

Graphical Aspects of Causal Models (Draft Copy)

Thomas Sadanand Verma

September 1, 1992

Abstract

The principle of causation is an invaluable part of human cognition. People are very proficient at accurately generating and understanding causal explanations of physical mechanisms and processes. However, despite the efforts of many philosophers engaged in centuries of debate, a universally accepted definition of causality is lacking. As a modest contribution to this debate, this dissertation proposes a formal definition of a causal model and explores the ramifications of that definition. To this end, two fundamental questions are examined in detail:

- “Given that causal model C accurately describes the behavior of physical process P , what set I of conditional independence statements should be revealed from observation of P ? ”
- “Given that set I of conditional independence statements is revealed from observation of process P , what features, if any, do all causal models which accurately predict I have in common? ”

Contents

0	Introduction	1
0.1	The Causal Modeling Framework	4
0.1.1	Directed Acyclic Graphs	5
0.1.2	Assumptions	6
0.2	Preview of Technical Results	6
0.2.1	Bayesian Networks	7
0.2.2	Causal Models and Theories	7
0.2.3	Latent Structures	8
0.2.4	Inferring Causal Relationships from Observational Data	8
0.3	Basic Concepts	8
0.3.1	Probability Theory	9
0.3.2	Graph Theory	10
1	Bayesian Networks	12
1.1	Dependency Models	13
1.1.1	Graphoids	16
1.2	Directed Acyclic Graphs (dags)	18
1.2.1	Recursive Decomposition	18
1.3	Historical and Bibliographic Remarks	23
2	Causal Models and Theories	24
2.1	Formal Definitions	24
2.1.1	Probabilistic Semantics	26
2.2	Equivalence	29
2.3	Recovery	34
2.3.1	The DAG Construction Algorithm	35
2.3.2	Correctness	36

2.3.3	Complexity Analysis	40
2.3.4	Extensions and Improvements	40
2.3.5	Naive Representation of Graphoids	42
2.4	Historical and Bibliographic Remarks	44
3	Latent Structures	45
3.1	Formal Definition	45
3.1.1	Probabilistic Semantics	45
3.1.2	Dependency Equivalence versus Equivalence	46
3.1.3	Projections	47
3.2	Invariant Properties	52
3.2.1	Inducing Paths	52
3.2.2	Vee Structures	55
3.2.3	Kite Structures	56
3.2.4	Induced Graphs	57
3.3	Historical and Bibliographic Remarks	59
4	Inferring Causal Relationships from Observational Data	60
4.1	Recovering Latent Structures	62
4.2	Probabilistic Criteria for Causal Relations	64
4.3	Causal Intuition and Virtual Experiments	68
4.4	Non-Temporal Causation and Statistical Time	72
4.5	Conclusions	75

Chapter 0

Introduction

A formal understanding of the ability to reason with and about causal information is one of many prerequisites to a realistic modeling of rational human behavior. Policy analysis invariably requires distinguishing between cause and effect. No rational person would manipulate an effect in an attempt to modify the cause. However, it is not always feasible to ascertain cause and effect relationships from manipulative experiments, thus it is necessary to develop a theory with more passive means.

This work concentrates on the process by which a rational person would come to accept the truth of a causal rule especially when manipulative experiments are not available. For example, given the statistical finding: “Women who drink bottled water during pregnancy have a lower incidence of birth defects”, some people might jump to the conclusion that drinking bottled water has a causal influence upon the possibility of having a birth defect. Rationality would dictate otherwise. Given that statistic alone, there is no way to rule out another possible causal explanation — drinking bottled water and the incidence of birth defects are spuriously associated by some unknown cause, such as health consciousness which influences the incidence of both drinking bottled water and of birth defects, as shown in Figure 0.1.

Most AI works have given the term “cause” a procedural semantics, attempting to match the way people use it in reasoning tasks, but have not been concerned with the experience that prompts people to believe that “*a* causes *b*”, as opposed to, say, “*b* causes *a*” or “*c* causes both *a* and *b*.“ The question of choosing an appropriate causal ordering received some attention in qualitative physics, where certain interactions attain directionality despite

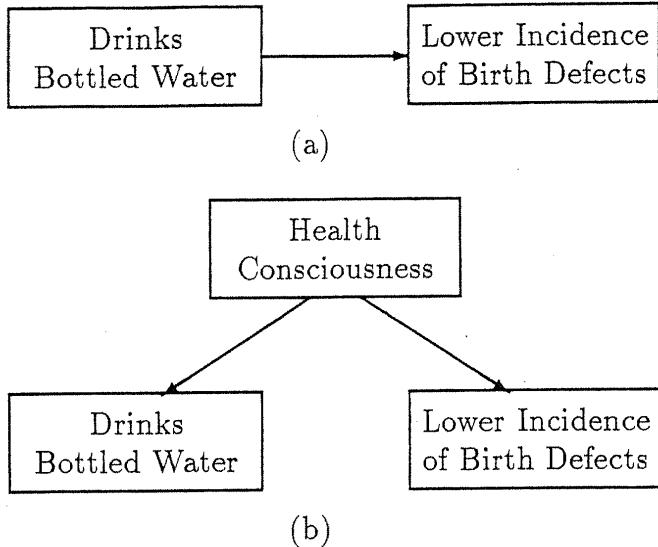


Figure 0.1: Two causal explanations for the observed correlation between drinking bottled water and birth defects.

the instantaneous and symmetrical nature of the underlying equations, as in “current causing a voltage drop across the resistor” [FG86]. In some systems causal ordering is defined as the ordering at which subsets of variables can be solved independently of others [IS86], in other systems it follows the way a disturbance is propagated from one variable to others [KB86]. Yet these choices are made, as a matter of convenience, to fit the structure of a given theory and do not reflect features of the empirical environment that compelled the formation of the theory.

An empirical semantics for causation is important for several reasons. First, an intelligent system attempting to build a workable model of its environment cannot rely exclusively on preprogrammed causal knowledge, but must be able to translate direct observations to cause-and-effect relationships. Second, by tracing empirical origins we stand to obtain an independent gauge for testing the soundness and completeness of the many logics proposed for causal reasoning while providing a proper account of causal utterances such as “*a* explains *b*”, “*a* suggests *b*”, “*a* tends to cause *b*”, and “*a* actually causes *b*”, etc.

While the notion of causation is conceptually associated with those of necessity and functional dependence, causal expressions often tolerate exceptions, primarily due to missing variables and coarse description. We say, for example, “reckless driving causes accidents” or “you will fail this course because of your laziness”. Suppes [Sup70] has argued convincingly that most causal utterances in ordinary conversation reflect probabilistic, not deterministic relations¹. Thus, probability theory should provide a natural language for capturing causation [Rei56, Goo83]. This is especially true when we attempt to infer causation from (noisy) observations – probability calculus remains an unchallenged formalism when it comes to translating statistical data into a system of revisable beliefs.

However, given that statistical analysis is driven by covariation, not causation, and assuming that most human knowledge derives from statistical observations, we must still identify the clues that prompt people to perceive causal relationships in the data, and we must find a computational model that emulates this perception.

Temporal precedence is normally assumed essential for defining causation, and it is undoubtedly one of the most important clues that people use to distinguish causal from other types of associations. Accordingly, most theories of causation invoke an explicit requirement that a cause precedes its effect in time [Goo83, Rei56, Sho88, Sup70]. Yet temporal information alone cannot distinguish genuine causation from spurious associations caused by unknown factors as evidenced by the classical fallacy — *post hoc ergo propter hoc*. In fact the statistical and philosophical literature has adamantly warned analysts that, unless one knows in advance all causally relevant factors, or unless one can carefully manipulate some variables, no genuine causal inferences are possible [Car89, Cli83, ES83, Fis53, Gar88, Hol86, Sky86]². Neither condition is realizable in normal learning environments, and the question remains how causal knowledge is ever acquired from experience.

A first step toward answering this question should be to analyze precisely what statistical relationships are to be expected from a phenomenon that is governed by a known causal model. To that end we first analize Bayesian networks (Chapter 1) which are used to provide a formal definition

¹See [DP90] for a treatment of causation in the context of deterministic data.

²Some of the popular quotes are: “No causation without manipulation”, [Hol86], “No causes in, no causes out”, [Car89] “No computer program can take account of variables that are not in the analysis”, [Cli83].

of causal models (Chapter 2). A complete characterization of the conditional independence relationships that are compelled by this a model are examined in Section 2.1. Subsequently, in Section 2.2 we identify properties that are common to all causal models that are consistent with a set of observed independencies. These properties lead to an algorithm that will determine if a set of independencies is consistent with a unique causal model, and if so generates it (Section 2.3). Latent variables are added to the model in Chapter 3 and this analysis culminates in Chapter 4 with the introduction of a minimal model semantics of causation.

Using this semantics we show that in many cases genuine causal influences can be distinguished from spurious covariations and, moreover, the direction of causal influences often can be determined without resorting to chronological information. (Although, when available, chronological information can significantly simplify the modeling task.) Such semantics should be applicable, therefore, to the organization of concurrent events or events whose chronological precedence cannot be determined with precision, (e.g. “old age explains disabilities”) in the spirit of Glymour [GSS87] and Simon [Sim54].

0.1 The Causal Modeling Framework

The task of causal modeling can be viewed as an identification game that scientists play against Nature. Relativity and quantum physics notwithstanding, Nature possesses stable causal mechanisms which on some fundamental level are deterministic functional relationships between variables, some of which are unobservable. These mechanisms are organized in the form of an acyclic schema which the scientist attempts to identify – a directed acyclic graph. The nodes of the dag correspond to the variables under analysis, while the directed arcs denote direct causal influences. A causal model serves as a blueprint for forming a “causal theory” – a precise specification of how each variable is influenced by its parents in the dag. Nature is at liberty to select arbitrary functional relationships between each effect and its causes and then to perturb these relationships by introducing arbitrary (yet mutually independent) disturbances. These disturbances reflect exceptions to the available deterministic theories which Nature chooses to govern by some undisclosed

probability function.³

Once a causal theory is formed, it defines a joint probability distribution over the variables in the system, and this distribution reflects some features of the causal model (e.g., each variable must be independent of its grandparents, given the values of its parents). Nature then permits the scientist to inspect a select subset of “observed” variables, and to ask questions about the probability distribution over the observables, but hides the underlying causal theory as well as the structure of the causal model.

0.1.1 Directed Acyclic Graphs

The choice of directed acyclic graphs (dags) as the basic language for causal models takes advantage of the ability of dags to naturally embody the notions of directionality and transitivity that are basic to human perception of causation. This choice follows a long tradition in statistics, philosophy (Reichenbach) [Rei56] economics and psychology (Duncan). Starting with the pioneering work of Sewal Wright [Wri21] who introduced path analysis to statistics, through the more recent developments of Bayesian networks and influence diagrams, dag structures have served for encoding causal influences between variables as well as between actions and variables.

Even statisticians who usually treat causality with extreme caution have found the structure of dags to be a useful model for explanatory purposes. N. Wermuth, for example, mentions several such advantages [Wer91]. First, the dag describes a stepwise stochastic process by which the data *could have been* generated and in this sense it may even “prove the basis for developing causal explanations” [Cox92]. Second, each parameter in the dag has a well understood meaning since it is a conditional probability, i.e., it measures the probability of the response variable given a particular configuration of the explanatory (parent) variables and all other variables being unspecified. Third, the task of estimating the parameters in the dag model can be decomposed into a sequence of local estimation analyses, each involving a variable and its parent set in the dag. Fourth, general results are available for reading all implied independencies directly off the dag (Chapter 1) [Ver86], [Pea88b], [LDL90] and for deciding from the topology of two given dags whether they are equivalent,

³The requirement of independence renders the disturbances “local” to each function; disturbances that affect several functions simultaneously will be treated explicitly as “latent” variables.

i.e., whether they specify the same set of independence-restrictions on the joint distribution (Chapter 2) [Fry90], [VP90], and whether one dag specifies more restrictions than the other (Chapter 2) [PGV89].

0.1.2 Assumptions

This formulation invokes several idealizations of the actual task of scientific discovery. The first assumption is that the scientist obtains the distribution directly, rather than from events sampled from the distribution. This assumption is justified when a large sample is available, sufficient to accurately reveal all the dependencies embedded in the distribution.

A more severe assumption is that the observed variables must actually appear in the original or underlying causal theory and are not some aggregate thereof. One major limitation due to this assumption concerns feed-back loops. The effect of this assumption is that the resolution of the samples must be fine enough that the feed-back loop can be “unwound”. For example, let a = “the popularity of a rock band”, and let b = “the record sales of a rock band”. There is an obvious feed-back loop between a and b , each has a positive influence upon the other. However, the value of a today does not really influence today’s value of b , but rather it will influence tomorrow’s, next week’s or maybe next year’s value of b (and vice-versa). The point is that a will not appear as a variable in the causal model, rather, for a couple relevant values of t , a_t will instead.

To overcome this assumption, a detailed analysis of how to correctly incorporate feedback loops into Bayesian networks is required.

0.2 Preview of Technical Results

There are two primary questions that this dissertation addresses:

- “Given that causal model C accurately describes the behavior of physical process P , what set I of conditional independence statements should be revealed from observation of P ? ”
- “Given that set I of conditional independence statements is revealed from observation of process P , what features, if any, do all causal models which accurately predict I have in common? ”

An answer to the first question will give the necessary statistical properties of causal models. An answer to the second question will identify those features of a causal model that can be identified by statistical observation. In particular, any feature that is shared by all causal models consistent with some observation must be a feature of the causal process underlying the data.

Along the path toward an answer to these two questions, several assumptions and definitions are made which lead to interesting questions and results.

The journey begins in Chapter 1 by exploring the theory of Bayesian networks. This is the natural starting point since causal models are a direct outgrowth of this theory. Chapter 2 departs from the general theory of Bayesian networks and examines the specific properties of causal models, answering the first of the original questions. Chapter 3 examines the impact of the latent variables upon the theory of causal modeling. Latent variables are those which are not directly observable. Chapter 4 returns to second original question which can be answered by the theory developed in Chapter 3, if some assumptions are made. The ramifications of these assumptions are examined in detail in this final chapter.

0.2.1 Bayesian Networks

This chapter answers the question: "How can dags be used to decompose a multi-variable function, relation or probability distribution?" In the process of answering that question a theory of dependency models and a special graphical criterion called *d-separation* are developed.

The theory of dependency models centers around the relationship between independence and decomposability. The way that a function can be decomposed is determined by the independences which hold for that function. The d-separation criterion identifies those independence statements that permit a recursive decomposition, so d-separation can be used to correctly infer independencies which hold in any recursively decomposable function.

0.2.2 Causal Models and Theories

This chapter develops the theory of causal models. It starts by identifying the statistical or observational semantics of a causal model. Fortunately, the statistical meaning of a causal model can be completely captured by a set of independence statements. The next question it addresses is that

of equivalence. Two criteria for equivalence are identified, one based upon dependency models and one is graphical. The graphical one is based upon those features which are invariant under equivalence. These invariant feature are used as the basis for an algorithm that can recover a causal model, up to statistical equivalence, from independence information alone.

0.2.3 Latent Structures

This chapter adds the notion of latent variables to the causal model. Latent variables have a severe impact upon the semantics of a causal model. While the probabilistic semantics are easily extended, dependency models are not rich enough to fully capture the semantics of causal models with latent variables. However, they can significantly constrain such causal models. Thus a weaker notion of equivalence, called dependency equivalence, is developed. Since it is weaker, properties which are invariant under dependency equivalence are guaranteed to be invariant under statistical equivalence. The next concept developed is that of a projection, a restricted form of causal model which has at most one latent variable per pair of observable variables. Every causal model, no matter how many variables are latent, has a dependency equivalent projection, which is a rather surprising and important result. Finally several invariant properties are identified.

0.2.4 Inferring Causal Relationships from Observational Data

This chapter addresses how to infer causal relationships from observational data. Definitions for genuine cause and spurious association are given in terms of a minimal model semantics. Then the theory developed in the previous chapters is used to develop criteria for determining genuine causes and spurious associations.

0.3 Basic Concepts

The foundations for the theory of causal modeling can be found in elementary probability and graph theory.

0.3.1 Probability Theory

Let U be a finite set of random variables over which there is a well-defined probability distribution $P(\cdot)$. The range of a random variable $a \in U$ will be denoted by R_a . The random variables in U are primitive in the sense that, in general, no assumptions will be made about their internal structure. They can be continuous or discrete, scalars or vectors.

Since U is finite, subsets of U can be treated as random vectors whose values range over the cross product of the ranges of the individual components. For example, if $A = \{a_1, a_2, \dots, a_k\} \subseteq U$ then $R_A = \times_{i=1}^k R_{a_i}$.

Singleton subsets of random variables are often abbreviated, e.g. $\{a\}$ is written as a . The union of sets of random variables is also abbreviated, e.g. $A \cup B$ is written as AB . Thus, for example, $abAB$ is equivalent to $\{a, b\} \cup A \cup B$.

If $A \subseteq U$ and $V_A \subseteq R_A$ is a measurable subset then $P(A \in V_A)$ denotes the probability that A takes on a value from the set V_A . If $A, B \subseteq U$ and $V_A \subseteq R_A$ and $V_B \subseteq R_B$ are measurable subsets of values then

$$P(A \in V_A | B \in V_B) \stackrel{\text{def}}{=} \frac{P(A \in V_A \text{ and } B \in V_B)}{P(B \in V_B)}$$

and denotes the conditional probability that A takes on a value from the set V_A given that the value of B is in the set V_B . Note that $P(A \in V_A | B \in V_B)$ is undefined whenever $P(B \in V_B) = 0$.

If $A \subseteq U$ then $P(A)$ denotes the probability distribution over A . If $A, B \subseteq U$ then $P(A|B)$ denotes the probability distribution over A conditioned upon B .

Marginalization is a process that removes random variables from a joint probability distribution. Summation is used to marginalize over discrete random scalars, and integration is used to marginalize over continuous random scalars. Marginalization of a vector is accomplished by recursively marginalizing over its components.

Conditional independence, $I(\cdot, \cdot | \cdot)$ is a three place relation over subsets of U . For any $A, B, C \subseteq U$

$$I(A, B | C) \iff P(AB | C) = P(A | C)P(B | C)$$

denotes that A is independent of B given C . An alternate definition is,

$I(A, B|C)$ iff for all $V_A \subseteq R_A, V_B \subseteq R_B$ and $V_C \subseteq R_C$, if $P(C \in V_C) \neq 0$ then

$$P(A \in V_A \text{ and } B \in V_B | C \in V_C) = P(A \in V_A | C \in V_C)P(B \in V_B | C \in V_C)$$

Except as explicitly stated, the three arguments of an I-statement will be assumed to be mutually exclusive. Thus, for example, in the statement $I(A, B|C)$, A , B and C are three mutually exclusive subsets of U .

Let the notation $I[P]$ denote the set of all I-statements which hold according to P .

0.3.2 Graph Theory

Most graph theoretical concepts will be explained as used. Standard notions associated with directed acyclic graphs will be used.

A dag is a pair $\langle U, E \rangle$ where U are the nodes and E are the directed edges or links of the dag. Two nodes are adjacent \bar{ab} iff there is a link between them (i.e. $a \rightarrow b$ or $a \leftarrow b$). A path is a sequence of adjacent nodes. A node is head to head on a path if both of its incident arcs point at it (converging arcs). A node is tail to tail on a path if neither of its incident arcs point at it (diverging arcs). A node is head to tail on a path if exactly one of its incident arcs point at it.

A path $\bar{p} = \langle p_1, p_2, \dots, p_k \rangle$ is directed from p_1 to p_k iff for all $1 \leq i < k$ $p_i \rightarrow p_{i+1}$. A path $\bar{p} = \langle p_1, p_2, \dots, p_k \rangle$ is directed from p_k to p_1 iff for all $1 \leq i < k$ $p_i \rightarrow p_{i+1}$.

A cycle is a path where the first and last nodes are identical. A dag has no directed cycles. A simple path has no cyclic subpaths.

$\overline{p_{i,j}}$ denotes a subpath of \bar{p} . Thus $\bar{p} = \overline{p_{1,|p|}}$. $\bar{p} \oplus \bar{q} = \langle p_1, p_2, \dots, p_{|p|}, q_2, q_3, \dots, q_{|q|} \rangle$ denotes the concatenation of two paths if $p_{|p|} = q_1$.

In some proofs, the over bar is omitted from the path notation.

A path $\bar{p} = \langle p_1, p_2, \dots, p_k \rangle$ ends pointing at p_1 if $p_1 \leftarrow p_2$. A path $\bar{p} = \langle p_1, p_2, \dots, p_k \rangle$ ends pointing at p_k if $p_{k-1} \rightarrow p_k$. If \bar{p} is a directed path from x to a and \bar{q} is a directed path from x to b the $\bar{p} \oplus \bar{q}$ is a divergent path between a and b . That is, a divergent path contains exactly one tail to tail node and no head to head nodes.

If there is a directed link from a to b then a is a parent of b and b is a child of a . If there is a directed path from a to b then a is an ancestor of b and b is a descendant of a . The notation \mathcal{P}_a denotes the set of parents of the

node a . The notation \mathcal{P}_{ab} denotes $\mathcal{P}_a \mathcal{P}_b \setminus ab$. The notation \mathcal{A}_a denotes the set of ancestors of the node a . The notation \mathcal{A}_{ab} denotes $\mathcal{A}_a \mathcal{A}_b \setminus ab$.

Chapter 1

Bayesian Networks

In many domains there exist solutions to problems which require the computation of a measure. A measure is a function that assigns a value to events or possible worlds. Examples of values assigned by commonly used measures include: probability, utility, expected value, likelihood, degree of belief, plausibility, possibility, necessity, abnormality, etc. Within these domains, a variety of reasoning tasks depend upon the efficient storage and manipulation of such a measure. However, when measures are naively represented, the time and storage requirements grow exponentially with the size of the measure domain.

This chapter focuses on a method that recursively decomposes a measure with a large domain into several measures over smaller domains. The key to this decomposition is the notion of conditional independence — a license to safely ignore information. By definition, one set of variables X is independent of a second set Y conditioned upon a third set Z if and only if there exists a loss-less decomposition of the measure along the set Z , i.e.

$$I(X, Y|Z) \stackrel{\text{def}}{\iff} F(XYZ) = F_1(XZ) \oplus F_2(YZ)$$

where F is a measure, F_1 and F_2 are restricted measures and \oplus is a function that combines the values of F_1 and F_2 to compute the value of F .

The first section of this chapter introduces dependency models, a generalization of the notion of probabilistic independence. Examination of several common dependency models leads to the definition of *graphoids*, a broad class of dependency models with a desirable set of properties.

The second section reveals that directed acyclic graphs (dags) offer a useful approximation for graphoids. The dag approximation has two major benefits:

1. The representation is polynomial in both time and space.
2. The independence statements represented by the dag are a subset of those which hold in the original graphoid. Thus each represented statement is guaranteed to be valid, which is very important in light of the fact that independence is a license to ignore information. This guarantee implies that the dag based decomposition will be exact in spite of the approximation used to represent the dependency model.

1.1 Dependency Models

1.1.0.1 Definition. A *Dependency Model* $M = \langle U, I \rangle$ consists of a set of objects U and a set of I-statements I , where each I-statement $(X, Y|Z) \in I$ is an ordered triple of mutually exclusive subsets of U . \square

The triplets in I represent independencies, that is, $(X, Y|Z) \in I$ asserts that X and Y interact only via Z , or, “ X is independent of Y given Z ”. This statement is also written as $I_M(X, Y|Z)$ or simply $I(X, Y|Z)$. Often it is convenient to abuse notation and write $(X, Y|Z) \in M$ rather than $(X, Y|Z) \in I_M$ or $M_1 \subseteq M_2$ rather than $I_{M_1} \subseteq I_{M_2}$.

The class of all dependency models is not some interesting. It simply contains a dependency model corresponding to each subset of the set of ordered triples of subsets of U . However, the various subclasses of dependency models prove to have very interesting properties.

1.1.0.2 Example. The class of *Probabilistic Dependency Models* (PD) contains every dependency model $M = \langle U, I \rangle$ such that U can be put into 1-1 correspondence with a set of random variables U' , there exists a probability distribution $P(\cdot)$ over U' and $I = I[P]$. \square

In other words, a dependency model is probabilistic if there exists a probability distribution that exhibits a pattern of conditional independence statements identical to that of M .

1.1.0.3 Example. The class of *Non-Extreme Probabilistic Dependency Models* (PD^+) contains only those probabilistic dependency models for which there exists a strictly positive probability distribution, $P(\cdot) > 0$ satisfying the conditions of the previous example. \square

1.1.0.4 Example. The class of *Qualitative Dependency Models* (QD) contains every dependency model $M = \langle U, I \rangle$ such that U can be put into 1-1 correspondence with a set of random variables U' , there exists a probability distribution $P(\cdot)$ over U' and for all $X, Y, Z \subseteq U$, $(X, Y|Z) \in I$ iff for all measurable subsets $V_X \subseteq R_X, V_Y \subseteq R_Y$ and $V_Z \subseteq R_Z$, if $P(C \in V_C) \neq 0$ then

$$P(A \in V_A \text{ and } B \in V_B | C \in V_C) > 0 \iff \begin{aligned} P(A \in V_A | C \in V_C) &> 0 \\ \text{and } P(B \in V_B | C \in V_C) &> 0 \end{aligned}$$

where $X', Y', Z' \subseteq U$ such that the elements of X correspond to the elements of X' , the elements of Y correspond to the elements of Y' and the elements of Z correspond to the elements of Z' . \square

In other words, a dependency model is qualitative if there exists a probability distribution which exhibits a corresponding pattern of *qualitative conditional independence statements*, where qualitative independence only checks that each spot in the Cartesian product space is not empty. Qualitative dependencies are frequently used in common-sense reasoning and relational database theory. For example, someone may claim:

There are rich white people, poor white people, rich black people and poor black people. Therefore skin color is *independent* of financial success.

Clearly, while there is no *functional* relationship between skin color and financial success, this analysis completely ignores relative frequencies, as the two parameters are unfortunately *probabilistically dependent*; at least at the present time.

Another class of dependency models which happens to be identical to QD is defined in terms of relational databases.

1.1.0.5 Example. The class of *Embedded Multivalued Dependency Models* (EMVD)[Fag77] is defined in terms of type of dependence that arises in the theory of relational databases. If X, Y and Z are three sets of attributes from some relational scheme, then the notation $Z \rightarrow X|Y$ denotes an embedded multivalued dependence which means that X and Y are independent when the values of Z are held fixed. In particular, if $v_X \in R_X, v'_X \in R_X, v_Y \in R_Y, v'_Y \in R'_Y$ and $v_Z \in R_Z$ then

$$\langle v_X, v_Y, v_Z \rangle \langle v'_X, v'_Y, v_Z \rangle \Rightarrow \langle v_X, v'_Y, v_Z \rangle$$

Where $\langle v_X, v_Y, v_Z \rangle$ means that there exists a tuple in the relation in which $X = v_X, Y = v_Y$ and $Z = v_Z$. In other words, for any given fixed value of Z , the full Cartesian product of the possible values of X and Y must exist.

□

The following class of dependency models is defined in terms of a Hilbert space (which is a Banach space whose norm is defined by an inner product [Ash72]).

1.1.0.6 Example. The class of *correlational graphoids*[PP86] (CG) is defined in terms of the Hilbert norm $\|x - y\|$ on the differences between two elements x and y of a Hilbert space:

$$I(X, Y|Z) \iff \min_{v_Z} \|v_X - v_Z\|^2 = \min_{v'_Z} \|v_X - v'_Z\|^2$$

where $v_X \in R_X, v_Z \in R_Z$ and $v'_Z \in R'_Y R_Z$.

□

It's name derives from the fact that this class is equivalent to the subclass of probabilistic dependency models when the distributions are restricted to be normal (i.e. correlational models). The term graphoid refers to a general but interesting class of dependency models defined in the next subsection.

Graphical separation gives rise to dependency models as well.

1.1.0.7 Example. The class of *Undirected Graph Dependency Models* (UGD) contains every dependency model $M = \langle U, I \rangle$ such that there exists an undirected graph over U , $G = \langle U, E \rangle$ and for all $X, Y, Z \subseteq U$, $(X, Y|Z) \in I$ iff Z is a cut-set separating X and Y .

This class of dependency models is based upon a graphical notion of separation. An alternate way to state the cutset criterion is that, X is separated from Y by Z if every path from a node in X to a node in Y contains a node in Z .

1.1.0.8 Example. The class of *Directed Acyclic Graph Dependency Models* (DAGD) contains every dependency model $M = \langle U, I \rangle$ such that there exists an directed graph over U , $G = \langle U, E \rangle$ and for all $X, Y, Z \subseteq U$, $(X, Y|Z) \in I$ iff there is no directed path between a node in X and a node in Y along which every node with converging arrows either is or has a descendent in Z and every other node is outside Z . \square

This criterion is called *d-separation* in [Pearl, 1986] and is explained in more detail in the next section.

1.1.1 Graphoids

Each class of dependency models has several properties. Some of those properties are common to all of the classes presented thus far. For example, every notion of independence presented here is symmetric: If X is probabilistically independent of Y conditioned upon Z then Y is probabilistically independent of X conditioned upon Z . This is a trivial consequence of the definition. The following four common properties define the class of graphoids.

1.1.1.1 Definition. A *graphoid* is any dependency model M which is closed under the following properties:

$$\text{Symmetry: } I(X, Y|Z) \Leftrightarrow I(Y, X|Z)$$

$$\text{Decomposition: } I(X, YW|Z) \Rightarrow I(X, Y|Z)$$

$$\text{Weak Union: } I(X, YW|Z) \Rightarrow I(X, Y|ZW)$$

$$\text{Contraction: } I(X, W|ZY) \& I(X, Y|Z) \Rightarrow I(X, YW|Z)$$

\square

It is straight forward to show that all the specialized classes of dependency models presented thus far are graphoids[PP86, PP80], and in view of this generality, these four properties are selected to represent the general notion of mediated dependence between items of information.

With the exception of UGD, none of the specialized dependency classes possesses complete parsimonious axiomatization similar to that of graphoids.

EMVD is known to be non-axiomatizable by a finite set of Horn clauses [Parker, 1980]. A similar but more restricted result has been reported for DAGD. [Geiger, 1987] has shown that there is no finite set of universally quantified horn axioms characterizing DAGD. However, [Verma, unpublished] has found a finite characterization of DAGD using existentially quantified axioms.

1.1.1.2 Definition. A *positive-graphoid* is any graphoid M which is also closed under the following property:

$$\text{Intersection: } I(X, W|ZY) \& I(X, Y|ZW) \Rightarrow I(X, YW|Z) \quad \square$$

It is straight forward to show that classes PD⁺, CG, UGD, and DAGD are all positive graphoids. Only the classes QD, EMVD and PD do not comply with this axiom.

1.1.1.3 Definition. Let $M = \langle U, I \rangle$ be a dependency model from some class \mathcal{M} of dependency models. A subset of triplets $B \subseteq I$ of triplets is a positive \mathcal{M} -basis of M iff for every model $M' = \langle U, I' \rangle \in \mathcal{M}$ if $B \subseteq I'$ then contains $I \subseteq I'$. \square

A basis provides a complete encoding of the information contained in M ; knowing B and \mathcal{M} it is possible, in principle, to decide what triplets belong to M .

1.1.1.4 Definition. An *I-map* of a dependency model M is any model M' such that $M' \subset M$. \square

For example, the undirected graph $WX - Z - Y$ is an *I-map* of the complete graph over the four variables.

1.1.1.5 Definition. A *D-map* of a dependency model M is any model M' such that $M' \supset M$. \square

For example, if a relation R contains all tuples having non-zero probability in P then M_R is a *D-map* of M_P .

1.1.1.6 Definition. A *Perfect-map* of a dependency model M is any model M' such that $M' = M$. For example, the undirected graph $X - Z - Y$ is a perfect map of the DAG $X \rightarrow Z \rightarrow Y$. \square

We will be primarily interested in mapping data dependencies into graphical structures, where the task of testing connectedness is easier than that of testing membership in the original model M . A D -map guarantees that vertices found to be connected are, indeed, dependent; however, it may occasionally display dependent variables as separated vertices. An I -map works that opposite way: it guarantees that vertices found to be separated always correspond to genuinely independent variables but does not guarantee that all those shown to be connected are, in fact, dependent. Empty graphs are trivial D -maps, while complete graphs are trivial I -maps.

1.2 Directed Acyclic Graphs (dags)

This section presents a recursive decomposition of graphoids then proves its correctness.

1.2.1 Recursive Decomposition

Suppose that a measure is decomposed via a dag D by the following formula:

$$F(U) = \bigoplus_i F_i(x_i, \mathcal{P}_{x_i})$$

where \mathcal{P}_x denotes the set of parents of x in D , and the x_i 's are in some topological ordering of D . An important question to answer is: “What constraints does this decomposition place upon F ?”. A related question is: “Can a given measure F be decomposed via D ?“

The decomposition is equivalent to the following set of independence assertions called a *recursive basis* of D :

$$\{I(x_i, U_i \setminus \mathcal{P}_{x_i} | \mathcal{P}_{x_i})\}$$

where U_i contains all nodes which precede x_i in the topological ordering. The proof of this is purely algebraic and depends upon two assumptions about

the structure of the measure. First there must be the combination function \oplus . This function is assumed to form an Abelian group over the values of the measure space. Second there must be a marginalization operation which is analogous to the summation and integration operations from probability theory and projection operation from relational database theory. The \oplus operation must distribute over the marginalization operation. For a more detailed analysis of the generalizations of these operations, see [SS86].

Since we are primarily concerned with probability measures, this proof will be specific to them, but note that the proof will generalize.

1.2.1.1 Theorem. [KSC84] *A probability distribution P decomposes according to:*

$$P(U) = \prod_i P(x_i | \mathcal{P}_{x_i})$$

if and only if:

$$\{I(x_i, U_i \setminus \mathcal{P}_{x_i} | \mathcal{P}_{x_i})\} \subseteq I[P]$$

Proof: Simply observe (1) that the marginalization of the decomposition equation over $U \setminus x_i U_i$ will give an equation that is equivalent to the definition of $I(x_i, U_i \setminus \mathcal{P}_{x_i} | \mathcal{P}_{x_i})$ and (2) that the product of the definitions of all the I-statements in the set yields the above equation.

1. For simplicity, we will use \sum_X to denote the operations that are needed to marginalize over all variables in X . This operation may encompass several summations and integrations. For any i , consider that:

$$\sum_{U \setminus x_i U_i} P(U) = \sum_{U \setminus x_i U_i} \prod_j P(x_j | \mathcal{P}_{x_j})$$

$$P(x_i U_i) = \prod_{j <= i} P(x_j | \mathcal{P}_{x_j})$$

$$P(x_i U_i) = P(x_i | \mathcal{P}_{x_i}) \prod_{j < i} P(x_j | \mathcal{P}_{x_j})$$

$$P(x_i U_i) = P(x_i | \mathcal{P}_{x_i}) P(U_i)$$

$$P(x_i U_i \setminus \mathcal{P}_{x_i} | \mathcal{P}_{x_i}) = P(x_i | \mathcal{P}_{x_i}) P(U_i \setminus \mathcal{P}_{x_i} | \mathcal{P}_{x_i})$$

Which is the definition that:

$$I(x_i, U_i \setminus \mathcal{P}_{x_i} | \mathcal{P}_{x_i})$$

2. Consider that:

$$I(x_i, U_i \setminus \mathcal{P}_{x_i} | \mathcal{P}_{x_i})$$

is defined as:

$$P(x_i, U_i \setminus \mathcal{P}_{x_i} | \mathcal{P}_{x_i}) = P(x_i | \mathcal{P}_{x_i}) P(U_i \setminus \mathcal{P}_{x_i} | \mathcal{P}_{x_i})$$

which is equivalent to:

$$P(x_i, U_i) = P(x_i | \mathcal{P}_{x_i}) P(U_i)$$

Take the product over all of these equations:

$$\prod_i P(x_i, U_i) = \prod_i P(x_i | \mathcal{P}_{x_i}) P(U_i)$$

The left-hand side of the i^{th} equation cancels with the second term of the right-hand side of the $i + 1^{\text{th}}$ for all $1 \leq i < |U|$. Thus it simplifies to:

$$P(U) = \prod_i P(x_i | \mathcal{P}_{x_i})$$

□

Thus the joint probability distribution over a set of variables is defined as the product of conditionals of each node given its parents.

The separation criterion for dags, called d-separation, was defined with the intention of capturing those independencies which follow from the recursive decomposition of a probability distribution. In fact, it soundly identifies independencies which follow from the decomposition of any measure whose corresponding class dependency models are graphoids as the following theorem will prove. But first lets review the definition in more detail.

1.2.1.2 Definition. A set of nodes X is **d-separated** from a second set Y given a third set Z iff every path¹ between any node in X and any node in Y is not activated by Z .

A path ρ is **activated** by a set Z if every head-to-head² node on ρ is activated by Z , and no other node of ρ is contained in Z .

A node is **activated** by a set if either it is in the set or it has a descendent in the set. □

¹A path is an ordered list of adjacent nodes; their directionality is irrelevant.

²A node b is head-to-head on ρ iff $a \rightarrow b \leftarrow c$ is a subpath of ρ

1.2.1.3 Theorem. *The d-separation criterion is sound for all graphoids.*

Proof: Let $D = \langle U, E \rangle$ be a dag and let X, Y and Z be three subsets of U . Soundness means that if X is d-separated from Y given Z in D then X is independent of Y given Z in every graphoid consistent with D , i.e. that D is an I-map of the graphoid closure of any basis of D .

Let M_D be the set of all d-separation statements which hold in D , let B be any basis of D , and let M_B be the graphoid closure of B . It remains to show that $M_D \subseteq M_B$.

The proof will be by induction on $|U|$, if $|U| = 1$ then $M_D = M_B$ because they contain only trivial statements.

If $|U| = k$ then let n be the last variable in the ordering imposed by B , let $D \setminus n$ be the dag formed by removing n and its incident links from D and let $B \setminus n$ be the set formed by removing the last triplet, B_k , from B . Since n is the last variable in the ordering of B , $B \setminus n$ is the basis of $D \setminus n$.

Now let $M_{D \setminus n}$ be the set of all d-separation statements which hold in $D \setminus n$, and $M_{B \setminus n}$ be the graphoid closure of $B \setminus n$. Since $D \setminus n$ has $k - 1$ variables, $M_{D \setminus n} \subseteq M_{B \setminus n}$ by the inductive hypothesis.

Each triplet T of M_D falls into one of four categories; either the variable n does not appear in T or it appears in the first, second or third entry of T . These will be treated separately as cases 1, 2, 3 and 4, respectively.

case-1:

If n does not appear in T then T must equal $(X, Y | Z)$, where X, Y and Z are three disjoint subsets of variables, none of which contain n . Suppose (r.a.a.) that $T \notin M_{D \setminus n}$, i.e. that $D \setminus n$ contains a Z -active path between some node in X and some node in Y . This path would also be active in D since the addition of nodes and links can not deactivate a path, but that would contradict the hypothesis that X and Y are d-separated given Z in D . Thus $T \in M_{D \setminus n} \subset M_{B \setminus n} \subset M_B$.

case-2:

If n appears in the first entry of the triplet, then $T = (Xn, Y | Z)$ with the same constraints on X, Y and Z as in case-1. Consider $B_k = (n, R | \Pi)$, the last triplet in B . Let Π_X, Π_Y, Π_Z and Π_0 be a partitioning of Π and R_X, R_Y ,

R_Z and R_0 be a partitioning of R such that $X = \Pi_X \cup R_X$, $Y = \Pi_Y \cup R_Y$ and $Z = \Pi_Z \cup R_Z$ as shown in Figure 1.

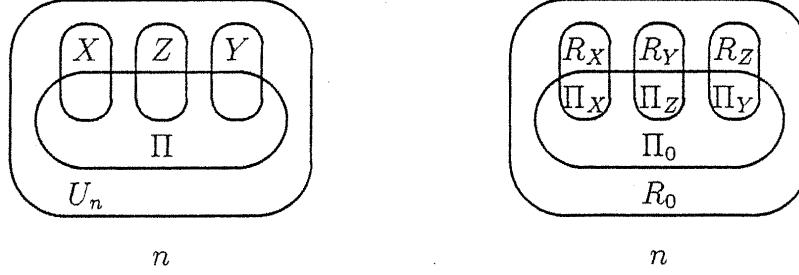


Figure 1.1: Partitioning of the sets

The set Π_Y must be empty because every node in Π is a parent of n (by the definition of a basis) and if Π_Y were not empty then there would be a parent of n in $Y \supset \Pi_Y$, which would contradict the hypothesis that $(Xn, Y|Z) \in M_D$. Since $\Pi_Y = \emptyset$ it follows that $B_k = (n, R_X R_Z Y R_0 | \Pi_X \Pi_0 \Pi_Z)$. And because $X = \Pi_X \cup R_X$, $Y = R_Y$ and M is a graphoid, decomposition and weak union imply that $(n, Y|X\Pi_0 Z) \in M$.

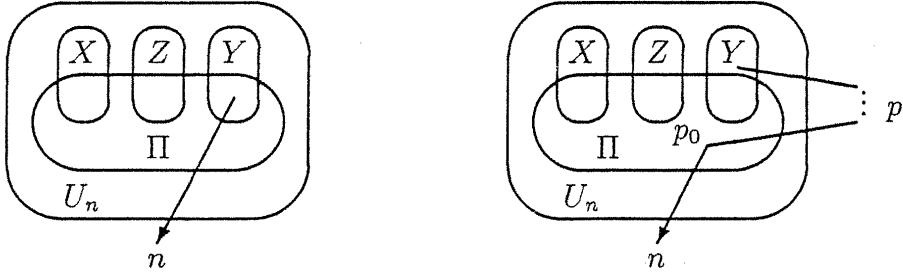


Figure 1.2: Partitioning of the sets

The set Π_0 must be d-separated from Y given Z in D . If the sets were not separated, then there would be a Z -active path, p , between a node of Π_0 and Y . Since $p_0 \in \Pi_0$ it would be a parent of n (see Figure 2b), thus the path $n \oplus p$ would be Z -active and would contradict the hypothesis that $T = (Xn, Y|Z) \in M_D$.

Since Π_0 and X are each separated from Y given Z in the dag D , their union must also be separated, i.e. $(X\Pi_0, Y|Z) \in M_D$. Since n is not in this triplet, the argument of case-1 above implies that $(X\Pi_0, Y|Z) \in M_B$. Using symmetry and contraction on this statement and $(n, Y|X\Pi_0Z) \in M_B$ it follows that $T = (Xn, Y|Z) \in M_B$.

case-3:

If n appears in the second entry, then by symmetry the triplet T is equivalent to one with n in the first entry, and the argument of case-2 above shows that $T \in M$.

case-4:

If n appears in the third entry then $T = (X, Y|Zn)$. Now it must be the case that X and Y are separated in G given only Z . Suppose (r.a.a) that they were not separated; there would be a Z -active path between some node in X and some node in Y . Since X and Y are separated given Zn , instantiating n would have to deactivate the path. But n is a sink and cannot serve to deactivate any path by being instantiated. Thus there is no such path and $(X, Y|Z) \in M_D$. This statement along with $(X, Y|Zn)$ imply that either $(Xn, Y|Z)$ or $(X, Yn|Z)$ is in M_D by the weak transitivity property of DAGs. Either way case-2 or case-3 would imply that the corresponding triplet must be in M_B , and either of these would imply that $T \in M_B$ by weak union. \square

1.3 Historical and Bibliographic Remarks

Chapter 2

Causal Models and Theories

Causal models represent the underlying causal mechanisms that govern the behavior of a set of variables. The utility of causal models lies in their ability to clearly state hypotheses about such causal mechanisms and thus invite standard scientific methods to test these hypotheses. Knowledge about the actual causal mechanisms may come from a variety of sources, none of which are direct. When systems are built from well understood parts, a causal understanding stems from the physics behind the interacting mechanisms. For example, pulling a car's gearshift lever causes the transmission to change gears or pushing on the throttle causes an increase in acceleration. Alternatively, if a physical knowledge of the phenomena is lacking, causal relations can be learned, either via the standard method of manipulative experimentation or via the non-manipulative methods described in this dissertation.

2.1 Formal Definitions

2.1.0.4 Definition. A causal model over a set of variables U is a dag $D = \langle U, E \rangle$ where the variables in U are the nodes of the dag D . \square

2.1.0.5 Example. The causal model in Figure (2.1a) represents the following sequence of events: a rolling soccer ball (b_0) is kicked (k) then it hits and possibly breaks a window (w). The node b_1 denotes the ball immediately after being kicked; its value depends based upon the ball's initial configu-

ration and the strength of the kick. This value directly affects the window which absorbs some energy and may or may not break. The final state of the window, in turn, affects the soccer ball which either bounces back, falls down or passes through the window if the window had broken. \square

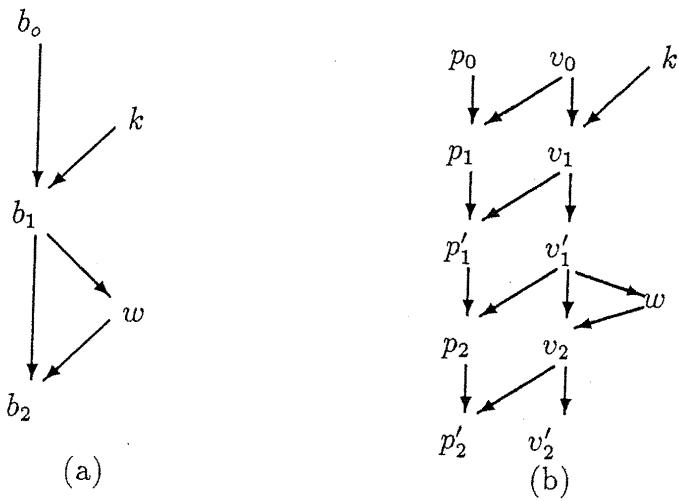


Figure 2.1: kicking a soccer ball

2.1.0.6 Example. The causal model in Figure (2.1b) offers a more detailed representation of the sequence of events of the previous example by explicating some of the internal variables associated with the ball. For example, the strength of the kick (k) does not affect the position of the ball at the instant that the ball is kicked (p_1), rather it affects the velocity of the ball (v_1) which, in turn, will affect the position of the ball at some later point in time (p'_1). The window is affected by both the velocity and position of the ball but only the velocity of the ball is directly affected by the window. \square

Quantifying the relationships of a causal model results in a causal theory – a precise specification of how each variable is influenced by its parents in

the dag. Ideally, the value of a variable should be a deterministic function of its causes. However, this would only be true if every possible causal influence were modeled, which is never the case. Thus the functional relationship between a set of causes and their effect includes a noise factor or arbitrary stochastic disturbance. These disturbances are recursively independent; a requirement which renders each disturbance to be local to one variable; disturbances which directly influence several variables will be treated explicitly as latent variables in Chapter 3.

2.1.0.7 Definition. A **causal theory** is a pair $T = \langle D, \Theta_D \rangle$ containing a causal model D and a set of parameters Θ_D compatible with D . Θ_D assigns a function $x_i = f_i[\text{pa}(x_i), \xi_i]$ and a probability measure g_i , to each $x_i \in U$, where $\text{pa}(x_i)$ are the parents of x_i in D and each g_i randomly distributes a disturbance vector ξ_i independently of every other ξ_j and of every $x_j \in U$ s.t. $0 < j < i$. \square

2.1.1 Probabilistic Semantics

Every causal theory T defines a unique joint probability distribution over its variables, $P[T]$. This distribution can be estimated by drawing samples and is the only experimentally meaningful facet of a causal theory. Since the observational meaning of a causal model is given by the statistical constraints that the model places upon its variables, the meaning of a particular causal model is given by the set of joint probability distributions consistent with that model.

2.1.1.1 Definition. A causal model D is **consistent** with a distribution P if D can accommodate some theory that generates P , i.e. there exists a Θ_D s.t. $P[\langle D, \Theta_D \rangle] = P$. \square

2.1.1.2 Example. Consider a causal model D which is a complete dag-over n variables. Since D is complete, it imposes a total ordering over the variables such that x_1 is the only root node, and the set of parents for any other node $\text{pa}(x_i)$ is equal to the set of nodes that precede it in the total ordering (i.e $\{x_1, \dots, x_{i-1}\}$). Now the chain rule for conditional probabilities states:

$$P(x_1, \dots, x_n) = P(x_1) \times P(x_2|x_1) \times P(x_3|x_1, x_2) \times \dots \times P(x_n|x_1 \dots x_{n-1})$$

This rule can be used to construct a set of parameters Θ_D for D such that the causal theory $\langle D, \Theta_D \rangle$ will generate any probability distribution P . Hence a complete dag is consistent with any distribution. \square

Proving consistency is easy since the proof is by example. However, at this point to prove inconsistency would entail non-existence proof which is considerably more difficult. In the next section the concept of conditional independence will be used to establish necessary and sufficient conditions for consistency which will significantly simplify this task.

2.1.1.3 Theorem. *A causal model D is consistent with a probability distribution P iff P can be decomposed as:*

$$P(U) = \prod_i P(x_i|\mathcal{P}_{x_i})$$

Proof: If D is consistent with P , then there exist a set of parameters such that can be used to generate P . The parameters include a deterministic function $f_i[\text{pa}(x_i), \xi_i]$ and a probability measure g_i for each x_i . Together these define a unique conditional probability distribution $P(x_i|\mathcal{P}_{x_i})$.

If $P(U)$ can be decomposed according to the above equation, then each conditional $P(x_i|\mathcal{P}_{x_i})$ in the decomposition can be used to define an appropriate function $f_i[\text{pa}(x_i), \xi_i]$ and probability measure g_i such that the set of all these functions and measures, Θ can be used to generate P . \square

2.1.1.4 Corollary. *A causal model D is consistent with a probability distribution P iff every independence statement in any recursive basis of D holds in P .*

Proof: This follows directly from Theorems 1.2.1.1 and 2.1.1.3. \square

2.1.1.5 Definition. One causal model D is preferred to another D' written, $D \preceq D'$ iff D' can mimic D , i.e. for every set of parameters Θ_D there exists a corresponding set $\Theta'_{D'}$ s.t. $P(\langle D', \Theta'_{D'} \rangle) = P(\langle D, \Theta_D \rangle)$. \square

A simple criterion for preference is given by Corollary 2.1.1.4. If every statement in the recursive basis of D holds as a d-separation statement in D' then $D \preceq D'$.

2.1.1.6 Example. Since a complete graph is consistent with any distribution, it can mimic any other model and hence any causal model is preferred to a complete graph. \square

2.1.1.7 Definition. Two causal models D_1 and D_2 are equivalent if for every theory $T_1 = \langle D_1, \Theta_1 \rangle$ there is a theory $T_2 = \langle D_2, \Theta_2 \rangle$ such that T_1 and T_2 define the same probability distribution, and vice versa. \square

A simple criterion for equivalence is given by Corollary 2.1.1.4. If every statement in the recursive basis of D holds as a d-separation statement in D' , and vice-versa, then $D \equiv D'$.

2.1.1.8 Example. Consider the four causal models in Figure 2.2. Their semantics are given by the following constraints on the joint distribution $P(abc)$, namely:

$$P(abc) = P(c|b)P(b|a)P(a)$$

$$P(abc) = P(a|b)P(c|b)P(b)$$

$$P(abc) = P(a|b)P(b|c)P(c)$$

$$P(abc) = P(b|ac)P(a)P(c)$$

Using the definition of conditional probabilities, it is straight forward to show that the first three equations imply each other, but not the forth. Therefore, the first three causal models all have the same semantics and are equivalent. \square

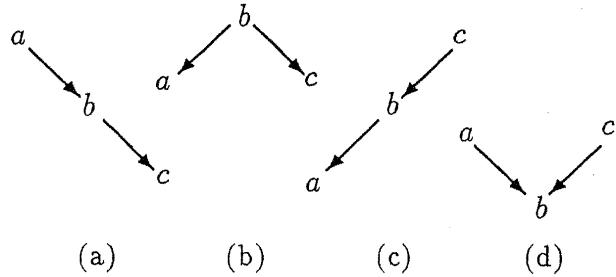


Figure 2.2: Three of the four models are equivalent

2.2 Equivalence

The following theorem demonstrates that adjacency and vee structures are properties of causal models which are invariant under equivalence. Furthermore, it demonstrates that these properties completely define a causal model up to statistical equivalence. Formally, two nodes are adjacent, written \overline{ab} if and only if there is a link between them. Three nodes from a vee structure, written \overline{abc} , iff $a \rightarrow b \leftarrow c$ and a is not adjacent to c .

2.2.0.9 Theorem. *Let M be any dag isomorphic dependency model, a dag D is consistent with M iff the following two conditions hold:*

1. \overline{ab} in D iff $\forall_S, I(a, b|S) \notin M$.
2. \overline{abc} in D iff \overline{abc} and $\neg\overline{ac}$ in D and \forall_S , if $I(a, c|S) \in M$ then $b \notin S$.

Proof: There are three basic parts to the proof, (1) that the first condition is necessary for consistency, (2) that the second condition is necessary, and (3) that both conditions together are sufficient.

Part 1: *If D is consistent with M then Condition 1 holds.*

Since D is consistent with M , independence in M is identical to that in D , so it is enough to show that two nodes are adjacent in D iff there is no way to d-separate them.

A link between two adjacent nodes is a path which cannot be deactivated, thus if \overline{ab} then there could not be any set S s.t. $I(a, b|S) \in M$.

It remains to show that if there is no set S s.t. $I(a, b|S) \in M$ then then a and b are adjacent. It suffices to consider $S = \{x \neq a, b : x \text{ is an ancestor of } a \text{ or } b\}$. Since, by assumption, a and b are not d-separated by any set, it must be the case that $I(a, b|S) \notin M$ thus there must be a path ρ connecting a and b in D which is active given S . Since ρ is S -active, every head to head node on ρ must be in or have a descendent in S . But by the definition of S , every node which has a descendant in S must be in S as well. Thus every head-to-head node on ρ must be in S . Every other node on ρ is an ancestor of a, b or one of the head to head nodes of the path. Hence every node on ρ must be in S with the exception of a and b . Thus every node of ρ , except a and b , must be a head-to-head node. There are only three paths satisfying this condition: $a \rightarrow b$, $a \leftarrow b$ and $a \rightarrow c \leftarrow b$. However the last case is not possible because c is in S so it must be an ancestor of either a or b and thus it cannot be common child of both a and b as well or there would be a directed cycle. Hence a and b are adjacent.

Part 2 If D is consistent with M then Condition 2 holds.

If b is head-to-head in between a and c then the two link path cannot be de-activated by any set containing b . The rest of the only-if portion of condition 2 follows trivially from the definition of a vee structure.

To complete the proof of Part 2, let \overline{abc} be a chain with $\neg\overline{ac}$. Furthermore, assume that for any set S , $I(a, b|S) \in M$ implies $b \notin S$. If b were not head-to-head on the path \overline{abc} then any set S for which $I(a, c|S) \in M$ would necessarily contain b in order to deactivate this path. Since $\neg\overline{ac}$, there must be a such an S , however by assumption for any such S , $b \notin S$. Thus b must be head-to-head on the path \overline{abc} , hence it must be the case that \overline{abc} .

Part 3 If Conditions 1 and 2 hold then D is consistent with M .

If M is dag isomorphic then there must exist a dag which is consistent with M , call it D^* . By Parts 1 and 2 above, D and D^* have the same skeletons and vee structures, so it is enough to prove Proposition 2.2.0.10:

2.2.0.10 Proposition. *If any two dags, D and E , have the same skeletons and vee structures then every active path in one dag corresponds to an active path in the other.*

Let ρ be an S -active path in D which is minimal in the following sense: if k is the number of nodes in ρ , ρ_1 is the first node and ρ_k is the last node then (1) there cannot exist an S -active path between ρ_1 and ρ_k with strictly fewer than k nodes and (2) there cannot exist a different S -active path ϕ between ρ_1 and ρ_k with exactly k nodes such that for all $1 < i < k$, either $\phi_i = \rho_i$ or ϕ_1 is a descendant of ρ_i .

Since D and E have the same links ρ must be a path in E . It can be shown by induction on the number of head-to-head nodes that ρ is S -active in E as well. By definition, a single nodes will be considered as an active path. The remainder of the proof has three sub-parts: the first part proves that if ρ contains no head-to-head nodes then it is S -active in E , the second part proves that if ρ contains at least one head-to-head node $x = \rho_i$ then ρ is S -active in E iff x is S -active in E , and the third part proves that x is S -active in E .

Sub-Part 1:

If ρ does not contain any head-to-head nodes in D then it would be S -active in E unless it contains a head-to-head node in E . It is enough to show that ρ cannot have any head-to-head nodes in E . Suppose that some node $x = \rho_i$ were head-to-head in E with parents $y = \rho_{i-1}$ and $z = \rho_{i+1}$, Figure 2.3 shows the possible configurations for D .

$$\begin{array}{ccc} D & & E \\ \rho_1 \cdots y \rightarrow x \rightarrow z \cdots \rho_k & & \rho_1 \cdots y \rightarrow x \leftarrow z \cdots \rho_k \\ \rho_1 \cdots y \leftarrow x \rightarrow z \cdots \rho_k & & \\ \rho_1 \cdots y \leftarrow x \leftarrow z \cdots \rho_k & & \end{array}$$

Figure 2.3: D has no head-to-head nodes, but E does.

The parents of ρ_i along ρ in E would be adjacent in both D and E since the two graphs share links and vee structures. But the sequence of nodes formed by removing ρ_i from ρ would be a path in D since its parents would be adjacent. Moreover this path would be S -active since it could contain

no head-to-head nodes (unless D contained a directed loop). But this path would contradict Condition 1 of the minimality of ρ in D . Therefore if ρ contains no head-to-head nodes in D then it is S -active in E .

Sub-Part 2:

Suppose that ρ contains at least one head-to-head node $x = \rho_i$ in D with parents $y = \rho_{i-1}$ and $z = \rho_{i+1}$ as shown in Figure 2.4. Let $\rho_{1,i-1}$ be the

$$\begin{array}{c} D \\ \rho_1 \cdots y \rightarrow x \leftarrow z \cdots \rho_k \end{array}$$

Figure 2.4: D has some head-to-head nodes.

subpath of ρ between a and y and $\rho_{i+1,k}$ be the subpath between z and b . Note that $i - 1$ may equal 1 and/or $i + 1$ may equal k , in which case the corresponding subpath(s) would be a single node. Both ρ_1 and ρ_2 are minimal S -active paths of D and both contain strictly fewer head-to-head nodes than ρ thus by the inductive hypothesis, they are S -active in E . If y and z were adjacent in D then since both nodes are both S -active in D (they are parents of an S -active node) and neither is in S (because neither is head-to-head on ρ in D), it follows that the path formed by removing x from ρ would be S -active. This path which would contradict Condition 1 of the minimality of ρ .

Therefore y and z cannot be adjacent in either graph and must be common parents of x in both. Since x is head-to-head on ρ in E and both the subpaths $\rho_{1,i-1}$ and $\rho_{i+1,k}$ are S -active in E it follows that ρ would be S -active in E iff x were S -active in E .

Sub-Part 3:

Since x is S -active in D there exists a directed path in D from x to some node w in S . Let ϕ be the shortest such path. It remains to show (by induction on the length l of ϕ) that ϕ is strictly directed from x to w in E . There are three cases, either $l = 0$, $l = 1$ or $l > 1$.

If $l = 0$ then $x = w$ and x is trivially S -active in E .

If $l = 1$ then ϕ is a single link. Consider the parents, y and z of x . If they were both adjacent to w as in Figure 2.5 then they would be common parents of w in D (or there would be a directed loop in D). Thus the sequence of nodes ρ' formed by replacing x with w in ρ would be an S -active path in D .

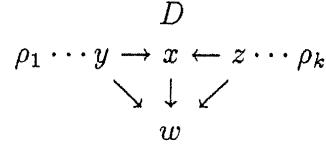


Figure 2.5: A single link descendant path.

This path would contradict Condition 2 of the minimality of ρ , so at least one parent of x must not be adjacent to w . Without loss of generality, assume y is not adjacent to w . Since y and w are not parents of x in D , they cannot both be parents of x in E as the two graphs share vee structures. Therefore x must be a parent of w in E and x would be S -active in E .

If $l > 1$ then ϕ contains at least two links. Consider the last two links of ϕ , shown in Figure 2.6 where $u = \phi_{l-2}$, $v = \phi_{l-1}$. Note that $l-2$ may equal 1 in which case $x = u$. The initial subpath $\phi_{1,l-1}$ must be directed from

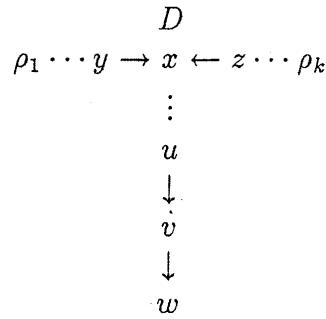


Figure 2.6: A multiple link descendant path.

x to v by induction. If u were adjacent to w then there would have been a shorter directed path from x to w in D , thus u and w are not adjacent and not parents of v in D so they cannot both be parents of v in E . Therefore v must be a parent of w in E and ϕ is S -active in E . \square

2.2.0.11 Corollary. *Two dags are equivalent iff they share the same set of links and same set of vee structures.*

Proof: This result follows directly from the proof of the previous lemma. \square

2.2.0.12 Lemma. *For any dag-isomorphic dependency model M and any chain \overline{abc} , if \exists_S s.t. $I(a, c|S) \in M$ and $b \notin S$ then $\forall_{S'} I(a, c|S') \in M$ implies $b \notin S'$.*

Proof: Suppose \overline{abc} and \exists_S s.t. $I(a, c|S) \in M$ and $b \notin S$. In order for S to d-separate a and c , it must be the case that $a \rightarrow b \leftarrow c$ — if b were not head-to-head then this two link path would be active given any set not containing b . Now since b is head-to-head it must be the case that any set S which contains b will activate this two link path, hence for any S if $I(a, b|S) \in M$ then $b \notin S$. \square

2.3 Recovery

Given a dependency model $M = \langle U, I \rangle$ as an explicit list of independence statements it is required to decide whether there exists a directed acyclic graph D that is consistent with M , and if so to generate one.

It is initially assumed that M is a graphoid, i.e. that the list of independence statements is closed under the graphoid axioms. The assumptions about M are varied in subsections 2.3.4 and 2.3.5.

- Phase 1 examines the independence statements in M and tries to construct a pdag, G with the following guarantees:
 1. If M is dag-isomorphic then every extension of G will be consistent with M .
 2. If Phase 1 fails to generate a pdag, then M is not dag-isomorphic.
- Phase 2 extends a pdag, G , into a dag D , if possible.
- Phase 3 verifies if D is consistent with M .

If D is found to be consistent with M then M is dag-isomorphic, by definition. If D is found to be inconsistent with M then M is not dag-isomorphic and (by definition) no dag can be consistent with M .

Additional improvements to this algorithm and extensions to the problem are discussed in Section 2.3.4.

2.3.1 The DAG Construction Algorithm

Phase 1

Generate a pdag G , from M , if possible.

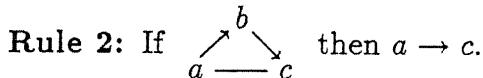
1. For each pair of variables, (a, b) , look through M for a statement of the form $I(a, b|S)$, where S is any set of variables (including \emptyset). Construct an undirected graph G where vertices a and b are connected by an edge iff a statement $I(a, b|S)$ is not found in M . Mark every pair of non-adjacent nodes in G with the first such set S found in M , call this set $S(a, b)$.
2. For every pair of non-adjacent nodes a and c in G , test if there is a node b not in $S(a, c)$ that is adjacent to both a and c . If there is such a node then direct the arcs $a \rightarrow b$ and $c \rightarrow b$ unless there already exists a directed path from b to a or from b to c , in which case Phase 1 FAILS.
3. If the orientation of Step 2 is completed then Phase 1 SUCCEEDS, and returns a partially directed graph, G .

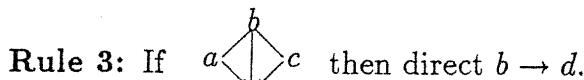
Phase 2

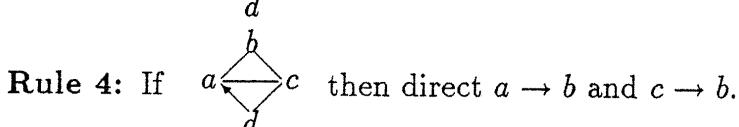
Extend G into a dag, D , if possible.

1. Initially let C be an empty stack and let D equal G .
2. While D contains any undirected arcs repeat 2a, 2b and 2c:
 - (a) Close D under the following four rules, if possible.

Rule 1: If $a \rightarrow b - c$ and a is not adjacent to c then direct $b \rightarrow c$.

Rule 2: If  then $a \rightarrow c$.

Rule 3: If  then direct $b \rightarrow d$.

Rule 4: If  then direct $a \rightarrow b$ and $c \rightarrow b$.

- (b) If the closure was successful, i.e. there are no directed cycles or new vee structures, then:
 - If D still contains any undirected arcs, select one and choose a direction for it, push the arc and a copy of D onto the stack C and continue the while loop (i.e. go back to 2a).
 - If G contains no more undirected arcs, then the while loop is completed, Phase 2 SUCCEEDS, and returns a directed acyclic graph D .
- (c) If the closure was unsuccessful and the stack is not empty, then discard the current value of D and pop the most recent copy off of the stack along with the selected arc. Reverse the chosen direction of the arc in D and continue the while loop (i.e. go back to 2a).
 If the closure was unsuccessful and the stack is empty then Phase 2 FAILS.

Phase 3

Check if D is consistent with M .

1. Test that every statement I in M holds in D (using the d-separation criterion)¹.
2. Pick any total ordering of the nodes which agrees with the directional-ity of the D and let U_a stand for the set of nodes which precede a in this ordering. For every node a in D , test if the statement $I(a, U_a \setminus \mathcal{P}_a | \mathcal{P}_a)$ is in M .
3. If both tests are confirmed, EXIT with SUCCESS, and return D ; else, EXIT with FAIL.

2.3.2 Correctness

Phase 1

This phase examines M and generates a graph, G subject to the above guarantees, if possible. That is, if M is dag-isomorphic then every extension of

¹A linear time algorithm for testing d-separation is reported in [GVP90].

G is consistent with M . The correctness of Step 1 of this phase follows from Lemma 2.2.0.9. This lemma is also the basis for the inference algorithm developed by Spirtes and Glymour [SG91].

The *only-if* portion of this lemma guarantees that:

1. If there exists some dag D^* which is consistent with M , then any dag D consistent with M must have the same skeleton as D^* .
2. Furthermore, every dag D , consistent with M must have the same vee structures as D^* .

The *if* part guarantees that every dag D which has the same skeleton and vee structures as D^* , is consistent with M . The first step of Phase 1 attempts to construct this invariant skeleton if M is dag-isomorphic. The arrowheads added in the second step identify the invariant vee structures, again, if M is dag-isomorphic.

Note however, that Step 2 of Phase 1 directs arcs immediately upon finding one set S satisfying condition 2 of the lemma. This decision is correct due to Lemma 2.2.0.12. This lemma permits the use of the first S found to orient the vee structures.

If M is not dag-isomorphic it would be possible for Phase 1 to build a graph that is not a pdag if it weren't for the failure condition in Step 2. The next example illustrates a failure resulting from an application of Phase 1 on a non-dag-isomorphic dependency model.

2.3.2.1 Example. Let $U = \{a, b, c, d\}$ and M be the closure of the set $\{I(a, c|\emptyset), I(a, d|\emptyset), I(b, d|\emptyset)\}$ under symmetry².

Step 1 of Phase 1 will construct the skeleton $a - b - c - d$, and $S(a, c) = S(a, d) = S(b, d) = \emptyset$. Since there is a chain \overline{abc} and $\neg\overline{ac}$ and $b \notin S(a, c)$ Step 2 could direct $a \rightarrow b \leftarrow c$. Similarly since \overline{bcd} and $\neg\overline{bd}$ and $c \notin S(b, d)$, Step 2 could direct $b \rightarrow c \leftarrow d$.

One of the two directions would be assigned first, then upon attempting the second the algorithm would FAIL. \square

²Symmetry states that $I(A, B|C)$ iff $I(B, A|C)$. Unless otherwise noted, dependency models are assumed to be closed under symmetry since this is a trivial operation.

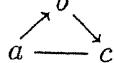
Phase 2

The task of Phase 2 is to find whether a pdag, G , has any extensions and to find one if such exists. This is a purely graph theoretic task; it does not involve M .

To prove that this phase of the construction is correct, it is sufficient to prove that each of the four rules is sound, namely, that the orientation choices dictated by these rules never need to be revoked.

- Rule 1: If $a \rightarrow b - c$ and a is not adjacent to c then direct $b \rightarrow c$.

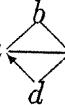
Directing $b - c$ as $b \leftarrow c$ would create a new vee structure, \overleftarrow{abc} , thus if there is a consistent extension it must contain $b \rightarrow c$.

- Rule 2: If  then $a \rightarrow c$.

Directing $a - c$ as $a \leftarrow c$ would create a directed cycle, $[abca]$, thus if there is a consistent extension it must contain $a \rightarrow c$.

- Rule 3: If  then direct $b \rightarrow d$.

Directing $b - d$ as $b \leftarrow d$ would imply that $a - b$ must be directed as $a \rightarrow b$ or else there would be a directed cycle, $[adba]$. Now if $b - c$ is directed as $b \rightarrow c$ then there is a directed cycle, $[bcdb]$, and if it is directed as $b \leftarrow c$ then there is a new vee structure, \overleftarrow{abc} . Thus if there is a consistent extension it must contain $b \rightarrow d$.

- Rule 4: If  then direct $a \rightarrow b \leftarrow c$.

First, $a - b$ must be directed as $a \rightarrow b$ or there would be a new vee structure, \overleftarrow{dab} . If $b - c$ is directed as $b \rightarrow c$ then $c - d$ cannot be directed as $c \rightarrow d$ or there would be a directed cycle, $[cdabc]$. Moreover, $c - d$ cannot be directed as $c \leftarrow d$ or there would be a new vee structure, \overleftarrow{bcd} . Thus if there is a consistent extension, then it must contain $a \rightarrow b \leftarrow c$.

Following are two simple examples of pdags which cannot be extended into dags.

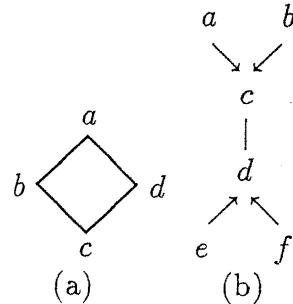


Figure 2.7: Two pdags which cannot be extended.

2.3.2.2 Example. Consider the graph of Figure 2.7.a. Initially, no rules apply, so the algorithm would select an arbitrary arc and direct it, without loss of generality assume it directs $a \rightarrow b$. Now Rule 1 will apply twice, directing $b \rightarrow c \rightarrow d$. However a third application to infer $d \rightarrow a$ would produce a directed cycle. It is easy to see that a cycle would result no matter which arc is initially chosen and no matter what initial directionality is assigned. Thus this graph has no dag extension. \square

2.3.2.3 Example. Consider the graph of Figure 2.7.b. Any application of Rule 1 to direct the arc $c - d$ would create a new vee structure. Hence this graph as well, has no dag extension. \square

Phase 3

The soundness of Step 1 follows from the definition of consistency; it simply checks if each and every independence statement of M is represented in D . The soundness of Step 2, namely that testing only statements of the form $I(a, U_a \setminus \mathcal{P}_a | \mathcal{P}_a)$ is sufficient follows from the proof of the soundness of d-separation[Ver86].

2.3.2.4 Example. Let $U = a, b, c$ and $M = \{I(a, b|\emptyset), I(a, c|\emptyset), I(b, c|\emptyset)\}$. Phase 1 will produce an empty graph which can trivially be extended into an empty dag. But every independence statement is true in an empty dag, including, e.g. $I(a, b|c)$ which is not in M . Thus M is not dag isomorphic.

□

2.3.3 Complexity Analysis

Phase 1 can be completed in $O(|M| + |U|^2)$ steps, as follows:

- Start with a complete graph G . For each statement, $I(A, B|S)$ in M , and for each pair of variables $a \in A$, and $b \in B$ remove the links $a - b$ from G and define $S(a, b) = S$.
- For each node a let $N(a) = \{b|a - b\}$ be the set of neighbors of a .
- For each separating set $S(a, b)$ defined above, note that $C(a, b) = N(a) \cup N(b) \setminus S(a, b)$ must be children of a and b so direct $a \rightarrow c \leftarrow b \forall c \in C(a, b)$.

Phase 2 may appear to require an exponential amount of time in the worst case due to possible backtracking in Step 2(c). However, we conjecture that if G is extendible, then Rules 1-4 are sufficient to guarantee that no choice will ever need to be revoked. Empirical studies have, so far, confirmed our conjecture. Furthermore, [Ver92] presents an alternative algorithm for Phase 2 based on the maximum cardinality search developed by [TY84], and which is provably a linear-time algorithm. This algorithm, however, is considerably more complicated and less intuitive than the one presented here.

If the conjecture is correct, it would be possible to replace the backtrack step with a definite failure, in which case the time complexity of this phase would be polynomial, no more than $O(|U|^4 * |E|)$. On the other hand, if it is not correct, the complexity could be exponential in $|E|$.

Phase 3 can be completed in $O(|M| * |E| + |M| * |U|)$ steps.

2.3.4 Extensions and Improvements

In general, the set of all independence statements which hold for a given domain will grow exponentially as the number of variables grows. Thus it

might be impractical to specify M by explicit enumeration of its I-statements. In such cases it may be desirable, instead, to specify a basis, L , such that M is the logical closure of L , (i.e. $M = CL(L)$), relative to some semantics, (e.g. the graphoid axioms, correlational graphoids axioms, or even probability theory).

The major difficulty in permitting the dependency model to be specified as the closure of some basis lies in solving the so called *membership problem*. Simply stated, the problem is to decide if a particular statement, I_0 , is contained in the closure, M , of a given list of statements, L . In general, membership problems are often undecidable, and of those that are decidable, many are NP-hard. In particular, the membership problems for both graphoids and probabilistic independence are unsolved [Gei90].

However, in spite of this difficulty, it may still be possible to have an efficient dag construction algorithm, because the queries required are of a special form. The algorithm makes four types of queries to M :

1. “Is there any S such that $I(a, b|S) \in CL(L)$?” (Phase 1, Step 1)
2. “Is b in any set S such that $I(a, c|S) \in CL(L)$?” (Phase 1, Step 2)
3. “Is every statement in $CL(L)$ represented in D ?” (Phase 3, Step 1)
4. “Is every statement represented in D in $CL(L)$?” (Phase 3, Step 2)

In the case that M is assumed to be the graphoid closure of L , queries of type 1, 2 and 3 are all manageable. The queries for Phase 1 can both be quickly answered due to the following lemma:

2.3.4.1 Lemma. *If \exists_S s.t. $I(a, b|S) \in CL(L)$ then $\exists_{A,B,C}$ s.t. $I(aA, bB|C) \in L$*

Proof: This can be proven by induction on the derivation of $I(a, b|S)$. If the derivation has length 0 then the lemma is trivial. If it is of length k then $I(a, b|S)$ must follow from one of the rules. Each rule has an antecedent with a separated from b in a manner satisfying the inductive hypothesis. Thus since this antecedent must have a derivation of length $< k$ the lemma holds. \square

Remark: Note that this simplification is possible due to the special form of these queries, namely that a and b are both singletons and any separating set will suffice.

Type 3 queries pose no particular problem since the axioms of graphoids hold for d-separation. Thus it is enough to check that each statement in L is represented in D to ensure that every statement in closure of L is represented in D .

However, to check that each statement represented in D is contained in $CL(L)$ it is necessary to make the $|U|$ membership queries explicated in Step 2 of Phase 3. Although these statements have a special form, it is yet unclear whether a lemma similar to 2.3.4.1 exists to simplify these queries.

Another possible source for simplification is to note that the dag D being tested in Step 2 of Phase 3 is not an arbitrary dag, but the output of the construction algorithm. While Example 2.3.2.4 demonstrates that it is possible for D to contain I-statements which are not in $CL(L)$, it may still be the case that any such I-statements must have either a certain form or some other property that would simplify the membership query.

2.3.5 Naive Representation of Graphoids

In the previous two sections various complexity results were given based upon different assumptions made about the representation of the input dependency model. Originally, it was assumed that the dependency model was an explicit list, and the analysis was in terms of the length of that list. The second analysis assumed that the given dependency model was the graphoid closure of a given list. In this section the number of queries to the dependency model needed for both recovery and the test for d-mapness will be shown to increase exponentially in the size of $|U|$ when no assumptions are made about the representation of the dependency model.

2.3.5.1 Example. Let $\bar{n} = \{n_1, \dots, n_k\}$, let $U = \{a, b\} \cup \bar{n}$, and let $E = \{n_i \rightarrow n_j | 1 \leq i < j \leq k\}$. For any $S \subseteq \bar{n}$, let $E_S = \{a \leftarrow n_i \rightarrow b | n_i \in S\} \cup \{a \rightarrow n_i \leftarrow b | n_i \notin S\}$, and let $D_S = \langle U, EE_S \rangle$.

Observation. $I(a, b|S)$ is the only d-separation which holds in D_S .

Since a and b are the only non-adjacent nodes in D_S , it follows that they are the only nodes which could be separated. Since every node in S is a parent of both a and b , it follows that any set which separates a from b

must contain every element of S . Finally, since a and b are both parents of every node outside of S , it follows that any separating set can only contain elements of S . Hence S is the only separating set. \square

2.3.5.2 Theorem. *The Naive recovery of a dag-isomorphic dependency model takes exponential time on average.*

Proof: In naive recovery, only queries about the truth of single I-statements are allowed. Consider $\overline{D} = \{D_S\}$, the set of dags constructed in Example 2.3.5.1 and $\overline{M} = \{M_S\}$ the corresponding set of dependency models. Recovery of the dag corresponding to a given $M \in \overline{M}$ is not harder than the general task of dag recovery since more information is given. However, naive recovery of the dag corresponding to M is the same as trying to guess which set $S \subset \overline{n}$ corresponds to M by asking only questions of the form “Is S' equal to S ?” Since there are an exponential number of possible sets and there is no way to rule out more than one set per question, this process will take an exponential amount of time on average. \square

2.3.5.3 Example. Let $\overline{n} = \{n_1, \dots, n_k\}$, let $U = \{a, b, c\} \cup \overline{n}$, $E = \{a \rightarrow x | x \in U \setminus a\}$, let $D = \langle U, E \rangle$ and $M_D = I[D]$. For any $S \subseteq \overline{n}$, let $M_S = M_D + I(b, c|S) + I(c, b|S)$.

Observation. *For every $S \subseteq \overline{n}$, M_S is a graphoid, D is an I-map of M_S and D is not a D-map M_S .*

Consider that every triple in M_D has the form $(T_1, T_2 | aT_3)$ where T_1 , T_2 and T_3 are mutually exclusive subsets of $U \setminus a$. Since $M_S = M_D + I(b, c|S) + I(c, b|S)$, it must be closed under the unary axioms of symmetry, decomposition and weak union because both M_D and $\{I(b, c|S), I(c, b|S)\}$ are individually closed, and the axioms are unary. It remains to show that M_S is closed under contraction. Consider the contraction axiom:

$$I(X, Y|Z) \text{ and } I(X, W, YZ) \Rightarrow I(X, YW, Z)$$

There are four ways for $I(b, c|S)$ and $I(c, b|S)$ to unify with this axiom:

- 1 $I(b, c|S)$ and $I(b, W|cS) \Rightarrow I(b, cW|S)$
- 2 $I(c, b|S)$ and $I(c, W|bS) \Rightarrow I(c, bW|S)$
- 3 $I(b, S_1|S_2)$ and $I(b, c|S) \Rightarrow I(b, cS_1|S_2)$
- 4 $I(c, S_1|S_2)$ and $I(c, b|S) \Rightarrow I(c, bS_1|S_2)$

In each case the axiom fails because the other part of the LHS of the axiom cannot unify with any element of M_S . Thus M_S is closed.

D is trivially an I-map of $M_S = M_D + I(b, c|S) + I(c, b|S)$, since it is a perfect map of M_D . It is also trivially not a D-map since b and c are not d-separated by any set which does not contain a . \square

2.3.5.4 Theorem. *The naive test for d-mapness takes exponential time on average.*

Proof: In the naive test, only queries about the truth of single I-statements are allowed. Consider D and $\overline{M} = \{M_S\}$, the dag and graphoids of Example 2.3.5.3. Deciding if a particular graphoid $M \in \overline{M} + M_D$ is a D-map of D is not harder than the general D-mapness decision problem because more information is given. However, deciding if M is a D-map of D is equivalent to guessing which, if any, S was chosen by asking only questions of the form "Is S' equal to S ?". Since there are an exponential number of possible sets and there is no way to rule out more than one set per question, this process will take an exponential amount of time on average. In fact, it will take an exponential amount of time for any positive answer since b and c must be dependent given every $S \subset \bar{n}$. \square

2.4 Historical and Bibliographic Remarks

Chapter 3

Latent Structures

In complex systems, it is usually the case that some relevant quantities are not directly observable. Thus it is necessary to consider the impact of these latent variables in the process of causal modeling.

3.1 Formal Definition

A latent structure is simply defined as a causal model which contains some latent variables, $U_{\mathcal{L}}$, and some observable variables, $U_{\mathcal{O}}$.

3.1.0.5 Definition. A latent structure is any pair $L = \langle D, U_{\mathcal{O}} \rangle$ where $D = \langle U, E \rangle$ is a causal model and $U_{\mathcal{O}} \subseteq U$ is a set of observable variables. The set $U_{\mathcal{L}} = U \setminus U_{\mathcal{O}}$ are the latent variables of L . \square

3.1.1 Probabilistic Semantics

The statistical meaning of a latent structure is given by the constraints that it places upon the joint distribution of its observable variables, namely:

$$P(U_{\mathcal{O}}) = \sum_{U_{\mathcal{L}}} \prod_x \in UP(x|\text{pa}(x))$$

Furthermore, concepts such as consistency, preference and equivalence are defined in terms of the observable characteristics of latent structures.

3.1.1.1 Definition. A latent structure $L = \langle D, U_O \rangle$ is **consistent** with a distribution P if it can accommodate some theory that generates P , i.e. there exists a set of parameters Θ_L s.t. $\sum_{U_O} P[\langle D, \Theta_D \rangle] = P$ \square

3.1.1.2 Definition. One latent structure $L = \langle D, U_O \rangle$ is **preferred** to another $L' = \langle D', U_O \rangle$ written, $L \preceq L'$ iff the first model D' can *mimic* the second D over U_O , i.e. for every set of parameters Θ_D there exists another set $\Theta'_{D'}$ s.t. $\sum_{U_O} P[\langle D', \Theta'_{D'} \rangle] = \sum_{U_O} P[\langle D, \Theta_D \rangle]$. \square

3.1.1.3 Definition. Two latent structures are **equivalent**, written $L' \equiv L$, iff $L \preceq L'$ and $L \succeq L'$. \square

3.1.2 Dependency Equivalence versus Equivalence

Figure 3.1 illustrates a special problem that embedded causal models pose (hidden variables are denoted by greek letters). Unlike simple causal models,

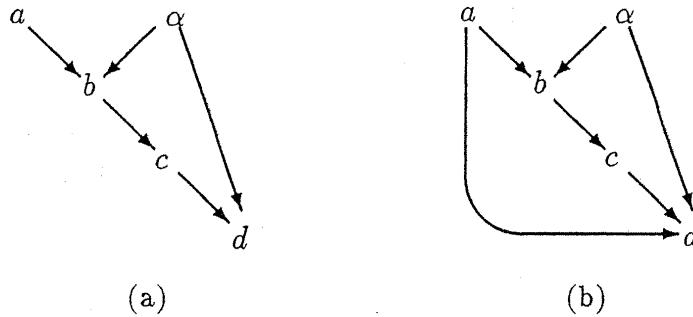


Figure 3.1: Two dependency equivalent embedded causal models which are not equivalent in general.

the statistical meaning of an embedded causal model cannot be completely characterized by dependency information alone; two dependency equivalent causal models need not be equivalent in the general sense. For example both embedded causal models of Figure 3.1 represent the dependency statement

$I(a, b, c)$, but the first model (a) imposes an additional constraint upon the set of distributions it can describe:

$$\sum_b P(b|a)P(d|abc) = f(c, d)$$

Fortunately, dependency equivalence is a tight enough necessary condition for equivalence that it permits many sound conclusions to be derived by graphical means.

3.1.3 Projections

3.1.3.1 Definition. The *projection* of a latent structure $L = \langle D, U_{\mathcal{O}} \rangle$ is any latent structure, $L' = \langle D', U'_{\mathcal{O}} \rangle$ where $D = \langle U, E \rangle$ and $D' = \langle U', E' \rangle$ with the following four properties:

1. $U'_{\mathcal{O}} = U_{\mathcal{O}}$
2. Every node in $U'_{\mathcal{L}}$ is a source node of D' .
3. Every node in $U'_{\mathcal{L}}$ has exactly two children, both of which are in $U_{\mathcal{O}}$.
4. $I[L] = I[L']$

□

Projection Algorithm

Input: A latent structure $L = \langle D, U_{\mathcal{O}} \rangle$, where $D = \langle U, E \rangle$.

Output: A latent structure $L' = \langle D', U'_{\mathcal{O}} \rangle$, where $D' = \langle U', E' \rangle$.

1. Let $U'_{\mathcal{O}} = U_{\mathcal{O}}$.
2. Let $U'_{\mathcal{L}} = \emptyset$.
3. Let $E' = \emptyset$.
4. For each pair of variables a and b check if there exists a directed path from a to b in E such that every internal node of the path is in $U_{\mathcal{L}}$.
If there is such a path, then add $a \rightarrow b$ to E' .

5. For each pair of variables a and b check if there exists a divergent path between a and b in E such that every internal node of the path is in U_L . If there is such a path, then add $v_{a,b}$ to U_L and add $a \leftarrow v_{a,b} \rightarrow b$ to E' .

3.1.3.2 Theorem. *The output of the projection algorithm is a projection of the input.*

Proof: Let L' be the output of the projection algorithm when given L as input. L' trivially obeys the first three conditions of a projection. To demonstrate that the fourth condition holds, it is enough to show, for any set S , that every S -active path between observable nodes in one dag has a corresponding S -active path in the other. This can be shown in eight steps:

1. *If there is a directed path \bar{p} from a to b in D then there is one, \bar{p}' in D' .*

If \bar{p} contains no internal observable nodes then $\bar{p}' = a \rightarrow b$ in D' . If \bar{p} contains k internal observable nodes then pick any one, p_i . The subpaths $\bar{p}_{1,i}$ and $\bar{p}_{i,|p|}$ each contain fewer than k internal observable nodes, hence by induction there exists directed paths from p_1 to p_i and from p_i to $p_{|p|}$ in D' . Let \bar{p}' be the concatenation of them.

Note that this result implies that for any set S , all S -active nodes in D are S -active in D' .

2. *If there is a path \bar{p} between a and b in D which contains no internal observable nodes and no head to head nodes, then there exists a corresponding path \bar{p}' in D' .*

Since \bar{p} contains no head to head nodes, it must either be directed from a to b , directed from b to a , or is a divergent path between a and b . The former two cases have already been covered. If there is a divergent path then $\bar{p}' = a \leftarrow v_{a,b} \rightarrow b$ in D' .

Note that a path with no internal observable nodes must be active given any set.

3. *For any set S , if there is an S -active path \bar{p} between a and b which contains no internal observable nodes and k head to head nodes, then there exists a corresponding S -active path \bar{p}' in D' .*

The case for $k = 0$ is already covered. If $k > 0$ then let p_i be any head to head node on \bar{p} . Since \bar{p} is S -active, p_i must have a descendant in S , let \bar{q} be the shortest path from p_i to some node in S , and let q_j be the first observable on this path. The paths $\overline{p_{1,i}} \oplus \overline{q_{1,k}}$ and $\overline{q_{1,k}} \oplus \overline{p_{i,|p|}}$ are both S -active paths between observables of D which contain no internal observable nodes and fewer than k head to head nodes. Thus by induction, there exist corresponding S -active paths \bar{r}' and \bar{s}' in D' . The path $\bar{p}' = \bar{q}' \oplus \bar{r}'$ is active since \bar{q}' and \bar{r}' are both active and they meet at the head to head node q_k which is active.

4. *For any set S , if there is an S -active path \bar{p} between a and b which contains k internal observable nodes, then there exists a corresponding S -active path \bar{p}' in D' .*

The case for $k = 0$ is already covered. If $k > 0$ then let p_i be any internal observable node on \bar{p} . The subpaths $\overline{p_{1,i}}$ and $\overline{p_{i,|p|}}$ are both S -active paths between observables of D with fewer than k internal observable nodes. Thus by induction, there exist corresponding S -active paths \bar{r}' and \bar{s}' in D' . The path $\bar{p}' = \bar{q}' \oplus \bar{r}'$ is active since \bar{q}' and \bar{r}' are both active and they meet at p_i which is head to head in D iff it is head to head in D' .

5. *If there is a directed path from a to b in D' then there is one in D .*

Suppose that there is a directed path \bar{p}' from a to b in D' . If \bar{p}' contains no internal observable nodes then it must be a direct link, which would imply the existence of a directed path from a to b in D . If \bar{p}' contains k internal observable nodes then pick any one, p'_i . The paths $\overline{p'_{1,i}}$ and $\overline{p'_{i,|p'|}}$ each contain fewer than k internal observable nodes, hence by induction there exists directed paths from p_1 to p_i and from p_i to $p_{|p|}$ in D' .

6. *If there is a path \bar{p} between a and b in D' which contains no internal observable nodes and no head to head nodes, then there exists a corresponding path in D .*

Since \bar{p} contains no head to head nodes, it must either be directed from a to b , directed from b to a , or is the path $a \leftarrow v_{a,b} \rightarrow b$. The former two cases have already been covered. In the latter case, there would be a divergent path between a and b in D .

7. For any set S , if there is an S -active path \bar{p}' between a and b which contains no internal observable nodes and k head to head nodes, then there exists a corresponding S -active path \bar{p} in D' .

The case for $k = 0$ is already covered. If $k > 0$ then let p'_i be any head to head node on \bar{p}' . The paths $\bar{p}_{1,i}$ and $\bar{p}_{i,|p'|}$ are both S -active paths between observables of D' which contain no internal observable nodes and fewer than k head to head nodes. Thus by induction, there exist corresponding S -active paths \bar{r} and \bar{s} in D . The path $\bar{p} = \bar{q} \oplus \bar{r}$ is active since \bar{q} and \bar{r} are both active and they meet at the head to head node p'_i which is active.

8. For any set S , if there is an S -active path \bar{p}' between a and b which contains k internal observable nodes, then there exists a corresponding S -active path \bar{p} in D' .

The case for $k = 0$ is already covered. If $k > 0$ then let p'_i be any internal observable node on \bar{p}' . The subpaths $\bar{p}'_{1,i}$ and $\bar{p}'_{i,|p'|}$ are both S -active paths between observables of D' with fewer than k internal observable nodes. Thus by induction, there exist corresponding S -active paths \bar{r} and \bar{s} in D . The path $\bar{p} = \bar{q} \oplus \bar{r}$ is active since \bar{q} and \bar{r} are both active and they meet at p'_i which is head to head in D' iff it is head to head in D .

□

Partially-directed graphs offer an excellent tool for describing equivalence classes of causal models; it would be desirable to find a corresponding tool for latent structures. Such a tool requires the ability to represent a direct non-causal correlation between two variables. In a simple causal model, whenever two variables are unseparable, there must be a directed link between them, dictating that either the first causes the second or the second causes the first. There is no way to represent the existence of an unknown common cause, as illustrated in the following latent structure (Figure 3.2 (a)). Assume a, b, c and d are the observables and α is unobservable. There is no dag that can represent the dependencies between a, b, c and d using these variables only. However, the *hybrid graph* (Figure 3.2 (b)) which contains a *bi-directional* link does represent these dependencies under a natural extension of d-separation called h-separation. The definition for h-separation is identical to that of

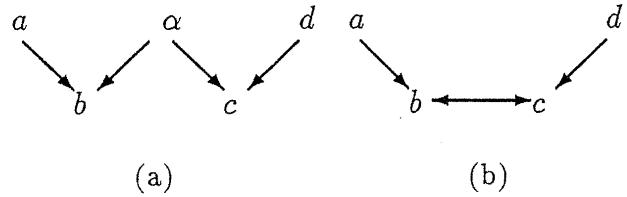


Figure 3.2: The representation of a hidden common cause.

d-separation as long as head-to-head is taken to mean $a \rightarrow b \leftarrow c$, $a \leftrightarrow b \leftarrow c$ or $a \rightarrow b \leftrightarrow c$ (i.e. b is head-to-head on all three paths).

For hybrid graphs, the notation \overrightarrow{ab} denotes the existence of a link with at least an arrow head pointing at b , namely either $a \rightarrow b$ or $a \leftrightarrow b$, while \overline{ab} denotes the existence of a link without any constraints on its orientation. Thus, for example, when applied to a dag, \overline{ab} means $a \rightarrow b$ or $a \leftarrow b$; while in hybrid graphs \overline{ab} denotes the existence of any of the four possible types of links, (namely, $a - b$, $a \rightarrow b$, $a \leftarrow b$ and $a \leftrightarrow b$). The notion of d-separation can now be extended to hybrid-graphs.

Note that hybrid-graphs are intended to be isomorphic to projections, thus they may be multiply connected. The definition of a projection does not preclude the possibility of two adjacent \overline{ab} nodes having a common latent parent, thus it may be the case that $a \leftrightarrow b$ and $a \rightarrow b$. In this case there are two paths with one link one between a and b , one of which points at both ends, and one that points only at a .

3.1.3.3 Definition. [h-separation] For any hybrid graph G , and any three disjoint sets of nodes, A , B and S , A is h-separated from B given S , written $I_G(A, B|S)$, if and only if there is no S -active path between any node in A and any node in B .

A path is S -active if and only if every head to head node of the path is S -active and every other node of the path is not an element of S .

A node is S -active if and only if there is a directed path from it to some element of S . \square

The terms S -active, “active given S ” and “activated by S ” are used interchangeably. The term **context** denotes a subset of nodes which are to be used as an activating set. When applied to a dag, h-separation is equivalent to d-separation. Two sets are **unseparable** if there is no set which can h-separate them.

Two graphs that represent the same set of h-separations will be called h-equivalent. Under the present interpretation of hybrid graphs, two latent structures are dependency equivalent iff their projections are h-equivalent.

3.1.3.4 Corollary. *Every S -active path has a corresponding simple S -active path.*

Proof: Assume that two nodes a and b are connected by at least one S -active path. Pick any such path p which has a minimal number k of cyclic subpaths. If $k > 0$ then p would not be simple and there would be some $i < j$ such that $p_i = p_j$. Let $q = p_{1i} \oplus p_{j|p|}$. Now other than q_i , every head-to-head node of q would be S -active and every non-head-to-head node would not be in S . If q_i is not head-to-head on q then either p_i or p_j was not head-to-head on p , thus $q_i \notin S$ and q would be S -active. If q_i is head-to-head and either p_i or p_j is head-to-head then both q_i and q must be S -active. Finally, if q_i is head-to-head and neither p_i nor p_j is head-to-head then the sub-path p_{ij} could not point at either of its end nodes and therefore must contain a head-to-head node. The head-to-head node closest to p_i must be an S -active descendent of p_i . Thus q_i and q must be S -active as well. Therefore in any case q is an S -active path. Furthermore, since every cyclic subpath of q would also be a subpath of p , and q would not contain the cycle p_{ij} it follows that q would have strictly fewer than k cycles. But, by assumption, p is an S -active path with a minimal number of cycles, hence k must be 0 and p is simple. \square

3.2 Invariant Properties

3.2.1 Inducing Paths

3.2.1.1 Definition. [Inducing Path] A path between two variables a and b of a latent structure is an **inducing path** iff every internal node of the path is both head-to-head on the path and an element of \mathcal{A}_{ab} . \square

3.2.1.2 Corollary. *For every inducing path there is a corresponding simple inducing path.*

Proof: Assume that two nodes a and b are connected by at least one inducing path. Pick any such path p which has a minimal number k of cyclic subpaths. If $k > 0$ then p would not be simple and there would be some $i < j$ such that $p_i = p_j$. Let $q = p_{1:i} \oplus p_{j|p|}$. Since every cyclic subpath of q would also be a subpath of p and q would not contain p_{ij} , q would have strictly fewer than k cycles. At the same time q would be an inducing path between a and b but by assumption, p had a minimal number of cycles. Therefore $k = 0$ and every inducing path has a corresponding simple path. \square

3.2.1.3 Corollary. *An inducing path can have at most one (singly) directed link.*

Proof: Let p be an inducing path. Since every interior node is head-to-head, every link must point at any incident interior nodes. Hence the only links which could possibly be singly directed are the first and last. Without loss of generality suppose that the first link were directed $p_1 \rightarrow p_2$. Since p_2 must be an ancestor of one of the ends of p it follows that both p_1 and p_2 are ancestors of $p_{|p|}$. Thus if the last arc were also singly directed, there would be a directed loop. Hence at most one link of p can be singly directed. \square

3.2.1.4 Lemma. *There exists an inducing path between two nodes of a latent structure iff the nodes are unseparable.*

Proof: To show that unseparability implies the existence of an inducing path, consider any two nodes a and b and their ancestor set \mathcal{A}_{ab} . There must exist an \mathcal{A}_{ab} -active path p between a and b , or the nodes would be separable. Since p is \mathcal{A}_{ab} -active it follows that every head-to-head node on p must be an ancestor of either a or b . Furthermore, since every internal non-head-to-head node of a path must be an ancestor of one of the path's end nodes or of one of the path's head-to-head nodes, it follows that every internal node of p is in \mathcal{A}_{ab} . Thus, if any internal node were not head-to-head, p would not be \mathcal{A}_{ab} -active. Therefore every internal node is head-to-head and p is an inducing path.

To complete the lemma, it remains to show that the existence of an inducing path p between two nodes a and b implies their unseparability. For

any given context S , it is enough to construct an S -active path p' between a and b . Such a p' can be constructed by dividing p into three pieces, p_{1i} , p_{ij} and $p_{j|p|}$ and replacing the first and last parts by two new paths q and r , where q is a directed path from p_i to a and r is a directed path from p_j to b .

First consider that every internal node of p is head-to-head by definition. Furthermore, any of these nodes may or may not be S -active, but every one of these nodes is in \mathcal{A}_{ab} . If every ancestor of b on p is S -active, then let $p_j = b$, i.e. $j = |p|$, otherwise let p_j be the inactive ancestor of b that is closest to a . Similarly, if every ancestor of a on the sub-path p_{1i} is S -active then let $p_i = a$ otherwise let p_i be the inactive ancestor of a that is closest to p_j . Let q be any directed path from p_i to a (if $p_i = a$ then $q = [a]$). Similarly, let r be any directed path from p_j to b .

Both q and r are S -active since they contain no head-to-head nodes and if either contained an element of S then p_i or p_j respectively would be an S -active head-to-head node on p . Further p_{ij} is S -active since it contains only active head-to-head nodes (if any) by the selection of i and j . Finally, p_i and p_j are not head-to-head on the path $p' = q \oplus p_{ij} \oplus r$. Therefore p' is an S -active path between a and b and a and b are unseparable. \square

3.2.1.5 Corollary. *If p is an inducing path between two nodes a and b of a latent structure L and p points at b but not a then L contains a directed path from a to b . Furthermore, for any context S there exists an S -active path in L which points at b but not a .*

Proof: If p contains no internal nodes, then it must be a single link $a \rightarrow b$ which is both the desired directed and S -active path.

Suppose p contains at least one internal node. Since p does not point at a it must be the case that $a \rightarrow p_2$ and since p is an inducing path, $p_2 \in \mathcal{A}_{ab}$. Therefore both a and p_2 are ancestors of b and there must be a simple directed path s from a to b .

Now consider the node p_j chosen in the proof of lemma 3.2.1.4. If $j = 2$ then directed path s is the desired S -active path. If $j \neq 2$ then p_2 must be active which implies that a and all of its ancestors are active thus $i = 1$ and $p' = p_{1j} \oplus r$ is the desired S -active path. \square

3.2.1.6 Corollary. *If p is an inducing path of a latent structure L and p points at both of its end nodes then for any context S there exists an S -active path in L which points at both of its end nodes.*

Proof: The path p' constructed in the proof of the lemma 3.2.1.4 will be S -active and will point at both of its end nodes since the only time a portion of the path is replaced, it is replaced by a directed path that points at the corresponding end node. \square

3.2.2 Vee Structures

3.2.2.1 Definition. [vee] A vee is a ternary relation, written $\nu(a, b, c)$, where a and c are the ends and b is the center of the vee. A latent structure contains a vee if both of its ends are connected to the center by inducing paths that point at the center and if the ends themselves are not connected by any inducing paths. \square

3.2.2.2 Lemma. *Three nodes form a vee in a latent structure iff the center node is unseparable from either end node and if the end nodes are separable by some set which doesn't contain the center node.*

Proof: If a latent structure L contains $\nu(a, b, c)$ then there must be inducing paths between a and b and between b and c . Thus by lemma 3.2.1.4 it follows that the center node is unseparable from either end node. Furthermore, by definition, there is no inducing path between a and c so there must exist a set S which separates them. Since both inducing paths point at b it follows from corollaries 3.2.1.5 and 3.2.1.6 that there exist S -active paths p and q which point at b and connect it with a and c respectively. Assume that these paths are simple (corollary 3.1.3.4), thus b occurs only once on $r = p \oplus q$ and is head-to-head on it which means that r is Sb -active. But since r connects a and c and they are h-separated by S it follows that $b \notin S$.

It remains to show that the conditions for the existence of a vee are complete. Suppose that b is unseparable from either a or c and that a and c are themselves separable by some set $S \subseteq U - abc$. Lemma 3.2.1.4 implies that there must be inducing paths between a and b and between b and c . If either of these inducing paths didn't point at b then corollaries 3.2.1.5 and 3.2.1.6 would imply the existence of simple S -active paths p and q which connect b with a and c respectively in such a manner that b would not be head-to-head on the path $r = p \oplus q$. This would mean that r is an S -active path connecting a and c . However, a and c are separated by S , hence both

inducing paths must point at b . Finally, the existence of a separating set S implies that there is no inducing path between a and c by lemma 3.2.1.4. \square

3.2.3 Kite Structures

3.2.3.1 Definition. [kite] A kite is a four place relation, written $\kappa(a, b, c, d)$, where a is the **tail**, b is the **center**, c is the **top** and d is the **bottom** of the kite. A latent structure L contains a kite iff there are inducing paths between (1) the tail and center, pointing at the center, (2) the center and the top, pointing at both, (3) the center and the bottom, pointing only at the bottom and (4) the top and bottom pointing at both and if there is no inducing path between the tail and the top nor the bottom. \square

3.2.3.2 Lemma. *Four nodes of a latent structure form a kite, $\kappa(a, b, c, d)$, iff the center node is unseparable from any of the other nodes and if the top and bottom nodes are unseparable and if there exists some set $R \subset U - abcd$ which separates the tail from the upper node and another set $S \subset U - acd$ which contains the center and separates the tail from the lower node.*

Proof: (soundness) If a latent structure L contains a kite $\kappa(a, b, c, d)$ then there are inducing paths connecting the center with the tail, top and bottom nodes and there is one connecting the top and bottom. These paths imply by lemma 3.2.1.4 that the center is unseparable from any other node and that the top and bottom nodes are unseparable.

Furthermore, there is no inducing path connecting the tail to the top, so there must exist a set R which separates the two nodes. The inducing paths imply that there are simple R -active paths connecting a with b and b with c . The concatenation of these two paths p will contain b only once and as a head-to-head node because the inducing paths both pointed at b . And p is Rb -active thus $b \notin R$ as the two nodes are separated by R alone. Furthermore, there is a directed path from b to d because of the inducing path connecting them and corollary 3.2.1.5. Thus p is Rd -active as well, which implies that $R \subseteq U - abcd$.

Finally, since a and d are separable, there must be a separating set S . The inducing paths connecting a with b and b with d imply the existence of corresponding simple S -active paths. The concatenation of these paths q also

contains b only once, but this time as a non-head-to-head node, thus $b \in S$ or else q would be S -active. Furthermore, there must be simple S -active paths connecting a with b , b with c and c with d . Let their concatenation be r . If c is on r more than once then c must be on the S -active path between a and b , in which case it must be an ancestor of either a or b by the construction used in lemma 3.2.1.4. Since b is a parent of d it follows that the sequence $[abcd]$ is an inducing path from a to d , but a and d are separable, thus c cannot be on the S -active path between a and b . Therefore c appears only once on r and is head-to-head on r , which implies that r is Sc -active. Hence $c \notin S$ and $S \subseteq U - acd$.

(completeness) Due to the unseparability constraints, lemma 3.2.1.4 implies the existence of inducing paths p between a and b , q between b and c , r between c and d and s between b and d . Further since a and c are separable by a set R which doesn't contain b it follows that both p and q must point at b otherwise there would exist R -active paths p' and q' whose concatenation would be R -active since b would not be head-to-head on $p' \oplus q'$. Similarly, since a and d are separated by a set S which contains b but not c , both q and r must point at c or there would exist S -active paths p' and q' and r' whose concatenation would be S -active since c would not be head-to-head on $p' \oplus q' \oplus r'$. And s cannot point at b otherwise there would be S -active paths p' and s' whose concatenation would be S -active since S contains b and b would be head-to-head on $p' \oplus s'$. Finally, r must point at d otherwise there would be a directed path from d to c , and hence from b to c . Thus the sequence $[abc]$ would be an inducing path between a and c . But a and c are separated by R . \square

3.2.4 Induced Graphs

3.2.4.1 Definition. [Pattern] The **rudimentary pattern** P of a latent structure L is the hybrid graph with fewest arrowheads satisfying the following conditions:

- (1) Two nodes are adjacent in P iff they are unseparable in L .
- (2) If $\nu(a, b, c) \in L$ then $\overset{\rightarrow}{ab} \in P$.
- (3) If $\kappa(a, b, c, d) \in L$ then $b \leftrightarrow c \leftrightarrow d \in P$.

\square

3.2.4.2 Corollary (Uniqueness). *Every latent structure has exactly one pattern.*

Proof: Let L be a latent structure, and P and Q be patterns of it. It is enough to show that P and Q have the same links and same arrowheads. If there is a link in one pattern, then there must be a corresponding inducing path in L and the link must exist in the other pattern. Similarly if there is an arrowhead on a link in one pattern, then that link must correspond to part of a vee or part of a kite in L , hence the corresponding link in the other pattern must have a corresponding arrowhead. \square

3.2.4.3 Corollary (Soundness). *h -Equivalent latent structures have the same pattern.*

Proof: The pattern is defined solely upon h -separation statements, hence if two latent structures are dependency-equivalent then they define the same set of h -separation statements and will have the same pattern. \square

In order to prove completeness, it is enough to show that every latent structure is h -equivalent to its pattern. This is done in two steps, first it is shown that the induced graph, i.e. the graph formed by adding induced arcs, is h -equivalent to the original and is a super graph of the pattern, then it is shown that removal of every arrowhead that is not part of a vee or kite does not affect the h -separation statements. Hence the pattern is h -equivalent.

The following result permits the definition of a unique induced graph.

3.2.4.4 Lemma. *If $\overset{\rightarrow}{ab}$ in the pattern of a latent structure L then $b \notin A_a$ in L .*

Proof: There are five ways for $\overset{\rightarrow}{ab}$ in the pattern P of L . Either (1) b is the center of a vee, (2) a is the center and b is the top of a kite, (3) a is the top and b is the center of a kite, (4) a is the top and b is the bottom of a kite, (4) a is the top and b is the bottom, or (5) a is the bottom and b is the top of a kite. It remains to show that in each of these cases, b cannot be an ancestor of a .

(1) If the center of the kite were an ancestor of either end, then the inducing path from one end to the center, when joined with the inducing path from the center to the other end would form an inducing path between the ends. But by definition the ends of a kite are separable.

(2) If the top of a kite were an ancestor of its center, then both would be ancestors of the bottom and there would be an inducing path from the tail to the bottom. But by definition, the tail and bottom of a kite are separable.

(3) If the center were an ancestor of the top then there would be an inducing path from the tail to the top which, by definition, are separable.

(4) If the bottom were an ancestor of the top then the center would be an ancestor of the top and there would be an inducing path between the two nodes which, by definition, are separable.

(5) If the top were an ancestor of the bottom then both the center and top would be ancestors of the bottom; once again there would be an inducing path from the tail to the bottom which, by definition, are separable. \square

3.3 Historical and Bibliographic Remarks

Chapter 4

Inferring Causal Relationships from Observational Data

The stage is now set to infer causal relations from statistical data. The first step will be to define that a causal relation holds with respect to a probability distribution iff it holds in all minimal latent structures consistent with that distribution.

4.0.0.5 Definition. Given a probability distribution P , a variable c has a **causal influence** on e iff there exists a directed path from c to e in every minimal latent structure consistent with P . \square

Two other causal notions are important as well. The first captures the ability to prove the lack of a causal relationship in one direction, the second captures the ability to prove the lack of a causal relationship in both directions.

4.0.0.6 Definition. Given a probability distribution P , a variable c has a **potential causal influence** on e iff there does not exist a directed path from c to e in any minimal latent structure consistent with P . \square

4.0.0.7 Definition. Given a probability distribution P , a variable c is **spurious associated** with e iff there exists an inducing between c and e which ends pointing at both c and e in every minimal latent structure consistent with P . \square

Spurious association means that every minimal consistent latent structure contains an S -active path between the nodes that ends pointing at both, for every context S .

It turns out that while the minimality principle is sufficient for forming a normative and operational theory of causation, it does not guarantee that the search through the vast space of minimal models would be computationally practical. In general there could be many minimal models, having totally disparate structures. To facilitate an effective proof theory, we rule out such eventualities, and impose a restriction on the distribution called “stability” (or “dag-isomorphism” in [Pea88b]). It conveys the assumption that all vanishing dependencies are structural, not formed by incidental equalities of numerical parameters¹.

4.0.0.8 Definition. Let $I[P]$ denote the set of all conditional independence relationships embodied in P . A causal theory $T = \langle D, \Theta_D \rangle$ generates a stable distribution iff it contains no extraneous independences, i.e. $I[P[\langle D, \Theta_D \rangle]] \subseteq I[P[\langle D, \Theta'_D \rangle]]$ for any set of parameters Θ'_D . \square

Stability insures that every observed independence corresponds to a d-separation in the original causal model, or h-separation in the projection of the latent structure. With the added assumption of stability, every minimal causal model of a distribution is statistically equivalent to every other one, as long as there are no hidden variables. When there are hidden variables, every minimal latent structure of a distribution is dependency equivalent to every other one. The reason is simple, since the dag underlying any theory must be an I-map of the generated distribution, and a P-map when the distribution is stable, any distribution generated from that dag must either be dependency equivalent to the stable distribution (another P-map) or must contain a superset of independences. Thus, any properties of causal models which are invariant under dependency equivalence must be properties of any minimal causal model consistent with a stable distribution. However note that some non-minimal models will be dependency equivalent to the minimal ones, so there may very well be properties that are invariant among all of the

¹It is possible to show that, if the parameters are chosen at random from any reasonable distribution, then any unstable distribution has measure zero [SGS89]. Stability precludes deterministic constraints. Less restrictive assumptions are treated in [GPP90].

statistically equivalent minimal models but not among all models that are dependency equivalent to the minimal ones. This analysis will fail to capture such properties.

The search for the minimal model then boils down to recovering the structure of the underlying dag from queries about the dependencies portrayed in that dag. This search is exponential in general, but simplifies significantly when the underlying structure is sparse (see [SG91, VP90] for such algorithms).

4.1 Recovering Latent Structures

IC-Algorithm (Inductive Causation)

Input: P a sampled distribution.

Output: $\text{core}(P)$ a marked hybrid acyclic graph.

1. For each pair of variables a and b , search for a set S_{ab} such that (a, S_{ab}, b) is in $I(P)$, namely a and b are independent in P , conditioned on S_{ab} . If there is no such S_{ab} , place an undirected link between the variables.
2. For each pair of non-adjacent variables a and b with a common neighbor c , check if $c \in S_{ab}$.
If it is, then continue.
If it is not, then add arrowheads pointing at c , (i.e. $a \rightarrow c \leftarrow b$).
3. Form $\text{core}(P)$ by recursively adding arrowheads according to the following two rules:²
If \overline{ab} and there is a strictly directed path from a to b then add an arrowhead at b .
If a and b are not adjacent but \overrightarrow{ac} and $c \rightarrow b$, then direct the link $c \rightarrow b$.
4. If a and b are not adjacent but \overrightarrow{ac} and $c \rightarrow b$, then mark the link $c \rightarrow b$.

² \overline{ab} denotes adjacency, i.e. $a - b$, $a \rightarrow b$, $a \leftarrow b$ or $a \leftrightarrow b$; \overrightarrow{ab} denotes either $a \rightarrow b$ or $a \leftrightarrow b$.

The result of this procedure is a substructure called $\text{core}(P)$ in which every marked uni-directed arrow $x \rightarrow y$ stands for the statement: “ x has a causal influence on y (in all minimal latent structures consistent with the data)”. We call these relationships “genuine” causal influences (e.g. $c \rightarrow d$ in previous Figure 1a).

4.1.0.9 Definition. For any latent structure L , $\text{core}(L)$ is defined as the hybrid graph³ satisfying (1) two nodes are adjacent in $\text{core}(L)$ iff they are adjacent or they have a common unobserved cause in every projection of L , and (2) a link between a and b has an arrowhead pointing at b iff $a \rightarrow b$ or a and b have a common unobserved cause in every projection of L . \square

4.1.0.10 Theorem. (soundness) For any latent structure $L = \langle D, O \rangle$ and an associated theory $T = \langle D, \Theta_D \rangle$ if $P(T)$ is stable then every arrowhead identified by IC is also in $\text{core}(L)$.

Proof: Suppose that $c \rightarrow b$ is a marked link identified by IC. Then consider the link \overrightarrow{ac} . The arrowhead pointing at c must have been added either in step 2 or 3.

If it was added in step 2, then \overrightarrow{ac} in the pattern of L , hence it cannot be the case that c is an ancestor of a in any latent structure that is dependency equivalent to L . So there must be an inducing path between a and c pointing at c in every such latent structure.

If it was added in step 3, then either it was added either to avoid a directed cycle, or because of some other link \overrightarrow{da} and d is not adjacent to c . In either case there must be a path between a and c which ends pointing at c in every dependency equivalent latent structure.

Now consider that $c \rightarrow b$ is identified by IC. It must be the case that there is an inducing path between c and b in every such latent structure as well. But since a and c are not adjacent it follows that if this inducing path ended pointing at c in any dependency equivalent latent structure, it would have to in all. Furthermore, it would have been identified in step 2 of the IC algorithm, and the link would have been $c \leftrightarrow b$. Thus the inducing path between c and b cannot point at c . Hence there is a directed path from c to

³In a hybrid graph links may be undirected, uni-directed or bi-directed.

b in all dependency equivalent latent structures. And the arrow would be in $\text{core}(L)$. \square

4.1.0.11 Corollary. *If every link of the directed path from c to e is marked in $\text{core}(P)$ then c has a causal influence on e according to P .*

Proof: This follows directly from the previous theorem and the fact that causal influence is transitive. \square

4.2 Probabilistic Criteria for Causal Relations

The IC-algorithm takes a distribution P and outputs a dag, some of its links are marked uni-directional (denoting genuine causation), some are unmarked uni-directional (denoting potential causation), some are bi-directional (denoting spurious association) and some are undirected (denoting relationships that remain undetermined). The conditions which give rise to these labelings constitute operational criteria for the various kinds of causal relationships. In this section we present explicit criteria for potential and genuine causation, as they emerge from Theorem 4.1.0.10 and the IC-algorithm. Note that in each case, the criterion for causation between two variables, x and y , will require that a third variable z exhibit a specific pattern of interactions with x and y . This is not surprising, since the very essence of causal claims is to stipulate the behavior of x and y under the influence of a third variable, one that corresponds to an external control of x . Therefore, our criteria are in line with the paradigm of “no causation without manipulation” [Hol86]). The difference is only that the variable z , acting as a virtual control of x , must be identified within the data itself. The IC-algorithm provides a systematic way of searching for variables z that qualify as virtual controls.

Detailed discussions of these criteria in terms of virtual control are given in Sections 4.3 and 4.4.

4.2.0.12 Theorem (Potential Cause). *A variable x has a potential causal influence on another variable y (inferable from P), if*

1. *x and y are dependent in every context.*

2. There exists a variable z and a context S such that

- (i) x and z are independent given S
- (ii) z and y are dependent given S

Proof: Suppose that there is not a potential causal influence, i.e. that there is a directed path \bar{p} from y to x in some minimal latent structure L consistent with P . In this structure, there must be an S -active path \bar{q} connecting z and y . Now, $\bar{p} \oplus \bar{q}$ would be an S -active path between z and x unless \bar{p} is not S -active because y could not be head to head on that path. But if \bar{p} is not S -active then there must be some other S -active path \bar{r} joining x and y . If y is not head to head on $\bar{q} \oplus \bar{r}$ then it is an S -active path between x and z . If y is head to head on $\bar{q} \oplus \bar{r}$ then it is still an S -active path between x and z because some node on \bar{p} must have been in S or it would have been S -active. But by assumption, x and z are independent given S . Therefore, x does have a potential causal influence on y . \square

Note that this criterion precludes a variable x from being a potential cause of itself or of any other variable which functionally determines x .

4.2.0.13 Theorem (Genuine Cause). *A variable x has a genuine causal influence on another variable y if there exists a variable z such that:*

1. x and y are dependent in any context and there exists a context S satisfying:

- (i) z is a potential cause of x
- (ii) z and y are dependent given S .
- (iii) z and y are independent given Sx ,

or,

2. x and y are in the transitive closure of rule 1.

Proof: It is enough to show that condition (1) implies a causal influence since the causal influences relation is closed under transitivity.

Suppose that there is no genuine causal influence. Then there would be some minimal latent structure consistent with P in which does not contain a directed path from x to y .

Since z is a potential cause of x , they are dependent in any context, and there cannot be a directed path from x to z . There must be an S -active path \bar{p} between z and y , which must not be Sx -active. Thus x must be on \bar{p} and not head-to-head. Thus the subpath $\bar{p}_{1,i}$ between z and x must be S -active and must end pointing at x . The subpath $\bar{p}_{i,|p|}$ between x and y must also be S -active. If $\bar{p}_{i,|p|}$ contains no head to head nodes, then it would be a directed path from x to y since \bar{p} is S active and $\bar{p}_{1,i}$ ends pointing at $x \notin S$. So the only case left is if $\bar{p}_{i,|p|}$ contains a head to head node. Since x is not head to head on \bar{p} , x must be an ancestor of the first head-to-head on $\bar{p}_{i,|p|}$, and hence an ancestor of some element of S . Now consider that x and y are dependent in every context. Thus there must be an inducing path between them. If this inducing path ends pointing at x then there would be an S -active path q which ends pointing at x . But then $\bar{p}_{i,|p|} \oplus q$ would be an Sx -active path between z and y . Thus the inducing path between x and y must not point at x . But this implies that there is a directed path from x to y . \square

4.2.0.14 Theorem (Spurious Association). *Two variables x and y are spuriously associated if they are dependent in every context and there exists two other variables z_1 and z_2 , and a context S such that:*

1. z_1 and x are dependent given S
2. z_2 and y are dependent given S
3. z_1 and y are independent given S
4. z_2 and x are independent given S

Proof: Since x and y are dependent in every context, there must be an inducing path between them. If this path doesn't end pointing at x , then there would be an S -active path \bar{p} that doesn't end pointing at x . There must also be an S -active path \bar{q} between z_1 and x . But $\bar{p} \oplus \bar{q}$ would be an S -active path between z_1 and y . Thus the inducing path must end pointing at x . A similar argument implies that the path must also end pointing at y . \square

Succinctly, using the predicates I and $\neg I$ to denote independence and dependence respectively, the conditions above can be written:

1. $\neg I(z_1, x | S)$

2. $\neg I(z_2, y|S)$
3. $I(z_1, y|S)$
4. $I(z_2, x|S)$

Theorem 4.2.0.12 was formulated in [Pea90] as a relation between events (rather than variables) with the added condition $P(y|x) > P(y)$ in the spirit of [Goo83, Rei56, Sup70]. Condition 1 in Theorem 4.2.0.13 may be established either by statistical methods (per Theorem 4.2.0.12) or by other sources of information e.g., experimental studies or temporal succession (i.e. that z precedes x in time).

When temporal information is available, as it is assumed in the most theories of causality ([Gra88, Spo83, Sup70]), then Theorem 4.2.0.13 and 4.2.0.14 simplify considerably because every variable preceding and adjacent to x now qualifies as a “potential cause” of x . Moreover, adjacency (i.e. condition 1 of Theorem 4.2.0.12) is not required as long as the context S is confined to be earlier than x . These considerations lead to simpler conditions distinguishing genuine from spurious causes as shown next.

4.2.0.15 Theorem. (Genuine Causation with temporal information)
A variable x has a causal influence on y if there is a third variable z and a context S , both occurring before x such that:

1. $\neg I(z, y|S)$
2. $I(z, y|Sx)$

Proof: There must be an S -active path joining y and z . Since z and y are independent given Sx , x must be on the path and not be head to head. Suppose that the subpath between z and x did not end pointing at x . There would be a head to head node on this path. Consider the one closest to x . Since the path is S -active, this node must be an ancestor of a node in S . But then x would also be an ancestor of that node in S , which would be impossible since S occurs before x . Therefore the subpath between z and x ends pointing at x . Now since x is not head to head on the path joining z and y , the subpath of the path between x and y must not end pointing at x . Thus this subpath cannot contain any head to head nodes, for the one nearest x would imply that x precedes S . Thus the subpath between x and y must be directed from x to y . \square

4.2.0.16 Theorem. (Spurious Association with temporal information) *Two variables x and y are spuriously associated if they are dependent in every context, both x and some context S precede y and there exists a variable z satisfying:*

1. $I(z, y|S)$

2. $\neg I(z, x|S)$

Proof: There is an inducing path between x and y . It must end pointing at y or there would be a directed path from y to x violating the assumption that x precedes y . It must end pointing at x or the S -active path between z and x could form an S -active path between z and y when conjoined with the S -active path induced between x and y . \square

4.3 Causal Intuition and Virtual Experiments

This section explains how the formulation introduced above conforms to common intuition about causation and, in particular, how symmetric probabilistic dependencies can be transformed into judgements about causal influences. We shall first uncover the intuition behind Theorem 4.2.0.15, assuming the availability of temporal information, then (in Section 4.4) generalize to non temporal data, per Theorem 4.2.0.13.

The common intuition about causation is captured by the heuristic definition [Rub89]: “ x is a cause for y if an external agent interfering only with x can affect y ”.

Thus, causal claims are much bolder than those made by probability statements; not only do they summarize relationships that hold in the distribution underlying the data, but they also predict relationships that should hold when the distribution undergoes changes, such as those inferable from external intervention. The claim “ x causes y ” asserts the existence of a *stable* dependence between x and y , one that cannot be attributed to some prior cause common to both, and one that should be preserved when an exogenous control is applied to x .

This intuition requires the formalization of three notions:

1. That the intervening agent be “external” (or “exogenous”)

2. That the agent can “affect” y
3. That the agent interferes “only” with x

If we label the behavior of the intervening agent by a variable z , then these notions can be given the following probabilistic explications:

1. **Externality of z :** Variations in z must be independent of any factors W which precede x , i.e.,

$$I(z, W) \quad \forall W : t_W < t_x \quad (4.1)$$

2. **Control:** For z to effect changes in y (via x) we require that z and y be dependent, written:

$$\neg I(z, y) \quad (4.2)$$

3. **Locality:** To ensure that z interferes “only” with x , i.e., that its entire effect on y is mediated by x , we use the conditional independence assertion:

$$I(z, y|x) \quad (4.3)$$

to read “ z is independent of y , given x ”.

Note that (4.1) and (4.2) imply (by the axioms of conditional independence [Pea88b]) that x and y are dependent, namely, $\neg I(x, y)$.

Conditions (1) through (3) constitute the traditional premises behind controlled statistical experiments, with (1) reflecting the requirement that units selected for the experiment be chosen at random from the population under study. They guarantee that any dependency observed between x and y cannot be explained away by holding fixed some factors W preceding x , hence it must be attributed to genuine causation. The sufficiency of these premises is clearly not a theorem of probability theory, as it relies on temporal relationships among the variables. However, it can be derived from probability theory together with Reichenbach’s principle [Rei56], stating that every dependence $\neg I(x, y)$ requires a causal explanation, namely either one of the variables causes the other, or there must be a variable W preceding x and y such that $I(x, y|W)$ (see Figure 2). Indeed, if there is no back path from z to y through W (Eq. (4.1)) and no direct path from z to y avoiding x (Eq.

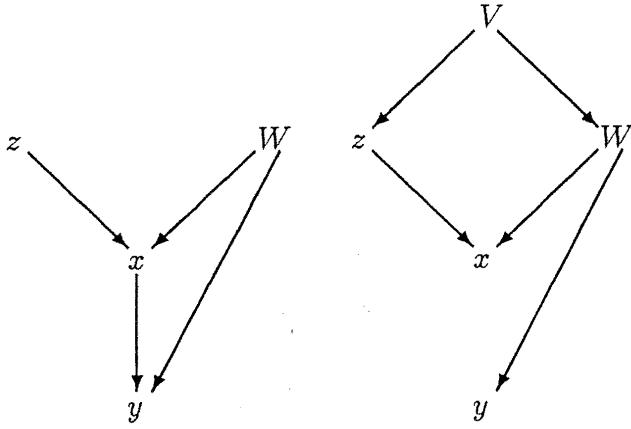


Figure 2

Figure 3

(4.3)) then there must be a causal path from x to y that is responsible for the dependence in Eq. (4.2)⁴.

In non-experimental situations it is not practical to detach x completely from its natural surrounding and to subject it to the exclusive control of an exogenous (and randomized) variable z . Instead, one could view some of x 's natural causes as “virtual controls” and, provided certain conditions are met, use the latter to reveal non-spurious causal relationship between x and y . In so doing we compromise, of course, condition (4.1), because we can no longer guarantee that those natural causes of x are not themselves affected by other causes which, in turn, might influence y (see Figure 3). However, it turns out that for stable distributions, conditions (4.2) and (4.3) are sufficient to guarantee that the association between x and y is non-spurious, thus intuitively justifying Theorem 4.2.0.15 for genuine causation.

The intuition goes as follows (see Figure 3): If the dependency between z and y (and similarly, between x and y) is spurious, namely, x and y are merely manifestations of some common cause W , there is no reason then for x to screen-off y from z , and condition (4.2) should be violated. In case condition (2) is accidentally satisfied by some strange combination of parameters, it is bound to be “unstable”, as it will be perturbed with any slight change of experimental conditions.

Conditions (4.2) and (4.3) are identical to those in Theorem 4.2.0.15,

⁴Cartwright [Car89] offers a sufficiency proof in the context of linear models.

save for the context S which is common to both. The inclusion of the fixed context S is legitimized by noting that if $P(x, y, z)$ is a marginal of a stable distribution, then so is the conditional distribution $P(x, y, z|S = s)$, as long as S corresponds to variables which precede x .

Theorem 4.2.0.15 constitutes an alternative way of recovering causal structures, more flexible than the IC-algorithm; we search the data for three variables z, x, y (in this temporal order) that satisfy the two conditions in some context $S = s$, and when such a triple is found, x is proclaimed to have a genuine causal influence on y . Clearly, permitting an arbitrary context S increases the number of genuine causal influences that can be identified in any given data; marginal independencies and even 1-place conditional independencies are rare phenomenon.

Note that failing to satisfy the test for genuine causation does not mean that such relationship is necessarily absent between the quantities under study. Rather, it means that the data available cannot substantiate the claim of genuine causation. To further test such claims one may need to either conduct experimental studies, or consult a richer data set where virtual control variables are found.

In testing this modeling scheme on real life data, we have examined the observations reported in Sewal Wright's seminal paper "Corn and Hog Correlations" [Wri21]. As expected, corn-price (x) can clearly be identified as a cause of hog-price (y), not the other way around. The reason lies in the existence of the variable corn-crop (z) that, by satisfying the conditions of Theorem 4.2.0.15 (with $S = \emptyset$), acts as a virtual control of x (see Figure 2). To test for the possibility of reciprocal causation, one can try to find a virtual controller for y , for example, the amount of hog-breeding (z'). However, it turns out that z' is not screened off from x by y (possibly because corn prices exert direct influence over farmer's decision to breed more hogs), hence, failing condition (iii), y disqualifies as a genuine cause of x . Such distinctions are important to policy makers in deciding, for example, which commodity, corn or hog, should be subsidized or taxed.

4.4 Non-Temporal Causation and Statistical Time

When temporal information is unavailable the condition that z precede x (Theorem 4.2.0.15) cannot be tested directly and must be replaced by an equivalent condition, based on dependence information. As it turns out, the only reason we had to require that z precede x is to rule out the possibility that z is a causal consequence of x ; if it were a consequence of x then the dependency between z and y could easily be explained away be a common cause W of x and y (see Figure 2).

The information that permits us to conclude that one variable is not a causal consequence of another comes in the form of an “intransitive triplet”, such as the variables a , b and c in Figure 1(a) satisfying: $I(a, b)$, $\neg I(a, c)$ and $\neg I(b, c)$. The argument goes as follows: If we create conditions (fixing S_{ab}) where two variables, a and b , are each correlated with a third variable c but are independent of each other, then the third variable cannot act as a cause of a or b , (recall that in stable distributions, common causes induce dependence among their effects); it must be either their common effect, $a \rightarrow c \leftarrow b$, or be associated with a and b via common causes, forming a pattern such as $a \leftrightarrow c \leftrightarrow b$. This is indeed the eventuality that permits our algorithm to begin orienting edges in the graph (step 2), and assign arrowheads pointing at c . It is also this intransitive pattern which is used to ensure that x is not a consequence of y (in Theorem 4.2.0.12) and that z is not a consequence of x (in Theorem 4.2.0.13). In Theorem 4.2.0.15 we have two intransitive triplets, (z_1, x, y) and (x, y, z_2) , thus ruling out direct causal influence between x and y , implying spurious associations as the only explanation for their dependence.

This interpretation of the intransitive triple is in line with the “virtual control” view of causation. For example, one of the reasons people insist that the rain causes the grass to become wet, and not the other way around, is that they can find other means of getting the grass wet, totally independent of the rain. Transferred to our chain $a - c - b$, we can preclude c from being a cause of a if we find another means (b) of potentially controlling c without affecting a [Pea88a, p. 396].

Determining the direction of causal influences from nontemporal data raises some interesting philosophical questions about the nature of time and

causal explanations. For example, can the orientation assigned to the arrow $x \rightarrow y$ in Theorem 4.2.0.13 ever clash with temporal information (say by a subsequent discovery that y precedes x)? Alternatively, since the rationale behind Theorem 4.2.0.13 is based on strong intuitions about how causal influences should behave (statistically), it is apparent that such clashes, if they occur, are rather rare. The question arises then, why? Why should orientations determined solely by statistical dependencies have anything to do with the flow of time?

In human discourse, causal explanations indeed carry two connotations, temporal and statistical. The temporal aspect is represented by the convention that a cause should precede its effect. The statistical aspect expects causal explanations (once accounted for) to screen off their effects, i.e., render their effects conditionally independent⁵. More generally, causal explanations are expected to obey many of the rules that govern paths in a directed acyclic graphs (e.g., the intransitive triplet criterion for potential causation, Section 4.3). This leads to the observation that, if agreement is to hold between the temporal and statistical aspects of causation, natural statistical phenomena must exhibit some basic temporal bias. Indeed, we often encounter phenomenon where knowledge of a present state renders the variables of the future state conditionally independent (e.g., multi-variables economic time series as in Eq. (4.4) below). We rarely find the converse phenomenon, where knowledge of the present state would render the components of the past state conditionally independent. The question arises whether there is any compelling reason for this temporal bias.

A convenient way to articulate this bias is through the notion of “Statistical Time”.

4.4.0.17 Definition. [Statistical Time] Given an empirical distribution P ,

⁵This principle, known as Reichenbach’s “conjunctive fork” or “common-cause” criterion [Rei56, SZ81] has been criticized by Salmon [Sal84], who showed that some events would qualify as causal explanations though they fail to meet Reichenbach’s criterion. Salmon admits, however, that when a conjunctive fork does occur, the screening off variable is expected to be the cause of the other two, not the effect [Sal84, p. 167]. He notes that it is difficult to find physically meaningful examples where a response variable renders its two causes conditionally independent (although this would not violate any axiom of probability theory). This asymmetry is further evidence that humans tend to reject causal theories that yield unstable distributions.

a statistical time of P is any ordering of the variables that agrees with at least one minimal causal model consistent with P . \square

We see, for example, that a scalar Markov-chain process has many statistical times; one coinciding with the physical time, one opposite to it and the others correspond to any time ordering of the variables away from some chosen variable. On the other hand a process governed by two coupled Markov chains,

$$\begin{aligned} x_t &= \alpha x_{t-1} + \beta y_{t-1} + \xi_t \\ y_t &= \gamma x_{t-1} + \delta y_{t-1} + \xi'_t, \end{aligned} \tag{4.4}$$

has only one statistical time – the one coinciding with the physical time⁶. Indeed, running the IC-algorithm on samples taken from such a process, while suppressing all temporal information, quickly identifies the components of x_{t-1} and y_{t-1} as genuine causes of x_t and y_t . This can be seen from Theorem 4.2.0.12, where x_{t-2} qualifies as a potential cause of x_{t-1} using $z = y_{t-2}$ and $S = \{x_{t-3}, y_{t-3}\}$, and Theorem 4.2.0.13, where x_{t-1} qualifies as a genuine cause of x_t using $z = x_{t-2}$ and $S = \{y_{t-1}\}$ of x_t .

The temporal bias postulated earlier can be expressed as follows:

4.4.0.18 Conjecture (Temporal Bias). *In most natural phenomenon, the physical time coincides with at least one statistical time.* \square

Reichenbach [Rei56] attributed the asymmetry associated with his conjunctive fork to the second law of thermodynamics. We are not sure at this point whether the second law can provide a full account of the temporal bias as defined above, since the influence of the external noise ξ_t and ξ'_t renders the process in (4.4) nonconservative⁷. What is clear, however, is that the temporal bias is *language dependent*. For example, expressing Eq.(4.4) in a different coordinate system (say, using a unitary transformation $(x'_t, y'_t) = U(x_t, y_t)$), it is possible to make the statistical time (in the (x', y') representation) run contrary to the physical time. This suggests that the apparent agreement between the physical and statistical times is a byproduct of human choice of linguistic primitives and, moreover, that the choice is compelled by a survival pressure to facilitate predictions at the expense of diagnosis and planning.

⁶ ξ_t and ξ'_t are assumed to be two independent, white noise time series. Also $\alpha \neq \delta$ and $\gamma \neq \beta$.

⁷We are grateful to Seth Lloyd for this observation.

4.5 Conclusions

The theory presented in this dissertation should dispel the belief that statistical analysis can never distinguish genuine causation from spurious covariation. This belief, shaped and nurtured by generations of statisticians [Fis53, Key39, Lin83, Nil22] has been a major hindrance in the way of developing a satisfactory, non-circular account of causation. In the words of Gardenfors [Gar88, page 193]:

In order to distinguish genuine from spurious causes, we must already know the causally relevant background factors. ... Further, the extra amount of information is substantial: In order to determine whether C is a cause of E, *all* causally relevant background factors must be available. It seems clear that we often have determinate beliefs about causal relations between events, even if we do not know exactly which factors are causally relevant to the events in question⁸.

This dissertation shows that such extra information is often unnecessary: Under the assumptions of model-minimality (and/or stability), there are patterns of dependencies that should be sufficient to uncover genuine causal relationships. These relationships cannot be attributed to hidden causes lest we violate one of the basic maxims of scientific methodology: the semantical version of Occam's razor. Adherence to this maxim explains why humans reach consensus regarding the directionality and nonspuriousness of causal relationships, in the face of opposing alternatives, perfectly consistent with experience. Echoing Cartwright [Car89] we summarize our claim with the slogan "No Causes In, Some Causes Out".

From a methodological viewpoint, our theory should settle some of the ongoing disputes regarding the validity of path-analytic approaches to causal modeling in the social sciences [Fre87, Lin83]. It shows that the basic philosophy governing path-analytic methods is legitimate, faithfully adhering to the traditional norms of scientific investigation. At the same time our results also explicate the assumptions upon which these methods are based, and the conditions that must be fulfilled before claims made by these methods can be accepted. Specifically, our analysis makes it clear that causal modeling must

⁸See also Cartwright [Car89] for a similar position, and for a survey of the literature.

begin with *vanishing* (*conditional*) *dependencies* (i.e. missing links in their graphical representations). Models that embody no vanishing dependencies contain no virtual control variables, hence, the causal component of their claims cannot be substantiated by observational studies. With such models, the data can be used only for estimating the parameters of the causal links once we are absolutely sure of the causal structure, but the structure itself, and especially the directionality of the links, cannot be inferred from the data. Unfortunately, such models are often employed in the social and behavioral sciences e.g. [Ken79].

On the practical side, we have shown that the assumption of model minimality, together with that of “stability” (no accidental independencies) lead to an effective algorithm of recovering causal structures, transparent as well as latent. Simulation studies conducted at our laboratory show that networks containing tens of variables require less than 5000 samples to have their structure recovered by the algorithm. For example, 1000 samples taken from the process shown in Eq. (5), each containing ten successive x, y pairs, were sufficient for recovering its double-chain structure (and the correct direction of time). The greater the noise, the quicker the recovery.

Another result of practical importance is the following: Given a proposed causal theory of some phenomenon, our algorithm can identify in linear time those causal relationships that could potentially be substantiated by observational studies, and those whose directionality and non-spuriousness can only be determined by controlled, manipulative experiments.

It should also be interesting to explore how the new criteria for causation could benefit current research in machine learning. In some sense, our method resembles a search through elements of a version space [Mit82], where each hypothesis stands for a causal theory. Unfortunately, this is where the resemblance ends. The prevailing paradigm in the machine learning literature has been to define each hypothesis (or theory, or concept) as a subset of observable instances; once we observe the entire extension of this subset, the hypothesis is defined unambiguously. This is not the case in causal modeling. Even if the training sample exhausts the hypothesis subset (in our case, this corresponds to observing P precisely), we are still left with a vast number of equivalent causal theories, each stipulating a drastically different set of causal claims. Fitness to data, therefore, is an insufficient criterion for validating causal theories. Whereas in traditional learning tasks we attempt to generalize from one set of instances to another, the causal modeling task

is to generalize from behavior under one set of conditions to behavior under another set. Causal models should therefore be chosen by a criterion that challenges their stability against changing conditions, and these show up in the data in the form of virtual control variables. Thus, the dependence patterns identified by Theorems 4.2.0.13 through 4.2.0.16 constitute islands of stability as well as virtual validation tests for causal models. It would be interesting to examine whether these criteria, when incorporated into existing machine learning programs would improve the stability of theories discovered by such programs.

Bibliography

- [Ash72] Robert B. Ash. *Real Analysis and Probability*. Academic Press, New York, 1972.
- [Car89] N. Cartwright. *Nature Capacities and Their Measurements*. Clarendon Press, Oxford, 1989.
- [Cli83] N. Cliff. “Some cautions concerning the application of causal modeling methods.” *Multivariate behavioral research*, **18**:115 – 126, 1983.
- [Cox92] D. R. Cox. “Causality; some statistical aspects.”, 1992. To appear in J. Roy. Statist. Soc. Ser. A.
- [DP90] R. Dechter and J. Pearl. “Directional Constraint Networks: A Relational Framework for Causal Modeling.” Technical Report R-153, UCLA Cognitive Systems Laboratory, 1990.
- [ES83] E. Eells and E. Sober. “Probabilistic Causality.” *Philosophy of Science*, **50**:35 – 57, 1983.
- [Fag77] Ronald Fagin. “Multivalued Dependencies and a New Form for Relational Databases.” *ACM Transactions on Database Systems*, **3**:262 – 278, 1977.
- [FG86] K. D. Forbus and D. Gentner. “Causal Reasoning about Quantities.” *Proceedings Cognitive Science Society*, pp. 196 – 207, 1986.
- [Fis53] R. A. Fisher. *Design of Experiments*. Oliver and Boyd, London, 1953.

- [Fre87] D. Freedman. “As Others See Us: A Case Study in Path Analysis (with discussion).” *Journal of Educational Statistics*, **12**:101 – 223, 1987.
- [Fry90] M. Frydenberg. “The Chain Graph Markov Property.” *Scand. J. Statist.*, **17**:333 – 353, 1990.
- [Gar88] P. Gardenfors. “Causation and the Dynamics of Belief.” In W.L. Harper and B. Skyrms, editors, *Causation in Decision, Belief Change and Statistics II*, pp. 85 – 104. Kluwer Academic Publishers, 1988.
- [Gei90] D. Geiger. *Graphoids – A Qualitative Framework for Probabilistic Inference*. PhD thesis, UCLA, 1990.
- [Goo83] I. J. Good. “A causal calculus.” *British Journal for Philosophy of Science*, **11** and **12** and **13**:305 – 328 and 43 – 51 and 88, 1983. reprinted as Ch. 21 in Good Thinking University of Minnesota Press, Minneapolis, MN.
- [GPP90] D. Geiger, A. Paz, and J. Pearl. “Learning Causal Trees from Dependence Information.” In *Proceedings, AAAI-90*, pp. 770 – 776, Boston, MA, 1990.
- [Gra88] C. W. J. Granger. “Causality Testing in a Decision Science.” In W.L. Harper and B. Skyrms, editors, *Causation in Decision, Belief Change and Statistics I*, pp. 1 – 20. Kluwer Academic Publishers, 1988.
- [GSS87] C. Glymour, R. Scheines, P. Spirtes, and K. Kelly. *Discovering Causal Structure*. Academic Press, New York, 1987.
- [GVP90] D. Geiger, T. S. Verma, and Judea Pearl. “Identifying Independence in Bayesian Networks.” *Networks*, **20**:507 – 534, 1990.
- [Hol86] Paul Holland. “Statistics and Causal Inference.” *Journal of the American Statistical Association*, **81**:945 – 960, 1986.
- [IS86] Y. Iwasaki and H. A. Simon. “Causality in Device Behavior.” *Artificial Intelligence*, **29**(1):3 – 32, 1986.

- [KB86] J. de Kleer and J. S. Brown. “Theories of causal ordering.” *Artificial Intelligence*, **29**(1):33 – 62, 1986.
- [Ken79] D. A. Kenny. *Correlation and Causality*. Wiley, New York, 1979.
- [Key39] J. M. Keynes. “Professor Tinbergen’s Method.” *Economic Journal*, **49**:560, 1939.
- [KSC84] H. Kiiveri, T. P. Speed, and J. B. Carlin. “Recursive Causal Models.” *Journal of Australian Mathematical Society*, **36**:30 – 52, 1984.
- [LDL90] S. L. Lauritzen, A. P. Dawid, B. Larsen, and H. G. Leimer. “Independence properties of directed Markov fields.” *Networks*, **20**:491–505, 1990.
- [Lin83] R. Ling. “Review of “Correlation and Causation” by D. Kenny.” *Journal of the American Statistical Association*, pp. 489 – 491, 1983.
- [Mit82] T. M. Mitchell. “Generalization as search.” *Artificial Intelligence*, **18**:203 – 226, 1982.
- [Nil22] H. E. Niles. “Correlation, causation, and Wright theory of “path coefficients”.” *Genetics*, **7**:258 – 273, 1922.
- [Pea88a] J. Pearl. “Embracing causality in formal reasoning.” *Artificial Intelligence*, **35**(2):259 – 71, 1988.
- [Pea88b] J. Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan-Kaufman, San Mateo, CA, 1988.
- [Pea90] J. Pearl. “Probabilistic and Qualitative Abduction.” In *Proceedings of AAAI Spring Symposium on Abduction*, pp. 155 – 158, Stanford, March 1990.
- [PGV89] J. Pearl, D. Geiger, and T. S. Verma. “The Logic of Influence Diagrams.” In R. M. Oliver and J. Q. Smith, editors, *Influence Diagrams, Belief Networks and Decision Analysis*, pp. 67 – 87. John Wiley and Sons, Ltd., Sussex, England, 1989.

- [PP80] Stott Parker and Kamran Parsay. “Inferences Involving Embedded Multivalued Dependencies and Transitive Dependencies.” In *Proceedings International Conference on Management of Data (ACM-SIGMOD)*, pp. 52 – 57, 1980.
- [PP86] Judea Pearl and Azaria Paz. “GRAPHOIDS: A Graph-based Logic for Reasoning about Relevance Relations.” In B. Du Boulay et al., editor, *Advances in Artificial Intelligence-II*, pp. 357–363. North Holland, Amsterdam, 1986.
- [Rei56] H. Reichenbach. *The Direction of Time*. University of California Press, Berkeley, 1956.
- [Rub89] H. Rubin. “Discussion of “The Logic of Influence Diagrams” by Pearl at al.” In R. M. Oliver and J. Q. Smith, editors, *Influence Diagrams, Belief Networks and Decision Analysis*, pp. 83 – 85. John Wiley and Sons, Ltd., Sussex, England, 1989.
- [Sal84] W. Salmon. *Scientific explanation and the causal structure of the world*. Princeton University Press., Princeton, 1984.
- [SG91] P. Spirtes and C. Glymour. “An Algorithm for Fast Recovery of Sparse Causal Graphs.” *Social Science Computer Review*, 9(1):62 – 72, 1991.
- [SGS89] P. Spirtes, C. Glymour, and R. Scheines. “Causality from probability.” Technical Report CMU-LCL-89-4, Department of Philosophy Carnegie-Mellon University, 1989.
- [Sho88] Y. Shoham. *Reasoning About Change*. MIT Press, Boston, MA, 1988.
- [Sim54] H. Simon. “Spurious correlations: A causal interpretation.” *Journal American Statistical Association*, 49:469 – 492, 1954.
- [Sky86] B. Skyrms. *Causal Necessity*. Yale University Press, New Haven, CT, 1986.
- [Spo83] W. Spohn. “Deterministic and Probabilistic Reasons and Causes.” *Erkenntnis*, 19:371 – 396, 1983.

- [SS86] Prakash P. Shenoy and Glen Shafer. “Propagating Belief Functions with Local Computations.” *IEEE Expert*, **1**(3):43–52, 1986.
- [Sup70] P. Suppes. *A Probabilistic Theory of Causation*. North Holland, Amsterdam, 1970.
- [SZ81] P. Suppes and M. Zaniotti. “When are Probabilistic Explanations Possible?” *Synthese*, **48**:191 – 199, 1981.
- [TY84] R. E. Tarjan and M. Yannakakis. “Simple Linear-Time Algorithms to Test Chordality of Grams, Test Acyclicity of Hypergraphs, and Selectively Reduce Acyclic Hypergraphs.” *SIAM Journal on Computing*, **13**(3):566 – 579, 1984.
- [Ver86] T.S. Verma. “Causal Networks: Semantics and Expressiveness.” Technical Report R-65, UCLA Cognitive Systems Laboratory, 1986. Also in: R. Shachter, T.S. Levitt and L.N. Kanal, editors, *Uncertainty in AI 4*, pages 325-359, Elsevier Science Publishers 1989.
- [Ver92] T.S. Verma. “A Linear-Time Algorithm for Finding a Consistent Extension of a Partially Oriented Graph.” Technical Report R-180, UCLA Cognitive Systems Laboratory, 1992.
- [VP90] T. S. Verma and J. Pearl. “Equivalence and Synthesis of Causal Models.” In *Proceedings 6th Conference on Uncertainty in AI*, pp. 220 – 227, 1990.
- [Wer91] N. Wermuth. “On Block-Recursive Linear Regression Equations.” Technical Report ISSN 0177-0098, Psychological Institute, University of Mainz, Mainz, FRG, September, 1991. Forthcoming in the Brazilian Journal of Probability and Statistics.
- [Wri21] S. Wright. “Correlation and Causation.” *J. Agricult. Res.*, **20**:557 – 585, 1921.