects as counterfactual formulas, and translate the identifiability conditions on the formula to conditions on the graph.

The following is the formalization of the notion of path-specific effect in terms of a modified causal model, as it appears in [Pearl, 2001]:

Definition 9 (path-specific effect) *Let G be the causal graph associated with model M, and let g be an edge-subgraph of G containing the paths selected for effect analysis (we will refer to g as the **effect subgraph**). The g-specific effect of z on Y (relative to reference z^*) is defined as the total effect of z on Y in a modified model M_g formed as follows. Let each parent set PA^i in G be partitioned into two parts $PA^i = \{PA^i(g), PA^i(\bar{g})\}$, where $PA^i(g)$ represents those members of PA^i that are linked to V^i in g, and $PA^i(\bar{g})$ represents the complementary set. We replace each function f^i in M with a new function f^i_g in M_g, defined as follows: for every set of instantiations $pa^i(g)$ of $PA^i(g)$, $f^i_g(pa^i(g), \boldsymbol{u}) = f^i(pa^i(g), pa^i(\bar{g})^*, \boldsymbol{u})$, where $pa^i(\bar{g})^*$ takes the value of $PA^i(\bar{g})_{z^*}(\boldsymbol{u})$ in M. The collection of modified functions forms a new model M_g. The g-specific effect of z on Y, denoted $SE_g(z, z^*; Y, \boldsymbol{u})_M$ is defined as the total effect (abbreviated as TE) of z on Y in the modified model:*

$$SE_g(z, z^*; Y, \boldsymbol{u})_M = TE(z, z^*; Y, \boldsymbol{u})_{M_g} \tag{2}$$

where $TE(z, z^*; Y, \boldsymbol{u})_{M_g} = Y_z(\boldsymbol{u})_{M_g} - Y_{z^*}(\boldsymbol{u})_{M_g}$.

If we wish to summarize the path-specific effect over all settings of $\boldsymbol{u}$, we should resort to the expectation of the above difference, or the expected path-specific effect. To identify this effect, we need to identify $P(y_z)$ and $P(y_{z^*})$ in M_g. For our purposes we can restrict our attention to $P(y_z)$, as the second term corresponds to the quantity $P(y_{z^*})$ in the original model M, and so is trivially P_*-identifiable.

In this paper we assume, without loss of generality, edges in $\bar{g} = G \setminus g$ are all along directed paths between Z and Y. The next theorem states that any path specific effect, expressed as a total effect in the modified model M_g, can be expressed as a counterfactual formula in the original model M.

Theorem 1 *Every path specific effect $P(y_z)_{M_g}$ has a corresponding counterfactual formula α in M s.t for every $\boldsymbol{u}$,*

$$M_g \models y_z(\boldsymbol{u}) \iff M \models \alpha(\boldsymbol{u})$$

Proof outline: The proof is for causal models with finite domains. Fix M, $\boldsymbol{u}$, y, z and g. To prove the theorem, we need to 'unroll' y_z and remove any implicit references to modified functions in M_g, while preserving the truth value of the statement. Our proof will use the axiom of composition, known to hold true for causal models under consideration. In our language, the axiom states that for any three variables Z, Y, W, and any settings $\boldsymbol{u}, z, w, y$, $(W_z = w \Rightarrow Y_{z,w} = Y_z)(\boldsymbol{u})$.

Fix $\boldsymbol{u}_1$. Let $\boldsymbol{S} = An(Y) \cap De(Z)$ Then by axiom of composition, $y_z(\boldsymbol{u}_1)$ has the same truth value as a conjunction of atomic formulas of the form $v^i_{pa^i(g)}$, where $V^i \in \boldsymbol{S}$, $PA^i(g)$ is the set of parents of V^i in M_g, and $pa^i(g)$ and v^i are suitably chosen constants. Denote this conjunction α_1.

For every term $v^i_{pa^i(g)}$ in α_1 corresponding to V^i with $PA^i(g) \subset PA^i$, replace it by $v^i_{pa^i(g), pa^i(\bar{g})^*} \wedge pa^i(\bar{g})^*_{z^*}$ in the conjunction, where $pa^i(\bar{g})^*$ takes the value of $PA^i(\bar{g})_{z^*}(\boldsymbol{u}_1)$ in M. Denote the result α^*_1. Note that α^*_1 is in M and $M_g \models y_z(\boldsymbol{u}_1) \iff M \models \alpha^*_1(\boldsymbol{u}_1)$. We construct a similar conjunction α^*_j for every instantiation $\boldsymbol{u}_j$ in M. Let $\alpha = \bigvee_j \alpha^*_j$. It's easy to see the claim holds for α by construction. □

An easy corollary of the theorem is, as before, that $P(y_{\boldsymbol{z}})_{M_g} = P(\alpha)_M$. Note that different α_i in the proof only differ in the values they assign to variables in $\boldsymbol{S}$. Since M is composed of functions, the values of variables in $\boldsymbol{S}$ are fixed given $\boldsymbol{u}$, and since $P(\alpha) = \sum_{\{u | M \models \bigvee_i \alpha_i(u)\}} P(u)$ by definition, we can express $P(\alpha)$ as a summation over the variables in $S \setminus \{Y\}$.

For instance, the first term of the path-specific effect in Fig.

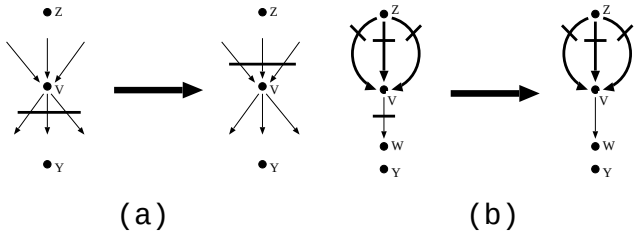

Figure 4: Bold edges represent directed paths (a) R_1 Rule (b) R_2 Rule

2 (a) can be expressed as

$$P(s_a)_{M_{g2a}} = \sum_{k,b,p,h} P(s_{k,b,p,a} \wedge k_h \wedge b_p \wedge p_{a*} \wedge h_{a*}) = \sum_{h,p} P(s_{a,h,p} \wedge h_{a*} \wedge p_{a*}) \quad (3)$$

which is just the direct effect. The more general case of Fig. 3 (a) can be expressed as:[1]

$$P(s_a)_{M_{g3a}} = \sum_{k,b,p,h} P(s_{k,b,p,a} \wedge k_h \wedge b_{a^*} \wedge p_a \wedge h_a) = \sum_{b} P(s_{a,b} \wedge b_{a^*}) \quad (4)$$

It looks as if the expressions in Eq. (3) and (4) for the two effects are very similar, moreover we know that direct effects are always P_*-identifiable in Markovian models. Surprisingly, the path specific effect of Fig. 3 (a) and Eq. (4) is not P_*-identifiable as we will show later.

We will find it useful to modify the effect subgraph g while preserving the value of the path-specific effect. We do so by means of the following two rules. Let M be a causal model with the graph G, g an effect subgraph of G, and $\bar{g} = G \setminus g$. For a node V, let $in(V)$ denote the set of edges incoming into V, and $out(V)$ denote the set of edges outgoing from V, in G.

R_1: If there is a node V in G such that $out(V) \subseteq \bar{g}$, then $R_1(g) = (g \setminus out(V)) \cup in(V)$. See Fig. 4 (a).

R_2: If there is an edge $e \in \bar{g}$, such that for all directed paths from Z to Y which include e, there exists another edge $e' \in \bar{g}$, which occurs 'upstream' from e, then $R_2(g) = g \setminus \{e\}$. See Fig. 4 (b).

Theorem 2 (Effect-Invariant Rules) *If R_1 is applicable the $R_1(g)$-specific effect is equal to the g-specific effect. If R_2 is applicable the $R_2(g)$-specific effect is equal to the g-specific effect.*

Proof outline: The proof is by induction on graph structure, and is an easy consequence of the definition of g-specific effect, and the R_1 and R_2 rules. □

Intuitively, R_1 'moves' the blocked edges closer to the manipulated variable Z, and R_2 removes redundant blocked edges. Thus, it is not surprising these two identities cannot be applied forever in a dag.

[1]Note that Eq (4) is different from $\sum_{b_{a^*}} P(s_{a,b} \wedge b_{a^*})$ which is just a marginalization over the counterfactual variable b_{a^*}

Table 1: The functions f_R^1 and f_R^2

Z	U_R	$R = f_R^1(z, u_R)$	$R = f_R^2(z, u_R)$
0	1	0	1
0	2	1	1
0	3	1	0
1	1	1	1
1	2	0	0
1	3	0	0

Lemma 1 *Let M be a causal model, g an effect subgraph. Then any sequence of applications of R_1 and R_2 to g will reach a fixed point g^*.*

4 Problematic Counterfactual Formulas

Identification of a distribution must precede its estimation, as there is certainly no hope of estimating a quantity not uniquely determined by the modeling assumptions. Furthermore, uniqueness frequently cannot be guaranteed in causal models. For instance, when identifying interventions from observational data, a particular graph structure, the 'bow-arc', has proven to be troublesome. Whenever the graph of a causal model contains the bow-arc, certain experiments become unidentifiable [Pearl, 2000]. Our investigation revealed that a similarly problematic structure exists for experimental identifiability, which we call the 'kite graph', due to its shape. The kite graph arises when we try to identity counterfactual probabilities of the form $P(r_{z^*} \wedge r'_z)$.

Lemma 2 *Let M be a causal model, let Z and R be variables such that Z is a parent of R. Then $P(r_{z^*} \wedge r'_z)$ is not P_*-identifiable if $z^* \neq z$.*

Proof outline: The proof is by counter example. We let $\alpha = r_{z^*} \wedge r'_z$, and construct two causal models M^1 and M^2 that agree on the interventional distribution set P_*, but disagree on $P(\alpha)$. In fact, we only need 2 variables. The two models agree on the following: Z is the parent of R, U_Z, Z and R are binary variables, U_R be a ternary variable, $f_Z = U_Z$, and $P(u_Z)$, and $P(u_R)$ are uniform. The two models only differ on the functions f_R, which are given by table 4. It's easy to verify our claim holds for the two models for any values $z^* \neq z$ of Z. □

The next theorem shows how a particular path-specific effect leads to problematic counterfactuals from the previous lemma.

Theorem 3 *The g-specific effect of Z on Y as described in Fig. 5 (a) is not P_*-identifiable.*

Proof: We extend models M^1 and M^2 from the previous proof with additional variables V, Y, and U_Y. We assume $P(u_Y)$ is uniform, and both $P(V, Y|R)$ and the functions which determine V and Y are the same in both models.

Note that since all variables are discrete, the conditional probability distributions can be represented as tables. If we require $|R| = |V|$ and $|Y| = |V| * |R|$, then the conditional

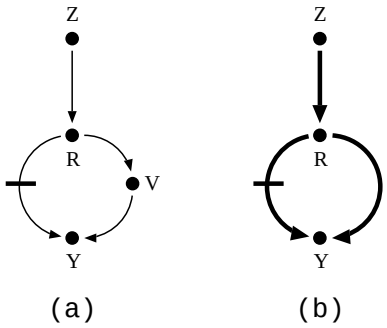

Figure 5: (a) Problematic effect (b) The kite graph

probabilities are representable as square matrices. We fix the functions f_V and f_Y, as well as the exogenous parents of V and Y such that the matrices corresponding to $P(V, Y|R)$ and $P(V|R)$ are matrices are invertible.

Call the extended models M^3 and M^4. Note that by construction, the two models are Markovian. Since M^1 and M^2 have the same P_*, and since the two extended models agree on all functions and distributions not in M^1 and M^2, they must also have the same P_*.

Consider the g-specific effect shown in Fig. 5 (a). From Theorem 1 we can express the path-specific effect in M^3_g in terms of M^3, In particular:

$$\begin{aligned} P(y_z)_{M^3_g} &= \sum_{rv} P(y_{rv} \wedge r_{z^*} \wedge v_z)_{M^3} \\ &= \sum_{r,v,r'} P(y_{rv} \wedge r_{z^*} \wedge v_{r'} \wedge r'_z)_{M^3} \\ &= \sum_{r,v,r'} P(y_{rv})_{M^3} P(v_{r'})_{M^3} P(r_{z^*}, r'_z)_{M^3} \end{aligned}$$

The last step is licensed by the independence assumptions encoded in the parallel worlds model of $y_{rv} \wedge r_{z^*} \wedge v_{r'} \wedge r'_z$. The same expression can be derived for $P(y_z)_{M^4_g}$. Note that since P_* is the same for both models they have the same values for the interventional distributions $P(y_{rv})$ and $P(v_{r'})$. Note that since $P(Y|R, V)$ and $P(V|R)$ are square matrices, the summing out of $P(Y|R, V)$ and $P(V|R)$ can be viewed as a $linear\ transformation$. Since the matrices are invertible, the transformations are one to one, and so if their composition. Since $P(y_{rv}) = P(y|r, v)$ and $P(v_{r'}) = P(v|r')$, and since $P(r_{z^*} \wedge r'_z)$ is different in the two models, we obtain that $P(y_z)_{M^3_g} \neq P(y_z)_{M^4_g}$. Since adding directed or bidirected edges to a graph cannot help identifiability, the result also holds in semi-Markovian models. □

5 Main Result

The main result of this section is that a simple sufficient and necessary (in Markovian models) graphical criterion exists. This condition is easily stated and can be derived from the effect subgraph g in linear time. By contrast, the only other methods known to us for obtaining identifiability results of probabilities of general counterfactual logic formulas are proof search procedures based on results in [Galles and Pearl, 1998], [Halpern, 2000]. Such procedures are far less intuitive, do not have running time bounds, and cannot be used to obtain non-identifiability proofs.

First let's define this criterion:

Definition 10 (Recanting witness criterion) *Let $R \neq Z$ be a node in G, such that there exists a directed path in g from Z to R, a directed path from R to Y in g, and a direct path from R to Y in G but not g. Then Z, Y, and g satisfy the recanting witness criterion with R as a witness*

The recanting witness criterion is illustrated graphically as the 'kite pattern' in Fig. 5 (b). The name 'recanting witness' comes from the behavior of the variable R in the center of the 'kite.' This variable, in some sense, 'tries to have it both ways.' Along one path from R to Y, R behaves as if the variable Z was set to one value, but along another path, R behaves as if Z was set to another value. This 'changing of the story' of R is what causes the problem, and as we will show it essentially leads to the the existence of a non P_*-identifiable expression of the type discussed in section 4.

To proceed, we must make use of the following helpful lemmas: Let g be an effect subgraph of G and g^* the fixed point of R_1 and R_2. Let $\overline{g^*} = G \setminus g^*$.

Lemma 3 *g^* satisfies the recanting witness criterion iff g does. Moreover, if g^* does satisfy the criterion, then there exists a witness R s.t $out(R) \cap \overline{g^*} \neq \emptyset$. If g^* does not, then $\overline{g^*} \subseteq out(Z)$.*

Lemma 3 states that repeated applications of rules R_1 and R_2 preserves the satisfaction of the recanting witness criterion. Moreover, if the witness exists in the fixed point g^*, then some outgoing edge from it is blocked. If the witness does not exist in g^*, then only root-emanating edges are blocked.

Lemma 4 *Assume the g^*-specific effect of Z on Y is P_*-identifiable. Let $\boldsymbol{E}$ be any set of edges in $\overline{g^*}$. Let $g' = \boldsymbol{E} \cup g^*$. Then the g'-specific effect of Z on Y is P_*-identifiable.*

Lemma 4 states that if a path specific effect is not identified, then adding blocked directed edges 'does not help,' in that the effect remains unidentified. Now we can state and prove the main results:

Theorem 4 *If g satisfies the recanting witness criterion, then the g-specific effect of Z on Y is not P_*-identifiable.*

Proof: Let M be our model and assume that g satisfies the recanting witness criterion. By Lemma 3 so does g^*, let R be the witness from the lemma s.t $e = R \rightarrow V$ is in $\overline{g^*}$. Assume the g-specific effect is identifiable, By Theorem 2 so is the g^*-specific effect. Let g' be the path specific effect obtained by adding all edges to g^*, but e. By Lemma 4 the g'-specific effect is also P_*-identifiable. Now by composing the functions in g' we can obtain a new model M' which is exactly the model of Fig. 5 (a)[2] and $P(y_z)_{M_{g'}} = P(y_z)_{M'_{g'}}$. From Theorem 3 we know that $P(y_z)_{M'_{g'}}$ is not P_*-identifiable, hence, neither is $P(y_z)_{M_{g'}}$ and the g'-specific effect is not P_*-identifiable. Contradiction. □ To illustrate the use of

[2] or a similar model where we "cut" the edge $R \rightarrow V$ and not the edge $R \rightarrow Y$

the theorem, consider the example in Eq. (4) from Section 3. The expression $\sum_b P(s_{a,b} \wedge b_{a^*}) =$

$$\begin{aligned} &= \sum_{b,p} P(s_{a,b} \wedge b_{p'} \wedge p'_{a^*}) \\ &= \sum_{b,p,p'} P(s_{a,b,p} \wedge b_{p'} \wedge p'_{a^*} \wedge p_a) \\ &= \sum_{b,p,p'} P(s_{a,b,p} \wedge b_{p'}) P(p'_{a^*} \wedge p_a) \end{aligned} \quad (5)$$

The first two steps are by definition, the last step is licensed by the parallel worlds model corresponding to the formula in Eq. 5. The theorem shows that, as in this example, non-identifiability arises because formulas of the form $p'_{a^*} \wedge p_a$ appear whenever the recanting witness criterion holds.

Theorem 5 *If g does not satisfy the recanting witness criterion, then the g-specific effect of Z on Y is P_*-identifiable in Markovian models.*

Proof: From theorem 2 we have that $P(y_z)_{M_{g^*}} = P(y_z)_{M_g}$. Since g does not satisfy the recanting witness criterion, by Lemma 3 all the edges in $\overline{g^*}$ emanate from Z. From Theorem 1 there is a formula $\alpha(g^*)$ corresponding to $P(y_z)_{M_{g^*}}$ that contains only atomic counterfactuals of the form $v^i_{pa^i}$. Since all blocked edges emanate from Z, it can be easily observed that for each two atomic counterfactuals in $\alpha(g^*)$, $v^i_{pa^i}, v^j_{pa^j}$, $i \neq j$. This follows, since we only introduce atomic counterfactuals with $do(z^*)$ where we cut edges. Now since in Markovian models any two different variables are independent if you set all their parents, all the atomic counterfactual in $\alpha(g^*)$ are independent of each other which makes the expression P_*-identifiable. □

For example, we stated earlier that the g specific effect of Fig 3 (b) is identifiable, this is true since g does not satisfy the recanting witness criterion. In particular the expression for the path-specific effect is:

$$\begin{aligned} P(s_a)_{M_{g3b}} &= \sum_{k,b,p,h} P(s_{k,b,p,a} \wedge k_h \wedge b_a \wedge p_a \wedge h_{a^*}) \\ &= \sum_h P(s_{h,a} \wedge h_{a^*}) \\ &= \sum_h P(s_{h,a}) P(h_{a^*}) \end{aligned} \quad (6)$$

As before, the first two steps are by definition, and the last step is licensed by the parallel worlds model corresponding to the formula in Eq. 6. But now note that $P(s_{h,a}), P(h_{a^*}) \in P_*$, therefore the above expression can be computed from experiments.

6 Conclusions

Our paper presented a sufficient and necessary graphical conditions for the experimental identifiability of path-specific effects, using tools from probability theory, graph theory, and counterfactual logic. We related identifiable path-specific effects to direct and indirect effects by showing that all such effects only block root-emanating edges.

While it is possible to give a sufficient condition for identifiability of general counterfactual formulas in our language, using induction on formula structure, this does not give a single necessary and sufficient condition for semi-Markovian models. The search for such a condition is a good direction for future work.

Another interesting direction is to consider special cases of causal models where path-specific effects can be identified even in the presence of the 'kite' – this is true in linear models, for instance.

Finally, our result assumes causal models with finite domains, and 'small' graphs. An interesting generalization is to consider causal models with 'large' or infinite graphs and infinite domains. Such models may require adding first-order features to the language.

7 Acknowledgements

The authors would like to thank Brian Gaeke and Paul Twohey for proofreading earlier versions of this paper.

References

[Balke and Pearl, 1994] Alexander Balke and Judea Pearl. Counterfactual probabilities: Computational methods, bounds and applications. In *Proceedings of UAI-94*, pages 46–54, 1994.

[Galles and Pearl, 1995] David Galles and Judea Pearl. Testing identifiability of causal effects. In *Proceedings of UAI-95*, pages 185–195, 1995.

[Galles and Pearl, 1998] David Galles and Judea Pearl. An axiomatic characterization of causal counterfactuals. *Foundation of Science*, 3:151–182, 1998.

[Halpern, 1990] Joseph Y. Halpern. An analysis of first-order logics of probability. *Artificial Intelligence*, 46(3):311–350, 1990.

[Halpern, 2000] Joseph Halpern. Axiomatizing causal reasoning. *Journal of A.I. Research*, pages 317–337, 2000.

[Pearl, 2000] Judea Pearl. *Causality: models, reasoning, and inference*. Cambridge University Press, 2000.

[Pearl, 2001] Judea Pearl. Direct and indirect effects. In *Proceedings of UAI-01*, pages 411–420, 2001.

[Robins and Greenland, 1992] James M. Robins and Sander Greenland. Identifiability and exchangeability of direct and indirect effects. *Epidemiology*, 3:143–155, 1992.

[Robins, 1997] James M. Robins. Causal inference from complex longitudinal data. In *Latent Variable Modeling and Applications to Causality*, volume 120, pages 69–117, 1997.

[Rubin, 1974] D. B. Rubin. Estimating causal effects of treatments in randomized and non-randomized studies. *Journal of Educational Psychology*, 66:688–701, 1974.