Causal Identifiability and Piecemeal Experimentation

Conor Mayo-Wilson

Abstract

In medicine and the social sciences, researchers often measure only a handful of variables simultaneously. The underlying assumption behind this methodology is that combining the results of dozens of smaller studies can, in principle, yield as much information as one large study, in which dozens of variables are measured simultaneously. Mayo-Wilson [2011, 2013] shows that assumption is false when causal theories are inferred from *observational* data. This paper extends Mayo-Wilson's results to cases in which *experimental* data is available. I prove several new theorems that show that, as the number of variables under investigation grows, experiments do not improve, in the worst-case, one's ability to identify the true causal model if one can measure only a few variables at a time. However, stronger statistical assumptions (e.g., Gaussianity) significantly aid causal discovery in piecemeal inquiry, even if such assumptions are unhelpful when all variables can be measured simultaneously.

Introduction

In medicine and the social sciences, researchers often measure only a handful of variables simultaneously. A sociologist investigating poverty, for example, might begin by studying race, education and income. If she is lucky, she will be able to integrate her data with psychologists' evidence about poverty, emotion and motivation. The sociologist might later incorporate economists' data about unemployment, inequality, and poverty. And so on. As she acquires more evidence, the sociologist will begin to construct a *causal collage,* i.e., a theoretical pastiche assembled from the results of dozens of studies of differing but causally-related phenomena.

The imaginary sociologist's methodology is common in medical research and the social sciences. Relatively rare is the longitudinal study that tracks dozens of variables. Rather, evidence is often collected *piecemeal,* i.e., a few variables at a time. Unfortunately, a neglected but critical assumption underlies all piecemeal inquiry. Namely, piecemeal inquiry assumes that one large study, in which dozens of variables are measured simultaneously, provides no more information than the sum of dozens of smaller studies. Mayo-Wilson [2011, 2013] shows that assumption is false. *Even if all potential confounding variables are measured in one study or another,* when data is collected piecemeal, causal theories might be dramatically underdetermined by evidence in ways they would not have been had more variables been measured simultaneously. Mayo-Wilson calls this *the problem of piecemeal induction.*

Mayo-Wilson [2011, 2013], however, focuses exclusively on causal inference from *observational* data. It is well-known that, under some assumptions, experiments/interventions[1] allow one to infer both (i) whether or not two variables are spuriously correlated due to an unmeasured common cause, and (ii) the direction of a causal relationship if one exists. Therefore, one might guess that experiments can reduce the extent of underdetermination in piecemeal causal inference. I argue this guess is half right.

After reviewing standard assumptions for causal discovery in section one, I argue in section two that the"problem of piecemeal induction" extends to experimental data: even if all potential confounding variables are measured in one *experiment* or another, causal theories can be dramatically underdetermined by evidence.[2] This problem raises three questions that I address in section two. First, what *type* of information is lost in the piecemeal construction of causal theories from experimental data, and how *much* is lost? Second, how *often* does the problem arise? Third, when, if ever, is no information lost in integrating the findings of many experiments? Most of my results are negative. Theorems 5 and 8 show that significant amounts of information about the existence of causal relations can be lost during piecemeal inquiry, even when experiments are available. Theorem 7 suggests that information loss might be extremely common.

Nonetheless, I do not endorse skepticism towards medical research and the social sciences, where piecemeal inquiry is ubiquitous. Rather, in section three, I argue that the theorems in section two and in [Mayo-Wilson, 2013] have a simple explanation: some common methods for causal discovery (namely, those that use conditional independence constraints alone) are far weaker in piecemeal inquiry. So in section three, I state three preliminary results that suggest that, by exploiting more fine-grained statistical information about the variables under investigation, piecemeal causal inquiry might be feasible.

Before beginning, I should clarify how my results below relate to ex-

[1] As standard, I use the word *experiment* to refer to settings in which at least one variable is manipulated. A randomized controlled trial (RCT) is a paradigm of an experiment.

[2] Proofs of all new theorems are in the technical appendix.

isting literature. There are already a number of algorithms for discovering causal structure from combinations of small observational studies and experiments.[3] In this paper, I characterize what, in principle, can be learned from such algorithms, when latent variables are not present. In formal terms, my theorems describe the equivalence classes of the outputs of these algorithms, which Tillman and Eberhardt [2014] admit they do not do because "in the case of overlapping datasets, the equivalence classes very quickly become very large."

1 Preliminaries

1.1 Causation and Probability

In the United States, the federal funds rate influences consumer mortgage rates, but one cannot "see" the causal relationship for at least three reasons. First, the causal relationship is noisy. Second, both rates are affected by other economic variables. Finally, mortgage rates always lag changes in the federal reserve's policies. In general, because of statistical noise, unmeasured confounders, and lack of spatiotemporal contiguity, scientists cannot observe all causal relationships directly. Instead, they must infer causes from *probabilistic regularities*. Probabilistic regularities come in many forms, but in the first two sections of the paper, I focus on facts about conditional independence.[4]

I represent causal relationships among a set of variables $\mathcal{V}$ using **directed, acyclic graphs** (DAGs), like the ones below.[5] This representation is now standard in much research on causation [Pearl, 2000, Spirtes et al., 2000]. An arrow $V \to U$ indicates that, *relative to* the variables in the graph, the variable V is a **direct cause** of U. If there is a sequence of arrows $V_1 \to V_2 \to \ldots V_n$ and no arrow between V_1 and V_n, then V_1 is called an **indirect cause** of V_n. For example, relative to the set of variables in the left graph in Figure 1, *Age* is an indirect cause of *Income*, but it is a direct cause in the right graph.

[3] See [Tillman and Spirtes, 2011], [Tsamardinos et al., 2012], and [Triantafillou and Tsamardinos, 2015] for algorithms that work with observational data in the presence of latent confounding. Section six of [Tillman and Eberhardt, 2014] extends these algorithms to experimental data.

[4] Two events A and B are **conditionally independent given** C if $P(A, B|C) = P(A|C) \cdot P(B|C)$. This definition extends to random variables in the obvious way.

[5] For the remainder of the paper, I use the uppercase letters G and H to denote DAGs, and the upper case letters U, V, and W to denote vertices in graphs. I use $\mathcal{V}$ to represent a causally sufficient set of variables under investigation, and I use calligraphic letters like $\mathcal{U}$ and $\mathcal{E}$ to denote subsets of $\mathcal{V}$. Finally, I use the scripted letters $\mathscr{U}, \mathscr{V}$ and $\mathscr{W}$ to denote subsets of the power set $\mathcal{P}(\mathcal{V})$ of $\mathcal{V}$.

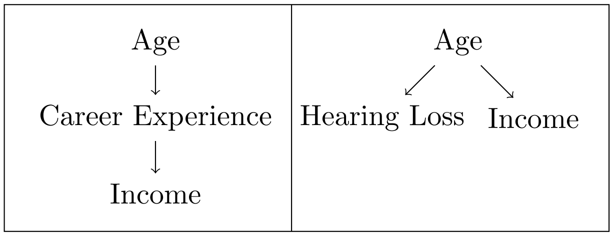


Figure 1: Directed, acyclic graphs

A set of variables $\mathcal{V}$ is **causally sufficient** if for any pair $V_1, V_2 \in \mathcal{V}$, if W is a common cause of V_1 and V_2, then $W \in \mathcal{V}$. For example, if the right graph in Figure 1 accurately represents the world, then the set of variables $\{Income,\ Hearing\ Loss\}$ is *not* causally sufficient because it omits a common cause, namely, *Age*. Given this background, I can state two standard assumptions about the relationship between causation and probability:

Causal Markov Condition (CMC): If $\mathcal{V}$ is causally sufficient, two variables in $\mathcal{V}$ are conditionally independent of their non-effects in $\mathcal{V}$ given their direct causes in $\mathcal{V}$.

Causal Faithfulness Condition (CFC): No two variables are conditionally independent unless CMC entails so.

These principles are not uncontroversial, but an enormous amount of research in causal discovery assumes one or both conditions.[6] As an example of CMC, consider the graph on the right in Figure 1. If the set of variables $\{Age,\ Income,\ Hearing\ Loss\}$ is causally sufficient, then CMC entails that *Income* is conditionally independent of *Hearing Loss* given *Age*, whereas CFC entails that *Income* and *Hearing Loss* are unconditionally dependent. So for example, Ada's income provides some information about whether Ada suffers from hearing loss because her income provides some information about her age. However, Ada's income would be irrelevant for predicting whether she suffers from hearing loss if one already knew her age.

The CMC and CFC are important because, if a set of variables $\mathcal{V}$ is causally sufficient, then one can associate every possible causal DAG over $\mathcal{V}$ with a set of assertions about conditional independence. For example, if the set $\{Age,\ Income,\ Hearing\ Loss\}$ were causally sufficient and if the rightmost graph in Figure 1 were true, then *Income*

[6] For an extensive discussion of both principles, see [Spirtes et al., 2000]. For further defenses of the CMC, see [Hausman and Woodward, 2004] and [Steel, 2005]; for criticisms, see [Cartwright, 2002, 2007]. For criticisms of CFC, see [Freedman and Humphreys, 1999] and [Cartwright, 2007].

would be independent of *Hearing Loss* given *Age*. In contrast, if *Income* were a direct cause of *Hearing Loss* or vice versa, then no such conditional independence would hold. So the CMC and CFC allow one to distinguish among competing causal theories given observed probabilistic regularities.

Unfortunately, even under the assumption that a set of variables is causally sufficient, not every pair of causal DAGs can be distinguished by conditional independence facts alone. Say two causal DAGs are **I-indistinguishable** if, by the CMC and CFC, they entail the same set of conditional independence facts.[7]

When are two causal DAGs I-indistinguishable? Verma and Pearl's theorem below provides a precise answer. Two definitions are necessary to understand the theorem. First, the **skeleton** of a causal DAG G is the undirected graph that results from ignoring the direction of the arrows in G. For instance, the graphs $V_1 \leftarrow V_2 \rightarrow V_3$ and $V_1 \rightarrow V_2 \leftarrow V_3$ both have the same skeletons. Second, three variables V_1, V_2, V_3 form an **unshielded collider** in a graph G if (i) V_1 and V_3 are direct causes of V_2, and (ii) neither V_1 nor V_3 is a direct cause of the other. Verma and Pearl show:

Theorem 1 *[Pearl and Verma [1995]] Two causal graphs are I-indistinguishable if and only if they have the same skeletons and unshielded colliders.*

Verma and Pearl's theorem can provide guidance about what observational studies to conduct. Suppose two different graphs plausibly describe the causal relationships among several variables. If the two graphs are I-distinguishable, then there is some conditional independence that is entailed by one but not the other. For example, one graph might entail that *Income* (V_1) is independent of *Hearing Loss* (V_2) given $\mathcal{U} = \{Age,\ Profession\}$. In principle, a researcher could then use a statistical test to determine whether V_1 is in fact independent of V_2 given $\mathcal{U}$ if she could **comeasure** the variables $\{V_1, V_2\} \cup \mathcal{U}$, i.e., if she could conduct a single study in which all variables in $\{V_1, V_2\} \cup \mathcal{U}$

[7] "I" stands for independence. Causal theories that cannot be distinguished by conditional independence facts might nonetheless be distinguishable using background knowledge, temporal information, and other statistical assumptions (e.g. that variables are non-Gaussian and linear combinations of their causes [Shimizu et al., 2006]). I discuss these issues further in Section 3. Formally, two graphs are what I call "I-indistinguishable" if their d-separation relations are identical (see Appendix), and so I-indistinguishability is a purely graph-theoretic relation that does not require any probabilistic notions. Nonetheless, it is easiest to explain I-indistinguishability using its relationship to a mathematically equivalent notion of "Markov equivalence," which is a relationship between *Bayesian Networks*, i.e., pairs of the form $\langle G, p\rangle$ where G is a DAG containing random variables as its vertices and p is a probability distribution over the variables in G satisfying particular conditional independence facts. Again, see Section 3.

are measured simultaneously. The result of the statistical test would then provide evidence for one theory over the other.

Unfortunately, it might be financially prohibitive, scientifically infeasible, or even unethical (e.g., for privacy reasons) to comeasure some sets of variables. Instead, researchers often conduct several studies and hope that the conditional independences among the comeasured variables are sufficient to distinguish among rival causal theories. Mayo-Wilson [2011, 2013] shows that, unfortunately, this isn't always possible: there are I-distinguishable causal graphs that cannot be distinguished using piecemeal data alone.

For example, consider the two DAGs in Figure 2. Assuming the CFC, both DAGs entail that all three variables are pairwise dependent; this is intuitive because V_1 is a cause of both V_2 and V_3 in both graphs, and so V_1 ought to be associated with both V_2 and V_3. Similarly, V_2 is a cause of V_3 in both graphs, and hence, the two variables should be dependent. Hence, if one only knew which pairs of variables were correlated, one would not be able to distinguish between the two graphs. To distinguish between the two graphs, one must know whether V_1 is conditionally independent of V_3 given V_2, but that requires comeasuring all three variables.

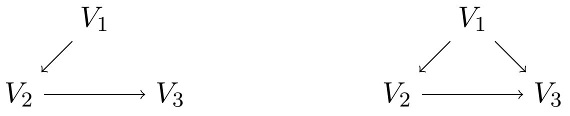


Figure 2: Graphs indistinguishable by passive observation of *pairs* of variables

Importantly, Mayo-Wilson [2011, 2013]'s arguments do not depend upon the existence of latent/unmeasured confounding variables. Rather, he assumes that, when social scientists and medical researchers integrate the results of many observational studies, the combined set of variables is causally sufficient. This assumption allows him to investigate how piecemeal measurement of variables affects what can be learned via causal inference, even if researchers find themselves in the very lucky position of having identified a causally sufficient set of variables. In this paper, I likewise investigate what can be learned from piecemeal inquiry when researchers have identified a causally sufficient set of variables, but I investigate what can be learned from *experimental* (rather than observational) data.

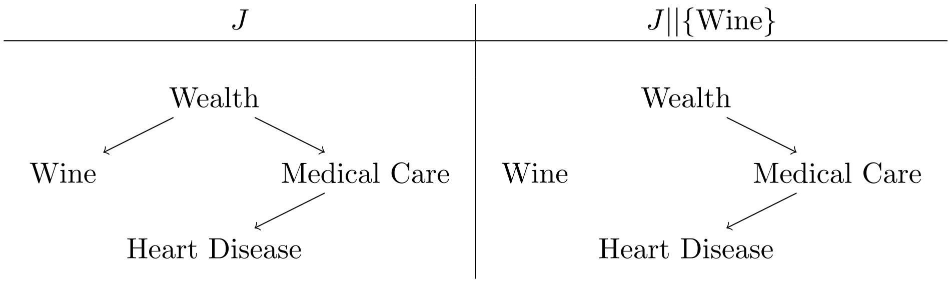
1.2 Experiments

What is the value of experiments? Let's consider an example. Suppose medical researchers are investigating whether drinking wine reduces the chances of developing heart disease. An observational study finds that wine consumption and incidence of heart disease are negatively correlated, but it's unknown whether the correlation is due to a causal connection or a latent cause (e.g., wealth).

To investigate further, researchers conduct a randomized controlled trial (RCT) in which half of the subjects are randomly selected to drink a glass of wine a day; the other half is asked to drink wine no more than once a week. When the experiment ends, it is found that the daily wine-drinkers develop heart disease with the same frequency as the weekly ones. Researchers conclude that drinking wine does not reduce the chances of developing heart disease and that there is likely some common cause (e.g., wealth) that affects both wine-drinking habits and one's chances of getting heart disease.

What justifies the researchers' conclusion?[8] And why was an experiment rather than an observational study necessary to reach this conclusion? To answer these questions, scientists must make assumptions about how experiments alter causal structure.

Let G be a causal theory like that one pictured below that describes (possible) causal relationships among heart disease, wine-drinking, wealth, and access to medical care. Next, suppose some variable V in G is manipulated, as for example, the wine-drinking habits of study participants in our example. Finally, define $G||V$ to be the graph obtained by taking G and eliminating all of the arrows that are directed *into* V. Two examples are Figure 3.



[8] I will not discuss either the ethics or epistemological necessity of randomization in treatment. See [Kadane and Seidenfeld, 1990] and [Worrall, 2007] for critical discussions of randomization.

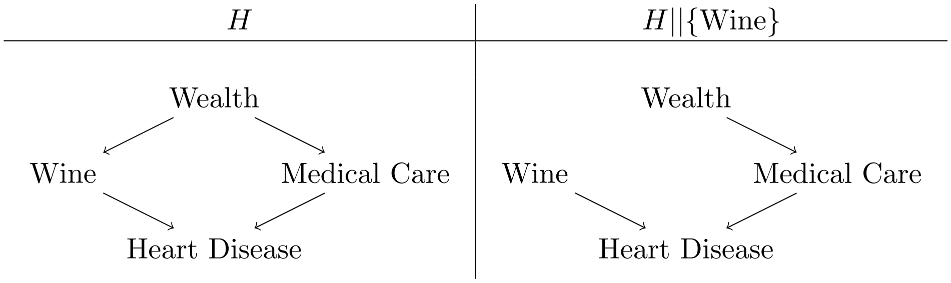


Figure 3: Graphical representation of experiments/interventions

The graph $G||V$ is intended to represent the causal relationships among the variables in G after an experiment is conducted in which researchers manipulate the variable V. Why does $G||V$ represent this? Since the variable V is controlled by experimenters, it is no longer affected by its causes.[9]

Now consider the two rival theories in the example above. The first postulates that drinking wine reduces the chances of developing heart disease, and the second postulates that the correlation between the two variables is due to at least one unmeasured common cause. These two theories are represented graphically by J and H respectively. Suppose the variable *Wine-drinking* (call it V) is manipulated an in experiment. Then the causal relationships between the variables in such an experiment would be represented by either the graph $J||V$ or the graph $H||V$, depending upon which is true.

Notice that the two graphs $J||V$ and $H||V$ differ in an important way. In $H||V$, *Wine-Drinking* is a direct cause of *Heart Disease*. Hence, if H were the true underlying causal structure, it would follow (by the CFC) that the two variables would be correlated in an experiment in which participants wine-drinking habits were manipulated. In contrast, in $J||V$, there is no path from wine-drinking to heart disease whatsoever. Therefore, if J were the true causal theory, then the two variables would be uncorrelated in the experiment (by the CMC). These results mirror the above intuitions exactly. Moreover, they show that arguments concerning the value of experiments can be made precise by using the CMC and CFC in conjunction with a particular representation

[9] The way in which experiments are represented here is most appropriate for modeling the effects of medical treatments in RCTs in which patients comply perfectly with treatment. In such (rare) trials, researchers' choices completely determine a patient's treatment. Such experiments are often called "hard" or "surgical" interventions. In this paper, I restrict my attention to hard interventions and ignore "soft" interventions, which introduce a new cause of a variable but fail to eliminate other causal influences. See [Nyberg and Korb, 2006], [Eberhardt, 2007], and [Eaton and Murphy, 2007] for discussions of soft interventions.

of experiments - namely, the representation in which particular edges are removed from a causal graph.

Perhaps most importantly, these techniques allow one to characterize what can be learned (i) from experiments in which several different variables are manipulated, and (ii) from sequences of experiments. Consider issue (i) first. Although medical researchers typically manipulate only one variable (namely, type and dosage of a treatment) in an RCT, it seems possible that more could be learned from experiments in which several variables are subject to an intervention. How should one model such complex interventions? One answer is to represent a complex intervention as a conjunction of simple ones.

More precisely, suppose the true causal theory among a collection of variables is represented by a causal graph G. Next, suppose an experiment is conducted in which several variables - call them $V_1, V_2, \ldots, V_k$ - are all subject to an intervention. One can represent the causal relationships among the variables in such an experiment by the causal graph $G||\{V_1, V_2, \ldots, V_k\}$, which is defined to be the result of removing from G all edges that point into any of the variables $V_1, V_2, \ldots V_k$. An example is below. In general, if $\mathcal{E}$ is a set of variables that are manipulated in an experiment, let $G||\mathcal{E}$ denote the causal graph obtained by deleting all edges into any of the variables in $\mathcal{E}$.

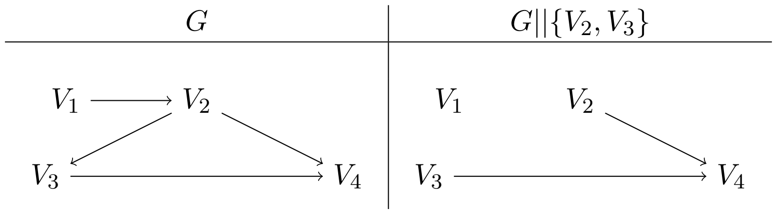


Figure 4: Causal graphs after multiple interventions

Social scientists and medical researchers often draw causal conclusions from several experiments, not just a single one. How can one represent what can be learned from a sequence of experiments? Recall, two causal graphs G and H are I-indistinguishable if they entail the same set of conditional independence facts. If an experiment is conducted in which the variables $\mathcal{E}$ are manipulated, therefore, two causal theories G and H are indistinguishable just in case $G||\mathcal{E}$ is I-indistinguishable from $H||\mathcal{E}$.

This extends naturally to sequences of experiments. For example, if variable set $\mathcal{E}_1$ is manipulated in one experiment and $\mathcal{E}_2$ is manipulated in a second, then G and H will be indistinguishable just in case $G||\mathcal{E}_1$ is I-indistinguishable from $H||\mathcal{E}_1$ and $G||\mathcal{E}_2$ is I-indistinguishable from $H||\mathcal{E}_2$. And so on for longer sequences of experiments. In general, given a series of experiments $\mathscr{E} = \{\mathcal{E}_1, \mathcal{E}_2, \ldots, \mathcal{E}_k\}$, define two graphs

G and H to be $\mathscr{E}$-experimentally indistinguishable just in case $G||\mathcal{E}$ is I-indistinguishable from $H||\mathcal{E}$ for all experimental sets $\mathcal{E}$ in $\mathscr{E}$. Several examples are below.

By appropriately choosing which experiments to conduct, one can learn far more from experiments than one could from passive observation alone. To see why, say a sequence of experiments $\mathcal{E}$ satisfies the **pair condition** just in case for any (unordered) pair of variables V and W, there is some experiment in $\mathcal{E}$ in which V is manipulated and W is passively observed. According to the following theorem, underdetermination is eliminated if one can conduct a sequence of experiments satisfying the pair condition:

Theorem 2 *[Eberhardt et al. [2006]] Let G and H be two distinct causal* DAG*s over a causally sufficient set of variables $\mathcal{V}$. Suppose $\mathcal{V}$ is passively observed, and then a series of experiments $\mathscr{E}$ is conducted. If $\mathscr{E}$ satisfies the pair condition, then G is $\mathscr{E}$-experimentally distinguishable from H. If $\mathscr{E}$ does not satisfy the pair condition, then there exist graphs that are $\mathscr{E}$-experimentally indistinguishable.*

In other words, if a set of variables is causally sufficient, one can learn the true causal graph from passive observation and any sequence of experiments satisfying the pair condition. Hence, in circumstances in which it is possible (and ethical) to perform a number of interventions, underdetermination of causal theories can be eliminated by experiment. Further, with a little work, Eberhardt's theorem provides an upper bound on the number of experiments that one needs to conduct in order to determine the true causal graph; it describes how quickly causal graphs can be learned from series of interventions.

Yet to apply Eberhardt's theorem all variables in $\mathcal{V}$ must be comeasured. Can one draw strong causal conclusions from experimental data if this is not the case?

2 Piecemeal Causal Inference from Experiments

2.1 An Example

Suppose medical researchers are investigating the effects of eating fast food on incidence of heart disease. Suppose that the true causal relationships are represented by the graph below. That is, eating fast food (indirectly) causes heart disease in two different ways: (1) it hardens one's arteries, and (2) it increase the amount of plaque in one's arteries. By Eberhardt's theorem, one could learn the true causal theory below under two assumptions, namely, that one can comeasure all four variables in an observational study, and that one can conduct a series of experiments satisfying the pair condition. In fact, if one conducted a single (unethical) experiment in which participants' fast-food eating habits were manipulated and all four variables were comeasured, then one would (with enough data) be certain the causal relationships among the four variables were represented by the graph below.

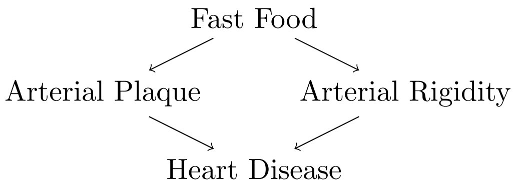
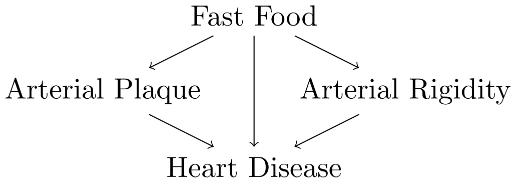


Figure 5: A causal graph underdetermined by comeasurement of all sets of three or fewer variables

However, suppose that only three variables can be comeasured in any particular experiment, and further suppose that any variable that is manipulated in an experiment is one of the three that is comeasured. For example, if an experiment is conducted in which subjects are administered medication that breaks up arterial plaque, then researchers can comeasure only two of the other variables under investigation (in addition to arterial plaque). If researchers consider complex interventions in which several variables can be manipulated simultaneously, then there are a total of 28 possible experiments and four (passive) observational studies that they might conduct.[10] Additionally, there are four observational studies in which no variables are subject to an intervention.

What can be learned from these 32 experiments and studies? In particular, can it be learned that consumption of fast food does not cause heart disease via another mechanism (perhaps directly) as shown in the diagram below?



[10] There are $\binom{4}{3}$ different ways to choose which three variables are observed. Given three variables, one can perform $3 = \binom{3}{1}$ many interventions on one variable, $3 = \binom{3}{2}$ on two variables, and $1 = \binom{3}{3}$ on all three variables. So there are $\binom{4}{3}(3 + 3 + 1) = 28$ possible experiments. Not all such experiments (e.g., intervening on all three variables simultaneously) will be informative.

Figure 6: A causal graph that is indistinguisable from the graph in Figure 5 if only three variables can be comeasured in any experiment

The answer is "no." To see why, consider an experiment in which fast food eating habits, arterial plaque, and heart disease are comeasured. Suppose that subjects' level of arterial plaque is manipulated by some medication in the experiment. In such an experiment, medical researchers would still find a correlation between incidence of heart disease and fast food eating habits. Why? By assumption, researchers do not manipulate the degree to which subjects' arteries are hardened, and so there is still a path by which eating fast food causes heart disease, namely, by hardening the arteries.

By symmetric reasoning, if another experiment were conducted in which subjects' arteries were made softer by medication, one would nonetheless detect a correlation between heart disease and fast food eating habits. Why? In such an experiment, subjects' levels of arterial plaque would be left unmanipulated, and so eating fast food would still exert causal influence on the development of heart disease.

The only experiment that would sever all causal paths between fast food eating habits and heart diseases is one in which both arterial plaque and rigidity are subject to an intervention. By assumption, however, researchers can comeasure at most three variables. Thus, if both arterial plaque and rigidity are manipulated (and hence measured), then researchers cannot observe both remaining variables. So researchers cannot learn that such a complex intervention eliminates the correlation between heart disease and fast food eating habits.

These considerations suggest that, if only three variables can be comeasured, then one cannot rule out the existence of a direct causal connection between heart disease and fast food eating habits *regardless of whether one can conduct experiments.* This can be proven formally: the so-called problem of piecemeal induction persists even when experimentation is possible. In the next section, therefore, I investigate three questions raised by this problem. First, what *type* of information is lost in the piecemeal construction of causal theories from experimental data, and how *much* is lost? Second, how *often* does the problem arise when experiments are available? Third, when, if ever, is no information lost in integrating the findings of many experiments?

2.2 The Problem of Piecemeal Induction for Experimental Data

In previous sections, I argued that a single experiment can be represented by a subset $\mathcal{E}$ of the variables under investigation; $\mathcal{E}$ represents which variables are subject to an intervention, and the remaining variables are assumed to be passively observed. If not all variables can be comeasured, however, then an experiment ought to be represented by a pair $\langle \mathcal{E}, \mathcal{U} \rangle$, where $\mathcal{E}$ represents those variables that are manipulated, and $\mathcal{U}$ represents all variables that are comeasured. I will assume that $\mathcal{E}$ is a subset of $\mathcal{U}$, so that, when researchers conduct an intervention, they know the effect of the intervention on the variables that are manipulated. A set of experiments $\mathscr{E}$, therefore, is just a set of pairs $\langle \mathcal{E}_1, \mathcal{U}_1 \rangle$, $\langle \mathcal{E}_2, \mathcal{U}_2 \rangle$, and so on.

With this representation of experiments, one can now make precise the notion of indistinguishability in piecemeal causal inquiry from experimental data. Given a set of experiments $\mathscr{E} = \{\langle \mathcal{E}_1, \mathcal{U}_1 \rangle, \langle \mathcal{E}_2, \mathcal{U}_2 \rangle, \ldots \langle \mathcal{E}_n, \mathcal{U}_n \rangle\}$, say two causal theories G and H are $\mathscr{E}$**-indistinguishable** just in case $G||\mathcal{E}$ is $\mathcal{U}$-indistinguishable from $H||\mathcal{E}$ for all experiments $\langle \mathcal{E}, \mathcal{U} \rangle$ in $\mathscr{E}$.[11] Here, the notion of $\mathcal{U}$-indistinguishability is like I-indistinguishability, except that one restricts one's attention to the variables in $\mathcal{U}$ only. That is, two graphs are $\mathcal{U}$-indistinguishable if they entail that the same set of conditional independences hold among the variables in $\mathcal{U}$. The next theorem shows that experiments do not eliminate the problem of piecemeal induction.

Theorem 3 *Let $\mathcal{V}$ be any set of variables of size at least two, and let $\mathscr{E}$ be any set of experiments such that $\mathcal{V}$ is never comeasured (i.e. there is no pair $\langle \mathcal{E}, \mathcal{V} \rangle$ in $\mathscr{E}$). Then there exist distinct causal theories G_1 and G_2 with different adjacencies that are $\mathscr{E}$-indistinguishable.*[12]

By Verma and Pearl's theorem, notice that the theories G_1 and G_2 in the above theorem are distinguishable by passive observation alone if all variables can be comeasured. Hence, the theorem raises the question: do experiments have any value in mitigating the problem of piecemeal induction? Intuitively, the answer is "yes", and this intuition can be made precise and justified. I will argue that, often, experiments provide evidence in two ways that passive observations cannot, namely, (1) by indicating the *direction* of a causal connection (whether direct or indirect), and (2) by indicating the *existence* (or absence) of a direct causal connection.

I should pause for a moment to discuss the importance of the second point. Under the assumption that all variables can be comeasured, Verma and Pearl's theorem guarantees that, if one has identified a causally sufficient set of variables, one can learn whether any two variables are directly causally connected by passive observation alone.

[11] If every pair in $\mathscr{E}$ is of the form $\langle \mathcal{E}, \mathcal{V} \rangle$ (i.e., all variables are observed in the intervention) and $\langle \emptyset, \mathcal{V} \rangle \in \mathscr{E}$ (i.e. all variables are passively observed simultaneously), then the $\mathscr{E}$-indistinguishability class is identical to the interventional equivalence class induced by $\mathcal{I} = \{\mathcal{E} : \langle \mathcal{E}, \mathcal{V} \rangle \in \mathscr{E}\}$ in the sense of [Hauser and Bühlmann, 2012].

[12] This theorem follows immediately from the next, and so only the latter is proven in the appendix.

Hence, one might conclude that, if one has identified a causally sufficient set of variables, experiments only play a role in indicating the direction of some causal connection (i.e., they are only useful for the first reason above). I will argue that this inference is hasty: even if a set of variables is causally sufficient, experiments can play a valuable role in indicating the existence or absence of causal connections precisely in the circumstances most frequently encountered by social scientists and medical researchers, namely, those in which causal theories must be constructed piecemeal.

I will now consider the two uses of experiments in order. It is fairly easy to see that experiments (represented as above) resolve all ambiguities concerning the direction of causation, *even when only two variables can be comeasured.* Why? Recall, by the CMC, two variables V and W are unconditionally dependent if and only if at least one of the following three conditions holds: (1) V is a cause of W, (2) W is a cause of V, and/or (3) the two share a common cause. I claim that researchers can learn which of the three options holds by conducting no more than two experiments in which only V and W are comeasured.

For example, consider the causal theory above concerning fast food habits, arterial plaque, arterial rigidity, and heart disease. Suppose medical researchers are interested in the relationship between arterial plaque and heart disease; they have noticed a correlation between the two, and they wish to know whether the correlation arises from a direct causal connection or from some common cause. In other words, researchers have conducted an observational study in which only arterial plaque and incidence of heart disease were comeasured, but they believe the discovered correlation may be attributable to a latent common cause. They decide to conduct an experiment.

In the experiment, researchers give some patients a pill that dissolves arterial plaque, and they give others a pill that increases plaque. After several years, researchers still observe a correlation between arterial plaque and heart disease. What can explain said correlation? In particular, can the correlation be explained by a common cause between plaque and heart disease? Or can it be explained by a theory in which heart disease is a (direct or indirect) cause of arterial plaque?

The answer to both questions is “no.” By assumption, the experiment breaks all edges into arterial plaque in the underlying causal graph. Thus, after the experiment, all causal paths between heart disease and arterial plaque are pointed “away” from the variable arterial plaque. In particular, the two variables share no common cause (as, if there were a common cause, there would be a path from the common cause “into” arterial plaque), and further, heart disease cannot be a cause of arterial plaque (as if it were, there would be a path from heart disease “into” arterial plaque). So the correlation in the experiment has only one explanation, namely, that arterial plaque is a cause of heart disease.[13] Further, if arterial plaque is a cause of heart disease in the experiment, it must likewise be a cause even when no intervention is performed.

Notice, this argument is completely general: if one knows that V and W are directly causally connected, then conducting one experiment in which V is manipulated is sufficient to reveal whether V causes W or vice versa. Combined with Verma and Pearl's theorem, this observation entails the first half of Eberhardt's theorem. Why? By Verma and Pearl's theorem, one can learn whether any two variables are directly causally connected via passive observation if all variables can be comeasured in a single study. Further, if one can conduct a series of experiments satisfying the pair condition, then the above argument shows that one can determine the direction of each direct causal connection. So one can learn the true underlying causal structure.

Importantly, the above argument is applicable even if one cannot (passively) comeasure all variables. Notice that, in the experiment involving arterial plaque and heart disease, researchers' conclusions depended only upon the observed correlation between heart disease and arterial plaque. That is, their conclusions depended in no way upon observations of any other variables, nor upon the previous observational study in which a correlation was discovered. This suggests that experiments are particularly powerful because they allow one to reach causal conclusions *even when no more than two variables can be comeasured.* Again, this intuition can be made precise.

Say two causal graphs G_1 and G_2 are ***k*, *j*-indistinguishable** if in every experiment in which at most k variables are comeasured and at most j are subject to an intervention, the theories G_1 and G_2 satisfy the same set of conditional independences. In the special case in which $j = 0$, say G_1 and G_2 are ***k*-indistinguishable**. To show that experiments mitigate the problem of piecemeal induction, therefore, one can examine the way in which k-indistinguishability and $k, 1$-indistinguishability differ. The argument just given is an informal proof of the following theorem:

Theorem 4 *Suppose G and H are $2, 1$-indistinguishable. If V is a (direct or indirect) cause of W in G, then it is likewise so in H.*

Theorem 4 shows that experiments can aid in detecting the direction of a causal connection, even when one can comeasure at most two variables in any study. To see how experiments aid in learning the *existence* (or absence) of a causal connection, it will be helpful to consider a toy example in which some variable V is a common cause of

[13] Importantly, this argument does not rule out the possibility that, if *no interventions are performed*, the two share a common cause. It simply shows that plaque is a cause of heart disease.

nine others - call them $W_1, W_2, \ldots W_9$. Suppose that none of the nine variables $W_1, W_2, \ldots W_9$ is a cause of any other, and so the causal relationships among the ten variables can be represented by the graph G in the figure below. Next, assume that researchers can passively observe at most two variables in any given study. Under this assumption, the true graph cannot be distinguished from any complete causal theory, i.e., a theory which states that all of the variables are directly causally connected. One such graph is depicted in the figure below. Here, the truth G differs from the rival complete graph H by 36 edges.

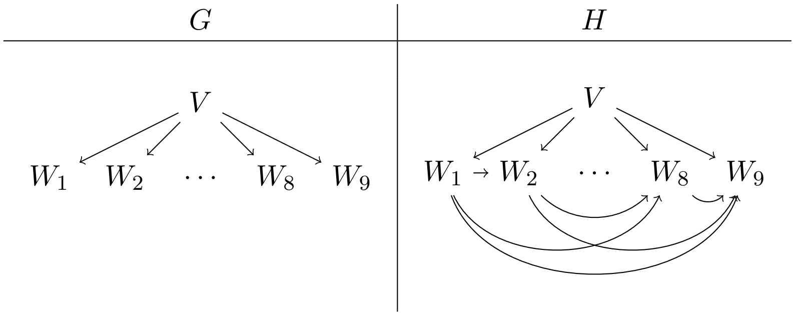


Figure 7: Graphs that are 2 indistinguishable, but $2, 1$ distinguishable

Importantly, if the true causal theory were represented by G in the above figure and no more than two variables could be passively observed in any study, then one could not tell whether *any* two variables were directly causally connected. That is, for any pair of variables X and Y under investigation, there is some causal theory indistinguishable from the truth in which X and Y are directly causally connected, and there is another theory indistinguishable from the truth in which they are not. Thus, absolutely nothing could be learned about the true causal structure G from passive observation of two variables at a time.

What could be learned if experiments were available? The answer is, “everything.” In the above example, although the true causal theory is 2-indistinguishable from many others (including complete graphs), it is $2, 1$ distinguishable from only itself. This follows almost immediately from Theorem 4. Why? Suppose some causal graph J is indistinguishable from the truth G. Then, by Theorem 4, J must entail that V is a cause of $W_1, \ldots, W_9$ as G does. Now, suppose for the sake of contradiction that W_i is a cause of W_j in J for some choice of i and j. Since J is indistinguishable from G, it follows from Theorem 4 that W_i is a cause of W_j in G as well. But no such causal connection exists in G, and so none must exist in J either. So the causal graph J is identical to G.

The above example suggests that two k, j-indistinguishable graphs might differ, in the worst-case, by fewer edges than do k-indistinguishable graphs. Unfortunately, this is not always the case. When $k = n - 1$ or $k = n - 2$, two causal theories might be k, j-indistinguishable, and yet one graph might contain one or $\binom{n-1}{2}$ edges respectively that the other does not; these are exactly the same number of edges as in the case of passive observation.[14] However, I conjecture that experiments do reduce the worst-case bounds in all "non-trivial" cases, i.e., in all cases which k is neither two nor $n - 1$.

By how many edges can two k, j-indistinguishable graphs differ? The next theorem provides a rather messy lower bound for when at most one variable is subject to an intervention (i.e. $j = 1$).

Theorem 5 *Suppose there are n variables under investigation and $k < n$. Let h be $\frac{n-2}{k-1}$ rounded down, and let $M = (n-1) - h(k-1)$. Then there exist graphs G and H such that G is $k, 1$-indistinguishable from H, and yet H contains $f(n, k)$ edges that G does not, where*

1. *$f(n, k) = n - k$ if $h = 1$,*
2. *$f(n, k) = (2n + 1) - 3k$ if $h = 2$, and*
3. *$f(n, k) = (k-1)^2\binom{h-1}{2} + (k-1)(h-1) + Mh$ if $h \geq 3$.*

I conjecture that each of these lower bounds is also an upper bound, but verifying that conjecture is future work. If this conjecture were true, then despite their messy and uninformative appearance, the bounds in Theorem 5 would show how experiments aid in discovering the presence or absence of causal connections. Why? If only passive observation were possible, two k-indistinguishable graphs might differ by $g(n, k) = \binom{n-k+1}{2}$ many edges.[15] Some quick algebra shows that g is a quadratic function of $n - k$, and so the number of edges by which two k-indistinguishable graphs might differ increases rapidly as either (a) the number of variables becomes large or (b) the number of variables one can comeasure in a study decreases. In contrast, notice that $f(n, k)$ is bounded above by a linear function of $n - k$ under certain conditions, and hence, f increases far less rapidly than g. Only when $h \geq 3$ is f a quadratic function of $n - k$.

The above theorems provide partial answers to the first question posed at the end of the last the section, namely, "what type of information is lost in the piecemeal construction of causal theories, and how much is lost?" Theorem 4 shows that, given experiments in which no more than two variables can be comeasured and at most one can be subject to an intervention, no information concerning the direction of causation is lost by the piecemeal construction of theories. In contrast,

[14] See Theorem 6 in [Mayo-Wilson, 2012].

[15] See Theorem 6 in [Mayo-Wilson, 2012].

Theorem 5 shows that, in the worst-case, quite a bit of information might still be lost concerning the presence or absence of causal connections, but that experiments still provide information that passive observation could not.

Now consider the second question ("when is no information lost?"). [Mayo-Wilson, 2013]'s Theorem 8 shows that, if the underlying true causal theory contains relatively few edges, then piecemeal inquiry leads to no loss of information. Specifically, if a DAG contains fewer than $2k - 2$ edges, then comeasuring k variables at a time is sufficient to determine its I-equivalence class. An analogous result holds when experiments are available, as is shown by the following theorem:

Theorem 6 *Suppose there are n variables under investigation. Assume that, in any given experiment, at most $k \leq n$ many variables can be comeasured and at most one variable can be manipulated. Finally, assume that the true causal theory postulates fewer than $2k - 2$ direct causal links. Then the true causal theory can be uniquely determined. In other words, the $k, 1$-indistinguishability class of any causal theory G, which has fewer than $2k - 2$ edges, contains only G itself.*

Notice that, by definition, if two theories are $k, 1$ distinguishable, then they are k, j distinguishable for all $j \geq 1$. So the above theorem shows also that, when the true causal graph is sufficiently sparse, then simple interventions/experiments are sufficient to discover it. This conclusion should be compared with results that show that, in the worst-case, complex interventions involving many variables may be necessary to discover certain causal structures.[16]

Thus far, I have characterized how much causal information is lost in the worst-case when many experiments are combined, and I also showed that, in the best-case (i.e. when the true causal theory is sufficiently simple), the truth can be determined uniquely. So one might ask, "How *often* does the problem of piecemeal induction arise?"

Say that a causal theory G_1 is $\boldsymbol{k, j}$ **underdetermined** if its k, j-indistinguishability class contains a theory G_2 such that G_1 and G_2 are not I-indistinguishable. In other words, G_1 and G_2 can be distinguished by passive observation alone if all variables can be comeasured, but they can't be distinguished if only k many variables can be comeasured at a time, even if up to j many of the observed variables may be subject to an intervention. Let $p_{k,j}(n)$ be the proportion of DAGs containing n many variables that are k, j underdetermined.

When only passive observation is possible (i.e., when $j = 0$), [Mayo-Wilson, 2013] (Theorem 9) shows that the proportion $p_{k,0}(n)$ of graphs over n many variables that are $k, 0$-undetermined approaches 1 as n approaches infinity. The following result shows that interventions do not

[16] Thanks to David Danks for suggesting this point.

reduce the frequency of underdetermination as the number of variables becomes large.

Theorem 7 *For any natural number k and any $j \leq k$, the proportion $p_{k,j}(n)$ of graphs over n many variables that are k, j-undetermined approaches 1 as n approaches infinity.*

The news gets worse. Given a causal theory G, define the **extent of $\boldsymbol{k, j}$-underdetermination** of G to be the maximum number of distinct causal theories G_1, $G_2, \ldots, G_m$ such that (i) G is k, j-indistinguishable from G_i for all $i \leq m$, and (ii) each pair of the theories $G, G_1, G_2, \ldots, G_m$ are I-distinguishable if all variables are comeasured. Let $E_{k,j}(n)$ be the average extent of k-underdetermination over all causal graphs concerning n variables. Informally, $E_{k,j}(n)$ measures how much piecemeal inquiry increases underdetermination, as it makes precise how many theories (on average) one can no longer distinguish.

When only passive observation is possible (i.e., when $j = 0$), [Mayo-Wilson, 2013] (Theorem 10) shows that the size of $k, 0$-indistinguishability classes $E_{k,0}(n)$ becomes arbitrarily large as n approaches infinity. The next result shows that interventions do not reduce the asymptotic size of indistinguishability classes as the number of variables grows.

Theorem 8 *For any natural numbers k and $j \leq k$, the average extent of k, j-underdetermination $E_{k,j}(n)$ becomes arbitrarily large as n approaches infinity.*

What is the philosophical upshot of these theorems? It's not clear. Theorem 7 says that, relative to a *uniform* distribution over the set of all DAGs consisting of n variables, the probability that a randomly chosen graph is k, j-underdetermined is high if n is large. However, it's rare that researchers will find *all* such graphs to be equally plausible given background theory. Of course, scientists are interested in large variable sets precisely in the circumstances in which the causal relationships among the variables are complex and varied, which is a reason to suspect the k, j-underdetermination might be common. But detailed case studies are necessary to determine the frequency of k, j-underdetermination in different empirical sciences.

Furthermore, although the fraction of k, j-underdetermined graphs approaches one as the number of variables is increased, neither the theorems above nor their proofs provide any information about how quickly that limit is reached. If n must be enormous in order for $p_{k,j}(n)$ to be significantly greater than zero, then researchers will know that the problem of piecemeal induction is sufficiently rare. Similar remarks apply to $E_{k,n}(n)$. A combination of simulations and detailed case studies might reveal that, in some fields, $p_{k,j}$ and $E_{k,j}(n)$ are small for realistic values of k and j.

Finally, the skeptical nature of my theorems thus far is, in part, a result of the fact that I have assumed researchers have a paucity of evidence. Researchers rarely know *only* facts about conditional independence. Social scientists and medical researchers often have substantial domain-specific knowledge that constrains plausible causal theories. But even if one restricts one's attention to purely statistical evidence, researchers will typically know more than just which variables are conditionally independent of which others. For example, researchers will know if a variable is continuous or if it can take only finitely-many values. They might know that some variables are normally-distributed whereas others are not. Researchers might have good reason to believe, given other evidence, that some variables are linear functions of others. And so on. Call these **distributional assumptions.**[17] The next section investigates what can be learned in causal inference when distributional assumptions are available.

3 Distributional Assumptions

In this section, I argue that distributional assumptions can mitigate the problem of piecemeal induction. Section 3.1 introduces a novel distinction between I-indistinguishability and what is typically called "Markov equivalence." I prove that, if not all variables can be comeasured simultaneously, then assuming the CMC and CFC allows one to distinguish causal graphs that *cannot* be distinguished by conditional independence facts alone.

In section 3.2, I investigate what can be learned using four types of stronger distributional assumptions, namely, the assumptions that the underlying causal model is (1) discrete multinomial, (2) noisy-or, (3) linear Gaussian, or (4) linear non-Gaussian. Here, I prove only two preliminary results, but the results highlight two important methodological points. First, although some distributional assumptions (e.g., that the true joint distribution is multivariate Gaussian) are known to be uninformative (in a sense to be clarified) when all variables are comeasured, those same assumptions are essential in piecemeal inquiry. Conversely, assumptions that are fruitful when all variables can be comeasured (e.g., that the true model is linear but non-Gaussian) may not be as helpful in piecemeal inquiry. Section 3.3 concludes with a brief discussion of what can be learned from a combination of experiments and distributional assumptions.

[17] So distributional assumptions include both parametric assumptions (e.g., that the true model is linear Gaussian) and non-parametric ones (e.g., that the model is non-Gaussian).

3.1 Markov Equivalence vs. I-Indistiguinshability

Fix a set of variables $\mathcal{V}$. Define a **causal model** over $\mathcal{V}$ to be a pair of the form $\langle G, p\rangle$, where G is a causal graph and p is a probability measure over the variables in G. Given a set of causal models $\boldsymbol{M}$, say a probability distribution p is **$\boldsymbol{M}$-compatible** with G if $\langle G, p\rangle \in \boldsymbol{M}$.

Say G is **$\boldsymbol{M}$-indistinguishable** from H if every probability distribution that is $\boldsymbol{M}$-compatible with G is $\boldsymbol{M}$-compatible with H and vice versa. Thus far, I have focused on when $\boldsymbol{M}$ consists of all pairs $\langle G, p\rangle$ such that the variables in G satisfy the CMC and CFC *with respect to* p.[18] Such causal models are called **Bayesian Networks**, and so let $\boldsymbol{M}_{BN}$ be the set of Bayesian networks.[19]

In the causal discovery literature, it is common to say two graphs G and H are **Markov equivalent** if they are $\boldsymbol{M}_{BN}$-indistinguishable. By definition, Markov-equivalence and I-indistinguishability are mathematically equivalent.[20] That mathematical equivalence, I believe, has led some practitioners to infer that, if one assumes only the CMC and CFC, one can distinguish between two causal graphs if and only if the graphs encode different conditional independence facts. In other words, one might infer that the CMC and CFC entail only that the observed data will (in the limit) satisfy certain conditional independence constraints. As I now show, that inference is valid only if one assumes that all variables can be comeasured simultaneously.

Given a set of variables $\mathcal{V}$, a subset $\mathcal{U} \subseteq \mathcal{V}$, and a probability distribution p over $\mathcal{V}$, the *marginal distribution* $p_{\mathcal{U}}$ describes the probabilities of events involving all and only the variables in $\mathcal{U}$. For example, if $\mathcal{V} = \{Wealth,\ Wine,\ Medical\ Care, Heart\ Disease\}$ and $\mathcal{U} = \{Wealth,\ Wine\}$, then $p_{\mathcal{U}}$ will specify the probability that a wealthy person drinks wine frequently, but it will not specify how probable it is that wealthy people have access to medical care.

If p describes the true underlying probability distribution among a set of variables $\mathcal{V}$, and if a researcher conducts a series of observational studies $\mathscr{U}$ in which one comeasures $\mathcal{U}_1$ in the first study, comeasures $\mathcal{U}_2$ in the second study, and so on, then she will acquire estimates of the marginal distributions $p_{\mathcal{U}_1}, p_{\mathcal{U}_2}$ and so on. Thus, given a set of causal models $\boldsymbol{M}$ and set of observational studies $\mathscr{U} = \{\mathcal{U}_1, \mathcal{U}_2, \ldots, \mathcal{U}_n\}$, say G is **$\mathscr{U}\boldsymbol{M}$-indistinguishable** from H if

1. For all distributions p that are $\boldsymbol{M}$-compatible with G, there is a

[18] Saying that G satisfies the CMC "with respect to p" means that every variable in G is conditionally independent *with respect to p* of its non-descendants given its ancestors.

[19] So "q is $\boldsymbol{M}_{BN}$ compatible with H" means "q is Markov and faithful to H."

[20] Note, I have defined "Bayesian network" using the Markov and faithfulness conditions. If one defines "Bayesian network" in terms of a condition that requires a probability distribution to factor in a particular way, the equivalence of Markov-equivalence and I-indistinguishability requires a proof. See [Lauritzen et al., 1990].

distribution q that is $\boldsymbol{M}$-compatible with H such that $p_{\mathcal{U}} = q_{\mathcal{U}}$ for all $\mathcal{U} \in \mathscr{U}$, AND

2. For all distributions p that are $\boldsymbol{M}$-compatible with H, there is a distribution q that is $\boldsymbol{M}$-compatible with G such that $p_{\mathcal{U}} = q_{\mathcal{U}}$ for all $\mathcal{U} \in \mathscr{U}$.

$\mathscr{U}\boldsymbol{M}$-indistinguishability formalizes what one could learn if one assumes that the true causal model belongs to $\boldsymbol{M}$ and that one can accurately estimate the marginal distribution $p_{\mathcal{U}}$ for each set of variables in $\mathcal{U} \in \mathscr{U}$. The obvious way to generalize the notion of Markov equivalence to piecemeal inquiry, therefore, is as follows. Say G is **$\mathscr{U}$-Markov-Equivalent** to H if the two graphs are $\mathscr{U}\boldsymbol{M}_{BN}$-indistinguishable.

Although Markov equivalence and I-indistinguishability are equivalent, their piecemeal analogs are not, as the next theorem shows. Before stating the result, a bit of motivation. Suppose $\langle G, p\rangle$ is the true causal model, and assume a researcher conducts a series of observational studies $\mathscr{U}$. Finally, imagine she makes no assumptions about the type of probability distribution generating the data other than it satisfies CMC and CFC with respect to the true DAG. Then if the researcher learns only the outcomes of conditional independence tests, then she will be unable to distinguish G from graphs H that are $\mathscr{U}$-indistinguishable from G. In contrast, if she pays attention to features of the marginal distributions she observes (in addition to the independence structure), she will be unable to distinguish G from graphs H only if G and H are $\mathscr{U}\boldsymbol{M}_{BN}$ indistinguishable.

Theorem 9 *If G and H are $\mathscr{U}$-Markov-Equivalent, then they are $\mathscr{U}$-indistinguishable. The converse is false in a fairly strong sense. Suppose $\mathcal{V} \notin \mathscr{U}$ and that $\mathcal{V}$ contains at least three variables. Then there exist $\mathscr{U}$-indistinguishable graphs G and H that are not $\mathscr{U}$-Markov-Equivalent.*

Theorem 9 says that, even when one makes no distributional assumptions other than the CMC and CFC, fine-grained statistical evidence – such as the probability that a person who drinks a glass of wine a day also develops heart disease – can sometimes help distinguish causal graphs during piecemeal inquiry in ways that conditional independence facts – like that wine-drinking is not independent of heart disease – cannot.

Why is Theorem 9 true? Suppose that $\langle G, p\rangle$ is the true underlying causal model and that a researcher has conducted a series of observational studies $\mathscr{U}$. Further, assume the graph H is $\mathscr{U}$-indistinguishable from G. That means there is probability distribution q that is $\boldsymbol{M}_{BN}$ compatible with H such that q entails the same conditional independence facts as p with respect to all the subsets of variables comeasured in the studies of $\mathscr{U}$. So if a researcher knew only the outcomes of conditional independence tests, she would be unable to rule out the possibility that $\langle H, q \rangle$ generated her data. However, although p and q might entail the same conditional independence facts, the marginal distributions of q over $\mathscr{U}$ might differ from those of p. So although q will be consistent with known conditional independence facts, it might conflict with other known features of p.

Why does the same reasoning not show that Markov equivalence and I-indistinguishability come apart? The answer is that if (1) p is $\boldsymbol{M}_{BN}$ compatible with G and (2) G is I-indistinguishable from H, then p *itself* is $\boldsymbol{M}_{BN}$ compatible with H, by definition of $\boldsymbol{M}_{BN}$-compatibility. In contrast, it is possible that (1) p is $\boldsymbol{M}_{BN}$ compatible with G, (2) G is $\mathscr{U}$-indistinguishable from H, and (3) p is *not* be $\boldsymbol{M}_{BN}$ compatible with H because p may not entail the set of conditional independences encoded by H over all of $\mathcal{V}$.

In short, theorem 9 suggests that attending to more than conditional independence can eliminate some underdetermination in piecemeal inquiry. In the next subsection, I show that a few stronger, but commonly satisfied, distributional assumptions are powerful tools in piecemeal inquiry.

3.2 Stronger Distributional Assumptions

If $\boldsymbol{M}$ represents the set of causal models that are possible *a priori*, then intuitively, the difficulty of causal discovery will vary with the size of $\boldsymbol{M}$: the larger $\boldsymbol{M}$ is, the more difficult discovery will be. In this section, I vindicate that intuition. I prove that, if it is known that the true causal model belongs to one of four classes of models that have been studied extensively, then causal discovery becomes easier.

Discrete Multinomial Models: A *discrete, multinomial model* (DMM) is a causal model in which all the variables take at most finitely many values. Let $\boldsymbol{D}$ denote the set of DMMs. Discrete models are ubiquitous in medical research and the social sciences. Let $\boldsymbol{D}^+$ denote the class of DMMs with a positive distribution, i.e., the class of pairs $\langle G, p \rangle$ such that $p(V_1 = r_1, V_2 = r_2, \ldots, V_n = r_n) > 0$ for all values of $V_1, \ldots, V_n$.

Noisy-Or Models: A special class of DMMs are called *noisy-or models*. Suppose Jane gets headaches routinely, and suppose that her headaches are often (but not always) caused by allergies or a cold (or both). So there are three variables under investigation, and all three are binary: either Jane has a headache or not, either she has a cold or not, and either she has allergies or not.

Suppose that if Jane has a cold, she does not always develop a headache and similarly for allergies. In each case, other factors must be present for Jane to develop a headache. For instance, Jane's cold might produce a headache only if she forgets to take an aspirin. Group all the relevant factors into a single variable B_{cold} such that, if Jane has a cold and the factors B_{cold} are present, then she will develop a headache. Similarly for allergies. Finally, suppose that, although Jane's headaches are often caused by allergies or a cold, there is some chance that she develops a headache even if she has neither a cold or allergies. In sum, Jane will develop a headache if and only if

- Jane has a cold and B_{cold} are present, *or*
- Jane has allergies and $B_{allergies}$ are present, *or*
- Some other unknown factor produces a headache.

In other words, the variable *Headache* is a logical disjunction of its causes, which is why the causal relationships described here are an example of a "noisy or" model. What makes the disjunction "noisy" is that the factors b_{cold} (or $b_{allergies}$) need to be present in order for the presence of a cold or allergies to have an effect.

In general, given a causal graph G among a collection of observable, binary variables $\mathcal{V}$, a noisy-or model is determined by a series of equations of the following sort. For each variable V, there is some unobservable "noise term" E_V such that, if E_V takes the value one, so does the variable V. Moreover, these error terms are independent. Similarly, for each direct causal connection $W \to V$, there is some unmeasured variable B_{WV} such that, if both W and B_{WV} take the value one, so does V. These variables B_{WV} are likewise independent, and they are also independent of the error terms. In symbols, one can write any variable V as a function of its parents $\textsc{pa}_G(V)$ as follows:

$$V = E_V \vee \bigvee_{W \in \textsc{pa}_G(V)} (B_{WV} \wedge W).$$

Let $\boldsymbol{N}\vee$ denote the class of noisy-or models.

Linear Gaussian: In noisy-or models, all variables take two values, and each variable is a logical disjunction of its parents. In linear Gaussian models, all variables are normally distributed, and each is a linear function of its parents. Formally, given a causal graph G among a collection of observable variables, a linear Gaussian model is determined by a set of real numbers and equations as follows. For each variable V in the graph, the model contains some normally distributed error variable E_V, and for each direct causal connection $W \to V$, there is some real number b_{WV} that represents the "strength" of the connection between W and V. So one can write any variable V as a function of its parents $\textsc{pa}_G(V)$ as follows:

$$V = E_V + \sum_{W \in \textsc{pa}_G(V)} b_{WV} \cdot W$$

Note the similarity between noisy-or and linear Gaussian models. If one systematically replaces the "or" with addition and "and" with multiplication in the equations of the noisy-or models, then one obtains a linear model. Let $\boldsymbol{LG}$ denote the class of all linear Gaussian models.

Linear Non-Gaussian: In a linear non-Gaussian model, variables are once again linear combinations of their parents and an independent noise term; the only difference from $\boldsymbol{LG}$ is that the error terms ϵ_v for each variable v are assumed to be *anything but* normally distributed. Let $\boldsymbol{Lingam}$ denote the class of all linear non-Gaussian models.

What can be learned when one assumes the true causal model belongs to one of these four groups? It turns out that for DMMs and linear Gaussian models, distributional assumptions seem to be of no use if all variables can be comeasured.

Theorem 10 *[Geiger et. al. 1990] Suppose G and H are I-indistinguishable. Let $\boldsymbol{M}$ be either the class of discrete models or linear Gaussian ones. Then G and H are also $\boldsymbol{M}$-indistinguishable.*[21]

However, the same assumption is tremendously powerful in piecemeal inquiry, and this is already implicitly widely-recognized. To see why, say G and H are $\boldsymbol{M}k$ indistinguishable if they are $\mathscr{U}\boldsymbol{M}$ indistinguishable and $\mathscr{U}$ consists of all subsets of size k.

Theorem 11 *Let $\boldsymbol{M}$ be the class of linear Gaussian models, and suppose that G and H are $\boldsymbol{M}2$-indistinguishable. Then G and H are $\boldsymbol{M}$-indistinguishable, and hence, I-indistinguishable.*

In other words, if the true causal model is known to be linear Gaussian, then comeasuring two variables at a time provides as much information about causal structure as passively observing all variables at once.[22] The proof of Theorem 11 is trivial,[23] but its importance,

[21] To my knowledge, theorem 10 was first proven by Geiger et al. [1990]. In the linear Gaussian case, the theorem was generalized by Richardson and Spirtes [2002] to include causal theories with latent variables in the presence of selection bias. An alternative proof in the discrete case was given by Meek [1995]. A constructive proof for the linear Gaussian case is given in [Mayo-Wilson, 2012]. Theorem 10 ought to be contrasted with results of [Shimizu et al., 2006], which shows that G is $\boldsymbol{M}$-indistinguishable from only itself if $\boldsymbol{M} = \boldsymbol{Lingam}$. [Hoyer et al., 2009] show that $\boldsymbol{M}$-indistinguishability differs from I-indistinguishability when $\boldsymbol{M}$ contains nonlinear additive noise models, but as far as I am aware, there is no general characterization of $\boldsymbol{M}$-indistinguishability for noisy-or models or for when $\boldsymbol{M}$ contains nonlinear additive noise models.

[22] At this point, I should reminder readers that all of the equivalence classes I have introduced characterize indistinguishability "in the limit" i.e., with arbitrarily large samples.

[23] Here's the proof, which was suggested to me by Frederick Eberhardt. If every pair of

I think, has been overlooked. Geiger's theorem might, on first glance, suggest that the assumption of normality is inert in causal inference. Theorem 11 shows this is the wrong interpretation. Rather, because the assumption that all variables can be comeasured is so strong, some distributional assumptions appear useless. But this is akin to arguing that a hammer is useless because, on some occasions, one has access to a set of power tools. Recall that if one jettisons the normality assumption and relies exclusively on conditional independence facts, then two causal theories over n many variables might postulate as many as $\binom{n-1}{2}$ different direct causal connections if only two variables can be comeasured in any study.

I conjecture that a result similar to theorem 11 is likewise true for linear non-Gaussian models.[24] Why? In such models, the joint effect of several variables is the sum of the individual effects: causes do not interact. So one might expect to be able to learn about direct causal connections by investigating pairwise interactions among the variables.

The case for noisy-or models is less clear. In the proof of Theorem 9, I construct two causal theories that are 2-indistinguishable but not $\boldsymbol{M}2$-indistinguishable if $\boldsymbol{M} = \boldsymbol{N}\vee$ is the set of noisy-or models. So the noisy-or distributional assumption provides information beyond what can be inferred from conditional independence constraints alone. That provides some optimism, as well as the structural similarity between noisy-or and linear models. But it is unclear how to use pairwise measurements of a noisy-or model to derive all conditional independence constraints; further research is necessary.

What about discrete models generally? Here is a very preliminary result.

Theorem 12 *Let $\boldsymbol{M} \subseteq \boldsymbol{D}^+$ be the class of positive* DMM*s with only binary variables. There exist causal theories G and H that are 2-indistinguishable but not $\boldsymbol{M}2$-indistinguishable.*

For reasons explained in the appendix, my proof of theorem 12 is not easily generalizable, and in particular, it does not help one understand the relationship between k-indistinguishability and $\boldsymbol{M}k$-indistinguishability if $k > 2$ or if $\boldsymbol{M}$ consists of models with dis-

variables is comeasured, then it follows that every variable V is measured in some study. So one can calculate the mean of each variable V. Similarly, if V and W are comeasured in an observational study, then one can calculate the covariance of V and W. Thus, if one can comeasure all pairs of variables, one can calculate the entire covariance matrix and mean vector. If the true model is linear Gaussian, then the unknown probability distribution is completely characterized by the mean vector and covariance matrix. Hence, any two models that are $\boldsymbol{M}2$-indistinguishable are likewise $\boldsymbol{M}$-indistinguishable.

[24] But one will need to eliminate the last clause, "and hence, I-indistinguishable", from the statement of the theorem, as $\boldsymbol{M}$-indistinguishability does not entail I-indistinguishability if $\boldsymbol{M} = \boldsymbol{Lingam}$.

crete variables that are *not* binary. So an open question is whether stronger distributional assumptions about discrete models might aid one in piecemeal causal inference.

3.3 Combining Experiments and distributional Assumptions

Of course, distributional assumptions can be combined with experiments. For example, by theorem 11, it follows that, if the true model is linear Gaussian, then any graph that is $2,0$-indistinguishable from the truth will have the same edges as the true graph. By Theorem 4, if researchers can manipulate any variable under investigation, then any graph indistinguishable from the truth will postulate that V is a (perhaps indirect) cause of W only if the true graph does. Hence, if the true model is known to be linear Gaussian *and* one can conduct any experiment in which two variables are comeasured and one is manipulated, then the true graph can be determined. That fact follows from a more general result in Hyttinen et al. [2010], where it is shown that the same fact remains true even if (i) the true causal graph is cyclic, (ii) there are latent variables, and/or (iv) the CFC fails. The moral is the combining experiments and distributional assumptions can yield informative conclusions even in piecemeal inquiry.

4 Conclusions and Future Work

I have argued that, if causal conclusions are inferred from conditional independence facts alone, piecemeal inquiry can dramatically increase underdetermination. However, theorem 9 suggests that finer-grained statistical information about the joint distribution - even in the absence of parametric or non-parametric assumptions - can help distinguish rival causal theories. Because my results are preliminary, the primary contribution of this paper, I would argue, is to highlight several important types of open questions about the piecemeal construction of causal theories from experimental data. Here, I discuss six categories of questions.

First, even when the problem of piecemeal induction is inevitable, it is possible that scientific institutions might be able to *plan* sequences of experiments so as to *minimize* the type of causal information that is lost. Eberhardt et al. [2006] proves a series of results that characterize how many experiments are necessary, in the worst-case, to discover the true causal graph. One can ask similar questions about the worst-case number of experiments necessary to determine that $k,1$-indistinguishability class of the true graph. Or, given existing studies and experiments, one can ask *which* larger subsets of variables ought to be comeasured next to reduce underdetermination.

Second, many of the above results assume that one can manipulate any variable under investigation and that every subset of variables of a fixed size can be comeasured. Neither of these assumptions is typically true. Even if some variables can be manipulated (e.g., arterial plaque through medication), intervening on others (e.g., to induce heart disease) might be unethical, practically impossible, or both. Similarly, not every subset of k many variables can be comeasured. Future research ought to characterize what can be learned from series of experiments that are not so "combinatorially nice."

Third, Theorem 9 shows that, even in the absence of parametric or non-parametric assumptions, the CMC and CFC entail that *there exist* rival causal models that can be distinguished by their marginal distributions, even if the marginal distributions of the two models entail the same conditional independence facts. Unfortunately, my proof does not provide a general method for determining which graphs are compatible with *arbitrarily given marginal distributions*, and so it cannot be used to devise a causal discovery algorithm from piecemeal data.

Fourth, all of my results assume CFC and that causal graphs are acyclic. What can be learned if one or more of these assumptions is dropped? Fifth, section three characterizes $\mathscr{U}\boldsymbol{M}$-indistinguishability classes only when $\boldsymbol{M}$ is the set of linear Gaussian models. For every other set of causal models $\boldsymbol{M}$, a precise characterization of $\mathscr{U}\boldsymbol{M}$-indistinguishability remains an open problem.

Finally, the above representation of experiments is not always realistic. Why? I have assumed that, when a variable V is manipulated, the causal influence that any other factors exert on V is eliminated. Such "hard" interventions are rarely possible in the social sciences.

Under the assumption that all variables can be comeasured in an experiment, Eberhardt [2007] and Eberhardt and Scheines [2007] investigate what can be learned from sequences of "soft" interventions, in which the experiment does not sever causal relationships between the manipulated variable and its causes. They show that soft interventions can be extremely informative when all variables are co-measured. But do soft interventions improve upon what can be learned by passive observation alone in piecemeal inquiry?

An example motivates optimism. Let G be the graph $V_1 \to V_2 \to V_3$. The 2-indistinguishability class of G contains all complete graphs and any graph with two edges and no unshielded colliders. The $2,1$-indistinguishability class of G contains only G and the graph obtained by adding an $X_1 \to X_3$ edge to G. Now suppose we define a j, k "soft intervention" class to be one in which we can introduce (and observe) j many new causes and observe k many variables in the original set. For instance, a soft intervention might allow one to introduce a new cause Z of V_2 and observe the relationship between Z, V_2 and V_3. It is easy to show that that the $2, 1$-"soft intervention" class of G is the same as the hard one. So soft interventions sometimes improve upon passive observation in piecemeal inquiry. Further research ought to characterize soft-interventional equivalence classes generally.

Acknowledgements

Thanks to David Danks, Clark Glymour, and Peter Spirtes for asking the questions that led to the results reported in this paper. While writing the paper, I benefited from several discussions with Frederick Eberhardt. Finally, thanks to three anonymous reviewers for their detailed comments and feedback.

References

N. Cartwright. Against modularity, the causal Markov condition, and any link between the two: Comments on Hausman and Woodward. *The British Journal for the Philosophy of Science*, 53(3):411–453, 2002.

N. Cartwright. *Hunting causes and using them: approaches in philosophy and economics*. Cambridge University Press, 2007.

D. Danks and C. Glymour. Linearity properties of Bayes nets with binary variables. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial intelligence*, pages 98–104, 2001.

Daniel Eaton and Kevin Murphy. Exact Bayesian structure learning from uncertain interventions. *Artificial Intelligence and Statistics*, 2007.

F. Eberhardt. Causation and intervention. *Unpublished doctoral dissertation, Carnegie Mellon University*, 2007.

F. Eberhardt and R. Scheines. Interventions and causal inference. *Philosophy of Science*, 74(5):981–995, 2007.

F. Eberhardt, C. Glymour, and R. Scheines. N-1 experiments suffice to determine the causal relations among n variables. *Innovations in machine learning*, pages 97–112, 2006.

D. Freedman and P. Humphreys. Are there algorithms that discover causal structure? *Synthese*, 121(1):29–54, 1999.

D. Geiger, T. Verma, and J. Pearl. Identifying independence in Bayesian networks. *Networks*, 20(5):507–534, 1990.

Alain Hauser and Peter Bühlmann. Characterization and greedy learning of interventional Markov equivalence classes of directed acyclic graphs. *Journal of Machine Learning Research*, 13(Aug):2409–2464, 2012.

D. Hausman and J. Woodward. Manipulation and the causal Markov condition. *Philosophy of Science*, 71(5):846–856, 2004.

Patrik O. Hoyer, Dominik Janzing, Joris M. Mooij, Jonas Peters, and Bernhard Schlkopf. Nonlinear causal discovery with additive noise models. In *Advances in Neural Information Processing Systems*, pages 689–696, 2009.

A. Hyttinen, F. Eberhardt, and P.O. Hoyer. Causal discovery for linear cyclic models with latent variables. *In Proceedings of the 5th European Workshop on Probabilistic Graphical Models (PGM 2010)*, 2010.

Joseph B. Kadane and Teddy Seidenfeld. Randomization in a Bayesian perspective. *Journal of Statistical Planning and Inference*, 25(3): 329–345, 1990.

S. L. Lauritzen, A. P. Dawid, B. N. Larsen, and H. G. Leimer. Independence properties of directed Markov fields. *Networks*, 20(5): 491–505, 1990.

C. Mayo-Wilson. The Problem of Piecemeal Induction. *Philosophy of Science*, 78(5):864–874, 2011.

C. Mayo-Wilson. *Combining Causal Theories and Dividing Scientific Labor*. Doctoral Dissertation. Carnegie Mellon University, 2012.

C. Mayo-Wilson. The Limits of Piecemeal Causal Inference. *The British Journal for the Philosophy of Science*, 2013. doi: 10.1093/bjps/axs030.

C. Meek. Strong completeness and faithfulness in Bayesian networks. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, pages 411–418, 1995.

E. Nyberg and K. Korb. Informative interventions. In *Causality and Probability in the Sciences*. College Publications, London, 2006.

J. Pearl. *Causality: models, reasoning, and inference*, volume 47. Cambridge Univ Press, 2000.

Judea Pearl and Thomas S. Verma. A theory of inferred causation. In Dag Prawitz, Brian Skyrms, and Dag Westersthl, editors, *Studies* *in Logic and the Foundations of Mathematics*, volume 134 of *Logic, Methodology and Philosophy of Science IX*, pages 789–811. Elsevier, January 1995.

T. Richardson and P. Spirtes. Ancestral graph Markov models. *The Annals of Statistics*, 30(4):962–1030, 2002.

S. Shimizu, P.O. Hoyer, A. Hyvrinen, and A. Kerminen. A linear non-Gaussian acyclic model for causal discovery. *The Journal of Machine Learning Research*, 7:2003–2030, 2006.

Peter Spirtes, Clark N. Glymour, and Richard Scheines. *Causation, Prediction, and Search.* The MIT Press, 2000.

D. Steel. Indeterminism and the causal Markov condition. *The British journal for the philosophy of science*, 56(1):3–26, 2005.

R.E. Tillman and F. Eberhardt. Learning causal structure from multiple datasets with similar variable sets. *Behaviormetrika*, 41(1): 41–64, 2014.

R.E. Tillman and P. Spirtes. Learning equivalence classes of acyclic models with latent and selection variables from multiple datasets with overlapping variables. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS 2011)*, 2011.

S. Triantafillou and I. Tsamardinos. Constraint-based causal discovery from multiple interventions over overlapping variable sets. *Journal of Machine Learning Research*, 16:2147–2205, 2015.

I. Tsamardinos, S. Triantafillou, and V. Lagani. Towards integrative causal analysis of heterogeneous data sets and studies. *Journal of Machine Learning Research*, 13:1097–1157., April 2012.

John Worrall. Why There's No Cause to Randomize. *The British Journal for the Philosophy of Science*, 58(3):451–488, September 2007.

Appendix

5 Directed Acyclic Digraphs

5.1 Notation and Definitions

For any finite set $\mathcal{V}$, let $\textsc{dag}_{\mathcal{V}}$ denote the set of all directed acyclic graphs (DAGs) that have the vertex set $\mathcal{V}$.

Let $G \in \textsc{dag}_{\mathcal{V}}$ and $V \to W$ be an edge in G. Then V is called a **parent** of W, and W is called a **child** of V. If V is either a parent or child of W, then the two vertices are said to be **adjacent** in G. Let $\textsc{pa}_G(V)$ denote the set of parents of V in G, and let $\textsc{ch}_G(V)$ denote its children. If $V_1 \to V_3 \leftarrow V_2 \in G$, then V_3 is called a **collider** with respect to V_1 and V_2. If V_3 is a collider with respect to V_1 and V_2 and, in addition, there is no edge between V_1 and V_2, then V_3 is called an **unshielded collider** with respect to V_1 and V_2.

A **path** π in G is a non-repeating sequence of vertices $\pi = \langle V_1, V_2, \ldots, V_n \rangle$ such that V_i and V_{i+1} are adjacent if $1 \leq i < n$. If V_i and V_{i+2} are both parents of V_{i+1} in G, then V_{i+1} is said to be a **collider on** π; if not, it is a **non-collider** on π. Endpoints of a path are, by definition, non-colliders on the path. A path π is called **directed** if V_i is a parent of V_{i+1} for all i. If there is a directed path from V to W, then V is said to be an **ancestor** of W, and W is said to be a **descendant** of V. For any vertex V, let $\textsc{desc}_G(V)$ denote the set of descendants of V in G.

Given a path $\pi = \langle V_1, V_2, \ldots, V_n \rangle$, let $\pi \downarrow V_i = \langle V_1, \ldots, V_i \rangle$, and call $\pi \downarrow V_i$ the **initial segment** of π that terminates with V_i. Similarly, let $\pi \uparrow V_i = \langle V_i, \ldots, V_n \rangle$, and call $\pi \uparrow V_i$ the **tail** of π that begins with V_i and terminates with the end of π. Given two paths π_1 and π_2 in a graph G such that the endpoint of π_1 is the starting point of π_2, let $\pi_1 \frown \pi_2$ denote the concatenation of the two paths.

In diagrams, I use straight lines to indicate the existence of an edge. Undirected paths are indicated by curves with no end markers (like that between V_2 and V_3), and a directed path is indicated by a curve with an arrow marker at one end (e.g. there is a directed path from V_4 to V_3).

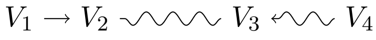


Figure 8: Edges, undirected paths, and directed paths

5.2 d-Separation

Fix a set $\mathcal{V}$ and $G \in \textsc{dag}_{\mathcal{V}}$. Let $V_1, V_2 \in \mathcal{V}$ be distinct vertices and $\mathcal{U} \subseteq \mathcal{V} \setminus \{V_1, V_2\}$. A path π between V_1 and V_2 is said to be **d-connecting** given $\mathcal{U}$ (or **active** given $\mathcal{U}$) if both of the following conditions hold:

1. Every non-collider on π is not in $\mathcal{U}$, and
2. Every collider on π is either in $\mathcal{U}$ or contains a descendant in $\mathcal{U}$.

Say V_1 and V_2 **d-connected** given $\mathcal{U}$ in G if there is a d-connecting path between the two, and say they are **d-separated** otherwise. Given three disjoint vertex sets $\mathcal{V}_1, \mathcal{V}_2, \mathcal{U} \in \mathcal{V}$, say that $\mathcal{V}_1$ and $\mathcal{V}_2$ are d-connected given $\mathcal{U}$ if there exists vertices $V_1 \in \mathcal{V}_1$ and $V_2 \in \mathcal{V}_2$ such that V_1 and V_2 are d-connected given $\mathcal{U}$; otherwise, say that $\mathcal{V}_1$ and $\mathcal{V}_2$ are d-separated given $\mathcal{U}$.

Let $\mathcal{V}$ be any set. Given $\mathscr{U} \subseteq \mathcal{P}(\mathcal{V})$, let $\mathsf{I}^{\mathscr{U}}$ denote the set of all pairwise disjoint triples $\langle \mathcal{V}_1, \mathcal{V}_2, \mathcal{W} \rangle$ such that $\mathcal{V}_1 \cup \mathcal{V}_2 \cup \mathcal{W} \subseteq \mathcal{U}$ for some $\mathcal{U} \in \mathscr{U}$. If $\mathcal{V} \in \mathscr{U}$, I will write $\mathsf{I}^{\mathcal{V}}$ instead of $\mathsf{I}^{\mathscr{U}}$ Let $|S|$ denote the cardinality of the set S. For each natural number $k \leq |\mathcal{V}|$, let $\mathscr{U}_k = \{\mathcal{U} \subseteq \mathcal{V} : |\mathcal{U}| \leq k\}$, and define $\mathsf{I}^k = \mathsf{I}^{\mathscr{U}_k}$.

Given $G \in \text{DAG}_{\mathcal{V}}$ and $\mathscr{U} \subseteq \mathcal{P}(\mathcal{V})$, let $\mathsf{I}^{\mathscr{U}}_G$ denote the set of all triples $\langle \mathcal{V}_1, \mathcal{V}_2, \mathcal{W} \rangle \in \mathsf{I}^{\mathscr{U}}$ such that $\mathcal{V}_1$ and $\mathcal{V}_2$ are d-separated in G given $\mathcal{W}$. Let $\mathsf{D}^{\mathscr{U}}_G$ denote the relative complement of $\mathsf{I}^{\mathscr{U}}_G$ in $\mathsf{I}^{\mathscr{U}}$. When $\mathcal{V} \in \mathscr{U}$, write I_G instead of $\mathsf{I}^{\mathscr{U}}_G$. If $\mathcal{V}_1$ and $\mathcal{V}_2$ are d-separated in G given $\mathcal{U}$, we say that G **satisfies** the triple $\langle \mathcal{V}_1, \mathcal{V}_2, \mathcal{U} \rangle$ (or that the triple **holds** in G). In the special case in which $\mathcal{V}_1$ and $\mathcal{V}_2$ are singletons $\{V_1\}$ and $\{V_2\}$ respectively, we write $\langle V_1, V_2, \mathcal{W} \rangle \in \mathsf{I}_G$ instead of $\langle \mathcal{V}_1, \mathcal{V}_2, \mathcal{W} \rangle \in \mathsf{I}_G$.

Typically, only *paths* are said to be d-connecting/active or not. In some of the theorems below, it will be helpful to consider active *variable sequences*, which may contain some vertex twice. Let α be a variable sequence with endpoints V_1 and V_2, and let $\mathcal{U} \subseteq \mathcal{V} \setminus \{V_1, V_2\}$. Then a vertex V_3 is **active on α in G given $\mathcal{U}$** just in case either

1. V_3 is not a collider on π and $V_3 \notin \mathcal{U}$
2. V_3 is a collider on π and either (i) $V_3 \in \mathcal{U}$ or (ii) there is $w \in Desc_G(V_3) \cap \mathcal{U}$ (or both).

The following lemma, which is a special case of Lemma 3.3.1. in Spirtes et al. [2000] (pp. 386), asserts that an active variable sequence indicates the existence of an active path with the same endpoints.

Lemma 1 *Let G be any* DAG*, and suppose that β is an active variable sequence given $\mathcal{U}$ with endpoints V and W. Then there is d-connecting path between V and W given $\mathcal{U}$.*

6 $\mathscr{E}$-equivalence

I say $G, H \in \text{DAG}_{\mathcal{V}}$ are **$\mathscr{U}$-equivalent** if $\mathsf{I}^{\mathscr{U}}_G = \mathsf{I}^{\mathscr{U}}_H$ and write $G \equiv_{\mathscr{U}} H$ in this case. When $\mathcal{V} \in \mathscr{U}$, I will write $G \equiv H$, and say that G and H are **I-equivalent**. Let $G \in \text{DAG}_{\mathcal{V}}$. Given a subset $\mathcal{E} \subseteq \mathcal{V}$, let $G \parallel \mathcal{E}$ be the graph obtained by removing from G all edges into the each variable in $\mathcal{E}$. I will use the script letter $\mathscr{E}$ to denote sets of pairs $\{\langle \mathcal{E}_1, \mathcal{U}_1 \rangle, \ldots, \langle \mathcal{E}_m, \mathcal{U}_m \rangle\}$ such that $\mathcal{E} \subseteq \mathcal{U} \subseteq \mathcal{V}$. I will call such pairs **experiments**, and I will say that $\mathcal{U}$ is **observed** and that the variables of $\mathcal{E}$ are subject to an **intervention**. Given $G, H \in \text{DAG}_{\mathcal{V}}$, write $G \equiv_{\mathscr{E}} H$ if $\mathsf{I}^{\mathcal{U}}_{G \parallel \mathcal{E}} = \mathsf{I}^{\mathcal{U}}_{H \parallel \mathcal{E}}$ for all $\langle \mathcal{E}, \mathcal{U} \rangle \in \mathscr{E}$. Let $[G]_{\mathscr{E}}$ denote the $\mathscr{E}$-equivalence class of G.

I will study the special case of $\mathscr{E}$-equivalence in which all subsets of k or fewer variables are observed, and all interventions of size $j \leq k$

are possible. To do so, define:

$$\mathscr{E}_{k,j} = \{\langle \mathcal{E}, \mathcal{U} \rangle \in \mathcal{V}^2 : \mathcal{E} \subseteq \mathcal{U} \text{ and } |\mathcal{U}| \leq k \text{ and } |\mathcal{E}| \leq j\}.$$

Write $G \equiv_{k,j} H$ if $G \equiv_{\mathscr{E}_{k,j}} H$. Similarly, let $[G]_{k,j}$ be the $\mathscr{E}_{k,j}$-equivalence class of G. Notice that $G \equiv_{k,j} H$ entails that $G \equiv_{k,l}$ for all $l \leq j$. When $j = 0$, I drop the subscript and write $G \equiv_k H$, $[G]_k$, and so on.

The following lemma will be essential. It is a generalization of Lemma 8 in [Mayo-Wilson, 2013].

Lemma 2 *Let $G \in \text{DAG}_{\mathcal{V}}$ be a graph with n vertices, and let $k < n$. Suppose that there are $k-1$ disjoint, directed paths from V_1 to V_2 in G. Moreover, suppose that V_1 is not a parent of V_2. Let H be the graph obtained by adding to G an edge from V_1 to V_2. Then $G \equiv_{k,j} H$ for all $j \leq k$.*

Proof: Let $\langle \mathcal{E}, \mathcal{U} \rangle$ be an experiment such that $|\mathcal{U}| \leq k$. It is necessary to show that $\mathsf{D}^{\mathcal{U}}_{G\|\mathcal{E}} = \mathsf{D}^{\mathcal{U}}_{H\|\mathcal{E}}$. First, suppose that $V_2 \in \mathcal{E}$. Then it follows that $G \parallel \mathcal{E} = H \parallel \mathcal{E}$, as the G and H differ by only one one edge that points into V_2. Hence, it immediately follows that $\mathsf{D}^{\mathcal{U}}_{G\|\mathcal{E}} = \mathsf{D}^{\mathcal{U}}_{H\|\mathcal{E}}$.

So suppose that if $V_2 \notin \mathcal{E}$. Then $H \parallel \mathcal{E}$ is the graph obtained by adding the $V_1 \to V_2$ edge to $G \parallel \mathcal{E}$. It follows that $\mathsf{D}^{k}_{G\|\mathcal{E}} \subseteq \mathsf{D}^{k}_{H\|\mathcal{E}}$. So it suffices to show that $\mathsf{D}^{k}_{H\|\mathcal{E}} \subseteq \mathsf{D}^{k}_{G\|\mathcal{E}}$. To this end, consider any triple $\langle Z_1, Z_2, \mathcal{W} \rangle$ in $\mathsf{D}^{k}_{H\|\mathcal{E}}$. By definition, there is a d-connecting path $\pi_{H\|\mathcal{E}}$ from Z_1 to Z_2 given $\mathcal{W}$ in $H \parallel \mathcal{E}$. I will construct a d-connecting path $\pi_{G\|\mathcal{E}}$ from Z_1 to Z_2 given $\mathcal{W}$ in $G \parallel \mathcal{E}$. To do so, let the $k-1$ distinct, directed paths from V_1 to V_2 be denoted δ_1 through δ_{k-1} respectively. The proof breaks into two cases, and each case has two subcases.

Case 1: Suppose there is some δ_i that contains no members of $\mathcal{E} \cup \mathcal{W}$. Now either $\pi_{H\|\mathcal{E}}$ is also a path in $G \parallel \mathcal{E}$ or it is not.

If $\pi_{H\|\mathcal{E}}$ is also a path in $G \parallel \mathcal{E}$. Then it's easy to show that $\pi_{G\|\mathcal{E}} = \pi_{H\|\mathcal{E}}$ is likewise d-connecting in $G \parallel \mathcal{E}$. Why? Every non-collider on $\pi_{G\|\mathcal{E}}$ is not a member of $\mathcal{W}$ because $\pi_{H\|\mathcal{E}}$ is active given $\mathcal{W}$ in $H \parallel \mathcal{E}$. Moreover, every collider C on $\pi_{G\|\mathcal{E}}$ is also a collider on $\pi_{H\|\mathcal{E}}$. Since $\pi_{H\|\mathcal{E}}$ is active given $\mathcal{W}$, it follows that either C or one of C's descendants in $H \parallel \mathcal{E}$ is a member of $\mathcal{W}$. But since $H \parallel \mathcal{E}$ is obtained from $G \parallel \mathcal{E}$ by adding an edge from V_1 to V_2, and V_1 is already an ancestor of V_2 in $G \parallel \mathcal{E}$ (as $G \parallel \mathcal{E}$ contains the directed path δ_i), it follows that the set of descendants of C in $H \parallel \mathcal{E}$ and in $G \parallel \mathcal{E}$ are identical.

If $\pi_{H\|\mathcal{E}}$ is not a path in $G \parallel \mathcal{E}$, then it must contain the edge $V_1 \to V_2$. Let β be the variable sequence obtained by replacing the edge $V_1 \to V_2$ with the directed path δ_i. In other words, define:

$$\beta = (\pi_{H\|\mathcal{E}} \downarrow V_1) \frown \delta_i \frown (\pi_{H\|\mathcal{E}} \uparrow V_2).$$

The sequence β is pictured in red below.

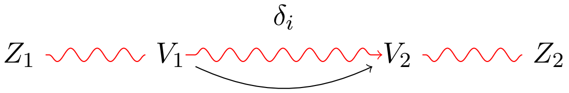


Figure 9:

Notice every variable on $(\pi_{H\|\mathcal{E}} \downarrow V_1)$ and $(\pi_{H\|\mathcal{E}} \uparrow V_2)$ other than V_1 and V_2 is active on β because it is active on $\pi_{H\|\mathcal{E}}$. Moreover, V_1 is a non-collider on both $\pi_{H\|\mathcal{E}}$ and β, and hence, it is active on both. Similarly, V_2 is a collider on β if and only if it is collider on $\pi_{H\|\mathcal{E}}$. Therefore, it is active given $\mathcal{W}$ on β because it is on $\pi_{H\|\mathcal{E}}$. So β is an active variable sequence given $\mathcal{W}$ between Z_1 and Z_2. By Lemma 1, there is an active path $\pi_{G\|\mathcal{E}}$ between Z_1 and Z_2 given $\mathcal{W}$.

Case 2: Suppose all of the paths $\delta_1, \delta_2, \ldots \delta_{k-1}$ contain some member of $\mathcal{E} \cup \mathcal{W}$. As the paths are disjoint by assumption, it follows that $\mathcal{E} \cup \mathcal{U}$ contains at least $k-1$ members. Since $Z_1, Z_2 \in \mathcal{U} \setminus \mathcal{W}$ and $\mathcal{E} \cup \mathcal{W} \subseteq \mathcal{U}$, it follows that either (a) $Z_1 \in \mathcal{E}$ and is a member of some path δ_i, or (b) $Z_2 \in \mathcal{E}$ and is a member of some path δ_i.

Case 2a: Suppose $Z_1 \in \mathcal{E}$ and is a member of some path δ_i. Then Z_1 is a descendant of V_1 in G. Again, either (i) $\pi_{H\|\mathcal{E}}$ is a path in $G \parallel \mathcal{E}$ or (ii) it is not.

Case 2ai: Suppose that $\pi_{H\|\mathcal{E}}$ is a path in $G \parallel \mathcal{E}$. I claim it is active given $\mathcal{W}$ in $G \parallel \mathcal{E}$. Suppose for the sake of contradiction it is not. Since $\pi_{H\|\mathcal{E}}$ is active given $\mathcal{W}$ in $H \parallel \mathcal{E}$, every non-collider on $\pi_{H\|\mathcal{E}}$ is not in $\mathcal{W}$. Hence, every such non-collider is also active given $\mathcal{W}$ in $G \parallel \mathcal{E}$. So if $\pi_{H\|\mathcal{E}}$ is not active, then there is some collider C on $\pi_{H\|\mathcal{E}}$ that is active in $H \parallel \mathcal{E}$ but not so in $G \parallel \mathcal{E}$. Let C be the collider closest to Z_1. Since $Z_1 \in \mathcal{E}$, it follows that the initial segment of $\pi_{H\|\mathcal{E}}$ points away from Z_1. Hence, as C is the closest collider to Z_1 on $\pi_{H\|\mathcal{E}}$, it follows that C is a descendant of Z_1 in $H \parallel \mathcal{E}$. As Z_1 occurs on a directed path from V_1 to V_2 in H by assumption of Case 2ai, it follows that C is also a descendant of V_1 in H.

Because C is active on $\pi_{H\|\mathcal{E}}$ in $H \parallel \mathcal{E}$ but not so in $G \parallel \mathcal{E}$, it follows that C has a descendant in $H \parallel \mathcal{E}$ that is not a descendant in $G \parallel \mathcal{E}$. Because $H \parallel \mathcal{E}$ is obtained from $G \parallel \mathcal{E}$ by adding the edge $V_1 \to V_2$, it must be the case that C is an ancestor of V_1. So C is an ancestor of V_1 in $H \parallel \mathcal{E}$, and hence, also in H. But I have already shown that C is a descendant of V_1 in H. So H contains a cycle, contradicting assumption.

Case 2aii: Now suppose $\pi_{H\|\mathcal{E}}$ is not a path in $G \parallel \mathcal{E}$. Then $\pi_{H\|\mathcal{E}}$ contains the edge $V_1 \to V_2$. Let $\beta = (\delta_i \uparrow Z_1) \frown (\pi_{H\|\mathcal{E}} \uparrow V_2)$. If β is active in $G \parallel \mathcal{E}$, then by Lemma 1, there is a d-connecting path given $\mathcal{W}$ between $G \parallel \mathcal{E}$.

If the variables in β are not adjacent in $G \parallel \mathcal{E}$, it follows that $\delta_i \uparrow Z_1$ contains some member W_i of $\mathcal{E} \cup \mathcal{W}$ other than Z_1. Thus:

$$W_i, Z_1 \in \mathcal{E} \subseteq \mathcal{E} \cup \mathcal{W} \subseteq \mathcal{U}.$$

Recall by assumption of Case 2, there are $k-2$ many other paths $\{\delta_j\}_{j \neq 2}$, each of which contains some member of $\mathcal{E} \cup \mathcal{W} \subseteq \mathcal{U}$. So $\mathcal{U}$ contains Z_1, Z_2, W_i and a distinct element W_j from each of the paths δ_j, where $j \neq i$. It follows that Z_2 is one of the W_l's, as otherwise $\mathcal{U}$ would contain $k+1$ many elements. Hence, Z_2 is an ancestor of V_2 and a descendant of V_1, and in particular, Z_2 does not equal V_1 or V_2.

Notice that V_2 is not equal to any of the elements of $\mathcal{U} = \{Z_1, Z_2, W_1, \ldots, W_{k-1}\}$. So, in particular, $V_2 \notin \mathcal{W} \subseteq \mathcal{U}$. Hence, V_2 is not a collider on $\pi_{H\|\mathcal{E}}$. Since $\pi_{H\|\mathcal{E}}$ contains the edge $V_1 \to V_2$, it follows that the path $\pi_{H\|\mathcal{E}} \uparrow V_2$ points away from V_2 and towards Z_2. Since Z_2 is an ancestor of V_2, the path $\pi_{H\|\mathcal{E}} \uparrow V_2$ is not directed from V_2 to Z_2. So it must contain a collider C which is closest to V_2; so C is a descendant of V_2. Because $\pi_{H\|\mathcal{E}}$ is active given $\mathcal{W}$ in $H \parallel \mathcal{E}$, it follows that either C or one of its descendants is in $\mathcal{W}$. Since $\mathcal{W} \subseteq \mathcal{U} \setminus \{Z_1, Z_2\}$ and $\mathcal{U} = \{Z_1, Z_2, W_1, \ldots, W_{k-1}\}$, it follows that C is either equal to W_j for some j, or is an ancestor of some W_j. In either case, C is an ancestor of V_2. But I have already shown that C is a descendant of V_2. So G contains a cycle, contradicting assumption.

Case 2b: Suppose $Z_2 \in \mathcal{E}$ and is a member of some path δ_i. Again, either (i) $\pi_{H\|\mathcal{E}}$ is a path in $G \parallel \mathcal{E}$ or (ii) it is not.

Case 2bi: This case is symmetric to Case 2ai.

Case 2bii: Suppose that $\pi_{H\|\mathcal{E}}$ is not a path in $G \parallel \mathcal{E}$, and hence, it contains the edge $V_1 \to V_2$. Since $Z_2 \in \mathcal{E}$, it follows that the tail of $\pi_{H\|\mathcal{E}}$ between V_1 and Z_2 points away from Z_2. Hence, there is a collider between V_1 and Z_2 on $\pi_{H\|\mathcal{E}}$. Let C be the collider closest to V_1, so that C is either V_2 or a descendant of V_2 in H.

Since C is active on $\pi_{H\|\mathcal{E}}$, it follows that C or one of its descendants is in $\mathcal{W}$ (and similarly, either V_2 or one of its descendants is in $\mathcal{W}$). Recall, by assumption of Case 2, each path δ_j contains at least one member from the set $\mathcal{E} \cup \mathcal{W}$. Since there are $k-2$ disjoint paths other than δ_i (i.e., the directed path containing Z_2), it follows that $\mathcal{U}$ contains a distinct element W_j from each path δ_j such that $j \neq i$. So $\mathcal{U}$ contains Z_1, Z_2, C and $k-2$ many elements W_j. It follows that

$Z_1 = W_j$ for some $j \neq i$, as otherwise, $\mathcal{U}$ would contain $k+1$ many elements.

Consider $\beta = (\delta_j \uparrow Z_1) \frown rev(\delta_i \uparrow Z_2)$, where rev reverses the order of the variables. Clearly, β has endpoints Z_1 and Z_2. Now if β is an active variable sequence given $\mathcal{U}$ in $G||\mathcal{E}$, then Lemma 1 entails the result.

Suppose for the sake of contradiction that β is not active. Since every variable on β other than V_2 is a non-collider, it follows that at least one of the four following cases holds: (1) at least one variable on $(\delta_j \uparrow Z_1)$, other than Z_1 and V_2 is a member of $\mathcal{E} \cup \mathcal{W}$, (2) at least one variable on $(\delta_i \uparrow Z_2)$ other than Z_2 and V_2 is a member of $\mathcal{E} \cup \mathcal{W}$, (3) $V_2 \in \mathcal{E}$, or (4) neither V_2 nor any of its descendants is a member of $\mathcal{W}$. I ruled out possibility (3) at the beginning of the proof, and possibility (4) contradicts the first sentence of the second to last paragraph.

So either (1) or (2) must hold. Consider (1) first, i.e., that $(\delta_j \uparrow Z_1)$ contains some member $W_l \in \mathcal{E} \cup S$ other than Z_1 or V_2. So $\mathcal{U}$ contains Z_1, Z_2, W_l, C and $k-3$ distinct elements W_m from each of the paths δ_m, where $m \neq i, j$. Since Z_1, Z_2 and W_l are pairwise distinct, it follows $C = W_m$ for some $m \neq i, j$, as otherwise $\mathcal{U}$ would contain at least $k+1$ many members. So C is on a directed path from V_1 to V_2 in H, and hence, an ancestor of V_2. But I have shown already that c is a descendant of V_2 in H. This is a contradiction. The proof that (2) leads to a contradiction is similar.

Theorem 4 *If $G \equiv_{2,1} H$ and there is a directed path from V_1 to V_2 in G, then there is likewise a directed path from V_1 to V_2 in H.*

Proof: Consider the experiment $\langle \mathcal{E}, \mathcal{U} \rangle := \langle \{V_1\}, \{V_1, V_2\} \rangle$. Since $G \equiv_{2,1} H$, it follows that $\mathsf{D}^{\mathcal{U}}_{G\|\mathcal{E}} = \mathsf{D}^{\mathcal{U}}_{H\|\mathcal{E}}$. By assumption, there is a directed path from V_1 to V_2 in G. So there is still a directed path from V_1 to V_2 in $G \parallel \mathcal{E}$, and that path is clearly d-connecting given the empty set. So $\langle V_1, V_2, \emptyset \rangle \in \mathsf{D}^{\mathcal{U}}_{G\|\mathcal{E}}$. Because $\mathsf{D}^{\mathcal{U}}_{G\|\mathcal{E}} = \mathsf{D}^{\mathcal{U}}_{H\|\mathcal{E}}$, it follows that there is a d-connecting path π between V_1 and V_2 in $H \parallel \mathcal{E}$ given the empty set. By definition of d-connecting, it follows that there are no colliders on π. Since all edges incident to V_1 in $H \parallel \mathcal{E}$ point "out of" V_1, it follows that the path π is out of V_1. Hence, because π contains no colliders and points away from V_1, it follows that π is a directed path from V_1 to V_2 as desired.

Theorem 5 *Let n, k be natural numbers such that $k < n$. Let $h = \lfloor \frac{n-2}{k-1} \rfloor$ and $M = (n-1) - h(k-1)$. Then there exist graphs G and H such that G is $k, 1$-equivalent to H, and yet H contains $f(n,k)$ edges that G does not, where*

1. *$f(n,k) = n - k$ if $h = 1$,*
2. *$f(n,k) = (2n+1) - 3k$ if $h = 2$, and*

3. $f(n,k) = (k-1)^2\binom{h-1}{2} + (k-1)(h-1) + Mh$ *if* $h \geq 3$.

Proof: Let $\mathcal{V}$ be a set with n many elements. Divide the vertices of $\mathcal{V}$ into groups as follows. Make h many groups of size $k-1$. Enumerate those groups as follows: $\{V_{1,1}, V_{1,2}, \ldots V_{1,k-1}\}$, $\{V_{2,1}, V_{2,2}, \ldots V_{2,k-1}\}, \ldots$ $\{V_{h,1}, \ldots, V_{h,k-1}\}$. There will be $n-(k-1)h = M+1$ vertices remaining. Denote one of those remaining vertices by $V_{0,1}$, and let the remainder be enumerated by $V_{h+1,1}, V_{h+1,2}, \ldots V_{h,M}$. To construct the graph G, place the variables in a matrix (as shown in Figure 10) so that $V_{r,c}$ is in the r^{th} row and c^{th} column. Draw an edge from every vertex in row $r+1$ to every vertex in row r. Call the resulting graph G. The graph H is a complete graph obtained by adding an edge from each vertex in row r to each vertex in row $s < r$ in G.

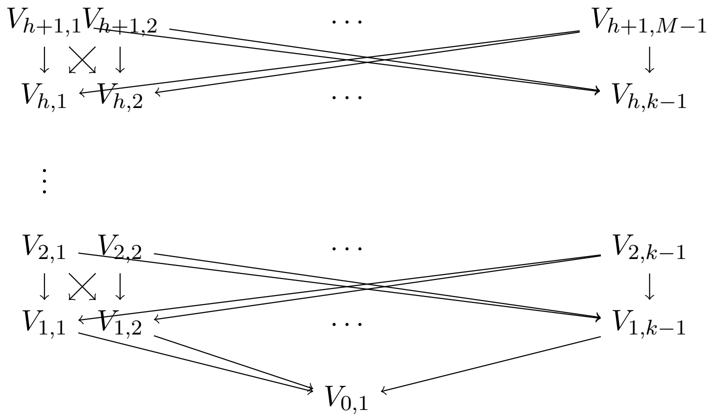


Figure 10: The graph described in Theorem 5

Notice that, if V is a vertex in row $r+2$ and W is a vertex in row r, then V and W are not adjacent and there are $k-1$ disjoint directed paths from V to W. Hence, by Lemma 2, the edge $V \to W$ can be added to G without breaking k,j-equivalence. Since H is the result of adding all such edges to G, it follows that $G \equiv_{k,j} H$.

□

Let G_k denote the DAG (pictured below) containing k vertices $\{X_1, X_2, W_1, \ldots W_k\}$ (1) an edge from X_1 to W_i for all $1 \leq i \leq k$, (2) an edge from W_i to X_2 for all $1 \leq i \leq k$, and (3) an edge from X_1 to X_2.

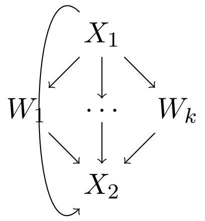


Figure 11: The graph denoted G_k in the remaining theorems

Theorem 6 *Suppose $k \geq 2$ and that G has fewer than $2k-2$ edges. Then $[G]_{k,1} = \{G\}$. Moreover, there is a graph H with exactly $2k-2$ edges such that $[H]_{k,j} = \{H_1, H_2\}$ for all $j \leq k$, and moreover, H_1 and H_2 are not I equivalent.*

Proof: Suppose $G \equiv_{k,1} H$ and G has fewer than $2k-2$ edges. Then $G \equiv_{k,0} H$, and so by Theorem 7 in [Mayo-Wilson, 2013], it follows that G and H are I-equivalent. Hence, they share the same skeleton by Verma and Pearl's theorem. So it suffices to show that all edges in G and H are oriented in the same direction. To do so, note that if $V_1 \to V_2$ is an edge in G, then trivially there is a directed path from V_1 to V_2 in G. By Theorem 4, it follows that there is a directed path from V_1 to V_2 in H. By acyclicity, it follows that the edge between V_1 and V_2 in H must be oriented as $V_1 \to V_2$.

For the second part of the theorem, let $H_1 = G_{k-1}$ and H_2 be the graph obtained by deleting the $X_1 \to X_2$ edge from H_1. By Lemma 2, it follows that $H_1 \equiv_{k,j} H_2$.

Theorem 7 *Fix a natural number k and any $j \leq g$, and let $p_{k,j}(n)$ be the fraction of* DAG*s G with $n \geq k$ many variables such that $[G]_{k,j}$ contains a graph that is not I-equivalent to G. Then $p_{k,j}(n) \to 1$ as $n \to \infty$.*

Proof: By Lemma 15 in [Mayo-Wilson, 2012], the proportion of DAGs containing an isomorphic copy of G_{k-1} approaches one as $n \to \infty$. If G contains a copy of G_{k-1}, then by Lemma 2, $[G]_{k,j}$ contains a graph that is not I-equivalent to G, namely, the graph in which the $X_1 \to X_2$ edge in G is removed. Hence, $p_{k,j}(n) \to 1$ as $n \to \infty$.

Theorem 8 *Fix any $k \in \mathbb{N}$ and any $j \leq k$. Let $E_{k,j}(G)$ be the maximum number m of* DAG*s $H_1, H_2, \ldots H_m$ such that $G \equiv_{k,j} H_i$ and $G \not\equiv H_i$ for all $i \leq m$. Let $E_{k,j}(n)$ be the average $E_{k,j}(G)$ for all* DAG*s G with n many variables. Then $E_{k,j}(n) \to \infty$ as $n \to \infty$.*

Proof: Let $m \in \mathbb{N}$. Let H be a DAG containing m disjoint copies of G_{k-1}, and let H_i be the result of removing the $X_1 \to X_2$ edge from the i^{th} copy. By Lemma 2, $G \equiv_{k,j} H_i$ for all $i \leq m$, but $G \not\equiv H_i$ as their skeletons differ. Thus, $E_{k,j}(H) \geq m$. By Lemma 15 in [Mayo-Wilson, 2012], the proportion of DAGs containing m disjoint copies of G_{k-1} approaches one as $n \to \infty$. Since m was arbitrary, $E_{k,j}(n) \to \infty$.

7 Bayesian Networks

7.1 Markov Equivalence

In this section, random variables will be denoted by the capital letters X, Y, and Z, and values of random variables will be denoted x, y, z. Vectors will be bolded. So $\boldsymbol{X}$ will represent a vector of random variables, and $\boldsymbol{x}$ is a vector of values of $\boldsymbol{X}$. Sets of random variables will be denoted in a scripted font, e.g., $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$. Given some ordering of the random variables $\mathcal{X}$, let $\boldsymbol{\mathcal{X}}$ denote the random vector obtained by ordering those variables.

Given a probability measure p, write $p \models \mathcal{X} \perp\!\!\!\perp \mathcal{Y} | \mathcal{Z}$ if the variables $\mathcal{X}$ and $\mathcal{Y}$ are conditionally independent given $\mathcal{Z}$ with respect to the measure p. Let $\mathcal{V}$ be any finite set and $\mathcal{X} = \{X_V\}_{V \in \mathcal{V}}$ be a collection of random variables indexed by $\mathcal{V}$. A **Bayesian network** over $\mathcal{X}$ is a pair $\langle G, p \rangle$, where $G \in \text{DAG}_{\mathcal{X}}$ such that $p \models \mathcal{W} \perp\!\!\!\perp \mathcal{Y} | \mathcal{Z}$ if and only if $\langle \mathcal{W}, \mathcal{Y}, \mathcal{Z} \rangle \in \mathsf{I}_G$. For any graph $G \in \text{DAG}_{\mathcal{V}}$, let $\mathbb{P}_G^{\mathcal{X}}$ be the set of probability measures p such that $\langle \mathcal{X}, G^{\mathcal{X}}, p \rangle$ is a Bayesian network, where $G^{\mathcal{X}}$ is the DAG obtained by replacing V with X_V for all $V \in \mathcal{V}$. Say that two graphs $G, H \in \text{DAG}_{\mathcal{V}}$ are **Markov equivalent** if $\mathbb{P}_G^{\mathcal{X}} = \mathbb{P}_H^{\mathcal{X}}$ for any collection of random variables $\mathcal{X}$ indexed by $\mathcal{V}$.

Given $p \in \mathbb{P}_G^{\mathcal{X}}$ and some $\mathcal{U} \subseteq \mathcal{V}$, let $p_{\mathcal{U}}$ denote the marginal distribution of p over $\{X_U \in \mathcal{X} : U \in \mathcal{U}\}$. Given $\mathscr{U} \subseteq \mathcal{P}(\mathcal{V})$, let $\mathbb{P}_G^{\mathscr{U}}$ be the set $\{\{p_{\mathcal{U}}\}_{\mathcal{U} \in \mathscr{U}} : p \in \mathbb{P}_G^{\mathcal{X}}\}$. Say $G, H \in \text{DAG}_{\mathcal{V}}$ are $\mathscr{U}$**Markov-equivalent** if $\mathbb{P}_G^{\mathscr{U}} = \mathbb{P}_H^{\mathscr{U}}$ with respect to all collections of random variables $\mathcal{X}$. Write $G \approx_{\mathscr{U}} H$ in this case.

Theorem 9 *If G and H are $\mathscr{U}$-Markov equivalent, they are also $\mathscr{U}$ equivalent. The converse is false in a fairly strong sense: for all $\mathcal{V}$ and all $\mathscr{U}$ not containing $\mathcal{V}$, there exist G and H such that G and H are $\mathscr{U}$-equivalent but not $\mathscr{U}$ Markov equivalent.*

Proof: The first claim is trivial. To show the converse is false, let $G = G_b$ be the graph pictured in Figure 11. Let H be the graph obtained by adding the $X_1 \to X_2$ edge to G.

Suppose $\mathcal{X}$ consists exclusively of binary random variables, and define a Bayesian network $\langle \mathcal{X}, G, p \rangle$ satisfying the following conditions:

- $p(X_1 = 0) = \frac{1}{2}$,

- $p(W_i = 0|X_1 = 0) = 1$ and $p(W_i = 0|X_1 = 1) = \frac{1}{2}$ for all $1 \leq i \leq n-2$, and
- $p(X_2 = 1|\boldsymbol{W} = \boldsymbol{w}) = \frac{2^{\Sigma \boldsymbol{w}} - 1}{2^{\Sigma \boldsymbol{w}}}$.

where $\boldsymbol{W} = \langle W_1, W_2, \ldots, W_{n-2} \rangle$ and $\Sigma \boldsymbol{w}$ is the number of non-zero coordinates in $\boldsymbol{w}$. Below, I show that the distribution p is faithful to G. Before doing so, let q be any distribution over $\mathcal{X}$ that agrees with p on all of the marginal distributions over any proper subset of variables of $\mathcal{X}$. I show that $q \models X_1 \perp\!\!\!\perp X_2 | \{W_1, \ldots, W_{n-2}\}$, and hence, q is not faithful to H. By definition, this entails that G and H are not $\mathscr{U}$Markov equivalent.

To prove $q \models X_1 \perp\!\!\!\perp X_2 | \{W_1, \ldots, W_{n-2}\}$, it is necessary to show

$$q(X_2 = x_2 | X_1 = x_1, \boldsymbol{W} = \boldsymbol{w}) = q(X_2 = x_2 | \boldsymbol{W} = \boldsymbol{w})$$

for any set of values $x_1, x_2, \boldsymbol{w}$. There are two cases to consider:

Case 1: Assume that at least one coordinate of $\boldsymbol{w}$ is equal to one.

$$\begin{aligned}
q(X_2 = x_2|\boldsymbol{W} = \boldsymbol{w}) &= q(X_2 = x_2|\boldsymbol{W} = \boldsymbol{w}, X_1 = 1) \cdot q(X_1 = x_1|\boldsymbol{W} = \boldsymbol{w}) \\
&+ q(X_2 = x_2|\boldsymbol{W} = \boldsymbol{w}, X_1 = 0) \cdot q(X_1 = x_0|\boldsymbol{W} = \boldsymbol{w}) \\
&\quad \text{by total probability} \\
&= q(X_2 = x_2|\boldsymbol{W} = \boldsymbol{w}, X_1 = 1) \cdot p(X_1 = 1|\boldsymbol{W} = \boldsymbol{w}) \\
&+ q(X_2 = x_2|\boldsymbol{W} = \boldsymbol{w}, X_1 = 0) \cdot p(X_1 = 0|\boldsymbol{W} = \boldsymbol{w}) \\
&\quad \text{as } p\&q \text{ agree on all marginal distributions over } \mathcal{X} \\
&= q(X_2 = x_2|\boldsymbol{W} = \boldsymbol{w}, X_1 = 1) \cdot 1 + q(X_2 = x_2|\boldsymbol{W} = \boldsymbol{w}, X_1 = 0) \cdot 0 \\
&\quad \text{by definition of } p \text{ as there is at least one coordinate of } \boldsymbol{w} \text{ equal to one} \\
&= q(X_2 = x_2|\boldsymbol{W} = \boldsymbol{w}, X_1 = 1)
\end{aligned}$$

Case 2: Assume that all coordinates of $\boldsymbol{w}$ are equal to zero, i.e., $\boldsymbol{w} = \mathbf{0}$
There are two subcases to consider.

Case 2a: Suppose $x_2 = 0$. Then:

$$\begin{aligned}
1 &= p(X_2 = x_2 = 0|\boldsymbol{W} = \mathbf{0}) = q(X_2 = x_2|\boldsymbol{W} = \mathbf{0}) \\
&\quad \text{as } p\&q \text{ agree on all marginal distributions over } \mathcal{X} \\
&= q(X_2 = x_2|\boldsymbol{W} = \boldsymbol{w}, X_1 = 1) \cdot q(X_1 = 1|\boldsymbol{W} = \mathbf{0}) \\
&+ q(X_2 = x_2|\boldsymbol{W} = \boldsymbol{w}, X_1 = 0) \cdot q(X_1 = 0|\boldsymbol{W} = \mathbf{0}) \\
&\quad \text{by total probability} \\
&= q(X_2 = x_2|\boldsymbol{W} = \boldsymbol{w}, X_1 = 1) \cdot q(X_1 = 1|\boldsymbol{W} = \mathbf{0}) \\
&+ q(X_2 = x_2|\boldsymbol{W} = \boldsymbol{w}, X_1 = 0) \cdot (1 - q(X_1 = 1|\boldsymbol{W} = \mathbf{0})) \\
&\quad \text{by properties of conditional probability} \\
&= q(X_2 = x_2|\boldsymbol{W} = \boldsymbol{w}, X_1 = 1) \cdot p(X_1 = 1|\boldsymbol{W} = \mathbf{0}) \\
&+ q(X_2 = x_2|\boldsymbol{W} = \boldsymbol{w}, X_1 = 0) \cdot (1 - p(X_1 = 1|\boldsymbol{W} = \mathbf{0})) \\
&\quad \text{as } p\&q \text{ agree on all marginal distributions over } \mathcal{X}
\end{aligned}$$

As $1 > p(X_1 = 1|\boldsymbol{W} = \mathbf{0}) > 0$, the last equation holds if and only if:

$$1 = q(X_2 = x_2|\boldsymbol{W} = \boldsymbol{w}, X_1 = 1) = q(X_2 = x_2|\boldsymbol{W} = \boldsymbol{w}, X_1 = 0)$$

and hence,

$$q(X_2 = x_2|\boldsymbol{W} = \mathbf{0}) = q(X_2 = x_2|\boldsymbol{W} = \boldsymbol{w}, X_1 = 1) = q(X_2 = x_2|\boldsymbol{W} = \boldsymbol{w}, X_1 = 0).$$

Case 2b: Suppose $x_1 = 0$. Then $p(X_2 = x_2 = 1|\boldsymbol{W} = \mathbf{0}) = 0$ by definition of p, and so

$$\begin{aligned}
0 &= p(X_2 = x_2 = 1|\boldsymbol{W} = \mathbf{0}) \\
&= q(X_2 = x_2 = 1|\boldsymbol{W} = \mathbf{0}) \\
&\quad \text{as } p\&q \text{ agree on all marginal distributions over } \mathcal{X} \\
&= q(X_2 = x_2|\boldsymbol{W} = \boldsymbol{w}, X_1 = 1) \cdot p(X_1 = 1|\boldsymbol{W} = \mathbf{0}) \\
&+ q(X_2 = x_2|\boldsymbol{W} = \boldsymbol{w}, X_1 = 0) \cdot (1 - p(X_1 = 1|\boldsymbol{W} = \mathbf{0})) \\
&\quad \text{ by the same reasoning as in Case 2a.}
\end{aligned}$$

As$1 > p(X_1 = 1|\boldsymbol{W} = \mathbf{0}) > 0$, the last equation holds if and only if:

$$0 = q(X_2 = x_2|\boldsymbol{W} = \boldsymbol{w}, X_1 = 1) = q(X_2 = x_2|\boldsymbol{W} = \boldsymbol{w}, X_1 = 0)$$

and hence,

$$q(X_2 = x_2|\boldsymbol{W} = \mathbf{0}) = q(X_2 = x_2|\boldsymbol{W} = \boldsymbol{w}, X_1 = 1) = q(X_2 = x_2|\boldsymbol{W} = \boldsymbol{w}, X_1 = 0).$$

To show p is faithful to G, note that all triples *not* entailed by G fall into one of the following five categories:

1. $X_1 \perp\!\!\!\perp X_2|\mathcal{Z}$ where $\mathcal{Z} \subsetneq \{W_1, W_2, \ldots W_{n-2}\}$,

2. $X_1 \perp\!\!\!\perp W_i|\mathcal{Z}$ where $\mathcal{Z} \subseteq \mathcal{X} \setminus \{X_1, W_i\}$ and $1 \leq i \leq n-2$,
3. $X_2 \perp\!\!\!\perp W_i|\mathcal{Z}$ where $\mathcal{Z} \subseteq \mathcal{X} \setminus \{X_2, W_i\}$ and $1 \leq i \leq n-2$,
4. $W_i \perp\!\!\!\perp W_j|\mathcal{Z}$ where $\mathcal{Z} \subseteq \mathcal{X} \setminus \{X_1, X_2, W_i, W_j\}$ and $1 \leq i, j \leq n-2$,
5. $W_i \perp\!\!\!\perp W_j|\mathcal{Z}$ where $X_2 \in \mathcal{Z} \subseteq \mathcal{X} \setminus \{W_i, W_j\}$ and $1 \leq i, j \leq n-2$.

I now show that p likewise does not entail any of the above conditional independences.

Category 1: It must be shown that $p \not\models X_1 \perp\!\!\!\perp X_2|\mathcal{Z}$ where $\mathcal{Z} \subsetneq \{W_1, W_2, \ldots W_{n-2}\}$. By definition, $p(X_2 = 0|X_1 = 0, \boldsymbol{\mathcal{Z}} = \mathbf{0}) = 1$ whereas $p(X_2 = 0|\boldsymbol{\mathcal{Z}} = \mathbf{0}) < 1$ because there is some $W_i \notin \mathcal{Z}$.

Category 2: It must be shown that $p \not\models X_1 \perp\!\!\!\perp W_i|\mathcal{Z}$ where $\mathcal{Z} \subseteq \mathcal{X} \setminus \{X_1, W_i\}$. This is similar to the last case. Note, by definition of p, it follows that $p(W_i = 0|X_1 = 0, \boldsymbol{\mathcal{Z}} = \mathbf{0}) = 1$, whereas $p(W_i = 0|\boldsymbol{\mathcal{Z}} = \mathbf{0}) < 1$.

Category 3: It must be shown that $p \not\models X_2 \perp\!\!\!\perp W_i|\mathcal{Z}$ where $\mathcal{Z} \subseteq \mathcal{X} \setminus \{X_2, W_i\}$. To do so, it suffices to show $p(X_2 = 1|W_i = 1, \boldsymbol{\mathcal{Z}} = \mathbf{0}) \neq p(X_2 = 1|\boldsymbol{\mathcal{Z}} = \mathbf{0})$. If $X_1 \in \mathcal{Z}$, this is trivial because the conditional probability on the left hand-side is undefined whereas that on the right is positive. If $X_1 \notin \mathcal{Z}$, it is tedious but routine to verify that $p(X_2 = 1|W_i = 1, \boldsymbol{\mathcal{Z}} = \mathbf{0}) > p(X_2 = 1|\boldsymbol{\mathcal{Z}} = \mathbf{0})$ using (1) the definition of p, (2) the the law of total probability, and (3) the fact that $p(\boldsymbol{\mathcal{W}} = \boldsymbol{w}|X_1 = 1) = \prod_{W_j \in \mathcal{W}} p(W_j = \boldsymbol{w}_j|X_1 = 1) = \frac{1}{2^{|\mathcal{W}|}}$ for all $\mathcal{W} \subseteq \{W_1, W_2, \ldots W_{n-2}\}$, which is an instance of the factorization property for Bayesian networks.

Category 4: We must show $p \not\models W_i \perp\!\!\!\perp W_j|\mathcal{Z}$ where $\mathcal{Z} \subseteq \mathcal{X} \setminus \{X_1, X_2, W_i, W_j\}$. To do so, use the three facts described in Category 3 to show $p(W_i = 0|W_j = 0, \boldsymbol{\mathcal{Z}} = \mathbf{0}) < p(W_i = 0|\boldsymbol{\mathcal{Z}} = \mathbf{0})$.

Category 5: It must be shown that $p \not\models W_i \perp\!\!\!\perp W_j|\mathcal{Z}$ where $X_2 \in \mathcal{Z}$ and $\mathcal{Z} \subseteq \mathcal{X} \setminus \{W_i, W_j\}$. To do this, use the three facts described in Category 3 $p(W_i = 1|W_j = 0, \boldsymbol{\mathcal{Z}} \setminus \boldsymbol{\{X_2\}} = \mathbf{0}, X_2 = 1) > p(W_i = 0|\boldsymbol{\mathcal{Z}} \setminus \boldsymbol{\{X_2\}} = \mathbf{0}, X_2 = 1)$.

7.2 Stronger Distributional Assumptions

Write $G \approx_{\mathscr{U}}^{\boldsymbol{M}} H$ if G and H are $\mathscr{U}\boldsymbol{M}$ indistinguishable. In the special case in which $\mathscr{U}$ is all subsets of $\mathcal{V}$ of a fixed size k, write $G \approx_{k}^{\boldsymbol{M}} H$. When $\mathcal{V} \in \mathscr{U}$, we drop the subscript $\mathscr{U}$ and write $G \approx^{\boldsymbol{M}} H$.

Theorem 12 *There are G and H such that $G \equiv_{2,0} H$ but $G \not\approx_{2}^{\boldsymbol{D}^+} H$.*

Proof: Let G be the graph $X \rightarrow Y \rightarrow Z$, and let H be the graph $X \rightarrow Z \rightarrow Y$. Clearly, $G \equiv_2 H$. Suppose for the sake of contradiction that $G \approx_2^{\boldsymbol{D}^+} H$.

Suppose X, Y, and Z are all binary random variables, and consider any two discrete, multinomial models $\langle G, p \rangle$ and $\langle H, q \rangle$. Since $G \approx_2^{\boldsymbol{D}^+} H$, the correlations of any pair of variables with respect to p and q are identical. Hence, I write ρ_{VW} to indicate the correlation between $V, W \in \{X, Y, Z\}$ in the two discrete models.

By theorem 2.12 of Danks and Glymour [2001], the correlation between any two variables in singly connected graphs containing only binary variables is the product of the correlations along the unique trek connecting them. Since G contains the trek $X \rightarrow Y \rightarrow Z$, it follows $\rho_{XZ} = \rho_{XY} \cdot \rho_{YZ}$. By the same theorem, since H contains the trek $X \rightarrow Z \rightarrow Y$, it follows that $\rho_{XY} = \rho_{XZ} \cdot \rho_{YZ}$. So $\rho_{YZ}^2 = 1$, or in other words, Y and Z are perfectly correlated. So the distributions p and q are not positive, contradicting assumption.

□

The last proof cannot be generalized straightforwardly to either (a) discrete variables taking more than two values, or (b) $\boldsymbol{D}^+k$-equivalence when $k > 2$. The former generalization is not straightforward because the the theorem that correlations can be multiplied along treks only holds for binary variables. The latter generalization is difficult because the same theorem only applies to singly connected networks, and two graphs that are $k, 0$-equivalent but not I-equivalent will often contain multiple treks between two variables (because they will generally contain different orientations).

Theorem 13 *For all $\mathcal{V}$ and all $\mathscr{U} \subseteq \mathcal{V}$ such that $\mathcal{V} \notin \mathscr{U}$, there exist G and H such that $G \equiv_{\mathscr{U}} H$ but $G \not\approx_{\mathscr{U}}^{\boldsymbol{N}^\vee} H$.*

Proof: The proof is the same as that of Theorem 9 because the distribution p in that proof is a noisy or parametrization. In greater detail, for a graph with n many variables, enumerate the variables $\mathcal{V} = \{X_1, X_2, W_1, \ldots W_{n-2}\}$. Consider the graph G_{n-2} with the latent "noise" terms $\{E_V : V \in \mathcal{V}\} \cup \{B_{U,V} : U \rightarrow V \text{ is an edge in } G_{n-2}\}$ as shown in **Figure 12**.

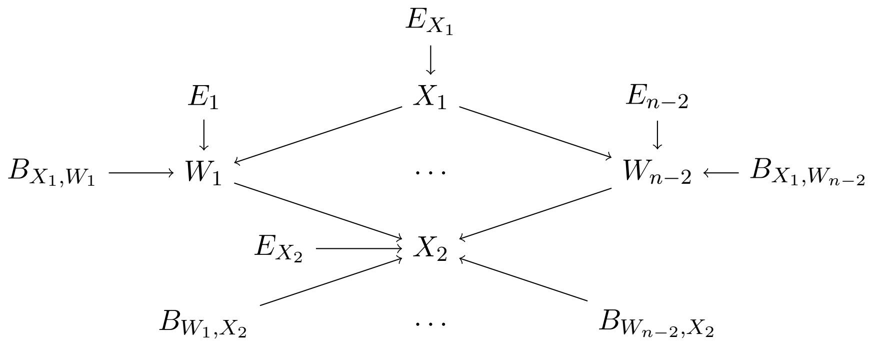


Figure 12: The graph denoted G_{n-2} with noise terms, as described in Theorem 13

Let p be the unique noisy-or parameterization of G_{n-2} such that (1) $p(E_{X_1}) = p(E_{X_2}) = 1/2$, (2) $p(E_{W_i} = 1) = 0$ for all W_i, and (3) $p(B_{U,V} = 1) = 1/2$ if $U \to V$ is an edge in G_{n-2}. Then

- $p(X_1 = 0) = \frac{1}{2}$ because $p(X_1 = 0) = p(E_{X_1} = 0) = \frac{1}{2}$.
- $p(W_i = 0|X_1 = 0) = 1$ and $p(W_i = 0|X_1 = 1) = \frac{1}{2}$ for all W_i. The former equation holds because $p(W_i = 0|X_1 = 0) = p(E_{W_i} = 0) = 1$ and the latter holds because $p(E_{W_i} = 1) = 0$ and $p(W_i = 0|X_1 = 1) = p(B_{X_1,W_i} = 1) = \frac{1}{2}$.

So to show p is the same distribution as in the proof of Theorem 9, we need to show only that $p(X_2 = 1|\boldsymbol{W} = \boldsymbol{w}) = \frac{2^{\Sigma \boldsymbol{w}}-1}{2^{\Sigma \boldsymbol{w}}}$. To do that, we rely on the following lemma, which can be proven by induction on k.

Lemma 3 $\sum_{m=1}^{k}(-1)^{m+1} \cdot \binom{k}{m}\frac{1}{2^m} = \frac{2^k-1}{2^k}$ *for all* $k \geq 1$.

With that lemma, let $\mathcal{Z}$ be the parents of X_2 including the error terms (i.e., $\mathcal{Z} = \{E_{X_2}\} \cup \{W_j : \boldsymbol{w}_j = 1\}$), and suppose J has size k. Given $m \leq k$, let $\mathcal{Z}_m = \{\mathcal{U} \subseteq \mathcal{Z} : |\mathcal{U}| = m\}$ be all subsets of $\mathcal{Z}$ of size m.

Then:

$$\begin{aligned}
p(X_2 = 1|\boldsymbol{W} = \boldsymbol{w}) &= p(\bigcup_{Z \in \mathcal{Z}} Z = 1) \\
&= \sum_{m \leq k} (-1)^{m+1} \cdot \sum_{\mathcal{U} \in \mathcal{Z}_m} p\left(\bigcap_{U \in \mathcal{U}} U = 1\right) \\
&\quad \text{by the inclusion exclusion principle} \\
&= \sum_{m \leq k+1} (-1)^{m+1} \sum_{\mathcal{U} \in \mathcal{Z}_m} \frac{1}{2^m} \\
&\quad \text{because } E_{X_2}, B_{W_1,X_2}, \ldots B_{W_{n-2},X_2} \text{ are mutually independent} \\
&= \sum_{m \leq k+1} (-1)^{m+1} \cdot \binom{k}{m} \cdot \frac{1}{2^m} \\
&= \frac{2^k - 1}{2^k} \text{ by the lemma}
\end{aligned}$$