

# The Limits of Piecemeal Causal Inference

Conor Mayo-Wilson

## Abstract

In medicine and the social sciences, researchers must frequently integrate the findings of many observational studies, which measure overlapping collections of variables. For instance, learning how to prevent obesity requires combining studies that investigate obesity and diet with others that investigate obesity and exercise. Recently developed causal discovery algorithms provide techniques for integrating many studies, but little is known about what can be learned from such algorithms. This paper argues that there are causal facts that one could learn by conducting a large study but which could not be learned by combining many smaller studies. Moreover, I characterize the frequency with which combining many studies increases underdetermination and exactly how much information is lost.

## Introduction

Since its inception, the *Journal of the American Medical Association* has published over 270,000 papers concerning the causes of heart disease. The enormous number of papers is, in part, a consequence of the enormous number of factors - diet, exercise, prescription drug use, and many others - that are potentially relevant to cardiovascular health. No single study or randomized controlled trial (RCT) could measure all the variables relevant to heart disease. Thus, heart disease must be investigated through a series of smaller studies or RCTs (e.g., one examining heart disease and obesity, another investigating heart disease and smoking, etc.). I refer to this practice of combining many small studies or RCTs as **the piecemeal construction of causal theories**. The piecemeal construction of causal theories raises a crucial question: how can so many observational studies and RCTs, each measuring a different set of variables, be integrated into a *single* causal theory concerning all variables under investigation?

Employing techniques from automated causal discovery, Danks (2002), Tillman, Danks, and Glymour (2008), Eberhardt, Hoyer, and Scheines (2010), Tillman and Spirtes (2011) develop algorithms that provide an answer to this crucial question: given a collection of data sets containing measurements of differing (but often overlapping) sets of variables, these algorithms output all causal hypotheses that are consistent with the data. Yet little is known about what can, in principle, be learned from such algorithms. For instance, are there any causal facts that one could learn by conducting a large study but which could not be learned by combining many smaller studies? If so, what *type* of causal information is lost in the piecemeal construction of causal theories? How *often* is such information lost? The purpose of this paper is to provide preliminary answers to these questions.

The structure of this paper is as follows. In the first section, I state and explain two principles, called the Causal Markov condition (CMC) and the Causal Faithfulness condition (CFC) respectively, that are frequently employed in automated causal discovery algorithms. In particular, both principles are assumed in all of the existing algorithms for the piecemeal construction of causal theories.

Although both principles are controversial,<sup>1</sup> I assume both without argument. Because my goal is to investigate what can be learned from the piecemeal construction of causal theories, I must make some assumptions concerning how causal hypotheses are inferred from statistical data. As the CMC and CFC are employed widely both in causal discovery and (often implicitly) in the sciences, it is important to characterize what can be learned via piecemeal methods when the two principles are assumed. Future work ought to characterize what can be learned from the piecemeal construction of causal theories under weaker or entirely different assumptions; some alternative principles are discussed in the final section of the paper.

In the second section, I argue that the piecemeal construction of causal theories can create a *problem of piecemeal induction*.<sup>2</sup> The problem is as follows: for any collection of variables, there are distinct causal theories  $T_1$  and  $T_2$  that can be distinguished by observational data if and only if *all*

---

<sup>1</sup>For defenses of the CMC, see Hausman and Woodward (2002, 2004) and Steel (2005); for criticisms, see Cartwright (2002, 2007). For criticisms of CFC, see Freedman and Humphreys (1999) and Cartwright (2007). The case study involving birth control and thrombosis, cited as a counterexample to CFC, is discussed in Hesslow (1976) and Cartwright (1989). Both the CMC and CFC are defended in Spirtes et. al. (2000).

<sup>2</sup>The problem was first discussed in Mayo-Wilson (2010).

variables are *simultaneously* measured. That is, for any collection of variables under investigation, measuring only subsets of the collection (no matter how many) can fail to reveal the full causal structure. I discuss the problem of piecemeal induction as it pertains to causal inference from *observational data*, but it is easy to see that analogous epistemological problems arise in inference from experimental data (e.g. like that obtained in RCTs).

Section 3 addresses three important questions that are raised by the problem of piecemeal induction. First, what *type* of information is lost in the piecemeal construction of causal theories, and how *much* is lost? Second, how *often* does the problem arise? That is, what types of causal structures require scientists to conduct large observational studies, and how frequently do researchers confront said structures? Third, when, if ever, is no information lost in integrating many observational studies? I state and explain six new theorems that provide partial answers to each of the three questions.<sup>3</sup> I conclude with a description of open problems that are important for piecemeal causal discovery.

Although this paper focuses on causal inference from many observational studies (in medicine and the social sciences, in particular), the six new theorems ought to be seen as part of a larger project investigating the frequency, extent, and character of underdetermination of theories in science more generally. The theorems show that, in causal inference, underdetermination is sometimes severe, in the sense that the theories underdetermined by evidence differ greatly and in important ways (see Theorems 6 and 7). In other circumstances, causal theories are *not* underdetermined in any significant sense (Theorem 5), if at all (Theorem 1). In domains in which the variables under investigation bear numerous intricate relationships to one another, underdetermination will be frequent (Theorem 9), but in other domains, scientists might never encounter underdetermination (Theorem 8). In short, sweeping arguments about the presence or absence of underdetermination in science need to be reconsidered.<sup>4</sup>

---

<sup>3</sup>Proofs of all theorems are available in Appendix B.

<sup>4</sup>Here, I am thinking of arguments that purport to show that any scientific theory has empirically equivalent rivals (e.g. see van Fraassen (1980)), or which attempt to categorically deny this claim (e.g. Laudan and Leplin (1991)). Of course, this paper discusses underdetermination in a context in which all alternative theories can be precisely enumerated, in which all possible future data/evidence can be precisely described, and when the relationship between said theories and evidence is mathematically determined. Anti-realists might plausibly argue that underdetermination is rampant in domains in which there are “unconceived alternatives” (Stanford, 2006), but of course, there are plenty

# 1 Causal Inference From Observational Data

The central problem for causal inference, made famous by Hume, is that causation is unobservable.<sup>5</sup> For example, one does not observe the event “overeating causing obesity.” Rather, one observes two distinct events - eating at some time  $t_0$  and weight gain at some later time  $t_1$  - and infers some connection between the two. Since there is no observable “causal event”, causal relationships must be inferred from *probabilistic/statistical* regularities between types of events, like overeating and weight gain.

However, not every statistical regularity indicates a causal relationship. Arm length and height, for instance, are clearly correlated, and yet neither is a cause of the other (as otherwise, one could get taller by stretching the arms or vice versa). Rather, the two quantities are correlated because they share several common (genetic and environmental) causes. So the central problem for causal inference reduces to another: characterize the types of correlations or probabilistic regularities that arise from genuinely causal relationships, and characterize those that arise from “spurious” factors, like chance or unmeasured common causes. Such a characterization would be useful even if there were “observable” causal events, as then one could still infer causal facts in the presence of unmeasured, confounding variables.

In a variety of scientific disciplines, two principles are often assumed to axiomatize (at least in part) the relationship between probability and causation, and hence, they have been employed extensively in drawing causal conclusions from probabilistic data.

- **Causal Markov Condition:** (CMC) Any variable is conditionally independent of its non-effects given its direct causes.
- **Causal Faithfulness Condition:** (CFC) No two variables are conditionally independent unless so entailed by the CMC.

Here, “conditional independence” refers to the standard probabilistic notion of independence. Informally, two events  $A$  and  $B$  are conditionally inde-

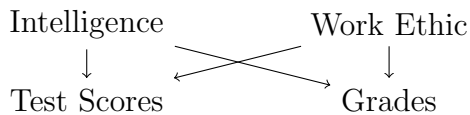
---

of cases in science in which the set of possible statistical models/theories is well-specified.

<sup>5</sup>Recent work in the psychology of “causal perception” suggests that Hume’s claim might need to be modified, as there might be a sense in which humans actually “perceive” causation between two objects that interact locally. Because not all causal relationships involve the interaction between spatio-temporally contiguous objects (e.g. money supply increases inflation), I assume that there are important causal relations (especially in medicine and the social sciences) that cannot be learned by direct observation alone.

pendent given  $C$  if, from a predictive standpoint,  $C$ 's occurrence renders  $B$  irrelevant in predicting whether or not  $A$  will occur.<sup>6</sup> For example, the event  $A$  “having blonde hair” is unconditionally dependent on  $B$  “having blue eyes” as one could more accurately predict an individual’s eye color if one were given his hair color (because, for example, blue eyes are more common among individuals with blonde hair). However, these two events are conditionally independent given a third event  $C$ , namely, having the recessive gene pair necessary for blue eyes. Why? Knowing that an individual has blonde hair is irrelevant to predicting his eye color if one already knows the individual has the genes responsible for blue eyes.

The CMC and CFC assert that conditional independence and causation are intimately connected. To see why, let’s consider an example. Suppose that work ethic and intelligence are causes both of one’s performance on standardized tests and also of one’s high school grades. These causal relationships can be represented in a *directed graph* like the one below, where an arrow of the form  $v \rightarrow v'$  indicates that  $v$  directly causes  $v'$ . In general, given a set of variables  $V$ , define a *causal theory* to be a directed graph with nodes from  $V$ . Of course, the causal theories of interest to scientists and policy-makers are more detailed than directed graphs, as graphs only indicate which variables cause which others. “Real” causal theories, for example, also tell one how *strong* the causal connection is between two variables (e.g. how many hours, on average, one needs to study in order to earn an “A” average). However, I will focus exclusively on learning causal graphs from data for two reasons. First, discovering which variables cause which others is a necessary first step in constructing causal theories, and second, there are additional difficulties with estimating the strength of a causal connection even once the graph has been correctly identified.



**Figure 1**

Returning to our example, because work ethic and intelligence are common causes of both test scores and grades, students with higher standardized

---

<sup>6</sup>Formally, two events  $A$  and  $B$  are said to be conditionally independent given  $C$  just in case  $P(A\&B|C) = P(A|C) \cdot P(B|C)$ .

test scores will typically also have better grades in high school (i.e. grades and test scores are positively correlated). So grades and test scores are dependent even though neither is a cause of the other. From a predictive standpoint, if one were asked to predict Mary's GPA, then it would be helpful to know Mary's SAT score, as the two are (however crudely) correlated.

However, knowing Mary's test scores is irrelevant to predicting her grades *if one already knows* that Mary is intelligent and works diligently. Why? Intuitively, the reason that standardized test scores might aid one in predicting a student's grades is that test scores are (albeit crude) indicators of the student's intelligence and work ethic, which are in turn indicators of the student's grades (because they are causes!). So if one already knows the student is intelligent and works diligently, then learning her SAT score will provide no new information about her grades.

In other words, the variable "standardized test scores" is independent of its non-effect "grades" given its direct causes "intelligence" and "work ethic." Notice this is just an instance of the CMC. The CMC, therefore, captures the important intuition that two variables might be correlated, and yet, they are "screened off" from one another when one conditions on all common causes.<sup>7</sup> The CMC also captures the intuition that *indirect* causes are screened off by more proximate ones, but space prevents a detailed discussion of this issue.

In general, the CMC says that if two variables are not causally related (i.e. neither is a cause of the other, nor do they share a common cause), then they are independent. The CFC essentially says the converse: informally, it says if two variables are causally related, then they are also dependent. For example, if a large study finds no correlation between the development of heart disease and smoking, then, in the absence of other information, one should conclude that there is no causal connection between the two variables.

One might be wary of drawing causal conclusions so quickly from the CMC and CFC, especially in scenarios like those involving social policy and medical recommendations concerning heart disease. In medicine and the social sciences, even the verdicts of "large" studies are often overturned by later, more comprehensive studies. Such a worry is legitimate, but it conflates (i) the error due to inherent uncertainty and "noise" in empirical data with

---

<sup>7</sup>Here one must be careful. Assuming the CMC and CFC, two variables might still be conditionally dependent given all *direct* common causes because of the existence of "M" structures. See Pearl (2000), pp. 186. The idea of "screening off" goes back at least to Reichenbach (1956), and understood in this way, the CMC is a generalization of Reichenbach's condition.

(ii) a suspicion about the validity of the CMC and CFC. Just as deductive rules of inference - *modus ponens*, *reductio ad absurdum*, and the like - may yield false conclusions from false premises, so too the CMC and CFC will fail to yield reliable causal conclusions unless one correctly identifies which variables are *truly* associated. And correctly identifying such associations is non-trivial.

For example, suppose two variables are directly causally connected, but they are only very weakly correlated. In a small study with limited data, researchers might fail to detect the weak correlation. Such a scenario is not uncommon in medicine (e.g. when a drug increases survival rates by only a fraction of a percent) or in the social sciences. In this case, employing the CFC as a rule of inference will lead one to erroneously conclude that the two variables are not directly causally connected. Analogous remarks apply to the use of the CMC.

The possibility of misidentifying which variables are associated, however, should not lead one to reject the CMC and CFC; any principles for inferring causal conclusions from associations will be subject to the same problem,. To reject the CMC and CFC for this reason would require rejecting the enterprise of causal inference from noisy, non-experimental data. For the remainder of the paper, therefore, I will focus on what can be learned from CMC and CFC under the assumption that the probabilistic relations (i.e. conditional independencies) among all variables have been correctly identified.

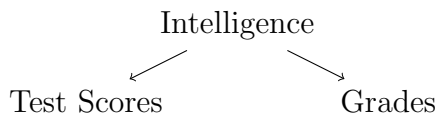
Given the CMC and CFC, one can define two causal theories/graphs to be *independence indistinguishable* (or *I*-indistinguishable for short) if they imply that the same conditional independencies for a set of variables.<sup>8</sup> Intuitively, two causal theories are *I*-indistinguishable if no amount of observational data could allow one to conclude which is correct, unless one employed domain-specific knowledge.<sup>9</sup>

---

<sup>8</sup>*I*-indistinguishability is typically called “Markov equivalence” in the causal discovery literature. I use a different term here for two reasons. First, I wish to explain why the relation is epistemologically important. The term “*I*-indistinguishable” does this, as it suggests that two *I*-indistinguishable graphs cannot be distinguished (in some important way) from data. Second, the term explains *in what way* graphs are indistinguishable, namely, that use of conditional independence information alone is insufficient to distinguish the graphs. This suggests the possibility (or rather, the fact) that other types of information might be used to distinguish *I*-indistinguishable theories.

<sup>9</sup>“Domain-specific knowledge” comes in at least three forms. The first is timing. For example, because developing a smoking habit typically precedes developing lung cancer, it is likely that having lung cancer does not cause one to smoke. Together with the CMC

For example, let's first consider a three-variable example like the one above. Suppose the true causal relationships among intelligence, standardized test scores, and grades is as below.



**Figure 2**

It turns out there are two other theories that are *I*-indistinguishable from the true one, which are depicted in the diagram below.



**Figure 3** : Two *I*-indistinguishable graphs from that in Figure 2

In this example, it seems implausible to assert that either standardized test scores or grades is a cause of intelligence. So domain-specific knowledge helps rule out alternatives that conditional independence facts cannot. Such domain-specific knowledge, however, may not be available in other areas of science, and thus, it is important to characterize what can be learned in the absence of such knowledge. Surprisingly, even with no background knowledge whatsoever, the CMC and CFC sometimes entail that there is a

---

and CFC, the assumption of “no backwards causation” allows one to discover the true causal graph given the timing of the variables and enough data (See corollary 3 in Pearl (1988)). Second, researchers often know particular mechanisms by which certain causal effects might be mediated. For example, it is known that smoking produces tar buildup in the lungs, which is one possible source of lung cancer. In contrast, it would be difficult to postulate a mechanism by which lung cancer causes one to smoke. So knowledge of possible mechanisms can constrain which causal theories are plausible given data, and such knowledge can be incorporated into well-known causal discovery procedures (See Spirtes, et. al. (2000), pp. 93). Third, parametric assumptions (e.g. that the variables are normally distributed) can be used to distinguish otherwise *I*-indistinguishable theories (Geiger and Heckerman, 1994), and similarly for assuming that the underlying probability distribution belongs to a special class of non-parametric distributions (e.g. linear, non-Gaussian, see Shimizu et. al (2006)).



*unique* causal structure compatible with the data. For example, there is no distinct causal theory that is  $I$ -indistinguishable from the theory concerning the four variables intelligence, work-ethic, grades, and test scores depicted in **Figure 1**. If the theory in **Figure 1** were true, assuming the CMC and CFC, the conditional independences among the variables would be sufficient to discover the truth.

How *often* is there a unique causal theory compatible with one’s data? Of course, the answer depends upon the scientific domain and type of causal system under investigation. However, if one assumes that all causal graphs are equally likely, then as the number of variables under investigation increases indefinitely, one in fourteen theories will be uniquely determined by probabilistic relations alone (Steinsky, 2004).

Formally, let  $n$  be a positive whole number, and suppose one is interested in the causal relationships among  $n$  many variables. Suppose one repeatedly guesses a causal theory at random (say, by picking pictures of causal graphs from a hat). The one can ask, “what proportion  $p_n$  of the randomly guessed causal theories over  $n$  variables are  $I$ -indistinguishable from themselves only?” The answer:

**Theorem 1** [Steinsky 2004]  $p_n$  approaches (about)  $\frac{1}{14}$  as  $n$  approaches infinity.<sup>10</sup>

This is good news, but it’s only the tip of the iceberg. Steinsky’s theorem asserts that some causal theories will be uniquely compatible with the data (given a sufficiently large sample). But even when the data undeterdetermines the truth, Pearl and Verma (1991) prove that the CMC and CFC entail that all theories compatible with the data will share important features. Two definitions are necessary to understand Pearl and Verma’s theorem. First, say  $X$  and  $Y$  are **adjacent** in a causal graph if either  $X$  directly causes  $Y$  or vice versa. Second, suppose  $X$  and  $Y$  both directly cause  $Z$ , but that neither

---

<sup>10</sup>In Bayesian terms, Steinsky’s theorem asserts that, with respect to a uniform prior distribution over all causal graphs with  $n$  many variables, the probability that the true, unknown graph is uniquely determined by conditional independence facts is about  $\frac{1}{14}$ . Another reasonable Bayesian prior might assign equal probability to every *equivalence class* of  $I$ -indistinguishable theories (i.e. the prior is uniform over Markov equivalence classes). Although a proof is unavailable, large computer simulations in Gillespie and Perlman (2001) suggest that about one in four  $I$ -indistinguishable classes contain only one member, and that, furthermore, the average size of  $I$ -indistinguishable classes (with respect to said prior) is approximately 3.73.

$X$  nor  $Y$  causes the other. Then say that the triple  $X, Y$ , and  $Z$  forms a *vee*, as the three variables can be arranged in a V-shape (i.e.  $X \rightarrow Y \leftarrow Z$ ). Pearl and Verma’s theorem asserts that

**Theorem 2 (Pearl and Verma (1991))** *Assuming the CMC and CFC, two causal graphs are  $I$ -indistinguishable if and only if they have the same adjacencies and vees.*

Pearl and Verma’s theorem is important because, in a wide variety of circumstances, it allows one to infer significant causal knowledge from observational data alone. For example, suppose that  $T_1$  and  $T_2$  are  $I$ -indistinguishable and that  $T_1$  entails that  $X$  and  $Y$  are adjacent, which recall, means one is a direct cause of the other. Then, according to Pearl and Verma’s theorem,  $T_2$  also entails that  $X$  and  $Y$  are adjacent. Similarly, if  $T_1$  and  $T_2$  disagree about whether  $X$  and  $Y$  are adjacent, then they are not  $I$ -indistinguishable. Thus, given enough data, Pearl and Verma’s theorem allows one to infer which variables are directly causally connected (under assumption of the CMC and CFC. of course). Why? In the large sample limit, the only theories that will be compatible with the data will also be  $I$ -indistinguishable from the true causal graph, and hence, they will entail the same adjacencies.

Applying Pearl and Verma’s theorem in practice, however, requires that all variables are, what I will call, **comeasured**. To explain the concept and its relevance to Verma and Pearl’s theorem, I will need to introduce one additional term. Thus far, I have described the causal relationships among variables, like weight, daily caloric intake, and so on. For my purposes, one can think of a variable as reporting a property of a **unit**. In medicine, the unit is almost always an individual human being, and a variable (e.g. weight or daily caloric intake) reports a property of that human being.

In the social sciences, units are also often individual human beings, and variables (e.g. standardized test scores) often report properties just like in medicine. That is not always the case. For example, economists might be interested in the relationship between business investment and taxes at two different times. In such cases, the appropriate unit is a *collection* of individuals (or perhaps an economy) at a fixed time, and the variables report properties of those individuals (e.g. how much the individuals were taxed at time  $t_1$  and how much they invested in businesses at time  $t_2$ ).

I will say that a collection  $V$  of variables is **comeasured** in a study when the “reports” of *every* variable in  $V$  are available for *every* unit in the

study. For example, if  $V$  is the set of variables consisting of intelligence, standardized test scores, work and ethic, then  $V$  is comeasured in a study if one has data concerning the intelligence, standardized test scores, etc. of every person considered in the study.<sup>11</sup>

Why does applying Verma and Pearl’s theorem require comeasurement of all variables? To apply the theorem, one must be able to rule out causal theories that are  $I$ -distinguishable (given sufficient data) from the true one. That’s harder than it sounds. If two theories are  $I$ -distinguishable, then there is *some* assertion of the form “ $v$  is independent of  $v'$  given variable set  $S$ ” that is entailed by one theory and not the other. Learning such an assertion, in general, requires comeasurement of  $v$ ,  $v'$ , and  $S$ . For example, I considered a causal theory in which standardized test scores (call it  $v$ ) is independent of grades (call this  $v'$ ) given both intelligence and work ethic (so  $S$  is the set {intelligence, work ethic}). To learn this independence, a researcher would need to compare test scores and grades among study participants of similar intelligence and work ethic. Without fixing these two other variables, a researcher would detect a correlation between test scores and grades, as they are (by assumption) effects of intelligence and work ethic. To compare two variables, while holding the two others fixed, however, requires that all four variables are comeasured.

Now, in the above example, the set  $S$  contains only two variables. This need not be the case in general. If the set  $S$  contains sufficiently many variables, it may be impossible, for experimental, practical, or ethical reasons,

---

<sup>11</sup>Although the definition of comeasurement seems strict, when I say a set of variables has not been comeasured, I will actually mean something stronger than what is required by the definition. For example, suppose one conducts a “single” study in which the height, weight, and age of a thousand participants is recorded. Suppose, unfortunately, the data file is corrupted, and the last participant’s weight is no longer available; all other data remain unchanged. Then, according to the strict definition of comeasurement, the three variables are not comeasured in the data set in question. However, one could treat the data set as being the result of two studies, one with 999 participants, and one with a single participant. In the former “study,” the three variables are comeasured, and in the second, they are not. Although it may seem odd to say one’s data came from “two” studies, there is nothing wrong with treating the data like this from a technical standpoint. Hence, when I say a set of variables  $V$  is not comeasured in any study, I will typically mean that the data points containing reports of all members of  $V$  is sufficiently sparse, even if it’s non-empty and is comeasured (in the strict sense) under some division of the available data into “studies.” I retain the strict definition because it is precise (as it does not contain the vague term “sufficiently sparse”) and provides one with the ability to describe the more important, but less precise meaning of comeasurement.

to comeasure all of  $S$  in one study. In general, there are three primary reasons why variables might not be comeasured.

The first is cost. In medicine, variables might report the outcomes of laboratory tests, x-rays, fMRI scans, etc. Such tests are expensive, and hence, it is rare that patients/study-participants are subjected to all of them. Similar remarks apply in the social sciences. For instance, gross domestic product, which measures the total value of goods exchanged in a country over a specified time (often a fiscal quarter or year), incorporates citizens' wages, corporate profits, government spending, and several other factors. To calculate gross domestic product, therefore, an economist needs data collected from millions of individuals, businesses, and governmental agencies. In sum, measuring a single variable can be costly, and so financial constraints often prevent comeasurement of many variables.

The second is lack of expertise or resources. Even with sufficient funding, social scientists might need an expert in survey-design to craft questionnaires; medical researchers might need many competent doctors to administer medical tests, and so on. Thus, measurement of some variable is rarely routine laboratory work; it often requires significant training and access to particular types of equipment. Lack of expertise or resources, therefore, can prevent comeasurement of variables even when sufficient funding is available.

The third obstacle to comeasurement is privacy. Recent work in machine learning shows that an American citizen's social security number can be predicted with approximately 87% accuracy using the person's date of birth, sex, and hometown alone (Sweeney (1997)). The upshot is that no superset of these three variables can be comeasured without, for all intents and purposes, sacrificing the anonymity of the individuals in a study. For instance, suppose a researcher is interested in the demographics of HIV infection and therefore conducts an "anonymous" survey in which individuals report their date of birth, their place of birth, their sex, and whether or not they have HIV. Because the first three variables can be used to accurately predict social security number, one could use the data from the study to identify individuals with HIV. Thus, researchers ought to be prohibited from collecting, or at the very least, publishing data sets containing comeasurement of variable sets like the one above.<sup>12</sup>

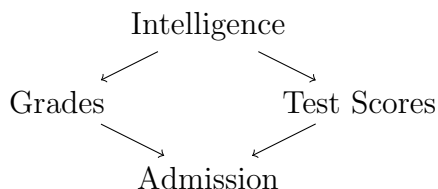
---

<sup>12</sup>In quantum mechanics, there is a fourth reason certain variables cannot be comeasured, namely, it is prohibited by Heisenberg's uncertainty principle. I do not discuss such limitations on comeasurement because I am primarily interested in causal inference in medicine and the social sciences. Moreover, there is substantial reason to doubt the

This discussion raises the question: can one draw strong causal conclusions, like those guaranteed by Steinsky and Pearl and Verma’s theorems, if all variables cannot be comeasured?

## 2 Piecemeal Causal Inference

Suppose that intelligence is an indirect cause of one’s chances of being admitted to college. Further, suppose there are two distinct ways in which intelligence affects college admissions, namely by increasing one’s standardized test scores and by improving one’s grades.



**Figure 4**

Consider what can be learned about the relationship between intelligence and college admissions in this fictitious example if one could only co-measure any proper subset of the four variables. In other words, suppose one tried to construct the above theory in a piecemeal fashion, by conducting four different observational studies in which all but one variable is measured. The variables measured in each of the four respective studies are listed in the table below. What would one learn?

Study 1	Study 2	Study 3	Study 4
Intelligence	Intelligence	Intelligence	Grades
Grades	Grades	Test Scores	Test Scores
Test Scores	Admission	Admission	Admission

**Table 1**

---

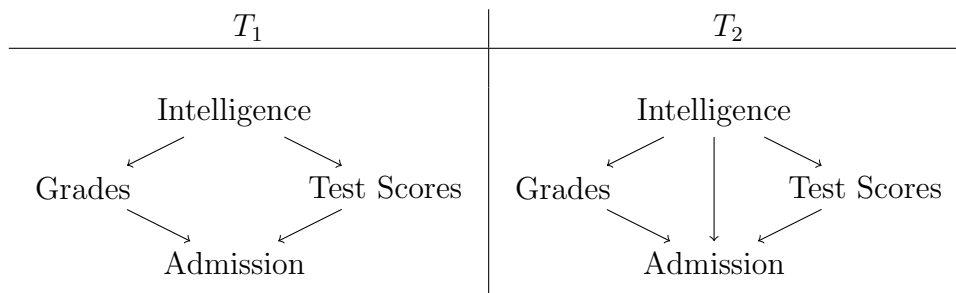
CMC and CFC are true at the quantum level.

Consider first Study 2 in which intelligence, grades, and college admission decisions are measured. As intelligence is an indirect cause of admission, one would find the two variables to be unconditionally dependent. Yet even if one conditioned on grades, one would still detect a dependence between the two variables. Why? Intuitively, if one conditions on a student's grades, one has failed to account for the other causal path by which intelligence affects admissions, namely, through test scores. This intuition can be proven formally assuming the CFC.

By symmetric reasoning, if one conducts Study 3 in which intelligence, test scores, and admissions are comeasured, one would find intelligence and admissions to be dependent, even conditional on test scores. And no other (three-variable) proper subset contains both intelligence and admissions. Therefore, if one measures only three variables at a time, the variables intelligence and admissions are dependent, regardless of which variables one puts in a conditioning set. Why is this observation important? One consequence of the CMC and CFC is the following:

**Theorem 3 (Spirtes, Glymour, Scheines (1993))** *Assuming CMC and CFC,  $X$  and  $Y$  are adjacent if and only if  $X$  and  $Y$  are dependent conditional on every set  $S$  not containing  $X$  and  $Y$ .*

Hence, by the above theorem, it appears that one cannot rule out a direct causal link between intelligence and admissions unless all four variables can be comeasured. This intuition is correct. Let  $T_1$  and  $T_2$  be the causal graphs in Figure 5 below. Notice that  $T_2$  is just like  $T_1$  except that  $T_2$  asserts that intelligence is also a direct cause of college admission decisions.



**Figure 5**

Assuming the CMC and CFC,  $T_1$  and  $T_2$  entail the same probabilistic relations involving three variables or fewer, and so  $T_1$  and  $T_2$  are indistinguishable

unless all four variables can be comeasured. Of course, because  $T_1$  and  $T_2$  do not postulate the same direct causal links, they are  $I$ -distinguishable if all variables are comeasured (by Pearl and Verma’s theorem).

The above example is fictitious, but the problem that it illustrates is not. Mayo-Wilson (2010) shows that the same theoretical problem arises in a real-world setting concerning learning the causal relations among ventilators and blood oxygen saturation in an intensive care unit. Given the sets of variables that were comeasured, one could not rule out the existence of particular direct causal links, even though one would be able to do so had all variables been comeasured.

Moreover, the problem does not disappear if one conducts more studies, and perhaps even more surprisingly, the problem is not a consequence of the existence of unmeasured “confounding” variables. The theorem below shows that, even if one knows *all potential confounding variables*, the piecemeal construction of causal theories can increase underdetermination of theory by evidence. To understand the theorem, it will help to introduce one definition. Let  $V$  be any collection of variables. For instance,  $V$  might be the set containing measurements of intelligence, work ethic, grades and standardized test scores, as in the above examples. One can think of an observational study  $U$  as a subset of the variables  $V$ . For example, an observational study that measures grades and standardized test scores can be represented by the pair consisting of those two variables.

A collection of observational studies  $\mathcal{U}$ , therefore, can be represented by a collection of subsets of  $V$ . In the above example,  $\mathcal{U}$  consisted of every three variable subset of the four variables intelligence, work ethic, grades and standardized test scores. Say two theories  $T_1$  and  $T_2$  are *piecemeal indistinguishable relative to  $\mathcal{U}$*  if the two theories entail the same independence facts would be learned if one conducted only the studies represented by  $\mathcal{U}$ . I will omit  $\mathcal{U}$  when the set of studies is clear from context. Notice that  $I$ -indistinguishability is simply a special case of piecemeal indistinguishability, namely, the case in which one conducts a study in which all variables are comeasured. Mayo-Wilson (2010) shows that, no matter how many observational studies are conducted, if not every variable is comeasured, some information might be lost in the piecemeal construction of causal theories.

**Theorem 4 (Mayo-Wilson (2010))** *For any set of variables containing at least two variables,*

1. *There exist distinct causal theories  $T_1$  and  $T_2$  with different adjacencies*

that are piecemeal distinguishable (given sufficient data) if and only if every variable in  $V$  is comeasured. In fact,  $T_1$  might contain strictly more causal links than  $T_2$ , or both  $T_1$  and  $T_2$  might contain direct causal links that the other does not.

2. If the set contains at least four variables, there exist distinct causal theories  $T_1$  and  $T_2$  with different vees that are piecemeal distinguishable (given sufficient data) if and only if every variable in  $V$  is comeasured.

In other words, Pearl and Verma's theorem goes out the window when causal theories are constructed piecemeal. The previous theorem entails that the piecemeal construction of causal theories always raises the possibility of underdetermination about both *adjacency* and *orientation* information, regardless of how many observational studies are conducted and how many variables are accounted for. Mayo-Wilson (2010) calls this the *problem of piecemeal induction*.

Does the problem of piecemeal induction undermine the possibility of significant causal discovery? It's not clear. Understanding the severity of the problem requires answering at least three other questions. First, how *much* information is lost in piecemeal causal inquiry? For instance, the above theorem entails that piecemeal indistinguishable graphs might possess different adjacencies and vees, but by how many adjacencies and vees might they differ? If the number is sufficiently small, then one might be able to infer important causal facts (like those guaranteed by Pearl and Verma's theorem) in piecemeal causal inquiry.

Second, how often does the problem of piecemeal induction arise? The previous theorem says that, if particular types of causal theories happen to be true, then piecemeal inquiry increases causal undetermination. Importantly, the theorem does not say that *all* causal theories create a problem of piecemeal induction, and it does not provide any information about how frequently the troublesome causal theories occur in nature. If the problem of piecemeal induction is rare, then scientists can safely ignore it in practice. Third, under what circumstances is no causal information lost in piecemeal inquiry? These three questions are taken up in turn in the next section.



### 3 The Extent and Frequency of the Problem of Piecemeal Induction

How much information is lost in piecemeal inquiry? A rough answer to this question would characterize how much information is lost in the “best-case” scenario, namely, when one can conduct a series of observational studies in which *every proper subset* of the variables under investigation is measured. In such a case, very little adjacency information is lost:

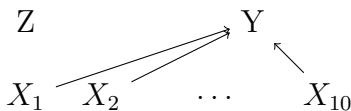
**Theorem 5** *Let  $V$  be any set of variables, and suppose every proper subset of  $V$  can be measured. If  $T_1$  and  $T_2$  are piecemeal indistinguishable causal graphs, then  $T_1$  contains no more than one adjacency that  $T_2$  does not. It follows that  $T_1$  and  $T_2$  differ by no more than two adjacencies (in the sense that  $X$  and  $Y$  may be adjacent in  $T_1$  but not  $T_2$ , whereas  $W$  and  $Z$  may be adjacent in  $T_2$  but not  $T_1$ ).*

Unfortunately, considerably more information might be lost concerning vees:

**Theorem 6** *Suppose  $V$  contains  $n \geq 3$  many variables, and suppose that every proper subset of  $V$  can be measured.*

1. *If  $T_1$  and  $T_2$  are piecemeal indistinguishable causal graphs, then  $T_1$  contains no more than  $n - 3$  vees that  $T_2$  does not. Moreover,*
2. *There are piecemeal indistinguishable graphs  $T_1$  and  $T_2$  such that  $T_1$  contains  $n - 3$  vees that  $T_2$  does not.*

One may wonder how  $T_1$  can contain so many more vees than  $T_2$  if it can contain no more than one additional adjacency. The answer is that, by adding one adjacency to  $T_2$ , one can add many vees at once. For example, suppose that a causal graph contains a direct causal link from each of ten variables  $X_1, X_2, \dots, X_{10}$  to a variable  $Y$ , but does not contain any other edges (see below). Then if one adds an edge from  $Z$  to  $Y$ , the resulting graph contains ten vees that the previous one did not, namely,  $X_i \rightarrow Y \leftarrow Z$  for all of the  $X_i$ 's.



## Figure 6

Thus, the previous two theorems provide some room for optimism. How so? Recall that Pearl and Verma’s theorem entails that if every variable under investigation is comeasured, then one can identify all of the direct causal links in the true theory. That is, one will not err in identifying which pairs of variables are directly causally linked. The above theorem says that, if every proper subset of the variables is measured, then one will err in identifying the direct causal links at most two times (and perhaps none).

However, that optimism should be tempered. Researchers rarely have the luxury of measuring *every proper subset* of the variables of interest. So it is natural to ask what can be learned from studies in which no more than a fixed, small number of variables are observed at a time. The next theorem says, unsurprisingly, that the smaller the studies, the more underdetermination increases concerning direct causal links:

**Theorem 7** *Suppose  $V$  contains  $n$  variables, and at most  $k < n$  variables can be observed at a time. Then there exist piecemeal indistinguishable causal theories  $T_1$  and  $T_2$  such that  $T_1$  contains  $\binom{n-k+1}{2}$  many direct causal links that  $T_2$  does not.*

I conjecture that the above theorem also describes the worst-case for underdetermination when at most  $k < n$  variables are comeasured. It is an open problem to determine the number of vees by which two indistinguishable causal theories can differ when at most  $k < n$  variables are measured.

Together, the above theorems provide preliminary answers to the question, “How *much* information might be lost in piecemeal inquiry?” An equally important question is, “How *frequently* is information lost in piecemeal inquiry?” In particular, when, if ever is no information lost?

Let’s begin with the last question. It is helpful to consider again the fictitious example involving intelligence, grades, test scores, and college admissions. In that example, one cannot discern whether intelligence is a direct cause of admissions or not. Therefore, the two theories  $T_1$  and  $T_2$  depicted in Figure 5, with four and five direct causal links respectively, are piecemeal indistinguishable. Notice that, in any causal theory involving four variables, there are no more than six direct causal links. Thus, in the fictitious example, the two piecemeal indistinguishable theories contain almost as many direct causal links as is theoretically possible.

Here's the insight: piecemeal inquiry seems to fail in the example because there are several different causal paths from intelligence to admissions. One might conjecture, therefore, that if the true causal theory is sufficiently *simple* (where simplicity is measured by number of direct causal links), then piecemeal inquiry might succeed in recovering all information that could have been learned had all variables been comeasured. This conjecture is borne out by the following theorem:

**Theorem 8** *Suppose there are  $n$  variables under investigation, and assume that  $k < n$  variables can be measured at a time. Further, assume that the true causal theory postulates fewer than  $2k - 2$  direct causal links. Then no information is lost in piecemeal inquiry.*

Theorem 8 has at least two important philosophical consequences. First, it highlights the relationship between two central features of scientific practice: the use of Ockham's razor and the piecemeal construction of scientific theories. Ockham's razor is the principle that, all other things being equal, it is rational to prefer simpler scientific theories to more complex ones. If Ockham's razor can be justified, then, by the above theorem, so might one justify piecemeal causal inquiry. Why? The above theorem entails that piecemeal inquiry succeeds under the assumption the true causal theory is sufficiently simple, and Ockham's razor asserts that a systematic preference for simpler theories is rationally justified. Hence, assuming Ockham's razor, it is rational to assume that the true causal theory can be constructed in a piecemeal fashion.<sup>13</sup>

Second, Theorem 8 also provides a rational justification for a robust pattern of scientific practice. Large observational studies are uncommon, and when they are conducted (e.g. the Framingham study on the causes of heart disease), researchers typically expect the causal connections among the variables to be intricate and numerous. If researchers are (at least implicitly) employing principles like the CMC and CFC, then the above theorem justifies why such large observational studies need to be conducted only when there is a large, dense causal graph under investigation: if the true causal theory postulates only a few connections among the variables, then a larger study is unnecessary to confirm it.

Thus far, I have characterized how much causal information is lost in the worst-case when many studies are combined, and I also showed that, in the

---

<sup>13</sup>For a defense of Ockham's razor in causal inference, see Kelly and Mayo-Wilson (2010).

best-case (i.e. when the true causal theory is sufficiently simple), no information is lost whatsoever. So one might ask, “How *often* does the problem of piecemeal induction arise?” We can make this question precise as follows. Suppose researchers can measure at most  $k$  variables at a time, and say a causal theory  $T$  is  $k$ -underdetermined just in case there is some theory  $T'$  that is distinguishable from  $T$  when all  $n$  variables are measured, but not so when only  $k$  variables can be comeasured. Let  $p_k(n)$  be the proportion of causal graphs over  $n$  variables that are  $k$ -underdetermined. In essence, the proportion  $p_k(n)$  describes how frequently one will lose at least some causal information if one assumes all causal theories over  $n$  variables are equally likely. If  $p_k(n)$  is close to zero and  $k$  many variables can be comeasured, then researchers can take consolation in the fact that the problem of piecemeal induction is rare. If  $p_k(n)$  is close to one and one can observe at most  $k$  variables at a time, then researchers will need to make use of domain-specific knowledge in order to combat potentially rampant underdetermination. Unfortunately, the news is quite bad:

**Theorem 9** *For any natural number  $k$ , the proportion  $p_k(n)$  of graphs over  $n \geq k$  variables that are  $k$ -undetermined approaches 1 as  $n$  approaches infinity.*

Compare the result with Steinsky’s theorem. According to Steinsky, if all variables can be comeasured, then about one in fourteen causal theories is uniquely determined by conditional independence facts. In contrast, by Theorem 9, if there is any finite bound  $k$  to the number of variables that can be comeasured (where  $k$  can be as large as one pleases), then the proportion of causal theories uniquely determined by conditional independence facts approaches zero as the number of variables under investigation increases. This is bad news, but it gets worse.

The proof of Theorem 9 can be used to show that the amount of information lost in piecemeal causal inquiry, on average, becomes arbitrarily bad as the number of variables increases. How so? Given a causal theory  $T$ , define the *extent of  $k$ -underdetermination* of  $T$  to be the maximum number of distinct causal theories  $T_1, T_2, \dots, T_m$  such that (i)  $T$  is piecemeal indistinguishable from  $T_i$  for all  $i \leq m$  if at most  $k$  many variables are measured at a time, and (ii) each pair of the theories  $T, T_1, T_2, \dots, T_m$  are  $I$ -distinguishable if all variables are comeasured. Let  $E_k(n)$  be the average extent of  $k$ -underdetermination over all causal theories concerning  $n$  variables. Informally,  $E_k(n)$  measures how much the piecemeal construction of

causal theories increases underdetermination, as it makes precise how many theories (on average) one can no longer distinguish due to the piecemeal construction of causal theories. Then:

**Theorem 10** *For any natural number  $k$ , the average extent of  $k$  underdetermination  $E_k(n)$  becomes arbitrarily large as  $n$  approaches infinity.*

One should be careful in drawing conclusions from the previous two theorems, however. In most applications, there is little reason to believe that every causal theory is equally likely. Some causal graphs are clearly implausible given what one knows about the variables of interest. If domain-specific knowledge guarantees that the true causal graph is sufficiently simple, then the above theorem should not cause researchers to abandon their faith in finding the true causal theory. On the other hand, if the true causal graph is known to contain a large number of adjacencies (and that is all that is known), then  $p_k(n)$  is a conservative estimate of the likelihood that information concerning direct causal links and their orientations will be lost, and  $E_k(n)$  is a conservative estimate of how much information is lost.

## 4 Conclusion

The last two theorems show that the problem of piecemeal induction is both serious and potentially common in various scientific domains. When the underlying causal truth is sufficiently complex, there is a significant possibility that a number of relevant causal facts are lost by trying to integrate the results of many observational studies. Yet the theorems in this paper leave open a number of important questions.

First, even if the probability of  $k$ -underdetermination approaches one as the number of variables under investigation approaches infinity, it is not clear *how fast* that limit is reached. In other words,  $p_k(n)$  may be very small unless  $n$  is very large. If that were the case, then unless researchers have reason to believe that the causal system under interest contains an enormous number of variables, they ought not worry about the problem of piecemeal induction. Similar remarks apply to how quickly the average extent of  $k$ -underdetermination increases as a function of number of variables.

Second, even when the problem of piecemeal induction is inevitable, it is possible that scientific institutions might be able to *plan* sequences of studies so as to *minimize* the type of causal information that is lost. Danks

(2005) presents two case studies that suggest precisely this point, but the study of sequential-planning of observational studies (from a causal discovery framework) remains essentially unexplored.

Third, the results above concern the problem of piecemeal induction as it pertains to causal inference from *observational* data. Although there exist methods for combining the results of several experiments (see Eberhardt, Hoyer, and Scheines (2010)), the correctness of such methods is generally assessed under the assumption that *any intervention* on the variables under investigation is possible. Such an assumption is often false because of ethical, financial, and experimental limitations, and hence, a different but analogous problem of piecemeal induction will arise in the experimental context when only certain interventions are possible. Future work ought to characterize the extent and frequency of that analogous problem.

Fourth, all of the above theorems assume the CMC and CFC, but several recently developed causal discovery algorithms work under entirely different assumptions. Recall the CMC and CFC intend to axiomatize the relationship between (i) probability and (ii) causation. When used as principles for causal inference, however, the principles make use of only a very particular type of probabilistic information, namely, which variables are conditionally independent of which others. But data often contains far more information than just which variables are independent or correlated. For instance, scientists often have good reason to believe that a particular variable can take only finitely many values or that it is continuous. Data might suggest that one variable is normally distributed, whereas another is not. When employed alone, the CMC and CFC make no use of such information, which can often be helpful in causal discovery.<sup>14</sup> Future work ought to characterize the severity and frequency of the problem of piecemeal induction under these alternative assumptions.

Fifth, the theorems above show that the problem of piecemeal induction is a *possibility* in some areas of causal discovery, and my arguments (as well as the case study in Mayo-Wilson (2010)) show that, at least in some scientific contexts, it is a *real* problem. However, a systematic series of case studies is still necessary in order to characterize the types of domains/systems in which the problem is most pernicious and frequent. The ability to hold particular

---

<sup>14</sup>See references in Footnote 8. Since submitting this paper for publication, the author has made significant progress in answering the third and fourth questions here. Contact the author for a summary of what can be learned from piecemeal causal inference from experimental data under various parametric assumptions.

variables fixed in an experimental context (as is typical in physics and chemistry) and available domain-specific knowledge might eliminate much of the underdetermination caused by the piecemeal construction of causal theories. Perhaps more importantly, existing algorithms for the piecemeal construction of causal theories ought to be improved to incorporate such domain-specific knowledge when available.

Finally, this paper takes preliminary steps in characterizing what can be learned from the piecemeal construction of *causal* theories. But it is obvious that the problem of piecemeal induction is a much broader phenomenon - one that likely occurs in all areas of science in which data from disparate sources needs to be integrated. Scientific theorizing increasingly requires synthesizing data from more and more areas, and scientists have become ever more reliant on statistical methods that allow them to analyze large and complex data sets. Thus, substantial work remains for philosophers of science and methodologists in characterizing the frequency and extent of the problem of piecemeal induction in new domains of empirical research.

Conor Mayo-Wilson  
Department of Philosophy  
Baker Hall 135  
Carnegie Mellon University  
Pittsburgh, PA 15213-3890

## Acknowledgements

Special thanks to David Danks for providing numerous comments and suggestions on earlier drafts of this paper. Clark Glymour prompted me to pursue questions concerning the frequency of the problem of piecemeal induction, which ultimately led to the discovery and proofs of Theorems 9 and 10. Thanks to Brendan McKay who, in personal correspondence, suggested a proof sketch of Lemma 9, which is one of two key lemmas in the proof of Theorems 9 and 10. Finally, two anonymous referees suggested several important clarifications to the main philosophical argument and its motivation.

## References

- Bender, E., Richmond, L., Robinson, R., and Wormald N. [1986]: “The asymptotic number of labelled acyclic digraphs.” *Combinatorica*. pp. 15-22.
- Cartwright, N. [1989]: “Natures Capacities and Their Measurement.” Oxford: Oxford University Press.
- Cartwright, N. [2002]: “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**, pp. 411-453.
- Cartwright, N. [2007]: *Hunting Causes and Using Them*. Cambridge: Cambridge University Press.
- Danks, D. [2002]: “Learning the causal structure of overlapping variable sets.” In S. Lange, K. Satoh, and C. H. Smith (Eds.), *Discovery science: Proceedings of the 5th international conference*. Berlin: Springer-Verlag. pp. 178-191.
- Danks, D. [2005]: “Scientific coherence and the fusion of experimental results.” *The British Journal for the Philosophy of Science*, 56. pp. 791-807.
- Eberhardt, F., Hoyer, P.O., and Scheines, R. [2010]. Combining Experiments to Discover Linear Cyclic Models with Latent Variables. In *Journal of Machine Learning, Workshop and Conference Proceedings (AISTATS 2010)*, **9**, pp. 185-192.
- Hausman, D. and Woodward, J. [2002]: “Manipulation and the Causal Markov Condition.” *Proceedings of the 2002 Biennial Meeting of The Philosophy of Science Association*, pp. 846-857.
- Hausman, D. and Woodward, J. [2004]: “Modularity and the Causal Markov Condition: A Restatement.” *British Journal Philosophy of Science*, **55**, pp. 147-161.
- Freedman, D. and Humphreys, P. [1999]: “Are There Algorithms That Discover Causal Structure?” *Synthese*. **121**, pp. 29-54.



- Geiger, D. and Heckerman, D. [1994]: “Learning Gaussian Networks.” *Proceedings of the Tenth Conference Annual Conference on Uncertainty in Artificial Intelligence*. pp. 235-243.
- Gillespie, S. and Perlman, M. [2001]: “Enumerating Markov equivalence classes of acyclic digraph models.” In Ed. Goldszmidt, M., Breese, J., and Koller, D. *Proceedings of the Seventh Conference on Uncertainty in Artificial Intelligence*. Seattle, WA. Morgan Kaufmann. pp. 171-177.
- Hesslow, G. [1976]: “Discussion: Two Notes on the Probabilistic Approach to Causality”, *Philosophy of Science*, **43**, pp. 290-292.
- Kelly, K. and Mayo-Wilson, C. [2010]: “Causal Conclusions That Flip Repeatedly and Their Justification.” *Proceedings of the Twenty Sixth Conference on Uncertainty in Artificial Intelligence*. Eds. P. Grunwald and P. Spirtes. pp. 277-286.
- Laudan, L. and Leplin, J. [1991]: “Empirical Equivalence and Underdetermination,” *Journal of Philosophy* **88**, pp. 449-472.
- Mayo-Wilson, C. [2011]: “The Problem of Piecemeal Induction.” *Philosophy of Science*. **78**(5). December, 2011. pp. 864-874.
- McKay, B. [1989]: “On the shape of a random acyclic digraph.” *The Mathematical Proceedings of the Cambridge Philosophical Society*. **106**. pp. 459-465.
- Pearl, J. [1988]: *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Francisco: Morgan Kaufmann.
- Pearl, J. [2000]: *Causality: Models, Reasoning, and Inference*. New York: Cambridge University Press.
- Pearl, J. and Verma, T. [1991]: “A Theory of Inferred Causation,” in J.A. Allen, R. Fikes, and E. Sandewall eds., *Principles of Knowledge Representation and Reasoning: Proceeding of the Second International Conference*, San Mateo, CA: Morgan Kaufmann. pp. 441-452.
- Reichenbach, H. [1956]: *The Direction of Time*. University of Los Angeles Press.

- Shimizu, S., Hoyer, P. O., Hyvärinen, A. and Kerminen A. J. [2006]: “A linear non-gaussian acyclic model for causal discovery.” *Journal of Machine Learning Research*. **7**. pp. 2003-2030.
- Spirtes, P., Glymour, C. and Scheines, R. [2000]: *Causation, Prediction, and Search*. Second Edition. Cambridge: MIT Press.
- Stanford, P. K., [2006]: *Exceeding Our Grasp: Science, History, and the Problem of Unconceived Alternatives*. New York: Oxford University Press.
- Steel, D. [2005]: “Indeterminism and the Causal Markov Condition.” *The British Journal for the Philosophy of Science*. **56**, pp. 3-26.
- Steinsky, B. [2004]: “Asymptotic Behaviour of the Number of Labelled Essential Acyclic Digraphs and Labelled Chain Graphs.” *Graphs and Combinatorics*. **20**, pp. 399-411.
- Sweeney, L. [1997]: Description of the software is available online at <http://dataprivacylab.org/people/sweeney/artifacts.html> }
- Tillman, R. E., Danks, D. and Glymour, C. [2008]: “Integrating Locally Learned Causal Structures with Overlapping Variables.” In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou (Eds.), *Advances in Neural Information Processing Systems*, **21**, pp. 1665-1672.
- Tillman, R. E. [2009]: “Structure learning with independent non-identically distributed data.” In *Proceedings of the 26th International Conference on Machine Learning (ICML 2009)*.
- Tillman, R. E. and Spirtes, P. [2011]: “Learning equivalence classes of acyclic models with latent and selection variables from multiple datasets with overlapping variables”. To appear in *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS 2011)*.
- van Fraassen, B. C. [1980]: *The Scientific Image*, Oxford: Clarendon Press.

## Appendix A - Definitions and Previous Results

Both appendices assume familiarity with Bayesian networks and their use in causal discovery. Here, I introduce some notational conventions and state some known results that will be employed in the proofs of the theorems.

For any finite set  $V$ , let  $\text{DAG}_V$  denote the set of all directed acyclic graphs (DAGs) that have the vertex set  $V$ . I use the uppercase letters  $G$  and  $H$  to denote members of  $\text{DAG}_V$ . For any graph  $G \in \text{DAG}_V$  and any vertex  $v \in V$ , let  $PA_G(v)$  denote the set of parents of  $v$  in  $G$ , and let  $Ch_G(v)$  denote its children. Let  $Desc_G(v)$  denote the set of descendants of  $v$  in  $G$ , and similarly let  $Anc_G(v)$  denote its ancestors. If  $v_1 \rightarrow v_3 \leftarrow v_2 \in G$ , then we say  $v_3$  is 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 we say  $v_3$  is an **unshielded collider** with respect to  $v_1$  and  $v_2$ . Accordingly, for any two variables  $v_1, v_2 \in V$  and any  $G \in \text{DAG}_V$ , define  $UC_G(v_1, v_2)$  to be the set of all  $v_3$  such that  $v_1 \rightarrow v_3 \leftarrow v_2$  is an unshielded collider in  $G$ . Finally, for any  $v \in V_G$ , define the **Markov Blanket**  $MB_G(v)$  of  $v$  in  $G$  to be the set of vertices  $w$  that are either adjacent to  $v$  or that form unshielded colliders  $v \rightarrow u \leftarrow w$  in  $G$ .

A **path**  $\pi$  in  $G$  is a non-repeating sequence of vertices  $\pi = \langle v_1, v_2, \dots, v_n \rangle$  such that  $v_i$  and  $v_{i+1}$  are adjacent if  $1 \leq i < n$ . The path  $\pi$  is called **directed** if  $v_i$  is a parent of  $v_{i+1}$  for all  $i$ . Given a path  $\pi = \langle v_1, v_2, \dots, v_n \rangle$ , let  $\pi \downarrow v_i = \langle v_1, \dots, v_i \rangle$ , and call  $\pi \downarrow v_i$  the **initial segment** of  $\pi$  that terminates with  $v_i$ . Similarly, let  $\pi \uparrow v_i = \langle v_i, \dots, v_n \rangle$ , and call  $\pi \uparrow v_i$  the **tail** of  $\pi$  that begins with  $v_i$  and terminates with the end of  $\pi$ . Given two paths  $\pi_1$  and  $\pi_2$  in a graph  $G$  such that the endpoint of  $\pi_1$  is the starting point of  $\pi_2$ , let  $\pi_1 \frown \pi_2$  denote the concatenation of the two paths. If  $\pi$  is a path between  $v$  and  $v'$ , and no variables on  $\pi$  are colliders on  $\pi$ , then we say  $\pi$  is a **trek**. In diagrams, I use straight lines to indicate the existence of an edge, and if I wish to indicate that the edge has a particular direction, then I will use an arrow marker (e.g. see the edge between  $v_1$  and  $v_2$  below). 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  $v_3$ ).

$$v_1 \rightarrow v_2 \rightsquigarrow v_3 \leftarrow \rightsquigarrow v_4$$

Figure 7

## Bayesian Networks and Markov Equivalence

Let  $V$  be a set of random variables on a measurable space  $(\Omega, \mathcal{F})$ . I will use the lowercase letters  $v, w, x, y, z$  to denote elements of  $V$ , the uppercase letters  $U$  and  $W$  to denote subsets of  $V$ , and scripted letters  $\mathcal{U}, \mathcal{W}$  to denote collections of subsets of  $V$ , i.e.  $\mathcal{U}, \mathcal{W} \subseteq 2^V$ .

For any probability measure  $p$  on  $(\Omega, \mathcal{F})$ , any two variables  $v, v' \in V$  and any  $U \subseteq V \setminus \{v, v'\}$ , write  $p \models v \perp\!\!\!\perp v' | U$  if, with respect to the measure induced by  $p$ , the variables  $v$  and  $v'$  are conditionally independent given  $U$ . The assertion  $v \perp\!\!\!\perp v' | U$  is called a **conditional independence constraint**. Define  $\text{CIC}^V$  to be the set of all such conditional independence constraints concerning the variables  $V$ . For any measure  $p$  define:

$$\text{CIC}_p^V := \{\varphi \in \text{CIC}^V : p \models \varphi\}$$

In other words,  $\text{CIC}_p^V$  contains the set of all true assertions about which variables of  $V$  are conditionally independent of others with respect to  $p$ .

For any  $G \in \text{DAG}_V$ , let  $\Phi(G)$  denote the set of conditional independence constraints of the form  $v \perp\!\!\!\perp v' | PA_G(v)$ , where  $v, v' \in V$ ,  $PA_G(v)$  is the set of parents of  $v$  in  $G$ , and  $v$  is not an ancestor of  $v'$  in  $G$ . Say that  $p$  is **Markov** to  $G$  if and only if  $\Phi(G) \subseteq \text{CIC}_p^V$ . Define:

$$\text{CIC}_G^V = \bigcap \{\text{CIC}_p^V : p \text{ is a measure on } (\Omega, \mathcal{F}) \text{ \& } \Phi(G) \subseteq \text{CIC}_p^V\}.$$

If  $\text{CIC}_p^V = \text{CIC}_G^V$ , then  $p$  is said to be **faithful** to  $G$ .

Although it is standard to study the set of conditional *independencies* that hold with respect to a Bayes net, in the following, it will be helpful to consider *dependencies* as well. Accordingly, let  $\text{CDC}^V$  be the set of all **conditional dependence constraints** over the variables in  $V$  (i.e. the set of negations of the assertions in  $\text{CIC}^V$ ), and similarly, define:

$$\begin{aligned} \text{CDC}_G^V &= \{\varphi \in \text{CDC}^V : \varphi \notin \text{CIC}_G^V\} \\ \text{CIDC}^V &= \text{CIC}^V \cup \text{CDC}^V \\ \text{CIDC}_G^V &= \text{CIC}_G^V \cup \text{CDC}_G^V \end{aligned}$$

If  $\text{CIC}_G^V = \text{CIC}_H^V$ , say  $G$  and  $H$  are **Markov equivalent**, and write  $G \equiv H$ . Denote the Markov equivalence class of  $G$  by  $[G]$ .

We will use the lowercase Greek letters  $\varphi$  and  $\psi$  and so on to denote elements of  $\text{CIDC}^V$ , and upper case Greek letters  $\Phi$  and  $\Psi$  to denote subsets of  $\text{CIDC}^V$ . For any  $\Phi \subseteq \text{CIDC}^V$ , we write  $G \models \Phi$  if  $\Phi \subseteq \text{CIDC}_G^V$ , and we say that  $G$  **satisfies**  $\Phi$ .

## Graphical Structure and Probabilistic Relations

The following theorems and corollaries characterize the relationship between (i) graphical properties (like the existence of paths, treks, colliders, etc.) in Bayesian networks, and (ii) the conditional independence structure represented by the graph. They are stated without proof and used frequently in the remainder of the proofs.

**Definition 1** Let  $G \in \text{DAG}_V$ ,  $v_1, v_2 \in V$ , and suppose  $v_1 \neq v_2$ . Let  $\pi$  be an undirected path between  $v_1$  and  $v_2$ , and  $U \subseteq V \setminus \{v_1, v_2\}$ . Then a vertex  $v_3$  is **active on  $\pi$  in  $G$  given  $U$**  just in case either

1.  $v_3$  is not a collider on  $\pi$  and  $v_3 \notin U$
2.  $v_3$  is a collider on  $\pi$  and either (i)  $v_3 \in U$  or (ii) there is  $w \in \text{Desc}_G(v_3) \cap U$  (or both).

Say a path  $\pi$  between  $v_1$  and  $v_2$  is **active given  $U$**  just in case every variable on  $\pi$  is active.

**Definition 2** Let  $G \in \text{DAG}_V$ ,  $v_1, v_2 \in V$ , and suppose  $v_1 \neq v_2$ . Let  $U \subseteq V \setminus \{v_1, v_2\}$ . Then  $v_1$  and  $v_2$  are **d-separated given  $U$**  if and only there is no undirected path  $\pi$  between  $v_1$  and  $v_2$  such that  $\pi$  is active relative to  $U$ . Say they are **d-connected given  $U$**  otherwise.

**Proposition 1 (Verma and Pearl)** Let  $G \in \text{DAG}_V$ . For all  $v_1, v_2 \in V$ , and  $U \subseteq V \setminus \{v_1, v_2\}$ :

$$v_1 \amalg v_2 | U \in \text{CIC}_G \text{ if and only if } v_1 \text{ and } v_2 \text{ are d-separated given } U.$$

Two useful corollaries of Pearl and Verma's theorem is below:

**Corollary 1** Let  $G \in \text{DAG}_V$  and  $v_1, v_2 \in V$ . Then  $v_1$  and  $v_2$  are adjacent in  $G$  if and only if  $v_1 \amalg v_2 | U \notin \text{CIC}_G$  for all  $U \subseteq V \setminus \{v_1, v_2\}$ .

**Corollary 2** Let  $G \in \text{DAG}_V$  and  $v_1, v_2, v_3 \in V$ . Then  $v_1 \rightarrow v_2 \leftarrow v_3$  is a collider in  $G$  if and only if  $v_1 \amalg v_3 | U \notin \text{CIC}_G$  for all  $U \subseteq V_G$  containing  $v_2$ .

**Proposition 2 (Verma and Pearl)** Let  $G, G' \in \text{DAG}_V$ . Then  $G \equiv G'$  if and only if  $G$  and  $G'$  possess the same adjacencies and unshielded colliders.

**Definition 3** Let  $G \in \text{DAG}_V$ ,  $v_1, v_2 \in V$ , and  $U \subseteq V \setminus \{v_1, v_2\}$ . A path  $\pi$  between  $v_1$  and  $v_2$  is called an **inducing path over  $U$**  if every member of  $U$  that appears on  $\pi$  is a collider, and every collider on  $\pi$  is an ancestor of either  $v_1$  or  $v_2$ .

**Proposition 3 (Verma and Pearl (1991))** Let  $G \in \text{DAG}_V$ ,  $v_1, v_2 \in V$ , and  $U \subseteq V \setminus \{v_1, v_2\}$ . Then  $v_1$  and  $v_2$  are  $d$ -connected given any subset  $U' \subseteq U$  if and only if there exists an inducing path between  $v_1$  and  $v_2$  over  $U$ .

## $\mathcal{U}$ -equivalence

Given  $\mathcal{U} \subseteq 2^V$ , define  $\text{CIC}_G^{\mathcal{U}} = \cup_{U \in \mathcal{U}} \text{CIC}_G^U$ . Here,  $\mathcal{U} \subseteq 2^V$  is intended to represent a collection of observational studies, and  $\text{CIC}_G^{\mathcal{U}}$  represents the set of conditional independence constraints one would endorse if one sampled from each of the sets  $U \in \mathcal{U}$  infinitely often. See Mayo-Wilson (2010) for more detailed discussions of these points. If  $\text{CIC}_G^{\mathcal{U}} = \text{CIC}_H^{\mathcal{U}}$ , say  $G$  and  $H$  are  **$\mathcal{U}$ -equivalent**, and write  $G \equiv_{\mathcal{U}} H$ . Denote the  $\mathcal{U}$ -equivalence class of  $G$  by  $[G]_{\mathcal{U}}$ . Two graphs are  $\mathcal{U}$ -equivalent if, in the absence of background information on the data generating process, no amount of observational data collected in the observational studies  $\mathcal{U}$  would allow one to know which of the two graphs truly describes the causal relations among the variables  $V$ . Such background information might include, among other things, constraints on the graph (e.g.  $v_1$  is a known cause of  $v_2$ ), parametric assumptions (e.g. the model is known to be linear Gaussian), or assumptions that the truth belongs to a large class of non-parametric models (e.g. the model is linear with non-Gaussian errors).

For each  $k \leq |V|$ , if  $\mathcal{U} = \{U \subseteq V : |U| \leq k\}$ , then we let  $\text{CIC}_G^k = \text{CIC}_G^{\mathcal{U}}$  and write  $G \equiv_k H$  if  $G \equiv_{\mathcal{U}} H$ . In this case, we say  $G$  and  $H$  are  **$k$ -equivalent**. Denote the  $k$ -equivalence class of  $G$  by  $[G]_k$ . The precise statement of Theorem 3 in the body of the paper, then, is as follows:

**Theorem 3 (Mayo-Wilson (2010))** For all natural numbers  $n \geq 2$  and all sets  $V$  such that  $|V| = n$

1. There exist  $G, H \in \text{DAG}_V$  such that  $G \equiv_{n-1} H$  but  $H$  contains strictly more edges than  $G$ . Moreover, there exist  $G, H \in \text{DAG}_V$  such that  $G \equiv_{n-1} H$  and both  $G$  and  $H$  contain edges that the other does not.

2. If  $n \geq 4$ , then there exist  $G, H \in \text{DAG}_V$  such that  $G \equiv_{n-1} H$  but  $H$  contains strictly more unshielded colliders than does  $G$ .

## Appendix B - Proofs of Theorems

We now precisely restate and prove the theorems in the body of the text.

**Theorem 4** *For any set  $V$ , if  $G \equiv_{n-1} H$  (where  $n = |V|$ ), then  $G$  contains **at most** one edge that  $H$  does not, and  $H$  can contain **at most** one edge that  $G$  does not. So it follows that  $G$  and  $H$  differ by at most two edges.*

**Lemma 1** *Let  $G$  and  $H$  be DAGs over  $n$ -variables and  $k < n$ . Suppose  $G \equiv_k H$  and that  $v_1$  and  $v_2$  are adjacent in  $H$  but not  $G$ . Then either  $v_1$  or  $v_2$  (or both) has at least  $k - 1$  parents in  $G$ .*

**Proof:** For the sake of contradiction, suppose that both  $v_1$  and  $v_2$  have fewer than  $k - 1$  parents in  $G$ . As  $G$  is acyclic, either  $v_1$  is not an ancestor of  $v_2$  or  $v_2$  is not an ancestor of  $v_1$  in  $G$ . Without loss of generality, assume  $v_1$  is not an ancestor of  $v_2$ . As  $v_1$  and  $v_2$  are not adjacent in  $G$ , it follows that  $v_2$  is not a parent of  $v_1$  in  $G$ . By the Markov condition,  $G \models v_1 \perp\!\!\!\perp v_2 | PA_G(v_1)$ . By assumption  $PA_G(v_1)$  contains no more than  $k - 2$  variables, and so  $v_1 \perp\!\!\!\perp v_2 | PA_G(v_1) \in \text{CDC}_G^k$ . Since  $G \equiv_k H$ , it follows that  $H \models v_1 \perp\!\!\!\perp v_2 | PA_G(v_1)$ , which is impossible as  $v_1$  and  $v_2$  are adjacent in  $H$  by supposition and hence, dependent given any set of variables by Lemma 1. □

**Corollary 3** *Let  $V$  be a set with  $n$  variables,  $G, H \in \text{DAG}_V$ , and suppose  $G \equiv_{n-1} H$ . Further, suppose that  $v_1$  and  $v_2$  are adjacent in  $H$  but not in  $G$ . Then every variable  $v_3 \in V \setminus \{v_1, v_2\}$  is a parent of  $v_1$  in  $G$ , or every such variable is a parent of  $v_2$  in  $G$ .*

We now prove Theorem 4. For the sake of contradiction, assume that there are two distinct pairs of variables  $\{v_1, v_2\}$  and  $\{v_3, v_4\}$  that are respectively adjacent in  $H$  but not in  $G$ . By corollary 3, because  $v_3$  and  $v_4$  are not adjacent in  $G$ , it follows that both  $v_1$  and  $v_2$  are parents of  $v_3$ , or both are parents of  $v_4$ . Without loss of generality, assume both are parents of  $v_3$ . Again, by corollary 3, because  $v_1$  and  $v_2$  are not adjacent in  $G$ , it follows that both  $v_3$  and  $v_4$  are parents of  $v_1$  or both are parents of  $v_2$ . In particular,  $v_3$  is a parent of  $v_1$  or it is a parent of  $v_2$ , thereby creating a cycle.

□

**Theorem 5** *Let  $V$  be any set of random variables and suppose that  $|V| = n$ .*

1. *For all  $G, H \in \text{DAG}_V$ , if  $G \equiv_{n-1} H$ , then  $H$  contains no more than  $n - 3$  unshielded colliders that  $G$  does not.*
2. *There exist  $G, H \in \text{DAG}_V$  such that  $G \equiv_{n-1} H$  and  $H$  contains  $n - 3$  unshielded colliders that  $G$  does not.*

**Proof of Theorem 5.1:** Let  $C$  be the set of triples  $l \rightarrow m \leftarrow r$  that form unshielded colliders in  $H$  but not  $G$ . Here, “l” stands for “left”, “m” stands for “middle”, and “r” stands for “right.” We break the proof into several smaller lemmas.

**Lemma 2** *For all triples  $l \rightarrow m \leftarrow r$  in  $C$ , the triple  $l, m, r$  does not form a trek in  $G$ .*

**Proof:** Suppose that, for the sake of contradiction,  $l - m - r$  is a trek in  $G$ . Since  $l$  and  $r$  are not adjacent in  $H$ , by Lemma 1, there is some set  $U \subseteq V \setminus \{l, r\}$  such that  $H \models l \amalg r | U$ . By Corollary 2, it must be the case that  $m \notin U$ . Since  $m \notin U$ , the trek  $l - m - r$  is active in  $G$  given  $U$ , and so  $G \not\models l \amalg r | U$ . Since  $U$  contains no more than  $n - 3$  elements (as  $l, m, r \notin U$ ), this contradicts the fact that  $G \equiv_{n-1} H$ . (End Proof Lemma 2)

**Lemma 3** *For all triples  $l \rightarrow m \leftarrow r$  in  $C$ , there is no  $l - r$  edge in  $G$ .*

**Proof:** Since  $l \rightarrow m \leftarrow r$  is an unshielded collider in  $H$ , there is some  $U \subseteq V \setminus \{l, m, r\}$  such that  $H \models l \amalg r | U$ . But by Lemma 1, if  $G$  contained an  $l - r$  edge, then  $G \not\models l \amalg r | U$ , contradicting the assumption that  $G \equiv_{n-1} H$ . (End Proof Lemma 3)

**Lemma 4** *For all triples  $l \rightarrow m \leftarrow r$  and  $l' \rightarrow m' \leftarrow r'$  in  $C$ , we have  $m = m'$ .*

**Proof:** Suppose not. Then there exist triples  $l \rightarrow m \leftarrow r$  and  $l' \rightarrow m' \leftarrow r'$  in  $C$  such that  $m \neq m'$ . Note that (i)  $G$  contains neither an  $l - r$  nor a  $l' - r'$  edge by Lemma 3, and (ii) the triples  $l, m, r$  and  $l', m', r'$  cannot be treks in  $G$  by Lemma 2. Since the triples are unshielded colliders in  $H$  but not in  $G$ , it follows from (i) and (ii) that:



1. in  $G$ , either  $l$  and  $m$  are not adjacent or  $m$  and  $r$  are not adjacent,  
AND
2. in  $G$ , either  $l'$  and  $m'$  are not adjacent or  $m'$  and  $r'$  are not adjacent

Since  $m \neq m'$ , it follows that  $H$  contains two edges that  $G$  does not, contradicting Theorem 4. (End Proof Lemma 4)

Next, we claim that we may write the triples in  $C$  so that for all  $l \rightarrow m \leftarrow r$  and  $l' \rightarrow m' \leftarrow r'$  in  $C$ , we have  $l = l'$ . The reasoning is analogous to that of Lemma 4. Suppose for the sake of contradiction there are triples  $l \rightarrow m \leftarrow r$  and  $l' \rightarrow m' \leftarrow r'$  in  $C$  such that  $l \neq l'$ ,  $l \neq r'$ ,  $l' \neq r$ , and  $r \neq r'$ . Using the same reasoning as in Lemma 4, one can show  $H$  contains at least two edges that  $G$  does not, contradicting Theorem 4.

Now we finish the proof of Theorem 5.1. Suppose for the sake of contradiction that  $H$  contains at least  $n - 2$  unshielded colliders  $l \rightarrow m \leftarrow r_i$ , where  $1 \leq i \leq n - 2$ . Notice that only the variable  $r_i$  contains the index  $i$ , as we may assume the middle and left element of the triples are identical by the reasoning above. By Lemma 2, none of the triples  $l \rightarrow m \leftarrow r_i$  are treks in  $G$ , and moreover, by Lemma 3,  $G$  contains no edges between  $l$  and  $r_i$  for all  $i \leq n - 2$ . So if  $l, m, r_i$  is not an unshielded collider in  $G$ , then it must be the case that one of the edges  $l - m$  or  $m - r_i$  is not present in  $G$ . Since this holds for all  $i \leq n - 2$  and  $H$  can contain at most one edge that  $G$  does not, it must be the case that  $l$  and  $m$  are not adjacent in  $G$ .

So we have shown that  $l$  is not adjacent to  $m$ , and  $l$  is not adjacent to  $r_i$  for all  $i \leq n - 2$  in  $G$ . In other words, there are no edges incident to  $l$  at all in  $G$ , and thus, there are no paths from  $l$  to any other variable in  $G$ . By Proposition 1, it follows that  $G \models l \amalg v | U$  for all  $v \in V$  and all  $U \subseteq V \setminus \{l, v\}$ . So  $G \models l \amalg m$ , but because  $l$  and  $m$  are adjacent in  $H$  we have  $H \not\models l \amalg m$  by Lemma 1. So  $G \not\equiv_{n-1} H$ , contradicting assumption.

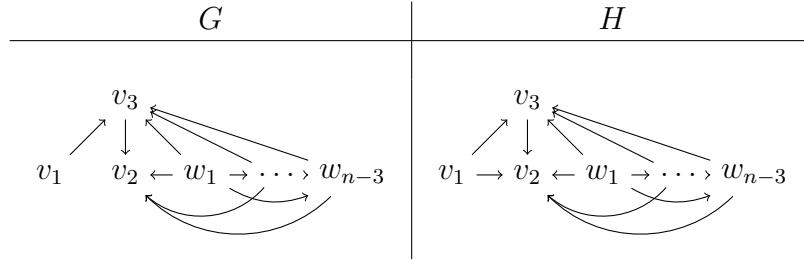
□

**Proof of Theorem 5.2:** Suppose  $V$  has  $n \geq 3$  elements, and let  $\{v_1, v_2, v_3, w_1, w_2, \dots, w_{n-3}\}$  be an enumeration of such elements. Let  $G$  be the graph containing all and only the following edges:

1. An edge from  $v_1$  and  $v_3$
2. An edge from  $w_i$  to  $v_2$  for all  $i \leq k$

3. An edge from  $v_3$  to  $v_2$
4. An edge from  $w_i$  to  $w_j$  for all  $i < j$

Let  $H$  be the result of adding the edge  $v_1 \rightarrow v_2$  to  $G_1$ . Note that  $v_1 \rightarrow v_2 \leftarrow w_i$  is an unshielded collider in  $H$  but not in  $G$  for all  $i \leq n-3$ . So  $H$  contains  $n-3$  unshielded colliders that  $G$  does not.



**Figure 8**

The only non-adjacent pairs of vertices in  $G$  consist of (i)  $v_1$  and  $v_2$ , and (ii)  $v_1$  and  $w_i$  for  $i \leq k$ . So the only conditional independence constraints that hold in  $G$  concern these pairs.

First, consider the pair consisting  $v_1$  and  $v_2$ . Since  $v_1 \rightarrow v_3 \rightarrow v_2$  is a trek, it follows that  $v_1$  and  $v_2$  are dependent on any set  $U$  not containing  $v_3$ . Consider any  $U \subseteq V$  containing  $v_3$  and no more than  $n-3$  variables. Since  $U$  contains no more than  $n-3$  variables and  $v_3 \in U$ , there is some  $w_i \notin U$ , and hence,  $v_1 \rightarrow v_3 \leftarrow w_i \rightarrow v_2$  is a  $d$ -connecting path between  $v_1$  and  $v_2$  given  $U$  in  $G$ . So there is no CIC  $\varphi$  of the form  $v_1 \perp\!\!\!\perp v_2 | U$  in  $\text{CIC}_G^{n-1}$ .

Next consider any pair consisting of  $v_1$  and  $w_i$  for  $i \leq n-3$ . In  $G$ , the only paths between  $v_1$  and  $w_i$  pass through  $v_2$  or  $v_3$  (or both). Moreover, on all such paths, either  $v_2$  or  $v_3$  is a collider. It follows that  $v_1$  and  $w_i$  are unconditionally independent, and moreover, that  $v_1$  and  $w_i$  are dependent given any set  $U$  containing either  $v_2$  or  $v_3$ . So the only other relevant CICs to consider are of the form  $v_1 \perp\!\!\!\perp w_i | W$ , where  $W$  is a collection of the  $w_j$ 's other than  $w_i$ . It is easy to see (again using the relation of  $d$ -connection and conditional dependence) that all such CICs are satisfied by  $G$ . In sum:

$$\text{CIC}_G^{n-1} = \bigcup \{v_1 \perp\!\!\!\perp w_i | W : W \subseteq \{w_1, \dots, w_{n-3}\} \setminus \{w_i\}\}_{i \leq n-3}$$

Now consider the graph  $H$ . The only non-adjacent pairs of vertices in  $H$  consist of  $v_1$  and some  $w_i$  for  $i \leq k$ . So the only conditional independence

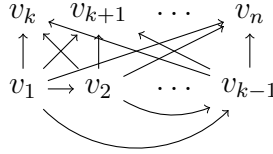
constraints that hold in  $H$  concern these pairs. By the same reasoning as with  $G$ , it is easy to see that

$$\text{CIC}_H^{n-1} = \bigcup \{v_i \amalg w_i | W : W \subseteq \{w_1, \dots, w_{n-3}\} \setminus \{w_i\}\}_{i \leq n-3}$$

So  $G \equiv_{\mathcal{U}} H$ , but  $H$  contains  $n - 3$  unshielded colliders that  $G$  does not. □

**Theorem 6** *For any  $V$  and any  $k \leq n = |V|$ . There exist  $G, H \in \text{DAG}_V$  such that  $G \equiv_k H$  but  $H$  contains  $\binom{n-k+1}{2}$  many edges that  $G$  does not.*

**Proof of 6:** Enumerate the vertices of  $V = \{v_1, v_2, \dots, v_n\}$ . Construct  $G$  as follows. For any  $i \leq k - 1$  and any  $j > i$ , draw an edge from  $v_i$  to  $v_j$ . The graph  $G$  contains no other edges (See Figure 9). Construct  $H$  by drawing an edge from  $v_i$  to  $v_j$  for all  $i < j \leq n$ , so that  $H$  is a complete graph. Notice  $G$  is a subgraph of  $H$  and that  $H$  contains  $\binom{n-k+1}{2}$  edges that  $G$  does not, as every pair  $\{v_i, v_j\}$  for  $k \leq i < j$  is adjacent in  $H$  but not  $G$ . We claim  $G \equiv_k H$ .



**Figure 9**

Since all variables in  $H$  are adjacent, by Lemma 1, it follows that  $\text{CDC}_H^V = \emptyset$ , and hence,  $\text{CIC}_H^k = \emptyset$ . To show  $G \equiv_k H$ , then, it is necessary to show that  $\text{CIC}_G^k = \emptyset$ . Since  $v_i$  and  $v_j$  are adjacent if  $i < j$  and  $i \leq k - 1$ , by Lemma 1, it follows that the only independences in  $G$  are of the form  $v_i \amalg v_j | U$  where  $i, j \geq k$ . So let  $i, j \geq k$ , and consider any arbitrary conditional independence constraint  $v_i \amalg v_j | U \in \text{CIC}^k$ . Since  $|U| \leq k - 2$ , it follows that there is some  $v_m \notin U$  such that  $m < k \leq i, j$ . By definition of  $G$ , the triple  $v_i \leftarrow v_m \rightarrow v_j$  is a trek in  $G$ , and since  $v_m \notin U$ , it follows that this trek is active given  $U$ . Hence,  $G \not\models v_i \amalg v_j | U$ , or in other words,  $v_i \amalg v_j | U \notin \text{CIC}_G^k$ . Since  $v_i \amalg v_j | U \in \text{CIC}^k$  was chosen arbitrarily, it follows that  $\text{CIC}_G^k = \emptyset$  as desired. □

Next we prove:

**Theorem 7** *Suppose  $G$  has fewer than  $2k - 2$  edges. Then  $[G] = [G]_k$ .*

Here's the idea of the proof. Call the set of edges in a graph its **skeleton**. We first show that if two  $k$ -equivalent graphs possess different skeletons, then they have at least  $2k - 2$  edges. So if a graph has fewer than  $2k - 2$  edges, then its skeleton is determined by conditional independence constraints involving  $k$  variables or fewer. We then show that  $k$ -equivalent graphs with the same skeletons and sufficiently few edges also share the same set of unshielded colliders, and hence, are Markov equivalent by Proposition 2.

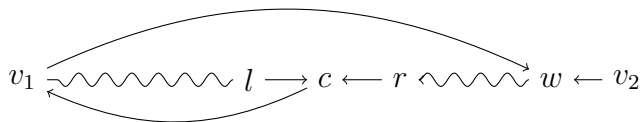
**Lemma 5** *Suppose  $v_1$  and  $v_2$  are not adjacent in  $G$ . Then there is  $U \subseteq MB_G(v_2) \setminus (UC_G(v_1, v_2) \cup \{v_1\})$  such that  $G \models v_1 \perp\!\!\!\perp v_2 | U$ .*

**Proof:** Suppose not for the sake of contradiction. Then, by Proposition 3, there is an inducing path  $\pi$  between  $v_1$  and  $v_2$  over  $MB_G(v_2) \setminus (UC_G(v_1, v_2) \cup \{v_1\})$ . Let  $w$  be the variable adjacent to  $v_2$  on  $\pi$ .

**Claim:**  $w$  is a collider on  $\pi$ .

If not, then  $w \notin MB_G(v_2) \setminus (UC_G(v_1, v_2) \cup \{v_1\})$  by definition of inducing path. But since  $w$  is adjacent to  $v_2$ , it is a member of the Markov Blanket of  $v_2$ . So  $w \in UC_G(v_1, v_2) \cup \{v_1\}$ . If  $w = v_1$ , then  $v_1$  and  $v_2$  would be adjacent, contradicting assumption. If  $w \in UC_G(v_1, v_2)$ , then  $w$  is a descendant of both  $v_1$  and  $v_2$ . I claim that  $w$  is also an ancestor of  $v_1$ , contradicting the fact that  $G$  is acyclic. Why?

Since  $w$  is a non-collider on  $\pi$  and a descendant of  $v_2$ , it follows that there is an edge “out of”  $w$  towards  $v_1$  on  $\pi$  (See **Figure 10** below). If the subpath of  $\pi$  is not directed from  $w$  to  $v_1$ , then it follows that there is some collider between  $w$  and  $v_1$  on  $\pi$ . Let  $c$  be the collider closest to  $w$ , and note that  $c$  is a descendant of  $w$  and  $v_2$ . Since  $\pi$  is an inducing path,  $c$  is an ancestor of either  $v_1$  or  $v_2$ . By acyclicity, it is not an ancestor of  $v_2$ . So  $c$  is an ancestor of  $v_1$ . Since  $w$  is an ancestor of  $c$ , it follows that  $w$  is an ancestor of  $v_1$ , as desired. This finishes the proof of the claim.



### Figure 10

So we have shown that  $w$  is a collider on  $\pi$ . Hence, it is an ancestor of either  $v_1$  or  $v_2$  by definition of inducing path. By acyclicity, it must be an ancestor of  $v_1$ . Let  $z$  be the variable (other than  $v_2$ ) adjacent to  $w$  on  $\pi$ . Notice that  $z$  is a parent of  $w$  as  $w$  is a collider on  $\pi$ . Moreover, since  $w$  is a collider on  $\pi$ , it follows that  $z$  is not a collider on  $\pi$ . Hence,  $z \notin MB_G(v_2) \setminus (UC_G(v_1, v_2) \cup \{v_1\})$ . But clearly  $z \in MB_G(v_2)$ , and so it follows that  $z \in UC_G(v_1, v_2) \cup \{v_1\}$  by definition of inducing path. If  $z = v_1$ , then  $w$  would be a descendant of  $v_1$ , which is impossible as  $w$  is an ancestor of  $v_1$ . So  $z \in UC_G(v_1, v_2)$ . So  $z$  is a descendant of  $v_1$ . Hence,  $w$  is a descendant of  $v_1$ , as  $w$  is a child of  $z$ . Again, this is a contradiction as  $w$  is an ancestor of  $v_1$ .

□

**Lemma 6** *Let  $G$  and  $H$  be DAGs over  $n$ -variables and  $k < n$ . Suppose  $G \equiv_k H$  and that  $v_1$  and  $v_2$  are adjacent in  $H$  but not  $G$ . Then  $G$  contains at least  $2k - 2$  edges.*

**Proof:** By Lemma 1, it follows that either  $v_1$  or  $v_2$  has  $k - 1$  parents in  $G$ . Without loss of generality, assume that  $v_1$  has at least  $k - 1$  parents in  $G$ . So if we let  $E_G(v_1)$  be the set of edges incident to  $v_1$ , then  $|E_G(v_1)| \geq k - 1$ .

By Lemma 5, there is some set  $U \subseteq MB_G(v_2) \setminus \{v_1\}$  such that  $G \models v_1 \amalg v_2 | U$ . As  $v_1$  and  $v_2$  are adjacent in  $H$ , by Corollary 1, we have that  $H \not\models v_1 \amalg v_2 | W$  for all  $W \subseteq V \setminus \{v_1, v_2\}$ . Since  $G \equiv_k H$  and  $G \models v_1 \amalg v_2 | U$ , it follows that  $|U| \geq k - 1$ . As  $U \subseteq MB_G(v_2) \setminus \{v_1\}$ , it must be the case that  $|MB_G(v_2) \setminus \{v_1\}| \geq k - 1$ . We claim we can define an injective function  $f$  from  $MB_G(v_2) \setminus \{v_1\}$  into  $E_G \setminus E_G(v_1)$ , where, recall,  $E_G$  is the set of edges in  $G$ . This would finish the proof of the theorem as then:

$$\begin{aligned} |E_G \setminus E_G(v_1)| &\geq |MB_G(v_2) \setminus \{v_1\}| \\ &\geq |U| \\ &\geq k - 1 \end{aligned}$$

which would then entail

$$|E_G| = |E_G \setminus E_G(v_1)| + |E_G(v_1)| \geq (k - 1) + (k - 1) \geq 2k - 2$$

How can one define  $f$ ? For each variable  $w$  that is adjacent to  $v_2$ , let  $f(w)$  be the edge that is incident to  $v_2$  and  $w$ . For each variable  $w$  that forms an unshielded collider  $v_2 \rightarrow z \leftarrow w$  with respect to  $v_2$ , let  $f(w)$  be the  $z \leftarrow w$  edge. Notice that  $f(w)$  is not a member of  $E_G(v_1)$  in either case as we have assumed that  $w \in MB_G(v_2) \setminus \{v_1\}$ , and in particular  $w \neq v_1$ . The function  $f$  is clearly injective, and so we are done.

□

**Lemma 7** *Suppose  $G$  has fewer than  $2k - 2$  edges and that  $G \equiv_k H$ . Then  $G$  and  $H$  have the same skeleton.*

**Proof:** We show that (i) any two variables that are adjacent in  $H$  are also adjacent in  $G$ , and (ii) any two variables adjacent in  $G$  are adjacent in  $H$ .

To show (i), suppose for the sake of contradiction that  $v_1$  and  $v_2$  were adjacent in  $H$  but not  $G$ . Then, by Lemma 6, it would follow that  $G$  contains at least  $2k - 2$  edges, contradicting assumption. So we have established (i).

Since any two variables adjacent in  $H$  are also adjacent in  $G$  and, moreover,  $G$  contains fewer than  $2k - 2$  edges, it follows that  $H$  contains fewer than  $2k - 2$  edges. By the same reasoning as above,  $G$  cannot contain an edge that  $H$  does not, as otherwise  $H$  would contain at least  $2k - 2$  edges.

□

**Proof of Theorem 7:** By Lemma 7, if  $G \equiv_k H$ , then  $G$  and  $H$  have identical skeletons. Hence, by Lemma 2, it suffices to show that if  $G \equiv_k H$ , then  $G$  and  $H$  have the same set of unshielded colliders. Suppose not. Then there is some triple  $v_1 \rightarrow v_2 \leftarrow v_3$  that is an unshielded collider in  $H$  but not in  $G$ , or vice versa. Without loss of generality, assume the unshielded collider occurs in  $H$  but not  $G$ . As  $G$  and  $H$  have the same skeleton, the triple  $v_1 - v_2 - v_3$  forms a trek in  $G$ .

By the Markov condition, there is some  $U_G$  such that  $v_1 \amalg v_3 | U_G \in \text{CIC}_G^V$  and either  $U_G \subseteq PA_G(v_1)$  or  $U_G \subseteq PA_G(v_2)$ . Without loss of generality, assume  $U_G \subseteq PA_G(v_1)$ . As the triple  $v_1 - v_2 - v_3$  is a trek in  $G$  by assumption, it follows that  $v_2 \in U_G$ . As  $G \equiv_k H$  and  $v_1 \rightarrow v_2 \leftarrow v_3$  is an active path from  $v_1$  to  $v_3$  given  $U_G$  in  $H$ , it follows that  $|U_G| \geq k - 1$ . Hence,  $v_1$  has at least  $k - 1$  parents in  $G$ . Let  $E_G(v_1)$  and  $E_H(v_1)$  be the set of edges incident to  $v_1$  in  $G$  and in  $H$  respectively. Because  $G$  has fewer than  $2k - 2$  edges and  $E_G(v_1)$  contains at least  $k - 1$  members, it follows that  $E_G \setminus E_G(v_1)$  contains

no more than  $k - 2$  edges. Because  $H$  and  $G$  have the same skeleton, it follows that  $E_H \setminus E_H(v_1)$  likewise has no more than  $k - 2$  edges.

As in Lemma 6, one can define an injection from  $MB_H(v_3) \setminus \{v_1\}$  into  $E_H \setminus E_H(v_1)$ , and so  $|MB_H(v_3) \setminus \{v_1\}| \leq k - 2$ . Moreover, by Lemma 5, there is some  $U_H \subseteq MB_H(v_3) \setminus (UC_H(v_1, v_3) \cup \{v_1\})$  such that  $v_1 \amalg v_3 | U_H \in \text{CIC}_H^V$ . Since  $U_H \subseteq MB_H(v_3) \setminus (UC_H(v_1, v_3) \cup \{v_1\})$  and  $|MB_H(v_3) \setminus \{v_1\}| \leq k - 2$ , it follows that  $|U_H| \leq k - 2$ . Hence,  $v_1 \amalg v_3 | U_H \in \text{CIC}_H^k$ . Because  $G \equiv_k H$ , it follows that  $v_1 \amalg v_3 | U_H \in \text{CIC}_G^k$ .

To obtain a contradiction, we note that, because the triple  $v_1 \rightarrow v_2 \leftarrow v_3$  is an unshielded collider in  $H$  and  $v_1 \amalg v_3 | U_H \in \text{CIC}_H^k$ , it must be the case that  $v_2 \notin U_H$ . However, because the triple forms a trek in  $G$  and  $v_1 \amalg v_3 | U_H \in \text{CIC}_G^k$ , it must be the  $v_2 \in U_H$ .

□

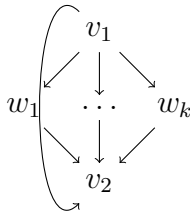
Let  $k$  be a fixed natural number,  $V$  be any set of random variables of size  $n \geq k$ , and  $G \in \text{DAG}_V$ . Say  $G$  is  $k$ -underdetermined if there exists  $H \in \text{DAG}_V$  such that  $G \not\equiv H$  but  $G \equiv_k H$ . Let  $p_k(n)$  be the proportion of DAGs over  $n$  vertices that are  $k$ -underdetermined. Let  $E_k(G)$  be the maximum number of DAGs  $H_1, H_2, \dots, H_m$  such that  $H_i \not\equiv G$  but  $H_i \equiv_k G$ . Finally, let  $E_k(n)$  be the average of  $E_k(H)$  over all graphs  $H$  over  $n$  vertices. Then:

**Theorem 8**  $p_k(n) \rightarrow 1$  as  $n \rightarrow \infty$

**Theorem 9**  $E_k(n) \rightarrow \infty$  as  $n \rightarrow \infty$

In the following proofs, let  $G_k$  denote the DAG (pictured below) containing  $k$  vertices  $\{v_1, v_2, w_1, \dots, w_k\}$  and the following edges.

1. An edge from  $v_1$  to  $w_i$  for all  $1 \leq i \leq k - 2$
2. An edge from  $w_i$  to  $v_2$  for all  $1 \leq i \leq k - 2$
3. An edge from  $v_1$  to  $v_2$



## Figure 11

Let  $G_{k,m}$  denote the DAG containing  $km$  vertices, with  $m$  disconnected copies of  $G_k$ . To prove Theorems 8 and 9, we employ two lemmas.

**Lemma 8** *Let  $n > k$  and  $G \in \text{DAG}_n$ . Suppose  $H$  contains an isomorphic copy of  $G_k$ , and let  $G$  be the graph obtained by deleting the edge from  $v_1$  to  $v_2$  in  $H$ . Then  $H \equiv_k G$ .*

**Lemma 9** *Let  $H$  be a DAG with  $k$  vertices. For each  $n \geq k$ , let  $p_n(H)$  be the proportion of DAGs with  $n$  vertices containing an isomorphic copy of  $H$  as a subgraph. Then  $\lim_{n \rightarrow \infty} p_n(H) = 1$*

We first prove the Theorems using these two lemmas. We then prove Lemma 8. A proof of Lemma 9 requires employing probabilistic combinatorial techniques and applying some fairly technical results from Bender et. al. (1986) and McKay (1989). Hence, a full proof is omitted because it is beyond the scope of this paper. Contact the author for a complete proof.

**Proof of Theorem 8:** By Lemma 9, it follows that  $p_n(G_k) \rightarrow 1$ . By Lemma 8, every graph containing  $G_k$  as a subgraph is  $k$ -underdetermined, as an edge can be removed from each such graph to obtain a  $k$ -equivalent graph. So it follows that  $p_k(n) \rightarrow 1$ .

□

**Proof of Theorem 9:** We must show that, for all  $m \in \mathbb{N}$ , there is some  $n_m \in \mathbb{N}$  such that  $E_k(n) \geq m$  for all  $n \geq n_m$ . To this end, note that by Lemma 9, it follows that  $p_n(G_{k,2m}) \rightarrow 1$ . So let  $n_m$  be the least natural number such that  $p_n(G_{k,2m}) > \frac{1}{2}$  for all  $n \geq n_m$ . By Lemma 8, one can remove  $2m$  many edges from every graph containing  $G_{k,2m}$ , and hence, if  $G$  contains  $G_{k,2m}$  as a subgraph, it follows that  $E_k(G) \geq 2m$ . Thus,  $E_k(n) > \frac{1}{2} \cdot 2m = m$  for all  $n \geq n_m$ , as desired.

□

**Proof of Lemma 8:** As  $H$  is obtained by adding an edge to  $G$ , it follows that  $\text{CDC}_G^k \subseteq \text{CDC}_H^k$ . So it suffices to show that  $\text{CDC}_H^k \subseteq \text{CDC}_G^k$ . To this end, let  $\varphi \in \text{CDC}_H^k$  be the assertion  $\neg z_1 \amalg z_2 | S$ . So by Proposition 1, it follows that there is a  $d$ -connecting path  $\pi_H$  from  $z_1$  to  $z_2$  given  $S$  in  $H$ . We construct a

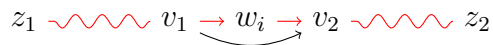


d-connecting path  $\pi_G$  from  $z_1$  to  $z_2$  given  $S$  in  $G$ .

**Case 1:** Suppose  $\pi_H$  is also a path in  $G$ . Then it's easy to show that  $\pi_G = \pi_H$  is likewise d-connecting in  $G$ . Why? Every non-collider on  $\pi_G$  is not a member of  $S$  because  $\pi_H$  is active given  $S$  in  $H$ . Moreover, every collider  $c$  on  $\pi_G$  is also a collider on  $\pi_H$ . Since  $\pi_H$  is active given  $S$ , it follows that either  $c$  or one of  $c$ 's descendants in  $H$  is a member of  $S$ . But since  $H$  is obtained from  $G$  by adding an edge from  $v_1$  to  $v_2$ , and  $v_1$  is already an ancestor of  $v_2$  in  $G$  (as  $G$  contains  $G_k$  as a subgraph), it follows that the set of descendants of  $c$  in  $H$  and in  $G$  are identical.

**Case 2:** Suppose  $\pi_H$  is not a path in  $G$ . As  $H$  is obtained from  $G$  by adding an edge from  $v_1$  to  $v_2$ , it follows that  $\pi_H$  contains the edge  $v_1 \rightarrow v_2$ .

Since  $|S| \leq k - 2$ , there is some  $i \leq k - 2$  such that  $w_i \notin S$ . Consider the variable sequence obtained by modifying  $\pi_H$  by removing the  $v_1 \rightarrow v_2$  edge, and replacing it with the trek  $v_1 \rightarrow w_i \rightarrow v_2$ , i.e. consider  $\alpha = (\pi_H \downarrow v_1) \frown \langle v_1, w_i, v_2 \rangle \frown (\pi_H \uparrow v_2)$ . The sequence  $\alpha$  is pictured below in red, and segments of  $\pi_H$  that are not also segments of  $\alpha$  are drawn in black.

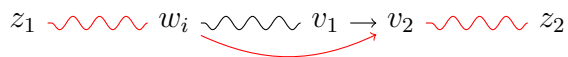


**Figure 12**

If  $\alpha$  is a path in  $G$ , then let  $\pi_G = \alpha$ . As  $\pi_H$  is active given  $S$ , both  $\pi_H \downarrow v_1$  and  $\pi_H \uparrow v_2$  are active given  $S$  in  $G$ . Moreover, since  $w_i \notin S$  and is a non-collider on  $\pi_G$ , it follows that  $w_i$  is active given  $S$  in  $G$ . Finally,  $v_1$  and  $v_2$  are both active given  $S$  in  $G$  because (i)  $v_1$  is a non-collider on both  $\pi_H$  and  $\pi_G$ , and hence, is active on one if and only if it's active on the other, and (ii)  $v_2$  is a collider on  $\pi_H$  if and only if it is a collider on  $\pi_G$ , and hence, is active on one if and only if it's active on the other.

If  $\alpha$  is not a path, then  $w_i$  already occurs on  $\pi_H$ . So there are two cases:

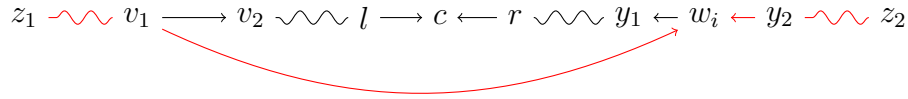
**Case 2a:** Suppose  $w_i$  occurs before  $v_1$  on  $\pi_H$ . Define  $\pi_G = (\pi_H \downarrow w_i) \frown \langle w_i, v_2 \rangle \frown (\pi_H \uparrow v_2)$ . The path  $\pi_G$  is indicated in red in the figure below, and segments of  $\pi_H$  are indicated by black lines and curves.



**Figure 13**

The sequence  $\pi_G$  is a path because  $\pi_H$  is a path, and it's d-connecting given  $S$  in  $G$  precisely because  $\pi_H$  is. Why? Since  $\pi_H$  is active given  $S$ , it follows that both of the segments  $\pi_H \downarrow w_i$  and  $\pi_H \uparrow v_2$  are active given  $S$  in  $G$ , and the variable  $w_i$  is active because it's a non-collider on  $\pi_G$  and (by choice) not an element of  $S$ .

**Case 2b:** Suppose  $w_i$  occurs after  $v_2$  on  $\pi_H$ . Define  $\pi_G = (\pi_H \downarrow v_1) \frown \langle v_1, w_i \rangle \frown (\pi_H \uparrow w_i)$ . The sequence  $\pi_G$  is a path because  $\pi_H$  is. To show it's d-connecting given  $S$  in  $G$ , note first that, since  $\pi_H$  is active given  $S$ , it follows that both of the segments  $\pi_H \downarrow v_1$  and  $\pi_H \uparrow w_i$  are active given  $S$  in  $G$ . The only thing to show is that  $w_i$  is active on  $\pi_G$ . Suppose not for the sake of contradiction. Then  $w_i$  is a collider on  $\pi_G$ , as is pictured below on the red path. Segments of  $\pi_H$  are indicated by black lines. Since  $w_i \notin S$  and  $\pi_H$  is active in  $H$ , we know that  $w_i$  is not a collider on  $\pi_H$ . So it follows that  $\pi_H$  contains a segment of the form  $y_1 \leftarrow w_i \leftarrow y_2$ .



**Figure 14**

Consider the segment of  $\pi_H$  lying between  $v_1$  and  $w_i$  in  $H$ . Since this segment contains edges directed out of both  $v_1$  and  $w_i$ , there must be some collider between  $v_1$  and  $w_i$  on  $\pi_H$ , and in particular, one such collider  $c$  is closest to  $w_i$  on  $\pi_H$ . So  $c$  is a descendant of  $w_i$  in both  $G$  and  $H$ . Since  $\pi_H$  is active given  $S$  in  $H$ , either  $c$  or one of its descendants in  $H$  is a member of  $S$ . Since  $c$ 's descendants in  $G$  and  $H$  are identical, it follows that either  $c$  or one of its descendants in  $G$  is a member of  $S$ . Since  $w_i$  is an ancestor of  $c$ , either  $w_i$  or one of its descendants in  $G$  is a member of  $S$ . But then  $w_i$  is active on  $\pi_G$ , contradicting assumption.

□