Chapter 3

MARKOV AND BAYESIAN NETWORKS:

Two Graphical Representations of Probabilistic Knowledge

Probability is not really about numbers; it is about the structure of reasoning.
— G. Shafer

In this chapter, we shall seek effective graphic representations of the dependencies embedded in probabilistic models. First, we will uncover a set of axioms for the probabilistic relation "X is independent of Y, given Z" and offer the set as a formal definition for the notion of informational dependency. Given an initial set of independence relationships, the axioms permit us to infer new independencies by nonnumeric, logical manipulations. Using this axiomatic basis, we will identify structural properties of probabilistic models that can be captured by graphical representations and compare two such representations, *Markov networks* and *Bayesian networks*. A Markov network is an undirected graph whose links represent symmetrical probabilistic dependencies, while a Bayesian network is a directed acyclic graph whose arrows represent causal influences or class-property relationships. After establishing formal semantics for both network types, we shall explore their power and limitations as knowledge representation schemes in inference systems.

3.1 FROM NUMERICAL TO GRAPHICAL REPRESENTATIONS

3.1.1 Introduction

Scholarly textbooks on probability theory have created the impression that to construct an adequate representation of probabilistic knowledge, we must literally define a *joint distribution function* $P(x_1,..., x_n)$ on all propositions and their combinations, this function serving as the primary basis for all inferred judgments. While useful for maintaining consistency and proving mathematical theorems, this view of probability theory is totally inadequate for representing human reasoning.

Consider, for example, the problem of encoding an arbitrary joint distribution, $P(x_1,..., x_n)$, for n propositional variables. To store $P(x_1,..., x_n)$ explicitly would require a table with 2^n entries, an unthinkably large number by any standard. Even if we found some economical way of storing $P(x_1,..., x_n)$ —or rules for generating it—there would remain the problem of computing from it the probabilities of propositions people consider interesting. For example, computing the marginal probability $P(x_i)$ would require summing $P(x_1,..., x_n)$ over all 2^{n-1} combinations of the remaining $n-1$ variables. Similarly, computing the conditional probability $P(x_i \mid x_j)$ via its textbook definition

$$P(x_i \mid x_j) = \frac{P(x_i, x_j)}{P(x_j)}$$

would entail dividing two marginal probabilities, each a result of summation over an exponentially large number of variable combinations. Human performance shows the opposite pattern of complexity: probabilistic judgments on a small number of propositions (especially two-component conditional statements such as the likelihood that a patient suffering from a given disease will develop a certain type of complication) are issued swiftly and reliably, while judging the likelihood of a conjunction of propositions entails much difficulty and hesitancy. This suggests that the elementary building blocks of human knowledge are not entries of a joint-distribution table. Rather, they are low-order marginal and conditional probabilities defined over small clusters of propositions.

Another problem with purely numerical representations of probabilistic information is their lack of *psychological meaningfulness*. The numerical representation can produce coherent probability measures for all propositional sentences, but it often leads to computations that a human reasoner would not use. As a result, the process leading from the premises to the conclusions cannot be followed, tested, or justified by the users, or even the designers, of the reasoning

system. Even simple tasks such as computing the impact of a piece of evidence $E = e$ on a hypothesis $H = h$ via

$$P(h \mid e) = \frac{P(h, e)}{P(e)} = \frac{\sum_{x_i : X_i \neq H,E} P(x_1, \ldots, x_n)}{\sum_{x_i : X_i \neq E} P(x_1, \ldots, x_n)}$$

require a horrendous number of meaningless arithmetic operations, unsupported by familiar mental processes.

THE QUALITATIVE NOTION OF DEPENDENCE

The most striking inadequacy of traditional theories of probability lies in the way these theories address the notion of independence. The traditional definition of independence uses equality of numerical quantities, as in $P(x, y) = P(x) \cdot P(y)$, suggesting that one must test whether the joint distribution of X and Y is equal to the product of their marginals in order to determine whether X and Y are independent. By contrast, people can easily and confidently detect dependencies, even though they may not be able to provide precise numerical estimates of probabilities.

A person who is reluctant to estimate the probability of being burglarized the next day or of having a nuclear war within five years can nevertheless state with ease whether the two events are dependent, namely, whether knowing the truth of one proposition will alter the belief in the other. Likewise, people tend to judge the three-place relationship of conditional dependency (i.e., X influences Y, given Z) with clarity, conviction, and consistency. For example, knowing the time of the last pickup from a bus stop is undeniably relevant for assessing how long we must wait for the next bus. However, once we learn the whereabouts of the next bus, the previous knowledge no longer provides useful information. These commonsense judgments are issued qualitatively, without reference to numerical probabilities, and could not possibly rely on arithmetic manipulation of precise probabilities.

Evidently, the notions of relevance and dependence are far more basic to human reasoning than the numerical values attached to probability judgments. In a commonsense reasoning system, therefore, the language used for representing probabilistic information should allow assertions about dependency relationships to be expressed qualitatively, directly, and explicitly. The verification of dependencies should not require lengthy numerical manipulations but should be accomplished swiftly with a few primitive operations on the salient features of the representation scheme. Once asserted, these dependency relationships should remain a part of the representation scheme, impervious to variations in numerical inputs. For example, one should be able to assert categorically that a nuclear disaster is independent of a home burglary; the system should retain and reaffirm

this independence despite changes in the estimated likelihoods of these and other events in the system.

Making effective use of information about dependencies is essential in reasoning. If we have acquired a body of knowledge K and now wish to assess the truth of proposition A, it is important to know whether it is worthwhile to consult another proposition B, which is not in K. In other words, before we examine B, we need to know if its truth value can generate new information that is relevant to A and is not available from K. Without this knowledge, an inference engine might spend precious time on derivations bearing no relevance to the task at hand. Relevance information, if available, can confine the engine's attention to derivations that truly are needed for the target conclusion. But how can we encode relevance information in a symbolic system?

Explicit encoding is clearly impractical; the number of (A, B, K) combinations needed is astronomical, because relevance and dependency are relationships that vary depending on the information available at any given time. Acquisition of new facts may destroy existing dependencies as well as create new ones. For example, learning a child's age destroys the dependency between height and reading ability, and learning that a patient suffers from a given symptom creates new dependencies among the diseases that could account for the symptom. The first kind of change will be called *normal* as it fits the normal picture that learning reduces dependencies, and the second will be called *induced* as it permits learned facts to induce new dependencies. What logic would facilitate these two modes of reasoning?

In probability theory, the notion of informational relevance is given quantitative underpinning through the device of *conditional independence,* which successfully captures our intuition about how dependencies should change in response to new facts. A proposition A is said to be independent of B, given the information K, if

$$P(A \mid B, K) = P(A \mid K),$$

namely, if once K is given, the probability of A will not be affected by the discovery of B. This formulation can represent both normal and induced dependencies: A and B could be marginally dependent (i.e., dependent when K is unknown) and become conditionally independent given K; conversely, A and B could be marginally independent and become dependent given K. Thus, in principle, probability theory could provide the machinery for identifying the propositions that are relevant to each other under a given state of knowledge.

But we have already argued that it is unreasonable to expect people or machines seeking relevance information to resort to numerical equality tests. Human behavior suggests that relevance information is inferred qualitatively from the organizational structure of human memory, not calculated from numerical values assigned to its components. Accordingly, it would be interesting to explore how assertions about relevance can be inferred qualitatively, and whether

assertions equivalent to those made about probabilistic dependencies can be derived *logically* without reference to numerical quantities. This task will be discussed in Section 3.1.2, which establishes an axiomatic basis for probabilistic dependencies and examines whether the set of axioms matches our intuitive notion of informational relevancy.

WHY GRAPHS?

A logic of dependency might be useful for verifying whether a set of dependencies asserted by an agent is consistent and whether a new dependency follows from the initial set. We could not guarantee, however, that the verification would be tractable or that any sequence of inferences would match mental steps taken by humans. To facilitate psychological meaningfulness, we must make sure most derivations in the logic correspond to simple local operations on structures depicting commonsense associations. We call such structures *dependency graphs.*

The nodes in these graphs represent propositional variables, and the arcs represent local dependencies among conceptually related propositions. Graph representations meet our earlier requirements of explicitness, saliency, and stability. The links in the graph permit us to express directly and qualitatively the dependence relationships, and the graph topology displays these relationships explicitly and preserves them, under any assignment of numerical parameters.

It is not surprising, therefore, that graphs are the most common metaphor for conceptual dependencies. Models of human memory are often portrayed in terms of associational graphs (e.g., semantic networks [Woods 1975], constraint networks [Montanari 1974], inference networks [Duda, Hart, and Nilsson 1976], conceptual dependencies [Schank 1972], and conceptual structures [Sowa 1984]). Graph concepts are so entrenched in our language (e.g., "threads of thoughts," "lines of reasoning," "connected ideas," "far-fetched arguments") that one wonders if people can reason any other way except by tracing links and arrows and paths in some mental representation of concepts and relations. The next question to ask is what aspects of informational relevance and probabilistic dependence can be represented graphically. In other words, what types of dependencies and independencies are deducible from the topological properties of a graph? This question will be addressed in Sections 3.2 (undirected graphs) and 3.3 (directed graphs).

Despite the prevailing use of graphs as metaphors for communicating and reasoning about dependencies, the task of capturing informational dependencies by graphs is not at all trivial. We have no problem configuring a graph which represents phenomena with explicit notions of neighborhood or adjacency (e.g., families, electronic circuits, communication networks). However, in modeling conceptual relations, such as causation, association, and relevance, it is often hard to distinguish direct neighbors from indirect neighbors; constructing a graph for the relation therefore becomes more delicate. The notion of conditional independence in probability theory is a perfect example. For a given probability

distribution P and any three variables X, Y, Z, it is straightforward to verify whether knowing Z renders X independent of Y, but P does not dictate which variables should be regarded as direct neighbors. Thus, many different topologies might be used to display P's dependencies. We shall also see that some useful properties of dependencies and relevancies cannot be represented graphically. The challenge is to devise graphical schemes that minimize these deficiencies; Markov and Bayesian networks are two such schemes.

CHAPTER OVERVIEW

This chapter is organized as follows: Section 3.1.2 uncovers a set of axioms for the probabilistic relation "X is independent of Y, given Z" and offers the set as a formal definition for the notion of informational dependency. Sections 3.1.3 and 3.1.4 examine those properties of dependencies that can be captured by graphical representations. Sections 3.2 and 3.3 compare two such representations, *Markov networks* and *Bayesian networks*. For both network types, we shall establish (1) a formal description of the dependencies portrayed by the networks, (2) an axiomatic description of the class of dependencies that can be captured by the network, (3) methods of constructing the network from either hard data or subjective judgments, and (4) a summary of properties relevant to the network's use as a knowledge representation scheme.

3.1.2 An Axiomatic Basis for Probabilistic Dependencies

NOTATION AND DEFINITIONS

We will consider a finite set $\boldsymbol{U}$ of discrete random variables (also called partitions or attributes), where each variable $X \in \boldsymbol{U}$ may take on values from a finite domain D_X. We will use capital letters for variable names (e.g., X, Y, Z) and lowercase letters (e.g., x, y, z) for specific values taken by variables. Sets of variables will be denoted by boldfaced capital letters (e.g., $\boldsymbol{X}$, $\boldsymbol{Y}$, $\boldsymbol{Z}$), and assignments of values to the variables in these sets (also called *configurations*), will be denoted by boldfaced lowercase letters (e.g., $\boldsymbol{x}$, $\boldsymbol{y}$, $\boldsymbol{z}$). For example, if $\boldsymbol{Z}$ stands for the set of variables $\{X, Y\}$, then $\boldsymbol{z}$ represents the configuration $\{x, y\}: x \in D_X, y \in D_Y$. When the distinction between variables and sets of variables requires special emphasis, Greek letters α, β, γ,... will be used to represent individual variables.

We shall repeatedly use the short notation $P(x)$ for the probabilities $P(X = x)$, $x \in D_X$, and we will write $P(z)$ for the set of variables $\boldsymbol{Z} = \{X, Y\}$, meaning

$$P(\boldsymbol{Z} = \boldsymbol{z}) = P(X = x, Y = y) \qquad x \in D_X, y \in D_Y.$$

(In the rare event that we run out of symbols, variable names will be used as arguments of probability statements, e.g., $P(X, Y)$, which is equivalent to $P(x, y)$.)

DEFINITION: *Let* $\mathbf{U} = \{\alpha, \beta, \ldots\}$ *be a finite set of variables with discrete values. Let* $P(\cdot)$ *be a joint probability function over the variables in* $\mathbf{U}$*, and let* $\mathbf{X}$*,* $\mathbf{Y}$*, and* $\mathbf{Z}$ *stand for any three subsets of variables in* $\mathbf{U}$*.* $\mathbf{X}$ *and* $\mathbf{Y}$ *are said to be* ***conditionally independent given*** $\mathbf{Z}$ *if*

$$P(\boldsymbol{x} \mid \boldsymbol{y}, \boldsymbol{z}) = P(\boldsymbol{x} \mid \boldsymbol{z}) \text{ whenever } P(\boldsymbol{y}, \boldsymbol{z}) > 0. \tag{3.1}$$

Eq. (3.1) is a terse way of saying the following: for any configuration $\boldsymbol{x}$ of the variables in the set $\boldsymbol{X}$ and for any configurations $\boldsymbol{y}$ and $\boldsymbol{z}$ of the variables in $\boldsymbol{Y}$ and $\boldsymbol{Z}$ satisfying $P(\boldsymbol{Y} = \boldsymbol{y}, \boldsymbol{Z} = \boldsymbol{z}) > 0$, we have

$$P(\boldsymbol{X} = \boldsymbol{x} \mid \boldsymbol{Y} = \boldsymbol{y}, \boldsymbol{Z} = \boldsymbol{z}) = P(\boldsymbol{X} = \boldsymbol{x} \mid \boldsymbol{Z} = \boldsymbol{z}). \tag{3.2}$$

We will use the notation $I(\boldsymbol{X},\boldsymbol{Z},\boldsymbol{Y})_P$ or simply $I(\boldsymbol{X},\boldsymbol{Z},\boldsymbol{Y})$ to denote the conditional independence of $\boldsymbol{X}$ and $\boldsymbol{Y}$ given $\boldsymbol{Z}$; thus,

$$I(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{Y})_P \text{ iff } P(\boldsymbol{x} \mid \boldsymbol{y}, \boldsymbol{z}) = P(\boldsymbol{x} \mid \boldsymbol{z}) \tag{3.3}$$

for all values $\boldsymbol{x}$, $\boldsymbol{y}$, and $\boldsymbol{z}$ such that $P(\boldsymbol{y}, \boldsymbol{z}) > 0$. Unconditional independence (also called *marginal independence*) will be denoted by $I(\boldsymbol{X}, \varnothing, \boldsymbol{Y})$, i.e.,

$$I(\boldsymbol{X}, \varnothing, \boldsymbol{Y}) \text{ iff } P(\boldsymbol{x} \mid \boldsymbol{y}) = P(\boldsymbol{x}) \text{ whenever } P(\boldsymbol{y}) > 0. \tag{3.4}$$

Note that $I(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{Y})$ implies the conditional independence of all pairs of variables $\alpha \in \boldsymbol{X}$ and $\beta \in \boldsymbol{Y}$, but the converse is not necessarily true.

The following is a partial list of (equivalent) properties satisfied by the conditional independence relation $I(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{Y})$ [Lauritzen 1982]:

$$I(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{Y}) \Longleftrightarrow P(\boldsymbol{x}, \boldsymbol{y} \mid \boldsymbol{z}) = P(\boldsymbol{x} \mid \boldsymbol{z})\, P(\boldsymbol{y} \mid \boldsymbol{z}), \tag{3.5a}$$

$$I(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{Y}) \Longleftrightarrow \exists f, g: P(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) = f(\boldsymbol{x}, \boldsymbol{z})\, g(\boldsymbol{y}, \boldsymbol{z}), \tag{3.5b}$$

$$I(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{Y}) \Longleftrightarrow P(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) = P(\boldsymbol{x} \mid \boldsymbol{z})\, P(\boldsymbol{y}, \boldsymbol{z}). \tag{3.5c}$$

The proof of these properties can be derived by elementary means from Eq. (3.3) and the basic axioms of probability theory. The properties are based on the numeric representation of P and therefore would be inadequate as an axiomatic system.

AXIOMATIC CHARACTERIZATION

We now ask what logical conditions, void of any reference to numerical forms, should constrain the relationship $I(X, Z, Y)$ if in some probability model P it stands for the statement "X is independent of Y, given that we know Z."

THEOREM 1: *Let X, Y, and Z be three disjoint subsets of variables from U. If $I(X, Z, Y)$ stands for the relation "X is independent of Y, given Z" in some probabilistic model P, then I must satisfy the following four independent conditions:*

- *Symmetry:*
$$I(X, Z, Y) \Longleftrightarrow I(Y, Z, X) \tag{3.6a}$$

- *Decomposition:*
$$I(X, Z, Y \cup W) \Longrightarrow I(X, Z, Y) \;\&\; I(X, Z, W) \tag{3.6b}$$

- *Weak Union:*
$$I(X, Z, Y \cup W) \Longrightarrow I(X, Z \cup W, Y) \tag{3.6c}$$

- *Contraction:*
$$I(X, Z, Y) \;\&\; I(X, Z \cup Y, W) \Longrightarrow I(X, Z, Y \cup W). \tag{3.6d}$$

If P is strictly positive, then a fifth condition holds:

- *Intersection:*
$$I(X, Z \cup W, Y) \;\&\; I(X, Z \cup Y, W) \Longrightarrow I(X, Z, Y \cup W). \tag{3.6e}$$

REMARKS:

1. The symbol $\cup$ in $Y \cup W$ represents a union of variable sets and should not be confused with logical disjunction. More specifically, it stands for the *conjunction* of events asserted by instantiating the set union $Y \cup W$. For example, $I(X, \varnothing, Y \cup W)$ stands for
$$P(X = x, Y = y, W = w) = P(X = x)\, P(Y = y, W = w) \;\; \forall\, x, y, w.$$
A simpler notation, $I(X, \varnothing, YW)$, will occasionally be used.

2. The requirement that the arguments of $I(\cdot)$ be disjoint was made for the sake of future clarity. Theorem 1 can be extended to include overlapping subsets as well, using an additional axiom,
$$I(X, Z, Z). \tag{3.6f}$$

From Eqs. (3.6*a*) through (3.6*d*) and Eq. (3.6*f*) one can prove the theorem

$$I(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{Y}) \Longleftrightarrow I(\boldsymbol{X} - \boldsymbol{Z}, \boldsymbol{Z}, \boldsymbol{Y} - \boldsymbol{Z}),$$

stating that the parts of $\boldsymbol{X}$ and $\boldsymbol{Y}$ that do not overlap $\boldsymbol{Z}$ are sufficient to determine whether $I(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{Y})$ holds. Thus, once $I(\cdot)$ is defined on the set of disjoint triplets $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$ it is also defined on the set of all triplets. Note that both $I(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{Z})$ and $I(\boldsymbol{X}, \boldsymbol{Z}, \varnothing)$ follow from Eq. (3.3).

3. The proof of Theorem 1 can be derived from Eq. (3.3) and from the basic axioms of probability theory [Dawid 1979]. That Eqs. (3.6*a*) through (3.6*e*) are logically independent can be demonstrated by letting $\boldsymbol{U}$ contain four elements and showing that it is always possible to contrive a subset I of triplets (from the subsets of $\boldsymbol{U}$) that violates one property and satisfies the other four.

INTUITIVE INTERPRETATION OF THE AXIOMS

Eqs. (3.6*a*) through (3.6*e*) can be interpreted as follows: The *symmetry* axiom states that in any state of knowledge $\boldsymbol{Z}$, if $\boldsymbol{Y}$ tells us nothing new about $\boldsymbol{X}$, then $\boldsymbol{X}$ tells us nothing new about $\boldsymbol{Y}$. The *decomposition* axiom asserts that if two combined items of information are judged irrelevant to $\boldsymbol{X}$, then each separate item is irrelevant as well. The *weak union* axiom states that learning irrelevant information $\boldsymbol{W}$ cannot help the irrelevant information $\boldsymbol{Y}$ become relevant to $\boldsymbol{X}$. The *contraction* axiom states that if we judge $\boldsymbol{W}$ irrelevant to $\boldsymbol{X}$ after learning some irrelevant information $\boldsymbol{Y}$, then $\boldsymbol{W}$ must have been irrelevant before we learned $\boldsymbol{Y}$. Together, the weak union and contraction properties mean that irrelevant information should not alter the relevance of other propositions in the system; what was relevant remains relevant, and what was irrelevant remains irrelevant. The *intersection* axiom states that unless $\boldsymbol{Y}$ affects $\boldsymbol{X}$ when $\boldsymbol{W}$ is held constant or $\boldsymbol{W}$ affects $\boldsymbol{X}$ when $\boldsymbol{Y}$ is held constant, neither $\boldsymbol{W}$ nor $\boldsymbol{Y}$ nor their combination can affect $\boldsymbol{X}$.

GRAPHICAL INTERPRETATIONS

The operational significance of these axioms and their role as inference rules can best be explained with a graph metaphor. Let $I(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{Y})$ stand for the phrase "$\boldsymbol{Z}$ separates $\boldsymbol{X}$ from $\boldsymbol{Y}$," i.e., "The removal of a set $\boldsymbol{Z}$ of nodes from the graph (together with their associated edges) would render the nodes in $\boldsymbol{X}$ disconnected from those in $\boldsymbol{Y}$." The validity of Eqs. (3.6*a*) through (3.6*e*) is clearly depicted by the chain X–Z–Y–W and by the schematics of Figure 3.1.

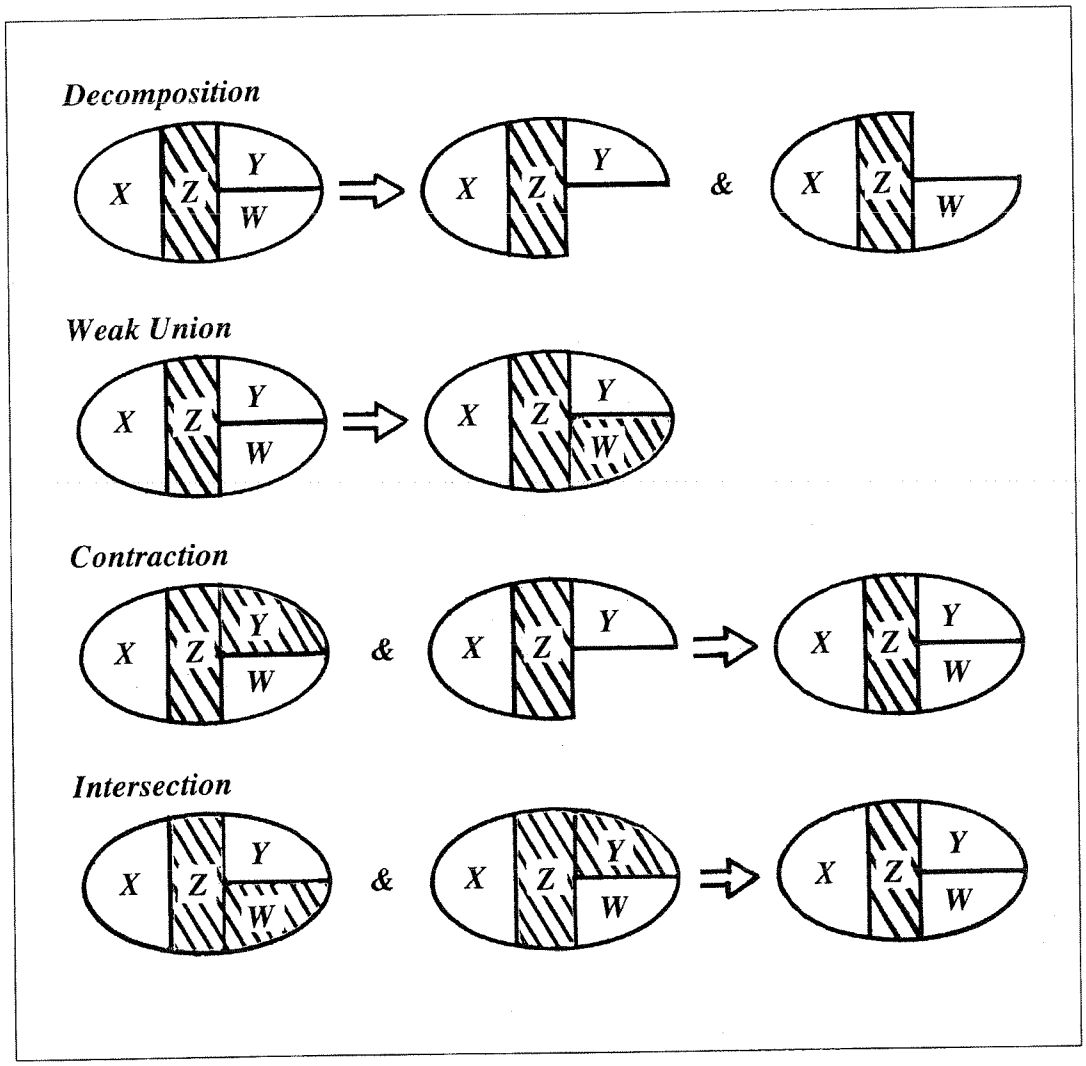


Figure 3.1. *Graphical interpretation of the axioms governing conditional independence.*

Symmetry simply states that if $\mathbf{Z}$ separates $\mathbf{X}$ from $\mathbf{Y}$, it also separates $\mathbf{Y}$ from $\mathbf{X}$. Decomposition asserts that if $\mathbf{Z}$ separates $\mathbf{X}$ from the compound set $\mathbf{S} = \mathbf{Y} \cup \mathbf{W}$, it also separates $\mathbf{X}$ from every subset of $\mathbf{S}$. Weak union provides conditions under which a separating set $\mathbf{Z}$ can be augmented by additional elements $\mathbf{W}$ and still separate $\mathbf{X}$ from $\mathbf{Y}$. The condition is that the added subset $\mathbf{W}$ should come from the section of space that was initially separated from $\mathbf{X}$ by $\mathbf{Z}$. Contraction provides conditions for reducing the size of the separating set; it permits the deletion of a subset $\mathbf{Y}$ from the separator $\mathbf{Z} \cup \mathbf{Y}$ if the remaining part, $\mathbf{Z}$, separates the deleted part, $\mathbf{Y}$, from $\mathbf{X}$. Intersection states that if within some set of variables $\mathbf{S} = \mathbf{X} \cup \mathbf{Y} \cup \mathbf{Z} \cup \mathbf{W}$, $\mathbf{X}$ can be separated from the rest of $\mathbf{S}$ by two different subsets, $\mathbf{S}_1$ and $\mathbf{S}_2$ (i.e., $\mathbf{S}_1 = \mathbf{Z} \cup \mathbf{Y}$ and $\mathbf{S}_2 = \mathbf{Z} \cup \mathbf{W}$), then the intersection of $\mathbf{S}_1$ and $\mathbf{S}_2$ is sufficient to separate $\mathbf{X}$ from the rest of $\mathbf{S}$.

THE INTERSECTION AXIOM AND STRICTLY POSITIVE DISTRIBUTIONS

The intersection axiom is the only one that requires $P(\boldsymbol{x}) > 0$ for all $\boldsymbol{x}$, and it will not hold if the variables in $\boldsymbol{U}$ are constrained by logical dependencies. For instance, if Y stands for the proposition "The water temperature is above freezing" and W stands for "The water temperature is above 32°F," then knowing the truth of either proposition clearly renders the other superfluous. Contrary to the intersection axiom, however, Y and W might still be relevant to a third proposition X ("We will enjoy swimming in that water," for example). The intersection axiom will hold if we regard these logical constraints as having some small probability ε of being violated.

The assumption $P \geq \varepsilon > 0$ means every event or combination of events, no matter how outrageous, has some chance of being true. When examining empirical facts, making this assumption is not as strange as it seems. For example, it is possible for the water temperature to be above freezing but below 32°F (if it is very salty, for instance). Once we accept such a possibility we must reject the statement that knowing either of these facts renders the other superfluous relative to any X. If X represents our concern about swimming in the water, then the temperature becomes the relevant fact, and whether it is frozen is irrelevant. On the other hand, if our interest is ice fishing, the frozenness, not the temperature, is relevant. This is exactly what Eq. (3.6*e*) claims: if two properties exert influence on X, then (at a sufficiently high level of detail) it is impossible that each of the two properties will render the other irrelevant. Such symmetrical exclusion is possible only with analytical or definitional properties (e.g., Y = "The water temperature is above 32°F," W = "The water temperature is not equal to or lower than 32°F") and not with properties defined by independent empirical tests.

GRAPHS VS. GRAPHOIDS

Decomposition and weak union are strikingly similar to vertex separation in graphs, but are much weaker. In graphs, two sets of vertices are said to be separated if there exists no path between an element of one set and an element of the other. The decomposition property (Eq. (3.6*b*)), on the other hand, reflects only one-way implication; a variable X may be independent of each individual variable in set $\boldsymbol{Y}$ and still be dependent on the entire set. For example, let $\boldsymbol{Y}$ be the outcomes of a set of fair coins, and let X be a variable that gets the value 1 whenever an even number of coins turn up ''heads'' and gets 0 otherwise. X is independent of every element and every proper subset of $\boldsymbol{Y}$, yet X is completely determined by the entire set $\boldsymbol{Y}$. Weak union is also weaker than vertex separation. If $\boldsymbol{Z}$ is a cutset of vertices that separates $\boldsymbol{X}$ from $\boldsymbol{Y}$ in some graph, then enlarging $\boldsymbol{Z}$ keeps $\boldsymbol{X}$ and $\boldsymbol{Y}$ separated. Weak union, on the other hand, severely restricts the conditions under which a separating set $\boldsymbol{Z}$ can be enlarged with elements $\boldsymbol{W}$; it

states that W should be chosen from a set that, like Y, is already separated from X by Z.

Any three-place relation $I(\cdot)$ that satisfies Eqs. (3.6*a*) through (3.6*d*) is called a *semi-graphoid.* If it also obeys Eq. (3.6*e*), it is called a *graphoid* [Pearl and Paz 1985]. Eqs. (3.6*a*) through (3.6*d*) are satisfied by many dependency models. Besides vertex separation in undirected graphs, they also hold in directed graphs (see Section 3.3), and they govern information dependencies based on partial correlations [Pearl and Paz 1985], embedded multi-valued dependencies (EMVDs) in relational databases [Fagin 1977], and qualitative constraints [Shafer, Shenoy, and Mellouli 1988]. Because of this generality, the semi-graphoid axioms have been proposed as the basis of information dependencies.

Qualitative formulations of dependencies are accompanied by extra properties, whereas the probabilistic formulation seems to be completely characterized by these four axioms and therefore is more general. This observation can be expressed more formally.

COMPLETENESS CONJECTURE [Pearl and Paz 1985]: *The set of axioms in Eqs. (3.6a) through (3.6d) is* ***complete*** *when I is interpreted as a conditional independence relation. In other words, for every three-place relation I satisfying Eqs. (3.6a) through (3.6d), there exists a probability model P such that*

$$P(x \mid y, z) = P(x \mid z) \quad \textit{iff} \quad I(X, Z, Y).$$

If the intersection axiom (Eqs. (3.6e)) also is satisfied, then there exists a positive P satisfying the above relation.

While no proof has yet been found for this conjecture, all known properties of conditional independence (those valid for all P) have been shown to be derivable from Eqs. (3.6*a*) through (3.6*d*). A thorough treatment of the completeness problem, as well as completeness results for special types of probabilistic dependencies, are given by Geiger and Pearl [1988a].

WHY AXIOMATIC CHARACTERIZATION?

Axiomatizing the notion of probabilistic dependence is useful for three reasons. First, it allows us to conjecture and derive interesting and powerful theorems that may or may not be obvious from the numerical representation of probabilities. For example, the chaining rule [Lauritzen 1982],

$$I(X, Y, Z) \,\&\, I(X \cup Y, Z, W) \Longrightarrow I(X, Y, W),$$

follows directly from Eqs. (3.6*a*) through (3.6*d*) and is important for recursively constructing directed graph representations (see Section 3.3). Another interesting theorem is the mixing rule [Dawid 1979],

$$I(X, Z, Y \cup W) \,\&\, I(Y, Z, W) \implies I(X \cup W, Z, Y),$$

which also follows from Eqs. (3.6*a*) through (3.6*d*). The mixing rule, with symmetry and decomposition, constitutes a complete axiomatization of marginal independencies, i.e., independence statements where the knowledge set $\boldsymbol{Z}$ is fixed [Geiger and Pearl 1988a]. The rule states that for each of the variables X, Y, W to be independent of the other two, it is enough that just one of them be independent of the other two and that the remaining pair be mutually independent. Generalizing recursively to n variables, the rule states that for n variables to be mutually independent, it is enough that one of them be independent of the other $n-1$, and that the remaining $n-1$ be mutually independent.

Second, the axioms can be viewed as qualitative inference rules used to derive new independencies from some initial set. For example, an expert might provide us with an initial set Σ of qualitative independence judgments in the form of triplets (X, Z, Y), and we may wish to test whether a new triplet $\sigma = (X', Z', Y')$ follows from Σ. This task, called the *membership problem* [Beeri 1980] may in principle be undecidable, because to test whether σ follows from Σ we must test whether σ holds in every distribution that satisfies Σ, and the number of distributions is infinite. If, however, we can derive σ by repeated application of sound axioms, we can guarantee that σ follows from Σ without searching the vast space of probability distributions. If, in addition, the set of axioms is complete, we are also guaranteed that every σ that follows from Σ eventually will be derived from Σ by repeated application of the axioms. In other words, the decidability of the membership problem hinges upon finding a complete set of axioms for conditional independence. Closely related to the membership problem is the task of verifying whether a mixed set Σ' of dependencies and independencies is *consistent*, namely, whether no subset of Σ' implies the negation of another. Thus, with a sound and efficient inference mechanism we can test and maintain consistency in a database of dependency information.

Finally, an axiomatic system provides a parsimonious and convenient code for comparing the features of several formalisms of dependency (e.g., probabilistic vs. qualitative) as well as the expressive power of various representations of such formalisms. In Sections 3.2 and 3.3, for example, we will use the axioms to compare the expressive powers of directed and undirected graphs, and to reveal what types of dependencies cannot be captured by graphical representations.

SUMMARY

The probabilistic relation of conditional independence possesses a set of qualitative properties that are consistent with our intuitive notion of "$\boldsymbol{X}$ is irrelevant to $\boldsymbol{Y}$, once we learn $\boldsymbol{Z}$." These properties, which are also satisfied by vertex separation in graphs, are captured by the axioms in Eq. (3.6). The defining axioms convey the idea that when we learn an irrelevant fact, the relevance relationships among other propositions remain unaltered; any information that was relevant remains relevant, and irrelevant information remains irrelevant. The

axioms established can be used as inference rules for deriving new independencies and for defining the common features among various formalisms of dependence.

3.1.3 On Representing Dependencies by Undirected Graphs

WHAT'S IN A MISSING LINK?

Suppose we have a collection $\boldsymbol{U} = \{\alpha, \beta, \ldots\}$ of interacting elements, and we decide to represent their interactions by an undirected graph G, in which the nodes correspond to individual elements of $\boldsymbol{U}$. Naturally, we would like to display independence between two elements as a lack of connection between their corresponding nodes in G; conversely, dependent elements should correspond to connected nodes in G. This requirement alone, however, does not take full advantage of the expressive power of graphical representation. It treats all connected components of G as equivalent and does not attribute any special significance to the structure of each connected component.

Clearly, if graph topology is to convey meaning beyond connectedness, a semantic distinction must be drawn between *direct* connection and *indirect* connection. This means that the absence of a direct link between two elements α and β should reflect an interaction that is *conditional*, i.e., it may become stronger, weaker, or zero, depending on the state of other elements in the system, especially those that lie on the paths connecting α and β and thus *mediate* between them.

As an example, consider a group of two males $\{M_1, M_2\}$ and two females $\{F_1, F_2\}$ who occasionally engage in pairwise heterosexual activities. The lack of direct contact between the two males and between the two females can be represented by the diamond-shaped graph of Figure 3.2, which can also be used to represent conditional dependencies between various propositions. For example, if by m_i (or f_i) we denote the proposition that male M_i (or female F_i) will carry a certain disease within a year, then the topology of the network in Figure 3.2 asserts that f_1 and f_2 are independent given m_1 and m_2, namely, once we know for sure whether M_1 and M_2 will carry the disease, knowing the truth of f_1 ought not change our belief in f_2.†

† This assumes, of course, that we are dealing with a known disease whose spreading mechanism is well understood. Otherwise, while we are still learning the disease characteristic, knowledge of f_1 may help decide the more basic question of whether the disease is contagious at all, and this information will and should have an effect on f_2.

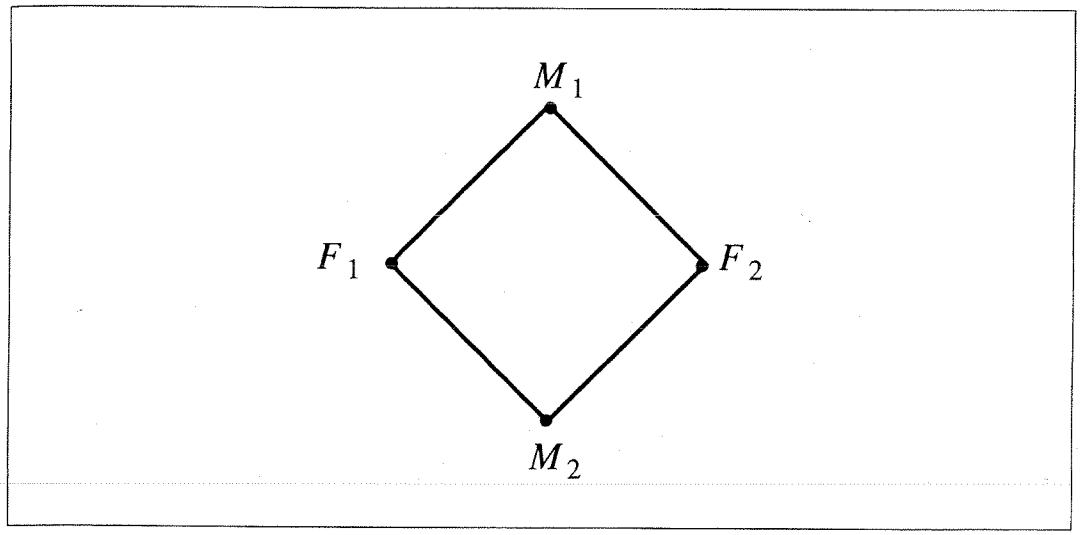


Figure 3.2. *An undirected graph representing interactions among four individuals.*

This conditional independence reflects a model whereby the disease spreads only by direct contact. Note that the links in this network are undirected, namely, either partner might be the originator of the disease. This does not exclude asymmetric interactions (e.g., the disease may be more easily transferable from males to females than the other way around). Such information, if available, will be contained in the numerical parameters that eventually will characterize the links in the network—they will be described in Section 3.2.3.

In summary, the semantics of the graph topology are defined by the meaning of the missing links, which tells us what other elements mediate the interactions between nonadjacent elements. This process of mediation will now be compared to the probabilistic relation of conditional independence $I(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{Y})$, Eq. (3.1), which formalizes the intuitive statement "Knowing $\boldsymbol{Y}$ tells me nothing new about $\boldsymbol{X}$ if I already know $\boldsymbol{Z}$."

DEPENDENCY MODELS AND DEPENDENCY MAPS

Let $\boldsymbol{U} = \{\alpha, \beta, \ldots\}$ be a finite set of elements (e.g., propositions or variables), and let $\boldsymbol{X}$, $\boldsymbol{Y}$, and $\boldsymbol{Z}$ stand for three disjoint subsets of elements in $\boldsymbol{U}$. Let M be a *dependency model,* that is to say, a rule that assigns truth values to the three-place predicate $I(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{Y})_M$, or in other words determines a subset I of triplets $(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{Y})$ for which the assertion "$\boldsymbol{X}$ is independent of $\boldsymbol{Y}$ given $\boldsymbol{Z}$" is true. Any probability distribution P is a dependency model, because for any triplet $(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{Y})$ we can test the validity of $I(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{Y})$ using Eq. (3.1). Our task is to characterize the set of dependency models capturable by graphs, including models that provide no explicit notion of adjacency. In other words, we are given the means to test whether a given subset $\boldsymbol{Z}$ of elements *intervenes* in a relation between the elements

of X and those of Y, but it is up to us to decide how to connect the elements in a graph that encodes these interventions.

An undirected graph $G = (V, E)$ is characterized by a set V of nodes (or vertices) and a set E of edges that connect certain pairs of nodes in V. By a *graphical representation* of a dependency model M, we mean a direct correspondence between the elements in U (of M) and the set of vertices in V (of G), such that the topology of G reflects some properties of M. When this correspondence is established, we will make no distinction between U and V but will write $G = (U, E)$.

Ideally, if a subset Z of nodes in a graph G intercepts all paths between the nodes of X and those of Y (written $< X \mid Z \mid Y >_G$), then this interception should correspond to conditional independence between X and Y given Z, namely,

$$< X \mid Z \mid Y >_G \implies I(X, Z, Y)_M,$$

and conversely,

$$I(X, Z, Y)_M \implies < X \mid Z \mid Y >_G.$$

This correspondence would provide a clear graphical representation for the notion that X does not affect Y directly, that the variables in Z mediate between them. Unfortunately, we are about to see that these two requirements are too strong; there often is no way of using vertex separation in a graph to display *all* dependencies and independencies embodied in a dependency model, even if the model portrays simple, everyday experiences.

DEFINITION: *An undirected graph G is a* ***dependency map*** *(or D-map) of M if there is a one-to-one correspondence between the elements of* U *and the nodes* V *of G, such that for all disjoint subsets* X, Y, Z *of elements we have*

$$I(X, Z, Y)_M \implies < X \mid Z \mid Y >_G. \tag{3.7}$$

Similarly, G is an ***independency map*** *(or I-map) of M if*

$$I(X, Z, Y)_M \impliedby < X \mid Z \mid Y >_G. \tag{3.8}$$

G is said to be a ***perfect map*** *of M if it is both a D-map and an I-map.*

A D-map guarantees that vertices found to be connected are indeed dependent in M (from the contrapositive form of Eq. (3.7)); it may, however, display a pair of dependent variables as a pair of separated vertices. An I-map, conversely, guarantees that vertices found to be separated correspond to independent variables but does not guarantee that all those shown to be connected are in fact dependent. Empty graphs are trivial D-maps, while complete graphs are trivial I-maps.

It is clear that many reasonable models of dependency have no perfect maps. An example is a model in which $I(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{Y})$ exhibits *induced dependencies*, i.e., totally unrelated propositions become relevant to each other when we learn new facts. Such a model, implying both $I(\boldsymbol{X}, \boldsymbol{Z}_1, \boldsymbol{Y})_M$ and $\neg I(\boldsymbol{X}, \boldsymbol{Z}_1 \cup \boldsymbol{Z}_2, \boldsymbol{Y})_M$, cannot have a graph representation that is both an I-map and a D-map, because graph separation always satisfies

$$< \boldsymbol{X} \mid \boldsymbol{Z}_1 \mid \boldsymbol{Y} >_G \Longrightarrow < \boldsymbol{X} \mid \boldsymbol{Z}_1 \cup \boldsymbol{Z}_2 \mid \boldsymbol{Y} >_G$$

for any two subsets $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$ of vertices. Thus, being a D-map requires G to display $\boldsymbol{Z}_1$ as a cutset separating $\boldsymbol{X}$ and $\boldsymbol{Y}$, while G's being an I-map prevents $\boldsymbol{Z}_1 \cup \boldsymbol{Z}_2$ from separating $\boldsymbol{X}$ and $\boldsymbol{Y}$. No graph can satisfy both requirements simultaneously.

This weakness in the expressive power of undirected graphs severely limits their ability to represent informational dependencies. Consider an experiment with two coins and a bell that rings whenever the outcomes of the two coins are the same. If we ignore the bell, the coin outcomes, X and Y, are mutually independent, i.e., $I(X, \varnothing, Y)$, but if we notice the bell (Z), then learning the outcome of one coin should change our opinion about the other coin, i.e., $\neg I(X, Z, Y)$. How can we graphically represent the simple dependencies between the coins and the bell, or between any two causes leading to a common consequence? If we take the naive approach and assign links to (Z, X) and (Z, Y), leaving X and Y unlinked, we get the graph X—Z—Y. This graph is not an I-map because it (wrongly) asserts that X and Y are independent given Z. If we add a link between X and Y we get the trivial I-map of a complete graph, which no longer reflects the obvious fact that the two coins are genuinely independent (the bell being a passive device that does not affect their interaction). In Section 3.3, we will show that such dependencies can be represented completely with the richer language of directed graphs. For now, let us further examine the representational capabilities of undirected graphs.

Our inability to provide graphical representations for some models of dependency (e.g., induced dependency) raises the need to delineate the class of models that *do* lend themselves to graphical representation. This we do in the following section by establishing an axiomatic characterization of the family of relations that are isomorphic to vertex separation in graphs.

3.1.4 Axiomatic Characterization of Graph-Isomorph Dependencies

DEFINITION: *A dependency model M is said to be a* ***graph-isomorph*** *if there exists an undirected graph $G = (\boldsymbol{U}, \boldsymbol{E})$ that is a perfect map of M, i.e., for every three disjoint subsets $\boldsymbol{X}, \boldsymbol{Y}$, and $\boldsymbol{Z}$ of $\boldsymbol{U}$, we have*

$$I(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{Y})_M \iff < \boldsymbol{X} \mid \boldsymbol{Z} \mid \boldsymbol{Y} >_G. \tag{3.9}$$

THEOREM 2 [Pearl and Paz 1985]: *A necessary and sufficient condition for a dependency model M to be a graph-isomorph is that* $I(X, Z, Y)_M$ *satisfies the following five independent axioms (the subscript M is dropped for clarity):*

- *Symmetry:*
$$I(X, Z, Y) \Longleftrightarrow I(Y, Z, X) \tag{3.10a}$$

- *Decomposition:*
$$I(X, Z, Y \cup W) \Longrightarrow I(X, Z, Y) \;\&\; I(X, Z, W) \tag{3.10b}$$

- *Intersection:*
$$I(X, Z \cup W, Y) \;\&\; I(X, Z \cup Y, W) \Longrightarrow I(X, Z, Y \cup W) \tag{3.10c}$$

- *Strong union:*
$$I(X, Z, Y) \Longrightarrow I(X, Z \cup W, Y) \tag{3.10d}$$

- *Transitivity:*
$$I(X, Z, Y) \Longrightarrow I(X, Z, \gamma) \text{ or } I(\gamma, Z, Y). \tag{3.10e}$$

REMARKS:

1. γ is a singleton element of U, and all three arguments of $I(\cdot)$ must represent disjoint subsets.

2. The axioms are clearly satisfied for vertex separation in graphs. Eq. (3.10*e*) is the contrapositive form of connectedness transitivity, stating that if X is connected to some vertex γ and γ is connected to Y, then X must also be connected to Y. Eq. (3.10*d*) states that if Z is a vertex cutset separating X from Y, then removing additional vertices W from the graph leaves X and Y still separated. Eq. (3.10*c*) states that if X is separated from W with Y removed and X is separated from Y with W removed, then X must be separated from both Y and W.

3. Eqs. (3.10*c*) and (3.10*d*) imply the converse of Eq. (3.10*b*), meaning I is completely defined by the set of triplets (α, Z, β) in which α and β are individual elements of U:
$$I(X, Z, Y) \Longleftrightarrow (\forall\, \alpha \in X)\,(\forall\, \beta \in Y)\, I(\alpha, Z, \beta).$$
Equivalently, we can express the axioms in Eq. (3.10) in terms of such triplets. Note that the union axiom, Eq. (3.10*d*), is unconditional and therefore stronger than Eq. (3.6*c*), which is required for probabilistic dependencies. Eq. (3.10*d*) provides a simple way to construct a unique graph G_0 that is an I-map of M: starting with a complete graph, we delete every edge (α, β) for which $I(\alpha, Z, \beta)$ holds.

Proof:

1. The "necessary" part follows from the observation that all five axioms are satisfied by vertex separation in graphs. The logical independence of the five axioms can be demonstrated by letting $\boldsymbol{U}$ contain four elements and showing that it is always possible to contrive a subset I of triplets that violates one axiom and satisfies the other four.

2. To prove sufficiency, we must show that for any set I of triplets $(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{Y})$ closed under Eqs. (3.10*a*) through (3.10*e*), there exists a graph G such that $(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{Y})$ is in I iff $\boldsymbol{Z}$ is a cutset in G that separates $\boldsymbol{X}$ from $\boldsymbol{Y}$. We show that $G_0 = (\boldsymbol{U}, \boldsymbol{E}_0)$ is such a graph, where $(\alpha, \beta) \notin \boldsymbol{E}_0$ iff $I(\alpha, \boldsymbol{Z}, \beta)$. In view of Remark 3 above, it is sufficient to show that

$$I(\alpha, \boldsymbol{S}, \beta) \implies <\alpha|\boldsymbol{S}|\beta>_{G_0} \text{ where } \alpha, \beta \in \boldsymbol{U} \text{ and } \boldsymbol{S} \subseteq \boldsymbol{U},$$

since the converse follows automatically from the construction of G_0.

This is proved by finite descending induction:

i. For $|\boldsymbol{S}| = n-2$, the theorem holds automatically, because of the way G_0 is constructed.

ii. Assume the theorem holds for all $\boldsymbol{S}$ of size $|\boldsymbol{S}| = k \leq n-2$. Let $\boldsymbol{S}'$ be any set of size $|\boldsymbol{S}'| = k-1$. For $k \leq n-2$, there exists an element γ outside $\boldsymbol{S}' \cup \alpha \cup \beta$, and using Eq. (3.10*d*), we have $I(\alpha, \boldsymbol{S}', \beta) \implies I(\alpha, \boldsymbol{S}' \cup \gamma, \beta)$.

iii. By Eq. (3.10*e*) we have either $I(\alpha, \boldsymbol{S}', \gamma)$ or $I(\gamma, \boldsymbol{S}', \beta)$.

iv. Applying Eq. (3.10*d*) to either alternative in (iii) gives $I(\alpha, \boldsymbol{S}' \cup \beta, \gamma)$.

v. The middle arguments $\boldsymbol{S}' \cup \gamma$ and $\boldsymbol{S}' \cup \beta$ in (ii) and (iv) are both of size k, so by the induction hypothesis we have $<\alpha|\boldsymbol{S}' \cup \gamma|\beta>_{G_0}$ and $<\alpha|\boldsymbol{S}' \cup \beta|\gamma>_{G_0}$.

vi. By Eq. (3.10*c*), the intersection property for vertex separation in graphs, (iv) and (v) imply $<\alpha|\boldsymbol{S}'|\beta>_{G_0}$. Q.E.D.

Having a complete characterization for vertex separation in graphs allows us to test whether a given model of dependency lends itself to graphical representation. In fact, it is now easy to show that probabilistic models may violate both of the last two axioms. Eq. (3.10*d*) is clearly violated in the coins and bell example of the preceding subsection. Transitivity (Eq. (3.10*e*)) is violated in the same example, for if one of the coins is not fair, the bell's response is dependent on the outcome of each coin separately; yet the two coins are independent of each other. Finally, Eq. (3.10*c*) is violated whenever $\boldsymbol{Y}$ and $\boldsymbol{W}$ logically constrain one another, as in the earlier water temperature example.

Having failed to provide isomorphic graphical representations for even the most elementary models of informational dependency, we settle for the following

compromise: instead of complete graph isomorphism, we will consider only *I*-maps, i.e., graphs that faithfully display every dependency. However, acknowledging that some independencies will escape representation, we shall insist that their number be kept at a minimum—in other words, that the graphs contain no superfluous edges.

3.2 MARKOV NETWORKS

When a connection is drawn between such seemingly unrelated objects as probability distributions and graphs, it is natural to raise the following three questions:

1. Given a probability distribution P, can we construct an I-map G of P that has the minimum number of edges?
2. Given a pair (P, G), can we test whether G is an I-map of P?
3. Given a graph G, can we construct a probability distribution P such that G is a perfect map of P?

The theory of Markov fields provides satisfactory answers to Question 2 for strictly positive P [Isham 1981; Lauritzen 1982]. This treatment is rather complex and relies heavily on the numerical representation of probabilities. We shall start with Question 1 and show the following:

- Question 1 has a simple unique solution for strictly positive distributions.
- The solution to Question 2 follows directly from the solution to Question 1.
- The solutions are obtained by nonnumerical analysis, based solely on Eqs. (3.6*a*) through (3.6*e*) in Section 3.1.2.

Question 3 recently was answered affirmatively [Geiger and Pearl 1988a] and will be treated briefly in Section 3.2.3. Sections 3.2.3 and 3.2.4 focus on finding a probabilistic interpretation for a graph G such that the dependencies shown in G reflect empirical knowledge about a given domain.

3.2.1 Definitions and Formal Properties

DEFINITION: *A graph G is a **minimal** I-map of a dependency model M if deleting any edge of G would make G cease to be an I-map. We call such a graph a **Markov network** of M.*

THEOREM 3 [Pearl and Paz 1985]: *Every dependency model M satisfying symmetry, decomposition, and intersection (Eq. (3.6)) has a unique minimal I-map* $G_0 = (\boldsymbol{U}, \boldsymbol{E}_0)$ *produced by deleting from the complete graph every edge* (α, β) *for which* $I(\alpha, \boldsymbol{U} - \alpha - \beta, \beta)_M$ *holds, i.e.,*

$$(\alpha, \beta) \notin \boldsymbol{E}_0 \quad \textit{iff} \quad I(\alpha, \boldsymbol{U} - \alpha - \beta, \beta)_M. \tag{3.11}$$

The proof is given in Appendix 3-A.

DEFINITION: *A* ***Markov blanket*** $\boldsymbol{BL}_I(\alpha)$ *of an element* $\alpha \in \boldsymbol{U}$ *is any subset* $\boldsymbol{S}$ *of elements for which*

$$I(\alpha, \boldsymbol{S}, \boldsymbol{U} - \boldsymbol{S} - \alpha) \quad \textit{and} \quad \alpha \notin \boldsymbol{S}. \tag{3.12}$$

A set is called a ***Markov boundary*** *of* α, *denoted* $\boldsymbol{B}_I(\alpha)$, *if it is a minimal Markov blanket of* α, *i.e., none of its proper subsets satisfy Eq. (3.12).*

The boundary $\boldsymbol{B}_I(\alpha)$ is to be interpreted as the smallest set of elements that shields α from the influence of all other elements. Note that $\boldsymbol{B}_I(\alpha)$ always exists because $I(\boldsymbol{X}, \boldsymbol{S}, \varnothing)$ guarantees that the set $\boldsymbol{S} = \boldsymbol{U} - \alpha$ satisfies Eq. (3.12).

THEOREM 4 [Pearl and Paz 1985]: *Every element* $\alpha \in \boldsymbol{U}$ *in a dependency model satisfying symmetry, decomposition, intersection, and weak union (Eq. (3.6)) has a unique Markov boundary* $\boldsymbol{B}_I(\alpha)$. *Moreover,* $\boldsymbol{B}_I(\alpha)$ *coincides with the set of vertices* $\boldsymbol{B}_{G_0}(\alpha)$ *adjacent to* α *in the minimal I-map* G_0.

The proof of Theorem 4 is given in Appendix 3-B. Since $\boldsymbol{B}_I(\alpha)$ coincides with $\boldsymbol{B}_{G_0}(\alpha)$, the following two interpretations of *direct neighbors* are identical: neighborhood as a blanket that shields α from the influence of all other variables, and neighborhood as a permanent bond of mutual influence between two variables, a bond that cannot be weakened by other elements in the system. Models satisfying the conditions of Theorem 4 are called *pseudo-graphoids*, i.e., graphoids lacking the contraction property (Eq. (3.6*d*)).

Since every strictly positive distribution defines a pseudo-graphoid, we can derive two corollaries.

COROLLARY 1: *The set of Markov boundaries* $\boldsymbol{B}_I(\alpha)$ *induced by a strictly positive probability distribution forms a* ***neighbor system****, i.e., a collection* $\boldsymbol{B}_I^* = \{\boldsymbol{B}_I(\alpha) : \alpha \in \boldsymbol{U}\}$ *of subsets of* $\boldsymbol{U}$ *such that for all pairs* $\alpha, \beta \in \boldsymbol{U}$ *we have*

(*i*) $\alpha \notin \boldsymbol{B}_I(\alpha)$ *and*

(*ii*) $\alpha \in \boldsymbol{B}_I(\beta)$ *iff* $\beta \in \boldsymbol{B}_I(\alpha)$.

COROLLARY 2: *The Markov network G_0 of any strictly positive distribution can be constructed by connecting each variable α to all members of its Markov boundary $\boldsymbol{B}_I(\alpha)$.*

Corollary 2 is useful because often it is the Markov boundaries $\boldsymbol{B}_I(\alpha)$ that are given to us when we request the factors that affect α most directly. These factors may be the immediate consequences of an event, the justifications for an action, or the salient properties that characterize a class of objects or a concept. Moreover, since either construction will yield an I-map, many global independence relationships can be validated by separation tests on graphs constructed from local information.

TESTING I-MAPNESS

We are now in a position to answer Question 2 from the beginning of this subsection: can we test whether a given graph G is an I-map of a distribution P (i.e., test the *I-mapness* of G)? We assume that P is not given explicitly but is represented by a procedure that answers queries of the type "Is $I(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{Y})$ true in P?"

THEOREM 5: *Given a strictly positive probability distribution P on $\boldsymbol{U}$ and a graph $G = (\boldsymbol{U}, \boldsymbol{E})$, the following three conditions are equivalent:*

i. *G is an I-map of P.*

ii. *G is a supergraph of the Markov network G_0 of P, i.e.,*

$$(\alpha, \beta) \in \boldsymbol{E} \quad \textit{whenever} \quad \neg I(\alpha, \boldsymbol{U} - \alpha - \beta, \beta).$$

iii. *G is locally Markov with respect to P, i.e., for every variable $\alpha \in \boldsymbol{U}$ we have $I(\alpha, \boldsymbol{B}_G(\alpha), \boldsymbol{U} - \alpha - \boldsymbol{B}_G(\alpha))$, where $\boldsymbol{B}_G(\alpha)$ is the set of vertices adjacent to α in G.*

Proof: The implication (ii) => (i) follows from the I-mapness of G_0 (Theorem 3), and (i) => (iii) follows from the definition of I-mapness. It remains to show (iii) => (ii), but this follows from the identity of $\boldsymbol{B}_I(\alpha)$ and $\boldsymbol{B}_{G_0}(\alpha)$ (Theorem 4). Q.E.D.

Properties (ii) and (iii) provide local procedures for testing I-mapness without examining every cutset in G. To show the essential role played by the assumption of strict positivity let us demonstrate the insufficiency of local tests when variables are subjected to functional constraints. Imagine four random variables constrained by equality, i.e., $X = Y = Z = W$. Any single variable is a Markov boundary of any other, because knowing the first variable determines the value of the second. Consequently, the graph shown in Figure 3.3*a* would qualify under the Markov

boundary condition (Property iii of Theorem 5). This graph is not an I-map of the distribution, however, because the pair (X, Y) is not independent of the pair (Z, W). Worse yet, since any pair of variables is rendered independent given the values of the other pair, $I(\alpha, \boldsymbol{U} - \alpha - \beta, \beta)$ holds for every pair (α, β). Thus, were we to construct G_0 by the edge-deletion method of Eq. (3.11), we would get an empty graph (Figure 3.3*b*), which obviously is not an I-map of the distribution.

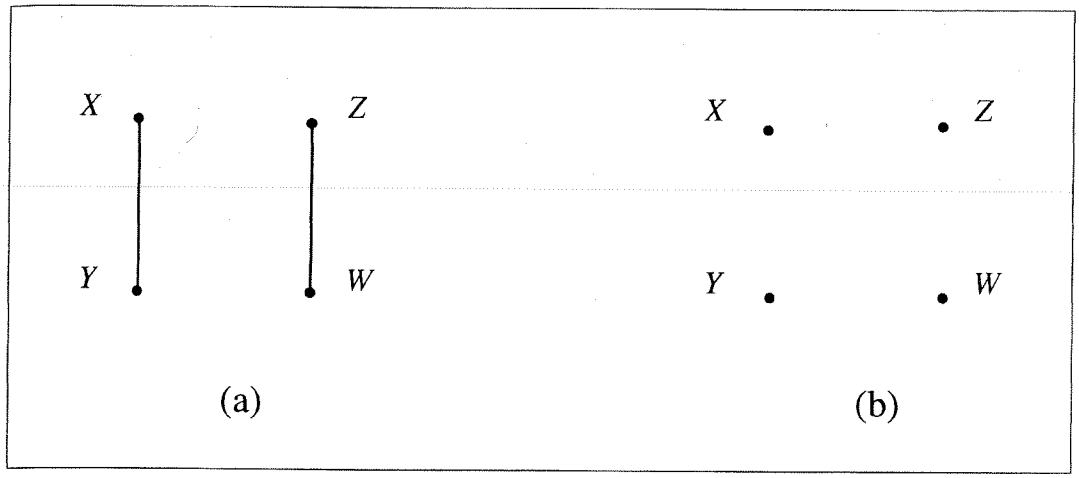


Figure 3.3. *Failure of local tests for I-mapness under equality constraints $X = Y = Z = W$. (a) A graph qualifying under the Markov boundary test. (b) An empty graph qualifying under the edge-deletion test (Eq. (3.11)).*

It can be shown that even if we connect each variable to the union of all its Markov boundaries, we will not get an I-map when categorical constraints are present. Thus, there appears to be no local test for I-mapness of undirected graphs that works for extreme probability distributions. We shall see in Section 3.3 that directed graphs do not suffer from this deficiency; local tests for I-mapness and minimal I-mapness exist even for distributions that reflect categorical constraints. It should be noted that the tests in (ii) and (iii), while local, still involve all the varibles in $\boldsymbol{U}$ and therefore may require exponentially complex procedures, especially when P is given as a table. Fortunately, in most practical applications we start with the graph representation G and use the probability model P merely as a theoretical abstraction to justify the operations conducted on G.

We see that representations of probabilistic independencies using undirected graphs rest heavily on the intersection and weak union axioms, Eqs. (3.6*e*) and (3.6*c*). In contrast, we shall see in Section 3.3 that directed graph representations rely on the contraction and weak union axioms, with intersection playing only a minor role.

3.2.2 Illustrations

GRAPHOIDS AND THEIR MARKOV NETWORKS

To see the roles of the various axioms of Eq. (3.6), consider a set of four integers $U = \{1, 2, 3, 4\}$, and let I be the set of twelve triplets listed below:

$$I = \{(1, 2, 3), (1, 3, 4), (2, 3, 4), (\{1, 2\}, 3, 4),$$
$$(1, \{2, 3\}, 4), (2, \{1, 3\}, 4)\ , \textit{symmetrical images}\}.$$

All other triplets are assumed to be dependent, i.e., outside I. It is easy to see that I satisfies the other axioms of Eq. (3.6) but does not satisfy contraction; I contains $(1, 2, 3)$ and $(1, \{2, 3\}, 4)$ but not $(1, 2, \{3, 4\})$. Thus, (from Theorem 1) I is supported by no probability model, but (from Theorem 3) it has a unique minimal I-map G_0, shown in Figure 3.4. Moreover, Theorem 4 ensures that G_0 can be constructed in two different ways, either by deleting the edges (1, 4) and (2, 4) from the complete graph, in accordance with Eq. (3.11), or by computing from I the Markov boundary of each element, in accordance with Eq. (3.12), yielding

$$\boldsymbol{B}_I(1) = \{2, 3\},\ \boldsymbol{B}_I(2) = \{1, 3\},\ \boldsymbol{B}_I(3) = \{1, 2, 4\},\ \boldsymbol{B}_I(4) = \{3\}.$$

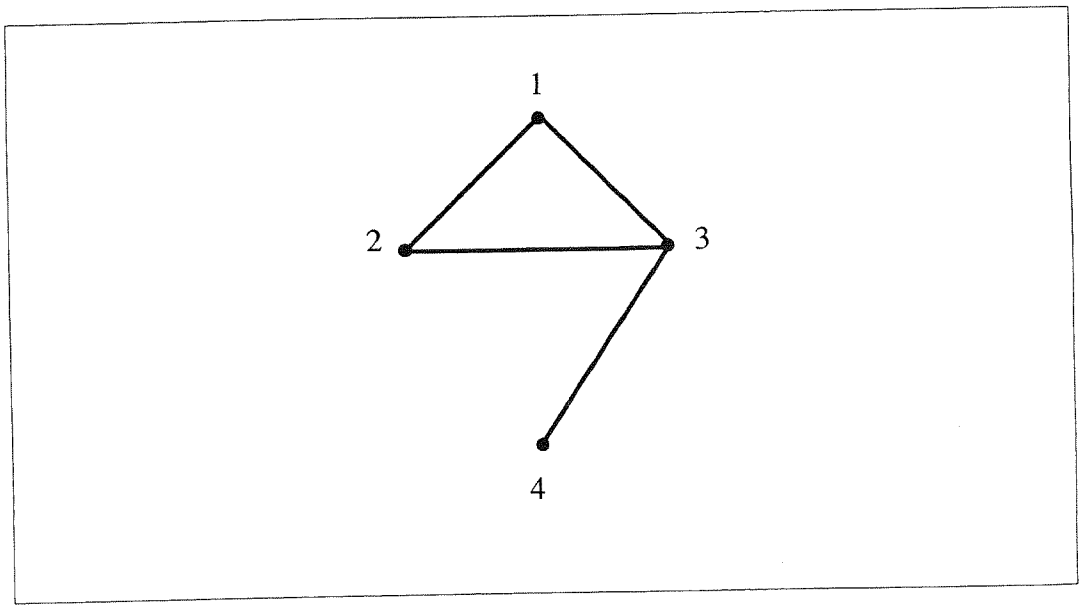


Figure 3.4. *The minimal I-map, G_0, of I.*

Now consider a modified list I' containing only the last two triplets of I (and their symmetrical images):

$$I' = \{(1, \{2, 3\}, 4), (2, \{1, 3\}, 4), \textit{symmetrical images}\}.$$

I' is a semi-graphoid (it satisfies Eqs. (3.6*a*) through (3.6*d*)) but not a graphoid, because the absence of the triplet ({1,2}, 3,4) violates the intersection axiom (Eq. (3.6*e*)). Hence, I' can represent a probability model but not a strictly positive one. Indeed, if we try to construct G_0 by the usual criterion of edge-deletion (Eq. (3.11)), we get the graph in Figure 3.4, but it is no longer an I-map of I'; it shows 3 separating 1 from 4, but (1, 3, 4) is not in I'. In fact, the only I-maps of I' are the three graphs in Figure 3.5, and the minimal I-map clearly is not unique.

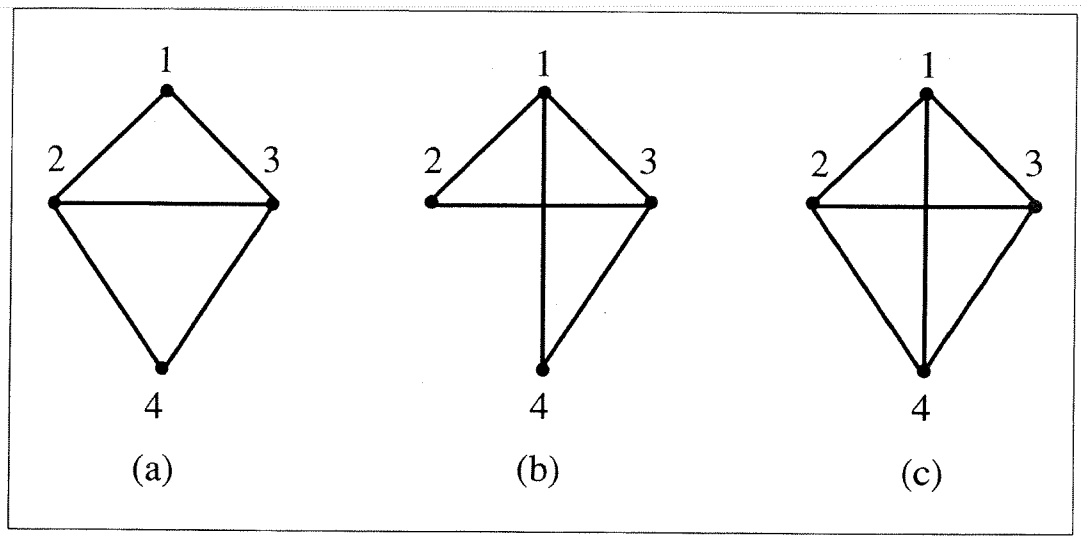


Figure 3.5. *The three I-maps of I′.*

Now consider the list

$$I'' = \{(1, 2, 3), (1, 3, 4), (2, 3, 4), (\{1, 2\}, 3, 4), \textit{symmetrical images}\}.$$

I'' satisfies Eqs. (3.6*a*), (3.6*b*), and (3.6*e*), but not the weak union axiom (Eq. (3.6*c*)). From Theorem 3 we can still construct a unique I-map for I'' using the edge-deletion method, but because no triplet of the form $(\alpha, \boldsymbol{U} - \alpha - \beta, \beta)$ appears in I'', the only I-map for this list is the complete graph. Moreover, the Markov boundaries of I'' do not form a neighbor set ($\boldsymbol{B}_{I''}(4) = 3$, $\boldsymbol{B}_{I''}(2) = \{1, 3, 4\}$, so $2 \notin \boldsymbol{B}_{I''}(4)$ while $4 \in \boldsymbol{B}_{I''}(2)$). Thus, we see that the lack of weak union prevents us from constructing an I-map by the Markov boundary method.

Since I does not obey the contraction property (Eq. (3.6*d*)), no probabilistic model can induce this set of independence relationships unless we add the triplet (1, 2, 4) to I. If I were a list of statements given by a domain expert, it would be

possible to invoke Eq. (3.6*a*) through (3.6*e*) to alert the expert to the inconsistency caused by the absence of (1, 2, 4). The incompleteness of I' and I'' would be easier to detect by graphical means because they interfere with the formation of G_0 and could be identified by a system attempting to construct it.

CONCEPTUAL DEPENDENCIES AND THEIR MARKOV NETWORKS

Consider the task of constructing a Markov network to represent the belief about whether agent A will be late for a meeting. Assume the agent identifies the following variables as having influence on the main question of being late to a meeting:

1. The time shown on the watch of Passerby 1.
2. The time shown on the watch of Passerby 2.
3. The correct time.
4. The time it takes to travel to the meeting place.
5. The arrival time at the meeting place.

The construction of G_0 can proceed by one of two methods:

- The *edge-deletion* method.
- The *Markov boundary* method.

Following Eq. (3.11), the first method requires that for every pair of variables (α, β) we determine whether fixing the values of all other variables in the system will render our belief in α sensitive to β. We know, for example, that the reading on Passerby 1's watch (1) will vary with the actual time (3) even if all other variables are known. On that basis, we can connect node 1 to node 3 and, by proceeding this way through all pairs of variables, construct the graph of Figure 3.6. The unusual edge (3, 4) reflects the reasoning that if we fix the arrival time (5), the travel time (4) must depend on the current time (3).

The Markov boundary method requires that for every variable α in the system, we identify a minimal set of variables sufficient to render the belief in α insensitive to all other variables in the system. It is a commonsense task, for instance, to decide that once we know the current time (3), no other variable can affect what we expect to read on passerby 1's watch (1). Similarly, to estimate our arrival time (5), we need only know the current time (3) and how long it takes to travel (4), independent of the watch readings (1) and (2). On the basis of these considerations, we can connect 1 to 3, 5 to 4 and 3, and so on. After we find

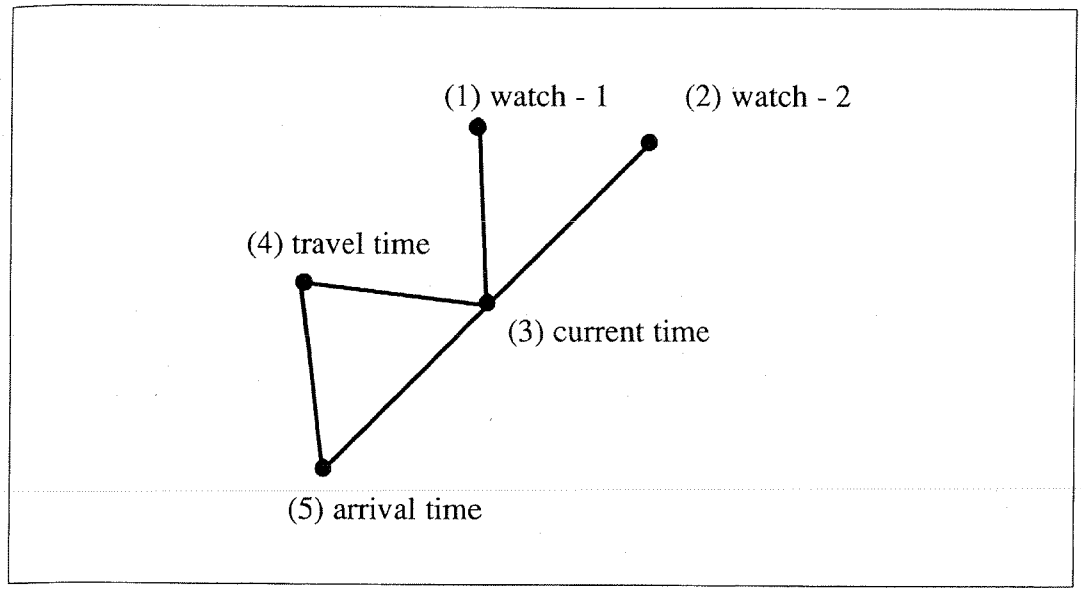


Figure 3.6. *The Markov network representing the prediction of A's arrival time.*

the immediate neighbors of any four variables in the system, the graph G_0 will emerge, identical to that of Figure 3.6.

Once established, G_0 can be used as an inference instrument. For example, we need not state explicitly that knowing the current time (3) renders the time on Passerby 1's watch (1) irrelevant for estimating the travel time (4) (i.e., $I(1,3,4)$); we can infer the information from the fact that 3 is a cutset in G_0, separating 1 from 4. Deriving such conclusions by syntactic manipulation of Eqs. (3.6*a*) through (3.6*e*) probably would be more complicated. Additionally, the graphical representation can help maintain consistency and completeness during the knowledge-building phase. One need ascertain only that the relevance boundaries identified by the expert form a neighbor system.

SUMMARY

The essential qualities of conditional independence are captured by five logical axioms: symmetry (Eq. (3.6*a*)), decomposition (Eq. (3.6*b*)), weak union (Eq. (3.6*c*)), contraction (Eq. (3.6*d*)), and intersection (Eq. (3.6*e*)). Intersection holds only for strictly positive distributions (i.e., reflecting no functional or definitional constraints) and is essential to the construction of undirected graphs. Symmetry, decomposition, and intersection enable us to construct a minimal graph G_0 (Markov network), in which every cutset corresponds to a genuine independence condition. The weak union axiom is needed to guarantee that the set of neighbors that G_0 assigns to each variable α is the smallest set required to shield α from the effects of all other variables.

The Markov network representation of conditional independence offers a sound inference mechanism for deducing, at any state of knowledge, which propositional variables are relevant to each other. If we identify the Markov boundaries associated with each proposition in the system and treat them as neighborhood relations defining a graph G_0, then we can correctly identify independence relationships by testing whether the set of known propositions constitutes a cutset in G_0.

Not all probabilistic dependencies can be captured by undirected graphs. For example, a dependency may be induced and non-transitive (see the coins and bell example of Section 3.1.3), but graph separation is strictly normal and transitive. For this reason directed graphs are finding wider application in reasoning systems [Duda, Hart, and Nilsson 1976; Howard and Matheson 1981; Pearl 1986c]. A systematic treatment of directed graph representations is given in Section 3.3.

3.2.3 Markov Network as a Knowledge Base

QUANTIFYING THE LINKS

So far, we have established the semantics of Markov networks in terms of the purely qualitative notion of conditional independence, i.e., a variable is proclaimed independent of all its non-neighbors once we know the values of its neighbors. However, if the network is to convey information useful for decisions and inference, we must also provide quantitative assessments of the strength of each link. In Figure 3.2, for example, if we know that the couple (M_1, F_2) meet less frequently than the couple (M_1, F_1), then the first link should be weaker than the second to show weaker dependency between the propositions m_1 and f_2.

The assigning of weights to the links of the graph must be handled with caution. If the weights are to be used in translating evidential data into meaningful probabilistic inferences, we must be certain that the model is both consistent and complete. Consistency guarantees that we do not overload the graph with too many parameters—overspecification can lead to contradictory conclusions, depending on which parameter is consulted first—and completeness protects us from underspecifying the model and thus guarantees that routines designed to generate conclusions will not get deadlocked for lack of information.

An attractive feature of the traditional joint-distribution representation of probabilities is the ease with which one can synthesize consistent probability models or detect inconsistencies in models. In this representation, to create a complete and consistent model, one need only assign to the elementary events (i.e., conjunctions of atomic propositions) nonnegative weights summing to one. The synthesis process in the graph representation is more hazardous. For example, assume that in Figure 3.2 we want to express the dependencies between the variables $\{M_1, M_2, F_1, F_2\}$ by specifying the four pairwise probabilities

$P(M_1, F_1)$, $P(F_1, M_2)$, $P(M_2, F_2)$, and $P(F_2, M_1)$. Unless the parameters given satisfy some nonobvious relationship, no probability model will support all four inputs, and we will get inconsistencies. Moreover, it is not clear that we can put all numerical inputs together without violating the qualitative dependency relationships shown in the graph. On the other hand, if we specify the pairwise probabilities of only three pairs, incompleteness will result; many models will conform to the input specification, and we will be unable to provide answers to many useful queries.

The theory of Markov fields [Isham 1981, Lauritzen 1982] provides a safe method (called *Gibbs' potential*) for constructing a complete and consistent quantitative model while preserving the dependency structure of an arbitrary graph G. The method consists of four steps:

1. Identify the cliques† of G, namely, the maximal subgraphs whose nodes are all adjacent to each other.

2. For each clique $\boldsymbol{C}_i$, assign a nonnegative compatibility function $g_i(\boldsymbol{c}_i)$, which measures the relative degree of compatibility associated with each value assignment $\boldsymbol{c}_i$ to the variables included in $\boldsymbol{C}_i$.

3. Form the product $\prod_i g_i(\boldsymbol{c}_i)$ of the compatibility functions over all the cliques.

4. Normalize the product over all possible value combinations of the variables in the system

$$P(x_1, \ldots, x_n) = K \prod_i g_i(\boldsymbol{c}_i), \tag{3.13}$$

where

$$K = [\sum_{x_1, \ldots, x_n} \prod_i g_i(\boldsymbol{c}_i)]^{-1}.$$

The normalized product P in Eq. (3.13) constitutes a joint distribution that embodies all the conditional independencies portrayed by the graph G, i.e., G is an I-map of P (see Theorem 6, below).

To illustrate the mechanics of this method, let us return to the example of Figure 3.2 and assume that the likelihood of two members of the i-th couple having the same state of disease is measured by a compatibility parameter α_i, and the likelihood that exactly one partner of the couple will carry the disease is assigned a

† We use the term *clique* for the more common term *maximal clique*.

compatibility parameter β_i. The dependency graph in this case has four cliques, corresponding to the four edges

$$\boldsymbol{C}_1 = \{M_1, F_1\}, \boldsymbol{C}_2 = \{M_1, F_2\},$$
$$\boldsymbol{C}_3 = \{M_2, F_1\}, \text{ and } \boldsymbol{C}_4 = \{M_2, F_2\},$$

and the compatibility functions g_i are given by

$$g_i(x_{i_1}, x_{i_2}) = \begin{cases} \alpha_i & \text{if } x_{i_1} = x_{i_2} \\ \beta_i & \text{if } x_{i_1} \neq x_{i_2}, \end{cases} \tag{3.14}$$

where x_{i_1} and x_{i_2} are the states of disease associated with the male and female, respectively, of couple $\boldsymbol{C}_i$. The overall probability distribution function is given by the normalized product

$$P(M_1, M_2, F_1, F_2) = K\, g_1(M_1, F_1)\, g_2(M_1, F_2)\, g_3(M_2, F_1)\, g_4(M_2, F_2)$$
$$= K \prod_i \beta_i^{|x_{i_1} - x_{i_2}|} \alpha_i^{1 - |x_{i_1} - x_{i_2}|}, \tag{3.15}$$

where K is a constant that makes P sum to unity over all states of the system, i.e.,

$$K^{-1} = \prod_i (\alpha_i + \beta_i) + \prod_i \alpha_i \sum_j \frac{\beta_j}{\alpha_j} + \prod_i \beta_i \sum_j \frac{\alpha_j}{\beta_j}. \tag{3.16}$$

For example, the state in which only the males carry the disease, $(m_1, \neg f_1, m_2, \neg f_2)$, will have a probability measure $K\beta_1\beta_2\beta_3\beta_4$ because the male and female of each couple are in unequal states of disease. The state $(m_1, f_1, \neg m_2, \neg f_2)$, on the other hand, has the probability $K\alpha_1\beta_2\beta_3\alpha_4$ because couples $\boldsymbol{C}_1$ and $\boldsymbol{C}_4$ are both homogeneous.

To show that P is consistent with the dependency structure of G, we note that any product of the form of Eq. (3.15) can be expressed either as the product $f(M_1, F_1, F_2)\ g(F_1, F_2, M_2)$ or as $f'(F_1, M_1, M_2)\ g'(M_1, M_2, F_2)$. Thus, invoking Eq. (3.5*b*), we conclude that $I(M_1, F_1 \cup F_2, M_2)_P$ and $I(F_1, M_1 \cup M_2, F_2)_P$.

The next theorem ensures the generality of this construction method.

THEOREM 6 [Hammersley and Clifford 1971]: *A probability function P formed by a normalized product of positive functions on the cliques of G is a Markov field relative to G, i.e., G is an I-map of P.*

Proof: G is guaranteed to be an I-map if P is locally Markov relative to G (Theorem 5). It is sufficient, therefore, to show that the neighbors in G of each variable α constitute a Markov blanket of α relative to P, i.e., that $I(\alpha, \boldsymbol{B}_G(\alpha), \boldsymbol{U} - \alpha - \boldsymbol{B}_G(\alpha))$ or (using Eq. (3.5*b*)) that

$$P(\alpha, \boldsymbol{B}_G(\alpha), \boldsymbol{U} - \alpha - \boldsymbol{B}_G(\alpha)) = f_1(\alpha, \boldsymbol{B}_G(\alpha))\, f_2(U - \alpha). \tag{3.17}$$

Let J_α stand for the set of indices marking all cliques in G that include α, $J_\alpha = \{j : \alpha \in \boldsymbol{C}_j\}$. Since P is in product form, we can write

$$P(\alpha, \beta, \ldots) = K \prod_j g_j(\boldsymbol{c}_j) = K \prod_{j \in J_\alpha} g_j(\boldsymbol{c}_j) \prod_{j \notin J_\alpha} g_j(\boldsymbol{c}_j). \tag{3.18}$$

The first product in Eq. (3.18) contains only variables that are adjacent to α in G; otherwise, $\boldsymbol{C}_j$ would not be a clique. According to the definition of J_α, the second product does not involve α. Thus, Eq. (3.17) is established. Q.E.D.

The converse of Theorem 6 also holds: any positive Markov field can be expressed in product form as in Eq. (3.13). The theorem, though not its converse (see Exercise 3.3), also holds for extreme probabilities. Theorem 6 still does not guarantee that every conditional *dependency* shown in the graph will be embodied in P if P is constructed by the product form of Eq. (3.13), but a more recent result gives us this guarantee, i.e., every undirected graph G has a distribution P such that G is a perfect map of P [Geiger and Pearl 1988a]. Thus, we can answer yes to Question 3 of the introduction to this section.

INTERPRETING THE LINK PARAMETERS

The preceding method of modeling guarantees consistency and completeness, but it leaves much to be desired. In particular, it is difficult to assign meanings to the parameters of the compatibility functions. If a model's parameters are to lead to meaningful inferences or decisions, they must come either from direct measurements or from an expert who can relate them to actual human experience. Both options encounter difficulties in the Markov network formulation.

Let us assume we have a huge record of medical tests conducted on homogeneous subjects, and the record includes a full account of their sexual habits. Can we extract from it the desired compatibility functions $g_i(M, F)$? The difficulty is that any disease pattern we observe on a given couple is a function not only of the relations between the male and female of this couple but also of interaction between this couple and the rest of the population. In other words, our measurements invariably are taken in a noisy environment; in our case, this means a large network of interactions surrounds the one that is tested.

To further appreciate the difficulties associated with context-dependent measurements, let us take an ideal case and assume that our record is based solely

on groups of four interacting individuals (as in Figure 3.2), with each group isolated from the rest of the world and all groups having the same sexual pattern. In other words, we are given the joint probability $P(M_1, F_1, F_2, M_2)$, or a close approximation to it, and we are asked to infer the compatibility functions g_i. Clearly it is not an easy task, even in this ideal case; using the data provided by P we must solve a set of simultaneous nonlinear equations for g_i, such as Eq. (3.13) or Eq. (3.15). In addition, the solution we obtain for g_i will not be applicable to new situations in which, say, the frequency of interaction is different. Thus, we see why the compatibility parameters cannot be given meaningful experiential interpretation.

For a parameter to be meaningful, it must be an abstraction of some invariant property of one's experience. In our example, the relation between frequency of contact and transference of the disease from one partner to another, under conditions of perfect isolation from the rest of the world, is meaningful. In probabilistic terminology, the quantities $P(f_1 \mid m_1, \neg m_2)$ and $P(f_1 \mid \neg m_1, \neg m_2)$ and their relations to the frequency of interaction of couple $\{M_1, F_1\}$ are perceived as invariant characteristics of the disease, generalizable across contexts. It is with these quantities, therefore, that an expert would choose to encode experiential knowledge, and it is these quantities that an expert is most willing to assess. Moreover, were we conducting a clean scientific experiment, these are the quantities we would choose to measure.

Unfortunately, the Markov network formulation does not allow the direct specification of such judgmental input. Judgments about low-order conditional probabilities (e.g., $P(m_1 \mid f_1, \neg m_2)$) can be taken only as constraints that the joint probability distribution (Eq. (3.13)) must satisfy; from them, we might be able to calculate the actual values of the compatibility parameters. But this is a rather tedious computation, especially if the number of variables is large (imagine a group of n interacting couples), and the computation must be performed at the knowledge-acquisition phase to ensure that the expert provides a consistent and complete set of constraints.

3.2.4 Decomposable Models

Some dependency models do not suffer from the quantification difficulty described in the preceding section; instead, the compatibility functions are directly related to the low-order marginal probabilities on the variables in each clique. Such *decomposable* models have the useful property that the cliques of their Markov networks form a tree.

MARKOV TREES

To understand why tree topologies have this desirable feature, let us consider a distribution P having a Markov network in the form of a chain

$$X_1—X_2—X_3—X_4.$$

From the chain rule of basic probability theory (Eq. (2.12)) we know that every distribution function $P(x_1,..., x_n)$ can be represented as a product:

$$P(x_1, ..., x_n) = P(x_1)\, P(x_2 \mid x_1) \ldots P(x_n \mid x_1, ..., x_{n-1}). \tag{3.19}$$

Thus, if we expand P in the order dictated by the chain, we can write

$$P(x_1, x_2, x_3, x_4) = P(x_1)\, P(x_2 \mid x_1)\, P(x_3 \mid x_1, x_2)\, P(x_4 \mid x_1, x_2, x_3),$$

and using the conditional independencies encoded in the chain, we obtain

$$P(x_1, x_2, x_3, x_4) = P(x_1)\, P(x_2 \mid x_1)\, P(x_3 \mid x_2)\, P(x_4 \mid x_3).$$

The joint probability P is expressible in terms of a product of three functions, each involving a pair of adjacent variables. Moreover, the functions are the very pairwise conditional probabilities that should carry conceptual meaning, according to our earlier discussion. This scheme leaves the choice of ordering quite flexible. For example, if we expand P in the order (X_3, X_2, X_4, X_1), we get

$$\begin{aligned} P(x_3, x_2, x_4, x_1) &= P(x_3)\, P(x_2 \mid x_3)\, P(x_4 \mid x_3, x_2)\, P(x_1 \mid x_3, x_2, x_4) \\ &= P(x_3)\, P(x_2 \mid x_3)\, P(x_4 \mid x_3)\, P(x_1 \mid x_2), \end{aligned}$$

again yielding a product of edge probabilities. The only requirement is this: as we order the variables from left to right, every variable except the first should have at least one of its graph neighbors to its left. The ordering (X_1, X_4, X_2, X_3), for example, would not yield the desired product form because X_4 is positioned to the left of its only neighbor, X_3.

Given a tree-structured Markov network, there are two ways to find its product-form distribution by inspection: *directed trees* and *product division*.

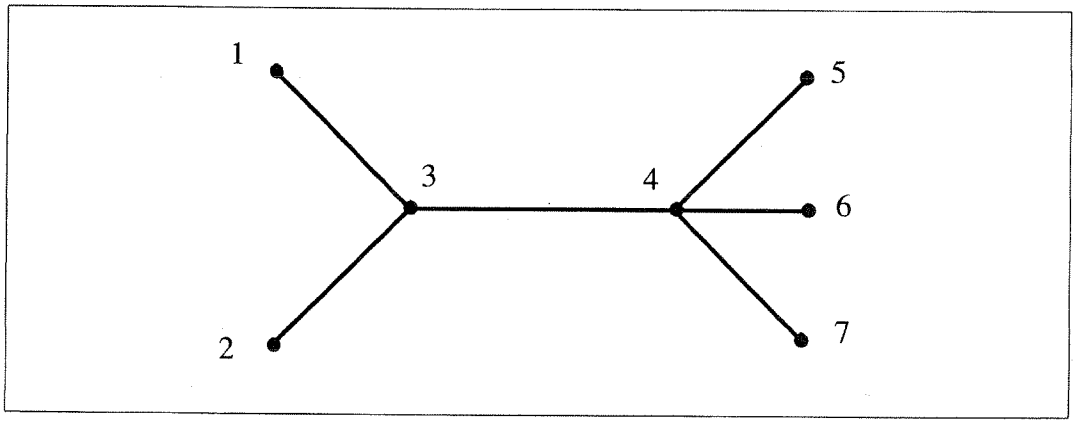


Figure 3.7. *An undirected tree of seven variables.*

Consider the tree of Figure 3.7, where the variables $X_1,...,X_7$ are marked 1,...,7 for short. If we arbitrarily choose node 3 as a root and assign to the links arrows pointing away from the root, we get the directed tree of Figure 3.8, where every non-root node has a single arrow coming from its unique parent. We can now write the product distribution by inspection, going from parents to children:

$$P(1,...,7) = P(3)\,P(1|3)\,P(2|3)\,P(4|3)\,P(5|4)\,P(6|4)\,P(7|4). \qquad \textbf{(3.20)}$$

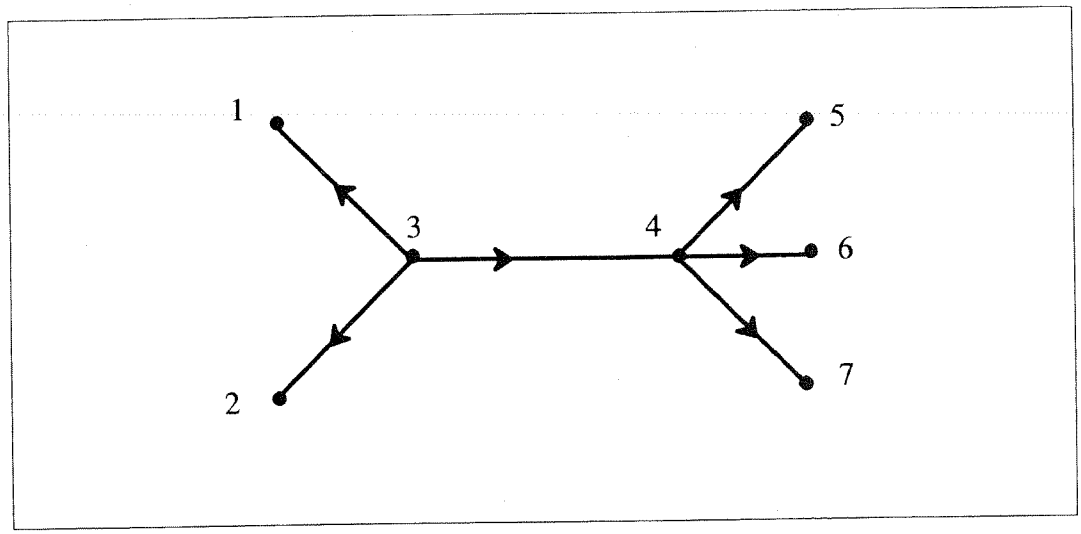


Figure 3.8. *A directed tree with root 3.*

The conditioning (right) variable in each term of the product is a direct parent of the conditioned (left) variable.

The second method for expressing the joint distribution is to divide the product of the marginal distributions on the edges (i.e., cliques) by the product of the distributions of the intermediate nodes (i.e., the intersections of the cliques). The distribution corresponding to the tree of Figure 3.8 will be written

$$P(1,...,7) = \frac{P(1,3)\,P(2,3)\,P(3,4)\,P(4,5)\,P(4,6)\,P(4,7)}{P(3)\quad P(3)\quad P(4)\quad P(4)\quad P(4)}, \qquad \textbf{(3.21)}$$

which is identical to Eq. (3.20). Each variable in the denominator appears one less time than it appears in the numerator.

JOIN TREES

Trees are not the only distributions amenable to product forms. Consider, for example, the structure of Figure 3.9*a*. Applying the chain rule in the order (A, B, C, D, E), and using the independencies embedded in the graph, we obtain

$$\begin{aligned} P(a, b, c, d, e) &= P(a)\,P(b \mid a)\,P(c \mid a, b)\,P(d \mid a, b, c)\,P(e \mid a, b, c, d) \\ &= P(a)\,P(b \mid a)\,P(c \mid a, b)\,P(d \mid b, c)\,P(e \mid c) \\ &= \frac{P(a, b, c)\,P(b, c, d)}{P(b, c)} \quad \frac{P(c, e)}{P(c)}. \end{aligned} \tag{3.22}$$

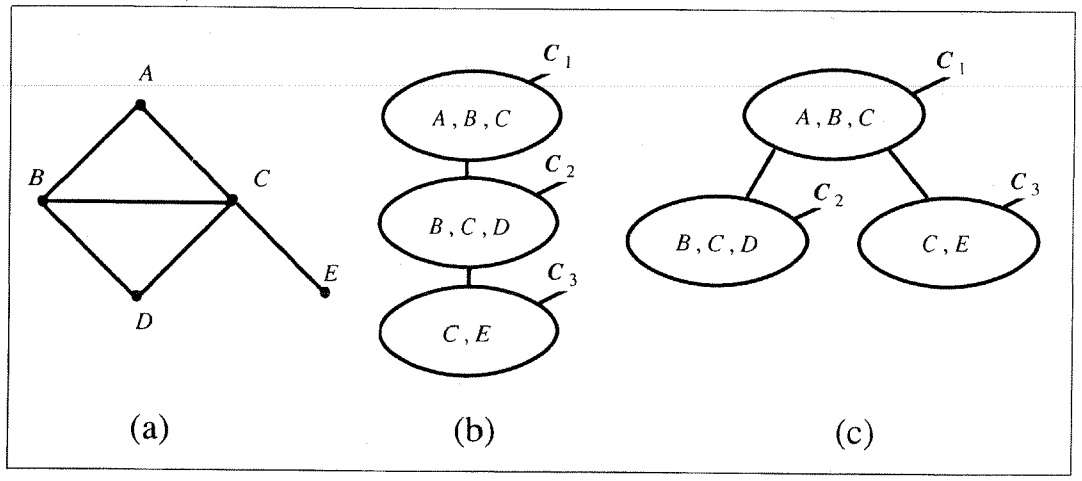


Figure 3.9. *Two join trees, (b) and (c), constructed from the cliques of the graph in (a).*

Eq. (3.22) again displays the same pattern as Eq. (3.21): the numerator is a product of the distributions of the cliques, and the denominator is a product of the distributions of their intersections. Note that C is a node common to all three cliques—$\{A, B, C\}$, $\{B, C, D\}$, and $\{C, E\}$— yet $P(c)$ appears only once in the denominator. The reason will become clear in the ensuing discussion, where we will justify the general formula for clique trees.

The unique feature of the graph in Figure 3.9*a* that enables us to obtain a product-form distribution is the fact that the cliques in this graph can be joined to form a tree, as seen in Figure 3.9*b* and Figure 3.9*c*. More precisely, there is a tree that is an I-map of P, with vertices corresponding to the cliques of G. Indeed, writing $\boldsymbol{C}_1 = \{A, B, C\}$, $\boldsymbol{C}_2 = \{B, C, D\}$, and $\boldsymbol{C}_3 = \{C, E\}$, we see that $\boldsymbol{C}_3$ and $\boldsymbol{C}_1$ are independent given $\boldsymbol{C}_2$, and we draw the I-map $\boldsymbol{C}_1$—$\boldsymbol{C}_2$—$\boldsymbol{C}_3$ of Figure 3.9*b*. Since $\boldsymbol{C}_3$ and $\boldsymbol{C}_2$ are independent given $\boldsymbol{C}_1$, we can also use the I-map $\boldsymbol{C}_2$—$\boldsymbol{C}_1$—$\boldsymbol{C}_3$ of Figure 3.9*c*. This nonuniqueness of the minimal I-maps, an apparent contradiction to Theorem 3, stems from the overlapping of $\boldsymbol{C}_1$, $\boldsymbol{C}_2$, and $\boldsymbol{C}_3$, which induces equality constraints and occasionally leads to violation of the intersection axiom (Eq. (3.6*e*)).

Now we shall present a theorem about *chordal graphs* [Beeri et. al. 1983] in order to further articulate the concept of a clique tree.

DEFINITION: *An undirected graph* $G = (\mathbf{V}, \mathbf{E})$ *is said to be* ***chordal*** *if every cycle of length four or more has at least one chord, i.e., an edge joining two nonconsecutive vertices along that cycle.*

THEOREM 7: *Let G be an undirected graph* $G = (V, E)$. *The following four conditions are equivalent:*

1. *G is chordal.*
2. *The edges of G can be directed acyclically so that every pair of converging arrows emanates from two adjacent vertices.*
3. *All vertices of G can be deleted by arranging them in separate piles, one for each clique, and then repeatedly applying the following two operations:*
   - *Delete a vertex that occurs in only one pile.*
   - *Delete a pile if all its vertices appear in another pile.*
4. *There is a tree T (called a join tree) with the cliques of G as vertices, such that for every vertex v of G, if we remove from T all cliques not containing v, the remaining subtree stays connected. In other words, any two cliques containing v are either adjacent in T or connected by a path made entirely of cliques that contain v.*

The four conditions of Theorem 7 are clearly satisfied in the graph of Figure 3.9*a*, and none are satisfied in the graph of Figure 3.2 (the diamond is the smallest nonchordal graph). Tarjan and Yannakakis [1984] offer an efficient two-step algorithm for both testing chordality of a graph and *triangulating* it (i.e., filling in the missing links that would make a non-chordal graph chordal).

GRAPH TRIANGULATION (FILL-IN) ALGORITHM

1. Compute an ordering for the nodes, using a *maximum cardinality search*, i.e., number vertices from 1 to $|V|$, in increasing order, always assigning the next number to the vertex having the largest set of previously numbered neighbors (breaking ties arbitrarily).
2. From $n = |V|$ to $n = 1$, recursively fill in edges between any two nonadjacent parents of n, i.e., neighbors of n having lower ranks than n

(including neighbors linked to n in previous steps). If no edges are added the graph is chordal; otherwise, the new filled graph is chordal.

Given a graph $G = (\mathbf{V}, \mathbf{E})$ we can construct a join tree using the following procedure, whose correctness is insured by property 4 of Theorem 7.

ASSEMBLING A JOIN TREE

1. Use the fill-in algorithm to generate a chordal graph G' (if G is chordal, $G = G'$).
2. Identify all cliques in G'. Since any vertex and its parent set (lower ranked nodes connected to it) form a clique in G', the maximum number of cliques is $|V|$.
3. Order the cliques $\boldsymbol{C}_1, \boldsymbol{C}_2, ..., \boldsymbol{C}_t$ by rank of the highest vertex in each clique.
4. Form the join tree by connecting each $\boldsymbol{C}_i$ to a predecessor $\boldsymbol{C}_j$ $(j < i)$ sharing the highest number of vertices with $\boldsymbol{C}_i$.

EXAMPLE: Consider the graph in Figure 3.9*a*. One maximum cardinality ordering is (A, B, C, D, E). Every vertex in this ordering has its preceding neighbors already connected, hence the graph is chordal and no edges need be added. The cliques are ranked $\boldsymbol{C}_1$, $\boldsymbol{C}_2$, and $\boldsymbol{C}_3$ as shown in Figure 3.9*b*. $\boldsymbol{C}_3 = \{C, E\}$ shares only vertex C with its predecessors $\boldsymbol{C}_2$ and $\boldsymbol{C}_1$, so either one can be chosen as the parent of $\boldsymbol{C}_3$. These two choices yield the join trees of Figures 3.9*b* and 3.9*c*.

Now suppose we wish to assemble a join tree for the same graph with the edge (B, C) missing. The ordering (A, B, C, D, E) is still a maximum cardinality ordering, but now when we discover that the preceeding neighbors of node D (i.e., B and C) are nonadjacent, we should fill in edge (B, C). This renders the graph chordal, and the rest of the procedure yields the same join trees as in Figures 3.9*b* and 3.9*c*.

DECOMPOSABLE DISTRIBUTIONS

DEFINITION: *A probability model P is said to be* ***decomposable*** *if it has a minimal I-map that is chordal. P is said to be* ***decomposable relative*** *to a graph G if the following two conditions are met:*

i. G is an I-map of P.

ii. G is chordal.

LEMMA 1: *If P is decomposable relative to G, then any join tree T of the cliques of G is an I-map relative to P. In other words, if* $\boldsymbol{C}_X$, $\boldsymbol{C}_Y$, *and* $\boldsymbol{C}_Z$ *are three disjoint sets of vertices in T, and* $\boldsymbol{X}$, $\boldsymbol{Y}$, *and* $\boldsymbol{Z}$ *are their corresponding sets of variables in G, then* $I(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{Y})_P$ *whenever* $\boldsymbol{C}_Z$ *separates* $\boldsymbol{C}_X$ *from* $\boldsymbol{C}_Y$ *in T (written* $< \boldsymbol{C}_X | \boldsymbol{C}_Z | \boldsymbol{C}_Y >_T$*).*

Proof: Since $(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{Y})$ may not be disjoint, we will prove $I(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{Y})_P$ by showing that $I(\boldsymbol{X}-\boldsymbol{Z}, \boldsymbol{Z}, \boldsymbol{Y}-\boldsymbol{Z})_P$ holds the two assertions are equivalent, according to Remark 2 of Theorem 1. Moreover, since G is an I-map of P, it is enough to show that $\boldsymbol{Z}$ is a cutset in G, separating $\boldsymbol{X}-\boldsymbol{Z}$ from $\boldsymbol{Y}-\boldsymbol{Z}$. Thus, we need to show

$$< \boldsymbol{C}_X | \boldsymbol{C}_Z | \boldsymbol{C}_Y >_T \implies < \boldsymbol{X}-\boldsymbol{Z} | \boldsymbol{Z} | \boldsymbol{Y}-\boldsymbol{Z} >_G, \tag{3.23}$$

which we shall prove by contradiction in two parts:

Part 1: If the right-hand side of Eq. (3.23) is false, then there exists a path $\alpha, \gamma_1, \gamma_2, \ldots, \gamma_n, \beta$ in G that goes from some element $\alpha \in \boldsymbol{X}-\boldsymbol{Z}$ to some element $\beta \in \boldsymbol{Y}-\boldsymbol{Z}$ without intersecting $\boldsymbol{Z}$, namely,

$$(\alpha, \gamma_1) \in \boldsymbol{E}\ (\gamma_i, \gamma_{i+1}) \in \boldsymbol{E},\ (\gamma_n, \beta) \in \boldsymbol{E} \text{ and } \gamma_i \notin \boldsymbol{Z}$$

for all $i = 1, 2, \ldots, n$.

Proof of Part 1: Let $\boldsymbol{C}_v$ denote the set of all cliques that contain some vertex v, and consider the set of cliques

$$S = \{\boldsymbol{C}_\alpha \cup_i \boldsymbol{C}_{\gamma_i} \cup \boldsymbol{C}_\beta - \boldsymbol{C}_Z\}.$$

We now argue that those vertices of T corresponding to the elements of S form a connected sub-tree. Indeed, T was constructed so that pulling out the variables in $\boldsymbol{C}_Z$ would leave the vertices of every $\boldsymbol{C}_{\gamma_i}$ connected. Moreover, the existence of an edge (γ_i, γ_{i+1}) in G guarantees that every clique containing γ_i shares an element (γ_i) with each clique containing (γ_i, γ_{i+1}); Each clique containing (γ_i, γ_{i+1}), in turn, shares an element (γ_{i+1}) with every clique containing γ_{i+1}. Consequently, the vertices corresponding to the elements of $\boldsymbol{C}_{\gamma_i}$ and $\boldsymbol{C}_{\gamma_{i+1}}$ are connected in T, even after the variables in $\boldsymbol{C}_Z$ are deleted.

Part 2: Part 1 asserts the existence of a path in T from some vertex in $\boldsymbol{C}_\alpha \subseteq \boldsymbol{C}_X$ to some vertex in $\boldsymbol{C}_\beta \subseteq \boldsymbol{C}_Y$, bypassing all vertices of $\boldsymbol{C}_Z$, thus contradicting the antecedent of Eq. (3.23). Q.E.D.

We are now in a position to demonstrate that decomposable models have joint distribution functions expressible in product form. Essentially, the demonstration relies on property iv of Theorem 7, which allows us to arrange the cliques of G as a tree and apply to them the chain rule formula (Eq. (3.19)), as we have done to the individual variables in Eq. (3.20).

THEOREM 8: *If P is decomposable relative to G, then the joint distribution of P can be written as a product of the distributions of the cliques of G divided by a product of the distributions of their intersections.*

Proof: Let T be the join tree of the cliques of G, and let $(\boldsymbol{C}_1, \boldsymbol{C}_2, ..., \boldsymbol{C}_i, ...)$ be an ordering of the cliques that is consistent with T, i.e., for every $i > j$ we have a unique predecessor $j(i) < i$ such that $\boldsymbol{C}_{j(i)}$ is adjacent to $\boldsymbol{C}_i$ in T. Clearly, $\boldsymbol{C}_{j(i)}$ separates $\boldsymbol{C}_i$ from $\boldsymbol{C}_1, \boldsymbol{C}_2, ..., \boldsymbol{C}_{i-1}$ in any such ordering. Applying the chain rule formula to the cliques of G, we obtain

$$P(x_1, x_2, ..., x_n) = \prod_i P(\boldsymbol{c}_i \mid \boldsymbol{c}_1, ..., \boldsymbol{c}_{i-1}) = \prod_i P(\boldsymbol{c}_i \mid \boldsymbol{c}_{j(i)}) \tag{3.24}$$

$$= \prod_i P(\boldsymbol{c}_i \mid \boldsymbol{c}_i \cap \boldsymbol{c}_{j(i)}) \tag{3.25}$$

$$= \prod_i \frac{P(\boldsymbol{c}_i)}{P(\boldsymbol{c}_i \cap \boldsymbol{c}_{j(i)})}. \tag{3.26}$$

Eq. (3.24) follows from the I-mapness of T (Lemma 1), and Eq. (3.25) follows from the I-mapness of G, since the variables that $\boldsymbol{C}_{j(i)}$ does not share with $\boldsymbol{C}_i$ are separated from those in $\boldsymbol{C}_i$ by the variables common to both $\boldsymbol{C}_i$ and $\boldsymbol{C}_{j(i)}$. In Figure 3.9*a*, for example, A is separated from D by $\{B, C\}$. Q.E.D.

To render P decomposable relative to some graph G, it is enough that G be any I-map of P; it need not be minimal. Thus, if we wish to express P as a product of marginal distributions of clusters of variables, and the Markov network G_0 of P happens to be non-chordal, it is possible to make G_0 chordal by filling in the missing chords and expressing P as a product of distributions defined on the cliques of the resulting graph. For example, if the Markov network of a certain model is given by the graph of Figure 3.9*a* with edge (BC) missing (as in Figure 3.2), G_0 is not chordal, and we cannot express P as a product of the pairwise distributions $P(a, b)$, $P(a, c)$, $P(c, d)$, $P(d, b)$, and $P(e, d)$. However, by filling in the link (B, C) we create a chordal I-map G of P (Theorem 5), and we can express P as a product of distributions on the cliques of G, as in Eq. (3.22). It is true that the condition $I(B, AD, C)$ is not explicit in the expression of Eq. (3.22) and can be encoded only by careful numerical crafting of the distributions $P(a, b, c)$ and $P(b, c, d)$. However, once encoded, the tree structure of the cliques of G facilitates convenient, recursive updating of probabilities [Lauritzen and Spiegelhalter 1988], as will be shown in Section 4.4.1. Moreover, in situations where the cluster distributions are obtained by statistical measurements, the graph triangulation method can help the experimenter select the right variable clusters for measurement [Goldman and Rivest 1986]. For example, in the model depicted by Figure 3.2, graph triangulation would prompt the experimenter to tabulate measurements of variable triplets (such as $\{M_1, F_1, F_2\}$ and $\{M_2, F_1, F_2\}$) as well as variable pairs.

3.3 BAYESIAN NETWORKS

The main weakness of Markov networks is their inability to represent induced and non-transitive dependencies; two independent variables will be directly connected by an edge, merely because some other variable depends on both. As a result, many useful independencies go unrepresented in the network. To overcome this deficiency, Bayesian networks use the richer language of *directed* graphs, where the directions of the arrows permit us to distinguish genuine dependencies from spurious dependencies induced by hypothetical observations. Reiterating the example of Section 3.1.3, if the sound of a bell is functionally determined by the outcomes of two coins, we will use the network *coin 1* $\rightarrow$ *bell* $\leftarrow$ *coin 2*, without connecting *coin 1* to *coin 2*. This network reflects the natural perception of causal influences; the arrows indicate that the sound of the bell is determined by the coin outcomes, which are mutually independent.

These arrows endow special status on paths that traverse converging arrows, like the path leading from *coin 1* to *coin 2* through *bell*. Such a path should not be interpreted as forming a connection between the variables at the tails of the arrows; the connection should be considered nonexistent, or *blocked*, until the variable *bell* (or any of its descendents) is instantiated. This direction-dependent criterion of connectivity, called *d-separation*, captures the induced dependency relationship among the three variables: the outcomes of the two coins are marginally independent, but they become mutually dependent when we learn the outcome of the bell (or any external evidence bearing on that outcome). The *d*-separation criterion is replaced by the usual cutset criterion of Markov networks whenever the arrows are diverging (*height* $\leftarrow$ *age* $\rightarrow$ *reading ability*) or cascaded (*weather* $\rightarrow$ *wheat crop* $\rightarrow$ *wheat price*).

A formal definition of the *d*-separation criterion for general directed acyclic graphs (DAGs) is given in Section 3.3.1. The criterion permits us to determine by inspection which sets of variables are considered independent of each other given a third set, thus making any DAG an unambiguous representation of dependency. In Section 3.3.2 we examine the possibility of using DAGs as minimal *I*-maps for probabilistic models, in much the same way that undirected graphs were used as minimal *I*-maps for Markov networks. Such minimal *I*-map DAGs will be called *Bayesian networks*.

In keeping with our treatment of Markov networks at the beginning of Section 3.2, we now address the following questions regarding Bayesian networks:

1. Given a probability distribution P, can we construct an edge-minimal DAG D that is an I-map of P?
2. Given a pair (P, D) can we test whether D is a (minimal) I-map of P?
3. Given a DAG D, can we construct a probability distribution P such that D is a perfect map of P?

Once again, the first two questions have simple solutions obtained by nonnumeric analysis and based solely on the axioms of conditional independence (Eq. (3.6)). This time, however, the semi-graphoid axioms, Eqs. (3.6*a*) through (3.6*d*), are used in the derivations, with the intersection axiom, Eq. (3.6*e*), playing only a minor role. Thus, the directionality of the arrows gives Bayesian networks another advantage over Markov networks; the requirement of strict positivity (i.e., the axiom of intersection) is no longer necessary for constructing an *I*-map from local dependencies. Hence, the network can serve as an inference instrument for logical and functional dependencies, too.

An even bigger advantage, perhaps, of the directed graph representations, is that they make it easy to quantify the links with local, conceptually meaningful parameters that turn the network as a whole into a globally consistent knowledge base. This feature is discussed in Section 3.3.2. Finally, in Section 3.3.3 we compare Bayesian networks with Markov networks for expressive power and range of applicability.

3.3.1 Dependence Semantics for Bayesian Networks

Bayesian networks are DAGs in which the nodes represent variables, the arcs signify the existence of direct causal influences between the linked variables, and the strengths of these influences are expressed by forward conditional probabilities.

The semantics of Bayesian networks demands a clear correspondence between the topology of a DAG and the dependence relationships portrayed by it. With Markov networks this correspondence was based on a simple separation criterion: If the removal of some subset $\mathbf{Z}$ of nodes from the network rendered nodes $\mathbf{X}$ and $\mathbf{Y}$ disconnected, then $\mathbf{X}$ and $\mathbf{Y}$ were proclaimed to be independent given $\mathbf{Z}$, i.e.,

$$< \mathbf{X} | \mathbf{Z} | \mathbf{Y} >_G \implies I(\mathbf{X}, \mathbf{Z}, \mathbf{Y}).$$

DAGs use a slightly more complex separability criterion, called *d*-separation, which takes into consideration the directionality of the arrows in the graph.

DEFINITION: *If* $\mathbf{X}$, $\mathbf{Y}$, *and* $\mathbf{Z}$ *are three disjoint subsets of nodes in a DAG* D, *then* $\mathbf{Z}$ *is said to* ***d-separate*** $\mathbf{X}$ *from* $\mathbf{Y}$, *denoted* $< \mathbf{X} | \mathbf{Z} | \mathbf{Y} >_D$, *if along every path between a node in* $\mathbf{X}$ *and a node in* $\mathbf{Y}$ *there is node* w *satisfying one of the following two conditions: (1)* w *has converging arrows and none of* w *or its descendants are in* $\mathbf{Z}$, *or (2)* w *does not have converging arrows and* w *is in* $\mathbf{Z}$.

If a path satisfies the condition above, it is said to be *blocked*; otherwise, it is said to be *activated* by $\mathbf{Z}$. In Figure 3.10, for example, $\boldsymbol{X} = \{2\}$ and $\boldsymbol{Y} = \{3\}$ are d-separated by $\mathbf{Z} = \{1\}$; the path $2 \leftarrow 1 \rightarrow 3$ is blocked by $1 \in \mathbf{Z}$, and the path $2 \rightarrow 4 \leftarrow 3$ is blocked because 4 and all its descendants are outside $\mathbf{Z}$. $\boldsymbol{X}$ and $\boldsymbol{Y}$ are not d-separated by $\mathbf{Z}' = \{1, 5\}$, however, because the path $2 \rightarrow 4 \leftarrow 3$ is rendered active: learning the value of the consequence 5 renders 5's causes, 2 and 3, dependent.

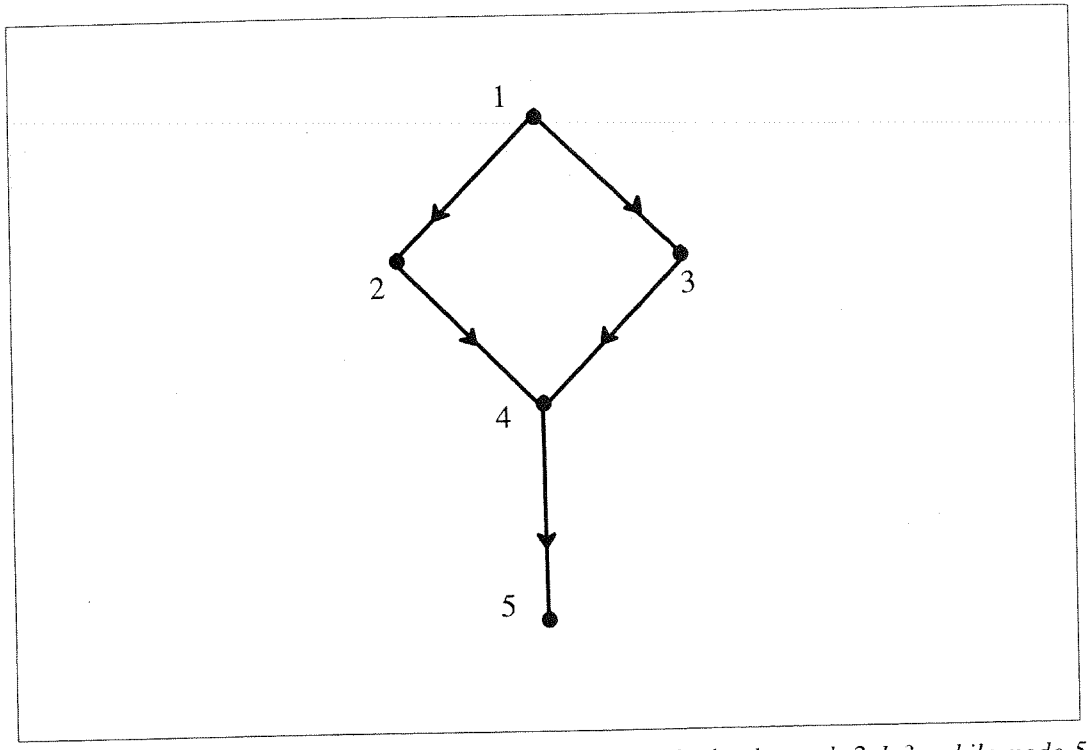


Figure 3.10. *A DAG depicting d-separation; node 1 blocks the path 2-1-3, while node 5 activates the path 2-4-3.*

The procedure for testing d-separation is only slightly more complicated than the conventional test for cutset separation in undirected graphs, and it can be handled by visual inspection. The only difference is that pathways along converging arrows, representing predicted events, are considered blocked until activated by evidential information. This is a basic pattern of diagnostic reasoning; for example, two inputs of a logic gate are presumed independent, but if the output becomes known, what we learn about one input has bearing on the other.

BAYESIAN NETWORKS AS I-MAPS

DEFINITION: *A DAG D is said to be an* ***I-map*** *of a dependency model M if every d-separation condition displayed in D corresponds to a valid conditional independence relationship in M, i.e., if for every three disjoint sets of vertices* $\mathbf{X}$, $\mathbf{Y}$, *and* $\mathbf{Z}$ *we have*

$$\langle \mathbf{X} | \mathbf{Z} | \mathbf{Y} \rangle_D \implies I(\mathbf{X}, \mathbf{Z}, \mathbf{Y})_M .$$

A DAG is a ***minimal*** *I-map of M if none of its arrows can be deleted without destroying its I-mapness.*

DEFINITION: *Given a probability distribution P on a set of variables* $\mathbf{U}$, *a DAG* $D = (\mathbf{U}, \vec{\mathbf{E}})$ *is called a* ***Bayesian network*** *of P iff D is a minimal I-map of P.*

We now address the task of constructing a Bayesian network for any given distribution P.

DEFINITION: *Let M be a dependency model defined on a set* $\mathbf{U} = \{X_1, X_2, ..., X_n\}$ *of elements, and let d be an ordering* $(X_1, X_2, ..., X_i, ...)$ *of the elements of* $\mathbf{U}$. *The* ***boundary strata*** *of M relative to d is an ordered set of subsets of* $\mathbf{U}$, $(\mathbf{B}_1, \mathbf{B}_2, ..., \mathbf{B}_i, ...)$, *such that each* $\mathbf{B}_i$ *is a Markov boundary of* X_i *with respect to the set* $\mathbf{U}_{(i)} = \{X_1, X_2, ..., X_{i-1}\}$, *i.e.,* $\mathbf{B}_i$ *is a minimal set satisfying* $\mathbf{B}_i \subseteq \mathbf{U}_{(i)}$ *and* $I(X_i, \mathbf{B}_i, \mathbf{U}_{(i)} - \mathbf{B}_i)$. *The DAG created by designating each* $\mathbf{B}_i$ *as parents of vertex* X_i *is called a* ***boundary DAG*** *of M relative to d.*

THEOREM 9: [Verma 1986]: *Let M be any semi-graphoid (i.e., any dependency model satisfying the axioms of Eqs. (3.6a) through (3.6d)). If D is a boundary DAG of M relative to any ordering d, then D is a minimal I-map of M.*

Theorem 9 is the key to constructing and testing Bayesian networks, as will be shown via three corollaries. The first corollary follows from the fact that every probability distribution P is a semi-graphoid (see Theorem 1).

COROLLARY 3: *Given a probability distribution* $P(x_1, x_2, ..., x_n)$ *and any ordering d of the variables, the DAG created by designating as parents of* X_i *any minimal set* $\mathbf{\Pi}_{X_i}$ *of predecessors satisfying*

$$P(x_i | \mathbf{\Pi}_{X_i}) = P(x_i | x_1, ..., x_{i-1}) , \quad \mathbf{\Pi}_{X_i} \subseteq \{X_1, X_2, ..., X_{i-1}\} \tag{3.27}$$

is a Bayesian network of P. If P is strictly positive, then all of the parent sets are unique (see Theorem 4) and the Bayesian network is unique (given d).

Although the structure of a Bayesian network depends strongly on the node ordering used in constructing it, each network nevertheless is 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 independent of the construction ordering. An immediate corollary of this observation yields an order-independent definition of Bayesian networks and a solution to Question 2 from the beginning of this section.

COROLLARY 4: *Given a DAG D and a probability distribution P, a necessary and sufficient condition for D to be a Bayesian network of P is that each variable X be conditionally independent of all its non-descendants, given its parents* $\mathbf{\Pi}_X$, *and that no proper subset of* $\mathbf{\Pi}_X$ *satisfy this condition.*

The "necessary" part holds because every parent set $\mathbf{\Pi}_X$ d-separates X from all its non-descendants. The "sufficient" part holds because X's independence of all its non-descendants means X is also independent of its predecessors in a particular ordering d (as required by Corollary 3).

COROLLARY 5: *If a Bayesian network D is constructed by the boundary-strata method in some ordering d, then any ordering d′ consistent with the direction of arrows in D will give rise to the same network topology.*

Corollary 5 follows from Corollary 4, which ensures that the set $\mathbf{\Pi}_{X_i}$ will satisfy Eq. (3.27) in any new ordering as long as the new set of X_i's predecessors does not contain any of X_i's old descendants. Thus, once the network is constructed, the original order can be forgotten; only the partial ordering displayed in the network matters.

Another interesting corollary of Theorem 9 is a generalization of the celebrated *Markov chain* property, which is used extensively in the probabilistic analysis of random walks, time-series data, and other stochastic processes [Feller 1968; Meditch 1969; Abend, Hartley, and Kanal 1965]. The property states the following: if in a sequence of n trials $X_1, X_2, \ldots, X_n$ the outcome of any trial X_k (where $2 \le k \le n$) depends only on the outcome of the directly preceding trial X_{k-1}, then, given all its predecessors and successors, the outcome of X_k depends on its adjacent outcomes, X_{k-1} and X_{k+1}. Formally,

$$I(X_k, X_{k-1}, X_1 \cdots X_{k-2}) \implies I(X_k, X_{k-1}\, X_{k+1}, X_1 \cdots X_{k-2}\, X_{k+2} \cdots X_n)\,.$$

(The converse holds only in strictly positive distributions, i.e., graphoids.) Theorem 9 generalizes the Markov chain property to non-probabilistic dependencies and to structures that are not chains, and, as the following corollary shows, the d-separation criterion uniquely determines a Markov blanket for any node X in a given Bayesian network (see Eq. (3.12)).

COROLLARY 6: *In any Bayesian network, the union of the following three types of neighbors is sufficient for forming a Markov blanket of a node X: the direct parents of X, the direct successors of X, and all direct parents of X's direct successors.*

Thus, if the network consists of a single path (i.e., is a Markov chain), the Markov blanket of any nonterminal node consists of its two immediate neighbors, as expected. In a tree, the Markov blanket consists of the (unique) father and the immediate successors. In Figure 3.10, however, the Markov blanket of node 3 is {1, 4, 2}. The reason the sets defined by Corollary 6 are Markov blankets but generally are not Markov boundaries is that alternative orderings might give X a different set of neighbors.

BAYESIAN NETWORKS AS A LOGIC OF DEPENDENCIES

A Bayesian network can be viewed as an inference instrument for deducing new independence relationships from those used in constructing the network. The topology of the network is assembled from a list of independence statements that comprise the boundary strata. This input list implies a host of additional statements, many of which can be deduced from the network by graphical criteria such as d-separation. For example, the network in Figure 3.10 was constructed from the boundary strata

$$(\boldsymbol{B}_2 = \{1\}, \boldsymbol{B}_3 = \{1\}, \boldsymbol{B}_4 = \{2, 3\}, \boldsymbol{B}_5 = \{4\}),$$

representing the independency list

$$L = \{I(2, 1, \varnothing), I(3, 1, 2), I(4, 23, 1), I(5, 4, 123)\}.$$

New independence relationships, all of them valid consequences of L, can be deduced from the network—e.g., $I(5, 23, 1)$ and $I(3, 124, 5)$. This raises the following questions:

1. Can d-separation be improved? Can a more sophisticated criterion reveal additional independencies that are valid consequences of the input information?
2. Are there valid consequences that escape graphical representation altogether?

The answer to both questions is no; every valid consequence of the input information L must show up as a d-separation condition in the DAG built from L. This follows from the next theorem.

THEOREM 10: [Geiger and Pearl 1988a]: *For any DAG D there exists a probability distribution P such that D is a perfect map of P relative to d-separation, i.e., P embodies all the independencies portrayed in D, and no others.*

Theorem 10 makes it impossible for some valid consequence σ of the input list to escape detection by *d*-separation. Any such σ is valid in all distributions that obey the input, and hence a probability P as specified in Theorem 10 (a probability that ought to violate σ) cannot exist.

COROLLARY 7: *Given a list L of independence relationships in the form of a boundary strata, a Bayesian network combined with the d-separation criterion constitutes a polynomially sound and complete inference mechanism relative to the closure of L, i.e., it identifies in polynomial time every conditional independence relationship that follows logically from those in L.*

Note, however, that a prerequisite of completeness is that the input be a boundary strata, i.e., that it identify recursively a Markov boundary for each element, in some order *d*. The tractability (and even the decidability) of the general membership problem, relative to an arbitrary noncausal input list of conditional independence statements, hinges upon the completeness conjecture stated in Section 3.1.2. Evidently, there are subtle computational advantages to organizing information in chronologically ordered strata. Whether this feature lends importance to causal schemata in knowledge organization is an interesting topic which we will leave for speculation.

3.3.2 Bayesian Network as a Knowledge Base

STRUCTURING THE NETWORK

In principle, given any joint distribution $P(x_1, ..., x_n)$ and an ordering d on the variables in $\boldsymbol{U}$, Corollary 4 prescribes a simple recursive procedure for constructing a Bayesian network. We start by choosing X_1 as a root and assign to it the marginal probability $P(x_1)$ dictated by $P(x_1, ..., x_n)$. Next, we form a node to represent X_2; if X_2 is dependent on X_1, a link from X_1 to X_2 is established and quantified by $P(x_2 | x_1)$. Otherwise, we leave X_1 and X_2 unconnected and assign the prior probability $P(x_2)$ to node X_2. At the i-th stage, we form the node X_i, draw a group of directed links to X_i from a parent set $\boldsymbol{\Pi}_{X_i}$ defined by Eq. (3.27), and quantify this group of links by the conditional probability $P(x_i | \boldsymbol{\pi}_{X_i})$. The result is a directed acyclic graph that represents many of the independencies embedded in $P(x_1, ..., x_n)$, i.e., all the independencies that follow logically from the definitions of the parent sets (Eq. (3.27)).

Conversely, the conditional probabilities $P(x_i \mid \Pi_{X_i})$ on the links of the DAG should contain all the information necessary for reconstructing the original distribution function. Writing the chain rule formula in the ordering d and using Eq. (3.27), we get the product

$$\begin{aligned} P(x_1, x_2, \ldots, x_n) &= P(x_n \mid x_{n-1}, \ldots, x_1)\, P(x_{n-1} \mid x_{n-2}, \ldots, x_1) \\ &\quad \cdots P(x_3 \mid x_2, x_1)\, P(x_2 \mid x_1)\, P(x_1) \\ &= \prod_i P(x_i \mid \Pi_{X_i}). \end{aligned} \tag{3.28}$$

For example, the distribution corresponding to the DAG of Figure 3.11 can be written by inspection:

$$\begin{aligned} &P(x_1, x_2, x_3, x_4, x_5, x_6) \\ &= P(x_6 \mid x_5)\, P(x_5 \mid x_2, x_3)\, P(x_4 \mid x_1, x_2) \cdot P(x_3 \mid x_1)\, P(x_2 \mid x_1)\, P(x_1). \end{aligned} \tag{3.29}$$

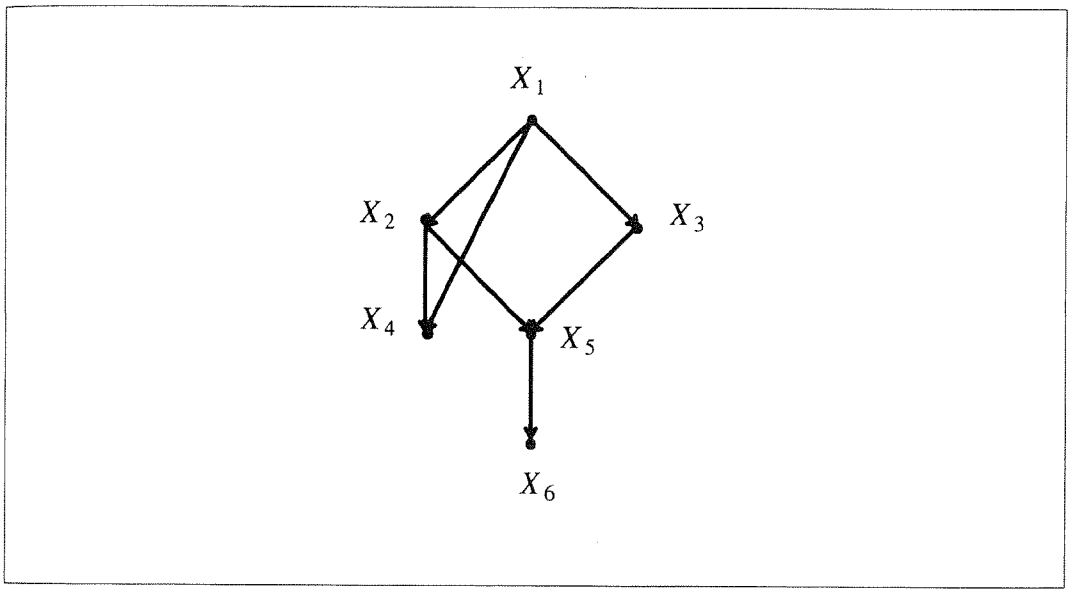


Figure 3.11. *A Bayesian network representing the distribution* $P(x_6 \mid x_5)\, P(x_5 \mid x_2, x_3)\, P(x_4 \mid x_1, x_2)\, P(x_3 \mid x_1)\, P(x_2 \mid x_1)\, P(x_1)$.

In practice, however, a numerical representation for $P(x_1, \ldots, x_n)$ is rarely available. Instead we normally have only intuitive understanding of the major constraints in the domain. The graph can still be configured as before, but the parent sets Π_{X_i} must be identified by human judgment.

The parents of X_i are those variables judged to be *direct causes* of X_i or to have *direct influence* on X_i. The informal notions of causation and influence replace the

formal notion of directional conditional independence as defined in Eq. (3.27). An important feature of the network representation is that it permits people to express directly the fundamental, qualitative relationships of direct influence; the network augments these with derived relationships of *indirect influence* and preserves them, even if the numerical assignments are just sloppy estimates. In Figure 3.11, for example, the model builder did not state that X_6 can tell us nothing new about X_1 once we know X_2 and X_3, but the relationship is logically implied by other inputs and will remain part of the model, regardless of the numbers assigned to the links.

The addition to the network of any new node Y requires that the knowledge provider identify a set $\mathbf{\Pi}_Y$ of variables that bear directly on Y, assess the strength of this relationship, and make no commitment regarding the effect of Y on variables outside $\mathbf{\Pi}_Y$. Even though each judgment is performed locally, their sum is guaranteed to be complete and consistent, as we shall see next.

QUANTIFYING THE LINKS

Suppose we are given a DAG D in which the arrows pointing to each node X_i emanate from a set $\mathbf{\Pi}_{X_i}$ of parent nodes judged to have direct influence on X_i. To specify consistently the strengths of these influences, one need only assess the conditional probabilities $P(x_i \mid \mathbf{\Pi}_{X_i})$ by some functions $F_i(x_i, \mathbf{\Pi}_{X_i})$ and make sure these assessments satisfy

$$\sum_{x_i} F_i(x_i, \mathbf{\Pi}_{X_i}) = 1 \,, \tag{3.30}$$

where $0 \le F_i(x_i, \mathbf{\Pi}_{X_i}) \le 1$ and the summation ranges over the domain of X_i. This specification is complete and consistent because the product form

$$P_a(x_1, \ldots, x_n) = \prod_i F_i(x_i, \mathbf{\Pi}_{X_i}) \tag{3.31}$$

constitutes a joint probability distribution that supports the assessed quantities. In other words, if we compute the conditional probabilities $P_a(x_i \mid \mathbf{\Pi}_{X_i})$ dictated by $P_a(x_1, \ldots, x_n)$, the original assessments $F_i(x_i, \mathbf{\Pi}_{X_i})$ will be recovered:

$$P_a(x_i \mid \mathbf{\Pi}_{X_i}) = \frac{P_a(x_i, \mathbf{\Pi}_{X_i})}{P_a(\mathbf{\Pi}_{X_i})} = \frac{\sum_{x_j \notin (x_i \cup \mathbf{\Pi}_{X_i})} P_a(x_1, \ldots, x_n)}{\sum_{x_j \notin \mathbf{\Pi}_{X_i}} P_a(x_1, \ldots, x_n)} = F_i(x_i, \mathbf{\Pi}_{X_i}). \tag{3.32}$$

Moreover, all the independencies dictated by the choices of $\mathbf{\Pi}_{X_i}$ (corresponding to those in Eq. (3.27)) are embodied in P_a.

Building models this way is much easier than quantifying Markov networks. The parameters requested from the model builder are the conditional probabilities that quantify many conceptual relationships in one's mind, e.g., cause-effect or frame-slot relations; they are psychologically meaningful and can be obtained by direct measurement. The thinking required for assessing the parameters of $P(x_i \mid \Pi_{X_i})$ is estimating the likelihood that the event $X_i = x_i$ will occur, given any instantiation of the variables in Π_{X_i} (for example, the likelihood that a patient will develop a certain symptom, assuming that he suffers from a given combination of disorders). These kinds of assessments are natural because they point to familiar frames (e.g., diseases) by which people organize empirical knowledge.

DAGs constructed by this method will be called *Bayesian belief networks* or *causal networks* interchangeably, the former emphasizing the judgmental origin and probabilistic nature of the quantifiers and the latter reflecting the directionality of the links. Such networks have a long and rich tradition, starting with the geneticist Sewal Wright in 1921. He developed a method called *path analysis* [Wright 1934], which later became an established representation of causal models in economics [Wold 1964], sociology [Blalock 1971; Kenny 1979], and psychology [Duncan 1975]. *Influence diagrams* represent another component in this tradition [Howard and Matheson 1981; Shachter 1986]; developed for decision analysis, they contain both event nodes and action nodes (see Chapter 6). *Recursive models* is the name given to such networks by statisticians seeking meaningful and effective decompositions of contingency tables [Lauritzen 1982; Wermuth and Lauritzen 1983; Kiiveri, Speed, and Carlin 1984].

In the strictest sense, Bayesian networks are not graphs but hypergraphs, because describing the dependency of a given node on its k parents requires a function of $k+1$ arguments; in general, it cannot be specified by k two-place functions on the individual links. Still, the directionality of the arrows and the fact that many parents remain unlinked convey important information that would be lost if we used the standard hypergraph representation and specified only the list of dependent subsets.

If the number of parents k is large, estimating $P(x \mid \Pi_{X_i})$ may be troublesome. In principle, it requires a table of size 2^k (for binary variables), but in practice (as noted in Section 2.2) people structure causal relationships into small prototypical clusters of variables; each requiring about k parameters. Common examples of such structures are noisy OR-gates (i.e., any variable is likely to trigger the effect), noisy AND-gates, and various enabling mechanisms (i.e., variables having no influence of their own except that they enable other influences to take effect). Detailed analysis of the noisy-OR-gate model is given in Section 4.3.2.

THE ROLE OF CAUSALITY

Note that the topology of a Bayesian network can be extremely sensitive to the node ordering d. What is a tree in one ordering might become a complete graph if

that ordering is reversed. For example, if $X_1, ..., X_n$ stands for the outcomes of n independent coins, and X_{n+1} represents the output of a detector triggered when any coin comes up heads, then the Bayesian network will be an inverted tree of n arrows pointing from each of the variables $X_1, ..., X_n$ to X_{n+1}. On the other hand, if the detector's outcome is chosen to be the first variable, say X_0, then the resulting Bayesian network will be a complete graph.

This sensitivity to order may seem paradoxical at first; d can be chosen arbitrarily, whereas people have fairly uniform conceptual structures, e.g., they agree on whether two propositions are directly or indirectly related. This consensus about the structure of dependencies shows the dominant role causality plays in the formation of these structures. In other words, the standard ordering imposed by the direction of causation indirectly induces identical topologies on the networks that people adopt to encode experiential knowledge. Were it not for the social convention of adopting a standard ordering of events that conforms to the flow of time and causation, human communication as we know it might be impossible. Why, then, do we use temporal ordering to organize our memory? It may be because information about temporal precedence is more readily available than other indexing information, or it may be that networks constructed with temporal ordering are inherently more parsimonious (i.e., they display more independencies.) Experience with expert systems applications does not entirely rule out the second possibility [Shachter and Heckerman 1987]. More on this subject can be found in Chapter 8.

3.3.3 How Expressive Are Bayesian Networks?

One might expect the introduction of directionality into the language of graphs to render directed graphs more expressive, i.e., capable of portraying more conditional independencies. We saw, indeed, that the d-separation criterion permits us to display induced and non-transitive dependencies that were excluded from the Markov network vocabulary. So we might ask how DAGs compare for expressive power with undirected graphs and probability models. Two questions arise:

1. Can all dependencies that are representable by a Markov network also be represented by a Bayesian network?
2. How well can Bayesian networks represent the type of dependencies induced by probabilistic models?

The answer to the first question is clearly no. For instance, the dependency structure of a diamond-shaped Markov network (e.g., Figure 3.2) with edges (AB), (AC), (CD), and (BD) asserts two independence relationships: $I(A, BC, D)$ and $I(B, AD, C)$. No Bayesian network can express these two relationships simultaneously and exclusively. If we direct the arrows from A to D, we get

$I(A, BC, D)$ but not $I(B, AD, C)$; if we direct the arrows from B to C, we get $I(B, AD, C)$ but not $I(A, BC, D)$. In view of property iv of Theorem 7, it is clear that this difficulty will always be encountered in non-chordal graphs. No matter how we direct the arrows, there will always be a pair of nonadjacent parents sharing a common child, a configuration that yields independence in Markov networks and dependence in Bayesian networks.

On the other hand, property iv of Theorem 7 also asserts that every chordal graph can be oriented so that the tails of every pair of converging arrows are adjacent. Hence, every dependency model that is isomorphic to a chordal graph is also isomorphic to a DAG. We conclude that the class of probabilistic dependencies that can be represented by both a DAG and an undirected graph are those that form decomposable models, i.e., probability distributions that have perfect maps in chordal graphs. These relationships are shown schematically in Figure 3.12.

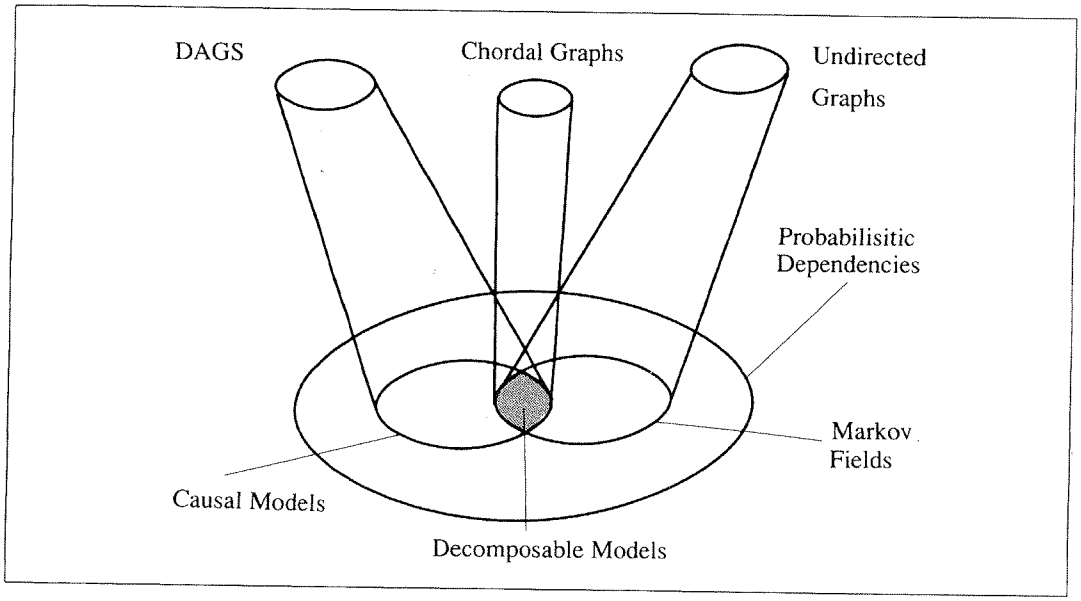


Figure 3.12. *Correspondence between probabilistic models and their graphical representations.*

The answer to Question 2 is also not encouraging. Clearly, no graphical representation can distinguish connectivity between sets from connectivity among their elements. In other word, in both directed and undirected graphs, separation between two sets of vertices is defined in terms of pairwise separation between their corresponding individual elements. In probability theory, on the other hand,

independence of elements does not imply independence of sets (see Eq. (3.6*b*)), as the coins and bell example demonstrated. When the two coins are fair, any two variables are mutually independent, but every variable is (deterministically) dependent on the other two.

CAUSAL MODELS AND THEIR DEPENDENCY STRUCTURE

Despite these shortcomings, we will see that the DAG representation is more flexible than the undirected graph representation, and it captures a larger set of probabilistic independencies, especially those that are conceptually meaningful. To show this, we offer a partial axiomatic characterization of DAG dependencies that indicates clearly where they differ from undirected graph dependencies (Eq. (3.10)) and from probabilistic dependencies (Eq. (3.6)).

DEFINITION: *A dependency model M is said to be* ***causal*** *(or a DAG isomorph) if there is a DAG D that is a perfect map of M relative to d-separation, i.e.,*

$$I(X, Z, Y)_M \iff <X|Z|Y>_D. \tag{3.33}$$

THEOREM 11: *A necessary condition for a dependency model M to be a DAG isomorph is that* $I(X, Z, Y)_M$ *satisfies the following independent axioms (the subscript M is dropped for clarity):*

Symmetry:

$$I(X, Z, Y) \iff I(Y, Z, X). \tag{3.34a}$$

Composition/Decomposition:

$$I(X, Z, Y \cup W) \iff I(X, Z, Y) \;\&\; I(X, Z, W). \tag{3.34b}$$

Intersection:

$$I(X, Z \cup W, Y) \;\&\; I(X, Z \cup Y, W) \Longrightarrow I(X, Z, Y \cup W). \tag{3.34c}$$

Weak union:

$$I(X, Z, Y \cup W) \Longrightarrow I(X, Z \cup W, Y). \tag{3.34d}$$

Contraction:

$$I(X, Z \cup Y, W) \;\&\; I(X, Z, Y) \Longrightarrow I(X, Z, Y \cup W). \tag{3.34e}$$

Weak transitivity:

$$I(X, Z, Y) \;\&\; I(X, Z \cup \gamma, Y) \Longrightarrow I(X, Z, \gamma) \text{ or } I(\gamma, Z, Y). \tag{3.34f}$$

Chordality:

$$I(\alpha, \gamma \cup \delta, \beta) \;\&\; I(\gamma, \alpha \cup \beta, \delta) \Longrightarrow I(\alpha, \gamma, \beta) \text{ or } I(\alpha, \delta, \beta). \tag{3.34g}$$

REMARKS:

1. Symmetry, intersection, weak union, and contraction are identical to the axioms governing probabilistic dependencies (Eq. (3.6)). Composition, weak transitivity, and chordality are constraints that go beyond Eq. (3.6). Thus, not every probabilistic model is a DAG isomorph.

2. Comparing Eq. (3.34) to the axioms defining separation in undirected graphs (Eq. (3.10)), we note that (Eq. (3.10)) implies all axioms in (Eq. (3.34)) except chordality (Eq. (3.34*g*)). Weak union is implied by strong union, composition and contraction are implied by intersection and strong union, and weak transitivity is implied by transitivity.

WEAK TRANSITIVITY

Weak transitivity (Eq. (3.34*f*)) means that if two sets of variables $\boldsymbol{X}$ and $\boldsymbol{Y}$, are both unconditionally and conditionally independent given a singleton variable γ, it is impossible for both $\boldsymbol{X}$ and $\boldsymbol{Y}$ to be dependent on γ. Contrapositively, if $\boldsymbol{X}$ and $\boldsymbol{Y}$ are each dependent on γ, then they must be dependent on each other in some way, either marginally or conditionally given γ. This restriction, which may be violated in some probability models, remains in effect when we associate independence with *d*-separation in DAGs.

THEOREM 12: *d-separation in DAGs is weakly transitive.*

Proof: If both $\boldsymbol{X}$ and $\boldsymbol{Y}$ are *d*-connected to γ in some DAG, then there must be an unblocked path from $\boldsymbol{X}$ to γ and an unblocked path from $\boldsymbol{Y}$ to γ. These two paths form at least one bidirected path from $\boldsymbol{X}$ to $\boldsymbol{Y}$ via γ. If that path traverses γ along converging arrows, it should be unblocked when we instantiate γ, so $\boldsymbol{X}$ and $\boldsymbol{Y}$ cannot be *d*-separated given γ. Conversely, if the arrows meeting at γ do not converge, the path from $\boldsymbol{X}$ to $\boldsymbol{Y}$ is unblocked when γ is uninstantiated, so $\boldsymbol{X}$ and $\boldsymbol{Y}$ cannot be marginally *d*-separated. Q.E.D.

Probability theory does not insist on weak transitivity, as it allows the following four conditions to exist simultaneously:

$$I(\boldsymbol{X}, \varnothing, \boldsymbol{Y})_P,\ I(\boldsymbol{X}, \gamma, \boldsymbol{Y})_P,\ \neg I(\boldsymbol{X}, \varnothing, \gamma)_P,\ \neg I(\boldsymbol{Y}, \varnothing, \gamma)_P.$$

For example, let X and Y be singleton binary variables, $X, Y \in \{TRUE, FALSE\}$, and let γ be a ternary variable, $\gamma \in \{1, 2, 3\}$. Choosing

$$P(x, y, \gamma) = P(x \mid \gamma)\, P(\gamma \mid y)\, P(y),$$

$$P(X = TRUE \mid \gamma) = (1/2,\ 1/4,\ 3/8),$$

$$P(\gamma \mid Y = TRUE) = (1/3,\ 1/3,\ 1/3),$$

$$P(\gamma \mid Y = FALSE) = (1/2,\ 1/2,\ 0)$$

renders γ dependent on both X and Y, yet X and Y are mutually independent, both conditionally (given γ) and unconditionally. Thus, although DAGs seem more capable than undirected graphs of displaying non-transitive dependencies, even DAGs require some weak form of transitivity and cannot capture totally non-transitive probabilistic dependencies. It can be shown, however, that if all variables are either normally distributed or binary, all probabilistic dependencies must be weakly transitive (see Exercise 3.10).

CHORDALITY AND AUXILIARY VARIABLES

The chordality axiom (Eq. (3.34*g*)) excludes dependency models that are isomorphic to non-chordal graphs (such as the one in Figure 3.13*a*), since these cannot be completely captured by DAGs (see Figure 3.12). In essence, Eq. (3.34*g*) insists that we either add the appropriate chords to any long cycle (length ≥ 4), thus disobeying the antecedent of Eq. (3.34*g*), or nullify some of its links, thus satisfying the consequent of Eq. (3.34*g*).

Though DAGs cannot represent non-chordal dependencies, this deficiency can be eliminated by introducing auxiliary variables. Consider the diamond-shaped graph of Figure 3.13*a*, which asserts two independence relationships: $I(A, BC, D)$ and $I(B, AD, C)$.

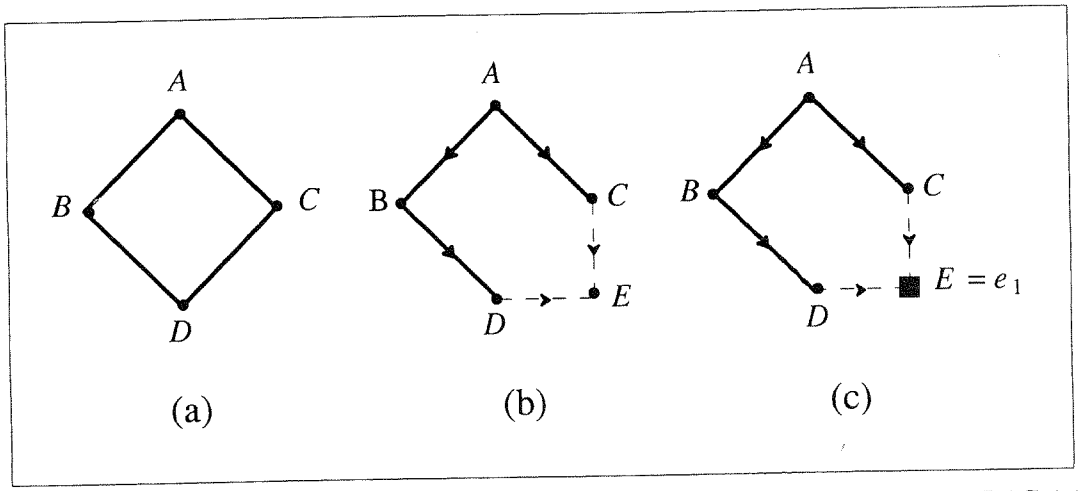


Figure 3.13. *The dependencies of an undirected graph (a) are represented by a DAG (c) using an auxilliary node E.*

Introducing an auxiliary variable E as shown in Figure 3.13*b* creates a DAG model of five variables whose dependencies are represented by the joint distribution function

$$P(a, b, c, d, e) = P(e \mid d, c)\, P(d \mid b)\, P(c \mid a)\, P(b \mid a)\, P(a).$$

Now imagine that we clamp the auxiliary variable E to some fixed value $E = e_1$, as in Figure 3.13*c*. The dependency structure that the clamped DAG induces on A, B, C, D is identical to the original structure (Figure 3.13*a*). Indeed, applying the d-separation criterion to Figure 3.12*c* uncovers the two original independencies: $I(A, BC, D)$ and $I(B, AD, C)$. The marginal distribution of the original variables conditioned upon $E = e_1$ is given by

$$P(a, b, c, d \mid E = e_1) = \frac{P(a, b, c, d, e_1)}{P(e_1)}$$

$$= K\, P(e_1 \mid d, c)\, P(d \mid b)\, P(c \mid a)\, P(b \mid a)\, P(a)$$

$$= g_1(d, c)\, g_2(d, b)\, g_3(a, c)\, g_4(a, b).$$

Using the analysis of Section 3.2.3, we see that this distribution is equivalent to the one portrayed by Figure 3.13*a*. Thus, the introduction of auxiliary variables permits us to dispose of the chordality restriction of Eq. (3.34*g*) and renders the DAG representation superior to that of undirected graphs; every dependency model expressible by the latter is also expressible by the former.

Weak transitivity and chordality are not the only dependencies that are sanctioned by probability theory but are not representable by DAGs. For example, one can show that the following axiom must hold in DAGs:

$$I(\mathbf{Y}, \mathbf{X}, \mathbf{Z}) \;\&\; I(\mathbf{Z}, \mathbf{Y}, \mathbf{X}) \;\&\; I(\mathbf{W}, \mathbf{Z}, \mathbf{X}) \Longrightarrow I(\mathbf{X}, \varnothing, \mathbf{W}).$$

But its denial,

$$I(\mathbf{Y}, \mathbf{X}, \mathbf{Z}) \;\&\; I(\mathbf{Z}, \mathbf{Y}, \mathbf{X}) \;\&\; I(\mathbf{W}, \mathbf{Z}, \mathbf{X}) \;\&\; \neg I(\mathbf{X}, \varnothing, \mathbf{W}),$$

is tolerated by probability theory (see Exercise 3.7). The question arises whether the class of properties specific to DAGs can be characterized axiomatically the way that of undirected graphs was (Theorem 2). The answer is probably no. The results of Geiger [1987] strongly suggest that the number of axioms required for a complete characterization of the d-separation in DAGs is unbounded.

3.4 BIBLIOGRAPHICAL AND HISTORICAL REMARKS

The idea of using graphical representations for probabilistic information can be traced to the geneticist Sewal Wright [1921], who developed the method of *path analysis* "as an aid in the biometric analysis of certain classes of data." The method came under severe attack (e.g., Niles [1922]) and was shunned by statisticians

during the first half of the 20th century (an era ruled by hard data and quantitative analysis), until it was discovered by economists, psychologists, and sociologists (see Section 7.2). The 1960s saw a reversal of this outlook, with statisticians such as Vorobev [1962], Goodman [1970], and Haberman [1974] realizing that some decomposition properties of statistical tables (called log-linear models) can best be expressed in graphical terms. These explorations led to an appreciation of the desirable properties of join trees, which were later recognized by database researchers [Beeri et al. 1983]. Lemmer [1983] has suggested the use of trees of local events groups (LEGs) for Bayesian updating, and Spiegelhalter [1986] proposed the fill-in algorithm to transform Bayesian networks into join trees. Other mathematical properties of chordal graphs are given in Golumbic [1980].

The development of Markov fields progressed in parallel, but from an opposite direction. Here, the network topology was presumed to be given (usually a geometrical arrangement of physical elements in space), and the problem was to characterize the probabilistic behavior of a system complying with the dependencies prescribed by the network. A survey of Markov fields can be found in Isham [1981]. Lauritzen [1982] applied the theory of Markov fields to the analysis of statistical tables and derived Theorems 3, 4, and 5 for independencies embedded in strictly positive probability distributions. Application of Markov fields to pattern recognition and vision are reported in Abend, Hartley, and Kanal [1965], Kanal [1981], and Geman and Geman [1984].

Since graphoids are a central theme of this chapter, and since the theory is still in its embryonic stage, we take the liberty now of presenting an extended history of this development.

The theory of graphoids was conceived in the summer of 1985, when Azaria Paz visited UCLA and he and I began collaborating on the problem of graphical representations. Inspired by Lauritzen's lecture notes on contingency tables [Lauritzen 1982], I sought axiomatic conditions on a dependency model M that would include probabilistic dependencies as a special case, such that the graph construction of Eq. (3.11) would yield an I-map of M. I posed the problem to Professor Paz, we labored for a few weeks, and he came up with a proof of what later became Theorem 3. Surprisingly, only three axioms were needed: symmetry, decomposition, and intersection. These, unfortunately, were not sufficient for Corollary 4, which Lauritzen listed among the properties of Markov fields. We then set out to discover what additional axioms were needed to ensure that the graph obtained by the edge-deletion method be identical to that built by the Markov boundary method. This led to Theorem 4, and to the identification of weak union as the final axiom we needed to fully characterize the graphical properties of Markov fields. The prospects of providing similar characterization for graph separation led to Theorem 2.

Strangely, the contraction axiom was not needed for Theorem 3 or for Theorem 4, but when added to the other four axioms of Eq. (3.6) it enabled us to derive all properties of probabilistic dependencies that we managed to dream up. Hence, we

posed the completeness of these axioms as a conjecture[1], and coined the name *graphoid*.

Around this time, Thomas Verma began examining the validity of *d*-separation in DAGs (Theorem 9). I had introduced this criterion without proof [Pearl 1986c], since my attempts to demonstrate its general validity got entangled in messy probability formulas. I therefore suggested that Tom try a "clean" proof, using the graphoid axioms only, and to our surprise he managed to do it without the intersection axioms [Verma 1986]. This led to semi-graphoids, and to directed graph representations of both probabilistic and logical dependencies; we finally understood how important the contraction property is for causal modeling. The generality of this result made us confident that dependency theorists dealing with databases and qualitative dependencies will eventually adopt DAGs as a representation scheme for their semi-graphoids, e.g., EMVD [Fagin 1977].

In December 1985, Glenn Shafer mentioned a possible connection between graphoids and previous work of A. P. Dawid. As it turned out, Dawid had presented axioms equivalent to Eqs. (3.6.a) through (3.6.d) as early as 1979 [Dawid 1979] but apparently was not concerned with their completeness or their relation to graphical representations. Smith [1989] has recognized the generality of Dawid's axioms and has used them to prove Corollaries 4 and 5 without resorting to probabilistic manipulations (unlike the treatment of Howard and Matheson [1981]).

The power of the *d*-separation criterion would have remained only partially appreciated without Geiger's proof of Theorem 10. Aside from showing that *d*-separation cannot be improved, the theorem legitimizes the use of DAGs as a representation scheme for probabilistic dependencies; a model builder who uses the language of DAGs to express dependencies is shielded from inconsistencies.

Recent advances in the theory of graphoids and Bayesian networks are reported in the references below.[2]–[6]

---

(1) The conjecture has recently been refuted by Studeny, M. "Conditional Independence Relations Have No Finite Complete Characterization," *Proc. of 11th Prague Conf. on Inf. Theory, Statist. Decision Funct. and Random Processes*, Prague, 1990. Also *Kybernetika*, 25(1-3), 1990, 72-79.

(2) R.M. Oliver and J.Q. Smith (Eds), *Influence Diagrams, Belief Nets and Decision Analysis*, Sussex, England: John Wiley & Sons, Ltd., 1990.

(3) D. Geiger, "Graphoids: A Qualitative Framework for Probabilistic Inference." Ph.D. Dissertation. University of California Los Angeles, Computer Science Dept. January 1990.

(4) Shachter, R., (ed.), Special Issue on Influence Diagrams, *Networks*, 20(5), 1990.

(5) D. Geiger, A. Paz, and J. Pearl, "Axioms and Algorithms for Inferences Involving Probabilistic Independence," *Information and Computation*, Vol. 91, No. 1, March 1991, 128-141.

(6) Geva, R., "Representation of Irrelevance Relations by Graphs," M.Sc. Thesis, Dept. of Computer Science, Technion, Haifa, Israel, 1989.

Exercises

3.1. Show that Eqs. (3.6*a*) through (3.6*d*) imply the chaining rule

$$I(X, Y, Z) \,\&\, I(XY, Z, W) \Longrightarrow I(X, Y, W)$$

and show that this rule cannot replace Eq. (3.6*d*) in the set of axioms.

3.2. Show which axioms of Eq. (3.6) are satisfied by the following dependency models:

a. Let U be the set of nodes in an undirected graph G, and let X, Y, and Z be three disjoint sets of nodes in G. $I(X, Z, Y)_{M_1}$ iff all shortest paths between a node $X \in X$ and a node $Y \in Y$ are intercepted by some node in Z.

b. Let U be the set of nodes in an undirected graph G, and let X, Y, and Z be three disjoint sets of nodes in G. $I(X, Z, Y)_{M_2}$ iff all shortest paths between the sets X and Y are intercepted by Z.

c. Let U be the set of points in a three-dimensional Euclidean space, and let X, Y, and Z be three disjoint regions of U. $I(X, Z, Y)_{M_3}$ iff every ray of light from a point in X to some point in Y is intercepted by Z.

d. Let U be the set of n-dimensional vectors, and let X, Y, and Z be three disjoint sets of such vectors. Denote by S_Z the linear subspace spanned by any set Z. $I(X, Z, Y)_{M_4}$ iff the closest distance between X and S_Z is equal to the closest distance between X and $S_{Z \cup Y}$.

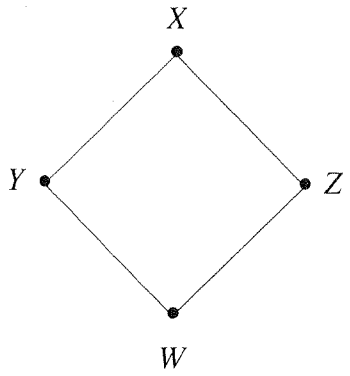
e. Let U be a set of random variables, let P be a probability distribution on those variables, and let X, Y, and Z be three disjoint subsets of U. $I(X, Z, Y)_{M_5}$ iff

$$P(x, z) > 0 \;\&\; P(y, z) > 0 \Longrightarrow P(x, y, z) > 0.$$

3.3. Let $U = \{X, Y, Z, W\}$, and let $P(x, y, z, w)$ be given by the following table:

X	Y	Z	W	P
1	1	1	1	⅓
1	2	2	2	⅓
2	2	1	3	⅓
all other tuples				0

a. Show that the graph G given below is a minimal I-map of P.



b. Show that P cannot be expressed as a product of functions on the cliques of G.

c. Find a tree I-map of P and express P as a product of functions on its edges.

d. Draw all the Bayesian networks of P in the orderings (X, Y, Z, W) and (W, X, Y, Z) and compute their parameters.

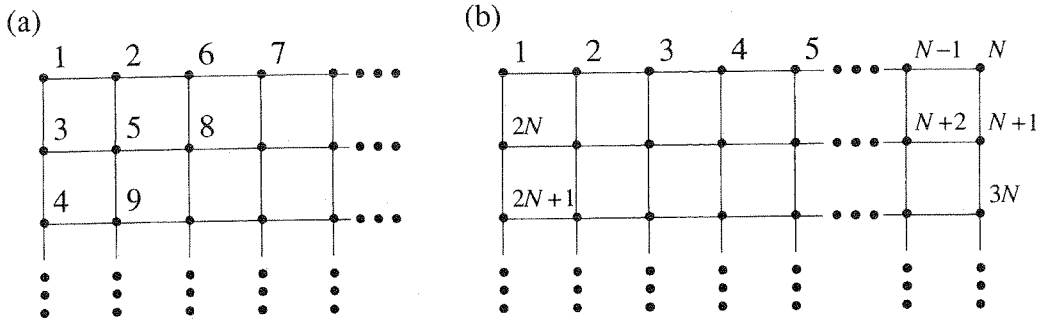
.4. **a.** Find the graphoid closure I^* of the set $I = \{(1, \varnothing, 2), (12, 3, 4)\}$.

b. Construct the Markov network of I^*.

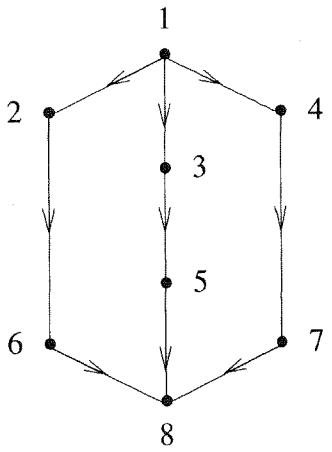
c. Construct the Bayesian networks of I^* corresponding to the following 3 orderings: (1, 2, 3, 4), (4, 3, 2, 1), (1, 4, 2, 3).

Note: The graphoid closure I^* is the smallest superset of I that is consistent with the axioms of Eqs. (3.6*a*) through (3.6*e*).

3.5. We wish to construct a Bayesian network for a Markov field of an $N \times N$ grid in the plane. Find the set of parents of a typical node (e.g., row 3, column 5), in the following two orderings:



3.6. **a.** Find the Markov network G_0 of a probabilistic model P for which the following DAG is a perfect-map:



b. Find an undirected graph G such that P (in problem (**a**)) is decomposable relative to G.

c. Draw a join tree of G.

d. Find an algebraic representation of P such that $P > 0$ for all events.

3.7. (After D. Geiger)

a. Prove that the following axiom holds for all DAGs:

$$I(\alpha_2, \alpha_1, \alpha_3) \;\&\; I(\alpha_3, \alpha_2, \alpha_4) \;\&\; I(\alpha_4, \alpha_3, \alpha_1) \Longrightarrow I(\alpha_1, \varnothing, \alpha_4)$$

(hint: use the definition of d-separation and prove by contradiction).

b. Generalize your arguments and prove that the following axiom holds as well:

$$I(\alpha_2, \alpha_1, \alpha_3) \,\&\, I(\alpha_3, \alpha_2, \alpha_4) \,\&\, \cdots \,\&\, I(\alpha_n, \alpha_{n-1}, \alpha_{n+1}) \,\&\, I(\alpha_{n+1}, \alpha_n, \alpha_1) \Longrightarrow I(\alpha_1, \varnothing\, \alpha_{n+1})$$

(where $n > 3$).

c. Construct a probability distribution that violates the axiom problem (**a**).

3.8. (After D. Geiger). Let P be a zero-mean normal distribution over n variables $X_1, \ldots, X_n$ with a covariance matrix $\Gamma = (\rho_{ij})$, where

$$\rho_{ij} = E[X_i \cdot X_j] \text{ and } E[X_i^2] = 1 \qquad (1 \le i,j \le n).$$

a. Prove the following propositions:

$$I(X_i, \varnothing, X_j)_P \Longleftrightarrow \rho_{ij} = 0,$$
$$I(X_i, X_k, X_j)_P \Longleftrightarrow \rho_{ij} = \rho_{ik} \cdot \rho_{jk},$$

b. In light of Exercises **3.7a** and **3.8a** construct a normal distribution P such that no DAG is a perfect map of P.

3.9. **a.** Show that the axioms of Eqs. (3.34*a*) through (3.34*g*) do not preclude the occurrence of

$$I(\boldsymbol{X}, \varnothing, \boldsymbol{Z}) \,\&\, \neg I(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}) \,\&\, I(\boldsymbol{Y}, \varnothing, \boldsymbol{W}) \,\&\, \neg I(\boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{W}).$$

b. Show a DAG where Eqs. (3.34*a*) through (3.34*g*) hold (in *d*-separation) and $\boldsymbol{X}$, $\boldsymbol{Y}$, $\boldsymbol{Z}$, and $\boldsymbol{W}$ are singleton nodes. (The DAG may have more than four nodes.)

c. Discuss the significance of problem (**b**) *vis a vis* the prospects of defining causal directionality in terms of dependencies.

3.10. Show that weak transitivity holds in
(a) every normal distribution, and
(b) every probability distribution over binary variables, relative to $\boldsymbol{Z}=\varnothing$.

3.11. A *recursive diagram* [Wermuth and Lauritzen 1983] is a DAG constructed as follows: the elements of $\boldsymbol{U}$ are ordered $X_1, \ldots, X_n$ and the

parents set S_i of each X_i is defined by $S_i = \{X_j : j < i \text{ and } \neg I(X_i, \{X_1, ..., X_{i-1}\} - X_j, X_j)\}$, namely, X_j is a parent of X_i if it remains dependent on X_i, given all other predecessors $\{X_1, ..., X_{i-1}\} - X_j$ of X_i.

a. Show that any recursive diagram constructed for a graphoid (i.e., a dependency model satisfying (3.6.*a*) to (3.6.*e*)) coincides with the Bayesian network constructed under the same ordering.

b. Show that for semi-graphoids (i.e., dependency models satisfying (3.6.*a*) to (3.6.*d*)) a recursive diagram is a subgraph of any Bayesian network constructed under the same ordering.

c. Find a probability distribution for which the Bayesian network is a chain but the recursive diagram has only one arc.

d. A recursive diagram R of a semi-graphoid M, has the shape of a linear chain of five nodes $1 \rightarrow 2 \rightarrow 3 \rightarrow 4 \rightarrow 5$. Using the same node ordering, draw all the DAGS that are guaranteed to be I-maps of M.

3.12. Given two DAGs, D_1 and D_2, on the same set of variables. Devise a polynomial time algorithm to test whether D_1 is an I-map of D_2 relative to d-separation.[1] What if D_2 contains a subset of the variables in D_1?[2]

3.13. A dag D is said to be a (causal) **model** of a probability function P if it is consistent with P, that is, the links of D can be annotated with conditional probabilities whose product equals P (see Eq. (3.28)). D is said to be a **minimal** model of P if the set of probabilities consistent with D is not a superset of those consistent with some other model of P.

a. Show that D is consistent with P iff it is an I-map of P.

Let a probability function P be isomorphic to a dag D:

b. Show that every minimal model of P must be a perfect map of D.

c. Show that every minimal model of P has the same arcs as D.

d. Identify the arcs in D whose orientation remains the same in all minimal models of P.[2]

---

(1) Pearl, J., D. Geiger & T. Verma, "The Logic and Influence Diagrams," in R.M. Oliver and J.Q. Smith (Eds.), *Influence Diagrams, Belief Nets and Decision Analysis*, Wiley, 1990, 67-87.

(2) T. Verma & J. Pearl, "Equivalence and Synthesis of Causal Models," in *Proc., Sixth Conf. on Uncertainty in AI*, Cambridge, Mass., 1990, 220-227. Also, North Holland, *UAI* 6, 1991, 255-268.

Appendix 3-A

Proof of Theorem 3

THEOREM 3 [Pearl and Paz 1985]: *Every dependency model M satisfying symmetry, decomposition, and intersection (see Eq. (3.6)) has a (unique) minimal I-map* $G_0 = (\boldsymbol{U}, \boldsymbol{E}_0)$ *produced by connecting only those pairs* (α, β) *for which* $I(\alpha, \boldsymbol{U} - \alpha - \beta, \beta)_M$ *is FALSE, i.e.,*

$$(\alpha, \beta) \notin \boldsymbol{E}_0 \quad \textit{iff } I(\alpha, \boldsymbol{U} - \alpha - \beta, \beta)_M. \tag{3.11}$$

Proof:

1. We first prove that G_0 is an I-map (i.e., $<\boldsymbol{X}|\boldsymbol{S}|\boldsymbol{Y}>_{G_0} \Rightarrow I(\boldsymbol{X}, \boldsymbol{S}, \boldsymbol{Y})$) using descending induction:

   i. Let $n = |\boldsymbol{U}|$. For $|\boldsymbol{S}| = n-2$ the I-mapness of G_0 is guaranteed by its method of construction, Eq. (3.11).

   ii. Assume the theorem holds for every $\boldsymbol{S}'$ with size $|\boldsymbol{S}'| = k \leq n-2$, and let $\boldsymbol{S}$ be any set such that $|\boldsymbol{S}| = k-1$ and $<\boldsymbol{X}|\boldsymbol{S}|\boldsymbol{Y}>_{G_0}$. We distinguish two subcases: $\boldsymbol{X} \cup \boldsymbol{S} \cup \boldsymbol{Y} = \boldsymbol{U}$ and $\boldsymbol{X} \cup \boldsymbol{S} \cup \boldsymbol{Y} \neq \boldsymbol{U}$.

   iii. If $\boldsymbol{X} \cup \boldsymbol{S} \cup \boldsymbol{Y} = \boldsymbol{U}$ then either $|\boldsymbol{X}| \geq 2$ or $|\boldsymbol{Y}| \geq 2$. Assume, without loss of generality, that $|\boldsymbol{Y}| \geq 2$, i.e. $\boldsymbol{Y} = \boldsymbol{Y}' \cup \gamma$. From $<\boldsymbol{X}|\boldsymbol{S}|\boldsymbol{Y}>_{G_0}$ and obvious properties of vertex separation in graphs, we conclude $<\boldsymbol{X}|\boldsymbol{S} \cup \gamma|\boldsymbol{Y}'>_{G_0}$ and $<\boldsymbol{X}|\boldsymbol{S} \cup \boldsymbol{Y}'|\gamma>_{G_0}$. The two separating sets, $\boldsymbol{S} \cup \gamma$ and $\boldsymbol{S} \cup \boldsymbol{Y}$, are at least $|\boldsymbol{S}| + 1 = k$ in size; therefore, by induction on the hypothesis,

   $$I(\boldsymbol{X}, \boldsymbol{S} \cup \gamma, \boldsymbol{Y}') \;\&\; I(\boldsymbol{X}, \boldsymbol{S} \cup \boldsymbol{Y}', \gamma).$$

   Applying the intersection property (Eq. (3.6*e*)) yields the desired result: $I(\boldsymbol{X}, \boldsymbol{S}, \boldsymbol{Y})$.

iv. If $X \cup S \cup Y \neq U$, then there exists at least one element δ that is not in $X \cup S \cup Y$, and for any such δ two obvious properties of graph separation hold:

$$<X|S \cup \delta|Y>_{G_0}$$

and

$$\text{either } <X|S \cup Y|\ \delta>_{G_0} \quad \text{or } <\delta|S \cup X|Y>_{G_0} \quad \text{or both.}$$

The separating sets above are at least $|S| + 1 = k$ in size; therefore, by induction on the hypothesis,

$$I(X, S \cup \delta, Y) \quad \& \quad I(X, S \cup Y, \delta)$$

or

$$I(X, S \cup \delta, Y) \quad \& \quad I(\delta, S \cup X, Y).$$

Applying the intersection property (Eq. (3.6*e*)) to either case yields $I(X, S, Y)$, which establishes the I-mapness of G_0.

2. Next we show that G_0 is edge-minimal and unique, i.e., that no edge can be deleted from G_0 without destroying its I-mapness. Indeed, deleting an edge $(\alpha,\beta) \in E_0$ leaves α separated from β by the complementary set $U-\alpha-\beta$, and if the resulting graph is still an I-map, we can conclude $I(\alpha, U-\alpha-\beta, \beta)$. However, from the method of constructing G_0 and from $(\alpha, \beta) \in E_0$ we know that $(\alpha, U-\alpha-\beta, \beta)$ is not in I. Thus, no edge can be removed from G_0, and its minimality and uniqueness are established. Q.E.D.

Note that the weak union property (Eq. (3.6*c*)) is not needed for the proof.

Appendix 3-B

Proof of Theorem 4

THEOREM 4: [Pearl and Paz 1985]: *Every element* $\alpha \in \boldsymbol{U}$ *in a dependency model satisfying symmetry, decomposition, intersection, and weak union (Eq. (3.6)) has a unique Markov boundary* $\boldsymbol{B}_I(\alpha)$. *Moreover,* $\boldsymbol{B}_I(\alpha)$ *coincides with the set of vertices* $\boldsymbol{B}_{G_0}(\alpha)$ *adjacent to* α *in the minimal I-map* G_0.

Proof:

i. Let $\boldsymbol{BL}^*(\alpha)$ stand for the set of all Markov blankets satisfying Eq. (3.12). $\boldsymbol{B}_I(\alpha)$ is unique because the intersection property (Eq. (3.6*e*)) renders $\boldsymbol{BL}^*(\alpha)$ closed under intersection. Moreover, $\boldsymbol{B}_I(\alpha)$ equals the intersection of all members of $\boldsymbol{BL}^*(\alpha)$.

ii. Conversely, every Markov blanket $\boldsymbol{BL} \in \boldsymbol{BL}^*(\alpha)$ remains in $\boldsymbol{BL}^*(\alpha)$ after we add to it an arbitrary set of elements S' not containing α. This follows from the weak union property (Eq. (3.6*c*)). In particular, if there is an element β outside $\boldsymbol{B}_I(\alpha) \cup \alpha$ then $\boldsymbol{U}-\alpha-\beta$ is in $\boldsymbol{BL}^*(\alpha)$.

iii. From (ii) we conclude that for every element $\beta \neq \alpha$ outside $\boldsymbol{B}_I(\alpha)$, we have $I(\alpha, \boldsymbol{U}-\alpha-\beta, \beta)$, meaning β cannot be connected to α in G_0. Thus,

$$\boldsymbol{B}_{G_0}(\alpha) \subseteq \boldsymbol{B}_I(\alpha).$$

iv. To prove that $\boldsymbol{B}_{G_0}(\alpha)$ actually coincides with $\boldsymbol{B}_I(\alpha)$ it is sufficient to show that $\boldsymbol{B}_{G_0}(\alpha)$ is in $\boldsymbol{BL}^*(\alpha)$, but this follows from the fact that G_0, as an I-map, must satisfy Eq. (3.12). Q.E.D.