Counterfactual Graphical Models for Mediation Analysis via Path-Specific Effects

Ilya Shpitser

Department of Epidemiology

Harvard School of Public Health

ishpitse@hsph.harvard.edu

May 2, 2012

Abstract

Potential outcome counterfactuals represent variation in the outcome of interest after a hypothetical treatment or intervention is performed. Causal graphical models are a concise, intuitive way of representing causal assumptions, including independence constraints among such counterfactuals. Much of modern causal inference is concerned with expressing cause effect relationships of interest in counterfactual form, showing how the resulting counterfactuals can be identified (that is expressed in terms of available data, using domain-specific causal assumptions), and subsequently estimated using statistical methods. In this paper we will use causal graphical models to analyze the identification problem of the so-called *path-specific effects*, that is effects of treatment on outcome along certain specified causal paths. Such effects arise in mediation analysis settings where it's important to distinguish direct and indirect effects of treatment. We review existing results on path-specific effects in the fully observable, static treatment setting, and extend them to settings with time-varying treatments, and latent variables.

1 Introduction

Human understanding of the natural world is often phrased in terms of cause effect relationships. Causal statements are a part of everyday speech, as well as legal, scientific, and philosophical vocabulary. Human beings reach an intuitive consensus on the meaning of many causal utterances, and there have been numerous attempts to formalize causality in a way that is faithful to this consensus [25], [7], [23], [5], [16], [12], [9].

The notion central to most of these formalizations is that of an intervention – a kind of idealized randomized experiment imposed on a "set of units," which can represent patients in a medical trial, a culture of cells, and so on. Many scientific questions can be phrased in terms of effects of such experiments. For instance, a standard question of interest in medicine or public health is the causal effect of a particular treatment, such as a drug, on patients. The gold standard for establishing causal effects is the randomized control trial, where patients are randomly divided into two groups, the control group which receives no treatment (or the placebo treatment), and the test group, which receives the actual treatment.

Randomized control trials are often difficult and expensive to perform. Thus trial data is often unavailable, or contains very small sample sizes. In many cases, trials are also unethical, if the effect in question is from a harmful exposure such as asbestos or smoking. An alternative approach which avoids the difficulties with trials is to use observational data to establish causal effects. What makes such an approach feasible is certain causal assumptions, which we will discuss in more detail later, which create the link between natural exposure in the observational data, such as data on habitual smokers who receive the harmful exposure voluntarily, and the hypothetical interventional exposure where people were forced to smoke.

This link between actual observed outcomes and hypothetical interventions is in fact an assumption expressed in a mathematical language of counterfactuals [7],[16] which augments the standard probability notation with notation representing po-

tential outcomes of interventions.

Aside from causal effects, sometimes called *total effects*, it is often of interest to distinguish direct effects of treatment from indirect effects through variables known as mediators. One such example, given by [9], is discrimination. Discrimination laws in the United States generally do not permit hiring decisions to directly depend on gender. However, it may easily happen that gender may influence certain secondary characteristics which make a potential hire more suitable. It is permissible to base hiring decisions on such characteristics, with a possible gender imbalance as a result. The legal question in a discrimination case, then, is whether gender had a *direct effect* on hiring or is all effect of gender mediated by secondary characteristics. [1]

Mediation analysis has a long history in the psychology literature, starting with Woodworth's stimulus-organism-response (S-O-R) model [24]. The important distinction between mediator variables, which serve as an observable manifestation of the causal mechanism through which the causal effect "flows," and the moderator variables, which serve to amplify or dampen the causal effect by (anti) synergistically interacting with the treatment has been made in [2]. Multiple examples of mediation analysis studies in psychology, agriculture, epidemiology, and other fields are given in [6].

Questions of this type can be phrased formally in the mathematical language of counterfactuals as a special type of causal effect called the *path-specific effect*, which operates only along certain "causal pathways." An additional difficulty with such effects is that they turn out to refer to multiple hypothetical worlds simultaneously, which makes it impossible to conduct randomized trials to establish such effects without additional assumptions. In this paper, we concentrate on identification of these path-specific effects, that is characterizing causal assumptions necessary to express such effects as a function of data obtainable from randomized trials.

The paper is organized as follows. In section 2, we will discuss the formal preliminaries of causal inference based on counterfactuals. In particular, we will discuss

[1] Naturally, it is not possible to intervene on gender. However it is possible to change the gender field on a job application, or use similar proxies for an actual intervention.

the potential outcome counterfactual notation, and introduce graphical models as a convenient tool for expressing causal assumptions in an intuitive form. We will also discuss causal effects, path-specific effects, and formally pose the identification problem. In section 3, we will give an overview of existing results on path-specific effects. In section 4, we present new results on path-specific effects in the presence of latent variables, and time-varying exposures. Section 5 will contain the discussion and our conclusions.

2 Graphs, Causal Effects and Counterfactuals

In the remainder of the paper, we will discuss vertices of graphs and random variables. Our notation convention will be as follows. A vertex in a graph $\mathcal{G}$ will be denoted by lower case letters: w. Sets of vertices will be denoted by upper case letters: W. A random variable corresponding to a vertex w will be denoted X_w, or sometimes subscripted: $X_1, \ldots X_n$. A set of random variables corresponding to a set of vertices W will be denoted by X_W. A value assignment to a random variable X_w will be denoted by x_w. A value assignment to a set of random variables X_W will be denoted by x_W.

We will represent settings of interest by a set of random variables $X_{v_1}, \ldots, X_{v_n}$, with a joint probability distribution $p(x_{v_1}, \ldots, x_{v_n})$ over these variables. This distribution represents "observational ground truth." In practice, we estimate this distribution from data sets obtained from observational studies.

2.1 Statistical Graphs, Causal Graphs, and Interventions

A major difficulty with probabilistic reasoning in general is that the space requirements needed to store (and numbers of parameters that represent) a joint distribution $p(x_{v_1}, \ldots, x_{v_n})$ grows exponentially with n. This difficulty is sometimes called *the curse of dimensionality.*

A popular approach to address the curse is to systematically exploit condi-

tional independence constraints in the joint distribution. These constraints can be exploited by utilizing statistical graphical models, sometimes known as Bayesian networks [8].

To discuss Bayesian networks, we need to introduce some graph theoretic terminology. A directed graph is a graph containing vertices (or nodes) and directed arrows connecting pairs of vertices. If vertices w, y in a graph $\mathcal{G}$ are connected by a directed edge $w \to y$, we say w is a parent of y and y is a child of w. A sequence of nodes such that every kth and $k+1$th node in the sequence are connected by an edge, and no node occurs more than once in a sequence is called a *path*. If vertices w, y are connected by a path of the form $w \to \ldots \to y$, then we say w is an ancestor of y and y is a descendant of w. A directed graph is acyclic if no node is its own ancestor. We abbreviate directed acyclic graphs as DAGs. For a given node w in $\mathcal{G}$, we denote its sets of parents, children, ancestors and descendants as $Pa_{\mathcal{G}}(w)$, $Ch_{\mathcal{G}}(w)$, $An_{\mathcal{G}}(w)$, $De_{\mathcal{G}}(w)$, respectively. The "genealogic relations" on sets of vertices are defined by taking unions, for instance for a set W, $An_{\mathcal{G}}(W) = \bigcup_{w_i \in W} An_{\mathcal{G}}(w_i)$.

A Bayesian network is a DAG $\mathcal{G}$ which contains n nodes, $\{v_1, \ldots v_n\} = V$, and a set of random variables $X_{v_1}, \ldots X_{v_n}$, one for each vertex in $\mathcal{G}$, forming a joint probability distribution $p(x_{v_1}, \ldots, x_{v_n})$ with a certain property linking the distribution and the graph. This property is called the Markov factorization property:

$$p(x_{v_1}, \ldots, x_{v_n}) = \prod_{i=1}^{n} p(x_{v_i} \mid x_{Pa_{\mathcal{G}}(v_i)})$$

This factorization is equivalent to the local Markov property which states that each X_{v_i} is independent of $X_{V \setminus (De_{\mathcal{G}}(v_i) \cup Pa_{\mathcal{G}}(v_i))}$ conditional on $X_{Pa_{\mathcal{G}}(v_i)}$, and in turn equivalent to the global Markov property defined by d-separation [8], which states, for any disjoint sets of vertices W, Y, Z in $\mathcal{G}$, that if all paths of a certain type from nodes in W to nodes in Y are "blocked" by nodes in Z in $\mathcal{G}$, then X_W is independent of X_Y given X_Z in $p(x_{v_1}, \ldots, x_{v_n})$. A Bayesian network is thus a statistical model in that its Markov properties define a set of probability distributions.

A statistical graph model can further be considered a causal model if we can meaningfully talk about interventions on variables. An intervention on A, denoted by $\text{do}(X_A = x_A)$ (which we will shorten to $\text{do}(x_A)$), is an operation that sets the variables X_A to values x_A, regardless of the usual behavior of X_A given by the observable joint distribution p. Effects of such interventions on other variables in the system will represent causal effects. A statistical model defined by a DAG $\mathcal{G}$ is causal if for every such $\text{do}(x_A)$,

$$p(x_{V \setminus A} \mid \text{do}(x_A)) = \prod_{v_i \notin A} p(x_{v_i} \mid x_{Pa_{\mathcal{G}}(v_i)})$$

Informally, this formula asserts that whenever we intervene on a set of variables X_A, we remove from the Markov factorization all terms $p(x_a \mid x_{Pa_{\mathcal{G}}(a)})$, for all $a \in A$. This is known as the truncation formula [20],[9], or the g-formula [15]. This formula implies, in particular, that for any X_a, $p(x_a \mid do(x_{Pa_{\mathcal{G}}(a)})) = p(x_a \mid x_{Pa_{\mathcal{G}}(a)})$.

The intuition for the g-formula is that in a causal model the parents of every variable are that variable's *direct causes*. These direct causes determine with what probability a variable assumes its values in the model. By intervening on a variable, we force it to attain a particular value, independently of the usual influence of direct causes. For this reason, we "drop out" the term which links the direct causes and the variable from the Markov factorization.

2.2 Counterfactuals, and Path-specific Effects

It is often useful to have a notation for individual variables after a particular intervention was performed. We denote a random variable X_y in a causal model after $do(x_A)$ has been performed by the notation $X_y(x_A)$. Such a variable is called a potential outcome, or a counterfactual variable.

An assumption very commonly made in causal inference is the so called consistency assumption, which states that if we observed variables X_A attain a value x_A, then for any X_Y, the variable sets X_Y and $X_Y(x_A)$ are the same. This assumption

is crucial in that it allows us to link outcomes under hypothetical interventions with outcomes seen in observational studies where no interventions were in fact performed.

Often of interest is the causal effect of $do(x_A)$ on a particular outcome X_Y, which we encode as an *interventional distribution* $p(x_Y \mid do(x_A))$, which can be computed in causal DAG models via the g-formula above. However, in many situations, it is desirable to distinguish direct and indirect effects of a treatment variable X_a on an outcome of interest X_y. Such effects can be encoded, independently of the parametric form chosen for $p(X_{v_1}, \ldots, X_{v_n})$, using the so called pure or natural effects [13], [10]. These effects are defined as follows. First, we choose for the treatment variable X_a two value levels, the reference value x_a^*, and the treatment value x_a. Then, we consider the distribution of the outcome X_y given that X_a was intervened on x_a while the variables $X_{Pa_{\mathcal{G}}(y)\setminus\{a\}}$ were intervened to take whatever value they would have attained had X_a been intervened to the reference value x_a^*.

A short hand notation for this counterfactual is $X_y(x_a, X_Z(x_a^*))$, where $Z = Pa_{\mathcal{G}}(y) \setminus \{a\}$. This notation stands for the counterfactual distribution

$$\sum_{x_Z} p(X_Y(x_a, x_Z), X_Z(x_a^*) = x_Z)$$

Note that this is a marginal obtained from a joint spanning two conflicting hypothetical worlds. In one of the worlds, X_a was intervened on to the value x_a, and in the other world, X_a was intervened on to the value x_a^*. Without additional assumptions, there is no way to estimate this joint distribution even with randomized trials, since it is not usually possible to simultaneously administer two different treatment levels to the same group of patients.

One common assumption is to assume independence of the counterfactual variable $X_y(x_a, x_Z)$ and a set of counterfactual variables $X_Z(x_a^*)$. If this assumption is true, then $X_y(x_a, X_Z(x_a^*)) = \sum_{X_Z = x_Z} p(X_Y(x_a, x_Z))p(X_Z(x_a^*) = x_Z)$. This formula contains two terms, both of which can be obtained from running randomized

trials. In addition, if running trials is not feasible, we can use the g-formula to estimate both of the interventional distributions involved namely $p(x_Y \mid \mathrm{do}(x_a, x_Z))$, and $p(x_Z \mid \mathrm{do}(x_a^*))$.

A common situation in estimating direct effects is shown in Fig. 1. Here we are interested in the direct effect of X_a on X_y (along the path shown in green), in the presence of some measured common causes of X_a, the mediator X_z and the outcome X_y. Assuming $X_y(x_a, x_z)$ is independent of $X_z(x_a^*)$ conditional on X_c, and applying the g-formula to estimating $X_y(x_a, x_z)$ and $X_z(x_a^*)$ yields $\sum_{x_{c,z}} p(x_y \mid x_{z,c}, x_a)p(x_z \mid x_a^*, x_c)p(x_c)$. If C is absent from the graph, the formula reduces to $\sum_{x_z} p(x_y \mid x_a, x_z)p(x_z \mid x_a^*)$, known as the *mediation formula* [4].

The natural indirect effect is defined similarly, except now the reference value x_a^* influences X_y along the direct path, while the treatment value influences X_y along the indirect path. The resulting counterfactual is $X_y(x_a^*, X_Z(x_a))$, which can also be estimated via the formulas above, with some value relabeling.

In order to express natural direct and indirect effects in terms of interventional distributions, it was necessary to assume independence of counterfactual variables which lie in different hypothetical worlds. This is not a testable assumption, since no possible experiment we could perform can falsify it. Nevertheless, there is one type of causal model that implies such assumptions in a plausible way. This causal model is the so called *non-parametric structural equation model* (NPSEM) [9]. An NPSEM is a graphical model which consists of a distribution $P(x_{v_1}, \ldots, x_{v_n})$ that factorizes according to a DAG $\mathcal{G}$ such that every intervention can be expressed in terms of the g-formula, and the consistency assumption is true for every counterfactual. In addition, we assume that every observable variable X_{v_i} is causally determined from its direct causes $X_{Pa_{\mathcal{G}}(v_i)}$ (plus possibly a single unobserved cause only of X_{v_i} and no other variable) via some unknown function or causal mechanism.

Because of these functions, an NPSEM can be viewed as a kind of "stochastic circuit" with variables representing wires, and functions representing logic gates that determine the voltage at a particular wire in terms of other wires in the circuit,

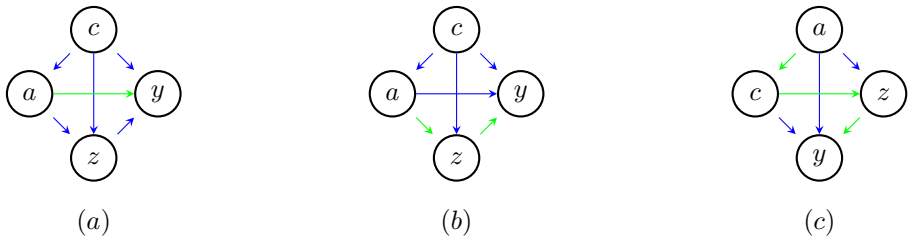

Figure 1: (a) A common case for estimating direct effect of X_a on X_y (path of interest shown in green). (b) The path of interest for the indirect effect of X_a on X_y. (c) A possible path of interest (shown in green) for a path-specific effect of X_a on X_y.

with a few specific wires are allowed to be randomized. On the one hand, the NPSEM may seem quite reasonable, since many data generating processes in Nature can be naturally thought of as such circuits (at least on the level of Newtonian physics). On the other hand, NPSEM implies that for every X_w, X_y, it is the case that $X_w(x_{Pa_{\mathcal{G}}(w)})$ is independent of $X_y(x_{Pa_{\mathcal{G}}(y)})$, for any value assignments $x_{Pa_{\mathcal{G}}(w)}$,$x_{Pa_{\mathcal{G}}(y)}$, even if $Pa_{\mathcal{G}}(w)$ and $Pa_{\mathcal{G}}(y)$ have nodes in common, and these nodes are set to conflicting values by $x_{Pa_{\mathcal{G}}(w)}$ and $x_{Pa_{\mathcal{G}}(y)}$. For this reason, NPSEMs are sometimes considered too strong [14]. In the remainder of this paper, we will assume our graphs represent NPSEMs, with a warning that without assuming such strong models directly, or at least certain cross-world independences such models imply, none of the identification results presented in this paper are valid.

Natural direct and indirect effects can be generalized by considering the effect of X_a on X_y along a specified subset of valid causal paths from a to y, rather than just the direct path from a to y, as in the direct effect case, or all paths but the direct path as in the indirect effect case. Effects along arbitrary specified causal paths are known as path-specific effects [10], [1]. Just as with direct effects, we will consider two values of the treatment variable X_a, the treatment value x_a, and the reference value x_a^*. We will show path-specific effects along a given set of paths π from a to y by showing every arrow along some path in π in green, and all other arrows as blue, for example see Fig. 1 (c). Note that a particular arrow may be

a part of two separate paths, one in π and one not in π (such an arrow would be green by our convention). For example, the arrow from z to y in Fig. 1 (c) is a part of the path $a \to c \to z \to y$ which we are interested in, and a path $a \to z \to y$ which we are not interested in.

We can translate the path of interest in a particular path-specific effect into counterfactual form. Assume we are interested in the path-specific effect of X_a on X_y along the path π specified in green in Fig. 1 (c). The resulting path-specific effect will be defined inductively, and will be denoted (with a slight abuse of notation) as $X_y(\pi(x_a), x_a^*)$. This path-specific effect can be thought of as the random variable X_y under the regime where X_a assumes value x_a for the purposes of the path bundle π, and the value x_a^* otherwise.

Let $V^* = An_{\mathcal{G}_{\not a}}(y)$, where $\mathcal{G}_{\not a}$ is a subgraph of $\mathcal{G}$ containing all vertices other than a, and all edges between these vertices which occur in $\mathcal{G}$. We will divide all nodes in V^* into three sets. All nodes $v \in Ch_{\mathcal{G}}(a)$, such that a, v are adjacent nodes on some path in π (in other words, the arrow $a \to v$ exists and is drawn green), will be denoted by the set $C_{\pi,a,y}$. All nodes $v \in Ch_{\mathcal{G}}(a)$, such that a, v are not adjacent nodes in any path in π (in other words, the arrow $a \to v$ exists and is drawn blue), will be denoted by the set $D_{\pi,a,y}$. Finally, all nodes $v \notin Ch_{\mathcal{G}}(a)$, will be denoted by the set $E_{\pi,a,y}$.

Fix a node s in V^*, and let $Pa_s = Pa_{\mathcal{G}}(s)$. Let B be the set of nodes $t \in Pa_s$ such that the arrow $t \to s$ is green. For each such t, let $X_t(\pi(x_a), x_a^*)$ be the inductively defined path-specific effect of X_a on X_t along π. Then the path-specific effect of X_a on X_s along π is defined as $X_s(X_B(\pi(x_a), x_a^*), X_{Pa_s \setminus (B \cup \{a\})}(x_a^*), x_a)$ if $s \in C_{\pi,a,y}$, $X_s(X_B(\pi(x_a), x_a^*), X_{Pa_s \setminus (B \cup \{a\})}(x_a^*), x_a^*)$ if $s \in D_{\pi,a,y}$, and
$X_s(X_B(\pi(x_a), x_a^*), X_{Pa_s \setminus (B \cup \{a\})}(x_a^*))$ if $s \in E_{\pi,a,y}$.

Applying this definition to the path shown in Fig. 1 (c), and "unrolling" the result, yields the following

$$X_y(\pi(x_a), x_a^*) = \sum_{x_c, x_c', x_z} p(X_y(x_c, x_z, x_a), X_z(x_a^*, x_c') = x_z, X_c(x_a) = x_c, X_c(x_a^*) = x_c')$$

Just as with natural direct and indirect effects, the resulting joint distribution refers to counterfactuals across multiple hypothetical worlds, which makes identifying such distributions from randomized trials difficult. In subsequent sections, we will characterize which NPSEMs imply the assumptions needed to identify such distributions.

2.3 Latent Variables, Latent Projections, and the Effect Identification Problem

Graphical causal models discussed so far assumed full observability, namely that if two variables of interest were observed, then any common causes of these variables were also observed. In practice, this assumption is too restrictive. In particular, in medical trials many common causes of treatments and outcomes are unrecorded.

It's possible to apply DAG model machinery to the latent variable case directly by simply labeling latent variables in a DAG model, and making parametric assumptions about those variable for modeling and inference. The difficulty with that approach is that multiple possible latent structures may lead to the same pattern of observable constraints, and parametric assumptions on the latents may be incorrect, leading to bias.

An alternative approach is to represent sets of possible DAG models with latents by a single graph called the *latent projection*. Such a graph contains two types of arrows, a directed arrow and a bidirected arrow. Given a DAG $\mathcal{G}$ with vertex set V, we attain a latent projection onto $S \subseteq V$, denoted $\mathcal{G}(S)$ as follows. $\mathcal{G}(S)$ is graph with vertex set S. Further, we connect any two vertices $w, y \in S$ by a directed arrow if there exists a directed path from a corresponding vertex w in $\mathcal{G}$ to a corresponding vertex y in $\mathcal{G}$ such that all intermediate nodes on the path are not

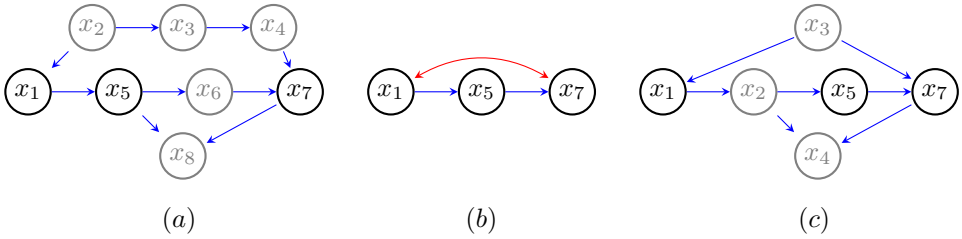

Figure 2: (a) A DAG with latent nodes marked in gray. (b) A latent projection of this DAG. Note that x_8 does not create a new path. (c) A different DAG resulting in the same latent projection.

in S (e.g. are latent). Similarly, we connect any two vertices w and y in $\mathcal{G}(S)$ by a bidirected arrow if there is a path in $\mathcal{G}$ with a first arrow pointing to w, the last arrow pointing to y, no intermediate node on the path with converging arrows, and all intermediate nodes are not in S. The resulting latent projection obtained from a DAG is an acyclic directed mixed graph (ADMG).

As an example, Fig. 2 (a) shows a DAG with latent nodes shaded in gray, and Fig. 2 (b) shows the corresponding latent projection. Fig. 2 (c) shows a DAG distinct from that shown in (a) but which results in the same latent projection.

Just as a particular DAG is associated with a set of distributions obeying the global Markov property, a particular latent projection ADMG is associated with a set of distributions obeying another global Markov property, defined by m-separation [11]. This property has a nice feature that if a distribution $p(x_V)$ satisfies the global Markov property (defined by d-separation) with respect to a DAG $\mathcal{G}$ with a vertex set V, and $\mathcal{G}(S)$ is a latent projection onto a set $S \subseteq V$, then the marginal distribution of $p(x_S)$ satisfies the global Markov property (defined by m-separation) with respect to $\mathcal{G}(S)$. In particular all conditional independences over the variable set X_S advertised by $\mathcal{G}$ are also advertised by $\mathcal{G}(S)$. In fact, certain additional independence constraints are preserved by latent projections, although the full treatment of this topic is beyond the scope of this paper.

Identification problems for causal effects and path-specific effects can be posed in latent projections. Unlike the fully observable DAG case, where every causal

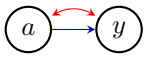

Figure 3: A latent projection where $p(X_y \mid \mathrm{do}(x_a))$ is not identifiable.

effect is identifiable by the g-formula, some causal effects are not identifiable in latent projections. Fig. 3 shows the simplest graph where $p(X_y \mid \mathrm{do}(x_a))$ is not identifiable.

The causal effects which are identifiable are characterized in [3], [18], [17]. The algorithms for their identification were given in [22], [17], [21], [19]. The algorithm works by a recursive application of the g-formula and marginalization operations. In the subsequent sections, we will consider the problem of identification of path-specific effects in latent projections.

3 Known Results on Path-Specific Effects

The identification of path-specific effects in DAG models was considered in [1]. In this section we summarize the results in that paper. The problem is to express a counterfactual distribution corresponding to the effect of X_a on X_y along a specified set of causal paths π in terms of either observed or interventional distributions. We will continue our convention that arrows on causal paths of interest are shown in green, and other arrows are shown in blue.

The main result is stated in terms of the following graphical criterion:

Definition 1 (recanting witness) *Let $\mathcal{G}$ be a DAG, a, y nodes in $\mathcal{G}$, and π a subset of directed paths from a to y in $\mathcal{G}$. Then a node w in $\mathcal{G}$ is called a* recanting witness *for the π-specific effect of X_a on Y_a if there exists a directed path in π containing w, and there exists a directed path from w to y which is not a subpath of any path in π.*

The presence of the recanting witness prevents identification of path-specific effects in terms of interventional (and thus observational) data:

Theorem 2 *Let $\mathcal{G}$ be a DAG, a, y nodes in $\mathcal{G}$, and π a subset of directed paths from a to y in $\mathcal{G}$. Then the π-specific effect of X_a on X_y is identifiable if and only if there does not exist a recanting witness for this effect.*

Note that according to this criterion, the path-specific effect corresponding to the paths shown in Fig. 1 (c) is not identifiable. If a path-specific effect of X_a on X_y is identifiable, it can be expressed via the *path-specific g-formula* as follows:

$$\begin{aligned}
\sum_{x_{V^*\setminus\{y\}}} \Big(\prod_{v\in C_{\pi,a,y}} p(X_v = x_v \mid \mathrm{do}(X_{M_v} = x_{M_v}, X_a = x_a)) \Big) \cdot \\
\Big(\prod_{v\in D_{\pi,a,y}} p(X_v = x_v \mid \mathrm{do}(X_{M_v} = x_{M_v}, X_a = x_a^*)) \Big) \cdot \\
\Big(\prod_{v\in E_{\pi,a,y}} p(X_v = x_v \mid \mathrm{do}(X_{M_v} = x_{M_v})) \Big) = \\
\sum_{x_{V^*\setminus\{y\}}} \Big(\prod_{v\in C_{\pi,a,y}} p(X_v = x_v \mid X_{M_v} = x_{M_v}, X_a = x_a) \Big) \cdot \\
\Big(\prod_{v\in D_{\pi,a,y}} p(X_v = x_v \mid X_{M_v} = x_{M_v}, X_a = x_a^*) \Big) \cdot \\
\Big(\prod_{v\in E_{\pi,a,y}} p(X_v = x_v \mid X_{M_v} = x_{M_v}) \Big)
\end{aligned}$$

where $V^* = An_{\mathcal{G}_{\bar{x}}}(y)$, for every node $v \in V^*$, $M_v = Pa_{\mathcal{G}}(v) \setminus \{a\}$, and $x_{V^*\setminus\{y\}}$ is consistent with $\bigcup_{v\in V^*}(x_{M_v} \cup x_v)$ and ranges over all possible assignments in the summation, and $C_{\pi,a,y}$, $D_{\pi,a,y}$ and $E_{\pi,a,y}$ are defined as in the previous section. This formula is a generalization of Pearl's mediation formula to identifiable path specific effects with a single treatment and single outcome in NPSEM models represented by DAGs.

The recanting witness criterion prevents identification in many seemingly reasonable cases. Consider the graph in Fig. 4. In this graph, we are interested in the effect of X_a on X_y along green colored causal paths, that is all causal paths not mediated by z. The reason this case is "seemingly reasonable" is that confounders

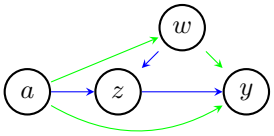

Figure 4: A graph where the natural direct effect of X_a on X_y (along the green paths) is not identifiable due to the recanting witness criterion (w is the witness).

(observable or otherwise) between the mediator z and the outcome y such as w are very common, and it is very common for such confounders to be affected by the treatment. Unfortunately, due to the recanting witness criterion, the presence of such confounders results in an effect that is not in general identifiable. This holds even if all observable nodes are binary.

4 General Path-Specific Effects

In this section we consider a general path-specific effects with multiple treatments and multiple outcomes in models represented by latent projection ADMGs.

We first generalize path-specific effects appropriately to the case of multiple treatments $\{a_1, \dots a_k\} = A$ and multiple outcomes $\{y_1, \dots y_m\} = Y$. Just as before, we will consider two sets of values for A, the treatment values $\{x_{a_1}, \dots x_{a_k}\} = x_A$ and reference values $\{x^*_{a_1}, \dots x^*_{a_k}\} = x^*_A$. We will be interested in effect of the set X_A on the set X_Y along a set π of directed paths from nodes in A to nodes in Y. As before, our convention will be that if an arrow is a part of some path in π, it is shown in green, otherwise it is shown in blue.

We now show how to translate path-specific effects along a given set of paths into counterfactual form. The resulting counterfactual, denoted by $X_Y(\pi(x_A), x^*_A)$, will be defined inductively.

Let $V^* = An(Y)_{\mathcal{G}_{\not A}}$, where $\mathcal{G}_{\not A}$ is a subgraph of $\mathcal{G}$ containing all vertices other than A, and all edges between these vertices which occur in $\mathcal{G}$. As before, we will partition all nodes in V^* into three sets. All nodes $v \in Ch_{\mathcal{G}}(a)$ for any $a \in A$,

such that a, v are adjacent nodes on some path in π (in other words, the arrow $a \to v$ exists and is drawn green), will be denoted by the set $C_{\pi,A,Y}$. All nodes $v \in Ch_{\mathcal{G}}(a)$ for any $a \in A$, such that a, v are not adjacent nodes in any path in π (in other words, the arrow $a \to v$ exists and is drawn blue), will be denoted by the set $D_{\pi,A,Y}$. Finally, all nodes $v \notin Ch_{\mathcal{G}}(A)$, will be denoted by the set $E_{\pi,A,Y}$.

Fix a node s in V^*, and let $Pa_s = Pa_{\mathcal{G}}(s)$. Let B be the set of nodes $t \in Pa_s$ such that the arrow $t \to s$ is green. For each such t, let $X_t(\pi(x_A), x^*_A)$ be the inductively defined path-specific effect of X_A on X_t along π. Then the path-specific effect of X_A on X_s along π is defined as $X_s(X_B(\pi(x_A), x^*_A), X_{Pa_s \setminus (B \cup A)}(x^*_A), x_A)$ if $s \in C_{\pi,A,Y}$, $X_s(X_B(\pi(x_A), x^*_A), X_{Pa_s \setminus (B \cup A)}(x^*_A), x^*_A)$ if $s \in D_{\pi,A,Y}$, and $X_s(X_B(\pi(x_A), x^*_A), X_{Pa_s \setminus (B \cup A)}(x^*_A))$ if $s \in E_{\pi,A,Y}$. The path-specific effect $X_Y(\pi(x_A), x^*_A)$ is defined to be the joint distribution over $X_{y_1}(\pi(x_A), x^*_A), \ldots, X_{y_m}(\pi(x_A), x^*_A)$.

We first consider identification of such path-specific effects in the fully observable case.

Definition 3 (generalized recanting witness) *Let $\mathcal{G}$ be a DAG, A, Y sets of nodes in $\mathcal{G}$, and π a subset of directed paths which start with a node in A and end in a node in Y in $\mathcal{G}$. Then a node w in $\mathcal{G}$ is called a* generalized recanting witness *for the π-specific effect of X_A on X_Y if there exists a directed path in π containing w that does not intersect A except at the original endpoint, and there exists a directed path from w to an element in Y which is not a subpath of any path in π, and which similarly does not intersect A.*

The presence of the generalized recanting witness prevents identification of path-specific effects with multiple treatments and multiple outcomes:

Theorem 4 *Let $\mathcal{G}$ be a DAG, A, Y sets of nodes nodes in $\mathcal{G}$, and π a subset of directed paths which start with a node in A and end in a node in Y in $\mathcal{G}$. Then the π-specific effect of X_A on X_Y is identifiable if and only if there does not exists a generalized recanting witness for this effect.*

As before, if a path-specific effect of X_A on X_Y is identifiable, it can be expressed via the *path-specific g-formula*:

$$\sum_{x_{V^* \setminus Y}} \Big(\prod_{v \in C_{\pi,A,Y}} p(X_v = x_v \mid \mathrm{do}(X_{M_v} = x_{M_v}, X_{A_v} = x_{A_v})) \Big) \cdot \Big(\prod_{v \in D_{\pi,A,Y}} p(X_v = x_v \mid \mathrm{do}(X_{M_v} = x_{M_v}, X_{A_v} = x^*_{A_v})) \Big) \cdot \Big(\prod_{v \in E_{\pi,A,Y}} p(X_v = x_v \mid \mathrm{do}(X_{M_v} = x_{M_v})) \Big) = \sum_{x_{V^* \setminus \{y\}}} \Big(\prod_{v \in C_{\pi,A,Y}} p(X_v = x_v \mid X_{M_v} = x_{M_v}, X_{A_v} = x_{A_v}) \Big) \cdot \Big(\prod_{v \in D_{\pi,A,Y}} p(X_v = x_v \mid X_{M_v} = x_{M_v}, X_{A_v} = x^*_{A_v}) \Big) \cdot \Big(\prod_{v \in E_{\pi,A,Y}} p(X_v = x_v \mid X_{M_v} = x_{M_v}) \Big)$$

where $V^* = An_{\mathcal{G}_{\not A}}(y)$, for every node $v \in V^*$, $M_v = Pa_{\mathcal{G}}(v) \setminus A$ and $A_v = A \cap Pa_{\mathcal{G}}(v)$, and $x_{V^* \setminus \{y\}}$ is consistent with $\bigcup_{v \in V^*}(x_{M_v} \cup x_v)$ and ranges over all possible assignments in the summation.

We illustrate this criterion with the following example. Consider the sequential treatment setting, where a patient visits a doctor periodically, let us say monthly, and the doctor prescribes treatment based on patient vitals taken during the visit. The situation is shown in Fig. 5 (a). For simplicity we consider a treatment regime which lasts two months, but the example generalizes for treatments of arbitrary length. In this example, the nodes z_i represent patient vitals, the nodes a_i represent treatments administered every month, and the node y is the outcome. We assume observational data, where doctors follow whatever policy they wish in administering treatment, such that the policy on X_{a_i} depends on all patient vitals at the same month and previous months: $X_{z_0}, \ldots X_{z_{i-1}}$. We are interested in the path-specific effect of the treatment regime $do(X_{a_0} = x_{a_0}, \ldots, X_{a_k} = x_{a_k})$ on X_y along all causal paths not through z_i. In our case of treatment that lasts two months, we wish to

Figure 5: (a) A graph representing a sequential treatment regime. We are interested in the path-specific effect of X_{a_1,a_2} on X_y along all paths not through z_1. (b) A similar graph where the path-specific effect of X_{a_1,a_2} on X_y along all paths not through z_1 is not identifiable.

exclude the path $x_1 \rightarrow z_1 \rightarrow y$. According to the generalized recanting witness criterion, this path-specific effect is identifiable, and given by the path-specific g-formula, specifically:

$$\sum_{x_{z_1}, x_{z_0}} p(x_y \mid x_{z_1}, x_{a_2}, x_{z_0}, x_{a_1}) p(x_{z_1} \mid x_{a_1^*}, x_{z_0}) p(x_{z_0})$$

Note that just as in the previous example, the presence of common causes of z_i and y which are influenced by treatments prevents identification, as shown in Fig. 5 (b).

We now extend our results to causal models with latent variables, represented by latent projections. We first introduce some terminology. In an ADMG $\mathcal{G}$ a path from a to y is called bidirected if every edge on this path is bidirected.

Definition 5 (district) *Let $\mathcal{G}$ be an ADMG. Then for any node a, the set of nodes in $\mathcal{G}$ reachable from a by bidirected paths is called the district of a, written $Dis_{\mathcal{G}}(a)$.*

For an ADMG $\mathcal{G}$ with the vertex set V, we denote by $\mathcal{G}_A$ a restriction of $\mathcal{G}$ to $A \subseteq V$, that is $\mathcal{G}_A$ is a graph which contains only vertices A, and only edges in $\mathcal{G}$ which connect vertices in A.

Definition 6 (recanting district criterion) *Let $\mathcal{G}$ be an ADMG, A, Y sets of nodes in $\mathcal{G}$, and π a subset of directed paths which start with a node in A and end in a node in Y in $\mathcal{G}$. Let V^* be the set of nodes not in A which are ancestral of*

Y via a directed path which does not intersect A. Then a district D in an ADMG $\mathcal{G}_{V^*}$ is called a recanting district *for the π-specific effect of X_A on X_Y if there exist nodes $z_i, z_j \in D$ (possibly $z_i = z_j$), $a_i \in A$, and $y_i, y_j \in Y$ (possibly $y_i = y_j$) such that there is a directed path $a_i \to z_i \to \ldots \to y_i$ in π, and a directed path $a_i \to z_j \to \ldots \to y_j$ not in π, such that neither of these paths intersect A except at their origin vertex.*

The existence of a recanting district prevents identification of path-specific effects with multiple treatments and multiple outcomes in models with latent variables, due to the following theorem.

Theorem 7 *Let $\mathcal{G}$ be an ADMG, A, Y sets of nodes nodes in $\mathcal{G}$, and π a subset of directed paths which start with a node in A and end in a node in Y in $\mathcal{G}$. Then the π-specific effect of X_A on X_Y is identifiable if and only if there does not exists a recanting district for this effect.*

This theorem, which clearly subsumes Theorem 4, is proven for NPSEMs in the Appendix. If a path-specific effect of X_A on X_Y is identifiable, it can be expressed via a generalization of the path-specific g-formula for ADMGs.

To give this formula, we first define V^*, as before, as equal to $An_{\mathcal{G}_{\bar{A}}}(y)$. We then partition $\mathcal{D}(\mathcal{G}_{V^*})$ into three sets: $D \in \mathcal{C}_{\pi,A,Y}$ if $A \cap Pa_{\mathcal{G}}(D) \neq \emptyset$ and for every $a \in A$, $d \in D$ such that $a \in Ch_{\mathcal{G}}(d)$, there exists a path in π where a, d are adjacent (in other words, $a \to d$ is green for every such a); $D \in \mathcal{D}_{\pi,A,Y}$ if $A \cap Pa_{\mathcal{G}}(D) \neq \emptyset$ and for every $a \in A$, $d \in D$ such that $a \in Ch_{\mathcal{G}}(d)$, there is no path in π where a, d are adjacent (in other words, $a \to d$ is blue for every such a); $D \in \mathcal{E}_{\pi,A,Y}$ if $A \cap Pa_{\mathcal{G}}(D) = \emptyset$. Note that these three sets only form a partition of $\mathcal{D}(\mathcal{G}_{V^*})$ if the effect of X_A on X_Y along paths π is identifiable. The formula is then:

$$\sum_{x_{V^* \setminus Y}} \Big(\prod_{D \in \mathcal{C}_{\pi,A,Y}} p(X_D = x_D \mid \mathrm{do}(X_{M_D} = x_{M_D}, x_{A_D} = x_{A_D})) \Big) \cdot$$
$$\Big(\prod_{D \in \mathcal{D}_{\pi,A,Y}} p(X_D = x_D \mid \mathrm{do}(X_{M_D} = x_{M_D}, x_{A_D} = x^*_{A_D})) \Big) \cdot$$
$$\Big(\prod_{D \in \mathcal{E}_{\pi,A,Y}} p(X_D = x_D \mid \mathrm{do}(X_{M_D} = x_{M_D})) \Big)$$

where for every $D \in \mathcal{D}(\mathcal{G}_{V^*})$, $M_D = Pa_{\mathcal{G}}(D) \setminus (D \cup A)$, $A_D = A \cap Pa_{\mathcal{G}}(D)$, and and $x_{V^* \setminus Y}$ is consistent with $\bigcup_{D \in \mathcal{D}(\mathcal{G}_{V^*})}(x_{M_D} \cup x_D)$ and ranges over all possible assignments in the summation.

Unlike the cases where we identified path-specific effects in DAGs, the identifying formula is in terms of interventional rather than observational distributions. This is because in DAG models, every interventional distribution is identifiable from the observational distribution, whereas in ADMG model, some interventional distributions are not. However, in some cases it is possible to express every interventional distribution in the above formula in terms of the observational distribution, as in the example shown in Fig. 6 (c).

We now illustrate Theorem 7 with a number of examples, shown in Fig. 6. For clarity, we will show bidirected arrows in red (this is merely to distinguish these arrows from directed arrows). In the graph shown in Fig. 6 (a), the effect of X_a on X_{y_1,y_2} along the green path is identifiable, and equal to $\sum_{x_w} p(X_{y_1,y_2} \mid \mathrm{do}(x_w, x^*_a))p(X_w = x_w \mid \mathrm{do}(x_a))$.

On the other hand, the path-specific effect of X_{a_1,a_2} on X_y in the graph shown on Fig. 6 (b) is not identifiable. This graph is almost identical to one show in Fig. 5 (a), except there is an additional bidirected arrow connecting z_1 and y. This single change is sufficient to prevent identification. This example illustrates that unobserved confounders between mediators and outcome prevent identification.

Finally, in the graph shown in Fig. 6 (c), the effect of X_2 on X_6 along the green

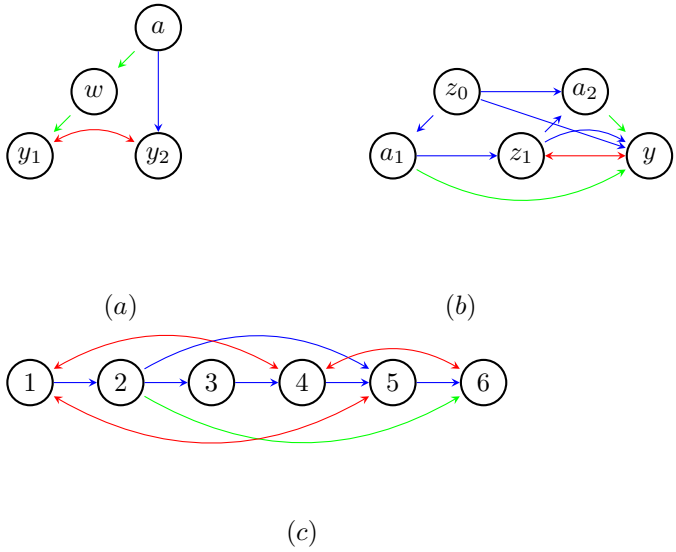

Figure 6: (a) The path-specific effect of X_a on X_{y_1,y_2} is identifiable. (b) The path-specific effect of X_{a_1,a_2} on X_y is not identifiable. (c) The path-specific effect of X_2 on X_6 is identifiable.

arrow is identifiable, and equal to

$$\sum_{x_3,x_4,x_5} p(x_6, x_4 \mid do(x_5, x_3, x_2))p(x_5 \mid do(x_4, x_2^*))p(x_3 \mid do(x_2^*))$$

In fact, each of the interventional distributions are themselves identified in the given graph in terms of observational data: $p(x_3 \mid do(x_2^*)) = p(x_3 \mid x_2^*)$, $p(x_5 \mid do(x_4, x_2^*)) = q^1_{x_2^*,x_3}(x_5 \mid x_4)$, where $q^1_{x_2^*,x_3}(x_5, x_4) = \sum_{x_1} p(x_5, x_4 \mid x_3, x_2^*, x_1)p(x_1)$, and finally $p(x_6, x_4 \mid do(x_5, x_3, x_2)) = q^2_{x_2,x_3}(x_6 \mid x_5, x_4)q^2_{x_2,x_3}(x_4)$, where $q^2_{x_2,x_3}(x_4, x_5, x_6) = \sum_{x_1} p(x_6, x_5, x_4 \mid x_3, x_2, x_1)p(x_1)$.

5 Conclusion

In this paper, we have used the language of causal graphical models and potential outcome counterfactuals to analyze the problem of identifying path-specific effects, that is the effect of a set of hypothetical interventions on a set of outcomes of

interest along certain causal pathways. To obtain our identification results, we relied on a particular type of causal model, the non-parametric structural equation model (NPSEM), which makes strong assumptions on causal mechanisms which entail independences among cross-world counterfactuals.

We have generalized existing results on single treatment single outcome path-specific effects [1] to the case of multiple treatments, multiple outcomes, and partial observability. In particular, our results give a complete characterization of identification of path-specific effects in the time-varying treatment setting. Identifiable path-specific effects are expressed in terms of the *path-specific g-formula*, which is a generalization of the mediation formula [4]. The natural next step is to extend existing statistical estimation techniques for the g-formula functionals arising from identifying total effects to the path-specific g-formula.

6 Appendix

We now give a proof of Theorem 7. We denote by $(X_Y \perp\!\!\!\perp X_W \mid X_Z)_p$ a statement that for sets of random variables X_Y, X_W, X_Z in the joint distribution p, X_Y is independent of X_W conditioned on X_Z. Fix an ADMG $\mathcal{G}$ with vertices V, subsets A, Y of V, and a set π of directed paths from nodes in A to nodes in Y.

6.1 The Soundness Proof

We first show soundness, namely that if a recanting district does not exist, then the path-specific effect is identifiable from interventional distributions via a generalization of the path-specific g-formula. To show this, we must express path-specific effects in nested counterfactual form as this generalized formula.

Let $\{v_1, \ldots v_m\} = V^* = An(Y)_{\mathcal{G}_{\bar{A}}}$. Consider the counterfactual joint distribution $p(X_{y_1}(\pi(x_A), x_A^*), \ldots, X_{y_m}(\pi(x_A), x_A^*))$, representing the path-specific effect of A on Y along paths in π.

"Unrolling" this counterfactual, we get the following formula:

$$\sum_{x_{V^* \setminus Y}} p(X_{v_1}(x_{Pa_{\mathcal{G}_{V^*}}(v_1)}), \ldots X_{v_m}(x_{Pa_{\mathcal{G}_{V^*}}(v_m)})) \tag{1}$$

where each value assignment $x_{Pa_{\mathcal{G}_{V^*}}(v_i)}$ is consistent with $x_{V^* \setminus Y}$, and the values of X_A given by the effect definition (that is if there is a green arrow from $a \in A$ to v_i, then $x_{Pa_{\mathcal{G}_{V^*}}(v_i)}$ assigns to X_a the treatment value x_a rather than the reference value x_a^*).

One of the assumptions that NPSEM DAG models make is that absence of a directed arrow from a to y implies fixing all observable parents of X_y renders the resulting counterfactual $X_y(x_{Pa_{\mathcal{G}}(y)})$ independent of any counterfactual $X_a(.)$, and that fixing X_a will not change $X_y(x_{Pa_{\mathcal{G}}(y)})$.

This in turn implies that in a marginal of a DAG NPSEM represented by an ADMG $\mathcal{G}$, for any two counterfactuals $X_z(x_{Pa_{\mathcal{G}}(z)})$, $X_w(x_{Pa_{\mathcal{G}}(w)})$, if there is no bidirected arrow from z to w in $\mathcal{G}$, then $p(X_z(x_{Pa_{\mathcal{G}}(z)}), X_w(x_{Pa_{\mathcal{G}}(w)})) = p(X_z(x_{Pa_{\mathcal{G}}(z)}))\cdot p(X_w(x_{Pa_{\mathcal{G}}(w)}))$. Further, NPSEMs obey a property called compositionality, which states that for any sets of counterfactual variables $X_A(x_{S_A})$,$X_Y(x_{S_Y})$,$X_Z(x_{S_Z})$,$X_W(x_{S_W})$ if both $(X_A(x_{S_A}) \perp\!\!\!\perp X_Y(x_{S_Y}) \mid X_Z(x_{S_Z}))$ and $(X_W(x_{S_W}) \perp\!\!\!\perp X_Y(x_{S_Y}) \mid X_Z(x_{S_Z}))$ hold, then $(X_A(x_{S_A}) \cup X_W(x_{S_W}) \perp\!\!\!\perp X_Y(x_{S_Y}) \mid X_Z(x_{S_Z}))$ also holds.

These properties imply the that formula 1 is equivalent to the following formula

$$\sum_{x_{V^* \setminus Y}} \prod_{v_1,\ldots,v_k \in D \in \mathcal{D}(\mathcal{G}_{V^*})} p(X_{v_1}(x_{Pa_{\mathcal{G}_{V^*}}(v_1)}), \ldots X_{v_k}(x_{Pa_{\mathcal{G}_{V^*}}(v_k)})) \tag{2}$$

which is a decomposition of formula 1 into a set of terms, one for each district in $\mathcal{G}_{V^*}$.

Finally, since all subscripts not involving value assignments to A are consistent with $v^* \setminus x$, we can use the consistency assumption to conclude formula 2 is equivalent to formula 3.

$$\sum_{x_{V^* \setminus Y}} \prod_{D \in \mathcal{D}(\mathcal{G}_{V^*})} p(X_D = x_D \mid \mathrm{do}(x_{Pa_{\mathcal{G}_{V^*}}(D) \setminus D})) \tag{3}$$

where $x_{Pa_{\mathcal{G}_{V^*}}(D)\setminus D}$ is a value assignment to $Pa(D)_{\mathcal{G}_{V^*}}\setminus D$ consistent with v^* and assignments to X given by the effect. In particular, if $D \in \mathcal{C}_{\pi,A,Y}$, $x_{Pa_{\mathcal{G}_{V^*}}(D)\setminus D}$ assigns to $X_{A\cap(Pa(D)_{\mathcal{G}_{V^*}}\setminus D)}$ the treatment values, if $D \in \mathcal{D}_{\pi,A,Y}$, $x_{Pa_{\mathcal{G}_{V^*}}(D)\setminus D}$ assigns to $X_{A\cap(Pa(D)_{\mathcal{G}_{V^*}}\setminus D)}$ the reference values.

This establishes our result.

6.2 The Completeness Proof

Next, we show completeness, namely that if a recanting district exists, then the path-specific effect given by a counterfactual distribution $p(X_{y_1}(\pi(x_A), x^*_A), \ldots, X_{y_m}(\pi(x_A), x^*_A))$ is not identifiable. The proof will proceed as follows.

We will first show if there exists a recanting district D (for a particular $a \in A$) then the following counterfactual γ_1 is not identifiable from $P_* = \{p(X_{V\setminus W} \mid \text{do}(x_w)) \mid W \subseteq V\}$ in the graph $\mathcal{G}_{D\cup\{a\}}$:

$$\gamma_1 = \sum_{x_{v_i}:v_i\in D\setminus rh(D)_{\mathcal{G}_{D\cup\{a\}}}} p(X_{v_1}(x_{Pa_{\mathcal{G}_{D\cup\{a\}}}(v_1)}), \ldots X_{v_k}(x_{Pa_{\mathcal{G}_{D\cup\{a\}}}(v_k)}) \qquad (4)$$

where $\{v_1, \ldots v_k\} = D$, $rh(D)_{\mathcal{G}_{D\cup\{a\}}} = \{v_i \in D \mid Ch(D)_{\mathcal{G}_{D\cup\{a\}}} \cap D = \emptyset\}$, and $x_{Pa_{\mathcal{G}_{D\cup\{a\}}}(v_i)}$ for every $v^i \in D$, is a value assignment defined as follows.

It's an assignment of values to $Pa(v_i)_{\mathcal{G}_{D\cup\{a\}}}$ that are consistent with x_{v_i} (values being summed) for nodes in $Pa(v_i)_{\mathcal{G}_{D\cup\{a\}}}\setminus\{a\}$. If $a \in Pa_{\mathcal{G}_{D\cup\{X_i\}}(v_i)}$, the assignment assigns to a the treatment value x_a if the arrow from a to v_i is green, and the reference value x^*_a otherwise (note that by assumption there exists both a green arrow from a to a node in D, and a blue arrow from a to a node in D).

After showing the non-identifiability of γ_1, we show the non-identifiability of a related counterfactual γ_2, defined as follows.

Fix $Y' \subseteq Y$, such that for all nodes in $rh(D)_{\mathcal{G}_{D\cup\{a\}}}$ are ancestral of Y' in $\mathcal{G}_{V^*}$, and for no subset of Y' is this true. For every node r in $rh(D)_{\mathcal{G}_{D\cup\{a\}}}$ pick a node $y_r \in Y'$ such that there is a directed path π_r from r to y_r. Let the set of nodes in

every such path be equal to W^*. Let $\mathcal{G}^*$ be a subgraph of $\mathcal{G}_{V^*}$ containing nodes in $D \cup W^*$. We will then show that:

$$\gamma_2 = \sum_{x_{v_i} : v_i \in (D \cup W^*) \setminus Y'} p(X_{v_1}(x_{Pa_{\mathcal{G}^*}(v_1)}), \ldots, X_{v_l}(x_{Pa_{\mathcal{G}^*}(v_l)}) \tag{5}$$

where $x_{Pa_{\mathcal{G}^*}(v_i)}$ is defined as before, is not identifiable from P_* in $\mathcal{G}^*$.

Having shown γ_2 is not identifiable in $\mathcal{G}^*$ from P_*, we then have two models M_1, M_2 which agree on P_* but disagree on γ^*. We then note that augmenting M_1, M_2 with additional variables can result in models M_1', M_2' that induce $\mathcal{G}$, and such that γ_2 is a marginal distribution of the counterfactual γ in these models. This will imply γ is not identifiable from P_* in $\mathcal{G}$, which was what we wanted to show.

Lemma 8 *The counterfactual γ_1 given in equation 4 is not identifiable from P_* in $\mathcal{G}_{D \cup \{a\}}$.*

Proof: Pick two nodes in D, v_1, v_2 such that a has a green arrow to v_1, and a blue arrow to v_2. Assume without loss of generality that a only affects those two nodes in D. Assume, also without loss of generality, that every node in D has at most one child (other arrows are vacuous).

We now construct two NPSEM models, $\mathcal{M}_1$ and $\mathcal{M}_2$, which both agree on P_*, both induce $\mathcal{G}_{D \cup \{a\}}$, but which disagree on γ_1 as defined. In these models, every observable variable is binary. Every bidirected arc corresponds to an unobserved binary variable with two children. In $\mathcal{M}_1$, for every observable node, its value is determined by the bit parity function of its parents (both observed and unobserved). For $\mathcal{M}_2$, for every observable node, its value is determined by the bit parity function of its parents, except the functions determining the values of v_1, v_2 do not take the value of a into account. The distributions over unobserved nodes is the same in both models, and is uniform.

We now show the two models have the desired properties. That both models induce $\mathcal{G}_{D \cup \{a\}}$ is clear. Next, we show $\mathcal{M}_1$ and $\mathcal{M}_2$ agree on P_*.

By construction, both models agree on $p(x_a)$. We next show both models agree on $p(x_D \mid \text{do}(x_a))$. It's not difficult to show (following the proof of Theorem 17 in [19]) that $p(x_D \mid \text{do}(x_a)) = p(x_D)$ is a uniform distribution in $\mathcal{M}_2$ over assignments x_D such that $x_{rh(D)_{\mathcal{G}_{D\cup\{a\}}}}$ has even bit parity. In fact, the same proof shows the same for $p(x_D \mid \text{do}(a))$ in $\mathcal{M}_1$. This implies that $p(x_{D\cup\{a\}}) = p(x_D \mid \text{do}(x_a))p(x_a)$ is the same in $\mathcal{M}_1$ and $\mathcal{M}_2$. Furthermore, for any partition (Z_1, Z_2) of $Z = (D \cup \{a\})$, it is either the case that $Z_1 \subset D$, or $p(x_{Z_1} \mid \text{do}(x_{Z_2})) = p(x_{Z_1\setminus\{a\}} \mid \text{do}(x_{Z_2\cup\{a\}}))p(x_a)$. In the former case we have two causal submodels derived from $\mathcal{M}_1, \mathcal{M}_2$ which agree on functions and distributions of unobserved variables, and which have the observed distribution $p(x_D \mid \text{do}(x_a))$. This implies $\mathcal{M}_1$ and $\mathcal{M}_2$ must agree on $p(x_{Z_1} \mid \text{do}(x_{Z_2}))$. In the latter case, the decomposition of the effect, and the previous argument implies our conclusion.

Finally, we must show $\mathcal{M}_1$ and $\mathcal{M}_2$ disagree on γ_1.

In $\mathcal{M}_2$, γ_1 is a distribution over nodes in $R = rh(D)_{\mathcal{G}_{D\cup\{a\}}}$. By assumption, the values of the variables in set X_R can be viewed as giving the bit parity of each unobserved value, counted twice. This implies γ_1 assigns probability 0 to any value assignment to X_R with odd bit parity, and a uniform distribution to even bit parity assignments. What we must now show is that γ_1 is a different distribution in $\mathcal{M}_1$. Indeed, in $\mathcal{M}_1$ the values of the variables in set X_R can be viewed as giving the bit parity of each unobserved value, counted twice, plus 1 (because a has exactly one directed path to R in $\mathcal{G}_{D\cup\{a\}}$ where a takes on value $x_a = 1$, and exactly one directed path to R in $\mathcal{G}_{D\cup\{a\}}$ where a takes on value $x_a = 0$). This implies γ_1 assigns probability 0 to any value assignment to X_R with even bit parity, and a uniform distribution to even bit parity assignments.

The constructed models $\mathcal{M}_1, \mathcal{M}_2$ induce non-positive probabilities $p(x_{D\cup\{a\}})$. It is not difficult to augment these models to create a pair of new models $\mathcal{M}'_1, \mathcal{M}'_2$ such that $p(x_{D\cup\{a\}})$ in the new models is positive, and the models agree on P'_* (the set of interventional distributions in these new models) and disagree on γ_1.

We construct $\mathcal{M}'_1, \mathcal{M}'_2$ by adding a new unobserved binary parent for every node

in R, with a distribution $\{\epsilon, 1-\epsilon\}$, where ϵ is a very small positive real number. Clearly, $\mathcal{M}'_1, \mathcal{M}'_2$ agree on any member of P'_* involving nodes in $(D \cup \{a\}) \setminus R$. Note that any member P'_j of P'_* involving nodes $R' \subseteq R$ in $\mathcal{M}'_1, \mathcal{M}'_2$ is a function of some interventional distribution over parents of R', the distribution $P(x_{U_R})$ over unobserved parents U_R of R added to $\mathcal{M}'_1, \mathcal{M}'_2$, the functions determining the values of R in $\mathcal{M}'_1, \mathcal{M}'_2$, and the distribution over original unobserved nodes in $\mathcal{M}'_1, \mathcal{M}'_2$. Since $\mathcal{M}'_1, \mathcal{M}'_2$ agree on all these objects, they must agree on P'_*.

By construction, the probability of γ_1 in $\mathcal{M}'_2$ assigns low but non zero probabilities to odd bit parity assignments to X_R, while the probability of γ_1 in $\mathcal{M}'_1$ assigns low but non zero probabilities to even parity assignments to X_R. Since ϵ can be made arbitrarily small, this implies $\mathcal{M}'_1, \mathcal{M}'_2$ disagree on γ_1.

This concludes our proof. □

Lemma 9 *The counterfactual γ_2 shown in equation 5 is not identifiable from P_* in $\mathcal{G}^*$.*

Proof: Without loss of generality, assume every node in $\mathcal{G}^*$ has at most one child. Then we augment $\mathcal{M}'_1, \mathcal{M}'_2$ constructed in the proof of Lemma 8 by adding a binary node for every vertex in $\mathcal{G}^*$, but not $\mathcal{G}_{D\cup\{a\}}$. We let each such node obtain its value from the bit parity of its parents in $\mathcal{G}^*$ (without adding unobserved parents). Call the resulting models $\mathcal{M}"_1, \mathcal{M}"_2$.

Every node added to $\mathcal{M}"_1, \mathcal{M}"_2$ forms its own district, and for every such node w, the distribution $p(x_w \mid \text{do}(x_{Pa(w)_{\mathcal{G}^*}}))$ is the same in $\mathcal{M}"_1$ and $\mathcal{M}"_2$ by construction. This implies $\mathcal{M}"_1, \mathcal{M}"_2$ agree on $P"_*$. But by construction we also obtain that $M"_1, M"_2$ disagree on γ_2.

As before, the constructed models $M"_1, M"_2$ do not yield positive observable distributions. We augment our models and create a new pair of models $\mathcal{M}^*_1, \mathcal{M}^*_2$ which induce positive observable distributions, which agree on P_* and disagree on γ_2. To do so, we add for every node in $W^* \setminus rh(D)_{\mathcal{G}^*}$ a new binary unobserved parent with probabilities $\{\epsilon, 1-\epsilon\}$, where ϵ is a very small positive real number.

Since every node w in $W^* \setminus rh(D)_{\mathcal{G}^*}$ is its own district, by construction $\mathcal{M}_1^*, \mathcal{M}_2^*$ agree on $p(x_w \mid \text{do}(x_{Pa_{\mathcal{G}^*}(w)}))$, which implies $\mathcal{M}_1^*, \mathcal{M}_2^*$ agree on P_*.

The probability of γ_2 in $\mathcal{M}_2^*$ then assigns a small but positive probability to any even bit parity assignment to Y', while the probability of γ_2 in $\mathcal{M}_1^*$ assigns a small but positive probability to any odd bit parity assignment to Y'. Since ϵ can be made arbitrarily small, this implies $\mathcal{M}_1^*, \mathcal{M}_2^*$ disagree on γ_2.

This establishes our result. □

Lemma 10 *The counterfactual γ is not identifiable from P_* in $\mathcal{G}$.*

Proof: This can be easily established by augmenting models $\mathcal{M}_1^*, \mathcal{M}_2^*$ constructed in the previous Lemma with enough extra nodes to enlarge $\mathcal{G}^*$ to $\mathcal{G}$. These extra nodes will be fully jointly independent of each other and nodes in $\mathcal{G}^*$. (That is, any edge connecting to such nodes in $\mathcal{G}$ will be vacuous in our augmentation of $\mathcal{M}_1^*, \mathcal{M}_2^*$. It's clear from this construction that the resulting augmented models agree on P_*, disagree on γ, and induce a positive observable distribution.

This establishes completeness of the criterion. □

References

[1] Chen Avin, Ilya Shpitser, and Judea Pearl. Identifiability of path-specific effects. In *International Joint Conference on Artificial Intelligence*, volume 19, pages 357–363, 2005.

[2] Reuben M. Baron and David A. Kenny. The moderator-mediator variable distinction in social psychology research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, 51:1173–1182, 1986.

[3] Yimin Huang and Marco Valtorta. Pearl's calculus of interventions is complete. In *Twenty Second Conference On Uncertainty in Artificial Intelligence*, 2006.

[4] judea pearl. The causal mediation formula – a guide to the assessment of pathways and mechanisms. Technical Report R-379, Cognitive Systems Laboratory, University of California, Los Angeles, 2011.

[5] D. Lewis. *Counterfactuals*. Cambridge, MA: Harvard University Press, 1973.

[6] David P. Mackinnon. *Statistical Mediation Analysis*. New York: Tailor and Francisc Group, 2008.

[7] J. Neyman. Sur les applications de la thar des probabilities aux experiences agaricales: Essay des principle. excerpts reprinted (1990) in english. *Statistical Science*, 5:463–472, 1923.

[8] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan and Kaufmann, San Mateo, 1988.

[9] Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2000.

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

[11] Thomas Richardson and Peter Spirtes. Ancestral graph Markov models. *Annals of Statistics*, 30:962–1030, 2002.

[12] J. M. Robins. A graphical approach to the identification and estimation of causal parameters in mortality studies with sustained exposure periods. *Journal of Chronic Disease*, 2:139–161, 1987.

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

[14] James M. Robins and Thomas S. Richardson. Alternative graphical causal models and the identification of direct effects. *Causality and Psychopathology: Finding the Determinants of Disorders and their Cures*, 2010.

[15] J.M. Robins. A new approach to causal inference in mortality studies with sustained exposure periods – application to control of the healthy worker survivor effect. *Mathematical Modeling*, 7:1393–1512, 1986.

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

[17] Ilya Shpitser and Judea Pearl. Identification of conditional interventional distributions. In *Uncertainty in Artificial Intelligence*, volume 22, 2006.

[18] Ilya Shpitser and Judea Pearl. Identification of joint interventional distributions in recursive semi-markovian causal models. In *Twenty-First National Conference on Artificial Intelligence*, 2006.

[19] Ilya Shpitser and Judea Pearl. Complete identification methods for the causal hierarchy. Technical Report R–336, UCLA Cognitive Systems Laboratory, 2007.

[20] P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. Springer Verlag, New York, 1993.

[21] Jin Tian. Identifying conditional causal effects. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2004.

[22] Jin Tian and Judea Pearl. A general identification condition for causal effects. In *Eighteenth National Conference on Artificial Intelligence*, pages 567–573, 2002.

[23] J. Tinbergen. *An Econometric Approach to Business Cycle Problems*. Hermann, Paris, 1937.

[24] R. S. Woodworth. Dynamic psychology. *Psychologies of 1925. C. Murchison (ed.)*, 1928.

[25] S. Wright. Correlation and causation. *Journal of Agricultural Research*, 20:557–585, 1921.