Graphoids:
A Qualitative Framework for Probabilistic Inference

Dan Geiger

January, 1990

Technical Report R-142
Cognitive Systems Laboratory
Department of Computer Science
UCLA
Los Angeles, CA 90024

This report reproduces a dissertation submitted to UCLA in partial satisfaction of the requirements for the degree Doctor of Philosophy in Computer Science. This work was partially supported by the National Science Foundation Grant #IRI-8821444, and the Navy Research Laboratory Grant #N00014-89-J-2007.

Acknowledgements

I owe a great debt to Judea Pearl for his excellent guidance. Judea Pearl's knowledge and scholarship and enthusiasm for the subject matter of this dissertation have greatly affected this writing as well as my scientific interests. I also owe a debt to Tom Verma and Azaria Paz, with whom I had many discussions whose impact can be traced throughout this dissertation. I thank David Heckerman for his contribution to the subject of separable graphoids and for the lengthy discussions that we had had. I have benefited from suggestions of Norman Dalkey, Eli Gafni, Hector Geffner, Thomas Ferguson, Itay Meiri, Yiannis Moschovakis, Stott Parker and Amir Weinstain. The comments of Fahiem Bacchus, Ronald Fagin, Ross Shachter and Glenn Shafer on earlier partial drafts of this dissertation considerably improved its presentation. This dissertation was illustrated and proofed by Gina George. Last but not least, I would like to thank my family; my wife Ora, who patiently supported me while developing her own career and caring for our children, and my parents, Yuda and Ruth, who initiated, encouraged and supported my education.

Table of Contents

page

1 The Graphoid Framework 1
1.1 Introduction 1
1.2 Probabilistic Expert Systems 4
1.3 Dependency Models 9
1.4 Probabilistic Networks 15
1.5 The Main Contributions 21

2 A Study of Three Independence Relations 24
2.1 Introduction 24
2.2 Preliminary Definitions 25
2.3 Armstrong Relations 30
2.4 Saturated Independence 36
2.5 Marginal Independence 41
2.6 Summary and Conclusion 47

3 The Logic of Probabilistic Networks 50
3.1 Introduction 50
3.2 Networks with Deterministic Nodes 53
3.3 A Linear Time Algorithm for Identifying Independence 68
3.4 Consistency of Probabilistic Networks 75
3.5 Discussion 77

4 Discovering Causality from Statistical Data 79
4.1 Introduction 79
4.2 Reconstructing Singly Connected Causal Networks 82
4.3 Proofs 86
4.4 Discussion 89

5 An Approach to Knowledge Acquisition 91
5.1 Introduction 91
5.2 Separable Graphoids 93
5.3 Separable Distributions and Refined Dependency Models 97
5.4 Discussion 102

6 References 103

Abstract

Graphoids: A Qualitative Framework for Probabilistic Inference

Dan Geiger

This dissertation investigates properties of conditional independence in relation to the elicitation, organization and inference of probabilistic expert systems. Qualitative notions of interaction, connectedness, mediation and causation are given formal probabilistic underpinning: graph-based representations and algorithms are developed for processing these notions.

A partial axiomatic characterization is established of the predicate $I(X, Z, Y)$ to read: "X is conditionally independent of Y, given Z". This characterization facilitates both a graphical representation of dependence information and a solution to the implication problem, of deciding whether an arbitrary independence statement $I(X, Z, Y)$ logically follows from a given set Σ of such statements. The solution of the implication problem is the key for identifying what information is unnecessary for performing a given computation. An algorithm is developed that identifies this information in probabilistic networks. The algorithm's correctness and optimality stems from the soundness and completeness of probabilistic networks with respect to probability theory. An enhanced version of the algorithm extends its applicability to networks that encode functional dependencies. Probabilistic dependence is also used to formalize the notion of interactions among variables; a class of distributions is identified for which this formal definition exhibits qualitative properties normally attributed to the word ''interact''. Finally, the problem is addressed of deciding whether a given distribution can be represented as a graph of certain structure. Conditions are identified for the existence of a unique solution, an efficient algorithm is developed to find this solution, and a relationship to the problem of discovering causality from statistical data is discussed.

CHAPTER 1
The Graphoid Framework

Generally people are bad probability assessors and even worse probability manipulators, yet they manage successfully to control environments full of uncertainties. A computer system attempting to emulate human behavior in such environments must therefore be guided by principles that permit qualitative reasoning about uncertainty. This chapter singles out conditional independence and its graph-based representations as the most fundamental relationships in such reasoning, reviews previous work on these concepts, and outlines the contribution of this dissertation.

1.1 Introduction

The volume of information needed in typical inference tasks, such as mineral exploration, weather prediction and medical diagnosis, is so high that reasoning would become unmanageable without making many assumptions of independence. Not surprisingly, the design of every system that emulates expert's behavior in these and other domains relies heavily on such assumptions (e.g., [1, 5, 7, 10, 16, 31, 43]). It is therefore vital to set forth these assumptions and to provide a means of testing whether they are suitable. The practical significance of conditional independence is reflected in three processes that are supported by expert systems: encoding the experience of an expert (elicitation), drawing conclusions (inference) and, communicating the system's recommendations to the user (explanation). In eliciting probabilistic models from human experts, qualitative dependencies among variables can often be asserted with confidence, while numerical assessments are subject to a great deal of hesitancy. For example, an expert may willingly state that cancer is related to both smoking habits and asbestos exposure, however, he would not provide a numeric quantification of these relationships unless he has rich experience with cancer patients or is aware of a reliable statistical survey that estimates the strength of these relationships. Developing a direct representation scheme for judgments about dependencies, which is a major theme of this dissertation, facilitates a qualitative organization of knowledge in a manner that is amenable to a human expert and guards the model builder from assigning numerical values that lead to conceptually unrealistic dependencies.

Knowledge about independence saves space when storing distribution functions and saves time when computing and updating the probability of an event; if we ignore independencies, then representing a discrete distribution function would require exponential size tables and calculating $P(x \text{ is } \mathit{true})$ would require a lengthy summation over the other variables in the table. Recognizing the independencies among the variables enables us to encode the table with fewer parameters and to considerably reduce the computations. Furthermore, if we choose to represent and process random variables by probabilistic networks, and we will argue that this is a plausible choice, then the topology of such networks, as well as the set of transformations that we are permitted to apply to them are determined by the rules that govern conditional independence.

Finally, a qualitative characterization of conditional independence in terms of logical axioms that do not refer to numerical quantities highlights plausible lines of reasoning that would otherwise be hidden in numerical calculations. Such axioms could serve as building blocks of systems that provide qualitative explanations as to why certain facts were or were not taken into account in a given computation. For example, the axiom (1.5d) below can be phrased to read: ''if two items together are judged to be unnecessary for a computation, then learning one of them leaves the other still unnecessary''. By contrast, a numeric representation of this argument, would involve complicated equations that hide the intuition behind it. Thus a logical characterization is preferable, and is pursued in Chapter 2.

Whereas conditions of total independence are rarely encountered in the real world (due to the existence of weak genuine dependencies), independence assumptions nevertheless approximate reality extremely well, and allow us to draw meaningful conclusions in reasonable time. For example, Markov models that assert the past is irrelevant for the future conditioned on the present state of the world, constitute a very powerful model of temporal data analysis. Structural equations which embody a rich set of independence assumptions, are the cornerstone of modeling and analysis in the behavioral sciences. A salient characteristic of these modeling methodologies is that they deploy a two phase strategy in which qualitative modeling is followed by numeric analysis. Similar separation between the two phases is also useful in modeling human experts; medical diagnosis systems, for example, are rendered much more reliable if qualitative relationships are first carefully structured [32]. For example, a more accurate model of the relationship between cancer and smoking habits would reveal some genuine independencies because there are many types of cancer only some of which, such as lung cancer, are actually dependent on one's smoking habits, while the rest are independent. This type of independence, called *subset independence,* was used by Heckerman [32] both to improve the accuracy of his medical diagnostic system as well as to speed up the process of knowledge elicitation. His system is based on networks called *probabilistic similarity networks,* which is another example of a model based on realistic assumptions of conditional independence.

A common feature of Markov models, structural equations and similarity networks is that they all express assumptions of independence in a graph-based formalism. This is not surprising. The terms ''dependence'' and ''connectedness'' are often related in our language; idioms such as ''threads of reasoning'', ''lines of reasoning'' and ''connected ideas'' suggest that an expert would conveniently express dependencies between variables by visualizing trails in some graph-based representation. However, formalizing a precise relationship between links in a graph and dependence between variables is not as easy as it seems; when we deal with informational dependencies it is often hard to distinguish between direct dependencies and indirect dependencies. Even worst, not every trail in a graph should necessarily represent a dependency; for example the three variables ''asbestos exposure'' (a), ''smoking habits'' (s) and ''cancer'' (c) are best represented by a graph of the form $a \rightarrow c \leftarrow s$. The dependence of cancer in the other two variables is represented by two directed links. However, ''smoking'' and ''asbestos exposure'' which seem connected, can reasonably be assumed independent. On the other hand, when a patient is known to have cancer, s and a become dependent variables; evidence that the patient is a heavy smoker decreases our belief that the patient has been exposed to asbestos because smoking is an alternative explanation to cancer. Thus, the chain between a, s and c represents the fact that a and s are independent, but become dependent given that we know the patient suffers from cancer.

Another intuitive interpretation of the direct links emanating from ''asbestos exposure'' and ''smoking habits'' into ''cancer'' is that the former two variables represent *causes* of cancer. The relationships between probabilistic dependencies (or correlations) and causation has long been debated. Clearly, a dependence found between two variables is not sufficient for us to assert that one causes the other, however, causal relationships are accompanied with special patterns of dependence. The basic pattern is that two independent causes become dependent once their common effect is known, as is well demonstrated by the cancer example. This observation is the basis for our attempt to formulate a procedure that recovers causal relationships from information about dependencies. Although this task is short of being complete, the association between dependence information, causality and graphical representations provides us with plausible conditions that must be satisfied if one ever aspires to deduce causal relationships from statistical data (see Chapter 4).

Causation is not the only high level concept that is associated with conditional independence. Functional dependencies, interaction between variables and the concept of information relevance are each characterized by distinct patterns of conditional independence. A variable that is function of a set variables is conditionally independent from any variable in the system, given that set. Two variables display an *interaction* between them only if they are dependent in some context. If two variables are independent, information gathered on one will be *irrelevant* for learning anything about the other. This observation is the most immediate reason to concentrate on the notion of dependence; any system, and in particular systems that operate under uncertainty, must be able to distinguish between facts that are relevant and those that are not, because otherwise the system would spend precious time in processing facts that have no bearing to the task at hand.

1.2 Probabilistic Expert Systems

The first step in constructing an expert system based on probabilistic networks is to identify the variables of interest, their relationships, and a topology that reflects these relationships. This is best illustrated by a simple example borrowed from Pearl [49], originally by Cooper [10]:

> Metastatic cancer (a_1) is a possible cause of a brain tumor (a_3) and is also an explanation for increased total serum calcium (a_2). In turn, either of these could explain a patient falling into a coma (a_4). Severe headache (a_5) is also possibly associated with a brain tumor.

In constructing the network, the expert associates a link between a perceived cause into its direct consequences. The resulting network for this example is given in Figure 1.1. Validation of the graph topology can be performed by asking the expert questions regarding the independencies that are represented in the network. For example, we may ask: If a patient is known to suffer from a Brain tumor, would his complaining of severe headaches change your belief about the possibility that he will fall into a Coma? To comply with the network's topology (Figure 1.1), the expert's answer is expected to be negative because the network shows that the node corresponding to ''Brain tumor'' blocks all trails between ''Severe headaches'' and ''Coma'', thus asserting that the latter two variables are conditionally independent given the former.

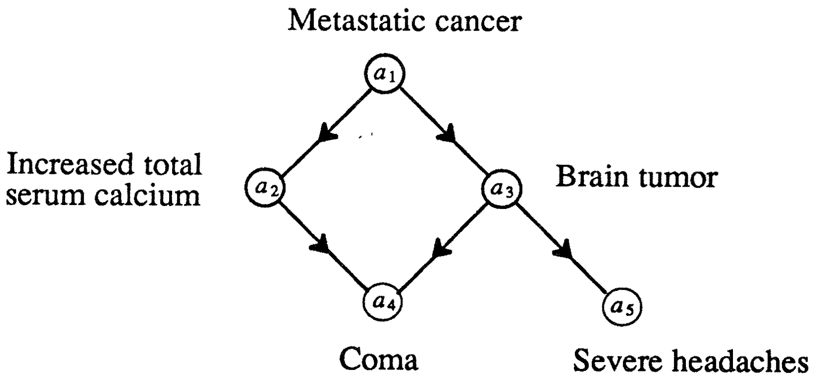


Figure 1.1

The next step is to let the expert estimate for each node a in the network a conditional distribution $P(a \mid \pi(a))$ of its values given any combination of its parents' values $\pi(a)$. The outcome is a *Bayesian network,* which represents a distribution over all possible values of the variables in the system. The distribution decomposes into:

$$P(a_1, \cdots, a_n) = \prod_{i=1}^{n} P(a_i \mid \pi(a_i)) ; \tag{1.1}$$

where $P(a_i \mid \pi(a_i)) = P(a_i)$ if a_i has no parents. This product form reflects the independencies encoded in the topology of the network given by the expert. In the Metastatic cancer example, the following parameters could be elicited from an expert:

$P(a_1)$:	$P(+a_1) = .20$	
$P(a_2 \mid a_1)$:	$P(+a_2 \mid +a_1) = .80$	$P(+a_2 \mid -a_1) = .20$
$P(a_3 \mid a_1)$:	$P(+a_3 \mid +a_1) = .20$	$P(+a_3 \mid -a_1) = .05$
$P(a_4 \mid a_2, a_3)$:	$P(+a_4 \mid +a_2, +a_3) = .80$	$P(+a_4 \mid -a_2, +a_3) = .80$
	$P(+a_4 \mid +a_2, -a_3) = .80$	$P(+a_4 \mid -a_2, -a_3) = .05$
$P(a_5 \mid a_3)$:	$P(+a_5 \mid +a_3) = .80$	$P(+a_5 \mid -a_3) = .60$

where $+a_i$ and $-a_i$ are the positive and negative outcomes of a_i, respectively. It has been noticed that this two-phase strategy, where first a qualitative model is constructed and only then parameters are elicited considerably improves the reliability of the system.

A standard query for a Bayesian network is to find the current belief distribution of a hypothesis x, given a composite evidence set $Y = Y$ i.e., to compute $P(x \mid Y = Y)$. For example, we might want to compute the probability of a patient suffering of Metastatic cancer, given that he complains on severe headaches and given that his level of serum calcium has increased. The answer to such queries can, in principle, be computed directly from Eq. (1.1) because this equation defines a full probability distribution over all variables. However, unless one exploits the independence relationships encoded in the network, this can be very inefficient both in time and space requirements. Efficient algorithms have been developed that do rely on these independencies [33, 41, 48, 55]. Two of these algorithms are demonstrated below. It should be emphasized, however, that our toy example serves only as an illustration of the way probability assessments are incorporated into the Bayesian network formulation and how these networks are used for inference. For a description of a real system consult [1].

Consider the query: ''What is the probability that a patient suffering a metastatic cancer and a brain tumor will have severe headaches''(i.e., $P(+a_5 \mid +a_1, +a_3)$) ? The answer can immediately be obtained from the network of Figure 1.1. We are interested in the quantity $P(+a_5 \mid +a_1, +a_3)$, however, in this case the answer is simply the entry $P(+a_5 \mid +a_3)$ of the conditional distribution associated with node a_5. Note that the fact that the patient suffers from metastatic cancer ($+a_1$) plays no role in this computation because the network (and the expert) asserts that this information is irrelevant for this computation; severe headaches are caused directly by a brain tumor and this is the only mechanism that associates headaches with metastatic cancer. As a less trivial example, consider the query: ''What is the probability of having Brain tumor given that the level of serum calcium has increased ?'' ($P(+a_3 \mid +a_2)$). This query cannot be answered by observation but requires an inference algorithm. We shall next describe a number of algorithms for this purpose.

Shachter has developed an algorithm based on two transformations of a network: *node-removal* and *arc-reversal* [55]. To compute $P(x \mid Y)$, these transformations change the network until the parents of node x are all in Y. First, the algorithm removes nodes that have no descendants in $\{x\} \cup Y$. Then, the algorithm picks a node a–parent of x that is not in Y, reorient all a's adjacent links into a using the transformation of ''arc reversal'' and then, when a has no children, it is removed using the ''node-removal'' transformation. In each step, new parameters of the transformed network are computed. This process is repeated until only nodes in Y are parents of x, in which case the distribution $P(x \mid Y)$ has been computed. The two transformations are listed below:

Node removal: A node that has no children is removed.

Arc reversal: Let D be a network, $a \rightarrow b$ be a link in D, S_a be the parents of a, and S_b be the parents of b. In the transformed network, the link between a and b is reversed and a link is added between any node in $S_a \cup S_b$ into nodes a and b (if such a link was missing in the original network).

To compute $P(+a_3 \mid +a_2)$, this algorithm, first removes nodes which have no children in $\{a_2, a_3\}$. Thus, node a_4 and a_5 are removed and the chain $a_2 \leftarrow a_1 \rightarrow a_3$ is left. The correctness of this step stems from the fact that the product $P(a_1)\,P(a_3 \mid a_1)\,P(a_2 \mid a_1)$, represented by this chain, results from summing Eq. (1.1) over the variables a_4 and a_5.

The termination condition has not been met yet since the parent set $\{a_1\}$ of a_3 is not a subset of $\{a_2\}$. The algorithm now removes a parent of a_3. The only parent is a_1. First, link $a_1 \rightarrow a_2$ is reversed and the resulting network becomes $a_2 \rightarrow a_1 \rightarrow a_3$. Notice that this transformation is the graphical equivalent of Bayes rule. The two parameters $\{P(a_1), P(a_2 \mid a_1)\}$ are replaced with $\{P(a_2), P(a_1 \mid a_2)\}$ using Bayes' formula:

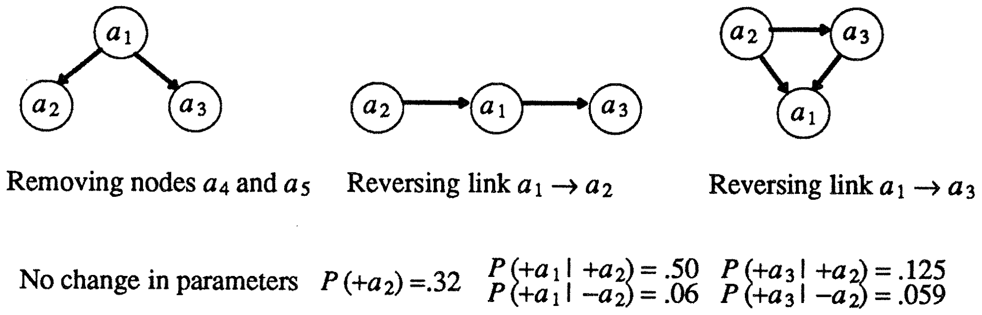
$$P(a_1 \mid a_2) = \frac{P(a_2 \mid a_1) P(a_1)}{P(a_2)}$$

where

$$P(a_2) = \sum_{a_1} P(a_2 \mid a_1)\, P(a_1).$$

The last two transformations reverse the link $a_1 \rightarrow a_3$ and remove node a_1. The first of these adds a link from a_2 to a_3 while the latter removes the node a_1 from the network, leaving a single link $a_2 \rightarrow a_3$. The distribution $P(a_3 \mid a_2)$ is computed using Bayes' rule and the entry $P(+a_3 \mid +a_2)$ of this distribution is the desired outcome.

Figure 1.2 summarizes the transformations and the corresponding changes in the parameters. At most two conditional distributions are changed by each transformation.



Removing nodes a_4 and a_5 Reversing link $a_1 \rightarrow a_2$ Reversing link $a_1 \rightarrow a_3$

No change in parameters $P(+a_2) = .32$ $P(+a_1 \mid +a_2) = .50$ $P(+a_1 \mid -a_2) = .06$ $P(+a_3 \mid +a_2) = .125$ $P(+a_3 \mid -a_2) = .059$

Figure 1.2

Note that in the last step, $P(a_1 \mid a_2, a_3)$ need not be computed because node a_1 is removed in the next step. This is always the case; when the last link is reversed into a node that is to be removed, then that node's conditional distribution need not be computed. The parameters of each transformed network in this sequence are computed in this example directly from the distribution that is defined by the previous network via Eq. (1.1). However, a simple closed form formula has been developed by Olmsted [44] and Shachter [55] which provides an efficient method for calculating these parameters. Interestingly, the only requirement for correctness of this algorithm is that in computing $P(a \mid Y)$, the distribution $P(a, Y)$ represented by the original network must remain unchanged by the transformations. This requirement is met because the transformed network never introduces any new independence assertions.

This algorithm has several disadvantages; it computes in each step the distribution function over all instances of Y, although we may be interested in computing the distribution only for one specific instance Y of Y. Its complexity depends on the order by which links are reversed, it is not aimed towards producing explanations as to how the outcome is reached and it is not incremental, namely, when new information becomes available, the algorithm must start the computations from scratch. On the other hand, it is conceptually simple and it operates on arbitrary networks.

Pearl developed an algorithm for singly-connected networks, namely, networks in which every two nodes are connected with at most one trail. His algorithm is linear in the number of variables, produces explanations that are meaningful to a human observer, is easily implementable by parallel architectures and is incremental [48]. Unfortunately, it works only for singly-connected networks. Adjustments were suggested to make the algorithm applicable to general networks [49], but many of its advantages vanish, in particular, its linear complexity. This is not surprising in light of Cooper's result showing that inference in Bayesian networks is NP-hard [12]. This realization motivated Henrion to introduce the method of *stochastic relaxation* where exact calculations are abandoned in favor of answers with preselected precision [33]. Chavez and Cooper analyzed a variant of this method [8, 9]. The main problem of these type of algorithms is their convergence rate which is unbounded when the distribution represented by the network uses parameters close to zero or one.

Another important algorithm is Lauritzen and Spiegelhalter's, which initially compiles a given network to a tree representation [41]. This compilation may be time and space consuming but it is performed only once. Afterwards, every query can be answered directly from the tree representation in time exponential in the size of the largest clique. First, the algorithm adds links to the networks until the networks becomes *chordal,* namely, every undirected cycle of length at least four contains a chord. In the cancer example, this step results in adding, for example, a link between a_2 and a_3. The problem of adding minimum number of links to obtain a chordal graph is NP-complete, but efficient algorithms are known that achieve near-to-optimal behavior

[69]. The next step in this algorithm is to form clusters of nodes according to the cliques of the resulting graph G and to organize them in a structure called *join-tree*; a tree of clusters in which every two clusters sharing a variable are connected via a path through clusters that contain this variable. A join-tree is guaranteed to be formed whenever G is chordal. For example, a join-tree for the metastatic example is given in Figure 1.3 below.

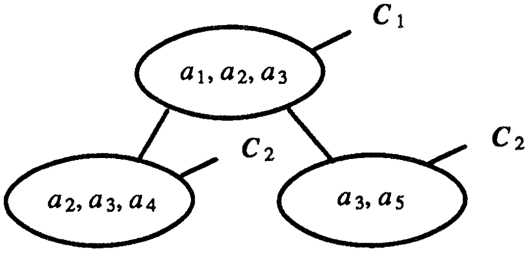


Figure 1.3

This graph asserts, for example, that a_4 and a_5 are independent given a_1, a_2 and a_3, because node C_1 blocks the path between C_2 and C_3. This independence assertion also holds in the original network of Figure 1.1. Indeed, this is always the case; the join-tree never introduces independence assertions not supported by the original network [49]. However, some independence assertions might escape explicit representation in the join-tree; these are encoded in the numeric parameters computed for the cliques. The precise formation and manipulation of these probabilistic parameters is omitted; see [41] and [49] for details.

1.3 Dependency Models

The concept of conditional independence plays a major role in probabilistic expert systems because it provides a mechanism for determining what information is unnecessary for performing a given computation. We say that two sets of variables X and Y are conditionally independent given Z, in some probability distribution P, if

$$P(X \mid Z, Y) = P(X \mid Z) \quad \text{whenever } P(Y, Z) > 0 \qquad (1.2)$$

for every instance X, Y and Z of X, Y and Z, respectively. This definition conveys the idea that once Z is known, the value of Y is irrelevant for calculating the probability of X namely, propositions represented by the set X are judged to be irrelevant to the propositions represented by the set Y, once we know Z. This definition captures our intuition about how dependencies are changed when learning new facts. For example, it permits two independent variables to become dependent upon learning a new fact, as in the ''asbestos exposure'' and ''smoking habits'' example, and it also renders dependent variables independent once we learn a fact that *mediates* between them, as in the case of two variables representing ''rain'' and ''slipping on a pavement'' which are dependent but become independent upon learning that the pavement is covered. The former type of dependence is called *induced dependence* and the later is called *mediated dependence* [49]. Thus, probabilistic conditional independence is sufficiently flexible to represent changes in dependencies and can in principle be employed for identifying which propositions are needed for a computation at any given state of knowledge.

Dependency models are formal ways for qualitatively representing such dependencies. A *dependency model* is as a truth assignment rule for the predicate $I(X, Z, Y)$, where I stands for ''X is independent of Y once Z is known''. Equivalently, M can be regarded as a list of triplets (X, Z, Y) for which $I(X, Z, Y)$ holds. Every distribution defines a dependency model through Eq. (1.2). However, a dependency model can encode the dependencies among variables without necessarily referring to probability distributions. In particular we are interested in graph-based models which determine these dependencies by tracing paths in a graph whose nodes represent the variables of interest. In dependency models based on undirected graphs, two sets of variables X and Y are said to be conditionally independent given Z if all paths between nodes corresponding to X and nodes corresponding to Y must traverse Z, i.e., if Z is a cutset separating X from Y (*). A trivial example of such a dependency model is the empty graph over n nodes, which represents a set of mutually independent variables. Another simple example is a chain representing a Markov process, say a language where the probability of the i-th letter is determined solely by the (i-1)-th letter via $P(l_i \mid l_{i-1})$. The dependencies embedded in the distribution function can be represented by the Markov chain of Figure 1.4.

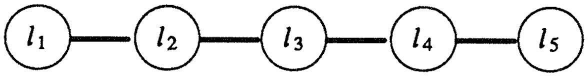


Figure 1.4

This graph asserts, for example, that variables l_1 and l_3 are conditionally independent given l_2, since node l_2 blocks all paths from l_1 to l_3. More generally, for every three disjoint sets X, Y and Z of nodes in a graph G, we define the predicate $I(X, Z, Y)_G$ by:

(*) The correspondence between vertex separation and conditional independence is the basis for Markov-fields theory [14, 37, 39].

$$I(X, Z, Y)_{\mathcal{G}} <\equiv> Z \text{ separates } X \text{ from } Y \text{ in } G.$$

We say that G represents the dependencies of P if there exists a 1-1 correspondence between the variables in P and the vertices of G such that:

$$I(X, Z, Y)_{\mathcal{G}} \Rightarrow I(X, Z, Y)_{\mathcal{P}} \tag{1.3}$$

Such a graph is called an *independence-map (I*-map) of P. When the implication in equation (1.3) is bi-directional, we say that G *perfectly* represents the dependencies of P and such a graph is called a *perfect-map* of P.

Undirected graphs have several disadvantages in representing dependencies. First, induced dependencies have no perfect representation because two nodes that represent independent variables will always remain independent when new variables are learned; paths that were blocked by a set of nodes S remain blocked when S is augmented with new nodes (representing additional pieces of information). Directed acyclic graphs, on the other hand, are well suited for representing induced dependencies. Furthermore, links in dags can be quantified compatibly with the independencies encoded in the dag in a way amenable to human experts (see Eq (1.1) and the metastatic cancer example). Undirected graphs lack such a simple quantification procedure. Despite these two disadvantages, undirected graphs are used in practice because they provide a conceptually simple representation of dependencies and because they facilitate efficient inference procedures [41]. In particular, undirected graphs are useful as an internal representation for independencies, such as join-trees in Lauritzen and Spiegelhalter's inference algorithm, while directed acyclic graphs are useful for eliciting the knowledge from an expert.

The discussion above is summarized by the definitions of dependency models, perfect maps and I-maps. These definitions were developed by Pearl and Paz [51].

Definition [51]: A *dependency model* M over a finite set of elements U is any subset of triplets (X, Z, Y) where X, Y and Z are disjoint subsets of U.

Definition [51]: Let U be a finite set of variables. Let $domain(u_i)$, $u_i \in U$, be countable sets, called the domain of u_i. A *Probabilistic Dependency Model* M_P is defined in terms of a discrete probability distribution P with a sample space $\times_{u_i \in U} domain(u_i)$. If X, Y and Z are three disjoint subsets of U, and $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{Z}$ are any instances from the domains of the variables in these subsets, then by definition $(X, Z, Y) \in M_P$ iff

$$P(X, Y, Z)\, P(Z) = P(X, Z)\, P(Y, Z). \tag{1.4}$$

A Probabilistic Dependency Model is said to be *non-extreme* if the range of P is restricted to the positive real numbers, (i.e., excluding 0's and 1's). Note that Eq. (1.4) is equivalent to equation (1.2) whenever $P(Y, Z) > 0$.

Definition [51]: An *Undirected Graph Dependency Model M_G* is defined in terms of an undirected graph G. If X, Y and Z are three disjoint subsets of nodes in G, then by definition $I(X, Z, Y)_G$ iff every path between nodes in X and Y contains at least one node in Z.

When speaking about dependency models, we use both set notations and logic notations. If $(X, Z, Y) \in M$, we say that the *independence statement* $I(X, Z, Y)_{\mathcal{P}}$ holds for M. The subscript $\mathcal{P}$ is omitted when the type of independence is not important. Similarly, we say that M contains a triplet (X, Z, Y) or that M satisfies a statement $I(X, Z, Y)$. An independence statement $I(X, Z, Y)$ is called an *independency* and its negation is called a *dependency. The* notation $P(Z)$, stands for $P(\mathbf{Z})$ for all instances $\mathbf{Z}$ of Z, or more explicitly, it stands for $P(z_1 = \mathbf{z}_1, \cdots, z_n = \mathbf{z}_n)$ where the z_i's are the variables in Z, and $\mathbf{z}_i$'s are arbitrary instances of z_i's.

Definition [51]: An *I*-map of a dependency model M is any model M' such that $M' \subseteq M$. A *perfect map* of M is any model M'' such that $M'' = M$. A graph G is a *minimal-edge I*-map if the dependency model defined by G is an *I*-map of M, and G ceases to be an *I*-map if any link is removed.

The graph of figure 1.4, for example, is an *I*-map of the distribution defined by the transition probabilities $P(l_{i+1} \mid l_i)$ of the corresponding Markov process. This graph, however, is not necessarily a perfect map because a degenerate Markov process might embody independencies that cannot be read from the topology of a chain, for example, $I(l_2, \varnothing, l_3)$. Interestingly, it has been shown that if a chain is a minimal-edge *I*-map of a Markov process, all variables are binary, and every combination of letters is possible, then the chain must be a perfect map [25].

There are two important types of dependency models that have not been mentioned so far: *Relational* and *Correlational*. A *Relational Dependency Model M_R* is defined in terms of a discrete probability distribution R. A triplet (X, Z, Y) belongs to M_R if once Z is fixed, the range of values permitted for X is not restricted by the choice of Y. A *Correlational Dependency Model M_C* is defined in terms of a collection C of random variables. A triplet (X, Z, Y) belongs to M_C if the linear estimation error of the variables in X using measurements on Z would not be reduced by adding measurements of the variables in Y, hence making Y irrelevant to the estimation of X. Precise definitions of these independence relations is given in Chapter 2.

These three types of independence: probabilistic, relational and correlational provide different formalisms for the notion of irrelevance, each capturing different aspects of the word ``irrelevant''. The similarity between these models is summarized axiomatically by the following definition of graphoids.

Definition: [51] A *graphoid* is any dependency model M which is closed under the following inference rules, considered as *axioms:* [(*)]

- Trivial Independence:
  $$I(X, Z, \varnothing) \tag{1.5a}$$
- Symmetry:
  $$I(X, Z, Y) \Rightarrow I(Y, Z, X) \tag{1.5b}$$
- Decomposition:
  $$I(X, Z, Y \cup W) \Rightarrow I(X, Z, Y) \tag{1.5c}$$
- Weak union:
  $$I(X, Z, Y \cup W) \Rightarrow I(X, Z \cup W, Y) \tag{1.5d}$$
- Contraction:
  $$I(X, Z, Y) \;\&\; I(X, Z \cup Y, W) \Rightarrow I(X, Z, Y \cup W) \tag{1.5e}$$

Intuitively, the essence of these axioms lies in Eqs. (1.5d) and (1.5e) asserting that when we learn an irrelevant fact, all relevance relationships among other variables in the system should remain unaltered; any information that was relevant remains relevant and that which was irrelevant remains irrelevant. These axioms, are very similar to those assembled by Dawid [15] for probabilistic conditional independence, those proposed by Smith [61] for *Generalized Conditional Independence* and those used by Spohn [64] in his exploration of *causal independence*. We shall henceforth call axioms (1.5a) through (1.5e) *graphoid axioms*. It can readily be shown that all the specialized classes of dependency models presented thus far are graphoids, and in view of this generality, these axioms are selected to represent the notion of mediated dependence between items of information [49]. A proof that the graphoid axioms hold for conditional independence when U, the set of variables, is countable and the cardinality of the domain of each variable is unrestricted, can be found in [64]. Our reason for avoiding infinite number of variables and non-countable domains is the technical burden imposed by removing these restrictions which will severely distract our discussion while adding no insight.

[(*)] This definition differs slightly from that given in Pearl and Paz [51] where axioms (1.5b) through (1.5e) define semi-graphoid and dependency models obeying also (1.6) are called graphoids. Axiom (1.5a) is added for future clarity.

Non-extreme probabilistic dependency models enjoy additional properties that do not hold for arbitrary probabilistic models. Two such properties are listed below:

- Intersection:

$$I(X,Z \cup W,Y) \;\&\; I(X,Z \cup Y,W) \Rightarrow I(X,Z,Y \cup W) \qquad (1.6)$$

- Strong Intersection:

$$I(X,Z \cup W,Y) \;\&\; I(X,Z \cup Y=\mathbf{Y},W) \Rightarrow I(X,Z,Y \cup W) \qquad (1.7)$$

The intersection axiom is best visualized using the undirected graph interpretation. If $Z \cup W$ is a set of nodes that shield X from Y and $Z \cup Y$ is a set of nodes that shield X from W, then their intersection Z shields X from both Y and W. Intersection holds for non-extreme distributions but does not hold for correlational or relational dependency models. A graphoid satisfying intersection is called an *intersectional graphoid*. In modeling empirical knowledge (e.g., mineral exploration and weather prediction), it is reasonable to assume that every combinations of facts has some non-zero probability of occurring, which renders the intersection axiom valid.

Strong intersection differs from the other properties of dependence models in that it refers to an independence assertion that holds only for one instance of a variable (i.e., $Y=\mathbf{Y}$), thus being *asymmetric*. A definition of *refined dependency models* is developed in Chapter 5, where reference to asymmetric independence is needed.

That strong intersection and intersection hold for non-extreme distributions is shown by the proof below:

$I(X,Z\cup W,Y)_{\mathcal{P}}$ implies $P(X \mid Z,W,Y)=P(X \mid Z,W)$,which in particular implies that $P(X \mid Z,W,Y=\mathbf{Y})=\mathrm{P(X \mid Z,W)}$. The first term of these is equal also to $P(X \mid Z,Y=\mathbf{Y})$ due to $I(X,Z\cup Y=\mathbf{Y},\mathrm{W})_{\mathcal{P}}$. Thus, $P(X \mid Z,W)=P(X \mid Z,Y=\mathbf{Y})$. $P(X \mid Z)=\sum_{W} P(X \mid Z,W)\,P(W \mid Z)$. Plugging the previous equality in the latter sum yields that $P(X \mid Z)$ equals $P(X \mid Z,Y=\mathbf{Y})$ which has been shown to equal $P(X \mid Z,W)$. In other words, $I(X,Z,W)_{\mathcal{P}}$ must hold (Eq. 1.2). This independence statement together with $I(X,Z\cup W,Y)_{\mathcal{P}}$, using contraction (1.5e), yields $I(X,Z,Y \cup W)_{\mathcal{P}}$, which is the desired consequence. □

That intersection does not hold in general can be seen from the following example; if x, y and w are three variables constrained by equality and $z=\emptyset$, then the two antecedents of the intersection axiom hold but the consequence is violated.

Note that strong intersection implies intersection but not vice versa. This is an important observation because it indicates that intersection does not summarize the difference between extreme and non-extreme probabilistic dependency models; there are many extreme probabilistic models that satisfy intersection. These distributions enjoy the computational advantage that intersection offers, namely, an efficient construction of an undirected graph representation. This construction is discussed below.

1.4 Probabilistic Networks

An important tool in representing probabilistic information is the construction of an appropriate graph representation, directed or undirected, for the dependencies in the domain. Ideally, to graphically represent all independencies of some distribution P by a graph G, we would like to require that every independence of P would belong to M_G, the dependency model defined by G, and vice versa, every triplet in M_G would represents an independence that holds in P. In other words, that G be a perfect map of P. This would provide a clear graphical representation of all variables that are conditionally independent. Unfortunately, this requirement is often too strong because there are many distributions that have no perfect map in a graphs. The spectrum of probabilistic dependencies is in fact so rich that it cannot be cast into any representation scheme that uses a polynomial amount of storage (*). Thus, the topology of a graph alone cannot always represent all the independencies and dependencies of a given distribution. Being unable to obtain a graphical representation that displays all independencies we compromise this requirement and allow some independencies to escape representation. Naturally, we seek a graph that displays only genuine independencies of P and which maximizes the number of such displayed independencies, namely, we require that G be a minimal-edge I-map of P. The resulting graph is called a *Markov network* of P.

Definition [51]: A graph G is called a *Markov network* of a dependency model M, if G is a minimal-edge I-map of M, namely, deleting any edge of G would make G cease to be an I-map of M.

The definition of a Markov network suggests a naive algorithm for constructing such a network from a given dependency model M; starting with a complete graph, where every node corresponds to a variable of M, remove any link as long as the remaining graph is an I-map. There are two difficulties with this algorithm. First, the resulting network may depend on the order by which links are removed and second,

(*) This claim is established by showing that the number of probabilistic dependency models over U is at least $O(2\exp\{2^{|U|}\})$, thus requiring, on the average, exponential amount of storage to represent an arbitrary dependency model [71]).

checking that the remaining graph is an I-map of M may require an exponential number of steps, one for each independence statement encoded in G. Both problems are resolved if M satisfies symmetry, decomposition and intersection (e.g., if M is a non-extreme probability model). These properties guarantee that the Markov network G_0 of M is unique and can be obtained by removing a link (a, b) whenever $(a, U-\{a, b\}, b) \in M$, namely, whenever a and b are independent given the rest of the variables. Thus, the order by which links are removed is immaterial. Alternatively, if M also satisfies weak union, as every probability distribution does, then G_0 can also be obtained by connecting each node a to a *minimal* set of nodes $N(a)$, such that $(a, N(a), U - N(a) - \{a\}) \in M$, namely, the neighbors of a in G_0 correspond to a minimal set of variables $N(a)$ that make a conditionally independent of the rest of the variables of M. A minimal set of variables that makes a variable a conditionally independent of all other variables is called a *boundary* of a (it is minimal if no node can be removed without destroying this property). The statements, $I(a, N(a), U-N(a)-\{a\})$, one for each variable of M, are called the *neighborhood basis of* M. Similarly, the statements $I(a, U-\{a, b\}, b)$ that hold for M are called the *pairwise basis*. The name *basis* comes from the fact that these statements are sufficient to identify a graph uniquely. The two types of bases have a common structure; they both consists of *saturated statements,* namely statements $I(X, Z, Y)$ where $X \cup Y \cup Z = U$ and U is the set of all variables [42]. In Chapter 2, we show that the graphoid axioms are *sound* and *complete* for saturated independence, namely, that these axioms characterize conditional independence if we limit ourselves to saturated statements. In particular, soundness implies that every separation statement $I(X, Z, Y)_{\mathcal{G}}$ that holds in a graph G is *derivable* by successive applications of the graphoid axioms from either of its bases.

Pearl and Paz [51] have shown that the boundary of each variable is unique whenever M satisfies symmetry, decomposition, intersection and weak union. These conditions, however, are a bit too strong. The axiom below is sufficient to guarantee unique boundaries. This axiom is implied both from symmetry, decomposition, intersection and weak union, and from symmetry, decomposition, intersection and contraction, but it is weaker than these two sets of axioms.

$$I(a, Z \cup V_1, W \cup V_2) \;\&\; I(a, Z \cup V_2, W \cup V_1) \Rightarrow I(a, Z, W \cup V_1 \cup V_2) \qquad (1.8)$$

Proof: Axiom (1.8) states that if $S_1 = Z \cup V_1$ and $S_2 = Z \cup V_2$ are two boundaries of a, then their intersection Z is also a boundary. Thus, the intersection of all boundaries is a boundary, and is therefore unique.

The construction of Markov networks is summarized by the following theorem:

Theorem 1.1 [51]: Every dependency model M satisfying symmetry (1.5b), decomposition (1.5c) and intersection (1.6) has a unique minimal I-map $G_0 = (U, E_0)$ produced by connecting those nodes a and b for which $(a, U - \{a, b\}, b) \notin M$, i.e.,

$$(a, b) \notin E_0 \leftrightarrow (a, U - \{a, b\}, b) \in M.$$

If M further satisfies weak union (1.5d) then G_0 equals to the graph produced by connecting each node a to a minimal set of nodes $N(a)$ such that $(a, N(a), U - N(a) - \{a\}) \in M$.

This local construction of an undirected graph representation can be applied to any distribution that excludes 0's and 1's but is not guaranteed for distributions that do not satisfy intersection. For example, If P is a distribution of three variables x, y and z that are constrained to be equal, then each two variables in P are conditionally independent given the third. Thus, both methods offered by Theorem 1.1 for constructing a Markov network yield an empty graph. This graph is not an I-map of P because it shows that x, y and z are independent while, in fact, they are dependent because they must be equal (a correct minimal-edge I-map would be *any* chain that connects these three variables). Nevertheless, a construction of a graphical representation for arbitrary graphoids is available when using the language of directed acyclic graphs (dags) and the *directional*-separation (d-separation) criteria.

The definition of d-separation is best motivated by regarding directed acyclic graphs as a representation of causal relationships. Designating a node for every variable and assigning a link between every cause to each of its direct consequences defines a graphical representation of a causal hierarchy. For example, the propositions ''It is raining'' (r), ''the pavement is wet'' (w) and ''John slipped on the pavement'' (s) are well represented by a three node chain, from r through w to s ; it indicates that rain and wet pavement could cause slipping, yet wet pavement is designated as the *direct cause;* rain could cause someone to slip if it wets the pavement, but not if the pavement is covered. Moreover, knowing the condition of the pavement renders ''slipping'' and ''raining'' independent, and this is represented graphically by showing node r and s separated from each other by node w. This configuration represents a mediated dependence. Furthermore, if we assume that ''broken pipe'' (b) is another direct cause for wet pavement, as in Figure 1.5, then an induced dependency exists between the two events that may cause the pavement to get wet: ''rain'' and ''broken pipe''. Although they appear connected in Figure 1.5, these propositions are marginally independent and become dependent once we learn that the pavement is wet or that someone broke his leg. An increase in our belief in either cause would decrease our belief in the other as it would ``explain away'' the observation.

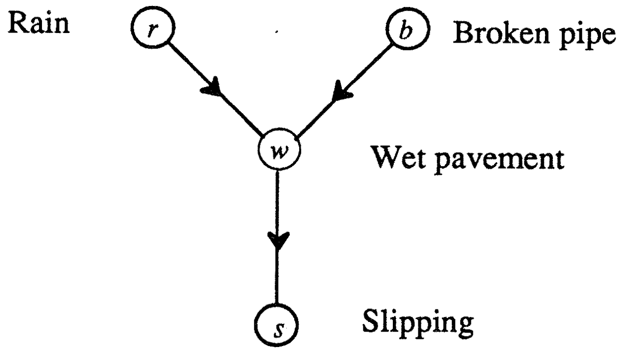


Figure 1.5

The following definition of *d*-separation permits us to graphically identify such induced dependencies from the network. A preliminary definition is needed.

Definition: A *trail* in a dag is a sequence of links that form a path in the underlying undirected graph. A trail is said to contain the nodes adjacent to its links. A node b is called a head-to-head node with respect to a trail t if there are two consecutive links $a \rightarrow b$ and $b \leftarrow c$ on t. A node that starts or ends a trail t is not a head-to-head node with respect to t.

The definitions of undirected graphs, acyclic graphs, trees, spanning trees, cliques, paths, adjacent links and nodes can be found in any text on graph algorithms (e.g., [18]).

Definition [49]: If X, Y, and Z are three disjoint subsets of nodes in a dag D, then Z is said to d-separate X from Y, denoted $I(X, Z, Y)_{\mathcal{D}}$, iff there exists no trail t between a node in X and a node in Y along which (1) every head-to-head node (wrt t) either is or has a descendent in Z and (2) every node that delivers an arrow along t is outside Z. A trail satisfying the two conditions above is said to be *active*. Otherwise, it is said to be *blocked* (by Z).

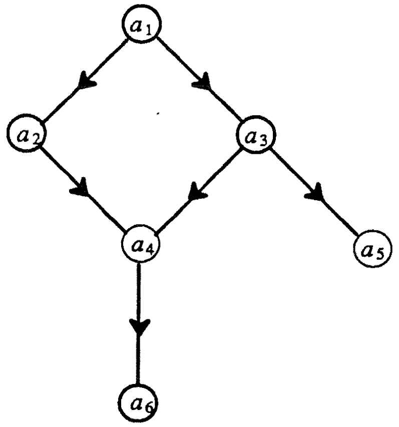


Figure 1.6

In Figure 1.6, for example, $X = \{a_2\}$ and $Y = \{a_3\}$ are d-separated by $Z = \{a_1\}$; the trail $a_2 \leftarrow a_1 \rightarrow a_3$ is blocked by $a_1 \in Z$ while the trail $a_2 \rightarrow a_4 \leftarrow a_3$ is blocked because a_4 and all its descendants are outside Z. Thus $I(a_2, a_1, a_3)_{\mathcal{D}}$ holds in D. However, X and Y are not d-separated by $Z' = \{a_1, a_6\}$ because the trail $a_2 \rightarrow a_4 \leftarrow a_3$ is rendered active: learning the value of the consequence a_6, renders its causes a_2 and a_3 dependent, like opening a pathway along the converging arrows at a_4. Consequently, $I(a_2, \{a_1, a_6\}, a_3)_{\mathcal{D}}$ does not hold in D. Note that if a dag contains no head-to-head nodes, then separation and d-separation are equivalent.

Definition: A *Dag Dependency Model* M_D is defined in terms of a directed acyclic graph D. If X, Y and Z are three disjoint sets of nodes in D, then, by definition, $(X, Z, Y) \in M_D$ iff there is no active trail by Z between nodes in X and Y.

The task of finding a dag which is a minimal-edge I-map of a given distribution P was solved in [53, 72]. The algorithm consists of the following steps: assign a total ordering d to the variables of P. For each variable a_i of P, identify a minimal set of predecessors $\pi(a_i)$ that renders a_i independent of all its other predecessors in the ordering of the first step. Assign a direct link from every variable in $\pi(a_i)$ to a_i. The resulting dag is an I-map of P, and is minimal in the sense that no edge can be deleted without destroying its I-mapness. The input L for this construction consists of n conditional independence statements, one for each variable, all of the form $I(a_i, \pi(a_i), U(a_i) - \pi(a_i))$ where $U(a_i)$ is the set of predecessors of a_i and $\pi(a_i)$ is a subset of $U(a_i)$ that renders a_i conditionally independent of all its other predecessors. This set of conditional independence statements is said to *generate* a dag and is called a *recursive basis* drawn from P.

Definition [49]: A dag D is called a *Bayesian network* of a dependency model M, if D is a minimal-edge I-map of M, namely, deleting any edge of D would make D cease to be an I-map of M.

The two theorems below summarize the discussion above.

Theorem 1.2 (soundness) [72]: If M is a graphoid, and L is any recursive basis drawn from M, then the dag generated by L is an I-map of M.

Theorem 1.3 (closure) [72]: Let D be a dag generated by a recursive basis L. Then M_D, the dependency model defined by D, is exactly the closure of L under axioms (1.5a) through (1.5e)

The significance of Theorems 1.2 and 1.3 is two fold. From a practitioner's view point, it allows us to reason about the structure of one's problem without the need to specify the model numerically. From a researcher's view point, it allows us to formalize and derive arguments about independencies by simple steps of logic deductions. Notably, the main difference between the construction of Bayesian networks and Markov networks, aside from the intersection axiom that is needed for the latter, is the stratification required for the construction of Bayesian networks. If a new variable is added to a Bayesian network, only local changes need to be incorporated, while the construction of a Markov network must start form scratch because each statement in its basis depends on the set of all variables U and when U changes, the entire graph needs to be revised.

Although the structure of the network depends strongly on the node ordering used in its construction, each network is nevertheless an I-map of the underlying distribution P. This means that all conditional independencies portrayed in the network (via d-separation) are valid in P and hence, are order independent. An immediate corollary of this observation yields an order-independent test for minimal I-mapness.

Corollary 1.4: Given a DAG D and a probability distribution P, a necessary and sufficient condition for D to be a minimal I-map (hence a Bayesian network) of P is that each variable X_i be conditionally independent of all its non-descendants, given its parents S_i, and no proper subset of S_i satisfies this condition.

The necessary part follows from the fact that every parent-set S_i d-separates X_i from all its non-descendants. The sufficient part holds because X_i's independence of all its non-descendants entails X_i's independence of its predecessors in a particular ordering d.

The construction of a Bayesian network for M requires that the parent set of each node be minimal, namely, no arc from a parent to a child a_i can be removed without violating the condition $(a_i, \pi(a_i), U(a_i) - \pi(a_i)) \in M$. Bayesian networks are not unique; first, they are very sensitive to the order of M's variables–one order may yield a complete graph while the other would produce a tree. Second, even when the order is fixed, a minimal set of variables that makes a_i independent of its predecessors in that order may not be unique. However, if M is an intersectional graphoid, or more precisely, if axiom (1.8) holds, then, once the order is fixed, the resulting dag is unique.

In Chapter 3, the dag representation scheme is extended to include a representation of *deterministic variables,* namely, variables that are *functions* of the variables corresponding to their parents [55]. In this scheme, more independencies are recorded in the dag, in particular, given its parents values, a deterministic node is conditionally independent of its descendants as well as of its non-descendants.

Other types of probabilistic networks are discussed in [38, 73]. Review of the use of probabilistic networks and decision theory in expert systems can be found in [11, 34].

1.5 The Main Contributions

The research reported in this dissertation establishes a suitable theoretical framework for expert systems founded on probability theory. Previous approaches to expert systems construction were mostly ad hoc. These methods, most notably the certainty factor paradigm upon which MYCIN [58] is based, were invented to overcome the so-called ''impracticality'' of probability theory, which if naively applied, seem to require vast amount of data and insurmountable amounts of computations. The use of certainty factors, although computationally very efficient, turned out to produce conceptually unacceptable conclusions even for very simple diagnostic problems. When certainty factors were found to work correctly, the computations were equivalent to those performed in a Bayesian network of a tree topology [30]. This topology is very restrictive, implying assumptions of independence, that are unlikely to be met in practice. Consequently, it has been realized that the key to constructing practical expert systems is to strike a practical balance between computational tractability and semantical-adequacy [49]. Conditional independence plays a major role in achieving such balance.

This dissertation investigates properties of conditional independence in relation to the elicitation, organization and inference of probabilistic expert systems. Qualitative notions of interaction, connectedness, mediation and causation are given formal probabilistic underpinning, and graph-based representations and algorithms are developed for processing these notions. Its main contributions are the formal characterization of conditional independence, the development of algorithms that derive such dependencies and identify them in graphical representations, the development of an algorithm that recovers the topology of a Bayesian network from statistical data, and, finally, a formalization of the concept of *interaction* among variables; a tool for organizing probabilistic assessments elicited from a human expert.

Chapter 2 develops a partial axiomatic characterization of conditional independence; the graphoid axioms are shown to be complete for special types of independence such as *marginal independence* and *saturated independence*. Conditional independence is shown to be a relation that is completely characterized by Horn type axioms. This property facilitates a proof that any undirected graph and any dag is suitable for perfectly representing the dependencies embedded in a probability distribution.

Chapter 3 investigates the relationships between conditional independence and its graph-based representations. In particular, the graphoid axioms are shown to be powerful enough to fully characterize the independencies that logically follow from the topology of a network. A new graphical criteria, *D*-separation, is introduced; it allows us to detect a maximal set of independencies that are encoded in a Bayesian network for which some variables are known to be functions of their parents' variables.

Chapter 4 associates the directionality of a link $a \rightarrow b$ in a Bayesian network with the sentence ''a is a cause of b'', an association that is implicit in the definition of d-separation. This association is potentially justified provided that the direction of a link is not sensitive to the specific order chosen to construct the network. Conditions are provided under which the directionality of some links is uniquely recoverable; an essential prerequisite for the recovery of causal relationships from statistical data. An efficient algorithm is developed that recovers these links whenever possible.

Chapter 5 concentrates on the problem of organizing probabilistic assessments. Two variables are said to *interact* if there exists a context in which they are dependent. The notion of *interaction* is useful for partitioning a set of variables to clusters that reflect independent subdomains. It is shown that for a large class of distributions, called *separable,* interaction induces a partition that coincides with the connected components of the corresponding Bayesian network and argues that it is a plausible choice to use separable distributions for the construction of these networks. Normal and strictly-positive binary are example of separable distributions.

Aside of the contributions of each individual chapter, and perhaps even more important, this dissertation provides a qualitative framework for research in a field that has been mostly governed by numerical techniques. The axiomatic approach developed in the first three chapters resting on the graphoid axioms, helps focus attention on the structure of probability arguments and on the organization of probabilistic expert systems. Chapter 4 and 5 present the first fruits of this approach; results which would be almost impossible to obtain without the axiomatic approach.

CHAPTER 2
A Study of Three Independence Relations

Three independence relations: correlational, relational and probabilistic are examined. The traditional numeric exploration of these concepts is abandoned in favor of a logical characterization. It is shown that the graphoid axioms are powerful enough to characterize meaningful subsets of these three independence relations. Furthermore, it is shown that Horn-axioms are sufficient for describing probabilistic and relational independence, but are not sufficient for correlational independence, which requires disjunctive axioms.

2.1 Introduction

The traditional definition of conditional independence suggests that in order to verify whether X is independent of Y, given Z, one must possess at hand a distribution $P(X, Y, Z)$ and test whether this distribution satisfies a set of equalities (Eq. 1.4). This definition stands in sharp contrast to the expert's ability to identify independencies easily and confidently, while having difficulties assessing the required distribution. Thus, if one wishes to create a framework in which an expert feels comfortable to express knowledge, one must provide a language in which information about dependence can be expressed without reference to numeric distributions. On the other hand, one must be certain that this language allows only the specification of independence and dependence assertions that are *consistent,* namely, assertions that can be realized simultaneously by some probability distribution.

A simple example serves to illustrate the problem: Are the following two statements consistent: ‘‘a is independent of both b and c’’ and ‘‘a is dependent on b, once c is known’’ ? Could these two assertions be realized simultaneously? The answer is negative. Whenever the former statement is realized, the negation of the latter is implied. This observation is phrased axiomatically by the weak union axiom (1.5d), or more explicitly by the axiom below, showing that the negation of the second assertion is implied from the first:

$$I(a, \emptyset, \{b, c\}) \Rightarrow I(a, c, b).$$

A critical step in assessing whether a given mixed set of dependencies and independencies is consistent is the solution of the *implication problem*: Given a set Σ of independencies and a single independence statement σ, decide whether σ logically follows from Σ, namely whether every distribution that satisfies Σ must satisfy σ as well. If the answer is negative, then the negation of σ is consistent with Σ, otherwise it is inconsistent. The basis for the solution of the implication problem is a *complete* set of inference rules. A complete set of inference rules is guaranteed to generate all independence statements that logically follow from a given set of statements. Such inference rules also serve to answer whether a set of independencies Σ^+ and a set of dependencies Σ^- are consistent: For each member of Σ^- determine, using the implication algorithm, whether its negation logically follows from Σ^+. If the answer is negative for all members of Σ^-, then the two sets are consistent, otherwise they are inconsistent.

The correctness of this algorithm is a major result of this chapter. It stems from the fact that if each member of Σ^- is individually consistent with Σ^+, then the entire set Σ^- is consistent with Σ^+. Independence relations that possess this property are called *Armstrong relations* (Section 2.3). These relations are characterized by the fact that if a disjunction of independence statements logically follows from a given set of such statements, then at least one disjunct must follow in itself. Thus, in order to check consistency of a set of assertions articulated by an expert, it suffices to check consistency of each individual dependency with the set of independencies.

This chapter is organized as follows: Section 2.2 presents the basic notions of soundness and completeness of a set of axioms. Section 2.3 shows that probabilistic and relational independence are Armstrong relations while correlational independence is not. Section 2.4 and 2.5 establish a complete set of axioms for marginal and saturated statements drawn from three distinct types of independence relations: correlational, relational and probabilistic. Section 2.6 summarizes most results in three tables.

2.2 Preliminary Definitions

The following notations are employed: σ, possibly subscripted, denotes a statement, Σ denotes a set of statements and $\mathcal{P}$ denotes a class of distributions. Among these are strictly positive discrete distributions ($\mathcal{PD}^+$), non-degenerated normal distributions ($\mathcal{PN}$), distributions over binary variables ($\mathcal{PB}$) and the class of all discrete probability distributions ($\mathcal{PD}$). A distribution $P(X)$ in $\mathcal{PB}$ is any joint distribution of the set of variables X in which each $x_i \in X$ has a domain of size two (e.g., $\{0, 1\}$). A distribution in $\mathcal{PN}$ is defined below. Variable symbols are drawn from a finite set $U = \{u_1, u_2, \cdots\}$. Letters x, y, z, u, v, w, possibly subscripted, denote variables, and X, Y, Z, V, W denote sets of variables. The set union symbol is dropped from complicated expressions: XY is written instead of $X \cup Y$ and Xa is written instead of $X \cup \{a\}$. The phrase "with respect to" is abbreviated to "wrt". Three classes of independence statements are considered:

i) General statements of independence called *probabilistic (independence) statements,* are denoted by $I(X, Z, Y)_{\mathcal{P}}$, where X, Y and Z are finite disjoint sets of variables, and $I(X, Z, Y)_{\mathcal{P}}$ is defined by Eq. (1.2). The following are equivalent definitions for $I(X, Z, Y)_{\mathcal{P}}$ [39]:

$$I(X,Z,Y)_{\mathcal{P}} \iff P(X,Y,Z) = P(X,Z)\,P(Y \mid Z) \text{ whenever } P(Z) > 0 \qquad (2.1a)$$

$$I(X,Z,Y)_{\mathcal{P}} \iff \underset{functions\ f,\,g}{\exists} P(X,Y,Z) = f(X,Z) \cdot g(Y,Z) \qquad (2.1b)$$

ii) *Saturated probabilistic statements* (or *saturated statements),* also denoted $I(X, Z, Y)_{\mathcal{P}}$, are a special case of probabilistic statements where $X \cup Y \cup Z$ must sum to a fixed finite set of variables U.

iii) Statements of marginal independence *(marginal probabilistic statements)* are denoted $I(X, \emptyset, Y)_{\mathcal{P}}$, where $\emptyset$ means that the value of no variable in U is known.

Two types of independence relations, other than probabilistic, are examined in this chapter: *relational* and *correlational* Their definition and the dependency models they induce are given below.

Definition [51]: Let U be a finite set of variables and let $domain(u_i)$, $u_i \in U$, be countable sets. A *Relational Dependency Model M_R* is defined in terms of a discrete probability distribution R with a sample space $\underset{u_i \in U}{\times} domain(u_i)$. If X, Y and Z are three disjoint subsets of U, and $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{Z}$ are any instances of the variables in these subsets, then by definition $(X, Z, Y) \in M_R$ iff

$$R(\mathbf{X},\mathbf{Z}) > 0 \ \& \ R(\mathbf{Y},\mathbf{Z}) > 0 \Rightarrow R(\mathbf{X},\mathbf{Y},\mathbf{Z}) > 0 . \qquad (2.2)$$

When the implication above holds we say that $I(X, Z, Y)_{\mathcal{R}}$ holds for M_R. The relation $I_{\mathcal{R}}$ is called *relational independence.* This definition views a distribution as defining a relation that contains all instances that have non-zero probability. It conveys the idea that, once Z is fixed, knowing Y cannot further restrict the range of values permitted for X. (*)

Saturated relational statements and marginal relational statements are defined analogously to saturated and marginal probabilistic statements. Readers familiar with the literature of relational databases should recognize the similarities between relational statements and embedded multi-valued dependencies (*EMVD* s) [20], as well as the resemblance between saturated relational statements and multi-valued dependencies (*MVD*) [4]. When appropriate, concepts borrowed from database theory are identified.

Definition [51]: A *Correlational Dependency Model* M_C is defined in terms of a finite collection of random variables U having non-zero finite variances and finite means. If X, Y and Z are three disjoint subsets of U, then by definition $(X, Z, Y) \in M_C$ iff $\rho_{ab.Z} = 0$ for every $a \in X$ and $b \in Y$ where $\rho_{ab.Z}$ is the partial correlation of a and b, defined recursively by the following equation [13]:

$$\rho_{ab.Z \cup \{c\}} = \frac{\rho_{ab.Z} - \rho_{ac.Z}\rho_{bc.Z}}{(1-\rho_{ac.Z}^2)^{\frac{1}{2}}\,(1-\rho_{bc.Z}^2)^{\frac{1}{2}}} \tag{2.3}$$

and

$$\rho_{ab} \triangleq \rho_{ab.\varnothing} = \frac{E\,[ab - E\,[ab\,]]}{(E\,[a - E\,[a\,]])^{\frac{1}{2}}\,(E\,[b - E\,[b\,]])^{\frac{1}{2}}}$$

where $E\,[x\,]$ is the mean of x.

When the equation above holds we say that $I\,(X, Z, Y\,)_C$ holds for M_C. The relation I_C is called *correlational independence*. This definition conveys the idea that the linear estimation error of the variables in X using measurements on Z would not be reduced by adding measurements on variables in Y, hence making Y irrelevant to the estimation of X. The numerator of equation (2.3) shows the change made in the correlation between a and b once c is taken into account. The denominator is a normalization factor that keeps the range of $\rho_{ab.Zc}$ between -1 and 1.

A (non-degenerate) normal distribution is characterized by the multivariate density function of the form

$$f\,(X\,) = \frac{1}{(2\pi)^{n/2}\,|\Lambda|^{\frac{1}{2}}}\ \exp\left\{ -\frac{1}{2\,|\Lambda|}(x - m\,)\,\Lambda^{-1}(X - m\,)^t \right\}$$

where X is a vector of variables, $m = E\,[X\,]$ is a vector of averages, $\Lambda = E\,[(X - m\,)\,(X - m\,)^t\,]$ is the covariance matrix, and all variables have non-zero

(*) This type of independence has been defined by Fagin [20] in the context of relational databases and was named qualitative independence in [57].

finite variances and finite means. The predicate $I(X, Z, Y)_{\mathcal{P}}$ holds for a normal distribution f, if:

$$f(X \mid Z) = f(X \mid Y, Z)$$

where f stands for the conditional densities given Z. These densities are known to exists for non-degenerate normal distributions and they are normal [13]. Normal distributions enforce many properties on the predicate I of which the following is somewhat surprising:

$$I(X, Z = \mathbf{Z}, Y)_{\mathcal{P}} \Rightarrow I(X, Z, Y)_{\mathcal{P}}.$$

If two variables are independent given one instance of Z, then these variables are independent for any instance of Z. This property is due to the fact that a conditional normal distribution is always normal and its covariance matrix is not sensitive to the value of the conditioned variables, only its means are [68, pp. 115]. This property relies on the convention that a normal distribution is defined by the density function of the form above. If the density is changed for some zero-measurable sets of points, then the resulting distribution is no longer called normal.

Correlation and dependence are identical notions for normal distributions [13], namely,

$$I(X, Z, Y)_{\mathcal{P}} \leftrightarrow I(X, Z, Y)_C .$$

In other words, normal distributions portray only linear dependencies between variables; our investigation into correlational independence is therefore facilitated by examining probabilistic independence wrt the class of normal distributions.

The term *statement* will be used to denote any of the nine types of statements defined above, namely, saturated, marginal, and unrestricted statements drawn from one of three types of independence: probabilistic, relational and correlational. Unless otherwise written, a statement $I(X, Z, Y)$ stands for a *non-trivial* statement, i.e., where $X \neq \emptyset$ and $Y \neq \emptyset$. In other words, axiom (1.5a), $I(X, Z, \emptyset)$, will be assumed to hold and will not be stated explicitly.

Definition : An *axiom*

$$\sigma_1 \,\&\, \sigma_2 \,\&\, \cdots \,\&\, \sigma_n \Rightarrow \sigma$$

is *sound* wrt a class of distributions $\mathcal{P}$ if every distribution $P \in \mathcal{P}$ that satisfies the antecedents of the axiom also satisfies σ. Axioms (1.a) through (1.e) are examples of sound axioms wrt $\mathcal{PD}^+$, $\mathcal{PB}$, $\mathcal{PN}$, and $\mathcal{PD}$.

Definition: A statement σ is *logically implied (logically follows)* by Σ, denoted $\Sigma \models_{\mathcal{P}} \sigma$, iff every distribution in $\mathcal{P}$ that satisfies Σ also satisfies σ. The set of statements that logically follow from Σ (in $\mathcal{P}$) is called the *logical closure* of Σ and is denoted by Σ^*.

Definition: Let $\mathcal{A}$ be a set of axioms. We say that σ is derivable from Σ using $\mathcal{A}$, denoted $\Sigma \vdash_{\mathcal{A}} \sigma$ or $\sigma \in cl_A(\Sigma)$, if there exists a *derivation chain* $\sigma_1, \ldots, \sigma_n = \sigma$ such that for each σ_j, either $\sigma_j \in \Sigma$, or σ_j is derived by an axiom in $\mathcal{A}$ from previous statements.

Definition: A set of axioms $\mathcal{A}$ is *sound* wrt $\mathcal{P}$ iff for every statement σ and every set of statements Σ

$$\text{if } \Sigma \vdash_{\mathcal{A}} \sigma \text{ then } \Sigma \models_{\mathcal{P}} \sigma$$

The set $\mathcal{A}$ is *complete* wrt $\mathcal{P}$ iff

$$\text{if } \Sigma \models_{\mathcal{P}} \sigma \text{ then } \Sigma \vdash_{\mathcal{A}} \sigma$$

Proposition 2.1: A set of axioms is sound wrt $\mathcal{P}$ iff each axiom in the set is sound wrt $\mathcal{P}$.
The proof is achieved by induction on the length of a derivation.

Proposition 2.2 (After Fagin [19]): A set of axioms $\mathcal{A}$ is complete wrt $\mathcal{P}$ iff for every set of statements Σ and every statement $\sigma \notin cl_{\mathcal{A}}(\Sigma)$ there exists a distribution P_σ in $\mathcal{P}$ that satisfies Σ and does not satisfy σ.

Proof: This is the contra-positive form of the completeness definition: if $\Sigma \nvdash_{\mathcal{A}} \sigma$ (equivalently, $\sigma \notin cl_A(\Sigma)$) then $\Sigma \not\models_{\mathcal{P}} \sigma$. □

A complete set of axioms does not provide sufficient means for deriving all the information that is implied by a given set of statements. For example, assume that the set $\Sigma = \{I(X, a, Y)_{\mathcal{P}}, I(X, \emptyset, Y)_{\mathcal{P}}\}$ is given, where a is a single variable and all variables are bi-valued i.e., drawn from $\mathcal{PB}$. It can be shown that the disjunction $I(X, \emptyset, a)_{\mathcal{P}}$ *or* $I(Y, \emptyset, a)_{\mathcal{P}}$ logically follows from Σ [49]. Yet this disjunction cannot necessarily be derived by a complete set of axioms; A complete set only guarantees to reveal, correctly, that neither of the disjuncts is logically implied by Σ but would not show that one of the two statements must hold. To obtain all disjunctions, a stronger set of axioms is needed.

Definition (after [4, 19]): A set of axioms $\mathcal{A}$ is *strongly complete* wrt a class of distributions $\mathcal{P}$, if for every set of statements Σ and for every set of single statements $\{\sigma_i \mid i = 1, \cdots, n\}$ the following relation holds:

$$\Sigma \models_{\mathcal{P}} \sigma_1 \; or \; \cdots \; or \; \sigma_n \;. \;\; \leftrightarrow \;\; \Sigma \vdash_{\mathcal{A}} \sigma_1 \; or \; \cdots \; or \; \sigma_n$$

Similar to Proposition 2.2, the following holds:

Proposition 2.3 (After [4, 19]: A set of axioms A is *strongly complete* iff for every set of statements Σ closed under axioms $\mathcal{A}$, there exists a distribution P in $\mathcal{P}$ that satisfies all statements in Σ and none other. (*)

Note that a strongly complete set of axioms is complete but the converse is not always true [19].

2.3 Armstrong Relations

The concept of Armstrong relations has evolved in the theory of relational databases and has been stated by Fagin in rather general terminology that makes it applicable to probabilistic, relational and correlational independence [21]. We will use this property to show that completeness and strong completeness are identical concepts when speaking about probabilistic independence but are distinct when speaking about correlational independence. Furthermore, this property is used to justify the algorithm for checking consistency which was described in section 2.1.

Fagin's general setting consists of a class of *models*, which in our case is a class of probability distributions, a class of *sentences* $\mathcal{S}$ (for our purposes independence statements) and a relationship *Holds* that states whether a sentence holds in a given model. Holds(P, σ) means that σ *holds for* P or that P *satisfies* σ. σ is a logical consequence of Σ, written $\Sigma \models \sigma$, if every model that satisfies the set of sentences Σ satisfies the sentence σ as well. $\Sigma^* \triangleq \{\sigma \mid \Sigma \models \sigma\}$. A set of sentences Σ is *consistent* if there exists a model that satisfies every sentence in Σ.

Theorem 2.4 [21]: Let $\mathcal{S}$ be a set of sentences. The following properties of $\mathcal{S}$ are equivalent.

(a) **Existence of a faithful operator.** There exists an operator $\otimes$ that maps

(*) Strong completeness is the analog of completeness in logic. Our completeness definition is a weaker version of the standard definition. We use this terminology to emphasize the importance of the latter, as reflected in Chapter 3.

nonempty families of models into models, such that if σ is a sentence in $\mathcal{S}$ and $<P_i : i \in I>$ is a nonempty family of models, then σ holds for $\otimes <P_i : i \in I>$ if and only if σ holds for each P_i.

(b) **Existence of Armstrong models.** Whenever Σ is a consistent subset of $\mathcal{S}$ and Σ^* is the set of sentences in $\mathcal{S}$ that are logical consequences of Σ, then there is a model (an "Armstrong model") that obeys Σ^* and no other sentences in $\mathcal{S}$.

(c) **Splitting of disjunctions.** Whenever Σ is a subset of $\mathcal{S}$ and $\{\sigma_i : i \in I\}$ is a nonempty subset of $\mathcal{S}$, then $\Sigma \models \bigvee\{\sigma_i : i \in I\}$ if and only if there is some i in I such that $\Sigma \models \sigma_i$.

Both b) and c) are important characterizations of Armstrong relations. Part b) guarantees that for every set of statements Σ, one can find a model that satisfies all the statements logically implied by Σ and none other. Thus, since probabilistic independence is an Armstrong relation wrt $\mathcal{PD}$ (see Theorem 2.5), Theorem 2.4 assures the existence of a distribution P that satisfies exactly the logical closure Σ^* of any given set Σ of probabilistic independence statements. We stress that the existence of such a distribution is unrelated to whether a finite complete set of axioms exists for probabilistic independence wrt $\mathcal{PD}$; The Armstrong property stands in itself. Part c) states that if a disjunction of statements is logically implied by Σ, then at least one of the disjuncts must be implied in itself. This property is useful in its contra-positive form; we will use it to prove that probabilistic independence is not an Armstrong relation wrt $\mathcal{PN}$ or wrt $\mathcal{PB}$.

Fagin provides several applications for his theorem and this dissertation provides an additional one. We first show the existence of a faithful operator for probabilistic independence. We note that while the theorem holds for any cardinality of the index set I, we use it only for finite nonempty I. Next we concentrate on two families of distributions - all discrete distributions, $\mathcal{PD}$, and strictly positive discrete distributions, $\mathcal{PD}^+$. Probabilistic independence is shown to be an Armstrong relation wrt both families. It should be emphasized that an independence relation can be an Armstrong relation wrt one class of distributions and not wrt another. For example, probabilistic independence is shown to be an Armstrong relation wrt $\mathcal{PD}$ but it is not an Armstrong relation wrt $\mathcal{PB}$; $\{I(a, \emptyset, b), I(a, c, b)\}$ logically implies the disjunction $I(a, \emptyset, c)$ or $I(c, \emptyset, b)$ wrt $\mathcal{PB}$ but neither of the disjuncts alone follows, thus violating condition (c) of Theorem 2.4.

Theorem 2.5: Probabilistic independence is an Armstrong relation wrt $\mathcal{PD}$.

Proof: We construct the operation $\otimes$ for probabilistic independence using a binary operation $\otimes'$ such that if $P = P_1 \otimes' P_2$, then for every probabilistic independence statement we obtain:

$$\otimes' P_i \text{ satisfies } I(X, Z, Y)_{\mathcal{P}} \leftrightarrow P_1 \text{ satisfies } I(X, Z, Y)_{\mathcal{P}} \text{ and } P_2 \text{ satisfies } I(X, Z, Y)_{\mathcal{P}} . \qquad (2.4)$$

The operation $\otimes$ is recursively defined in terms of $\otimes'$ as follows:

$$\otimes \; \{P_i \mid i{=}1..n\} = ((P_1 \otimes' P_2) \otimes' P_3) \otimes' \cdots P_n).$$

Clearly, if $\otimes'$ satisfies Eq. (2.4), then $\otimes$ satisfies the requirement of an Armstrong relation, i.e.,

$$P \text{ satisfies } I(X, Z, Y)_{\mathcal{P}} \;\leftrightarrow\; \underset{i}{\forall} P_i \text{ satisfies } I(X, Z, Y)_{\mathcal{P}} .$$

Therefore, it suffices to show that $\otimes'$ satisfies (2.4) (note that since $\otimes'$ is associative, $\otimes$ is well-defined).

Let $P_1(x_1, \cdots, x_n)$ and $P_2(x_1, \cdots, x_n)$ be two distributions sharing the same set of variables. Let $A_1, \cdots, A_n$ be the domains of $x_1, \cdots, x_n$ in P_1 and let $\alpha_1, \cdots, \alpha_n$ be an instance of these variables. Similarly, let $B_1, \cdots, B_n$ be the domains of $x_1, \cdots, x_n$ in P_2 and $\beta_1, \cdots, \beta_n$ an instance of these variables. Let the domain of $P = P_1 \otimes' P_2$ be the product domain $A_1 \times B_1, \cdots, A_n \times B_n$ and denote an instance of the variables of P by $(\alpha_1, \beta_1), \cdots, (\alpha_n, \beta_n)$, or more condensed by, $\alpha_1\beta_1, \cdots, \alpha_n\beta_n$. Define $P_1 \otimes' P_2$ by the following equation:

$$P(\alpha_1\beta_1, \alpha_2\beta_2, \cdots, \alpha_n\beta_n) = P_1(\alpha_1, \alpha_2, \cdots, \alpha_n) \cdot P_2(\beta_1, \beta_2, \cdots, \beta_n).$$

If P_1 and P_2 are proper probability distributions, then so is P. The distribution P is called the *direct product* of P_1 and P_2. Note that variables' domain is being altered by the product distribution; these are not random variables in the common discourse of probability theory. The latter are usually defined for a fixed domain.

To prove that $\oplus'$ satisfies the required conditions, we first show that a similar equation holds for every subset $\{x_{i_1}, \cdots, x_{i_l}\}$ of the variables of P, namely that,

$$P(\alpha_{i_1}\beta_{i_1}, \alpha_{i_2}\beta_{i_2}, \cdots, \alpha_{i_l}\beta_{i_l}) = P_1(\alpha_{i_1}, \alpha_{i_2}, \cdots, \alpha_{i_l}) \cdot P_2(\beta_{i_1}, \beta_{i_2}, \cdots, \beta_{i_l}) . \qquad (2.5)$$

We start by validating Eq. (2.5) for $i_1 = 1, i_2 = 2, \cdots i_l = l$. When $l{=}n$ this equation is identical to Eq. (2.4). We proceed by descending induction. Assume Eq. (2.5) holds for $l = k < n$, then,

$$P(\alpha_1\beta_1, \cdots, \alpha_{k-1}\beta_{k-1}) = \sum_{x_k} P(\alpha_1\beta_1, \cdots, \alpha_{k-1}\beta_{k-1}, x_k)$$

$$= \sum_{(\alpha_k, \beta_k) \in A_k B_k} P_1(\alpha_1, \cdots, \alpha_{k-1}, \alpha_k) \cdot P_2(\beta_1, \cdots, \beta_{k-1}, \beta_k)$$

$$= \left[\sum_{\alpha_k \in A_k} P_1(\alpha_1, \cdots, \alpha_{k-1}, \alpha_k) \right] \cdot \left[\sum_{\beta_k \in B_k} P_2(\beta_1, \cdots, \beta_{k-1}, \beta_k) \right]$$

$$= P_1(\alpha_1, \cdots, \alpha_{k-1}) \cdot P_2(\beta_1, \cdots, \beta_{k-1})$$

The proof of Eq. (2.5) is completed by repeating the induction step for the $n!$ orderings of $\{x_1, \cdots, x_n\}$.

It is left to show that for every statement $I(X, Z, Y)_{\mathcal{P}}$ we have

$I(X, Z, Y)_{\mathcal{P}}$ holds for P iff $I(X, Z, Y)_{\mathcal{P}}$ holds for both P_1 and P_2.

More explicitly, we show that for every instance of X, Y, Z for which $P(Z) > 0$, the equation below holds:

$$P(X, Y, Z) = P(X, Z) \cdot P(Y \mid Z) \text{ iff } P_1(X, Y, Z) = P_1(X, Z) \cdot P_1(Y \mid Z) \text{ and}$$

$$P_2(X, Y, Z) = P_2(X, Z) \cdot P_2(Y \mid Z) \qquad (2.6)$$

Let $\alpha_x, \alpha_y, \alpha_z$ be an instance of X, Y, Z in P_1 and $\beta_x, \beta_y, \beta_z$ be an instance of X, Y, Z in P_2. The *if* part of (2.6) is proved as follows:

$$P(\alpha_x\beta_x, \alpha_y\beta_y, \alpha_z\beta_z) = P_1(\alpha_x, \alpha_y, \alpha_z) \cdot P_2(\beta_x, \beta_y, \beta_z)$$

$$= P_1(\alpha_x, \alpha_z) \cdot P_1(\alpha_y \mid \alpha_z) \cdot P_2(\beta_x, \beta_z) \cdot P_2(\beta_y \mid \beta_z) =$$

$$= P(\alpha_x\beta_x, \alpha_z\beta_z) \cdot \left[\frac{P_1(\alpha_y, \alpha_z) \cdot P_2(\beta_y, \beta_z)}{P_1(\alpha_z) \cdot P_2(\beta_z)} \right] =$$

$$= P(\alpha_x\beta_x, \alpha_z\beta_z) \cdot \left[\frac{P(\alpha_y\beta_y, \alpha_z\beta_z)}{P(\alpha_z\beta_z)} \right] = P(\alpha_x\beta_x, \alpha_z\beta_z) \cdot P(\alpha_y\beta_y \mid \alpha_z\beta_z)$$

The *only if* part of (2.6) follows from:

$$P_1(\alpha_x, \alpha_y, \alpha_z)\, P_2(\beta_x, \beta_y, \beta_z) \;=\; P(\alpha_x \beta_x, \alpha_y \beta_y, \alpha_z \beta_z)$$

$$= \frac{P(\alpha_x \beta_x, \alpha_z \beta_z) \cdot P(\alpha_y \beta_y, \alpha_z \beta_z)}{P(\alpha_z \beta_z)} =$$

$$= \left[\frac{P_1(\alpha_x, \alpha_z) \cdot P_1(\alpha_y, \alpha_z)}{P_1(\alpha_z)} \right] \cdot \left[\frac{P_2(\beta_x, \beta_z) \cdot P_2(\beta_y, \beta_z)}{P_2(\beta_z)} \right]$$

$$= \left[P_1(\alpha_x, \alpha_z) \cdot P_1(\alpha_y \mid \alpha_z) \right] \cdot \left[P_2(\beta_x, \beta_z) \cdot P_2(\beta_y \mid \beta_z) \right]$$

By summing once over α_x and once over β_x we get that $I(X, Z, Y)_{\mathcal{P}}$ holds both in P_2 and P_1, respectively.

If P_1 and P_2 are defined over different sets of variables and $X = \{x_1, \cdots, x_n\}$ are their common variables, then, instead of applying $\otimes$ directly on P_1 and P_2, we form their projections $P_1{}'$, $P_2{}'$ on X and define

$$P = P_1 \otimes P_2 = P_1{}' \otimes P_2{}'.$$

Clearly, every statement satisfied by P is satisfied by $P_1{}'$ and $P_2{}'$ and therefore also by P_1 and P_2. The other direction holds as well; a statement $I(Y, V, W)$ that holds both for P_1 and for P_2 must satisfy $YVW \subseteq X$. This implies that $I(Y, V, W)$ holds in $P_1{}'$ and in $P_2{}'$ and therefore, by our construction, it holds in P as well. □

The direct product construction is also applicable wrt $\mathcal{PD}^+$ because $\otimes$ produces a strictly positive distribution whenever the input distributions are strictly positive. Furthermore, since saturated statements and marginal statements are both subclasses of independence statements, the construction of $\otimes$ assures that marginal and saturated probabilistic independence are Armstrong relations both wrt $\mathcal{PD}$ and wrt $\mathcal{PD}^+$. These considerations are summarized in the following corollary.

Corollary 2.6: Marginal, saturated, and unrestricted probabilistic independence are Armstrong relations both wrt $\mathcal{PD}$ and wrt $\mathcal{PD}^+$.

The operation $\otimes$ constructed for probabilistic independence (Theorem 2.5) also satisfies the requirement of an Armstrong operation for relational independence.

Corollary 2.7: Marginal, saturated, and unrestricted relational independence are Armstrong relations both wrt $\mathcal{PD}$ and wrt $\mathcal{PD}^+$.

Proof: Let $\otimes$ and $\otimes'$ be the operations defined by Theorem 2.5. If R_1 and R_2 are two discrete distributions then $R = R_1 \otimes R_2$ satisfies a relational statement iff both R_1 and R_2 satisfy this statement; the proof is analogous to that given for Theorem 2.5. □

The class of normal distributions exemplifies a case where marginal probabilistic independence is an Armstrong relation, while probabilistic independence is not. In $\mathcal{PN}$, $P(a,b) = P(a) \cdot P(b)$ iff $\rho_{ab} = 0$, where ρ_{ab} is the correlation factor of the variables a and b. Given a set of normal distributions we construct the normal standard distribution $\oplus P_i$ by assigning $\rho_{ab} = 0$ in $\oplus P_i$ iff $\rho_{ab} = 0$ in every P_i. All other correlation factors are assigned a non-zero quantity ρ, where ρ satisfies $n \cdot \rho^2 < 1$ to assure that the covariance matrix of $\oplus P_i$ is positive definite. Therefore, since $\oplus$ satisfies the requirements of the Armstrong definition, marginal independence is an Armstrong relation in $\mathcal{PN}$. Probabilistic independence in $\mathcal{PN}$ is not an Armstrong relation because although $\{I(a,\emptyset,b), I(a,c,b)\}$ logically implies the disjunction $I(a,\emptyset,c)$ or $I(c,\emptyset,b)$ (for jointly normally distributed variables a,b,c, the antecedents can be written as $\rho_{ac} \cdot \rho_{cb} = \rho_{ab} = 0$, hence either $\rho_{ac} = 0$ or $\rho_{cb} = 0$.), neither of the disjuncts alone follow, thus violating condition (c) of Theorem 2.4.

This example suggests, that to fully characterize probabilistic independence one should start by considering *disjunctive axioms,* i.e., axioms of the form,

$$s_1 \ \& \ s_2 \ \& \ \cdots \ s_n \Rightarrow \sigma_1 \ or \ \cdots \ or \ \sigma_m$$

(s_i, σ_i are statements) and not merely *Horn axioms* where $m = 1$. However, the Armstrong property of probabilistic independence (Theorem 2.5, Theorem 2.4 part c) assures that this is unnecessary — Horn axioms are sufficient to derive all disjunctions (in $\mathcal{PD}$ and $\mathcal{PD}^+$). This example also shows that correlational independence, which identifies with probabilistic independence in the class of normal distributions, is not an Armstrong relation.

Corollary 2.8: Marginal, saturated and unrestricted correlational independence are not Armstrong relations in $\mathcal{PD}^+$.

Section 2.1 suggested that in order to verify that a set of independencies Σ^+ and a set of dependencies Σ^- are consistent, we merely need to verify that each member of Σ^- is consistent with Σ^+. The correctness of this method is now clear. If a dependency $\neg\sigma$ is consistent with Σ^+, then there exists a distribution P_σ that satisfies Σ^+ and the dependency $\neg\sigma$. If such a distribution exists for each dependency in Σ^-, then the distribution $P = \otimes \{P_\sigma \mid \neg\sigma \in \Sigma^-\}$ guarantees the consistency of the the two sets; it satisfies all the dependencies of Σ^- and the independencies in Σ^+.

2.4 Saturated Independence

This section characterizes probabilistic, relational and correlational *saturated independence*. The interest in these statements stems from their serving as a *basis* for Markov networks (see section 1.3 and Theorem 3.16)

Theorem 2.9 (completeness for saturated independence): For every set Σ of saturated probabilistic (or relational) statements closed under the axioms:

Symmetry $\quad I(X,Z,Y) \Rightarrow I(Y,Z,X) \quad$ (2.7a)

Weak union $\quad I(X,Z,YW) \Rightarrow I(X,ZY,W) \quad$ (2.7b)

Weak contraction $\quad I(XY,Z,W) \; \& \; I(X,ZW,Y) \Rightarrow I(X,Z,YW) \quad$ (2.7c)

there exists a probability distribution P_σ that satisfies all statements in Σ and does not satisfy any saturated statement outside Σ.

Remark: Axioms (2.7) are sound wrt both relational and probabilistic independence.

Proof: Let $\sigma = I(X,Z,Y)$ be an arbitrary saturated statement (i.e., $U = XYZ$) not in Σ. We show that without loss of generality one can assume that for all sets $X'X''$ and $Y'Y''$ partitioning X and Y respectively, the statement $I(X', Z\,X''Y'', Y')$ is a member of Σ. A saturated statement satisfying this property is called a *maximal dependency*. If $\sigma = I(X,Z,Y)$ is not a maximal dependency, then we identify a maximal dependency σ' of the form $I(X', ZX''Y'', Y')$. Clearly, σ' always exists because Z is augmented by elements of X and Y until the desired property is obtained or until both X' and Y' become singletons, in which case, trivially, σ' is a maximal dependency. We will construct $P_{\sigma'}$ that satisfies Σ and violates σ'. Due to axioms (2.7a, 2.7b), which hold for all distributions, we know that any distribution that violates σ', violates σ as well. In particular, $P_{\sigma'}$ violates σ (while satisfying Σ), and therefore satisfies the conditions of the theorem.

Let U be a set of binary variables and $\sigma = I(X,Z,Y)$ be a maximal dependency. Denote all variables in X by $\{x_1, x_2 \cdots x_l\}$, those in Y by $\{y_1, y_2 \cdots y_m\}$ and those in Z by $\{z_1, z_2 \cdots z_k\}$. The construction of P_σ is obtained by forcing all variables in X and Y to be equal to one another and letting each z_i represent the outcome of an independent fair coin. The resulting distribution follows:

$$P_\sigma(X, Y, Z) = \prod_{z_i \in Z} P_\sigma(z_i) \cdot \begin{Bmatrix} \frac{1}{2} & \text{if all variables in } XY \text{ are equal to } 0 \\ \frac{1}{2} & \text{if all variables in } XY \text{ are equal to } 1 \\ 0 & \text{otherwise} \end{Bmatrix}$$

Clearly, P_σ does not satisfy σ because

$$P_\sigma(X = 0, Y = 1 \mid Z = 0) \neq P_\sigma(X = 0 \mid Z = 0) \cdot P_\sigma(Y = 1 \mid Z = 0)$$

It remains to be shown that every saturated statement in Σ holds for P_σ, or equivalently that every saturated statement either holds for P_σ or does not belong to Σ. Any saturated statement γ can be written as $I(X_1Y_1Z_1, X_2Y_2Z_2, X_3Y_3Z_3)$ where $X = X_1X_2X_3$, $Y = Y_1Y_2Y_3$ and $Z = Z_1Z_2Z_3$. If $X_2Y_2 \neq \emptyset$ then γ holds in P_σ because every instance of $X_1Y_1Z_1$ and of $X_3Y_3Z_3$ that is consistent with the values of X_2Y_2 has the same probability of occurring, namely $\frac{1}{2}^{|Z_1|} \cdot \frac{1}{2}^{|Z_3|}$. If $X_1Y_1 = \emptyset$ (symmetrically when $X_3Y_3 = \emptyset$) then again γ holds in P_σ because Z_1 (Z_3) is marginally and conditionally independent of any other set of variables of P_σ. Otherwise γ is of the form $I(X_1Y_1Z_1, Z_2, X_3Y_3Z_3)$ where $X_1Y_1 \neq \emptyset$ and $X_3Y_3 \neq \emptyset$. We continue by contradiction and show that in this case γ does not belong to Σ.

Assume $I(X_1Y_1Z_1, Z_2, X_3Y_3Z_3)$ does belong to Σ. Σ is closed under weak-union and symmetry. Therefore, $I(X_1Y_1, Z, X_3Y_3) \in \Sigma$. To reach a contradiction we show that this statement implies that σ must have been in Σ, contradicting our selection of σ. The proof uses the weak-contraction and symmetry axioms to infer $I(X_1X_3, Z, Y_1Y_3)$, or σ, from $I(X_1Y_1, Z, X_3Y_3)$ by "pushing" all the X's to one side and all Y's to the other side. We further assume that X_1, X_3, Y_1, Y_3 are non-empty sets. If some of these sets are empty, not all the derivations that follow need to be performed to reach the contradicting conclusion that $\sigma \in \Sigma$. The following is a derivation of σ.

First, $I(X_1, ZX_3Y_3, Y_1)$ belongs to Σ because $I(X, Z, Y)$ is a maximal dependency. Due to the weak-contraction axiom

$$I(X_1Y_1, Z, X_3Y_3) \;\&\; I(X_1, ZX_3Y_3, Y_1) \Rightarrow I(X_1, Z, Y_1X_3Y_3),$$

we conclude that $I(X_1, Z, YX_3) \in \Sigma$. Due to the symmetry axiom we conclude $I(YX_3, Z, X_1) \in \Sigma$ as well. $I(X_3, ZX_1, Y) \in \Sigma$ because σ is a maximal dependency and therefore (by symmetry) $I(Y, ZX_1, X_3)$ is also a member of Σ. Using weak-contraction again, we obtain:

$$I(YX_3, Z, X_1) \;\&\; I(Y, ZX_1, X_3) \Rightarrow I(Y, Z, X_1X_3).$$

This result leads to the conclusion that $I(Y, Z, X) \in \Sigma$, and thus, by symmetry, $I(X, Z, Y) \in \Sigma$, a contradiction. Thus, axioms (2.7) are complete for saturated probabilistic independence.

Next, we show that axioms (2.7) are also complete for saturated relational independence.

Let σ be an arbitrary saturated relational statement. Since axioms (2.7) can readily be shown to hold for relational independence, we can repeat the argument showing that σ can be assumed a maximal dependency. Consider the distribution P_σ as constructed in the first part of the proof. Clearly, σ does not hold in P_σ because

$$P_\sigma(x=0\,, z=0) > 0 \;\&\; P_\sigma(y=1\,, z=0) > 0 \;\not\Rightarrow\; P_\sigma(x=0, y=1, z=0) > 0.$$

In addition, P_σ satisfies all statements Σ (interpreted as relational independence) because $I(Y, W, V)_{\mathcal{P}}$ implies $I(Y, V, W)_{\mathcal{R}}$, and therefore satisfies the requirement of the theorem. □

Theorem 2.9 guarantees that by repeated application of axioms (2.7) on a set Σ of saturated statements (relational as well as probabilistic), any saturated statement that logically follows from Σ will eventually be derived. The distribution P_σ as constructed in the proof has an additional property; each combination of values for XYZ has either zero probability or a constant probability of $(½)^{|z|+1}$. Thus, the distribution P_σ can be viewed as a *database*, categorically distinguishing between two sets of tuples. This observation implies that the proof of Theorem 2.9 can be modified to show that axioms (2.7) are complete for Multi-Valued-Dependencies (*MVD* s) [20]. Indeed, the difference between these axioms and the ones governing *MVD* s [4] pertains only to the case of overlapping sets X, Y and Z in $I(X, Z, Y)_{\mathcal{R}}$. This equivalence permits the employment of a polynomial implication algorithm devised for *MVD* s [3] to determine whether a saturated statement logically follows from a set of such statements.

The next theorem establishes an axiomatization of probabilistic independence for two narrower classes of distributions: strictly positive and normal. For these classes, relational independence is a trivial relation, i.e., every relational statement holds in all distributions of $\mathcal{PD}^+$ and $\mathcal{PN}$.

Theorem 2.10: For every set of saturated statements Σ closed under the following axioms:

Symmetry $\qquad I(X, Z, Y) \Rightarrow I(Y, Z, X) \qquad$ (2.8a)

Weak union $\quad I(X,Z,YW) \Rightarrow I(X,ZY,W)$ (2.8b)

Intersection $\quad I(X,ZY,W) \ \& \ I(X,ZW,Y) \Rightarrow I(X,Z,YW)$ (2.8c)

and for every saturated statement σ not in Σ there exists a *non-degenerate* normal distribution $P_\sigma \in \mathcal{PN}$ that satisfies all statements in $cl(\Sigma)$ and does not satisfy σ.

Theorem 2.10 ensures the completeness of axioms (2.8) for saturated statements wrt $\mathcal{PD}^+$. Axioms (2.8) are implied by axioms (2.7) but not vice versa. Indeed, the additional knowledge that the distributions involved are positive results in more independencies being implied by Σ. However, further restricting the class to that of normal distributions, no longer has any effect.

Proof: We first show that without loss of generality one can assume σ is of the form $I(a,Z,b)$ where a and b are single variables. If σ is not of this form, say $\sigma = I(X,Z,Y)$ where Y is not a singleton, then for every element $b' \in Y$, either $I(X, Z(Y-\{b'\}), b')$ or $I(X, Z \cup \{b'\}, Y-\{b'\})$ is not a member of Σ. Otherwise, since Σ is closed under intersection (2.8c), this would imply that $I(X,Z,Y)$ is a member of Σ as well, contradicting our assumption. We repeat this process of augmenting Z by elements of Y until we obtain a statement of the form $I(X, Z(Y-\{b\}), b) \notin \Sigma$. This process is guaranteed to terminate because in each step the set Y is decreased by one element. A similar procedure is repeated on X to obtain a statement σ' of the form $I(a, Z(X-\{a\})(Y-\{b\}), b)$ which is not in Σ. Due to the weak-union (2.8b) and symmetry (2.8b) axioms, which hold for all distributions, any distribution that violates σ' must violate σ as well. Thus, it suffices to show the construction of P_σ for which X and Y are singletons. Consider the distribution

$$P_\sigma(a,b,Z) = f(a,b) \cdot \prod_{z_i \in U-ab} f(z_i) \ ,$$

where $f(\cdot)$ is the standard normal distribution and $f(\cdot\,,\cdot)$ is zero mean-normal distribution with a non-zero correlation factor between its two arguments. Clearly, in P_σ a and b are marginally and conditionally dependent and thus σ is not satisfied. It is left to show that P_σ satisfies Σ, or equivalently (in contra-positive form), that every saturated statement which P_σ does not satisfy cannot be a member of Σ. However, every saturated statement that does not hold in P_σ must be of the $I(\{a\} \cup Z', \hat{Z}, \{b\} \cup Z'')$ where $Z = Z'Z''\hat{Z}$. These statements cannot be members of Σ because each of them implies that σ is a member of Σ (by (2.8a) and (2.8b)), contradicting our selection of σ. □

Remark: Axioms (2.8) are complete for saturated independence in $\mathcal{PB}^+$ and $\mathcal{PN}$, since the selection of the functions f does not depend on the arity of a, b or Z.

Theorem 3.16 provides a graph-based proof of a similar theorem, stating that axioms (2.8), combined with decomposition (1.5c), are powerful enough to derive all statements, not merely saturated ones, that logically follow from an arbitrary set of saturated statements. Unlike the proof of Theorem 2.10, the proof of Theorem 3.16 is constructive and therefore provides an algorithm to answer the implication problem.

Next, we show that probabilistic and relational independence are equivalent in the sense that they induce the same dependency models.

Theorem 2.11 (equivalence of saturated probabilistic and relational independence):

1. For every distribution P there exists a distribution R such that for every disjoint sets of variables X, Y and Z, where $XYZ = U$, we have,

$$I(X, Z, Y)_{\mathcal{P}} \ \textit{iff} \ I(X, Z, Y)_{\mathcal{R}}$$

2. For every distribution R there exists a distribution P such that for every disjoint sets of variables X, Y and Z, where $XYZ = U$, we have,

$$I(X, Z, Y)_{\mathcal{R}} \ \textit{iff} \ I(X, Z, Y)_{\mathcal{P}}$$

Proof: Every distribution P satisfies axioms (2.7), thus by theorem 2.9, there exists a distribution R_σ that satisfies all statements in P and does not satisfy a given dependency. Since relational independence is an Armstrong relation, the direct product R of all R_σ's satisfies the requirements. The proof of part (b) is analogous. □

2.5 Marginal Independence

This section characterizes probabilistic, relational and correlational *marginal independence.*

Theorem 2.12 (Completeness for marginal probabilistic independence): Let Σ be a set of marginal probabilistic statements closed under the following axioms:

$$\textit{Symmetry} \qquad I(X, \varnothing, Y) \Rightarrow I(Y, \varnothing, X) \tag{2.9a}$$

$$\textit{Decomposition} \qquad I(X, \varnothing, YW) \Rightarrow I(X, \varnothing, Y) \tag{2.9b}$$

$$\textit{Mixing} \qquad I(X, \varnothing, Y) \ \& \ I(XY, \varnothing, W) \Rightarrow I(X, \varnothing, YW) \tag{2.9c}$$

Then there exists a distribution $P_\sigma \in \mathcal{PB}$ that satisfies all statements in Σ and no other marginal statement.

Proof: Let $\sigma = (X, \varnothing, Y)$ be an arbitrary marginal statement not in Σ. Without loss of generality we assume that for all non-empty sets X' and Y' obeying $X' \subseteq X$, $Y' \subseteq Y$ and $X'Y' \neq XY$ we have $(X', \varnothing, Y') \in \Sigma$. A statement obeying this property is called a *minimal statement.* If $\sigma = (X, \varnothing, Y)$ is not a minimal statement then we can always find a minimal statement $\sigma' = (X', \varnothing, Y')$ not in Σ, where $X' \subseteq X$ and $Y' \subseteq Y$, by deleting elements of X and Y until we obtain the desired property or until both X' and Y' become singletons, in which case, σ' is a minimal statement. For each such σ', we construct $P_{\sigma'}$ that satisfies Σ and violates σ'. Due the decomposition axiom (2.9b), which holds for all distributions, we know that any distribution that violates σ', violates σ as well. In particular, $P_{\sigma'}$ violates σ (while satisfying Σ), and therefore satisfies the conditions of the theorem.

For the rest of this proof, we shorten the notation for a marginal statement $(X, \varnothing, Y)$ to (X, Y). Let $\sigma = (X, Y)$ be a minimal statement where $X = \{x_1, x_2 \cdots x_l\}$, $Y = \{y_1, y_2 \cdots y_m\}$ and let $Z = \{z_1, z_2 \cdots z_k\}$ stand for the rest of the variables, namely, $U - XY$. Construct P_σ as follows: Let all variables, except x_1, be independent binary variables with probability ½ for each of their two values (e.g., fair coins), and let

$$x_1 = \sum_{i=2}^{l} x_i + \sum_{i=1}^{m} y_j \qquad (\,mod 2\,).$$

Clearly, P_σ has the product form:

$$P_\sigma(X\ Y\ Z) = P_\sigma(X\ Y) \cdot \prod_{z_i \in Z} P_\sigma(z_i). \tag{2.10}$$

We first show that $\sigma = (X, Y)$ does not hold in P_σ. Instantiate x_1 to one and all other variables in XY to zero. For this assignment of values we have

$$P_\sigma(x_1 \cdots x_l, y_1 \cdots y_m) \neq P_\sigma(x_1 \cdots x_l) \cdot P_\sigma(y_1 \cdots y_m) \tag{2.11}$$

because the LHS of Eq. (2.11) is equal to 0 whereas the RHS consists of a product of two non-zero quantities.

It is left to show that every statement in Σ holds in P_σ, or equivalently, that for an arbitrary statement (V, W) we have:

$$(V, W) \in \Sigma \Rightarrow P_\sigma(V, W) = P_\sigma(V) \cdot P_\sigma(W).$$

This is done by examining the statement (V, W) for every possible assignment of variables to the sets V and W and showing that either $P_\sigma(V, W) = P_\sigma(V) \cdot P_\sigma(W)$ or that $(V, W) \notin \Sigma$.

Case 1: Either V or W contain only elements of Z.
By Eq. (2.10), we get $P_\sigma(V, W) = P_\sigma(V) \cdot P_\sigma(W)$.

Case 2: Both V and W include an element of $X \cup Y$.

Case 2.1: $V \cup W$ does not include all the variables of $X \cup Y$.
To verify whether (V, W) holds in P_σ amounts to checking this statement in the projection of P_σ on the set $V \cup W$. Since the probability of every value assignment to a proper subset $S \subsetneqq X \cup Y$ is $(½)^{|S|}$, this projection assumes the product form $\prod_{w_i \in Y \cup V} P_\sigma(w_i)$. Hence, again, $P_\sigma(V, W) = P_\sigma(V) \cdot P_\sigma(W)$.

Case 2.2: $V \cup W$ includes all elements of $X \cup Y$.
This is the only case for which (V, W) is definitely not in Σ.

Let $V = X'Y'Z'$, $W = X''Y''Z''$ where $X = X'X''$, $Y = Y'Y''$ and $Z'Z'' \subseteq Z$. We continue by contradiction. Assume $(V, W) = (X'Y'Z', X''Y''Z'')$ belongs to Σ. Σ is closed under decomposition. Therefore, $(X'Y', X''Y'') \in \Sigma$. To reach a contradiction we show that this statement implies that σ must have been in Σ, contradicting our selection of σ. The proof uses the mixing and symmetry axioms to infer $(X'X'', Y'Y'')$ (i.e., σ) from $(X'Y', X''Y'')$ by "pushing" all the X's to one side and all Y's to the other side. The following is a derivation of σ.

First, (X', Y') belongs to Σ because (X, Y) is a minimal statement. Due to the mixing axiom

$$(X', Y') \;\&\; (X'Y', X''\,Y'') \Rightarrow (X', Y'X''Y'') .$$

We conclude that $(X', X''Y) \in \Sigma$. Due to symmetry $(X''Y, X') \in \Sigma$ as well. $(X'', Y) \in \Sigma$ because σ is a minimal statement and therefore (by symmetry) also (Y, X'') is a member of Σ. Using the mixing axiom again, we get,

$$(Y, X'') \;\&\; (YX'', X') \Rightarrow (Y, X'X'')$$

which leads to the conclusion that $(Y, X) \in \Sigma$, and by symmetry that (X, Y) is in Σ, contradiction (note that the derivation of σ remains valid when some of $X'\ X''\ Y'\ Y''$ are empty, as long as $X = X'X''$ and $Y = Y'Y''$). □

An $O(|\Sigma| \cdot n^2)$ implication algorithm based on these axioms has been developed by Paz [47]. This algorithm, presented below, uses the procedure *Find* to answer whether a statement σ is derivable from Σ by axioms (2.9a) through (2.9d). The notation, span(σ) stands for the set of elements represented in a statement σ, and similarly, span(Σ) denotes the set of elements represented in all the statements of Σ. For example, span($\{I(x_1, \emptyset, x_2)\ I(x_1, \emptyset, x_3)\}$) is $\{x_1, x_2, x_3\}$. The *projection* of σ on s, denoted $\sigma(s)$, is the statement derived from σ by removing all elements not in s from σ e.g., if $\sigma = (x_1x_2x_3, \emptyset, x_4x_5)$ then $\sigma(x_1x_2x_3) = (x_1x_2x_3, \emptyset, \emptyset)$ and $\sigma(x_1x_3x_4x_6) = (x_1x_3, \emptyset, x_4)$. Similarly, the projection of Σ on s, denoted $\Sigma(s)$, stands for $\{\sigma(s) \mid \sigma \in \Sigma\}$.

Implication Algorithm

Procedure Find (Σ, σ):

1. $\Sigma' := \Sigma(\text{span}(\sigma))$ { Σ' is the projection of Σ on the variables of the target statement σ }

2. If σ is trivial, or σ (or its symmetric image) belongs to Σ' then set Find(Σ, σ) := True and return.

3. Else if for all nontrivial $\sigma' \in \Sigma'$, span(σ') $\neq$ span(σ) then set Find(Σ, σ) := False.

4. Else there exists a statement $\sigma' \in \Sigma'$ such that span(σ') = span(σ), and up to symmetry, $\sigma' = (AP, BQ)$ and $\sigma = (AQ, BP)$ where one of the sets A, B, P, Q may be empty (If several such σ' exist, then choose one arbitrarily).

```
Set $\sigma_1 := (A, P), \sigma_2 = (B, Q)$,
Find $(\Sigma, \sigma) :=$ Find $(\Sigma', \sigma_1) \wedge$ Find $(\Sigma', \sigma_2)$.
return.
Begin {Membership}
    Input( $\Sigma, \sigma$)
    Print Find( $\Sigma, \sigma$)
End.
```

The correctness proof for this algorithm can be found in [47]. Since P_σ, constructed in the proof of Theorem 2.12, belongs to $\mathcal{PB}$ the implication algorithm and the axiomatization are also valid for $\mathcal{PB}$. A minor change in the construction of P_σ shows that axioms (2.9) are complete also wrt $\mathcal{PD}^+$. In $\mathcal{PD}$ and $\mathcal{PD}^+$ marginal independence is an Armstrong relation, hence these axioms are strongly complete. Moreover, these completeness results are also valid for relational independence (using a proof similar to that of Theorem 2.9).

The axioms that characterize marginal independence for normal distributions are stronger than those of (2.7); the mixing axiom is replaced by composition:

Theorem 2.13: The following axioms are strongly complete for marginal independence wrt $\mathcal{PN}$ (normal distributions).

$$\textit{Symmetry} \qquad I(X, \emptyset, Y) \Rightarrow I(Y, \emptyset, X) \tag{2.12a}$$

$$\textit{Decomposition} \qquad I(X, \emptyset, YW) \Rightarrow I(X, \emptyset, Y) \tag{2.12b}$$

$$\textit{Composition} \qquad I(X, \emptyset, Y) \ \& \ I(X, \emptyset, W) \Rightarrow I(X, \emptyset, YW) \tag{2.12c}$$

Proof: Let Σ be a set of marginal statements closed under Symmetry (2.12a), Decomposition (2.12b), Composition (2.12c). We construct a normal distribution P that satisfies Σ and no other marginal statement. Let $U = \{u_1, \cdots, u_n\}$ be all the variables appearing in statements of Σ and let P be a zero-mean normal distribution over the variables of U, with the following covariance matrix

$$\Gamma = (\rho_{i,j}) \qquad \textit{where} \quad \rho_{i,j} = \begin{cases} 0 & (u_i, \emptyset, u_j) \in \Sigma \\ \rho & \text{otherwise} \end{cases}$$

and $n \cdot \rho^2 < 1$ (To assure positive definiteness).

We need to show that P satisfies Σ and no other marginal statement, or equivalently that $I(X, \varnothing, Y) \in \Sigma$ if and only if $I(X, \varnothing, Y)$ holds in P. This is proven by the following chain of relationships:

$$I(X,\varnothing,Y) \in \Sigma \leftrightarrow \underset{\substack{u_i \in X \\ u_j \in Y}}{\forall} I(u_i,\varnothing,u_i) \in \Sigma \leftrightarrow \underset{\substack{u_i \in X \\ u_j \in Y}}{\forall} \rho_{i,j} = 0 \leftrightarrow$$

$$\underset{\substack{u_i \in X \\ u_j \in Y}}{\forall} I(u_i,\varnothing,u_j) \text{ holds} \in P \leftrightarrow I(X,\varnothing,Y) \text{ holds} \in P$$

The first and last relationships hold because axioms (2.6) hold both in Σ and in P making any statement $I(X, \varnothing, Y)$ completely determined by statements on singletons. The middle relationships trivially hold. □

The proof of Theorem 2.13 implies that the the matrix Γ constitutes a dense representation of the logical closure of a set of marginal statements wrt $\mathcal{PN}$. It requires $O(|\Sigma| \cdot n^2)$ steps, where n is the number of variables appearing in statements of Σ. To test whether a statement $I(X, \varnothing, Y)$ is in the closure amounts to verifying that $\rho_{i,j} = 0$ for every $u_i \in X$ and $u_j \in Y$ which is of order $O(n^2)$.

The difference between the implication algorithm for marginal statements wrt $\mathcal{PN}$ vs. $\mathcal{PD}$ is significant; the later requires $O(|\Sigma| \cdot n^2)$ operations to decide $\Sigma \models_{\mathcal{P}} \sigma$, while the former requires only $O(n^2)$ operations, regardless of $|\Sigma|$ (note that this is achieved at the cost of investing $O(|\Sigma| \cdot n^2)$ steps in constructing a condensed representation of the closure of Σ). This advantage, offered by normal distributions, could be significant since $|\Sigma|$ might be exponential in n.

The question arises whether it is possible to encode the closure (wrt $\mathcal{PD}$) of an arbitrary set of marginal statements Σ in space polynomial in n, such that each implication query, $\Sigma \models \sigma$, would be answered in time polynomial in n? The answer is no. The following argument shows that the closure of Σ requires, on the average, $O(2^n)$ bits of storage.

Following Verma [71], we construct $O(2^{c^n})$ $(c > 1)$ *distinct* probability distributions over the variables $\{x_1, \cdots, x_n\}$, each inducing a different set of marginal independencies. This implies that, on the average, an individual set of independencies requires at least $O(c^n)$ bits of storage which is, therefore, a lower bound for the storage required for an arbitrary closure of marginal statements. Consider the following set of marginal statements:

$$B = \{I(x_1, \varnothing, C) \mid C \text{ contains exactly} \lfloor n/2 \rfloor \textit{ variables}\}.$$

We show that an arbitrary subset Σ of B defines a probabilistic model P, i.e., P satisfies Σ and no other statement in B. This would complete the proof, because there are $O(2^{2^{n/2}})$ subsets for B and therefore at least that many distinct probabilistic models. We construct P as follows: let $\sigma = I(x_1, \emptyset, \{x_{i_1} \cdots x_{i_k}\})$ be an arbitrary marginal statement in $B - \Sigma$. Construct P_σ that represents the functional dependency:

$$x_1 = \sum_{j=1}^{k} x_{i_j} \quad (mod 2)$$

where x_{i_j} $j=1..k$ are the outcomes of independent fair coins. The model P_σ satisfies every marginal statement in B except σ. Hence, the model $\otimes\{P_\sigma \mid \sigma \in B - \Sigma\}$ satisfies Σ and no other statement of B.

Theorem 2.14: The number of distinct distributions over n variables (i.e., that embody a different set of marginal independencies) is greater than $2^{2^{n/2}}$.

It is interesting to note that it is the composition axiom (2.12c) in $\mathcal{PN}$ which severely restricts the number of distinct normal distributions to exactly $2^{n(n-1)/2}$ (which is not a super exponential growth). Thus, the algorithm offered by Theorem 2.13 is optimal in space since it uses only $O(n(n-1)/2)$ bits to encode the symmetric matrix Γ.

2.6 Summary and Conclusion

Probabilistic, relational and correlational independence have been investigated; the results are summarized in Table 1, 2 and 3 respectively. Table 1 displays properties of classes of distributions versus classes of probabilistic statements. The second row shows, for example, that for the class $\mathcal{PD}^+$ of non-extreme distributions we have established a strongly complete set of axioms for marginal and saturated statements, while the existence of a complete axiomatization for unrestricted probabilistic independence remains an open question. For marginal probabilistic independence in binary distributions we have not been able to determine the Armstrong property. This is the reason that axioms (2.9) are stated to be complete and not strongly complete. A question mark means that the problem remains unresolved as of the writing of this dissertation.

	PROPERTIES	MARGINAL PROBABILISTIC INDEPENDENCE	SATURATED PROBABILISTIC INDEPENDENCE	UNRESTRICTED PROBABILISTIC INDEPENDENCE
$\mathcal{PD}$	Complete axiomatization	axioms (2.9)	axioms (2.7)	?
	Strongly complete axiomatization	axioms (2.9)	axioms (2.7)	?
	Polynomial implication algorithm	Yes	Yes	?
	Armstrong relation	Yes	Yes	Yes
$\mathcal{PB}$	Complete axiomatization	axioms (2.9)	axioms (2.7)	?
	Strongly complete axiomatization	?	?	?
	Polynomial implication algorithm	Yes	Yes	?
	Armstrong relation	?	?	No
$\mathcal{PD}^+$	Complete axiomatization	axioms (2.9)	axioms (2.8)	?
	Strongly complete axiomatization	axioms (2.9)	axioms (2.8)	?
	Polynomial implication algorithm	Yes	Yes	?
	Armstrong relation	Yes	Yes	Yes

Table 1: Probabilistic independence

Table 2 summarizes our knowledge regarding relational independence.

	PROPERTIES	MARGINAL RELATIONAL INDEPENDENCE	SATURATED RELATIONAL INDEPENDENCE	UNRESTRICTED RELATIONAL INDEPENDENCE
$\mathcal{PD}$	Complete axiomatization	axioms (2.9)	axioms (2.7)	Does not exist [45, 54]
	Strongly complete axiomatization	axioms (2.9)	axioms (2.7)	Does not exist [45, 54]
	Polynomial implication algorithm	Yes	Yes	?
	Armstrong relation	Yes	Yes	Yes
$\mathcal{PB}$	Complete axiomatization	axioms (2.9)	axioms (2.7)	?
	Strongly complete axiomatization	?	?	?
	Polynomial implication algorithm	Yes	Yes	?
	Armstrong relation	?	?	?

Table 2: Relational independence

The striking similarities between Table 1 and 2 lead us to conjecture that analogously to relational independence, probabilistic independence is not finitely axiomatizable. This does not exclude, however, the existence of an efficient implication algorithm for both probabilistic and relational independence; for example, *Z−EMVD* is a non-axiomatizable type of relational independence for which an efficient polynomial implication algorithm has been found [54].

Table 3 below summarizes the axiomatization of correlational independence. These results were established by considering conditional independence in normal distributions where probabilistic independence and correlational independence coincide.

	PROPERTIES	MARGINAL CORRELATIONAL INDEPENDENCE	SATURATED CORRELATIONAL INDEPENDENCE	UNRESTRICTED CORRELATIONAL INDEPENDENCE
$\mathcal{PN}$	Complete axiomatization	axioms (2.12)	axioms (2.8)	?
	Strongly complete axiomatization	axioms (2.12)	axioms (2.8)	?
	Polynomial implication algorithm	Yes	Yes	?
	Armstrong relation	Yes	Yes	No

Table 3: Correlational independence

Interestingly, none of these three independence relations seem to be axiomatizable and possess many properties that are not implied from the graphoid axioms [45, 67]. The graphoid axioms, nevertheless, are common to these independence relations and seem to summarize their common properties. It has been shown that all *unary* axioms, namely, axioms having one statement in their antecedents are logically implied from symmetry (1.5b), decomposition (1.c) and weak union (1.5d) [27]. Thus, the graphoid axioms entail all properties expressible as unary axioms. We conjecture that the graphoid axioms also entail all *simple* axioms, namely axioms with at most two statements in their antecedents (such as all graphoid axioms).

These results indicate that if expert's judgments of dependencies is not expressible in a limited language for dependence assertions, then checking consistency might be very hard or even undecidable. We must therefore devise a language in which consistency of the expert's input is guaranteed apriori, yet its expressive power is not too limited. Probabilistic networks offer such a language.

CHAPTER 3
The Logic of Probabilistic Networks

An important feature of probabilistic networks is that they facilitate explicit encoding of information about independencies in the domain, information that is indispensable for efficient inferencing. This chapter characterizes all independence assertions that logically follow from the topology of a network, and develops a linear time algorithm that identifies these assertions. Furthermore, it shows that any probabilistic network, directed as well as undirected, is consistent, namely, it is a perfect map of some probability distribution.

3.1 Introduction

Qualitative information about dependencies can be specified by a list of independence statements of the form $I(X, Z, Y)$ to read:"X is independent of Y, given Z". Two questions arise: first, how to obtain this information from empirical data and, second, given a partial list of independence statements, how to complete the list by inferring new statements from the ones given. We concentrate on the second question. We assume that some independencies have already been established either by statistical analysis or by conceptual understanding of the phenomena, and that our task is to infer additional independence statements without resorting to numerical calculations.

The task of finding the set of all independence statements that logically follow from a given set is particularly important in light of the realization that a variable that is independent in a given context is irrelevant for some computations. Unfortunately, this task is hard and might be undecidable (Chapter 2). Nevertheless, when a set of independencies is given in terms of a probabilistic network, then its logical closure can be computed efficiently. In the construction of Bayesian networks, for example, we encode explicitly n statements of independence, one for each variable in the domain. This set of assertions is called the *recursive basis* and is said to *generates* the network (Section 1.4). Inference algorithms such as Shachter's make use of these and additional independence assertions that are implied by this basis. Thus, to explore the limits of such inference algorithms one must characterize all independencies that are implied from the recursive basis and make sure that they are fully exploited.

In other words, to take full advantage of the Bayesian network formulation, the following two problems must be examined: Given a set of variables Y, a Bayesian network D and the task of computing $P(\boldsymbol{x} \mid Y)$, determine, without resorting to numeric calculations: (1) whether the answer to the query is sensitive to the value of a variable c, and (2) whether the answer to the query is sensitive to the *parameters* $p_c = P(c \mid \pi(c))$ stored at node c. The answer to these questions is given in terms of conditional independence: the value of c would not affect the query $P(\boldsymbol{x} \mid Y)$ if $P(\boldsymbol{x} \mid Y) = P(\boldsymbol{x} \mid Y, \boldsymbol{c})$ for all instances $\boldsymbol{x}$, Y and $\boldsymbol{c}$ of x, Y and c, or equivalently, if X and c are conditionally independent given Y, denoted by $I(x, Y, c)_{\mathcal{P}}$. Similarly, the parameters p_c stored at node c would not affect this query if x is conditionally independent of p_c given Y. The latter problem is important in particular for efficient elicitation. For example, computing P(*Brain tumor* | *Increased level of serum calcium*) in the metastatic cancer example, can be performed without consulting the parameters associated with ‘‘Comma’’ and ‘‘Headeaches’’, hence these parameters need not be known at the time of the computation.

This chapter shows that the logical closure L^* of a recursive basis L can be detected directly from the topology of the network, by merely examining the trails along which x, Y and c are connected. The results of Verma and Pearl, reported in section 1.4 provide a partial solution along this line. They have shown that *d*-separation identifies only genuine independencies and that no additional independencies can be derived through repeated applications of the graphoid axioms on L. We show that the graphoid axioms are also *complete* in the sense that they are powerful enough to derive the entire logical closure of L, as defined semantically. In other words, this criteria is maximal; it can not be strengthen to reveal more independencies. Moreover, this result holds for three independence relations: probabilistic, relational and correlational.

The theorem below summarizes this discussion.

Theorem 3.1: Let D be a dag generated by a recursive basis L drawn from a graphoid M. Let L^* be the logical closure of L, namely the set of all independencies that logically follow from L wrt probabilistic, or relational or correlational independence. Let $cl(L)$ be the set of statements derivable from L by the graphoid axioms. Let M_D be the dependency model defined by D. Then $M_D = cl(L) = L^*$.

The first equality characterizes the statements identified by *d*-separation as being exactly the statements derivable from a recursive basis L via the graphoid axioms; it guarantees that *d*-separation identifies only independencies that hold in the original graphoid. This equality is due to Verma and Pearl (see Section 1.4). The second equality assures that a dag displays all statements that logically follow from L, that is, the graphoid axioms are capable of deriving the entire logical closure of a recursive basis wrt probabilistic, relational and correlational definitions of independence. Thus, Theorem 3.1 implies that for any independence assertion not displayed by d-separation there exists a distribution that satisfies L and violates this assertion, therefore we cannot hope to improve the d-separation criterion to display more independencies. Moreover, since a statement in a dag can be verified in linear time (Section 3.3), Theorem 3.1 provides a complete polynomial inference mechanism for deriving all independence statements that logically follow from a recursive basis. A generalized version of this theorem is proven in Section 3.2. Analogous results are obtained in Section 3.4 for Markov networks.

We conclude by showing how these results can be employed as an inference mechanism. Assume an expert has identified the following independencies between variables denoted a_1 through a_5:

$$L = \{I(a_2, a_1, \varnothing),\ I(a_3, a_1, a_2),\ I(a_4, \{a_2, a_3\}, a_1),\ I(a_5, a_3, \{a_1, a_2, a_4\})\}$$

(the first statement in L is trivial). We pose two questions. First, what is the logical closure of L ? Second, in particular, does $I(a_{3,}\ \{a_1, a_5\}, a_2)$ logically follow from L ? For general lists of independencies the answer to such questions may be undecidable but, since L is a recursive basis, it defines a dag that graphically displays each and every independence of L^*. The dag is the one shown in Figure 1.1. This dag constitutes a dense representation of the logical closure of L. To answer the second question, we simply observe that $I(a_{3,}\ \{a_1, a_5\}, a_2)_{\mathcal{D}}$ holds in D.

The rest of this Chapter is organized as follows: Section 3.2 extends the analysis of independence to networks in which deterministic variables are present, namely, variables that are functionally determined by its' parents. It provides a new graphical criteria, called D-separation, which is shown to be maximal for such networks. Additionally, Section 3.2 also provides a characterization of these independencies for correlational and relational interpretations. Section 3.3 employs the declarative definition of D-separation as the basis for an efficient linear-time algorithm that identifies both relevant variables and relevant parameters. Section 3.4 provides a characterization of independence relations in Markov networks and establishes the consistency of Markov and Bayesian networks.

3.2 Networks with Deterministic Nodes

The analysis of independence in Bayesian networks assumes that the information given by the expert is summarized by a recursive basis L, containing only statements of the form $I(a, \pi(a), U(a)-\pi(a))$ where $U(a)$ are the variables preceding a in some total order of the network's variables and $\pi(a)$ is a subset of $U(a)$. Occasionly, however, we are in possession of stronger forms of independence relationships, in which case, additional statements should be read off the dag. A common example is the case of a variable that is functionally dependent on its corresponding parents in the dag (*deterministic variable,* [55]). The existence of each such variable a could be encoded in L by a statement of *global* independence $I(a, \pi(a), U-\pi(a))$ asserting that, conditioned on $\pi(a)$, a is independent of all other variables, not merely of its predecessors. The independencies that are implied by the modified basis can be read from the dag using an enhanced version of d-separation, named, D-separation.

A node that corresponds to a deterministic variable is called a *deterministic node* and is depicted by a double circle. Other nodes are called *chance nodes.* For example in Figure 3.1 below, node a_5 is a deterministic node; the value of the corresponding variable a_5 is a function of a_3 and a_4's values.

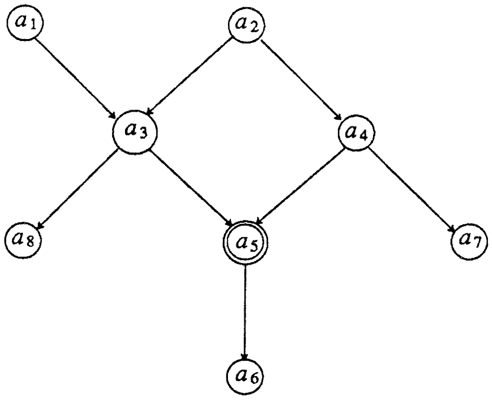


Figure 3.1

Definition: A node b is called a *tail-to-tail node with respect* to a trail t if there are two consecutive links $a \leftarrow b$ and $b \rightarrow c$ on t. A node that starts or ends a trail t is called a tail-to-tail node if it delivers an arrow.

Definition: A node a is *(functionally) determined* by Z iff $a \in Z$ or a is a deterministic node and all its parents are functionally determined by Z. If a is a deterministic node with no parents then it is functionally determined by Z. A set of nodes is determined by Z if each of its members is determined by Z.

Definition: If X, Y, and Z are three disjoint subsets of nodes in a dag D, then Z is said to D-separate X from Y, iff there is no trail t between a node in X and a node in Y along which (1) every node with converging arrows either is or has a descendant in Z, (2) every other node is outside Z, and (3) no tail-to-tail node on t is functionally determined by Z. A trail satisfying the three conditions above is said to be *active,* otherwise it is said to be *blocked*(*) (by Z).

The new criterion certifies all independencies that are revealed by d-separation plus additional ones due to condition 3 of the definition. In the dag of Figure 3.1, for example, the independence $I(a_5, \{a_3, a_4\}, a_6)_{\mathcal{D}}$ holds in D (by definition of D-separation) conveying the idea that once a node (a_5) is functionally determined, its value becomes independent of the rest of the network, independent even of its immediate successors. It should be noted that the definition of D-separation can be condensed without altering its meaning. This is shown by the following lemma.

Lemma 3.2: The following assertions are equivalent.

a. A trail t is activated by Z. Namely, t is a trail along which (a1) every node with converging arrows either is in Z or has a descendant in Z (a2), every other node is outside Z, and (a3) no tail-to-tail node (wrt t) is functionally determined by Z.

b. t is a trail along which (b1) every node with converging arrows either is in Z or has a descendant in Z and (b2) no other node is functionally determined by Z.

Proof: Let t be a trail connecting a and b that satisfies the three conditions in (a). Assume, by contradiction, that condition (b2) is violated, namely, that there exists a node a_1 on t that is not a head-to-head node, yet is determined by Z. By (a3) a_1 cannot be a tail-to-tail node. Examine a link in t that points towards a_1, say the link $a_2 \rightarrow a_1$. Since a_1 is determined by Z, either a_2 is in Z, in which case t violates condition (a2) or a_2 is determined by Z. We repeat the same argument for a_2 and obtain the chain $a_3 \rightarrow a_2 \rightarrow a_1$ where either a_3 is in Z or determined by Z. Eventually (since the number of nodes is finite), we either reach a node that is in Z, thus violating condition (a2) or we reach a tail-to-tail node that is determined by Z, in which case the trail violates condition (a3). Thus both cases contradict our assumption that t satisfies the three conditions stated in (a). The other direction, is immediate. Condition (a1) follows from (b1), and (a3) follows from (b2). Condition (a2) follows

(*) For the rest of this chapter, the terms *active* and *blocked* will refer to D-separation and not to d-separation as defined in Section 1.3.

from (b2) because a node that is not determined by Z must be outside Z. □

Note that in principle, to check whether Z D-separates X and Y, the definition requires us to examine all trails connecting a node in X and a node in Y, including trails that form loops. For example, in Figure 3.1, to check whether $X = \{a_1\}$ and $Y = \{a_8\}$ are D-separated by $Z = \{a_6\}$ would require checking trails such as $a_1, a_3, a_5, a_4, a_2, a_3, a_8$, and many others. The next lemma shows that such trails need not be examined because whenever there is an active trail with a loop there is an active *simple* trail as well, i.e. a trail that forms no cycles in the underling undirected graph. In the previous example, the trail a_1, a_3 and a_8 is the simple active trail (by $\{a_6\}$), guaranteed by Lemma 3.3.

Lemma 3.3: Let Z be a set of nodes in a dag D, and let $a, b \notin Z$ be two additional nodes of D. Then a and b are connected via an active trail (by Z) only if a and b are connected via a simple active trail (by Z).

Proof: Assume the converse holds. Let $q = (x_1, \cdots, x_n)$ be the shortest active trail (by Z) which is not simple and which connects a and b ($a \overset{\Delta}{=} x_1$ and $b \overset{\Delta}{=} x_n$). Since q has a cycle, some nodes on q are repeated. Let $x_i, x_{i+1}, \cdots, x_{i+l}$ be a portion of the trail that is repeated and let x_j be a node where $j > i+l$ such that $x_{j+m} = x_{i+m}$ for $m = 0 \cdots l$. Let $y_1, \cdots, y_k$ be the nodes between x_{i+l} and x_j on q. Let q' be the trail formed from q by removing nodes $y_1, ..., y_k$ and x_j,,..., x_{j+l} from q. The resulting trail is shorter than q and is shown to be active by Z, contradicting our selection of q. We consider two cases: $l > 0$ and $l = 0$. If $l > 0$ then every pair of adjacent links on q' is also adjacent on q. Thus, q' is active. If $l = 0$, then q' contains one new pair of links, $x_{i-1} - x_i - x_{j+1}$, that are not adjacent in q (see Figure 3.2).

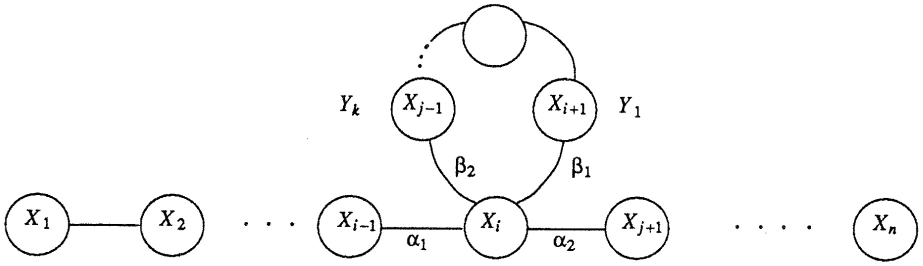


Figure 3.2

The trail q' is active only if the following three assertions hold:

1. If x_i is a head-to-head node on q', then x_i is or has a descendent in Z.
2. If x_i is not a head-to-head node on q' then $x_i \notin Z$
3. x_i is not determined by Z.

Let α_1, α_2, β_1, β_2 stand for the links $x_{i-1} - x_i$, $x_j - x_{j+1}$, $x_{j-1} - x_j$ and $x_i - x_{i+1}$.

If x_i is a head-to-head node on q', then both links α_1 and α_2 are directed towards x_i. If either links β_1 or β_2 is also directed towards x_i, then x_i is a head-to-head node on q as well, in which case, since q is active by Z, x_i must be or have a descendent in Z. If both links β_1 and β_2 are directed away from x_i then the trail $x_i, y_1, \cdots, y_k, x_j$ must contain a head-to-head node because otherwise this trail closes a cycle and D is acyclic. Let c be the closest such node to x_i. There exists a directed path from x_i to c. Node c is on q and q is active. Thus, c is or has a descendent in Z. Hence x_i has a descendent in Z.

If x_i is not a head-to-head node on q' then one of the links α_1 or α_2 is directed away from x_i. Without loss of generality assume α_1 is directed away from x_i. Consequently, x_i is not a head-to-head node on q and therefore it cannot be in Z.

If x_i were determined by Z, q would not have been active, by condition (3) of the definition of D-separation. □

The next lemma states that D-separation defines a graphoid.

Lemma 3.4: The predicate $I(X, Z, Y)_{\mathcal{D}}$ satisfies the following axioms:

- Trivial Independence:
$$I(X, Z, \emptyset) \tag{3.1a}$$
- Symmetry:
$$I(X, Z, Y) \Rightarrow I(Y, Z, X) \tag{3.1b}$$
- Decomposition:
$$I(X, Z, Y \cup W) \Rightarrow I(X, Z, Y) \tag{3.1c}$$
- Composition:
$$I(X, Z, Y) \ \& \ I(X, Z, W) \Rightarrow I(X, Z, Y \cup W) \tag{3.1d}$$
- Weak union:
$$I(X, Z, Y \cup W) \Rightarrow I(X, Z \cup W, Y) \tag{3.1e}$$
- Contraction:
$$I(X, Z, Y) \ \& \ I(X, Z \cup Y, W) \Rightarrow I(X, Z, Y \cup W) \tag{3.1f}$$

- Weak Transitivity:

$$I(X,Z,Y) \ \& \ I(X,Z \cup \{c\},Y) \Rightarrow I(X,Z,c) \text{ or } I(c,Z,Y) \qquad (3.1g)$$

We will prove axiom (3.1g). The first six axioms are proven similarly. The proof is the same as for *d*-separation [49, pp. 129]:

> **Proof:** If both X and Y are not D-separated from c in some dag, then there must be an unblocked trail between X and c and an unblocked trail between Y and c. These two trails form a trail between X and Y via c. If that trail traverses c along converging links, it should be unblocked when we instantiate c, so X and Y cannot be D-separated by c. Conversely, if the arrows meeting at c do not converge, then the trail between X and Y is unblocked when γ is uninstantiated, so X and Y cannot be marginally D-separated. □

Parallel to the discussion of Bayesian networks without deterministic nodes, we define a new basis and prove soundness and completeness of D-separation with respect to this basis.

Definition: An *enhanced basis* L drawn from a dependency model M in an ordering $a_1, \ldots, a_n$ of M's variables, is a set of n independence statements (i.e., triplets) $(a_i, \pi(a_i), W(a_i)) \in M$, $i = 1..n$, where $W(a_i)$ is either $U(a_i) - \pi(a_i) - \{a_i\}$ or $U - \pi(a_i) - \{a_i\}$, $U(a_i) = \{a_1, \ldots, a_{i-1}\}$ and $\pi(a_i) \subseteq U(a_i)$. An enhanced basis is said to *generate* a dag over n nodes where each node a_i corresponds to a variable a_i and its parents are those nodes corresponding to the variables in $\pi(a_i)$. When the i-th statement is a global independence, namely, $W(a_i) = U - \pi(a_i) - \{a_i\}$, then node a_i is a deterministic node, otherwise it is a chance node.

The next lemma is needed for establishing the soundness of D-separation; it explicates those independencies that are implied from the deterministic nodes alone. Recall that the union symbol is omitted from complicated expressions and that Xa denotes $X \cup \{a\}$.

Lemma 3.5: Let M be a graphoid over U, and L be an enhanced basis drawn from M. If Z and X are disjoint subsets of U and Z functionally determines X in the dag generated by L, then $(X, Z, U-XZ) \in M$.

> **Proof:** We prove the lemma by induction on the highest index l of an element in X as determined by the ordering of L. If the highest index is 1, then X is a singleton that has no parents in the dag. It is therefore determined by Z only if it is a deterministic

node in which case, $(X, \varnothing, U-X)$ is a member of L (and thus a member in M). By weak union, $(X, Z, U-XZ) \in M$ follows. Otherwise, Let $X = X'a$, where a has the highest index in X. Since Z determines X', by the induction hypothesis, the triplet $(X', Z, U-X'Z) \in M$.

We will show that the triplet $(a, Z, U-Za)$ is also in M and that the last two triplets together imply that $(X, Z, U-XZ) \in M$. Z determines a and $a \notin Z$, therefore Z determines the parents of a, denoted by V, and a is a deterministic node. Since all elements of V have a smaller index than a, by the induction hypothesis, $(V, Z, U-VZ) \in M$ $(\triangleq J_1)$. The triplet $(a, V, U-Va) \in L$ $(\triangleq J_2)$ because a is a deterministic node and therefore this triplet is also a member of M. Letting $W = U-VZa$, J_1 and J_2 are written as

$$(V, Z, Wa) \in M \quad \text{and} \quad (a, V, WZ) \in M .$$

The triplets $(a, Z, V) \in M$ and $(a, ZV, W) \in M$ are derived from the previous ones respectively by symmetry, decomposition, weak union. By using contraction on the latter triplets, it follows that $(a, Z, VW) \in M$. Substituting $U-VZa$ for W, we obtain that $(a, Z, U-aZ) \in M$.

It remains to be shown that $(X', Z, U-X'Z) \in M$ and $(a, Z, U-aZ) \in M$ imply that $(X, Z, U-XZ) \in M$. Letting $W = U-ZX'a$, we show that

$$(X', Z, Wa) \in M \;\; \& \;\; (a, Z, WX') \in M \;\Rightarrow\; (X'a, Z, W) \in M$$

follows from the graphoid axioms. The two triplets $(W, Z, X') \in M$ & $(W, ZX', a) \in M$ are derived from the antecedents by symmetry, decomposition and weak union. Using contraction on the resulting two triplets and then symmetry yields that $(X'a, Z, W) \in M$. Substituting $U-ZX'a$ for W and X for $X'a$, yields the desired conclusion $(X, Z, U-ZX) \in M$. □

Theorem 3.6 (soundness): If M is a graphoid, and L is any enhanced basis drawn from M, then the dag D generated by L is an I-map of M.

Proof: (*) Induct on the number of elements in the graphoid. For graphoids of one variable it is obvious that the single node dag generated is an I-map. Suppose for graphoids with fewer than k elements that the dag generated is an I-map. Let M have k elements, let u be the last element in the ordering of L, let $M[u]$ be the graphoid formed by removing u and all triplets involving u from M and let $D[u]$ be the dag formed by removing u and all its incident links from D. Additionally, let $L[u]$ be the set of triplets formed from L by removing the last triplet and deleting the element u

(*) This proof is essentially taken from [72]. The main change is in case 2 where lemma 3.5 is applied.

from the remaining triplets, namely $L[u]$ is equal to $\{(j, B, R-u) \mid j \neq u, (j, B, R) \in L\}$. The set $L[u]$ is an enhanced basis of $M[u]$ and it generates the dag $D[u]$. Thus, since $M[u]$ has $k-1$ elements, by the induction hypothesis, $D[u]$ is an I-map of it. Let M_D be the dependency model corresponding to the dag D, and $M_{D[u]}$ correspond to $D[u]$, (i.e. M_D contains all triplets (X, Z, Y) for which X and Y are D-separated by Z in D). Each triplet T of M_D falls into one of three categories: (1) u does not appear in T, (2) u appears on the first or third entry of T or (3) u appears in the second entry of T. These will be treated separately as cases 1, 2 and 3, respectively. For each case we will show that $T \in M_D$ implies that $T \in M$, thus proving that D is an I-map of M.

case-1: If u does not appear in T then T must be (X, Z, Y) with X, Y and Z three disjoint subsets of elements, none of which contain u. Since T is in M_D it must also be in $M_{D[u]}$ for if it were not then there would be an active trail (by Z) in $D[u]$ between a node in X and a node in Y. But if this trail were active in $D[u]$ then it would also be active in any dag containing $D[u]$ as a subgraph, specifically this trail would have remained active in D. By the induction hypothesis, $M_{D[u]}$ is a subset of $M[u]$, thus T must be an element of $M[u]$. $M[u]$ is a subset of M, so T is in M.

case-2: If u appears in the first entry of the triplet, then $T = (Xu, Z, Y)$ with X, Y and Z three disjoint subsets of elements, none of which contain u. Let (u, B, R) be the last triplet in L, B_X, B_Y, B_Z and B_0 be a partitioning of B and R_X, R_Y, R_Z and R_0 be a partitioning of R such that $X = B_X \cup R_X$, $Y = B_Y \cup R_Y$ and $Z = B_Z \cup R_Z$ as in Figure 3.3. We first show that $(Y, ZXB_0, u) \in M$. Then we will show that $(Y, Z, XB_0) \in M$. Since M is a graphoid containing these two statements, it will follow by contraction that $(Y, Z, XB_0u) \in M$ and by decomposition and symmetry that $T = (Xu, Z, Y) \in M$. This will complete the proof of case 2 because if u appears in the third entry of T, namely, $T = (Y, Z, Xu)$ then (Xu, Z, Y) be a member of M_D which would imply that $(Xu, Z, Y) \in M$ and since M is closed under symmetry T would be a member of M as well.

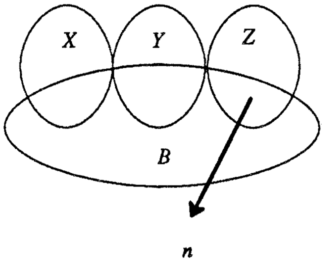


Figure 3.3

We now show that $(Y, ZXB_0, u) \in M$. Consider the set B_Y, any node in this set is directly linked to u, thus in order for Y to be D-separated from u given Z, B_Y must be determined by Z in D (or be empty, in which case, by our definitions, it is determined by Z). By Lemma 3.5, $(B_Y, Z, \dot{U} - B_Y Z) \in M$. Using weak union and decomposition, it follows that (B_Y, ZXB_0, u) is in M. The triplet (u, B, R) in L implies (by decomposition and weak union) that $(u, XZB_0B_Y, R_Y) \in M$. The last triplet together with (B_Y, ZXB_0, u) imply (by symmetry and contraction) that $(Y, ZXB_0, u) \in M$.

It remains to show that $(Y, Z, XB_0) \in M$. The triplet (Y, Z, B_0) must belong to M_D since otherwise there would have been an active trail between a node in Y and a node in B_0 which could have been augmented to form an active trail (by Z) between Y to u, by using the link that connects any element in B_0 to u (pointing to u). This would contradict the assumption that $(u, Z, Y) \in M_D$, as implied by decomposition from $(Xu, Z, Y) \in M_D$. Thus $(Y, Z, B_0) \in M_D$. $(Y, Z, X) \in M_D$ because it is implied by decomposition from the fact that (Xu, Z, Y) is in M_D. By the definition of D-separation two sets are D-separated iff each of their individual elements is D-separated. Therefore, $(Y, Z, X) \in M$ and $(Y, Z, B_0) \in M$ imply that (Y, Z, XB_0) must also be in M_D. The last triplet does not contain u, thus by the argument of case-1, $(Y, Z, XB_0) \in M$.

case-3: If u appears in the second entry then $T \in M_D$ has the form (X, Zu, Y). The triplet (X, Z, Y) must be a member of M_D as well for if there were an active trail (by Z) between a node in X and a node in Y, this trail would have remained activated by Zu because u is a sink on that trail. This would contradict our assumption that $(X, Zu, Y) \in M_D$. The triplets $(X, Z, Y) \in M_D$ and $(X, Zu, Y) \in M_D$ imply by the definition of D-separation that either $(X, Z, u) \in M_D$ or $(u, Z, Y) \in M_D$ (Lemma 3.4). By definition of D-separation, two sets are D-separated iff each of their individual elements is D-separated. Therefore, $(X, Z, Y) \in M_D$ and the disjunction above imply that either $(X, Z, Yu) \in M_D$ or $(Xu, Z, Y) \in M_D$. By the argument of case 2, it follows that either $(X, Z, Yu) \in M$ or $(Xu, Z, Y) \in M$. In both case, it follows by weak union and symmetry that $(X, Zu, Y) \in M$. □

Theorem 3.7 (closure): If L is an enhanced basis drawn from an arbitrary dependency model M, the dag dependency model M_D generated from L is a perfect map of the closure $cl(L)$ of L under the graphoid axioms. In other words, a triplet belongs to M_D if and only if it can be derived from the triplets of L using the graphoid axioms (1.5).

Proof: By Theorem 3.6, $M_D \subseteq cl(L)$. It remains to show that $cl(L) \subseteq M_D$. We will show, instead, that $L \subseteq M_D$. This will imply that $cl(L) \subseteq cl(M_D)$ and since every dag dependency model M_D satisfies the graphoid axioms (Lemma 3.4), it must be the case that $cl(M_D) = M_D$ which will complete the proof. Let (u, B, R) be a triplet in L. There are two cases: R does not contain any successor of u, in which case u is a chance node, or R contains all of u's successors, in which case u is a deterministic node. If u is a deterministic node, then its parents B, D-separate it from any other node, thus $(u, B, R) \in M_D$. If u is a chance node, then u is D-separated from R given B in the dag (i.e., $(u, B, R) \in M_D$), for if not, there would be a trail from a node in R to u which is active given B. But since every link into u is from B the trail must lead out of u into some node which was placed after u. Since, in the case of a chance node, every node in R was placed before u, this trail must contain a head-to-head node that was placed after u. But this trail cannot be activated by B since B contains no nodes placed after u, and thus, B would D-separate u from R in the dag. □

The completeness proof for D-separation requires the following lemma.

Lemma 3.8: For every dag D and a triplet $T = (a, Z, b) \notin M_D$, there exists a dag D' with the following properties:

1. $D' = (E', V)$ is a subgraph of $D = (E, V)$ i.e., $E' \subseteq E$.
2. $(a, Z, b) \notin M_{D'}$
3. The links of D' consist exclusively of the following three sets:
    a. A trail q between a and b.
    b. A single directed path p_i from every head-to-head node h_i on q to a distinct member z_i of Z. The paths p_j's do not share any node with each other and each p_i intersects q only at h_i.
    c. For each functional tail-to-tail node t_i on q, D' contains a directed path r_i from some chance node l_i to t_i such that l_i is the only chance node on r_i and the entire path is disjoint of Z. The paths r_j's do not share any node with each other or with any p_i and each path r_i intersects q only at node t_i.

A dag satisfying the three conditions above is called an *ab-trail dag*.

Proof: We first construct the dag D' and then prove it satisfies the requirements. Let q be an active trail (by Z) between a and b with a minimum number of head-to-head nodes denoted, from a to b, $h_1, h_2, \ldots, h_k$. Such a trail exists because $T \notin M_D$. Due to Lemma 3.3 we can assume q is a simple trail. Let z_i be the closest (wrt path length) descendent of h_i in Z and let p_i be a directed path from h_i to z_i (if $h_i \in Z$ then $z_i = h_i$). Let $t_1 \ldots t_l$ be all deterministic tail-to-tail nodes on q. Let l_i be the closest chance node, not in Z, that is an ancestor of t_i such that q_i, a directed path connecting l_i and t_i is entirely disjoint of Z. The paths p_i's exist because the trail q is active only if every h-h node on it is or has a descendent in Z. The paths q_i's exist in D because otherwise t_i would have been functionally determined by Z and the trail q would not have been active. Let $D' = (E', V)$ where E' consist exclusively of the links contained in q, p_i's and q_i's (e.g., Figure 3.4).

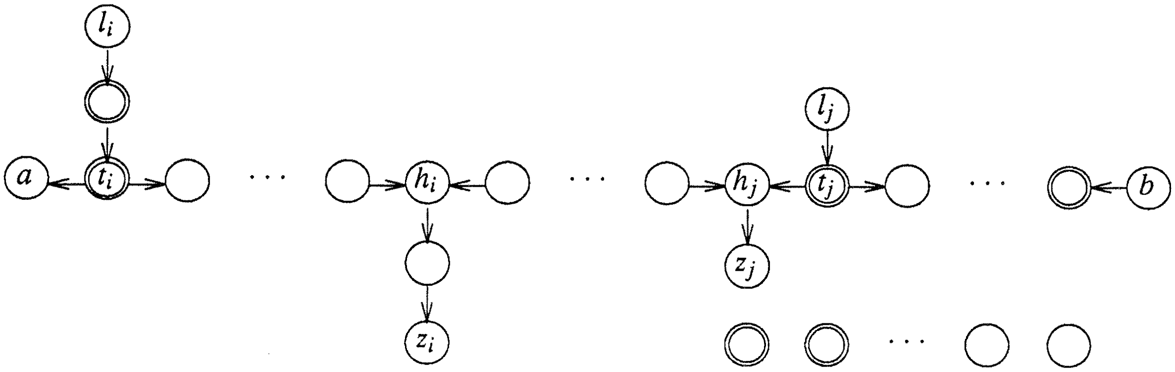


Figure 3.4

By our construction, D' satisfies conditions 1,2 and 3.a. Next we prove it satisfies requirement 3.b. First we claim that the path p_i's are distinct. Assume, by contradiction, that there are two paths p_i and p_j $(i < j)$ with a common node c (Figure 3.5). Under this assumption, we find an active trail between a and b that has fewer head-to-head nodes then q, contradicting the minimality of the latter. If c, the common node, is neither h_i nor h_j then the trail (a, h_i, c, h_j, b) is an active trail (by Z); Each of its head-to-head nodes is or has a descendent in Z because it is either c or a head-to-head node of q and every node that had been added is not determined by z. Every other node d lies either on the active trail q and therefore is not determined by Z (Lemma 3.2) or it lies on either p_i or p_j. In either of the last two cases, since z_i or z_j are the closest descendants of h_i or h_j respectively, d must be outside Z. The resulting active trail contradicts the minimality of q since both h_i and h_j are no longer head-to-head nodes while c is the only newly introduced head-to-head node. If $c = h_j$ then the trail (a, h_i, c, h_j, b) shrinks to be (a, h_i, c, b), which, using similar arguments, has fewer head-to-head nodes than q and is activated by Z (the case $c = h_i$ is similar).

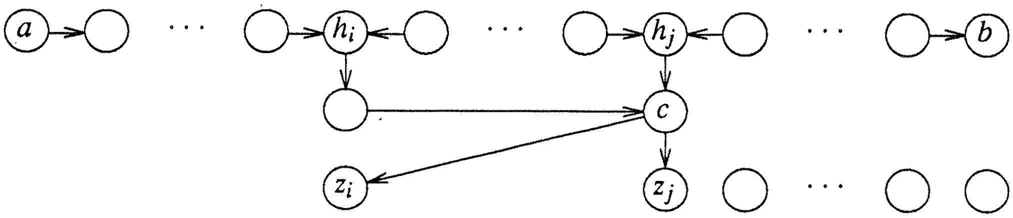


Figure 3.5

We complete the proof of 3.b by showing that each path p_i intersects q only at node h_i. Assume, by contradiction, that p_i and q have in common a node c other then h_i and assume that it lies between h_i and b (the case were c lies between h_i and a is similar) (Figure 3.6). It has been shown that p_i is distinct from all other p_j's therefore, in particular, node c is not a head-to-head node on q. Thus, c cannot belong to Z because otherwise q were blocked by c and thus would not have been active. Hence, the trail $q' = (a, h_{i,} c, b)$ is activated by Z. The trail q' contradicts the minimality of q because h_i is no longer a head-to-head node on q' while no new head-to-head nodes are introduced.

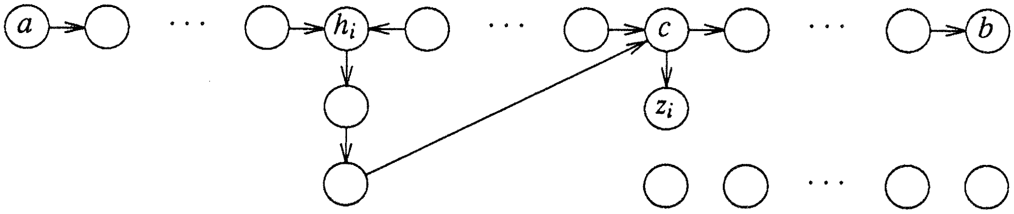


Figure 3.6

To prove 3.c we use similar arguments. Assume paths r_i and r_j have a common node c (e.g, Figure 3.7). Then the trail (a, t_i, c, t_j, b), depicted in Figure 3.7, is an active trail that contains fewer head-to-head nodes than q because the fragment of q between any two tail-to-tail nodes t_i's must contain a head-to-head node while the new bypass does not contain any. The new trail is active because no node on the bypass is determined by Z. Thus the new trail is active and therefore contradicts the minimality of q.

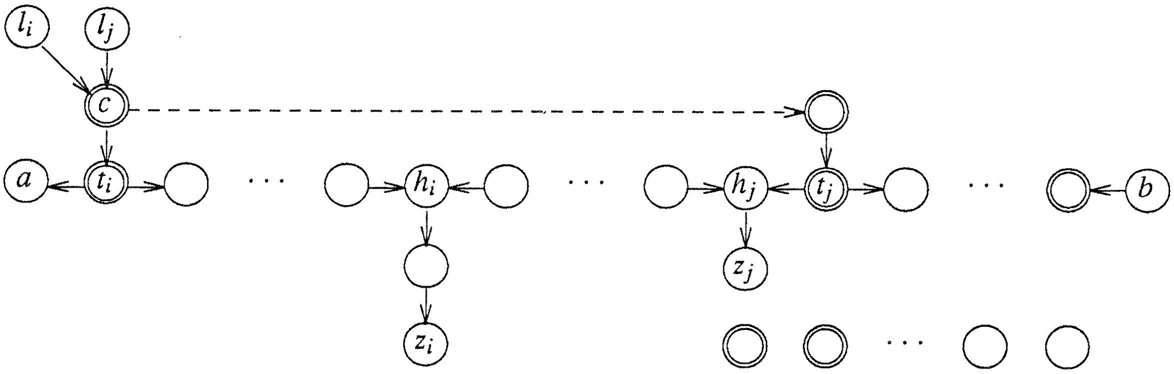


Figure 3.7

If a node c is shared by r_i and p_j then the trail (a, t_i, c, h_j, b) (e.g, Figure 3.8) contradicts the minimality of q; It contains fewer head-to-head nodes than q because h_j is no longer a head-to-head node and no new head-to-head nodes are added. It is active since neither of the nodes on the bypass is in Z or is determined by Z.

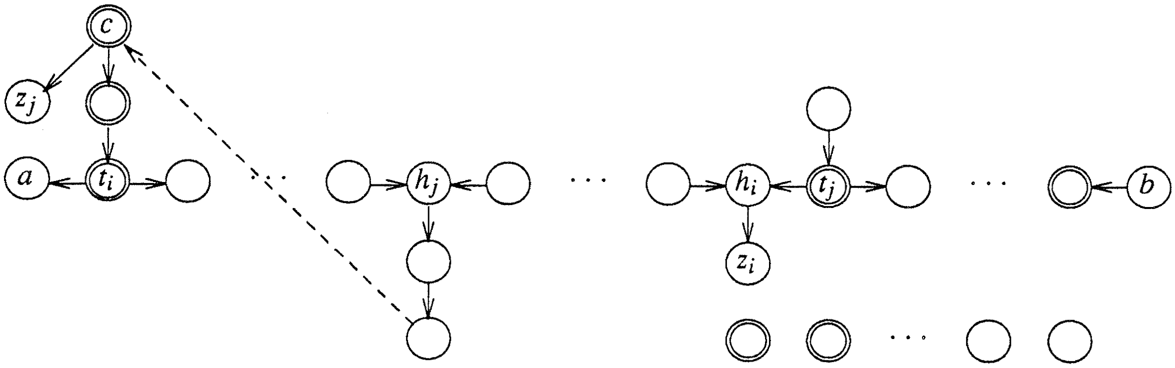


Figure 3.8

If c is shared by r_i and q then the new trail (a, t_i, c, b) (e.g, Figure 3.9) contradicts the minimality of q; It contains fewer head-to-head nodes than q because no new head-to-head nodes are added while the fragment of q between t_i and c must contain a head-to head node for otherwise D would have contained a circle. The new trail is active since no node on the bypass is in Z or is determined by Z.

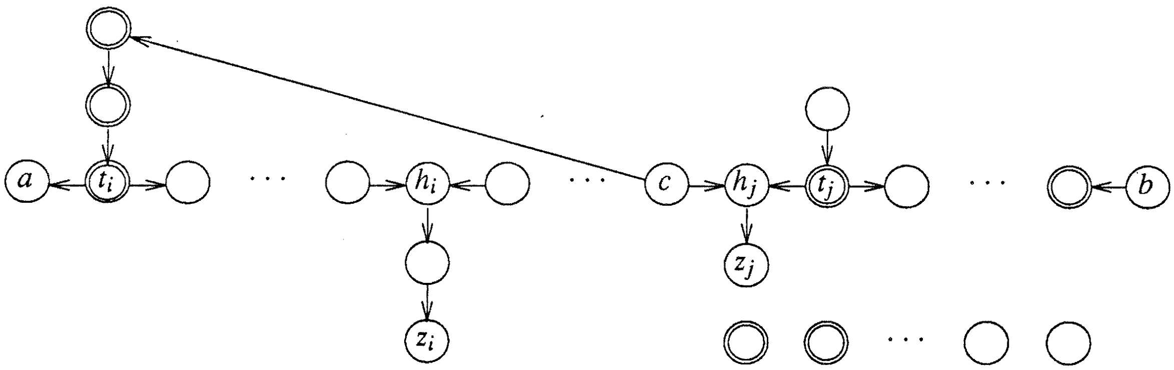


Figure 3.9

Thus D' satisfies all the requirements of the lemma. □

Theorem 3.7 states that $M_D = cl(L)$. Thus, $M_D \subseteq L^*$, the logical closure of L wrt probabilistic, relational and correlational independence. Theorem 3.9 below states the converse for probabilistic dependency models, namely, every statement in L^* wrt probabilistic independence is an independency in M_D. Similar results for correlational and relational dependency models are given as corollaries; they follow from the proof of Theorem 3.9.

Theorem 3.9 (completeness) : Let D be a dag generated by an enhanced basis L drawn from a probabilistic dependency model M_P. Then $M_D \subseteq L^*$ wrt probabilistic independence.

Proof: Let $T = (X, Z, Y)$ be an arbitrary triplet not in M_D (we assume that $XZY \subseteq U$ and that U is finite). We construct a distribution P_T whose dependency model P_T (*) contains all triplets of L and does not contain T. This distribution precludes T from being a semantic consequence of L and therefore, as the theorem claims, every semantic consequence of L must be a member in M_D.

The triplet $(X, Z, Y) \notin M_D$. Hence the definition of D-separation guarantees the existent of an active trail between a node a in X and a node b in Y that is not D-separated by Z. Constructing a distribution P_T that does not contain the triplet (a, Z, b), denoted T', guarantees also that $(X, Z, Y) \notin P_T$ because any distribution

(*) The symbol P_T is overloaded- sometimes it denotes a distribution and sometimes it denotes the dependency model defined by that distribution. The meaning will be clear from the context.

that renders X and Y conditionally independent must render each of their individual variables independent as well (decomposition (3.1c)). The triplet $(a, Z, b) \notin M_D$. Hence by Lemma 3.8, there exists an ab-trail dag D' (Figure 3.4). We will construct a distribution P_T whose dependency model contains all triplets of $M_{D'}$(i.e., $M_{D'} \subseteq P_T$) and which does not contain T. This will complete the proof; By property (i) of Lemma 3.8, $M_D \subseteq M_{D'}$. By the definition of D-separation, $L \subseteq M_D$. Thus $L \subseteq P_T$, as required by the theorem.

P_T is defined as follows: Each chance node with no parents corresponds to an independent fair binary coin. Every other node corresponds to a variable that is the sum modulo 2 of the variables corresponding to its parents. A deterministic node with no parents (a degenerate configuration) corresponds to a binary variable whose value is known with certainty. It remains to show that P_T satisfies the requirements. Variables a and b are conditionally dependent given Z in P_T because constraining a and Z to some specific values determines a value for b via the single trail q that connects them in D'. It remains to show that $M_{D'} \subseteq P_T$. Let L' be an enhanced basis that generates D' in the following ordering of the nodes of D': All nodes which have no parents appear first in the ordering followed by the rest of the nodes in any order compatible with the partial order defined by D' (e.g., a must precede b if $a \rightarrow b$ is a link in D'). The basis L' is contained in P_T because all chance nodes with no parents correspond to mutually independent variables and every other variable in P_T is a function of the variables corresponding to its parents and therefore, it must be independent from all its other predecessors and successors in the ordering of L'. Thus $L' \subseteq P_T$. Taking the closure under the graphoid axioms on both sides yields $cl(L') \subseteq cl(P_T)$. However, $P_T = cl(P_T)$ because P_T is a graphoid and $cl(L') = M_{D'}$ (by Theorem 3.7). Thus, $M_{D'} \subseteq P_T$. □

Dags have been used also as a representation scheme for structural equations [6, 17, 74, 75]. Each node represents a variable that is the linear combination of the variables corresponding to its parents, and a term representing noise. The noise sources are assumed to be independent, normally distributed, and have zero means and non-zero variances. Thus, the variable corresponding to node x is given by

$$x = a_1 z_1 + \cdots a_k z_k + z, \tag{3.2}$$

where $z_1 \cdots z_k$ are the variables corresponding to the parents of x, and z is the noise term. In this interpretation, two variables are ''independent'' given a set of variables Z, denoted by $I(a, Z, b)_C$ iff the correlation between a and b vanishes when the influence of Z is removed (see section 3.2, for the precise definition). Each structural equation *corresponds* to one independence statement. For example, Eq. (3.2) asserts that $I(x, \{z_1 \dots z_k\}, U(x) - \{z_1, \dots, z_k\})_c$ where $U(x)$ are the variables preceding x. The soundness of D-separation in this representation follows because I_C, being identical to conditional independence for normal distributions, must satisfy the graphoid axioms. Completeness is shown by the following corollary of the proof of Theorem 3.9. A similar, but more restricted, completeness theorem for $I(a, Z, b)_C$, where Z is a singleton, is given in [28]. We remark that no functional dependencies exist in this interpretation of dags, because all noise terms have non-zero variances, thus D-separation coincides with d-separation.

Corollary 3.10 (completeness): Let D be a dag generated by a set of structural equations and let L be the corresponding enhanced basis. Then, $M_D \subseteq L^*$ wrt correlational independence.

Proof: In normal distributions, two variables a and b are conditionally independent given Z iff the partial correlation $\rho_{ab.Z} = 0$ [13]. Therefore, it suffice to construct a normal distribution N_T with the same properties P_T had in the proof of Theorem 3.9. We define N_T as follows: Each chance node with no parents corresponds to the outcome of an independent normal variable. Every other node corresponds to a variable that is a (noisy) sum of the variables corresponding to its parents. Deterministic nodes are not present. N_T is a multivalued normal distribution. The proof that N_T fulfills the requirements is the same as in the proof of Theorem 3.9. □

The following corollary shows that D-separation is also complete when dags represent relational dependency models.

Corollary 3.11 (completeness): Let D be a dag generated by an enhanced basis L. Then, $M_D \subseteq L^*$ wrt relational independence.

Proof: Let P_T be the distribution constructed in the proof of Theorem 3.9, and let R_T be the relation defined to be the set of tuples for which P_T has a positive probability. R_T (viewed as a dependency model) contains P_T. P_T contains $M_{D'}$ where D' is an ab-trail dag constructed in the proof of Theorem 3.9. Thus R_T contains $M_{D'}$. Variables a and b are conditionally dependent given Z in R_T because constraining a and Z to some specific values determines a value for b via the single trail q that connects them in D'. □

Hunter [36] provides similar soundness and completeness results, with respect to an independence relation based on Spohn's ordinal conditional functions [65].

3.3 A Linear Time Algorithm for Identifying Independence

Conditional independence assertions encoded in Bayesian networks can be used for identifying what information and which parameters may be needed for performing a given computation. The analysis of the previous section guarantees that the D-separation criteria could identify the maximal set of variables that are independent of a set of variables X, given another set Z, without resorting to numerical calculations. However, it does not provide an efficient algorithm to do so. The algorithm we develop in this section is a variant of the well known Breath First Search algorithm; it finds all nodes reachable from X through an active trail (by Z), hence the maximal set of nodes Y satisfying $I(X, Z, Y)_{\mathcal{D}}$. This task can be viewed as an instance of a more general problem of finding a path in a directed graph for which some specified pairs of links are restricted not to appear consecutively. In this context, D-separation serves to specify such restrictions. For example, two links $u \rightarrow v$, $v \leftarrow w$ cannot appear consecutively in an active trail unless $v \in Z$ or v has a descendent in Z. The following notations are employed: $D = (V, E)$ is a directed graph, not necessarily acyclic, where V is a set of nodes, $E \subseteq V \times V$ is the set of directed links and $F \subseteq E \times E$ is a list of pairs of adjacent links that cannot appear consecutively (F-connotes fail). We say that an ordered pair of links (e_1, e_2) is *legal* iff $(e_1, e_2) \notin F$, and that a path is *legal* iff every pair of adjacent links on it is legal. We emphasize that by ''path'' we mean a directed path, not a trail.

First we devise a simple algorithm for the following problem: Given a finite directed graph $D = (V, E)$, a subset $F \subseteq E \times E$ of illegal pairs of links, and a set of nodes X, find all nodes reachable from X via a legal path in D. The algorithm and its proof of correctness are a slight modification of those found in [18].

Algorithm 1

Input: A directed graph $D = (V, E)$, a set of illegal pairs of links F and a set of nodes X.

Output: A labeling of the nodes such that a node is labeled with R (connoting "reachable") iff it is reachable from X via a legal path.

(i) Add a new node s to V and for each $a \in X$, add the link $s \rightarrow a$ to E and label them with 1. Label s and all $a \in X$ with R. Label all other nodes with "undefined."

(ii) $i := 1$

(iii) Find all unlabeled links $v \rightarrow w$ adjacent to at least one link $u \rightarrow v$ labeled i, such that $(u \rightarrow v, v \rightarrow w)$ is a legal pair. If no such link exists, stop.

(iv) Label each link $v \rightarrow w$ found in Step (iii) with $i+1$ and the corresponding node w with R.

(v) $i := i+1$, Goto Step (iii).

The main difference between this algorithm and BFS, a change which has been proposed by Gafni, (*) is the traversal of the graph according to a labeling of the links and not according to a labeling of nodes. This change is essential as the example in Figure 3.10 shows. Let F consist of one pair (α, γ). The path from a to c through links α, β and γ is legal while the path not traversing β is not legal because $(\alpha, \gamma) \in F$. BFS with node labeling would not reveal the legal path (a, b, b, c) connecting nodes a and c, because it visits every node, in particular b, only once.

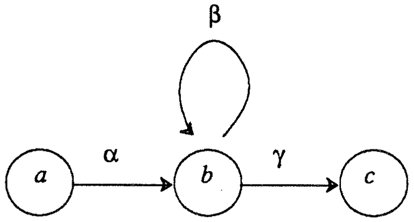


Figure 3.10

Lemma 3.12: Algorithm 1 labels with R all nodes that are reachable from s (and thus from X) via a legal path, and only those nodes.

Proof: First, we show that if a node a_l is labeled with R, then there exists a legal path from s to a_l. Let $a_{l-1} \rightarrow a_l$ be a link through which a_l has been labeled. We induct on the label l of the link $a_{l-1} \rightarrow a_l$. If $l = 1$ then $a_l \in X$ and is therefore reachable from s. If $l > 1$, then by step (iii), there exists a link $a_{l-2} \rightarrow a_{l-1}$ labeled with $l-1$ such that $(a_{l-2} \rightarrow a_{l-1}, a_{l-1} \rightarrow a_l)$ is a legal pair. Repeatedly applying this argu-

(*) personal communication

ment for $i = l...2$ yields a legal path $a_0 \rightarrow a_1 \rightarrow ... a_l$, where $a_0 \rightarrow a_l$ is labeled with 1. However, the only links labeled 1 emanate from s, hence the above path is the required legal path from s to a_l.

It remains to show that each node that is reachable from s via a legal path is labeled with R. Instead, we show that every link $a \rightarrow a_m$ that is reachable from s via a legal path (i.e., it participates in a legal path emanating from s) will eventually be labeled by the algorithm. The latter claim is stronger than the former because for every reachable node a_m there exists a reachable link $a \rightarrow a_m$ and by Step (iv), whenever $a \rightarrow a_m$ is labeled with some integer, a_m is labeled with R. We continue by contradiction. Let $l_m = a_{m-1} \rightarrow a_m$ be the closest link to s via a legal path that remains unlabeled. Let $p = s \rightarrow a_1 \rightarrow ...a_{m-1} \rightarrow a_m$ be the legal path emanating from s and terminating with the link l_m. The portion of this path that reaches the link $l_{m-1} = a_{m-2} \rightarrow a_{m-1}$ is shorter than p. Thus, by the induction hypothesis, l_{m-1} is labeled by the algorithm. Hence, the link l_m is labeled as well (by the next application of step (iv)), contradicting our assumption that it remains unlabeled. □

The complexity of Algorithm 1 for a general F is $O(|E| \cdot |V|)$. In the worst case, each of the $|V|$ nodes might be reached from $|V| - 1$ entry points and, for each entry, the remaining links may need to be examined afresh for reachability (For example, link c in the example of Figure 3.10 is examined twice). Thus, in the worst case, a link may be examined $|V-2|$ times before it is labeled, which leads to an $O(|E| \cdot |V|)$ complexity. However, for the special case where F is induced by the D-separation condition, we shall later see that each link is examined only a constant number of times, therefore the complexity reduces to $O(|E|)$.

Next we solve the problem of identifying the set of nodes that are D-separated from X by Z. For this aim, we will construct a directed graph D' with a set of legal pairs such that a node v is reachable from X via an **active trail** (by Z) in D iff v is reachable from X via a **legal path** in D'. Algorithm 1 is then applied to solve the latter problem. The following observations are the basis of our algorithm. First, any link on a trail can be traversed both ways. Therefore, to ensure that every active trail in D corresponds to a legal (directed) path, D' must consist of all links of D in their forward and reverse direction. Second, constructing a table that for each node that indicates whether it is determined by or has a descendent in Z, would facilitate a constant-time test for legal pairs in D'.

Algorithm 2

Input: A Bayesian network $D = (V, E)$ and two disjoint sets of nodes X and Z.

Data Structure: A list of incoming links (in-list) for each node $v \in V$.

Output: A set of nodes Y where $Y = \{b \mid I(X, Z, b)_{\mathcal{D}}\}$.

(i) Construct the following tables:

$$\text{determined}[v] = \begin{cases} \text{true} & \text{if } v \text{ is determined by Z} \\ \text{false} & \text{otherwise} \end{cases}$$

$$\text{descendent}[v] = \begin{cases} \text{true} & \text{if } v \text{ is or has a descendent in Z} \\ \text{false} & \text{otherwise} \end{cases}$$

(ii) Construct a directed graph $D' = (V, E')$ where
$E' = E \cup \{(u \rightarrow v) \mid (v \rightarrow u) \in E\}$

(iii) Using Algorithm 1, find the set of all nodes Y' which have a legal path from X in D', where a pair of links $(u \rightarrow v, v \rightarrow w)$ is legal iff $u \neq w$ and either (1) v is a head-to-head node on the trail u—v—w in D and descendent[v] = true or (2) v is not a head-to-head node on the trail u—v—w in D and determined[v] = false. (*)

(iv) $Y = V - (Y' \cup X \cup Z)$

Return (Y).

The correctness of this algorithm is established by the following theorem.

Theorem 3.13: The set Y returned by the algorithm is exactly $\{b \mid I(X, Z, b)_{\mathcal{D}}\}$.

Proof: The set Y' constructed in Step (iii) contains all nodes reachable from X via a legal path in D'. For any two nodes $a_0 \in X$ and $b \notin X \cup Z$, if $a_0 - a_1 \ldots b$ is an active trail (by Z) in D, then the directed path $a_0 \rightarrow a_1 \rightarrow \ldots b$ is a legal path in D', and vice versa. Thus Y' contains all nodes not in $X \cup Z$ that are reachable from X via an active trail (by Z) in D. By definition of D-separation, $I(X, Z, b)_{\mathcal{D}}$ holds iff $b \notin X \cup Z$ and b is not reachable from X (by an active trail by Z). Thus, $Y = V - (Y' \cup X \cup Z)$ is exactly the set $\{b \mid I(X, Z, b)_{\mathcal{D}}\}$. □

(*) Note that this step uses the alternative definition of D-separation, the one offered by Lemma 3.2.

Next, we show that the complexity of the algorithm is $O(|E|)$. The construction of *descendent*$[v]$ is implemented as follows: Initially mark all nodes of Z with true. Follow the incoming links of the nodes in Z to their parents and then to their parents and so on. This way, each link is examined at most once, hence this construction requires $O(|E|)$ operations. The construction of *determined*$[v]$ is similar and requires the same complexity. Step (ii) of Algorithm 2 requires the construction of a list for each node that specifies all the links that emanate from v in D (out-list). The in-list and the out-list completely and explicitly specify the topology of D'. This step also requires $O(|E|)$ steps. Using the two lists the task of finding a legal pair in step (iii) of Algorithm 1 requires only constant time; if $e_i = u \rightarrow v$ is labeled i then depending upon the direction of $u - v$ in D and whether v is determined by or has a descendent in Z, either all links of the out-list of v, or all links of the in-list of v, or both are selected. Thus, a constant number of operations per encountered link is performed. Hence, Step (iii) requires no more than $O(|E|)$ operation which is therefore the upper bound (assuming $|E| \geq |V|$) for the entire algorithm.

The above algorithm can also be employed to verify whether a specific statement $I(X, Z, Y)_{\mathcal{D}}$ holds in a dag D. Simply find the set $Y_{\max}$ of all nodes that are D-separated from X given Z and observe that $I(X, Z, Y)_{\mathcal{D}}$ holds in D iff $Y \subseteq Y_{\max}$. In fact, for this task, Algorithm 2 can slightly be improved by forcing termination once the condition $Y \subseteq Y_{\max}$ has been detected. Recently, another algorithm for the same task (for networks without deterministic nodes) has been proposed [40]. The algorithm consists of the following steps. First, form a dag D' by removing from D all nodes which are not ancestors of any node in $X \cup Y \cup Z$ (and removing their incident links). Second, form an undirected graph G, called the *moral graph,* by stripping the directionality of the links of D' and connecting any two nodes that have a common child in D' which is or has a descendent in Z. $I(X, Z, Y)_{\mathcal{D}}$ holds by the definition of *d*-separation iff all undirected paths between X and Y in G are intercepted by Z.

The complexity of the moral graph algorithm is $O(|V|^2)$ because the moral graph G may contain up to $|V|^2$ links, and, so, checking separation in G might require, in the worst case, $O(|V|^2)$ steps. Our algorithm requires O(|E|) steps, which yields significant gain in sparse graphs, namely, those having $|E| = O(|V|)$. If the maximal number of parents of each node is bounded by a constant, then the two algorithms achieve the same asymptotic behavior i.e, linear in $|E|$. On the other hand, the moral graph algorithm is conceptually simpler to communicate and, for small graphs, might offer computational advantages as well. (*) When the task is to find *all* nodes *d*-separated from X by Z, then a brute force application of the moral graph algorithm requires $O(|V|^3)$ steps, because for each node not in $X \cup Z$ the algorithm must construct a new moral graph. For this task, our algorithm offers a considerable improvement.

An area where Bayesian networks have been used extensively is decision analysis; an analyst elicits information from an expert about a decision problem, formulates the appropriate network and then, by an automated sequence of graphical and probabilistic manipulations an optimal decision is obtained [35, 44, 55]. When such a network is constructed it is important to determine what information is needed to answer a given query $P(x \mid Z)$ because eliciting irrelevant parameters may be a waste of effort [55]. Assuming that each node a_i stores the conditional distribution $P(a_i \mid \pi(a_i))$, the task is to identify the set Q of chance nodes that must be consulted in the process of computing $P(x \mid Z)$ or, alternatively, the set of chance nodes that can be assigned arbitrary conditional distributions without affecting the quantity $P(x \mid Z)$. The required set can also be identified by the D-separation criterion. We represent the parameters p_i of the distribution $P(a_i \mid \pi(a_i))$ as value in the domain of a dummy parent π_i of node a_i. This is clearly a legitimate representation complying with the format of Eq. (1), since for every chance node a_i, $P(a_i \mid \pi(a_i))$ can also be written as $P(a_i \mid \pi(a_i), p_i)$, so π_i can be regarded as a parent of a_i [62]. From Theorems 3.6 and 3.7, all dummy nodes that are D-separated from X by Z represent variables that are conditionally independent of X given Z and so, the information stored in these nodes can be ignored. Thus, the information required to compute $P(x \mid Z)$ resides in the set of dummy nodes which are not D-separated from X given Z. Moreover, the completeness of D-separation further implies that Q is minimal; no node in Q can be exempted from processing on purely topological grounds (i.e., without considering the numerical values of the probabilities involved). The algorithm below summarizes these considerations:

Algorithm 3

Input: A Bayesian network, two sets of nodes X and Z.

Output: A set of nodes Q that contains sufficient information to compute $P(x \mid Z)$

(i) Construct a dag D' by augmenting D with a dummy node v' for every chance node v in D and adding a link $v' \rightarrow v$.

(*) The average complexity of Algorithm 2 can be reduced by adapting the first step of the moral graph algorithm, but the worst case complexity would not be improved.

(ii) Use Algorithm 2 to compute the set Y' of nodes not D-separated from X by Z.

(iii) Q is the set of all dummy nodes v' that are included in Y'.

Note that the algorithm adds dummy nodes only to chance nodes. Hence, the algorithm should not be used to detect those functional relationships that could be ignored; it identifies, however, the set of probabilistic parameters that are sufficient for a computation of $P(x \mid Z)$. In order to identify the functional relationship that could be ignored, a more elaborated algorithm is required. This subtle point is illustrated in Shachter's example (8.d) and his algorithm addresses this task [56].

We conclude with an example. Consider the network D of Figure 3.11(a) and a query $P(a_3)$.

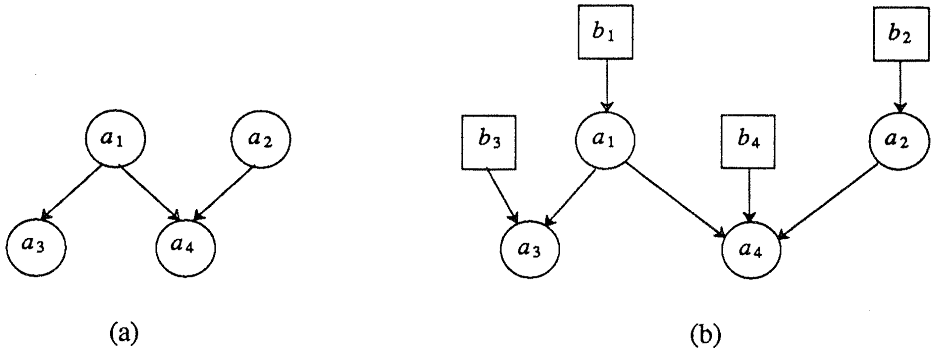


Figure 3.11

The computation of $P(a_3)$ requires only to multiply the matrices $P(a_3 \mid a_1)$ and $P(a_1)$ and to sum over the values of a_1. These two matrices are stored at the dummy nodes b_1 and b_3 of Figure 3.11(b), which indeed are the only dummy nodes not D-separated from node a_3 (given $\emptyset$). Thus, Algorithm 3 reveals the fact that the parameters represented by node b_2 and b_4 ($P(a_2)$, $P(a_4 \mid a_1, a_2)$) are not needed for the computation of $P(a_3)$. Note that the questions of the value of a node, or the parameters stored with a node influencing a given computation, may result in two different answers. For example, the value of a_4 might influence the computation of $P(a_3)$, because a_3 and a_4 could be dependent, while the parameters stored at node a_4 never affect this computation. Algorithm 3, by representing parameters as dummy variables, reveals this fact.

Shachter was the first to present an algorithm that identifies irrelevant parameters using transformations of arc-reversal and node-removal. A revised algorithm of Shachter [56] also detects irrelevant variables and it appears that the outcome of this algorithm is identical to ours. In our approach we maintain a clear distinction between the following two tasks: (1) declarative characterization of the independencies encoded in the network (i.e., the D-separation criterion) and (2) procedural implementation of the criterion defined in (1). Such separation facilitates a formal proof of the the algorithm's soundness, completeness and optimality. In Shachter's treatment, task (1) is inseparable from (2). The axiomatic basis upon which our method is grounded also provides means for extending the graphical criteria to other notions of independence, such as relational and correlational dependencies.

3.4 Consistency of Probabilistic Networks

The use of dags has been advocated in chapter 1 as a qualitative representation of judgments about dependencies and independencies in some domain. Is every dag consistent ? Is every dag realizable by a distribution ? Theorem 3.14 below provides an affirmative answer.

Theorem 3.14: For every dag D or an undirected graph G, there exists a non-extreme distribution P such that D is a perfect map of P.

Proof: Let Σ be the set of all statements that hold in a dag D (or in an undirected graph G). For every statement $\sigma \notin \Sigma$ in D, Theorem 3.9 guarantees the existence of a distribution P_σ that satisfies Σ and does not satisfy σ. Similarly, if G is an undirected graph, the construction of Theorem 3.9 applies as well; the trail between a and b would simply include no head-to-head nodes and no deterministic nodes ([2, 23] provide a direct proof of this assertion for undirected graphs). Let P be $\otimes\{P_\sigma \mid \sigma \notin \Sigma\}$ where $\otimes$ is the direct product operation, guaranteed by Theorem 2.5. P satisfies the statements in Σ and none other because these are the only statements that hold in all P_σ's. Therefore, P satisfies the requirements of the theorem. □

The construction presented in the proof of Theorem 3.14 leads to a rather complex distribution, where the domain of each variable is unrestricted. We conjecture, however, that the set of dependencies and independencies represented in a dag or an undirected graph can be realized in a more limited class of distributions, such as normal, or those defined on binary variables.

The next theorem provides the basis for an algorithm that determines whether an independence statement logically follows (wrt $\mathcal{PD}^+$) from a given set of saturated statements; it complements the results reported in section 2.4.

Theorem 3.15: Let Σ be a set of saturated statements and let $cl(\Sigma)$ be the closure of Σ under symmetry (1.5b), decomposition (1.5c), weak-union (1.5d), and intersection (1.6). Then, there exists a graph G and a strictly positive distribution P that satisfy exactly the statements in $cl(\Sigma)$.

Proof: Theorem 3.14 assures the existence of a strictly positive distribution that is a perfect map of any undirected graph. We shall construct an undirected graph G that satisfies exactly the statements in $cl(\Sigma)$. This will complete our proof. The graph G is constructed by first starting with the complete graph over U, the set of all variables. Then removing every edge (a,b), such that $a \in X$, $b \in Y$ for some statement $\sigma_i = I(X, Z, Y) \in \Sigma$ and only these edges. That G satisfies exactly the statements of $cl(\Sigma)$ can be shown as follows [46]:

Let k be the number of elements in Σ. Let $\sigma_1, \sigma_2 \cdots \sigma_k, \sigma_{k+1}, \cdots, \sigma_m$ be the list of all statements in $cl(\Sigma)$ ordered in a way such that each σ_i, $i > k$ is derived from previous statements in the list by one of the axioms. We show by finite induction that every statement in the list is represented in the graph: this proposition clearly holds for $\sigma_1, \cdots, \sigma_k$ because by our construction, G satisfies all statements of Σ. The truth for $j > k$ is implied by the induction hypothesis, and by the fact that vertex separation satisfies the graphoid axioms.

The other direction, namely that every statement that holds in G belongs to $cl(\Sigma)$ follows from theorem 1.1. This theorem constructs a graph G_0 in which a link (a, b) is removed from the complete graph if and only if $(a, U-\{a, b\}, b)$ belongs to $cl(\Sigma)$, and guarantees that every statement which holds in G_0 also holds in $cl(\Sigma)$. The weak-union axiom ensures that every edge removed in the construction of G will also be absent in G_0. Thus, G_0 is an edge-subgraph of G and, so, every statement in G holds in G_0 and therefore belongs to $cl(\Sigma)$. □

The implication algorithm for saturated statements is now clear: given Σ, construct G. For instance, consider the language where the probability of the i-th letter is determined solely by the (i-1)-th letter via $P(l_i \mid l_{i-1}) > 0$. Suppose that by sampling 5-letter words from this language the following two independencies were identified:

$$\Sigma = \{I(\{l_1, l_2\}, l_3, \{l_4, l_5\}),\ I(l_3, \{l_2\ l_4\}, \{l_1\ l_5\})\}$$

Are these statements sufficient to guarantee the Markov nature of the language and, moreover, is the chain a complete representation of all independencies that are implied by Σ ? The algorithm answers these questions affirmatively; it generates the chain of Figure 1.4 which faithfully represents each and every independence in the closure of Σ. The algorithm requires $O(k \cdot n^2)$ steps, where k is the size of Σ and n is the number of variables. To verify if a specific statement $\sigma = I(X, Z, Y)$ belongs to $cl(\Sigma)$ would require testing, in $O(n)$ steps, whether Z separates X from Y in G. The simplicity of the algorithm stems from the fact that saturated statements of graph separation and independence statements wrt $\mathcal{PD}^+$ possess identical axiomatic structure. This permits us to interpret Σ as statements about graph separation and construct a graph that embodies their closure.

Another interesting consequence of theorem 3.15 is given below.

Corollary 3.16: Let B be either a pairwise basis or a neighborhood basis of G. The dependency model defined by G is exactly the logical closure of B wrt $\mathcal{PD}^+$.

This follows from the observation that both the neighborhood and the pairwise bases consists of saturated statements. Corollary 3.16 provides the probabilistic semantics underlying vertex separation in Markov networks, and might explain the wide use of these networks.

3.5 Discussion

Researchers in relational database theory have invested much effort in characterizing dependencies between items of information [22, 70]. Their main goal has been the automatic construction of relational scheme, namely the construction of relational tables that facilitate the efficient representation and retrieval of information stored in a relational database. The initial paradigm has been to define several types of data dependencies, let the user specify those dependencies that govern the domain, and then use this information to design an efficient database scheme. This approach has been abandoned for several reasons. First, many types of dependencies were not common in real life and were therefore hard to elicit. Second, even when such dependencies were at hand, the computational difficulties of constructing optimal relational scheme turned out insurmountable.

The construction of expert systems based on Bayesian networks faces similar problems; a specification of independence and dependence assertions does not help in the automatic construction of optimal networks. It is even hard to determine what is logically implied by an arbitrary list of such assertions. Instead, this dissertation suggests an alternative approach. The expert is required to express knowledge about independencies in a graphical language dictated by the D-separation criteria, using informal guidelines of causation and time ordering. The resulting network possesses a precise semantics in terms of independence assertions and these can be used to verify whether the network faithfully represents the domain. The obvious loss is in generality, not everything can be specified; for example, there is no way to represent, in graphical language alone, all the dependencies and independencies that govern three variables constrained by equality–some independencies must be encoded numerically. The gain, however, is significant; the expert can express relationships between entities of interest in a convenient graphical format safe from contradictions and be assured that all conclusions are semantically valid.

CHAPTER 4
Discovering Causality from Statistical Data

The existence of some relationship between causality and the pattern of dependencies conveyed in Bayesian networks is obvious to anyone who constructs these networks from expert's judgments; whenever the construction ordering is consistent with the flow of causation the network ends up sparse and judgments, both structural and numerical, are produced consistently and reliably. However, the converse task, namely, inferring causal relationships from patterns of dependencies is far less understood. This chapter provides conditions under which the directionality of some links is not sensitive to the specific order chosen to construct the network; an essential prerequisite for associating a causal interpretation to these links. An efficient algorithm is developed that recovers these links from statistical data whenever possible.

4.1 Introduction

Most standard texts on the methodology of the social, behavioral and natural sciences warn the practitioner to refrain from inferring causal relationships from statistical dependence. This is not surprising in view of the wealth of examples showing the difficulty and presumably the impossibility of drawing such an inference. Several researchers in various schools of science have challenged this tradition with more modest goals in mind. The first of these has been to define causality in probabilistic terms such that the formal definition would 1) model as precisely as possible the common usage of the word ''causes'' and 2) would define the physical meaning of a cause [29, 64, 66, 68]. Contrary to Hume's deterministic view, the motivation to define causality in probabilistic terms stems from the observation that probabilistic causes are widely used in our language. The phrase ''reckless driving causes accidents'' provides an example where the relation between a cause and its effect is clearly probabilistic [68]. Since in all practical cases a model cannot reflect precisely the state of the world, the relationship between cause and effect must summarize entities outside the model and, hence become probabilistic. The second goal is computationally oriented. Given that causal models offer computational advantages of storage economy and retrieval time, one commits to cast statistical data in such models, and aspires to identify the most convenient causal structure feasible regardless of whether it corresponds to genuine causal mechanisms that ties together variables in the domain [28, 49, 52, 63].(*)

The lack of a precise definition of causality, despite the persistent attempts of philosophers throughout the years to obtain one, indicates that the second pragmatic approach might be found useful. Of course, one cannot claim to have discovered causal relationships if one does not possess an *operational definition* of causality, namely, a set of qualitative conditions which matches the way causality is used in natural discourse and which statistical data must satisfy before we are willing to accept the assertion ''a causes b''. Although such a definition should in principal be sensitive to the ''strength'' of the causal relationships under consideration, we will focus on basic qualitative features. For example, an operational definition may postulate that causality is transitive which perfectly complies with our intuition but may not always show in statistical data; the weather conditions of today may be regarded as a cause for tomorrow's weather while the weather in the turn of the century does not seem to influence tomorrow's weather [66]. Somewhere along the chain, causality has faded away contrary to our endorsement of transitivity.

Our attack on the problem of identifying causality is structured as follows; first, we pretend that Nature possesses ''true'' cause and effect relationships and that these relationships can be represented by a *causal network,* namely, a directed acyclic graph where each node represents a variable in the domain and the parents of that node denote directed causes of the corresponding variable. Next, we assume that Nature selects a joint distribution over the variables in such a way that direct causes of a variable render this variable statistically independent of all other variables except its consequences.(*) Then, we investigate the feasibility of recovering the network's topology efficiently and uniquely from the joint distribution. Computationally, solving this simplified problem is crucial if one aspires ever to deduce causal relationship from measurements; this is the main concern of this chapter. However, the solution is only partial because it does not offer a way of distinguishing between *spurious correlations* [59] and genuine causes, a distinction that is impossible within the confines of the close world assumption.

(*) The division of authors into two distinct categories is far of being sharp since the objectives are complementary. A summary of the different approaches to define causation can be found in [60].

(*) This matches exactly our formal description of parameterization of a dag D that forms a distribution for which D is a minimal-edge I-map (i.e., a Bayesian network).

It is not hard to see that if Nature were to assign totally arbitrary probabilities to the links, then some distributions would not enable us to uncover the structure of the network. However, by employing additional restrictions on the conditional distributions expressing properties we normally attribute to causal relationships, some structure could be recovered. The basic requirement is that two independent causes should become dependent once their effect is known [49]. For example, two independent inputs for an AND gate become dependent once the output is measured. This observation can be phrased axiomatically using conditional independence by the following property, called *Marginal Weak Transitivity:*

$$I(x,\emptyset,y) \ \& \ \neg I(x,\emptyset,a) \ \& \ \neg I(y,\emptyset,a) \Rightarrow \neg I(x,a,y)$$

This tells us that if two variables x and y are mutually independent, and each is dependent on their effect a, then x and y are conditionally dependent for at least one instance of a. If indeed x and y are perceived to be independent causes of a then people normally expect to find such relationship. Two additional properties are reasonable to attribute to causal interactions, and will be useful for recovering the causal network, intersection (1.6) and composition. Intersection is guaranteed if the distributions are strictly positive and is justified by the assumption that, to some extent, all observations are corrupted by noise. Composition is a property enforced, for example, by normal distributions, stating that two sets of variables X and Y are independent iff every $x \in X$ and $y \in Y$ are independent. This property is perhaps the most intuitive property of ''dependence'' in common discourse yet it is not enforced by all distributions.

The theory to be developed in the rest of the chapter addresses the following problem. We are given a distribution P and we know that P is represented as a *singly-connected* dag D whose structure is unknown. What properties of P allow the recovery of D ? It is shown that intersection composition and marginal weak transitivity are sufficient properties to ensure that the dag is uniquely recoverable (up to *isomorphism*) in polynomial time. The recovery algorithm developed considerably generalizes the method of Rebane & Pearl [49, Chapter 8] for the same task, as it does not assume the distribution is *dag-isomorph* (i.e., a distribution that is a perfect map of some dag). The algorithm implies, for example, that the assumption of a normal distribution is sufficient for a complete recovery of singly-connected dags.

4.2 Reconstructing Singly Connected Causal Networks

We investigate the feasibility of identifying whether or not a given distribution is induced from a singly-connected causal network and show how to identify the networks' topology when the distribution satisfies the following three properties:

- Intersection:
$$I(X, Z \cup Y, W) \ \& \ I(X, Z \cup W, Y) \Rightarrow I(X, Z, Y \cup W) \qquad (4.1)$$
- Composition:
$$I(X, Z, Y) \ \& \ I(X, Z, W) \Rightarrow I(X, Z, Y \cup W) \qquad (4.2)$$
- Marginal weak transitivity:
$$I(X, Z, Y) \ \& \ I(X, Z \cup \{c\}, W) \Rightarrow I(X, Z, c) \text{ or } I(c, Z, Y) \qquad (4.3)$$

The following definitions, some repeated from previous chapters, are found useful:

Definition: A graphoid is called *intersectional* if it satisfies (4.1), *semi-normal* if it satisfies (4.1) and (4.2), and *pseudo-normal* if it satisfies (4.1) through (4.3).

Definition: A *singly-connected* dag (or a *polytree)* is a directed acyclic graph with at most one trail connecting any two nodes. A dag is *non-triangular* if any two parents of a common node are never parents of each other. Polytrees are examples of non-triangular dags. The skeleton of a dag D, denoted *skeleton* (D), is the undirected graph obtained from D if the directionality of the links is ignored.

Definition: A Markov network G_0 of an intersectional graphoid M is the network formed by connecting two nodes, a and b, if and only if $(a, U - \{a, b\}, b) \notin M$. A *reduced graph* G_R of M is the graph obtained from G_0 by removing any edge $a - b$ for which $(a, \emptyset, b) \in M$.

Definition: A node b is called a head-to-head node with respect to a trail t in a dag D if there are two consecutive links $a \rightarrow b$ and $b \leftarrow c$ on t. Each occurrence of a head-to-head node wrt a trail is called a *head-to-head connection* of D. Each node of D may define several head-to-head connections, one with respect to each pair of its neighboring nodes.

Definition: Two dags D_1 and D_2 are *isomorphic* if the corresponding dependency models are equal.

Isomorphism draws the theoretical limitation of the ability to identify directionality of links using information on independence. For example, the two dags: $a \rightarrow b \rightarrow c$ and $a \leftarrow b \leftarrow c$, are indistinguishable in the sense that they portray the same set of independence assertions; these are isomorphic dags. On the other hand, the dag $a \rightarrow b \leftarrow c$ is distinguishable from the previous two because it portrays a new independence assertion, $I(a, \emptyset, c)$, which is not represented in either of the former dags. An immediate corollary of the definitions of d-separation and isomorphism is that any two polytrees sharing the same skeleton and the same head-to-head connections must be isomorphic. More generaly, it can be shown that two dags are isomorphic iff they share the same skeleton and the same head-to-head nodes emanating from non adjacent sources [50].

Lemma 4.1: Two polytrees T_1 and T_2 are *isomorphic* iff they share the same skeleton, and the same head-to-head connections.

> **Sufficiency:** If T_1 and T_2 share the same skeleton and the same head-to-head connections then every active trail in T_1 is an active trail in T_2 and vice versa. Thus, M_{T_1} and M_{T_2}, the dependency models corresponding to T_1 and T_2 respectively, are equal.
>
> **Necessity:** T_1 and T_2 must have the same set of nodes U, for otherwise their dependency models are not equal. If $a \rightarrow b$ is a link in T_1 and not in T_2, then the triplet $(a, U-\{a, b\}, b)$ is in M_{T_1} but not in M_{T_2}. Thus, if M_{T_1} and M_{T_2} are equal, then T_1 and T_2 must have the same skeleton. Assume T_1 and T_2 have the same skeleton and that $a \rightarrow c \leftarrow b$ is a head-to-head connection in T_1 but not in T_2. The trail $a - c - b$ is the only trail connecting a and b in T_2 because T_2 is singly-connected and it has the same skeleton as T_1. Since c is not a head-to-head node wrt this trail, $(a, c, b) \in M_{T_2}$. However, $(a, c, b) \notin M_{T_1}$ because the trail $a \rightarrow c \leftarrow b$ is activated by c. Thus, if M_{T_1} and M_{T_2} are equal, then T_1 and T_2 must have the same head-to-head connections. □

The algorithm below uses queries of the form $I(X, Z, Y)$ to decide whether a pseudo-normal graphoid M (e.g., a normal distribution) has a polytree I-map representation and if it does, then D's topology is identified. Axioms (4.1) through (4.3) are then used to prove that if D exists, then it is unique up to isomorphism. The algorithm is remarkably efficient; it requires only polynomial time, while a brute force approach would require checking $n!$ possible dags, one for each ordering of M's variables.

The Recovery Algorithm [*]

Input: Independence assertions of the form $I(X, Z, Y)$ drawn from a pseudo-normal graphoid M.

Output: A polytree I-map of M if such exists, or acknowledgment that no such I-map exists.

1. Start with a complete graph.

2. Construct the Markov network G_0 by removing every edge $i-j$ for which $(i, U-\{i, j\}, j)$ is in M.

3. Construct G_R by removing from G_0 any link $i-j$ for which $(i, \varnothing, j)$ is in M. If the resulting graph G_R has a cycle then answer ''NO''. Exit.

4. Orient every link $a-b$ in G_R towards b if b has a neighboring node c, such that $(a, \varnothing, c) \in M$ and $a-c$ is in G_0.

5. Orient the remaining links without introducing new head-to-head connections. If the resulting orientation is not feasible answer ''NO''. Exit.

6. Find an orientation that does not introduce new head-to-head connections. If the resulting polytree is not an I-map, answer ''NO''. Otherwise, this polytree is a minimal-edge I-map of M.

The following sequence of claims establishes the correctness of the algorithm and the uniqueness of the recovered network; full proofs are given in section 4.3.

Theorem 4.2: Let D be a non-triangular dag that is a minimal-edge I-map of an intersectional graphoid M. Then, for every link $a-b$ in D, $(a, U-\{a, b\}, b) \notin M$.

Theorem 4.2 ensures that every link in a minimal-edge polytree I-map (or more precisely, a link in a minimal-edge non-triangular dag I-map) must be a link in the Markov network G_0. Thus, we are guaranteed that Step 2 of the algorithm does not remove links that are needed for the construction of a minimal-edge polytree I-map.

Theorem 4.3: Let M be a semi-normal graphoid that has a minimal-edge polytree I-map T. Then, the reduced graph G_R of M equals *skeleton* (T).

[*] A variant of this algorithm has been suggested by Pearl (personal communication)

Corollary 4.4: All minimal-edge polytree I-maps of a semi-normal graphoid have the same skeleton (Since G_R is unique).

Theorem 4.3 shows that by computing G_R, the algorithm identifies the skeleton of any minimal-edge polytree I-map T, if such exists. Thus, if G_R has a cycle, then M has no polytree I-map and if M does have a polytree I-map, then it must be one of the orientations of G_R. Hence by checking all possible orientations of the links of the reduced graph one can decide whether a semi-normal graphoid has a minimal-edge polytree I-map. The next two theorems justify a more efficient way of establishing the orientations of G_R. Note that composition and intersection, which are properties of semi-normal graphoids, are sufficient to ensure that the skeleton of a polytree I-map of M is uniquely recoverable. Marginal weak transitivity, which is a property of pseudo-normal graphoids, is used to ensure that the algorithm orients the skeleton in a valid way. It is not clear, however, whether axioms (4.1) through (4.3) are indeed necessary for a unique recovery of polytrees.

Definition: Let M be a pseudo-normal graphoid for which the reduced graph G_R has no cycles. A *partially oriented polytree* P of M is a graph obtained form G_R by orienting a subset of the links of G_R using the following rule: A link $a \rightarrow b$ is in P if b has a neighboring node c, such that $(a, \emptyset, c) \in M$ and the link $a - c$ is in G_0. All other links in P are undirected.

Theorem 4.5: If M is a semi-normal graphoid that has a polytree I-map, then M defines a unique partially oriented polytree P.

Theorem 4.6: Let P be a partially oriented polytree of a semi-normal graphoid M. Then, every oriented link $a \rightarrow c$ of P is part of every minimal-edge polytree I-map of M.

Theorem 4.5 guarantees that the rule by which a partially oriented polytree is constructed cannot yield a conflicting orientation when M is pseudo-normal. Theorem 4.6 guarantees that the links that are oriented in P are oriented correctly, thus justifying Step 4.

We have thus shown that the algorithm identifies the right skeleton and that every link that is oriented must be oriented that way if a polytree I-map exists. It remains to orient the remaining links. Step 5 searches for an orientation that does not introduce new head-to-head connections. Theorem 4.7 below shows that no polytree I-map of M introduces new head-to-head connections. Lemma 4.1, shows that all orientations that do not introduce a head-to-head connection yield isomorphic dags because these polytrees share the same skeleton and the same head-to-head connections. Thus, in order to decide whether or not M has a polytree I-map, it is sufficient to examine merely a single polytree for I-mapness, as performed by Step 6.

Theorem 4.7: Let P be a partially oriented Polytree of a pseudo-normal graphoid M. Every orientation of the undirected links of P which introduces a new head-to-head connection to P yields a polytree that is not a minimal-edge I-map of M.

4.3 Proofs

Theorem 4.2: Let D be a non-triangular dag that is a minimal-edge I-map of an intersectional graphoid M. Then, for every link $a-b$ in D, $(a, U-\{a, b\}, b) \notin M$.

> **Proof:**$^{(*)}$ "Let 1 ... n be an ordering of the vertices of D. Let M_D be the dependency model corresponding to D. Let $i \rightarrow j$ be a link in D. If $j = n$ then $(i, U-\{i, n\}, n) \notin M$, for otherwise, D is not minimal. Assume that $i < j < n$ and, by contradiction, that $(i, U-\{i, j\}, j) \in M$. We will show that D cannot be minimal-edge. Nodes i and j cannot be both parents of n since this would imply the configuration $i \rightarrow n <- j$ with i connected to j in D contrary to its non-triangularity. Thus either $(i, U-\{i, n\}, n)$ or $(j, U-\{j, n\}, n)$ is in M_D which together with $(i, U-\{i, j\}, j) \in M$ imply by intersection (4.1), decomposition (1.5c) and symmetry (1.5b) that $(i, U-\{i, j, n\}, j) \in M$. Similarly, $n-1$ can not be a son of both i and j. Thus either $(i, U-\{i, n, n-1\}, n-1)$ or $(j, U-\{j, n, n-1\}, n-1)$ is in M_D which together with $(i, U-\{i, j, n\}, j) \in M$ (which is derived in the previous step) imply that $(i, U-\{i, j, n-1, n\}, j) \in M$. Continuing this way, by descending induction we get that the triplet (i, R_{ij}, j) is in M where R_{ij} are all vertices in D with indices less than j not including i. The link $i \rightarrow j$ is therefore redundant. This contradicts the minimality of D. □"

Theorem 4.3: Let M be a semi-normal graphoid that has a minimal-edge polytree I-map T. Then, the reduced graph G_R of M equals *skeleton* (T).

> **Proof:** Let $a-b$ be a link in *skeleton*(T) and let M_T be the dependency model defined by T. We show that $a-b$ must be a link in G_R. Since T is a polytree, T is non-triangular and therefore, by Theorem 4.2, the link $a-b$ is part of the Markov network G_0 of M. We will show that $(a, \emptyset, b) \notin M$. Thus the link $a-b$ is not re-

(*) Paz (personal communication)

moved from G_0. Consequently, $a-b$ is a link in G_R. Without loss of generality assume that $a \to b$ is a link in T (same argument when $b \to a$ is in T). Let A be the set of nodes connected to a with a trail not containing b, B be the set of b's descendants and C be the rest of the nodes in T. Being a polytree, A, B and C are disjoint. By definition of A, node a lies on the single trail connecting each node in A to b, and a is not a head-to-head node on none of these trails. Thus $(b, a, A) \in M_T$. T is an I-map of M. Hence (b, a, A) is a member of M as well. Assume, by contradiction, that $(b, \emptyset, a) \in M$. This triplet together with (b, a, A) imply by contraction (1.5e) that $(b, \emptyset, A \cup \{a\}) \in M$. By definition of C, all trails between C and $A \cup \{a\}$ contain at least one head-to head node, thus $(C, \emptyset, A \cup \{a\}) \in M_T$ and in M as well. This triplet together with $(b, \emptyset, A \cup \{a\})$ imply by composition that $(C \cup \{b\}, \emptyset, A \cup \{a\})$ must also be in M. By weak union, $(b, A \cup C, a) \in M$. Since $A \cup C$ is the set of all non-descendants of b, T is not minimal; link $b \to a$ should not be part of T, contradiction.

That the converse holds, namely, a link in G_R must be a link in *skeleton*(T), is shown as follows. Let a and b be two nodes not connected with a link in T. We show that the pair $a-b$ is not a link in the reduced graph G_R. There are three cases to consider. Either a is an ancestor of b (in T), b is an ancestor of a or neither is the case. In the first two cases there is a directed path from a to b or vice versa. The triplet $(a, U-\{a, b\}, b)$ is in M_T because $U-\{a, b\}$ includes a node that blocks this path. T is an I-map thus $(a, U-\{a, b\}, b) \in M$. Hence $a-b$ is not in G_0. Consequently, it is not in G_R either. If neither nodes is an ancestor of the other then $(a, \emptyset, b) \in M_T$ because each trail that connects a and b must contain a head-to-head node. Consequently, $(a, \emptyset, b) \in M$, and therefore $a-b$ is not a link in G_R. □.

Theorem 4.5: If M is a semi-normal graphoid that has a polytree I-map, then M defines a unique partially oriented polytree P.

Proof: Assume, by contradiction, that P is not unique, namely that there exists a conflicting orientation of some links of G_R (by Theorem 4.3, the skeleton of P is G_R). Let $a-b$ be a link that can be oriented both ways. Then, there exist a neighbor q of b for which $(a, \emptyset, q) \in M$ and $(a, U-\{a, q\}, q) \notin M$ that supports an orientation from a into b and there exists another node p, neighbor of a, for which $(b, \emptyset, p) \in M$ and $(b, U-\{b, p\}, p) \notin M$ that supports the reverse orientation. Thus, G_R must contain the chain $p-a-b-q$.

We reach a contradiction by showing that none of the eight possible orientations of the trail $p-a-b-q$ could be part of any minimal-edge polytree I-map of M. Since M has a polytree I-map, it has also a minimal-edge polytree I-map T. Consequently, the skeleton of T would not equal G_R, contradicting the assertion made by Theorem 4.3. If neither a nor b is a head-to-head node on this trail, then since $a-b-q$ is the only trail connecting a and q and this trail is blocked by b, which implies that $(a, U-\{a, q\}, q)$ must be a member of M, contradicting the selection of q. Otherwise, a or b are head-to-head nodes on this path. Assume b is a head-to-head node (the case where a is a head-to-head node is symmetric, by changing the roles of a and b). Then $p-a \rightarrow b \leftarrow q$ is part of T. In this case $(b, U-\{b, p\}, p) \in M_D \subseteq M$ contradicting the selection of p. □

Theorem 4.6: Let P be a partially oriented polytree of a semi-normal graphoid M. Then, every oriented link $a \rightarrow c$ of P is part of every minimal-edge polytree I-map of M.

Proof: By Theorem 4.6, P is unique and by Theorem 4.3, it has the same skeleton as any minimal-edge polytree I-map T of M. Since $a \rightarrow b$ is oriented in P, there must exist a node q, neighbor of b, for which $(a, \emptyset, q) \in M$ and $(a, U-\{a, q\}, q) \notin M$. Thus T, having the same skeleton of P, contains the trail $a-b-q$. Node b must be a head-to-head node on this trail in T. Otherwise, $(a, U-\{a, q\}, q) \in M_T$ because $U-\{a, q\}$ blocks the trail between a and b. Consequently, $(a, U-\{a, q\}, q)$ is in M as well, contradicting the selection of a and q. Thus b is a head-to-head node and therefore $a \rightarrow b$ is in T. □

Theorem 4.7: Let P be a partially oriented Polytree of a pseudo-normal graphoid M. Every orientation of the undirected links of P which introduces a new head-to-head connection to P yields a polytree that is not a minimal-edge I-map of M.

Proof: Assume, by contradiction that there exists an orientation of the undirected links of P that yields a minimal-edge polytree I-map T which introduces a new head-to-head connection. Let $a \rightarrow c \leftarrow b$ be a newly introduced head-to-head connection and let b be the node that is not a parent of c in P (namely, the link $c-b$ is not oriented in P). Let C be all parents of c in T, excluding a and b. Since T is singly-connected, $(C \cup \{a\}, \emptyset, b) \in M_T$, where M_T is the dependency model defined by T. T is an I-map, therefore $(C \cup \{a\}, \emptyset, b)$ is in M as well. We will show below that all paths between $C \cup \{a\}$ and b in G_0 must path through c. This will complete the proof; G_0 is an I-map, thus $(C \cup \{a\}, c, b) \in M$. This triplet, together with $(C \cup \{a\}, \varnothing, b)$ would imply, by marginal weak transitivity and contraction, that either $(C \cup \{a, c\}, \varnothing, b)$ or $(C \cup \{a\}, \varnothing, \{b, c\})$ are in M. These would imply, by weak union and symmetry, that either $(c, C \cup \{a\}, b)$ or $(c, C \cup \{b\}, a)$ are in M. Thus, either link $b \rightarrow c$ or $a \rightarrow c$ are redundant, contradicting the minimality of T.

It remains to show that all paths between $C \cup \{a\}$ and b in G_0 path through c. Let B be the set of nodes connected to b not through c (in T) and let A be the rest of the nodes excluding c. Thus the nodes of T consist of A, B and $\{c\}$, and these sets are disjoint. We will show that there is no link connecting a node in B and a node in A. Consequently, there exists no path between $C \cup \{a\} \subseteq A$ and $b \in B$ that does not path through c.

Any node $b' \in B$ is connected to a node $a' \in A$ in T only through the link $b \rightarrow c$ because T is singly connected. If $b' \neq b$, then $(b', U - \{a', b'\}, a') \in M_T \subseteq M$. Therefore, the pair $a' - b'$ is not a link in G_0. If $b' = b$, then it cannot be connected with a link to a parent a' of c because otherwise the link $b \rightarrow c$ would be oriented in P because the following two requirements would be met: $(b, U - \{a', b\}, a') \notin M$ and $(a', \varnothing, b) \in M$. Node b cannot be connected with a link to any other node $a' \in A$ because $(b, U - \{a', b\}, a') \in M$; c blocks the trail from b to each of c's descendants and the parents of a block the path from b to all of c's non-descendants. Thus, there exists no link connecting a node in A and a node in B. □

4.4 Discussion

In the absence of temporal information, the uniqueness of directionality is a prerequisite for inferring causal relationships from statistical information. This chapter provides conditions under which the directionality of some links is indeed uniquely recoverable. It is shown that if a distribution is generated from a singly connected causal network, then the topology of the network can be recovered provided that this distribution satisfies three properties: composition, intersection and marginal weak transitivity.

There is one difficulty with this approach; we are working within the confines of the closed world assumption, namely, we assume that the set of variables U adequately summarizes the domain and remains fixed throughout the structuring process. This assumption does not enable us to distinguish between genuine causes and spurious correlations; a link $a \rightarrow b$ that has been determined by our procedure may be represented by a chain $a \leftarrow c \rightarrow b$ where c is a variable not accounted for when the network is first constructed. Thus, the dependency between a and b which is marked as causal when $c \notin U$ is in fact spurious, and this is revealed when c becomes observable. This limitation, one must emphasize, is frequent in common discourse as well; whatever is perceived as a cause today may be changed when more accurate knowledge becomes available.

Future research is needed for incorporating variables outside U into the network in order to facilitate sparser dag representations. The addition of extra nodes often renders graphical representations sparser and more accurate. For example, a network that represents medical symptoms would show all nodes connected and would be of little use, but when a disease variable is added, the network becomes much sparser and more useful. Pearl and Tarsi provide an algorithm that constructs tree I-maps with added variables whenever possible [52]. An extension of this algorithm to polytrees and other topologies would be extremely valuable. It should be noted that the assumption of singly-connectedness may not be needed for a recovery algorithm. Theorem 1, which is the basic step of the recovery, assumes only non-triangularity. Thus an efficient recovery algorithm for non-triangular dags may be found as well.

CHAPTER 5
An Approach to Knowledge Acquisition

An important step in organizing a large body of knowledge is the grouping of related pieces of information into more or less independent chunks. In constructing large Bayesian networks from expert's judgments, this amounts to identifying the connected components of the network. Asking the expert directly whether variables x and y are connected may be a hard question to answer, since the expert may not have a clear global view of the network topology. However, the query: ''does the value of x ever tells you anything about the value of y ?'' should evoke more reliable judgment. This chapter identifies the class of distributions, called separable, for which the answer to this query can safely be interpreted as an assertion about the connectivity of x and y, and argues that it reasonable to assume these distributions in the construction of Bayesian networks. Normal and strictly-positive binary distributions are example of separable distributions.

5.1 Introduction

The construction of Bayesian networks as faithful representations of a given domain relies on the ease and confidence by which an expert can describe the relationships between variables in this domain. Explicating these relationships is often straight forward, however, difficulties may arise when variables have many instances. For example, in medical diagnosis, a variable corresponding to ''cancer'' may have dozens of possible values, each corresponding to a different type of cancer. An expert wishing to describe the relationship between the different symptoms, tests and treatments of cancer may find it rather confusing unless he first partitions the many types of cancer into several groups sharing common characteristics; in fact, the grouping of related pieces of information into more or less independent chunks is an important step in organizing any large body of knowledge.

The need to partitioning large knowledge bases has lead to the construction of *similarity networks,* which are an effective tool for eliciting Bayesian networks from experts [31]. This approach is summarized below: Let h be a distinguished variable designated for the disease ''hypothesis'', and let the instances of h, $\boldsymbol{h}_1, \cdots, \boldsymbol{h}_n$, stand for an exhaustive list of possible diseases. First, a connected undirected graph is constructed where each of the n nodes represents a different instance of h and each link represents a pair of ''similar'' diseases, namely diseases that are sometimes hard to discriminate. Next a spanning tree of this graph is formed. Then, for each link $\boldsymbol{h_i} - \boldsymbol{h_j}$ in the tree, a *local Bayesian network* is composed, assuming that either $h = \boldsymbol{h_i}$ or $h = \boldsymbol{h_j}$ must hold; it consists of a distinguished root node h whose instances are $\boldsymbol{h_i}$ and $\boldsymbol{h_j}$, additional nodes representing symptoms that help to discriminate between the two instances of h, and links representing the dependencies among symptoms and their relationship to the hypothesis node h. Finally, the global network is formed from the $n-1$ local networks; it consists of the union of all links and their adjacent nodes in the local networks.

The conceptual advantage of this approach is clear; the expert can focus his attention on two diseases at a time. This enables the expert to provide more accurate parameters which considerably improve the reliability of the resulting system. Moreover, although we face the task of constructing $n-1$ local networks instead of one, the task is performed faster in practice because it eases the estimation of individual parameters [32]. Furthermore, when we concentrate on two specific diseases at a time, many symptoms are not represented in the local network because they are irrelevant for discriminating these diseases. Heckerman has shown that under the assumption of strict-positiveness, namely that every combination of symptoms and diseases is feasible, the union of the connected components of node h in each local network generates a valid Bayesian network that faithfully represents the domain [32]. Technically, this means that if each local network is a minimal-edge I-map of a distribution P, then their graph union is also a minimal-edge I-map of P.

A difficulty with this approach is to identify the set of nodes that are connected to node h in each local network. We could consult the expert by asking him directly queries of the form: ''is node s (s connotes symptom) connected to node h, given that either $h = \boldsymbol{h_i}$ or $h = \boldsymbol{h_j}$ must hold ?'', however, this query may be inadequate because it refers to a graphical representation of the domain, a language with which the expert might not be familiar. On the other hand, the query ''does this symptom in any circumstances help you to discriminate between the two diseases $\boldsymbol{h_i}$ and $\boldsymbol{h_j}$?'' is much more appealing since it addresses exactly the specialty of the expert.

The main contribution of this chapter is identifying the class of distributions, called *separable,* for which the answer to these two queries is always identical. Strictly-positive distributions over binary variables and normal distributions are examples of separable distributions. Our analysis will be based on the notion of *interaction*, a stronger from of dependence.

5.2 Separable Graphoids

The definition of *interaction* formalizes the sentence: "does the value of a ever tells us anything about the value of b ?"; two variables are said to *interact* if there exists a context where they are dependent. More precisely, if U stands for a finite set of preselected variables of interest, then a and b *interact* if there exists a set $Z \subseteq U-\{a,b\}$ such that $\neg I(a,Z,b)$. One would expect that any two variables a and b that do not interact can always be placed in two disconnected components of a Bayesian network, indicating that they are totally unrelated. In other words, if a and b do not interact, one expects to find a partitioning of U into two sets U_a and U_b that contain a and b respectively, such that U_a and U_b are independent. Unfortunately, such a relationship between connectedness and interaction is not guaranteed by probability theory. For example, if U consists of three variables a, b and c, then it is possible that a and b do not interact, namely that a and b are both marginally independent and independent conditioned on c, and yet no variable is independent of the other two. This happens, for example, if a and b are the outcome of two independent fair coins and c is a variable whose domain is $\{head, tail\} \times \{head, tail\}$ and whose value is (i,j) if and only if the outcome of a is i and the outcome of b is j. Any Bayesian network representation of these variables will render node a and b connected, either by a direct link or via a trail through c, thus making it impossible to place a and b in two disconnected components of the network. A distribution that satisfies the condition that non-interaction implies non-connectedness (the converse always holds), is said to be *separable.* The definitions of interaction and separability are phrased below in the language of graphoids; recall that distributions are special types of graphoids.

Definition: Let M be a graphoid over a finite set of variables U. Two variables a and b of M are said to *interact,* denoted $interact(a,b)$, $\exists Z \subseteq U-\{a,b\}$ such that $(a,Z,b) \notin M$.

Definition: A graphoid M over a finite set of variables U is *separable* if for every two elements a and b that do not interact, there exists a partitioning U_a and U_b of U such that $a \in U_a$, $b \in U_b$ and $(U_a, \emptyset, U_b) \in M$.

The requirement that a distribution be separable can be cast in another appealing format; it is equivalent to the requirement that *interact* is a transitive relation, namely, that interact satisfies the following property:

- Transitivity:

$$interact(a,b) \ \& \ interact(b,c) \Rightarrow interact(a,c) \tag{5.1}$$

(Theorem 5.6 below). This property is so appealing to our intuition that we are tempted to speculate that all distributions not obeying this property are *epistemologically* *inadequate* for modeling a human reasoner, and that distributions that do satisfy this property are *natural* in the sense that they adequately represent the conventional properties of the word ``interact''. We note that transitivity can be viewed in a dual fashion: it could serve as a constraint that one would like to impose on an expert when information on interaction is elicited. Alternatively, it could be employed as a plausible argument to explain why a system views a and b as indirectly interacting although this fact was never explicitly stated by the expert. The following definitions and claims are needed to establish the equivalence between separability and transitivity.

Definition: Two nodes are *connected* in a dag D iff there exists a trail connecting them in D. A *connected component* in a dag (E, V) is a subgraph (E', V') (i.e., $E' \subseteq E$ and $V' \subseteq V$), for which any two nodes are connected.

Definition: Two elements of a graphoid M are said to be *related,* denoted *related*(a, b), if there exists a minimal-edge I-map dag of M in which the corresponding nodes are connected.

The next lemma and its corollary show that for any graphoid M the relation *related* can be determined from any single minimal-edge I-map of M; if two elements a and b are connected in some minimal-edge I-map of M then they are also connected in all such I-maps of M and if they are not connected in one, then they are not connected in none.

Lemma 5.1: The connected components of any two minimal-edge I-maps of a graphoid M induce the same partitioning on M's variables.

Proof: Let D_A and D_B be two minimal-edge I-maps of M. Let C_A and C_B be two connected components of D_A and D_B respectively. Let A and B be the nodes of C_A and C_B respectively. We show that either $A = B$ or $A \cap B = \emptyset$. This will complete the proof because for an arbitrary connected component C_A in D_A there must exists a connected component C_B that shares at least one node with C_A and thus, by the above claim, it must have exactly the same nodes as C_A. Thus each connected component of D_A shares the same nodes with exactly one connected component of D_B. Thus D_A and D_B induce the same partitioning on M's variables.

Since D_A is an I-map of M and C_A is a connected component of D_A, we must have $(A, \emptyset, U-A) \in M$, where U stands for M's variables. By symmetry (1.5b) and decomposition (1.5c), $(A \cap B, \emptyset, B-A) \in M$. Hence $A \cap B = \emptyset$ or $B-A = \emptyset$, for otherwise, due to the minimality of D_B, C_B would not be a connected component because it would consist of two non-empty independent sets: $A \cap B$ and $B-A = \emptyset$.

Similarly, $A \cap B = \varnothing$ or $A - B = \varnothing$, for otherwise C_A would not be a connected component. Thus, either $A = B$ or $A \cap B = \varnothing$ must hold. □

Corollary 5.2: The relation *related* defined by a graphoid is a transitive relation.

Proof: *related*(a, b) implies that a and b are connected in some minimal-edge *I*-map D. *related*(b, c) implies that b and c are connected in some other minimal-edge *I*-map D'. By Lemma 5.1, b and c are connected also in D. Thus a and c must be connected in D as well. Hence, *related*(a, c) holds for M. □

The definition of interaction can be extended to sets; this is found useful for showing that *interact* and *related* are equal relations for separable graphoids. Consequently, since *related* is a transitive relation, *interact* must be transitive as well.

Definition: Let M be a graphoid over a finite set of variables U. Two disjoint subsets A and B of U *interact,* denoted *interact*(A, B), iff $\exists Z \subseteq U - A \cup B$ such that $(A, Z, B) \notin M$. The relation $J(A, B)$ stands for $\neg$*interact*(A, B).

Interaction between sets, as the next lemma shows, is solely determined by the interactions between their individual elements. As is well known, this compositional property does not hold for probabilistic independence; two sets may be dependent although their individual elements are independent. For example, if a and b are two independent fair coins and c is their sum modulo 2 then c is dependent on $\{a, b\}$ but is independent of each individual.

Lemma 5.3: Let M be a graphoid over a finite set of variables U. Two subsets A and B of U do not interact iff every two variables $a \in A$ and $b \in B$ do not interact.

Proof: It is sufficient to show that for any three subsets A, B and C of U, $J(A, B)$ and $J(A, C)$ imply $J(A, B \cup C)$ and vice versa. That $J(A, B \cup C)$ implies $J(A, B)$ follows from decomposition (1.5c) of M. The converse is shown to follow from contraction (1.5e); $J(A, B)$ implies $(A, X, B) \in M$ and $J(A, C)$ implies $(A, X \cup B, C) \in M$. Together, these imply by contraction that $(A, X, B \cup C) \in M$. Since X is arbitrary, $J(A, B \cup C)$ holds in M. □

Next, we show that *interact* and *related* are two equal relations for separable graphoids.

Lemma 5.4: Let M be a separable graphoid. Then, for every two variables a, b of M,

$$interact(a, b) \leftrightarrow related(a, b)$$

Proof: If a and b do not interact then, since M is separable, there exists a partitioning of M's variables into two disjoint sets U_a, U_b such that $a \in U_a$, $b \in U_b$ and $(U_a, \emptyset, U_b) \in M$. Let D be a minimal-edge I-map dag of M formed by an ordering of M's variables that places all variables in U_a before those of U_b. The resulting I-map has no trail between U_a and U_b. Thus a and b are not connected in D. It follows that a and b are not connected in any minimal-edge I-map of M (Lemma 5.1). Thus a and b are not related in M.

If a and b interact then, by definition, there exists a set of variables Z such that $(a, Z, b) \notin M$. Let D be a minimal-edge dag I-map formed by an ordering of M's variables that starts with a, followed by the variables in Z, followed by b. In the resulting dag D, there exists a link connecting a and b because otherwise $(a, Z, b) \in M$. Thus, a and b are connected in D and therefore related in M. □

Corollary 5.5: Let M be a separable graphoid. Then, *interact* is a transitive relation.

Proof: By Lemma 5.4 the relations *interact* and *related* are equal. By Corollary 5.2, *related* is a transitive relation. Thus, *interact* is a transitive relation. □

Next, we establish the equivalence between separability and transitivity.

Theorem 5.6: Let M be a graphoid and *interact* the relation it defines. Then, M is separable iff *interact* is a transitive relation.

Proof: If M is separable, the relation *interact* is a transitive relation (Corollary 5.5). It remains to show the converse; transitivity implies separability. Let U stand for M's variables. Let a and b be two arbitrary elements of U that do not interact. We will show by induction on $|U|$ that if *interact* satisfies transitivity (5.1) then there exists a minimal-edge I-map D where a and b are not connected. Consequently, M is separable because $(U_a, \emptyset, U_b) \in M$ where U_a are the variables connected to a in

D and U_b are the rest of the variables.

We construct D in the ordering $u_1 \triangleq a, u_2 \triangleq b, u_3, \ldots, u_n \triangleq e$ of M's variables. Assume $n = 2$. Variables a and b do not interact, therefore $(a, \emptyset, b) \in M$. Thus, a and b are not connected. Otherwise, $n > 2$. Let D_e be a dag formed from M by the ordering $u_1, \ldots, u_{n-1}$ of M's variables. let A be the set of nodes connected to a and let B be the rest of the nodes in D_e. The dag D is formed from D_e by adding the last node e as a sink and letting its parents be a minimal set that makes e independent of all the rest in M (see the construction of Theorem 1.2). By the induction hypothesis, before e was added, A and B are disconnected. After node e is added, a trail through e might exists that connects a node in A and a node in B. We will show that there is none; if the parent set of e is indeed minimal, then either e has no parents in A or it has no parents in B, rendering a and b disconnected.

Since a and b do not interact and since M satisfies transitivity (5.1), it follows that either a or b do not interact with e. Without loss of generality assume that a and e do not interact. Let a' be an arbitrary node in A. By transitivity it follows that either a or e do not interact with a', for otherwise, a and e would interact, contrary to our selection of a. If a and a' do not interact, then by the induction hypothesis, A can be partitioned into two independent subsets, thus A would not be connected in the minimal-edge I-map D_e, contradicting our selection of A. Thus, every variable $a' \in A$ does not interact with e. It follows that the entire set A does not interact with e (Lemma 5.3). Thus, in particular, $(A, \hat{B}, e) \in M$ where $\hat{B}$ are the parents of e in B. Consequently, e has no parents in A because otherwise D were not minimal. Hence, a and b are on two different connected components of D. □

5.3 Separable Distributions and Refined Dependency Models

Two important classes of distributions, normal and positive binary, are shown to be separable. Hence, in dealing with these distributions one is guaranteed that interaction is transitive. We first examine a property of independence, called *propositional transitivity:*

$$I(A_1A_2A_3A_4, \emptyset, B_1B_2B_3B_4) \ \& \ I(A_1A_2B_3B_4, e=e', B_1B_2A_3A_4) \ \& \ I(A_1A_3B_2B_4, e=e'', B_1B_3A_2A_4) \Rightarrow$$

$$I(A_1, \emptyset, e\ A_2A_3A_4B_1B_2B_3B_4) \text{ or } I(B_1, \emptyset, e\ A_1A_2A_3A_4B_2B_3B_4) \qquad (5.2)$$

Our plan is to show that this axiom, which is satisfied by strictly-positive binary distributions and normal distributions, implies separability. The first antecedent of this axiom states that two sets of variables A and B are marginally independent where each set is the union of four possibly-empty subsets, $A_1A_2A_3A_4$ and $B_1B_2B_3B_4$, respectively. The second antecedent states that there exists a partitioning $A_1A_2B_3B_4$ and $B_1B_2A_3A_4$ of $A \cup B$ such that the two sets are independent given $e = e'$. Another partitioning, given $e = e''$, is stated by the third antecedent. The two statements in the consequence of (5.2) assert that either A_1 or B_1 are independent of all other variables, including e. Note that each statement in the antecedents has one less uninstantiated variable than each statement of the consequence; this observation is the basis of the inductive proof showing that propositional transitivity implies separability. Note also that when all sets aside from A_1 and B_1 are empty and e is a binary variable with a domain $\{e', e''\}$, then propositional transitivity reduces to the following known property of binary distributions:

$$I(A_1, \emptyset, B_1) \;\&\; I(A_1, e, B_1) \Rightarrow I(A_1e, \emptyset, B_1) \text{ or } I(A_1, \emptyset, eB_2)$$

The proof that propositional transitivity holds for normal distributions is given below. The proof that it holds for strictly-positive distributions over binary variables can be found in [24]. We conjecture that propositional transitivity holds for all binary distributions as well.

Lemma 5.7: [24] Propositional transitivity holds for any strictly-positive distribution over binary variables.

The definition of dependency models and graphoids of section 1.3 concentrates on independence between variables and not between particular instances of these variables. Thus, to allow the incorporation of propositional transitivity which refers to specific instances, a refinement of these definitions is needed in which the domain of each variable becomes explicit.

Definition: Let U be a set of variables each associated with a set of possible outcomes, called the domain of u. A member in the domain of u is called an instance of u. An instance $\mathbf{X}$ of a set of variables X is a member in the Cartesian product $\times_{x \in X} \mathit{domain}(x)$ where $\mathit{domain}(x)$ is the domain of x. A *refined dependency model M over U* is a set of triplets $(\mathbf{X}, \mathbf{Z}, \mathbf{Y})$ where X, Y and Z are disjoint subsets of U, and $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{Z}$ are their instances respectively.

Clearly, every refined dependency model M_R, defines a dependency model M in the sense of section 1.4; a triplet (X, Z, Y) is in M iff $(\mathbf{X}, \mathbf{Z}, \mathbf{Y})$ is in M_R for every instance $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$ of X, Y and Z, respectively. In particular, every probability distribution defines a refined dependency model.

Definition: A *refined graphoid* is any refined dependency model that satisfies axioms (1.5a) through (1.5e) (the graphoid axioms).

Next, we abstract the notion of conditional distributions.

Definition: Let $M(U)$ be a refined dependency model over a finite set of variables $U = \{u_1, \cdots, u_n\}$. The *conditional of* $M(U)$ *on* $u_n = \mathbf{u}_n$, denoted $M(\{u_1, \cdots, u_{n-1}\} \mid u_n = \mathbf{u}_n)$, is a refined dependency model that contains a triplet $(\mathbf{X}, \mathbf{Z}, \mathbf{Y})$ iff $(\mathbf{X}, \mathbf{Z} \cup \{\mathbf{u}_n\}, \mathbf{Y}) \in M(U)$

Theorem 5.8: Every refined graphoid satisfying propositional transitivity is separable.

> **Proof:** Let M be a graphoid and let U be its variables. Let a and b be two arbitrary elements of U that do not interact. We will show by induction on $|U|$ that if M satisfies propositional transitivity then there exists a minimal-edge I-map D where a and b are not connected. Consequently, M is separable because $(U_a, \varnothing, U_b) \in M$ where U_a are the variables connected to a in D and U_b are the rest of the variables.
>
> We construct D in the ordering $u_1 \triangleq a, u_2 \triangleq b, u_3, \ldots, u_n \triangleq e$ of M's variables. Assume $n = 2$. Variables a and b do not interact, therefore $(a, \varnothing, b) \in M$. Thus, a and b are not connected. Otherwise, $n > 2$. Let D_e be a dag formed from M by the ordering $u_1, \ldots, u_{n-1}$ of M's variables. let A be the set of nodes connected to a and let B be the rest of the nodes in D_e. Since a and b do not interact, by the induction hypothesis, $A \cap B = \varnothing$. Thus, $(A, \varnothing, B) \in M$ $(\triangleq I_1)$. The dag D is formed from D_e by adding the last node e as a sink and letting its parents be a minimal set that makes e independent of the rest of M's (see the construction of Theorem 1.2). Let $D_{e'}$ and $D_{e''}$ be minimal-edge I-maps of the conditional graphoids $M(A \cup B \mid e = e')$ $(\triangleq M_{e'})$ and $M(A \cup B \mid e = e'')$ $(\triangleq M_{e''})$ respectively, formed in the ordering $u_1, \ldots, u_{n-1}$. Since both $M_{e'}$ and $M_{e''}$ are subsets of M, a and b do not interact in neither of these dependency models. By the induction hypothesis $M_{e'}$ and $M_{e''}$ are separable. Hence there exists a partitioning $A_{e'}, \hat{A}_{e'}, B_{e'}$ and $\hat{B}_{e'}$ of $A \cup B$ where $A = A_{e'}\hat{A}_{e'}$, $B = B_{e'}\hat{B}_{e'}$, $a \in A_{e'}$ and $b \in B_{e'}$, such that $(A_{e'}\hat{B}_{e'}, \varnothing, B_{e'}\hat{A}_{e'}) \in M_{e'}$ $(\triangleq I_2)$. Similarly, there exists another partitioning of $A \cup B$ that satisfies, $A_{e''}, \hat{A}_{e''}$, $B_{e''}$ and $\hat{B}_{e''}$ of $A \cup B$ where $A = A_{e''}\hat{A}_{e''}$, $B = B_{e''}\hat{B}_{e''}$, $a \in A_{e''}$ and $b \in B_{e''}$, such that $(A_{e''}\hat{B}_{e''}, \varnothing, B_{e''}\hat{A}_{e''}) \in M_{e''}$ $(\triangleq I_3)$. In

other words, each of the two instances of e induces a partitioning of A and B into two independent subsets. There are at most eight disjoint subsets formed by the two partitioning. These include: $A_1 \triangleq A_{e'} \cap A_{e''}$, $A_2 \triangleq \hat{A}_{e'} \cap A_{e''}$, $A_3 \triangleq A_{e'} \cap \hat{A}_{e''}$, $A_4 \triangleq \hat{A}_{e'} \cap \hat{A}_{e''}$, $B_1 \triangleq B_{e'} \cap B_{e''}$, $B_2 \triangleq \hat{B}_{e'} \cap B_{e''}$, $B_3 \triangleq B_{e'} \cap \hat{B}_{e''}$ and $B_4 \triangleq \hat{B}_{e'} \cap \hat{B}_{e''}$. These definitions yield the following relationships: $A = A_1A_2A_3A_4$, $A_{e'} = A_1A_3$, $\hat{A}_{e'} = A_2A_4$, $A_{e''} = A_1A_2$, $\hat{A}_{e''} = A_3A_4$, $B = B_1B_2B_3B_4$, $B_{e'} = B_1B_3$, $\hat{B}_{e'} = B_2B_4$, $B_{e''} = B_1B_2$ and $\hat{B}_{e''} = B_3B_4$. Rewriting assertions I_1, I_2 and I_3 with these notations yields $(A_1A_2A_3A_4, \varnothing, B_1B_2B_3B_4) \in M$, $(A_1A_3B_2B_4, e = e', B_1B_3A_2A_4) \in M$ and $(A_1A_2B_3B_4, e = e'', B_1B_2A_3A_4) \in M$ which are the three antecedents of propositional transitivity (5.1). Since M is closed under this axiom, it follows that either $(A_1, \varnothing, e\ A_2A_3A_4B_1B_2B_3B_4) \in M$ or $(B_1, \varnothing, e\ A_1A_2A_3A_4B_2B_3B_4) \in M$. Since $a \in A_1$ and $b \in B_1$ were chosen arbitrarily, M is separable. □

Corollary 5.9: Every strictly-positive distribution over binary variables is separable.

The proof above also shows that normal distributions are separable because they satisfy propositional transitivity. That propositional transitivity holds for normal distributions stems from the following axioms which hold for these distributions:

- Composition [13]:
$$I(X, Z, Y) \Rightarrow I(Y, Z, X) \tag{5.3a}$$
- Unification [68]:
$$I(X, Z = Z, Y) \Rightarrow I(X, Z, Y) \tag{5.3b}$$
- Marginal weak transitivity [49]:
$$I(X, Z, Y)\ \&\ I(X, Z \cup Y, W) \Rightarrow I(X, Z, Y \cup W) \tag{5.3c}$$

Lemma 5.10: Propositional transitivity holds for any refined graphoid satisfying axioms (5.3a) through (5.3c) (e.g. normal distributions).

Proof: The three statements in the antecedents of propositional transitivity are listed below:

$$I(A_1A_2A_3A_4, \varnothing, B_1B_2B_3B_4), I(A_1A_2B_3B_4, e = e', B_1B_2A_3A_4), I(A_1A_3B_2B_4, e = e'', B_1B_3A_2A_4)$$

Applying the unification axiom yields the following assertions:

$$I(A_1A_2A_3A_4, \varnothing, B_1B_2B_3B_4),\ I(A_1A_2B_3B_4, e, B_1B_2A_3A_4),\ I(A_1A_3B_2B_4, e, B_1B_3A_2A_4),$$

denoted by I_1, I_2 and I_3 respectively. The following three implications are needed for the proof:

$$I(A_1A_2, \varnothing, B_1B_2)\ \&\ I(A_1A_2, e, B_1B_2) \Rightarrow I(A_1A_2, \varnothing, e) \text{ or } I(e, \varnothing, B_1B_2)$$

(5.4a)

$$I(A_1A_2, e, A_3A_4) \;\&\; I(A_1A_2, \emptyset, e) \Rightarrow I(A_1A_2, \emptyset, eA_3A_4) \qquad (5.4b)$$

$$I(A_1A_2, \emptyset, B_1B_2B_3B_4) \;\&\; I(A_1A_2, \emptyset, eA_3A_4) \Rightarrow I(A_1A_2, \emptyset, eA_3A_4B_1B_2B_3B_4) \qquad (5.4c)$$

The first follows from marginal weak transitivity (5.3c), the second from contraction (1.5e) and the third from composition (5.3a).

Next, we show that right hand side of propositional transitivity (5.2),

$I(A_1, \emptyset, e\ A_2A_3A_4B_1B_2B_3B_4)$ or $I(B_1, \emptyset, e\ A_1A_2A_3A_4B_2B_3B_4)$,

follows from I_1, I_2 and I_3; this will complete the proof. The two antecedents of (5.4a) are derived, by decomposition, from I_1 and I_2 respectively. Thus, either $I(A_1A_2, \emptyset, e)$ or $I(B_1B_2, \emptyset, e)$ is implied. Assume the first disjunct holds. The first antecedent of (5.4b) follows from I_2 by decomposition. Thus, (5.4b) yields $I(A_1A_2, \emptyset, eA_3A_4)$. The first antecedent of (5.4c) follows form I_1, thus (5.4c) yields $I(A_1A_2, \emptyset, eA_3A_4B_1B_2B_3B_4)$.

Assume the second disjunct $I(B_1B_2, \emptyset, e)$ holds. A similar derivation where the roles of A and B are switched yields that $I(B_1B_2, \emptyset, eB_3B_4A_1A_2A_3A_4)$ must hold. Consequently, we have thus shown that, I_1 and I_2 imply that either

$$I(A_1A_2, \emptyset, eA_3A_4B_1B_2B_3B_4)\ (\triangleq J_1) \quad or \quad I(B_1B_2, \emptyset, eB_3B_4A_1A_2A_3A_4)\ (\triangleq J_2)$$

Similarly, from I_1 and I_3 we obtain that either

$$I(A_1A_3, \emptyset, eA_2A_4B_1B_2B_3B_4)\ (\triangleq J_3) \quad or \quad I(B_1B_3, \emptyset, eB_2B_4A_1A_2A_3A_4)\ (\triangleq J_4)$$

hold in M (by switching the roles of A_2 with A_3 and B_2 with B_3). Thus, there are four cases to consider, by choosing one statement of each of the two disjunctions above. If J_1 and J_3 hold, then from J_1, by decomposition, $I(A_1, \emptyset, eA_3A_4B_1B_2B_3B_4)$ follows and from J_3, $I(A_1, \emptyset, A_2)$ follows. Together, by composition, $I(A_1, \emptyset, eA_2A_3A_4B_1B_2B_3B_4)$ is implied. Similarly, when J_2 and J_4 hold, $I(B_1, \emptyset, eA_1A_2A_3A_4B_2B_3B_4)$ must hold. If J_1 and J_4 hold, then by decomposition on J_4, $I(B_3, \emptyset, e)$ is obtained. $I(B_3, e, B_1)$ is implied from I_2 by decomposition. Together, the two statements yield, by contraction and decomposition, that $I(B_3, \emptyset, B_1)$ must hold. This statement combined with $I(B_1, \emptyset, eB_2B_4A_1A_2A_3A_4)$, which follows from J_4 by decomposition, yield, using composition, that $I(B_1, \emptyset, eA_1A_2A_3A_4B_2B_3B_4)$ holds in M. The case where J_2 and J_3 hold is symmetric to the case where J_1 and J_4 hold (by exchanging A 's with B 's), thus yielding $I(A_1, \emptyset, eA_2A_3A_4B_1B_2B_3B_4)$. □

The proof of separability of normal distributions is more complex than needed. Normal distributions satisfy stronger axioms than propositional transitivity which could have been used to show separability. One such property is the following:

$$I(C_1C_2, \emptyset, D_1D_2) \ \& \ I(C_1D_2, e{=}e, D_1C_2) \Rightarrow I(C_1, \emptyset, eC_2D_1D_2) \text{ or } I(D_1, \emptyset, eD_2C_1C_2).$$

We have chosen, however, to prove propositional transitivity for normal distributions because this choice allows us to unify the separability proof for two quite different classes of distributions, thus demonstrating the axiomatic approach.

5.4 Discussion

Our analysis of interaction, involved three steps; the probabilistic definition of the concept, the formalization of a requirement to be satisfied (Eq. 5.1), and finally, the identification of distributions for which the formal definition of interaction satisfies the requirement. We call these distributions *natural wrt interaction,* in the sense that they adequately represent the conventional properties of the word ''interact''.

Chapter 4 followed a similar line of reasoning; recoverability of causal relationships is formally defined, three properties: intersection, composition and marginal weak transitivity are found sufficient for the recovery of causal relationships, and distributions that satisfy these requirements are identified (e.g., normal); these distributions are considered natural in the sense that their structure can be recovered.

These examples indicate that the language of probabilistic dependencies is too weak to enforce properties that we normally attribute to a human reasoner. Consequently, when we employ probability for modeling a human reasoner, we must carefully select subsets of probability distributions that, on one hand, correctly represent high level concepts such as causation and interaction and, on the other hand, are sufficiently rich to express one's knowledge. Chapter 4 and 5 show that conditional independence together with graph-based representations provide appropriate tools for delineating these distributions.

REFERENCES

[1] S. Andreassen, M. Woldbye, B. Falck and S.K. Andersen. MUNIN - A causal probabilistic network for interpretation of electromyographic findings. *Proceedings of IJCAI.* Milan, Italy (1987) 366-372.

[2] S. Asmussen and D. Edwards. Collapsibility and response variables in contingency tables. *Biometrika* **70** (1983) 567-578.

[3] C. Beeri. On the membership problem for functional and multivalued dependencies in relational databases. *ACM Trans. Database Syst.* **5(3)** (1980) 241-249.

[4] C. Beeri, R. Fagin and J.H. Howard. A complete axiomatization of functional dependencies and multi valued dependencies in database relations. *Proceedings of the 1977 ACM SIGMOD Int. Conf. on Mgmt of Data.* Toronto, Canada (1977) 47-61.

[5] M. Ben-Bassat, R.W. Carlson, V.K. Puri, E. Lipnic, L.D. Portigal and M.H. Weil. Pattern-based interactive diagnosis of multiple disorders: The MEDAS system. *IEEE Trans. on Pattern Analysis & Machine Intelligence, (PAMI).* **2** (1980) 148-160.

[6] H.M. Blalock. *Causal Models in The Social Sciences.* London: Macmillan (1971).

[7] J.M. Bradshaw, S.P. Covington, P.J. Russo, and J.H. Boose. Knowledge acquisition techniques for intelligent decision systems: Integrating *Axotl* and *Aquinas* in *DDUCKS. Technical Report,* Advanced Technology Center, Boeing Computer Services, Seattle, Washington, 98124.

[8] R.M. Chavez, Hypermedia and randomized algorithms for probabilistic expert systems. Ph.D. thesis proposal, Knowledge systems laboratory, Stanford university, Stanford, CA. To appear in *networks.*

[9] R.M. Chavez, G.F. Cooper. An empirical evaluation of a randomized algorithm for probabilistic inference. *Proceedings of the 5th Workshop on Uncertainty in AI.* Windsor, Ontario (1989) 60-70.

[10] G.F. Cooper. NESTOR: A computer-based medical diagnostic aid that integrates causal and probabilistic knowledge. Ph.D. dissertation, Department of Computer Science, Stanford University (1984).

[11] G.F. Cooper. Current research directions in the development of expert systems based on belief networks. To appear in *Applied Stochastic Models and Data Analysis* (1989).

[12] G.F. Cooper. The computational complexity of probabilistic inference using belief networks. To appear in *Artificial Intelligence,* (1990).

[13] H.M. Cramér. *Mathematical Methods of Statistics.* Princeton: Princeton University Press (1946).

[14] J.N. Darroch, S.L. Lauritzen and T.P. Speed. Markov fields and loglinear interaction models for contingency tables. *Ann. Statist.* **8** (1980) 522-539.

[15] A.P. Dawid. Conditional independence in statistical theory. *J.R. Statist. Soc. B* **41:1** (1979) 1-31.

[16] R.O. Duda, P.E. Hart, P. Barnett, J. Gaschnig, K. Konolige, R. Reboh, and J. Slocum. Development of the PROSPECTOR consultant system for mineral exploration. Final Report for SRI Projects 5821 and 6915, Artificial Intelligence Center, SRI International (1978).

[17] O.D. Duncan. *Introduction to Structural Equation Models.* New York: Academic Press (1975).

[18] S. Even. *Graph algorithms.* Potomac, Md.: Computer Science Press (1979).

[19] R. Fagin. Functional dependencies in relational databases and propositional logic. Revision of IBM Research report RJ 1776, San Jose, California (1976).

[20] R. Fagin. Multivalued dependencies and a new normal form for relational databases. *ACM Trans. Database Syst.* **2 (3)** (1978) 262-278.

[21] R. Fagin. Horn clauses and database dependencies. *JACM* **29(4)** (1982) 952-985.

[22] R. Fagin and M. Vardi. The theory of data dependencies - a survey. *Proceedings of the Symposium in Applied Math.* **34** (1986).

[23] M. Frydenberg. Marginalization and collapsibility in graphical association models, Res. Rep. No. 166, Department of Theoretical Statistics, Arhus University (1988).

[24] D. Geiger and D. Heckerman. Interaction models. UCLA Cognitive Systems Laboratory, *Technical Report R-141.* In preparation.

[25] D. Geiger, A. Paz and J. Pearl. Identifying polytrees of compositional graphoids. UCLA, Department of Computer Science, Cognitive Systems Laboratory, *Technical Report R-140.* In preparation.

[26] D. Geiger, A. Paz and J. Pearl. Axioms and algorithms for inferences involving probabilistic independence. To appear in *Information and Computation* (1990).

[27] D. Geiger and J. Pearl. Logical and algorithmic properties of conditional independence. UCLA, Department of Computer Science, Cognitive Systems Laboratory, *Technical Report 870056 (R-97),* February (1988) (Submitted).

[28] C. Glymour, R. Scheines, P. Spirtes and K. Kelly. *Discovering Causal Structure.* New York: Academic Press (1987).

[29] I.J. Good. A causal calculus. *Philosophy of Science* **11** (1961) 305-18. Also in Good [1983].

[30] D. Heckerman. Probabilistic interpretations for MYCIN's certainty factors. In *Uncertainty in Artificial Intelligence,* ed. L. N. Kanal and J. F. Lemmer, (1986) 167-96. Amsterdam: North Holland.

[31] D. Heckerman. Probabilistic similarity networks. To appear in *Networks* (1990). Also in [32].

[32] D. Heckerman. Probabilistic similarity networks. Ph.d dissertation, Department of Computer Science and Medicine, Stanford University (1989). In preparation.

[33] M. Henrion. Propagating uncertainty by logic sampling in Bayes' networks. Technical Report, Department of Engineering and Public Policy, Carnegie-Mellon University (1986).

[34] E.J. Horovitz, J.S. Breese, and M. Henrion. Decision theory in expert systems and artificial intelligence. To appear in *Journal of Approximate Reasoning* (1990).

[35] R.A. Howard and J.E. Matheson. Influence diagrams. In *Principles and Applications of Decision Analysis.* **2** (1984) Menlo Park, Ca.: Strategic Decisions Group.

[36] D. Hunter. Graphoids, semi-graphoids and ordinal conditional independence. In preparation (1989).

[37] V. Isham. An introduction to spatial point processes and Markov random fields.

International Statistical Review **49** (1981) 21-43.

[38] H. Kiiveri, T.P. Speed and J.B. Carlin. Recursive causal models. *Journal of Australian Math Society* **36** (1984) 30-52.

[39] S. L. Lauritzen. *Lectures on Contingency Tables*. 2nd Ed., Aalborg, Denmark: University of Aalborg Press (1982).

[40] S.L. Lauritzen, A.P. Dawid, B.N., Larsen and H.G. Leimer. Independence properties of directed Markov fields. *Technical Report R 88-32*, Aalborg Universitets center, Aalborg Denmark , October (1988). To appear in *Networks*.

[41] S.L. Lauritzen and D.J. Spiegelhalter. Local computations with probabilities on graphical structures and their applications to expert systems. *J. Royal Statist. Soc., ser. B.* **50(2)** (1988) 154-227.

[42] Y.G. Lee and R.J. Buehler. Independence relationships for multivariate distributions. University of Minnesota, *Technical Report No. 464*. April (1986).

[43] T.S. Levitt. A Bayesian inference for radar imagery based surveillance. *Uncertainty in Artificial Intelligence 2,* Amsterdam: North-Holland (1988).

[44] S. M. Olmsted. On representing and solving decision problems. Ph.D. dissertation, EES Dept., Stanford University (1983).

[45] D.S. Parker and K. Parsaye-Ghomi. Inferences involving embedded multivalued dependencies and transitive dependencies. *Proceedings of the 1980 ACM SIGMOD Int. Conf. on Mgmt. of Data.* (1980) 52-57.

[46] A. Paz. A full characterization of pseudo-graphoids in terms of families of undirected graphs. UCLA, Cognitive Systems Laboratory, *Technical Report R-95* (1987).

[47] A. Paz. Membership algorithm for marginal independencies, UCLA Cognitive Systems Laboratory, *Technical Report R-117* (1988). Also in [26].

[48] J. Pearl. Fusion, propagation and structuring in belief networks, *Artificial Intelligence* **29(3)** (1986a) 241-288. Also in [52].

[49] J. Pearl. *Probabilistic Reasoning in Intelligent Systems.* San Mateo: Morgan-Kaufmann (1988).

[50] J. Pearl, D. Geiger, and T.S. Verma. The Logic of Influence Diagrams. *Proceedings* of the Berkeley Conference on Influence Diagrams, New York:

John Wiley & Sons Ltd (1989). Also a shorter version in *Kybernetica,* **25(2)**, (1989) 33-44.

[51] J. Pearl and A. Paz. Graphoids: A graph-based logic for reasoning about relevance relations. *Advances in Artificial Intelligence-II.* B. Du Boulay et al. (eds). Amsterdam: North-Holland (1987) 357-363.

[52] J. Pearl and M. Tarsi. Structuring causal trees. *Journal of Complexity* **2** (1986) 60-77.

[53] J. Pearl and T. Verma. The logic of representing dependencies by directed acyclic graphs. *Proceedings of the AAAI.* Seattle, Washington, July (1987) 374-379.

[54] Y. Sagiv and S. Walecka. Subset dependencies and completeness results for a subset of *EMVD* . *JACM* **29(1)** (1982) 103-117.

[55] R.D. Shachter. Probabilistic inference and influence diagrams. *Operations Research* **36** (1988) 589-604.

[56] R.D. Shachter. An ordered examination of influence diagrams. To appear in *Networks.*

[57] G. Shafer, P.P. Shenoy and K. Mellouli. Propagating belief functions in qualitative Markov trees. *International Journal of Approximate Reasoning* **1:4** (1988) 349-400.

[58] E. H. Shortliffe. *Computer-based medical consultation: MYCIN.* New York: Elsevier (1976).

[59] H. Simon. Spurious correlations: A causal interpretation. *J. Amer. Statist. Assoc.* **49** (1954) 469-92.

[60] B. Skyrms. Probability and causation. *Journal of Econometrics* **39** (1988) 53-68.

[61] J.Q. Smith. Influence diagrams for statistical modeling, *Annals of Statistics* **17(2)** (1989) 654-672.

[62] D.J. Spiegelhalter and S.L. Lauritzen. Sequential updating of conditional probabilities on directed graphical structures. To appear in *Networks.*

[63] P. Spirtes., C. Glymour and R. Scheines. Causality from probability. Department of Philosophy, Carnegie Mellon University, *Technical report CMU-LCL-*

89-4 (1989).

[64] W. Spohn. Stochastic independence, causal independence, and shieldability. In *Journal of Philosophical Logic* **9** (1980) 73-99.

[65] W. Spohn. Ordinal conditional functions: A dynamic theory of epistemic states. *Causation in Decision, Belief Change, and Statistics, II.* W. L. Harper and B. Skyrms (eds.) (1988) 105-134.

[66] W. Spohn. Direct and indirect causes. To appear in *Topoi* **9** (1990).

[67] M. Studeny. Attempts at axiomatic description of conditional independence. *Workshop on Uncertainty in Expert Systems.* Alsovice, Czechoslovakia, June 20-23 (1988).

[68] P. Suppes *A probabilistic theory of causation.* Amsterdam: North Holland (1970).

[69] R.E. Tarjan and M. Yannakakis. Simple linear-time algorithms to test chordality of graphs, test acyclicity of hypergraphs and selectively reduce acyclic hypergraphs. *SIAM J. Computing* **13** (1984) 566-79.

[70] M. Vardi. Fundamentals of dependency theory. In *Trends in Theoretical Computer Science.* E. Borger (ed). Rockville, Md: Computer Science Press (1988) 171-224.

[71] T. S. Verma. Some mathematical properties of dependency models. UCLA Cognitive Systems Laboratory, *Technical Report R-103* (1987).

[72] T.S. Verma and J. Pearl. Causal networks: semantics and expressiveness. *Proceedings of the 4th Workshop on Uncertainty in AI.* St. Paul, Minnesota (1988) 352-359.

[73] N. Wermuth. and S.L. Lauritzen. Graphical and recursive models for contingency tables. *Biometrika* **70** (1983) 537-552.

[74] H. Wold. *Econometric Model Building.* Amsterdam: North-Holland (1964).

[75] S. Wright. The method of path coefficients. *Ann. Math. Statist.* **5** (1934) 161-215.