MARKOV AND BAYES NETWORKS: A COMPARISON OF TWO GRAPHICAL REPRESENTATIONS OF PROBABILISTIC KNOWLEDGE

Judea Pearl

October 1986
CSD-860024

TECHNICAL REPORT
CSD 860024
R-46-I
October 1986

MARKOV AND BAYES NETWORKS:
a Comparison of Two Graphical Representations of Probabilistic Knowledge

Judea Pearl

Cognitive Systems Laboratory
Computer Science Department
University of California
Los Angeles, CA 90024

ABSTRACT

This paper deals with the task of configuring effective graphical representation of dependencies embedded in probabilistic models. It first uncovers the axiomatic basis for the probabilistic relation ''x is independent of y, given z,'' and offers it as a formal definition for the qualitative notion of informational dependency. Given an initial set of such independence relationships, the axioms established permit us to infer new independencies by non-numeric, logical manipulations. Using this axiomatic basis, the paper exposes those properties of probabilistic models that can be captured by graphical representations and compares the characteristics of two such representations, *Markov Networks* and *Bayes Networks*. A Markov network is an undirected graph where the links represent symmetrical probabilistic dependencies, while a Bayes network is a directed acyclic graph where the arrows represent causal influences or object-property relationships. For each of these two network types, we establish: 1) a formal semantic of the dependencies portrayed by the networks, 2) an axiomatic characterization of the class of dependencies capturable by the network, 3) a method of constructing the network from either hard data or expert judgments and 4) a summary of properties relevant to its use as a knowledge representation scheme in inference systems.

* This work was supported in part by the National Science Foundation Grant #DCR 85-01234.

1. INTRODUCTION: from Numerical to Graphical Representations

Scholarly textbooks on probability theory have created the impression that, to construct an adequate representation of probabilistic knowledge, we must start, literally, by defining a *joint distribution function* $P(x_1, ..., x_n)$ on all propositions and their combinations and that this function should serve as the sole basis for all inferred judgments. While useful for some purposes (e.g., 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. Moreover, even if we found some economical way of storing $P(x_1, ..., x_n)$ (or rules for generating it), there would still remain the problem of manipulating it to compute the probabilities of those 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 $P(x_i \mid x_j)$ from its textbook definition $P(x_i \mid x_j) = \frac{P(x_i, x_j)}{P(x_j)}$ would involve dividing two marginal probabilities, each a result of summation over an exponentially large number of variable combinations. Human performance, by contrast, exhibits a different complexity ordering, i.e., probabilistic judgments on a small number of propositions (especially 2-place 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 many propositions entails a great degree of difficulty and hesitancy. This suggests that the elementary building blocks of human knowledge are not the entries of a joint-distribution table. Rather, they are the lower-order marginal and conditional probabilities defined over small clusters of propositions.

Another problem with purely numerical representations of probabilistic information involves the issue of *psychological meaningfulness*. While capable of computing coherent probability measures for all propositional sentences, the numerical representation often leads to computation procedures in which intermediate steps are totally different from those used by a human reasoner. 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. For example, even simple tasks such as computing the impact of a piece of evidence e on a hypothesis h via

$$P(h \mid e) = \frac{P(h, e)}{P(e)} = \frac{\sum_{x_i \neq h, e} P(x_1, ..., x_n)}{\sum_{x_i \neq e} P(x_1, ..., x_n)}$$

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

However, 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 involves equality of numerical quantities, e.g., $P(h, e) = P(h) \cdot P(e)$, suggesting that, to verify whether h and e are independent, one needs to test whether the joint distribution of e and h is equal to the product of their marginals. Contrast this with the ease and conviction with which people identify independencies while, at the same time, being unwilling to provide precise numerical estimates of probabilities.

Whereas a person may show reluctance to giving a numerical estimate for the probability of being burglarized the next day or of having a nuclear war in five years' time, that person can usually state with ease whether the two events are dependent or independent, namely, whether knowing the truth of one proposition will alter the belief in the other. Likewise, people tend to judge the 3-place relationships of conditional dependency (i.e., x_i influences x_j given x_k) with clarity, conviction and consistency. For example, it is undisputed common knowledge that knowing the departure time of the last bus is relevant for assessing how long we are about to wait for the next bus. However, once we learn the current whereabouts of the next bus, the former no longer provides useful information. These common-sensical judgments are issued qualitatively with not the slightest reference to numerical probabilities and could not possibly rely on arithmetic operations with precise probabilities.

This suggests that the notions of *relevance* and *dependence* are far more basic to human reasoning than the numerical values attached to probability judgments. Consequently, if one aspires to construct common-sensical reasoning systems, it is important that the language used for representing probabilistic information should allow assertions about dependency relationships to be expressed qualitatively, directly and explicitly. Unlike the case of numerical representations, the verification of dependencies should not await lengthy numerical manipulations but be accomplished swiftly by a few primitive operations on the salient features of the representation scheme. Moreover, once asserted, these dependency relationships should remain a stable part of the representation scheme, impervious to variations in numerical inputs. For example, one should be able to assert, categorically, that the event of nuclear disaster is independent of encountering a home burglary; the system should retain and reaffirm this independence even after one changes the estimated likelihoods of these and other events in the system.

Making effective use of information about dependencies is a computational necessity, essential in any reasoning. If we have acquired a body of knowledge z and now wish to assess the truth of proposition x, it is important to know whether it would be worthwhile to consult another proposition y, which is not in z. In other words, before we examine y, we need to know if its truth value can potentially generate new information relative to x, not available from z. In the absence of such information, an inference engine would spend precious time on derivations bearing no relevance to the task at hand. Relevance information, if available, can guide and focus the derivations in such a way that only those truly needed for the target conclusion get activated. But how would relevance information be encoded in a symbolic system?

Explicit encoding is clearly impractical because the number of (x, y, z) combinations needed for reasoning tasks is astronomical. Relevance or dependencies are relationships which change dynamically as a function of the information available at any given time. Acquiring new facts may destroy existing dependencies as well as create new ones. The former change will be called *monotonic* as it narrows the scope of propositions relevant to the target conclusion, and the latter will be called *nonmonotonic* as the scope of relevant propositions widens. For example, in trying to predict whether I am going to be late for a meeting, it is normally a good idea to ask somebody on the street for the time. However, once I establish the precise time by listening to the radio, asking people for the time becomes superfluous, and their responses would be irrelevant, thus demonstrating monotonic change of dependencies. For an example of a nonmonotonic relationship, consider the following: Normally, knowing the color of X's car tells me nothing about the color of Y's, but if X were to tell me that he almost mistook Y's car for his own, a new dependency is created between the two color variables -- whatever I learn about the color of X's car will have bearing on what I believe the color of Y's car to be. What logic would facilitate these two modes of reasoning?

In probability theory, the notion of informational relevance is given precise quantitative underpinning using the device of *conditional independence*, which successfully captures our intuition about how dependencies should change with learning new facts. A variable x is said to be independent of y, given the information z, if

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

Accordingly, if x and y are marginally dependent (i.e., dependent, when z is unknown) and become conditionally independent given z, a monotonic relationship holds. Conversely, if x and y are marginally independent and become dependent upon learning the value of z, a nonmonotonic relationship between x, y and z is captured. Thus, in principle, probability theory could provide the machinery for identifying which propositions are relevant to each other with any given state of knowledge.

However, we have already argued that it is flatly unreasonable to expect people or machines to resort to numerical verification of equalities in order to extract relevance information. Human behavior suggests that such information is inferred qualitatively from the organizational structure of human memory, not from manipulating 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 of probabilistic dependencies can be derived *logically* without references to numerical quantities. This task is dealt with in Section 1, which establishes an axiomatic characterization of probabilistic dependencies and examines whether the set of axioms matches our intuitive notion of informational relevancy.

Having a logic of dependency would be useful for testing whether a set of dependencies asserted by an expert is self-consistent and would also allow us to infer new dependencies from a given initial set of such relationships. However, such logic would not, in itself, guarantee that any sequence of inferences would be psychologically meaningful, i.e., correlated with familiar mental steps taken by humans. To facilitate this latter feature, we must also make sure that most derivational steps in that logic correspond to simple local operations on structures depicting common-sensical 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 are perfectly suited for meeting our earlier requirements of explicitness, saliency and stability, i.e., the links in the graph permit us to directly and categorically express the essential dependence relationships, and the graph topology displays these relationships explicitly and preserves them, in fact, under any assignment of numerical parameters.

It is not surprising, therefore, that graphs constitute the most common metaphor for describing conceptual dependencies. Models for human memory are often portrayed in terms of associational graphs (e.g., semantic networks [Woods, 1975], constraint networks [Montanari, 1974], inference nets [Duda, Hart and Nilsson, 1976] and conceptual dependencies [Schank 1972]). Graph-related concepts are so entrenched in our language (e.g. "threads of thoughts," "lines of reasoning," "connected ideas," "far-fetched arguments" etc.) that one wonders whether people can, in fact, reason any other way except by tracing links and arrows and paths in some mental representation of concepts and relations. Therefore, a natural question to ask is whether the informal notion of informational relevancy or the more technical notion of probabilistic dependencies can be captured by graphical representation, in the sense that all dependencies and independencies in a given model would be deducible from the topological properties of some network. This question will be addressed in Sections 2 and 3.

Despite the prevailing use of graphs as metaphors for communicating and reasoning about dependencies, the task of capturing dependencies by graphs is not at all trivial. When we deal with a phenomenon where the notion of neighborhood or connectedness is explicit (e.g., family relations, electronic circuits, communication networks, etc.), we have no problem configuring a graph which represents the main features of the phenomenon. However, in modeling conceptual relations such as causation, association and relevance, it is often hard to distinguish direct neighbors from indirect neighbors; so, the task of constructing a graph representation then becomes more delicate. The notion of conditional independence in probability theory is a perfect example of such a relational structure. For a given probability distribution P and any three variables x, y, z, while it is fairly easy to verify whether knowing z renders x independent of y, P does not dictate which variables should be regarded as direct neighbors. Thus, many different topologies might be used to display the dependencies embodied in P. We shall also see that some useful properties of dependencies and relevancies cannot be represented graphically. Markov and Bayes networks represent two approaches to minimizing such deficiencies.

This paper is organized as follows: Section 1 uncovers the axiomatic basis for the probabilistic relation "x is independent of y, given z" and offers it as a formal definition for the qualitative notion of informational dependency. Given an initial set of such independence relationships, the axioms established permit us to infer new independencies by non-numeric, logical manipulations. Sections 2 and 3 examine those properties of probabilistic models that can be captured by graphical representations and compare the properties of two such representations: *Markov Networks* (Section 2) and *Bayes Networks* (Section 3). A Markov network is an undirected graph where the links represent probabilistic dependencies, while a Bayes network is a directed acyclic graph where the arrows represent causal influences or frame-slot relationships. For each of these two network types we establish:

(1) a formal semantic of the dependencies portrayed by the networks,
(2) an axiomatic characterization of the class of dependencies capturable by the network,
(3) a method of constructing the network from either hard data or expert judgments and
(4) a summary of properties relevant to its use as a knowledge representation scheme in inference systems.

2. AN AXIOMATIC BASIS FOR PROBABILISTIC DEPENDENCIES

Definition: Let $U = \{\alpha, \beta, ...\}$ be a finite set of discrete-valued variables (i.e., partitions or attributes) characterized by a joint probability function $P(\cdot)$, and let x, y and z stand for any three subsets of variables in U. x and y are said to be *conditionally independent given* z if

$$P(x, y \mid z) = P(x \mid z)\, P(y \mid z) \qquad when\, P(z) > 0 \tag{1}$$

Eq. (1) is a terse notation for the assertion that, for any instantiation z_k of the variables in z and for any instantiations x_i and y_j of x and y, we have

$$P(x{=}x_i \;\; and \;\; y{=}y_j \mid z{=}z_k) = P(x{=}x_i \mid z{=}z_k)\, P(y{=}y_j \mid z{=}z_k) \tag{2}$$

The requirement $P(z) > 0$ guarantees that all the conditional probabilities are well defined, and we shall henceforth assume that $P > 0$ for any instantiation of the variables in U. This rules out logical and functional dependencies among the variables, a case which would require special treatment.

We use the notation $I(x,z,y)_P$ or simply $I(x,z,y)$ to denote the independence of x and y given z; thus,

$$I(x, z, y)_P \;\; iff \;\; P(x, y \mid z) = P(x \mid z)\, P(y \mid z) \tag{3}$$

Unconditional independence (also called *marginal* independence) will be denoted by $I(x, \emptyset, y)$, i.e.,

$$I(x, \emptyset, y)_P \;\; iff \;\; P(x, y) = P(x)\, P(y)$$

Note that $I(x, z, y)$ implies the conditional independence of all pairs of variables $\alpha \in x$ and $\beta \in y$, but the converse is not necessarily true.

The conditional-independence relation $I(x,z,y)$ satisfies the following set of properties [Lauritzen, 1982]:

$$I(x, z, y) \Longleftrightarrow P(x \mid y, z) = P(x \mid z) \tag{4.a}$$

$$I(x, z, y) \Longleftrightarrow P(x, z \mid y) = P(x \mid z)\, P(z \mid y) \tag{4.b}$$

$$I(x, z, y) \Longleftrightarrow \exists f, g : P(x, y, z) = f(x, z) g(y, z) \tag{4.c}$$

$$I(x, z, y) \Longleftrightarrow P(x, y, z) = P(x \mid z)\, P(y, z) \tag{4.d}$$

$$I(x, z, y) \Longrightarrow I(x, (z, f(y)), y) \tag{5.a}$$

$$I(x, z, y) \Longrightarrow I(f(x, z), z, y) \tag{5.b}$$

The proof of these properties can be derived by elementary means from the definition (3) and the basic axioms of probability theory. These properties are based on the numeric representation of P and, therefore, would not be adequate as an axiomatic system.

We now ask what logical conditions, void of any reference to numerical forms, should constrain the relationship $I(x, z, y)$ if it stands for the statement "x is independent of y, given that we know z" in some probability model P. The next theorem establishes such a logical basis:

Theorem 1: Let x, y and z be three disjoint subsets of variables from U, and let $I(x, z, y)$ stand for the relation "x is independent of y, given z" in some probabilistic model P, then I must satisfy the following set of five independent conditions:

Symmetry (6.a)

$$I(x,z,y) \Longleftrightarrow I(y,z,x)$$

Decomposition (6.b)

$$I(x,z,y \cup w) \Rightarrow I(x,z,y) \;\&\; I(x,z,w)$$

Intersection (6.c)

$$I(x,z \cup w,y) \;\&\; I(x,z \cup y,w) \Rightarrow I(x,z,y \cup w)$$

Weak Union (6.d)

$$I(x,z,y \cup w) \Rightarrow I(x,z \cup w,y)$$

Contraction (6.e)

$$I(x,z \cup y,\ w) \,\&\, I(x,z,y) \Rightarrow I(x,z,y \cup w)$$

Remarks: The symbol $\cup$ in $y \cup w$ should not be confused with logical disjunction. Rather, 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

$$P(x=x_1 \;\&\; y=y_j \;\&\; w=w_k) = P(x=x_i)\,P(y=y_j \;\&\; w=w_k) \quad \forall i,j,k$$

When convenience prevails, an alternative, simpler notation, $I(x,\varnothing,yw)$, will be used.

Restricting the arguments of $I(\cdot)$ to disjoint subsets does not affect the generality of Theorem 1. Once I is defined on the set of disjoint triplets x, y, z it is also defined on the set of all triplets. This is seen from Eq.(5.b). Which, by proper choice of f, implies

$$I(x,z,y) \Longleftrightarrow I(x-z,z,y)$$

For technical convenience we shall also adopt the convention that every variable is independent of the null set, i.e., $I(x,z,\varnothing)$.

The intuitive interpretation of Eqs. (6.c) through (6.e) follows. (6.c) states that, if y does not affect x when w is held constant and if, simultaneously, w does not affect x when y is held constant, then neither w nor y can affect x. (6.d) states that learning an irrelevant fact (w) cannot help another irrelevant fact (y) become relevant to x. (6.e) can be interpreted to state that, if we judge w to be irrelevant (to x) after learning some irrelevant facts y, then w must have been irrelevant before learning y. Together, the weak union and contraction properties mean that learning irrelevant facts should not alter the relevance status of other propositions in the system; whatever was relevant remains relevant, and what was irrelevant remains irrelevant.

The operational significance of axioms (6.a)-(6.e) and their role as inference rules can best be explained by employing a graph metaphor, letting $I(x,z,y)$ stand for the phrase "z separates x from y" or, in other words, "the removal of a set z of nodes from the graph would render the nodes in x disconnected from those in y." The validity of (6.a) through (6.e) if clearly depicted by the chain $x-z-y-w$ and in the schematics of Appendix I.

Symmetry (6.a) simply states that if z separates x from y then it also separates y from x. The *decomposition* axiom (6.b) asserts that if z separates x from the compound set $S = y \cup w$ then it also separates x from every subset of S. The *intersection* axiom (6.c) states that if within some set of variables $S = x \cup y \cup z \cup w$, x is separated from the rest of S by two different subsets, S_1 and S_2, (i.e., $S_1 = z \cup y$ and $S_2 = z \cup w$) then the intersection of S_1 and S_2 would also separate x from the rest of S.

The *weak union* axiom (6.d) provides the conditions under which a separating set z can be augmented by additional elements (w) and still separate x from y. The condition is that the added subset w, must come from that part of the space which was initially separated from x by z. The *contraction* axiom (6.e) provides conditions for reducing the size of the separating set; it permits the deletion of a subset (y) from the separator ($z \cup y$) if the remaining part, z, separates the deleted part y from x. Figure 0 provides schematic descriptions of these rules.

The proof of Theorem 1 can be derived by elementary means from the definition (3) and from the basic axioms of probability theory. The proof that Eqs. (6.a) through (6.e) are logically independent can be derived by letting U contain four elements and showing that it is always possible to contrive a subset I of triplets (from the subsets of U), which violates one property and satisfies the other four.

The intersection property is the only one which requires the assumption $P(x) > 0$ and will not hold when the variables in 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^oF," then, clearly, knowing the truth of either one of them renders the other superfluous. Yet, contrary to (6.c), this should not render both y and w irrelevant to a third proposition x, say, whether we will enjoy swimming in that water. In such a case, Theorem 1 will still retain its validity if we regard each logical constraint as having some small probability ε of being violated, and let $\varepsilon \rightarrow 0$.

The assumption $P(E) \geq \varepsilon \geq 0$ amounts to stating that every event or combination of event, no matter how outrageous, has some chance of being true. As strange as it sounds, this is not an unreasonable assumption to make while talking about empirical facts. For example, it is not completely impossible for the water temperature to be above freezing and below 32° F (e.g., if it is very salty) and, now, that we accept such a possibility we must also denounce the statement that knowing any one of these two facts renders the other superfluous relative to any x. If x represents our concern about swimming in that water then the temperature becomes the relevant fact, rendering its freezing status irrelevant. If, on the other hand, our interest lies in ice formation, it is the freezing point, not the temperature, that is relevant. This is exactly what axiom (6.c) claims; if two properties exert influence on x, then (at a sufficiently fine level of detail) it is impossible that either one of them, interchangeably, would render the other irrelevant. Symmetrical exclusion is only possible when we are dealing with definitional properties (e.g., y: "The water temperature is above 32^oF; w: "The water temperature is not equal or lower than 32^oF,") but not with properties subject to independent empirical tests.

Despite their striking similarity to vertex separation in graphs, properties (4.b) and (4.d) are much weaker than their graph counterparts. In graphs, two sets of vertices are said to be separated if there exists no path between their individual elements. The composition property (4.b), on the other hand, contains only one-way implication; a variable α may be independent of each and every individual variable in set y and still be dependent on the entire set. For example, let y be the outcomes of a set of fair coins, and let α be a variable that attains the value 1 whenever an even number of coins turn up heads and the value 0 otherwise. α is statistically independent on every element as well as any proper subset of y; yet, α is completely determined by the entire set y.

Property (6.d) is also weaker than its corresponding property in graphs. If z is a cutset of vertices which separates x from y in some graph, then augmenting z by additional elements always keeps x and y separated. (6.d), on the other hand, severely restricts the conditions under which a separating set z can be enlarged by additional elements w -- w must be chosen from a set which, together with y is already separated from x by z.

Completeness Conjecture: The set of axioms (6.a) through (6.e) is *complete* when I is interpreted as a conditional-independence relation. In other words, for every 3-place relation I satisfying (6.a) through (6.e), there exists a probability model P such that

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

Although we have not been able to establish a general proof of completeness, we were not able to find any general property of conditional independence, valid for all P, which is not implied by (6.a) through (6.e).

The usefulness of axiomatizing the notion of probabilistic dependence is three-fold. First, it allows us to conjecture and derive interesting and powerful theorems which may or may not be obvious in the numerical representation of probabilities. For example, the chaining rule [Lauritzen, 1982]

$$I(x, y, z) \;\&\; I(x \cup y, z, w) \Rightarrow I(x, y, w)$$

follows directly from Eqs. (6.d) and (6.e) and is important for recursively constructing directed graph representations (See section 4.). Another interesting theorem is the ''mixing rule'' [Dalkey, 1986]

$$I(x, z, yw) \;\&\; I(y, z, w) \;==>\; I(xw, z, y)$$

which also follows from Eqs.(6.d) and (6.e).

Second, the set of axioms (6.a) to (6.e) can be viewed as qualitative inference rules which can be used to derive new independencies from some initial set of instances. For example, if an expert provides us with an initial set, S, of qualitative independence judgments in the form of triplets (x, z, y), we can use axioms (6.a)-(6.e) to generate the closure of S or, alternatively, to test whether a given additional triplet (x', z', y') logically follows from S. In this fashion, one can test and maintain consistency in the database as well as prevent reasoning systems from spending inordinate effort on variables proven irrelevant to the target hypotheses. Third, the axiomatic system provides a parsimonious and convenient code for comparing the features of several dependency models as well as expressive power for various representations of these models. In sections 3 and 4, for example, we will use the axiomatic characterization to compare the expressive powers of directed vs. undirected graphs and to reveal what type of dependencies are not capturable by graphical representations.

3. MARKOV NETS

What's in a missing link?

Suppose we have a collection U of interacting elements and we decide to represent their interactions by an undirected graph G in which the nodes correspond to individual elements of U. Naturally, we would like to display independence between elements by a lack of connectivity between their corresponding nodes in G and, conversely, dependent elements should correspond to connected nodes in G. This requirement alone, however, does not take full advantage of the expressive power of graph representation. It treats all connected components of G as equivalence classes and does not attribute any special significance to the topological configuration within each connected component of G.

Clearly, if graph topology is to convey meaning beyond its connectedness, a semantic distinction must be made between "direct connection" and "indirect connection" in the sense that arbitrarily adding a link between indirectly connected elements should correspond to a totally different state of dependency. This means that the absence of a direct link between two elements α and β should reflect the fact that their interaction is not basic but *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 fact that there is no direct contact between the two males or the two females can be represented by the diamond-shaped graph of Fig. 1, which may also be used to represent conditional dependencies between various propositions. For example, if by m_i (and f_i) we denote the proposition that male M_i (respectively, F_i) will carry a certain venereal disease within the next year, then the topology of the network in Fig. 1 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 to change our belief in f_2[1]. This conditional independence information reflects a model whereby the disease spreads only by direct sexual contact. Note that the links in this network are undirected, namely, either partner may be the originator of the disease. This does not exclude asymmetric interactions, e.g., if the disease is more easily transferable from males to females than the other way around. Such information, if available, will be contained in the numerical parameters which will eventually characterize the links in the network and will be described in a later section.

In summary, the semantic of the graph topology is defined by the meaning of the missing links which specify what other elements mediate the interaction between non-adjacent elements. This process of meditation will now be compared to the probabilistic relation of *conditional independence* $I(x, z, y)$, Eq. (1), which formalizes the intuitive statement: "Knowing y would tell me nothing new about x if I already know z."

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

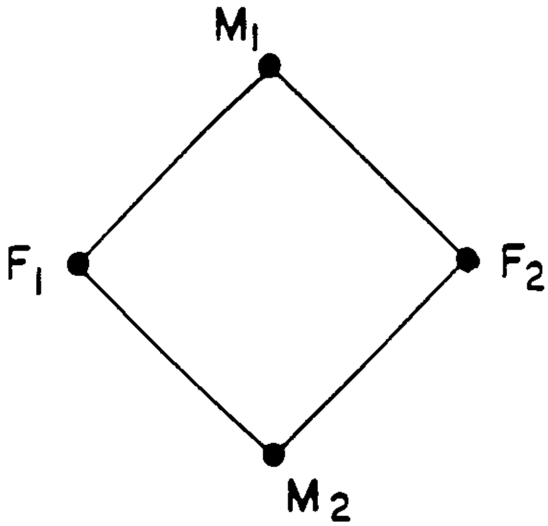


Figure 1

3.1 Graph Separation and Conditional Independence

Let $U = \{\alpha, \beta, ...\}$ be a finite set of elements (e.g., propositions, variables, etc.), and let x, y and z stand for three disjoint subsets of elements in U. Let M be a *dependency model* which assigns truth values to the 3-place predicate $I(x, z, y)$ or, in other words, M determines a subset I of triplets (x, z, y) for which the assertion "x is independent of y given z" is true. Any probability distribution P constitutes such a model because, for any triplet (x, z, y), we can test the validity of $I(x, z, y)$ using Eq. (1). Our task is to characterize the set of models capturable by graphs, assuming that the model does not provide direct tests for "adjacency". In other words, we are given the means to test whether any given subset S of elements *intervenes in a relation between* elements x and y, but it remains up to us to decide how to connect the elements together in a graph that encodes these interventions.

Ideally, we would like to require that if the removal of some subset S of nodes from the graph G renders nodes x and y disconnected (written $<x \mid S \mid y>_G$), then this separation should correspond to conditional independence between x and y given S, namely,

$$<x \mid S \mid y>_G \Rightarrow I(x, S, y)$$

and, conversely,

$$I(x, S, y) \Rightarrow <x \mid S \mid y>_G$$

This would provide a clear graphical representation for the notion that x does not affect y directly, that its influence is mediated by the variables in S. Unfortunately, we shall next see that these two requirements are too strong; there is often no way of using vertex separation in a graph to display *all* dependencies and independencies embodied in some probabilistic models, even those portraying simple, everyday experiences.

Definition: An undirected graph G is a *dependency map* (D-map) of M if there is a one-to-one correspondence between the elements of U and the nodes of G, such that for all disjoint subsets, x, y, z, of elements we have:

$$I\ (x, z, y)_M \Rightarrow \langle x \mid z \mid y \rangle_G \qquad (7)$$

Similarly, G is an *Independency map* (I-map) of M if:

$$I\ (x, z, y)_M \Leftarrow \langle x \mid z \mid y \rangle_G \qquad (8)$$

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

A D-map guarantees that vertices found to be connected are, indeed, dependent; however, it may occasionally display dependent variables as separated vertices. An I-map works the opposite way: it guarantees that vertices found to be separated always correspond to genuinely independent variables but does not guarantee that all those shown to be connected are, in fact, dependent. Empty graphs are trivial D-maps, while complete graphs are trivial I-maps.

It is not hard to see that many reasonable models of dependency have no perfect maps. This occurs, for example, in models where $I(x, z, y)$ exhibits *nonmonotonic* behavior; totally unrelated propositions become relevant to each other upon learning new facts. A nonmonotonic model M, implying both $I(x, z_1, y)_M$ and $\neg I(x, z_1 \cup z_2, y)_M$ cannot have a graph representation which is both an I-map and a D-map, because graph separation always satisfies $\langle x \mid z_1 \mid y \rangle_G \Rightarrow \langle x \mid z_1 \cup z_2 \mid y \rangle_G$ for any two subsets z_1 and z_2 of vertices. Thus, D-mapness forces G to display z_1 as a cutset separating x and y, while I-mapness prevents $z_1 \cup z_2$ from separating x and y. No graph can satisfy these two requirements simultaneously.

This weakness in the expressive power of undirected graphs severely limits their ability to represent probabilistic dependencies. A simple example illustrating this point is 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., $I(x, \emptyset, y)$. However, if we notice the bell (z), then learning the outcome of one coin should change our opinion about the other coin, namely, $\neg I(x, z, y)$.

How can we graphically represent these simple dependencies between the coins and the bell or, in general, between a set of multiple 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 asserts that x and y are independent given z, which is wrong. If we add a link between x and y as well, we get the trivial I-map of a complete graph, which no longer alerts us to the obvious fact that the two coins are genuinely independent since the bell is merely a passive device which does not affect their interaction. In Section 4, we will show that such dependencies can be represented completely by using the richer language of directed graph. In this section, however, we will continue to examine the representational capabilities of undirected graphs.

Being unable to provide graphical representations to some (e.g., nonmonotonic) models of dependency, raises the question of whether we can formally delineate the class of models which *do* lend themselves to graphical representation. This is accomplished in the following subsection by establishing an axiomatic characterization of the family of relations which are isomorphic to vertex separation in graphs.

3.2 Axiomatic Characterization of Graph-Isomorph Dependencies

Definition: A dependency model M is said to be a *graph-isomorph* if there exists a graph $G = (U, E)$ which is a perfect map of M, i.e., for every three disjoint subsets x, y and z of U, we have:

$$I(x, z, y)_M <\!\!==> < x \mid z \mid y >_G \tag{9}$$

Theorem 2: [Pearl & Paz, 1985] A necessary and sufficient condition for a dependency model M to be graph-isomorph is that $I(x, z, y)_M$ satisfies the following five independent axioms (the subscript M dropped for clarity):

(symmetry)

$$I(x, z, y) <\!\!==> I(y, z, x) \tag{10.a}$$

(decomposition)

$$I(x, z, y \cup w) \Rightarrow I(x, z, y) \;\&\; I(x, z, w) \tag{10.b}$$

(intersection)

$$I(x, z \cup w, y) \;\&\; I(x, z \cup y, w) \Rightarrow I(x, z, y \cup w) \tag{10.c}$$

(strong union)

$$I(x, z, y) \Rightarrow I(x, z \cup w, y) \qquad \forall\; w \subseteq U - x \cup z \cup y \tag{10.d}$$

(transitivity)

$$I(x, z, y) \Rightarrow I(x, z, \gamma) \; or \; I(\gamma, z, y) \qquad \forall\; \gamma \notin x \cup z \cup y \tag{10.e}$$

Remark-1: The axioms in (10) are clearly satisfied for vertex separation in graphs. (10.e) is the counterpositive form of connectedness transitivity, stating that, if x is connected to γ and γ is connected to y, then x must also be connected to y. (10.d) states that, if z is a vertex cutset separating x from y, then removing additional vertices w from the graph still leaves x and y separated. (10.c) claims that, if x is separated from w with y removed and, simultaneously, x is separated from y with w removed, then x must be separated from both y and w.

Remark-2: (10.c) and (10.d) imply the converse of (10.b), which makes I completely defined by the set of triplets (x, z, y) in which x and y are individual elements of U. Equivalently, we an express the axioms in (10) in terms of such triplets. Note, also, that the union axiom (10.d) is unconditional and, therefore, stronger than (10.d), the one required for probabilistic dependencies. It provides a simple method of constructing the unique graph G_0, which is a perfect map of M -- starting with a complete graph, we simply delete every edge (α, β) for which a triplet of the form (α, z, β) appears in I.

Proof:

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

2. To prove sufficiency, we need to show that, for any set I of triplets (x, z, y) satisfying (10.a) through (10.e), there exists a graph G such that x, z, y is in I *iff* Z is a cutset G that separates

x from y. We show that $G_0 = (U, E_0)$ is such a graph, where $(\alpha, \beta) \notin E_0$ *iff* $I(\alpha, z, \beta)$. In view of remark-2 above, it is sufficient to show that

$$I(\alpha, S, \beta) \Rightarrow \ <\alpha | S | \beta>_{G_0} \qquad \alpha, \beta \in U, S, \subseteq U$$

This is proved by finite induction:

i. For $|S| = n-2$, the theorem holds automatically, due to the way G_0 is constructed.

ii. Assume the theorem holds for and S with size $|S| = k =< n-2$. Let S' be any set of size $|S'| = k-1$. For $k =< n-2$, there exists an element γ outside $S' \cup \alpha \cup \beta$ and, using (10.d), we have: $I(\alpha, S', \beta) \Rightarrow I(\alpha, S' \cup \gamma, \beta)$.

iii. By (10.e) we have either $I(\alpha, S', \gamma)$ or $I(\gamma, S', \beta)$.

iv. Choosing the first alternative in (iv) (the latter giving an identical result), and applying (10.d), gives $I(\alpha, S' \cup \beta, \gamma)$.

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

vi. By the intersection property (10.c) for vertex-separation in graphs, these two assertions imply $<\alpha | S' | \beta>_{G_0}$. *Q.E.D.*

Having a complete characterization for vertex separation in graphs makes it easy 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 each of the last two axioms. Axiom (10.d) is clearly violated in the non-monotonic coins-and-bell example of the preceding subsection. Transitivity (10.e) is violated by that same example because, if one of the coins is not fair, then the bell's response is dependent on the outcome of each coin separately; yet, the two coins are independent of each other. Finally, (10.c) is violated in contexts where the propositions y and w logically constrain one another, as in the earlier example of the water temperature.

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 which faithfully display each and every dependency. However, acknowledging the fact that some independencies will escape representation, we shall insist that their number be kept at a minimum or, in other words, that the graphs in those maps contain no superfluous edges.

3.3 Markov Net: the Minimal *I*-Map of P

Whenever a correspondence is defined between such seemingly unrelated objects as probability distributions and graphs, it it natural to raise the following three questions:

1. Given a graph G, can we construct a probability distribution P such that G is a perfect map of P?

2. Given a pair (P, G), can we test if G is an I-map of P?

3. Given a probability distribution P, can we construct an I-map G of P which has the minimum number of edges?

The first two problems have been given satisfactory answers by the theory of Markov Fields [Isham, 1981], [Lauritzen, 1982] and [Geman & Geman, 1984]. This treatment is rather complex and relies heavily on the numerical representation of probabilities. We shall focus on the third problem and show that:

- Problem 3 has a simple unique solution.
- The solution to 2 follows directly from the solution to Problem 3.
- The solutions are obtained by non-numeric analysis,
  based solely on axioms (6.a) through (6.d) of Section 2.

Question 1 will be treated in Section 4.

3.1 Basic Definitions and Properties

Definition: A graph G is a *minimal* I-map of dependency model M if no edge of G can be deleted without destroying its I-mapness. We call such a graph a *Markov-Net* of M.

Theorem 3: [Pearl & Paz, 1985]. Every dependency model M satisfying (6.a)-(6.c) has a (unique) minimal I-map $G_0 = (U, E_0)$ produced by connecting *only* pairs (α, β) for which $I(\alpha, U - \alpha - \beta, \beta)_M$ is *FALSE, i.e.,*

$$(\alpha, \beta) \notin E_0 \quad \textit{iff} \quad I(\alpha, U - \alpha - \beta, \beta)_M \tag{11}$$

The proof is given in Section 5.

Definition: A *relevance blanket* $R_I(\alpha)$ of a variable $\alpha \in U$ is any subset S of variables for which

$$I(\alpha, S, U - S - \alpha) \text{ and } \alpha \notin S \tag{12}$$

Let $R_I^*(\alpha)$ stand for the set of all relevance blankets of α. A set is called a *relevance boundary* of α, denoted $B_I(\alpha)$, if it is in $R_I^*(\alpha)$ and if, in addition, none of its proper subsets are in $R_I^*(\alpha)$.

$B_I(\alpha)$ is to be interpreted as the smallest set of variables that "shields" α from the influence of all other variables. Note that $R_I(\alpha)$ is non-empty because $I(x, z, \emptyset)$ guarantees that the set $S = U - \alpha$ satisfies (12).

Theorem 4: [Pearl & Paz 1985]. Every variable $\alpha \in U$ in a probabilistic model P has a unique relevance boundary $B_I(\alpha)$ called the *Markov boundary* of α. $B_I(\alpha)$ coincides with the set of vertices $B_{G_0}(\alpha)$ adjacent to α in the Markov net G_0.

The proof of Theorem 3 (See Section 5.) also makes use of the weak-union property (6.d).

Corollary 1: The set of Markov boundaries $B_I(\alpha)$ forms a *neighbor system,* i.e., a collection $B_I^* = \{B_I(\alpha) : \alpha \in U\}$ of subsets of U such that

(i) $\alpha \notin B_I(\alpha)$, and

(ii) $\alpha \in B_I(\beta)$ iff $\beta \in B_I(\alpha)$, $\alpha, \beta \in U$

Corollary 2: The Markov net G_0 can be constructed by connecting each α to all members of its Markov boundary $B_I(\alpha)$.

The usefulness of this corollary lies in the fact that, in many cases, it is the Markov boundaries $B_I(\alpha)$ that define the organizational structure of human memory. People find it natural to identify the immediate consequences and/or justifications of each action or event [Doyle, 1979], and these relationships constitute the neighborhood semantics for inference nets used in expert systems [Duda et al., 1976]. The fact that $B_I(\alpha)$ coincides with $B_{G_0}(\alpha)$ guarantees that many global independence relationships can be validated by separation tests on graphs constructed from local information.

We are now in a position to answer the I-map recognition question mentioned at the beginning of this subsection (question 2), i.e., can we test whether a given graph G is an I-map of a distribution P.

Theorem 5: Given a probability distribution P on U and a graph $G = (U, E)$, the following three conditions are equivalent:

i. G is an I-map of P

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

$$(\alpha, \beta) \notin E \qquad \textit{only if} \qquad I(\alpha, U - \alpha - \beta, \beta)$$

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

Proof: The implication (ii) => (i) follows from the I-mapness of G_0. (i) => (iii) follows from the definition of I-mapness. It remains to show (iii) => (ii), but this follows from the uniqueness and minimality of G_0 (Theorem 3). Q.E.D.

Properties (ii) and (iii) provide procedures for testing I-mapness without exhaustively examining every cutset in G. They still require, though, tests which involve all the variables in U and, therefore, may lead to exponential complexity, especially when P is given as a table. Fortunately, in most practical applications, it is the graph representation G that we start with; the probability model P is used merely as a theoretical abstraction used to justify the operations conducted on G.

Thus we see that the major graphical properties of probabilistic independencies are consequences of the intersection and union axioms (6.c) and (6.d). Axioms (6.a) through (6.d) were chosen, therefore, as the definition of a general class of dependency models called *Graphoids* [Pearl, Paz 1985], which possess graphical representations similar to those of Markov nets (See Section 5.). The contraction axiom (6.e) is necessary for constructing directed-graph representations (to be treated in Section 4).

3.2 Illustration 1 (abstract)

To illustrate the role of the conditional-independence axioms (Eq.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 (6.a) through (6.d) but not (6.e), because it does not contain (1, 2, 4). Thus, (from Theorem 1) there is no probability model supporting I; yet, from Theorem 3, it has a unique minimal I-map G_0, shown in Figure 2.

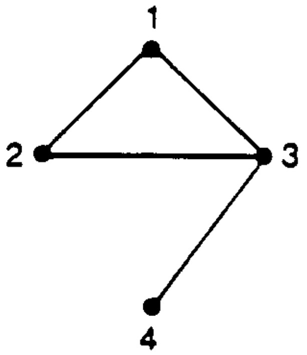


Figure 2: The Minimal I-Map, G_0, of I

This graph can be constructed either by deleting the edges (1, 4) and (2, 4) from the complete graph or by computing, from I, the relevance boundary of each element, i.e.,

$$B_I(1) = \{2, 3\},\ B_I(2) = \{1, 3\},\ B_I(3) = \{1, 2, 4\},\ B_I(4) = \{3\}.$$

Suppose that the list contained only the last two triplets (and their symmetrical images):

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

I' is clearly not a probabilistic independence relation because the absence of the triplets (1, 3, 4) and (2, 3, 4) violates the intersection axiom (6.c). Indeed, if we try to construct G_0 by the usual criterion of edge deletion, the graph in Figure 2 ensues, but it is no longer an I-map of I'; it shows 3 separating 1 from 4, while (1, 3, 4) is not in I'. In fact, the only I-maps of I' are the three graphs in Figure 3, and the edge-minimum graph is clearly not unique.

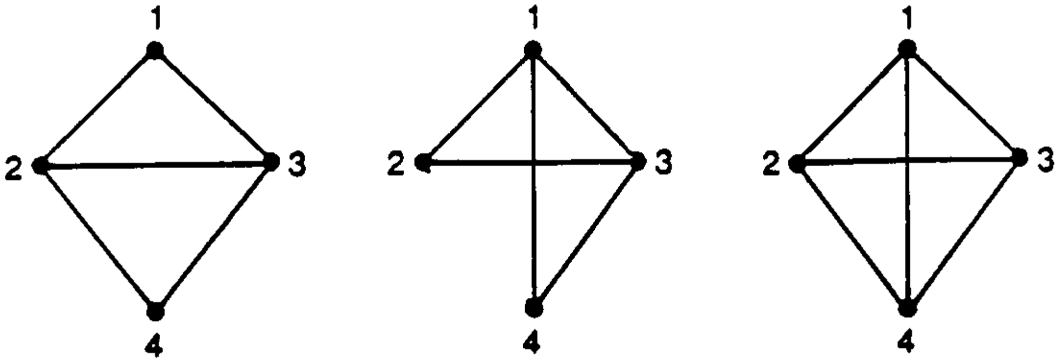


Figure 3: The Three I-Maps of I′

Now consider the list

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

I'' satisfies the first three axioms, (6.a) through (6.c) but not the union axiom (6.d). Since no triplet of the form (α, $U - \alpha - \beta$, β) appears in I'', the only I-map for this list is the complete graph. Moreover, the relevance boundaries of I'' do not form a neighbor set; e.g., $B_{I''}(4) = 3$, $B_{I''}(2) = \{1, 3, 4\}$; so, $2 \notin B_{I''}(4)$ while $4 \in B_{I''}(2)$.

Note that I does not possess the contraction property (6.e); therefore, there is no probabilistic model capable of inducing this set of independence relationships unless we also add the triplet (1, 2, 4) to I. Had I been a list of inputs given by a domain expert, it would be simple to invoke axioms (6.a) through (6.e) to alert the expert to inconsistency in the data, pointing to the absence of (1, 2, 4). However, the discrepancies in I' and I'' would be easier to detect because they interfere with the formation of G_0 and so could be identified by a system which attempts to construct it.

3.3 Illustration 2 (application)

Consider the task of constructing a Markov net to represent the belief whether or not an agent A is about to be late for a meeting. Assume that 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 current time;
4. the time A will show up at the meeting place;
5. the time A 's partner plans to show up;
6. the time A 's partner will actually show up;
7. whether A will be late for the meeting (i.e., will arrive after his partner).

The construction of G_0 can proceed by two methods:

1. the *complementary set* method; and
2. the *relevance-boundary* method.

The first method requires that, for every pair of variables (α, β), we determine whether fixing the value of all other variables in the system will render our belief in α sensitive to the value of 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 held constant. On that basis, we may connect node 1 to node 3 and, proceeding in that fashion through all pairs of variables, the graph of Figure 4 may be constructed.

The relevance-boundary method is more direct; for every variable α in the system we identify the minimal set of variables sufficient to render the belief in α insensitive to all other variables in the system. It is a common-sense task, for instance, to decide that, once we know the current time (3), no other variable may affect what we expect to read on passerby-1's watch (1). Similarly, to estimate our arrival time (4), it is sufficient that we know the current time (3), whether we are determined to be late (7) and when our partner will actually show up (6), independent of our partner's intentions (5). On the basis of these

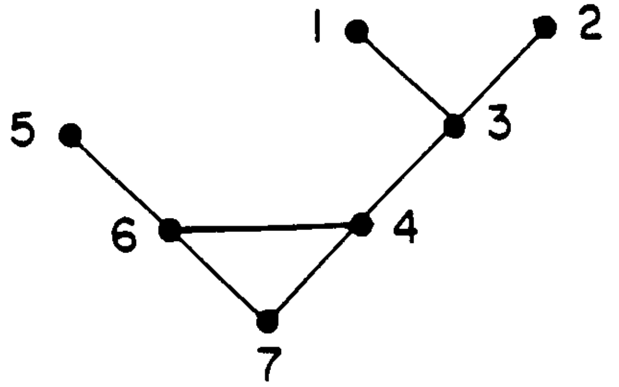


Figure 4

considerations, we may connect 1 to 3; 4 to 6, 7 to 3; and so on. After finding the immediate neighbors of any six variables in the system, the graph G_0 will emerge, identical to that of Figure 4.

Once established, G_0 can be used as an inference instrument. For example, the fact that knowing A 's arrival time (4) renders the time on passerby-1's watch (1) irrelevant for deciding whether A will be late (7) (i.e., $I(1,4,7)$) need not be stated explicitly; it can be inferred from the fact that 4 is a cutset in G_0, separating 1 from 7. Deriving this conclusion by syntactic manipulations of axioms (6.a) through (6.e) would probably be more complicated. Additionally, the graphical representation can be used to help maintain consistency and completeness during the knowledge-building phase. One need ascertain only that the relevance boundaries identified by the knowledge provider (e.g., the expert) form a neighbor system.

3.4 Summary

We have shown that the essential qualities characterizing the probabilistic notion of conditional independence are captured by five logical axioms: symmetry (6.a), decomposition (6.b), intersection (6.c), weak union (6.d) and contraction (6.e). The first three axioms enable us to construct an edge-minimum graph in which every cutset corresponds to a genuine independence condition. The fourth axiom is needed to guarantee that the set of neighbors which G_0 assigns to each variable α is actually the smallest set required to shield α from the effects of all other variables.

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

The probabilistic relation of conditional independence is shown to possess a rather plausible set of qualitative properties, consistent with our intuitive notion of "x being irrelevant to y, once we learn z." Reducing these properties to a set of logical axioms permits us to test whether other calculi of uncertainty also yield facilities for connecting relevance to knowledge. Moreover, the axioms established can be viewed as inference rules for deriving new independencies from some initial set.

Not all properties of probabilistic dependence can be captured by undirected graphs. For example, the former is non-monotonic and non-transitive (see 'coins and bell' example after proof of lemma), while graph separation is both monotonic and transitive. It is for these reasons that directed graphs such as *inference nets* [Duda et al., 1976], *influence diagrams* [Howard & Matheson 1984] and *Bayesian belief nets* [Pearl, 1986] are finding a wider application in reasoning systems. A systematic treatment of these graphical representations is given in Section 4.

4. MARKOV NET AS A KNOWLEDGE BASE

4.1 Quantifying the Links

So far, we have established the semantics of Markov networks in terms of the purely qualitative relationships, that is, 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 the links. In Figure 1, for example, we may know for a fact that the couple (M_1, F_2) meet less frequently than the couple (M_1, F_1); so, the former link should be weaker than the latter, implying weaker dependency between the propositions m_1 and f_2.

The task of assigning 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 first attend to two problems: *consistency* and *completeness*. Consistency guarantees that we do not overload the graph with an excessive number of parameters; overspecification may lead to contradictory conclusions, depending on which parameter is consulted first. Completeness protects us from underspecifying the graph dependencies and guarantees that our conclusion-generating routine will not get deadlocked for lack of information.

One of the attractive features of the traditional joint-distribution representation of probabilities is the transparency by which one can synthesize consistent probability models or detect inconsistencies therein. In this representation, to create a complete model, free of inconsistencies, one need only assign to the atomic events in the space (i.e., conjunctions of propositions) non-negative weights summing to one. By contrast, the synthesis process in the graph representation is more hazardous. For example, assume that in Figure 1 we want to express the dependencies between the variables $\{M_1, M_2, F_1, F_2\}$ by specifying the four pairwise probabilities $P(M_1, F_1)$, $P(F_1, M_2)$, $P(M_2, F_2)$, $P(F_2, M_1)$. It turns out that this, normally, will lead to inconsistencies; unless the parameters given satisfy some non-obvious relationship, there exists no probability model that will support all four inputs. Moreover, it is not at all clear how to put all numerical inputs together without violating the qualitative dependence relationships shown in the graph. By contrast, if we specify the marginal probabilities on only three pairs, incompleteness results; many models exist which conform to the input specification, and we will not be able to provide answers to many useful queries.

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

1. Identify the cliques* of G, namely, the largest subgraphs in which the nodes are all adjacent to each other.

2. For each clique C_i, assign a non-negative *compatibility* function $g_i(C_i)$, which measures the relative compatibility of all possible value assignments to the variables included in C_i.

3. Form the product $\prod_i g_i(C_i)$ of the compatibility functions over all the cliques.

* We use the term "clique" to denote what is termed in most of the literature "maximal clique."

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

$$P(x_1, \cdots, x_n) = K \prod_i g_i(C_i) \tag{13}$$

where

$$K = \Big[\sum_{x_1, \cdots, x_n} \prod_i g_i(C_i) \Big]^{-1}$$

The normalized product P in Eq.(13) constitutes a joint distribution which embodies all the conditional independencies portrayed by the graph G, i.e., G is an I-map of P.

To illustrate the mechanics of this method, let us return to the example of Figure 1 and assume that for the i-th couple the likelihood that the two members end up with the same state of disease is measured by a compatibility parameter α_i while the likelihood that exactly one partner of the couple remains unaffected by the disease (while the other carries it) is assigned a compatibility parameter β_i. The dependency graph in this case has four cliques, corresponding to the four edges:

$$C_1 = \{M_1, F_1\} \qquad C_2 = \{M_1, F_2\}$$

$$C_3 = \{M_2, F_1\} \qquad C_4 = \{M_2, F_2\}$$

and the compatibility functions g_i are given by

$$g_i(x_{i_1}, x_{i_2}) = \begin{cases} \alpha_i & \text{if } x_{i_1} = x_{i_2} \\ \beta_i & \text{if } x_{i_1} \neq x_{i_2} \end{cases} \tag{14}$$

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

$$P(M_1, M_2, F_1, F_2) = K\ g_1(M_1, F_1) g_2(M_1, F_2) g_3(M_2, F_1) g_4(M_2, F_2) \tag{15}$$

$$= K \prod_i \beta_i^{|x_{i_1} - x_{i_2}|} \alpha_i^{1 - |x_{i_1} - x_{i_2}|}$$

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

$$K^{-1} = \prod_i (\alpha_i + \beta_i) + \prod_i \alpha_i \sum_j \frac{\beta_j}{\alpha_j} + \prod_i \beta_i \sum \frac{\alpha_j}{\beta_j} \tag{16}$$

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 members in every couple are in unequal states of the 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 C_1 and 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 (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.(4.c), we conclude: $I(M_1, F_1 \cup F_2, M_2)_P$ and $I(F_1, M_1 \cup M_2, F_2)_P$.

It is easy to prove the generality of this construction method:

Theorem 6: A probability function P formed by the product of functions on the cliques of G is a *Markov field* relative to G, i.e., G is an I-map of P.

Proof: The I-mapness of G would be guaranteed if P is locally-Markov relative to G (Theorem 5). It is sufficient, therefore, to show that the G-neighbors of each variable α constitute a Markov blanket of α relative to P, i.e., that $I(\alpha, B_G(\alpha), U - \alpha - B_G(\alpha))$ or, using Eq.(4.c), that

$$P(\alpha, B_G(\alpha), U - \alpha - B_G(\alpha)) = f_1(\alpha, B_G(\alpha))\, f_2(U - \alpha) \tag{17}$$

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

$$P(\alpha, \beta \cdots) = K \prod_j g_j(C_j) = K \prod_{j \in J_\alpha} g_j(C_j) \prod_{j \notin J_\alpha} g_j(C_j) \tag{18}$$

The first product in (18) involves only variables which are adjacent to α in G, or else the C_j would not be cliques. The second product, according to the definition of J_α, does not involve α. Thus, (17) is established. Q.E.D.

4.2 Interpreting the Link Parameters

The preceding method of modeling, while guaranteeing consistency and completeness, leaves much to be desired. Its main deficiency lies in the difficulty of assigning meaningful semantics to the parameters of the compatibility functions. If a model's parameters are to lead to meaningful inferences or decisions, they must be obtained either from direct measurements or from an expert who can relate them to actual human experience. Both options encounter difficulties in the Markov nets formulation.

Assuming we have a *huge* record of medical tests conducted on homogeneous population of subjects, including a full account of their sexual habits, can we extract from such record the desired compatibility functions $g_i(M, F)$? The difficulty is that whatever disease pattern we observe on any given couple, that pattern is a function not only of the relations between this couple but also of interaction between this couple and the rest of the population. In other words, we are invariably limited to measurements taken in a noisy environment which, in our case, amounts to having a large network of interactions surrounding the one under test.

To 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 1) isolated from the rest of the world, all groups having the same sexual pattern. In other words, we are actually 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, this is not an easy task, even in such an ideal case; it involves solving a set of simultaneous nonlinear equations for g_i, in terms of data provided by P. If, in addition to this difficulty, we also face the problem that whatever solution we obtain for g_i will not be applicable to new situations, say where the frequency of interaction is different, we realize that it is no mere coincidence that 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, a meaningful invariant would be the relation between the frequency of sexual contact and the transference of the disease from one partner to another under conditions of perfect isolation from the rest of the world. In probabilistic terminology, the quantities $P(f_1 | m_1, \neg m_2)$ and $P(f_1 | \neg m_1, \neg m_2)$ and their relations to the frequency of interaction of couple $\{M_1, F_1\}$ is what we perceive to be an invariant characteristic of the disease, generalizable across contexts. It is with these quantities, therefore, that an expert would choose to encode experiential knowledge and which he/she would find most comfortable to assess. Moreover, were we to conduct a clean scientific experiment, these are precisely the types of quantities we would choose to measure.

Unfortunately, the Markov net formulation does not allow the direct acquisition of such judgmental input. The compatibility parameters appear totally meaningless to the expert, while judgments about low-order conditional probabilities (e.g., $P(m_1 | f_1, \neg m_2)$) can be taken only as constraints over the joint probability distribution with which one hopes to obtain the actual values of the compatibility parameters. This is a rather tedious computational procedure, especially if the number of variables is large (e.g., imagine a ring of n interacting couples) and one which must be performed at the knowledge-acquisition phase in order to ensure that the expert provides a consistent and complete set of constraints.

4.3 Decomposable Models

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

To understand why tree topologies have this desired feature, let us again consider the example of Figure 1 and assume that the interaction between the couple $\{M_2, F_1\}$ is non-existent, namely, the Markov net consists of the chain F_1—M_1—F_2—M_2.

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

$$P(x_1, \dots x_n) = P(x_1)\, P(x_2 | x_1), \dots P(x_n | x_1, \dots x_{n-1}) \tag{19}$$

Thus, if we order our four variables along the chain by (F_1, M_1, F_2, M_2), we can write:

$$P(F_1, M_1, F_2, M_2) = P(F_1)\, P(M_1 | F_1)\, P(F_2 | F_1, M_1)\, P(M_2 | F_1, M_1, F_2)$$

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

$$P(F_1, M_1, F_2, M_2) = P(F_1)\, P(M_1 | F_1)\, P(F_2 | M_1)\, P(M_2 | F_2)$$

Clearly, then, 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 none other than the pairwise conditional probabilities of the interacting variables which, following our earlier argument, should carry conceptual meaning.

This scheme leaves ample flexibility as to the choice of ordering. For example, if we take the order (F_2, M_1, M_2, F_1), we get

$$P(F_2, M_1, M_2, F_1) = P(F_2)\, P(M_1 | F_2)\, P(M_2 | F_2, M_1)\, P(F_1 | F_2, M_1, M_2)$$

$$= P(F_2)\, P(M_1 | F_2)\, P(M_2 | F_2)\, P(F_1 | M_1),$$

again yielding a product of edge probabilities. The only requirement is that, as we order the variables from left to right, every variable (except the leftmost) should have at least one of its graph-neighbors to its left. The ordering (F_1, M_2, M_1, F_2), for example, would not yield the desired product-form because M_2 is positioned to the left of its two neighbors.

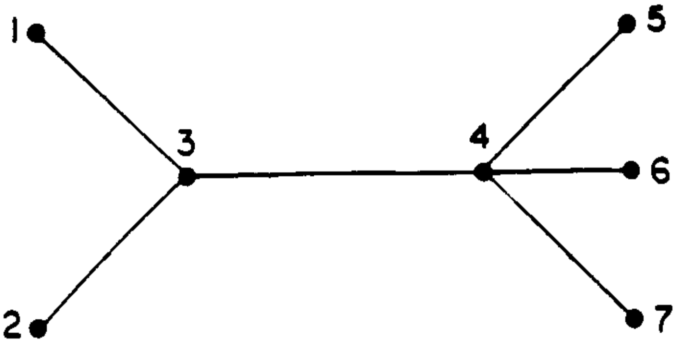


Figure 5(a)

Given a tree-structured Markov net, there are two ways by which one can write down its product-form distribution by inspection: *directed trees* and *product division.*

Consider the tree of Figure 5(a) where the variables $X_1, \ldots, X_7$ are marked 1, ... 7, for short. If we arbitrarily choose node 3 as a root and assign arrows to the links pointing "away" from the root, the directed tree of Figure 5(b) ensues, where every non-root node has a single incoming arrow designating its unique parent. We can now write the product distribution by inspection, going from parents to children:

$$P(1, \ldots, 7) = P(3)\, P(1 | 3)\, P(2 | 3)\, P(4 | 3)\, P(5 | 4)\, P(6 | 4)\, P(7 | 4) \qquad (20)$$

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 two products -- the product of the marginal distributions on the edges (cliques) divided by the product of the distributions of the intermediate nodes (the intersection of the cliques). The distribution corresponding to the tree of Figure 5(a) will be written

$$P(1, \ldots, 7) = \frac{P(1,3)\, P(2,3)\, P(3,4)\, P(4,5)\, P(4,6)\, P(4,7)}{P(3) \quad P(3) \quad P(4) \quad P(4) \quad P(4)}\quad, \qquad (21)$$

which is identical to that of Eq.(20).

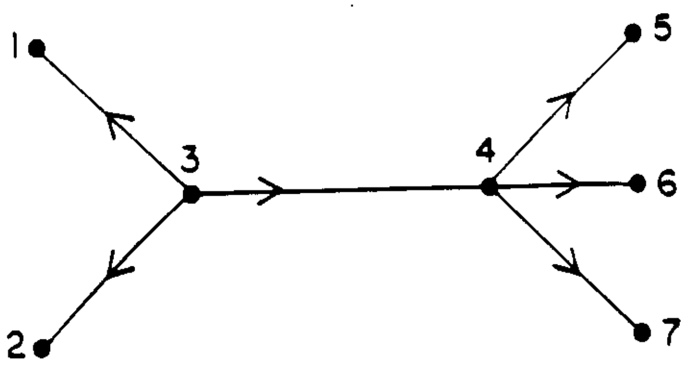
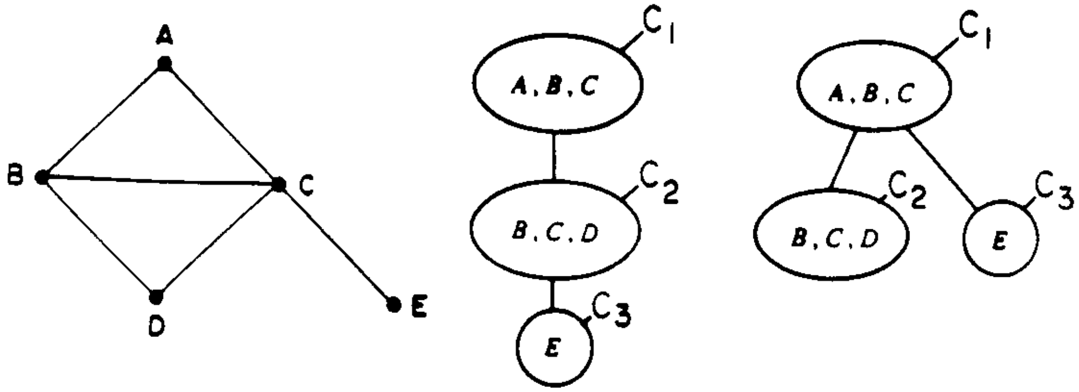


Figure 5(b)

Distributions amenable to product forms are not limited to trees. Consider, for example, the structure of Figure 6. Applying the chain rule in the order (A, B, C, D, E) and using the structural independencies of the graph, we obtain

$$\begin{aligned} P(A,B,C,D,E) &= P(A)P(B \mid A)P(C \mid A,B)P(D \mid A,B,C)P(E \mid A,B,C,D) \\ &= P(A)P(B \mid A)P(C \mid A,B)P(D \mid B,C)P(E \mid C) \\ &= \frac{P(A,B,C)P(B,C,D)}{P(B,C)} \ \frac{P(C\ E)}{P(C)} \qquad (22) \end{aligned}$$



(a) (b) (c)

Figure 6

Eq. (22) again displays the same pattern as in Eq.(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, it appears only once in the denominator. The reason for this will become clear in the ensuing discussion, where we shall justify the general formula for clique-trees.

The unique feature of the graph in Figure 6(a), which 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 in Figure 6(b) and 6(c). More precisely, there is a tree with vertices corresponding to the cliques of G which is an I-map of P. Indeed, writing $c_1 = \{A, B, C\}$, $c_2 = \{B, C, D\}$ and $c_3 = \{C, E\}$, we see that c_3 and c_1 are independent, given c_2, which yields the I-map $c_1 — c_2 — c_3$ of Figure 6(b). Alternatively, since c_3 and c_2 are independent given c_1, we can also use the I-map $c_2 — c_1 — c_3$ of Figure 6(c). This non-uniqueness of the minimal I-maps, an apparent contradiction to Theorem 3, stems from the non-disjointedness of c_1, c_2 and c_3 which, unlike dependencies among disjoint sets of variables, occasionally leads to violation of axiom (6.c).

The concept of a clique-tree is made more precise by the following Theorem [Beeri et al., 1981]; [Tarjan & Yannakakis, 1984] about *chordal graphs:*

Theorem 7: Let G be an undirected graph $G = (V, E)$ and let C be the set of maximal cliques of G. G is *chordal* if, and only if, either one of the following equivalent conditions holds:

1. Every cycle of length at least four has a chord, i.e., an edge joining two nonconsecutive vertices on the cycle.

2. The edges of G can be directed acyclically so that every pair of converging arrows emanates from two adjacent vertices.

3. All cliques of G can be deleted by repeatedly applying the following two operations:

   i) delete a vertex that occurs in only one clique;
   ii) delete a clique that is contained in another clique.

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 there is a path between them made entirely of cliques that also contain v.

The four conditions of the Theorem are clearly satisfied in the graph of Figure 6(a), and none is satisfied in that of Figure 1. (The diamond is the smallest nonchordal graph). Tarjan & Yannakakis [1984] offer an efficient algorithm for both testing chordality of graphs and for "filling out" the missing links that would turn non chordal graphs into chordal.

Definition: A probability model P is said to be *decomposable* if its Markov net is chordal. P is said to be *decomposable relative to a graph G* if

i) G is an I-map of P; and
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 C_x, C_y and C_z are three disjoint sets of vertices in T, and x, y, z their corresponding sets of variables in G, then $I(x, z, y)_P$ whenever C_z separates C_x from C_y in T (written $< C_x \mid C_z \mid C_y >_T$).

Proof: Since $(x, z, y)_P$ may not be disjoint, we will prove $I(x, z, y)_P$ by showing that

$I(x-z\,,z\,,y-z)_P$ holds -- the two assertions are equivalent, according to Eqs. (5.a) and (5.b). Moreover, since G is an I-map of P, it is enough to show that z is a cutset in G, separating $x-z$ from $y-z$, i.e., $<x-z \mid z \mid y-z>_G$. Thus, we need to show

$$<C_x \mid C_z \mid C_y>_T \Rightarrow <x-z \mid z \mid y-z>_G \tag{23}$$

which we shall prove by contradiction:

If the right-hand side of (23) is false, then there exists a path $\alpha, \gamma_1, \gamma_2,\ldots, \gamma_n, \beta$ in G from some element $\alpha \in x-z$ to some element $\beta \in y-z$ which does not intersect z, namely,

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

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

Let C_v denote the set of all cliques which contain some vertex v, and consider the set of cliques $S = \{C_\alpha \bigcup_i C_{\gamma_i} \bigcup C_\beta - C_z\}$. We now argue that those vertices of T corresponding to the elements of S form a connected sub-tree. Indeed, T was constructed in such a way that, pulling out the variables in C_z would leave the vertices of every C_{γ_i} connected and, 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 both (γ_i, γ_{i+1}) and the latter, in turn, shares an element (γ_{i+1}) with every clique containing γ_{i+1}. Consequently the vertices corresponding to the elements of C_{γ_i} and $C_{\gamma_{i+1}}$ are connected in T, even after deleting the variables in C_z. Q.E.D. This asserts the existence of a path in T, between some vertex in $C_\alpha \subseteq C_x$ and some vertex in $C_\beta \subseteq C_y$, which bypasses all vertices of C_z, thus contradicting the antecedent part of (20). Q.E.D.

We are now in a position to demonstrate that decomposable models have a joint distribution function expressible in product form. Essentially, the demonstration relies on property 4 of Theorem 7, which allows us to arrange the cliques of G in a tree-consistent ordering, and apply to them the chain-rule formula (19), as we have done to the individual variables in Eq.(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 in G and let $(C_1, C_2,\ldots C_i\ldots)$ be an ordering of the cliques which is consistent with T, i.e., for every $i > j$ we have a predecessor $j(i) < i$ for which $C_{j(i)}$ is adjacent to C_i in T and which separates C_i from $C_1, C_2,\ldots C_{i-1}$. Applying the chain-rule formula to the cliques of G, we obtain:

$$P(x_1, x_2\ldots x_n) = \prod_i P(C_i \mid C_1,\ldots, C_{i-1}) = \prod_i P(C_i \mid C_{j(i)}) \tag{24}$$

$$= \prod_i P(C_i \mid C_i \cap C_{j(i)}) \tag{25}$$

$$= \prod_i \frac{P\ (C_i)}{P\ \ (C_i \cap C_{j(i)})} \tag{26}$$

Eq.(24) follows from the I-mapness of T (Lemma 1), and Eq.(25) follows from the I-mapness of G since the variables which $C_{j(i)}$ does not share with C_i are separated from those in C_i by the variables common to both C_i and $C_{j(i)}$. In Figure 6(a), for example, A is separated from D by $\{B, C\}$.

Note that, to render P decomposable relative to some graph G, it is enough that G be any I-map of P, not necessarily minimal. That means that, if we desire to express P as a product of marginal distributions of clusters of variables and it so happened that the Markov net G_0 of P is nonchordal, it is possible to make G_0 chordal by "filling in" the missing chords and then expressing P as product of the cliques of the augmented graph. For example, if the Markov net of a certain model is given by the graph of Figure 6(a) with edge (BC) missing (e.g., as in Figure 1), 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 the cliques of G, as in Eq. (22). It is true that the independence $I(B, AD, C)$ is not explicit in the expression of (22) and can be encoded only by careful numerical crafting of the distributions $P(A, B, C)$ and $P(B, C, D)$. Once encoded, however, the tree structure of the cliques of G facilitates convenient, propagation-like updating of probabilities in response to new observations [Spiegelhalter, 1986]. Moreover, in situations where the cluster distributions are obtained by statistical measurements, the "filling in" method is useful in directing the experimenter toward selecting the right variable aggregates for measurement [Goldman & Rivest, 1986]. For example, in the model depicted by Figure 1, the "filling in" method would advise the experimenter to tabulate measurements of variable triplets (e.g., $\{M_1, F_1, F_2\}$ and $\{M_2, F_1, F_2\}$, not merely of variable pairs.

4. BAYESIAN BELIEF NETWORKS

The main weakness of Markov nets stems from their inability to represent nonmonotonic dependencies; two independent variables must be directly connected by an edge, merely because there exists some other variable that depends on both. As a result, many useful independencies remain unrepresented in the network. To overcome this deficiency, Bayesian networks make use of the richer language of *directed* graphs, where the directions of the arrows permit us to distinguish genuine dependencies from spurious dependencies induced by hypothetical future observations. For instance, the coins-and-bell example of Section 3 will be represented by the network $coin\,1 \rightarrow bell \leftarrow coin\,2$, which more naturally reflects the common perception of causal influences; the arrows clearly indicate that the sound of the bell is determined by the outcomes of the coins, not the other way around.

These arrows endow special status to the path between coin 1 and coin 2, reflecting the nonmonotonic dependency between the three variables. This path traverses two adjacent arrows converging head-to-head on a variable z = "bell sound." 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 until the variable z (or any of its descendents) is instantiated. This special criterion of *direction-conditional connectivity* perfectly matches the nonmonotonic dependency relationship among the three variables; the outcomes of the two coins are marginally independent but become dependent upon knowing the outcome of the bell (or any external evidence bearing on that outcome). The connectivity criterion reverts back to the usual cutset criterion of Markov nets whenever the arrows are either diverging e.g.,

$$height \leftarrow age \rightarrow reading\ ability$$

or cascaded, e.g.,

$$weather \rightarrow wheat\ crop \rightarrow wheat\ price\,.$$

A detailed discussion of this criterion in the context of general networks and an examination of its power of expression are contained in Section 4.2 and 4.3, respectively. First we introduce a definition of Bayesian networks and their methods of construction.

4.1 Constructing a Bayesian Network

Bayesian Belief networks are directed acyclic graphs ("dia-graphs") in which the nodes represent variables, the arcs signify the existence of direct causal influences between the linked variables and the strength of these influences are quantified by forward conditional probabilities.

Informally, if the nodes in the graph represent the variables $X_1, X_2 ... X_n$, the structure of a Bayesian network can be determined by a simple procedure: we assign each variable to a vertex in a graph and draw arrows toward each vertex X_i from a set S_i of vertices perceived to be "direct causes" of X_i. More formally, the notion of "direct cause" can be defined in terms of a probability distribution $P(x_1 ... x_n)$ and an ordering d on the variables. (In practice, the ordering d will not be arbitrary but will reflect one's perceived flow of causation, i.e., no variable should be preceded by any of its consequences). Given P and d, the *direct causes* of X_i are the smallest set of variables $S_i \subseteq \{X_1 ... X_{i-1}\}$ satisfying the condition

$$P(x_i \mid S_i) = P(x_i \mid x_{i-1}, ..., x_1) \tag{27}$$

In other words, S_i is the Markov boundary of X_i relative to the set $U_{(i)} = \{X_1, X_2, ..., X_i\}$ of variables, namely, the smallest set of variables that "shields" X_i from all its other predecessors. Since Markov boundaries are unique (Theorem 4), the set of parents S_i assigned to each variable is unique and the structure of the dia-graph is well defined.

This leads to a simple method of constructing a dia-graph representation, given any joint distribution $P(x_1 ... x_n)$ and an order d on the variables in U. We start by choosing X_1 as a root of the graph 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 $P(x_2 \mid x_1)$. Otherwise, we leave X_1 and X_2 unconnected and assign the prior $P(x_2)$ to node X_2. At the i-th stage, we form the node X_i and establish a group of directed links to X_i from the set S_i defined by Eq.(1), and quantify this group of arrows by the conditional probability $P(x_i \mid S_i)$. Thus, the distribution, $P(x_1, ..., x_n)$, together with the order d uniquely identify a directed acyclic graph which represents many of the independencies embedded in $P(x_1, ..., x_n)$.

The conditional probabilities $P(x_i \mid S_i)$ on the links of the dia-graph contain all the information necessary for reconstructing the original distribution function. Writing the chain-rule formula in the ordering d and using Eq.(27) leads to the product:

$$P(x_1, x_2, ... x_n) = P(x_n \mid x_{n-1} ... x_1)\, P(x_{n-1} \mid x_{n-2} ... x_1) ... P(x_3 \mid x_2, x_1)\, P(x_2 \mid x_1)\, P(x_1)$$

$$= \prod_i P(x_i \mid S_i) \tag{28}$$

So, for example, the distribution corresponding to the dia-graph of Figure 7 can be written by inspection:

$$P(x_1, x_2, x_3, x_4, x_5, x_6) = P(x_6 \mid x_5)\, P(x_5 \mid x_2, x_3)\, P(x_4 \mid x_1, x_2)\, P(x_3 \mid x_1)\, P(x_2 \mid x_1)\, P(x_1) \tag{29}$$

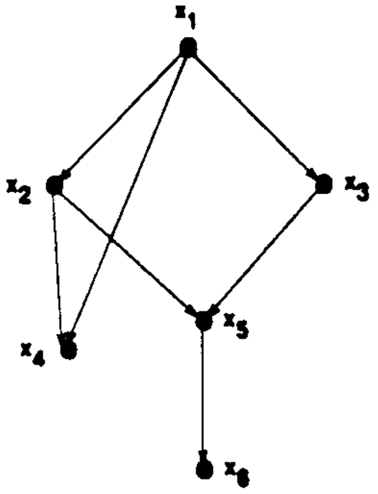


Figure 7

In expert-systems applications where, instead of a numerical representation for $P(x_1,...,x_n)$, we have only intuitive understanding of the major constraints in the domain, the graph can still be configured as before by a local method, except that the parent set S_i must be selected judgmentally. The addition to the network of any new node X_i requires only that the expert identify a set S_i of variables which "directly bear" on X_i, locally assess the strength of this relationship and make no commitment regarding the effect of X_i on other variables outside S_i. We shall next see that, even though each judgment is performed locally, their sum total is guaranteed to be complete and consistent.

Suppose we are given a directed acyclic graph G in which the arrows pointing to each node X_i emanate from a set S_i of parent nodes, and we wish to quantify the strengths of these influences in a complete and consistent fashion. Since, by direct parents we mean a set of variables which, once we fix their values, would shield X_i from the influence of all other predecessors of X_i (i.e., $P(x_i \mid S_i) = P(x_i \mid x_1,..., x_{i-1})$), the chain-rule formula (28) states that a separate assessment of each child-parents' relationship should suffice. We need only assess the conditional probabilities, $P(x_i \mid S_i)$, by some functions, $F_i(x_i, S_i)$, and make sure these assessments satisfy

$$\sum_{x_i} F_i(x_i, S_i) = 1 \qquad 0 \le F_i(x_i, S_i) \le 1 \tag{30}$$

where the summation ranges over all values of x_i. This specification is complete and consistent because the product form

$$P_a(x_1 ... x_n) = \prod_i F_i(x_i, S_i) \tag{31}$$

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

$$P_a(x_i \mid S_i) = \frac{P_a(x_i, S_i)}{P_a(S_i)} = \frac{\sum\limits_{x_j \notin (x_i \cup S_i)} P_a(x_1, \ldots, x_n)}{\sum\limits_{x_j \notin S_i} P_a(x_1, \ldots, x_n)} = F_i(x_i, S_i) \tag{32}$$

Contrasted with the difficulties of quantifying Markov-nets, this model-building process offers a significant advantage. The parameters requested from the model builder are exactly the forward conditional probabilities which quantify stable conceptual relationships in one's memory; they are both psychologically meaningful and can be obtained by direct measurements. The mental activity required for assessing the parameters of $P(x_i \mid S_i)$ involves estimating the likelihood that x_i will occur, given a condition specified by any instantiation of the variables in S_i. (For example, the likelihood that a patient will develop a certain symptom, assuming that he/she suffers from a given combination of disorders.) These kinds of assessments are natural because they relate to the primitive relationships people use to encode empirical knowledge.

Aside from supporting a consistent set of assessments, another dia-graph advantage is that it permits people to qualitatively express the essential causal relationships in the domain; the network augments these input relationships with additional independencies implied by the inputs and preserves them, despite sloppy assignments of numerical estimates. In Fig. 1, for example, the fact that X_6 can tell us nothing new about X_1 once we know X_2 and X_3 was not stated explicitly by the model builder. Yet, it is logically implied by other inputs and will remain part of the model, independent of how the numbers are assigned to the links.

Dia-graphs constructed by this method will be called *"Bayesian Belief Networks"* or *"Influence Networks"* interchangeably, the former to emphasize the judgmental origin and probabilistic nature of the quantifiers, the latter to reflect the directionality of the links. When the nature of the interactions is perceived to be causal, then the term *"Causal Network"* may also be appropriate. In general, however, an influence network may also represent associative or inferential dependencies, in which case the directionality of the arrows mainly provides computational convenience [Howard & Matheson, 1984].

In the strictest sense, Bayesian belief networks are not graphs but hypergraphs because to describe the dependency of a given node on its k parents requires a function of $k+1$ arguments which, in general, could not be specified by k two-place functions on the individual links. However, both the directionality of the arrows and the fact that many parents remain unlinked convey important information that would be lost, had we used the standard hypergraph representation specifying merely the list of dependent subsets.

If the number of parents k is large, estimating $P(x_i \mid S_i)$ may be troublesome because, in principle, it requires a table of size 2^k. In practice, however, people conceptualize causal relationships by forming hierarchies of small clusters of variables and, moreover, the interactions among the factors in each cluster are normally perceived to fall into one of a few prestored, prototypical structures, each requiring about k parameters. Common examples of such prototypical structures are: noisy OR gates (i.e., any one of the factors is likely to trigger the effect), noisy AND gates and various enabling mechanisms (i.e., factors identified as having no influence of their own except enabling other influences to become effective). Detailed analysis of the noisy OR-gate model is given in [Pearl, 1986(a)].

Note that the topology of a Bayesian network can be extremely sensitive to the node ordering d; a network with a tree structure in one ordering may turn into a complete graph if that ordering is reversed. For example, if $x_1, \ldots, x_n$ stands for the outcomes of n independent coins, and x_{n+1} represents the output of a detector triggered if any of the coins comes up HEADS, then the influence network will be an inverted tree of n arrows pointing from each of the variables $x_1, \ldots, x_n$ toward x_{n+1}. On the other hand, if the detector's outcome is chosen to be the first variable, say x_0, then the underlying influence network will be a complete graph.

This order sensitivity may seem paradoxical at first; d can be arbitrarily chosen, whereas people have fairly uniform conceptual structures, e.g., they agree on whether a pair of propositions are directly or indirectly related. The consensus about the structure of influence networks is indicative of the dominant role *causality* plays in the formation of these networks. In other words, the standard ordering imposed by the direction of causation indirectly induces identical topologies on the networks that people adopt for encoding experiential knowledge. It is tempting to speculate that, were it not for the social convention of adopting a standard ordering of events conforming to the flow of time and causation, human communication (as we now know it) would be impossible. It also raises the philosophical question of whether causality is not but psychological illusion created by computational needs, i.e., that the flow of causation we attribute to the external world is no other but an ordering found to lead to the most parsimonious and effective encoding of our experience. More on this subject can be found in [Pearl, 1986(b)].

4.2 Dia-Graph Separation and Conditional Independence

To facilitate the verification of dependencies among the variables in a Bayes network, we need to establish a clear correspondence between the topology of the network and the dependence relationships portrayed by it. In Markov nets this correspondence was based on a simple graph separation criterion: Should the removal of some subset S of nodes from the network render nodes X_i and X_j disconnected, X_i and X_j were proclaimed to be independent given S, i.e.,

$$< X_i, S, X_j >_G \Rightarrow I(X_i, S, X_j)$$

To serve as I-maps for nonmonotonic dependencies, Bayes networks are based on a slightly more complex criterion of separability, one which takes into consideration the directionality of the arrows in the graph. This criterion distinguishes between the three possible ways that a pair of arrows may join at some vertex X_2:

(1) tail-to-tail, $X_1 \leftarrow X_2 \rightarrow X_3$
(2) head-to-tail, $X_1 \rightarrow X_2 \rightarrow X_3$ or $X_1 \leftarrow X_2 \leftarrow X_3$
(3) head-to-head, $X_1 \rightarrow X_2 \leftarrow X_3$

If we assume that X_1, X_2, X_3 are the only variables involved, it is clear from the method of constructing the network that, in cases (1) and (2), X_1 and X_3 are conditionally independent, given X_2, while in case (3), X_1 and X_3 are marginally independent (i.e., $P(X_3 | X_1) = P(X_3)$) but may become dependent, given the value of X_2. Moreover, if X_2 in case (3) has descendants $X_4, X_5 \ldots$, then X_1 and X_3 may also become dependent if any one of those descendant variables is instantiated. These considerations motivate the definition of a qualified version of path connectivity, applicable to paths with directed links and sensitive to all the variables for which values are known at a given time.

Definition:

a. Two arrows meeting head-to-tail or tail-to-tail at node α are said to be *blocked by* a set S of vertices S if α is in S.

b. Two arrows meeting head-to-head at node α are *blocked by* S if neither α nor any of its descendants is in S.

Definition:

a. An undirected path P in a dia-graph G_d is said to be *d-separated* by a subset S of vertices if at least one pair of successive arrows along P is *blocked* by S.

b. Let x, y, and S be three disjoint sets of vertices in a dia-graph G_d. S is said to *d-separate* x from y if all paths between x and y are *d-separated* by S. Such separation will be denoted by $< x \mid S \mid y >_{G_d}$.

This modified definition of separation provides a valid test for conditional independence in dia-graphs:

Theorem 9: [Verma, 1986]: Let G_d be a dia-graph constructed from distribution P in some order d. If x, y and z are three disjoint subsets of vertices in G_d such that z d-separates x from y, then x and y are conditionally independent given z, in P. In other words, G_d is an I-map of P relative to d-separation:

$$< x \mid z \mid y >_{G_d} \Rightarrow I(x, z, y)_P$$

The proof of Theorem 9 uses the contraction axiom (4.e)

The procedure involved in testing d-separation is only slightly more complicated than the conventional test for cutset separation in undirected graphs and can be handled by visual inspection. In Figure 1, for example, one can easily verify that variables X_2 and X_3 are d-separated by $S_1 = \{X_1\}$ or $S_2 = \{X_1, X_4\}$ because the two paths between X_2 and X_3 are blocked by either one of these subsets. However, X_2 and X_3 are not separated by $S_3 = \{X_1, X_6\}$ because X_6, as a descendant of X_5, "unblocks" the head-to-head connection at X_5, thus opening a pathway between X_2 and X_3.

The d-separation criterion is used routinely by technicians involved in electronic troubleshooting. This is so because the functional dependencies between the inputs and output of electronic devices match the dependencies portrayed by dia-graphs -- 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.

Although the structure of Bayes networks, together with the directionality of its links, depends strongly on the node ordering used in the network construction, conditional independence is a property of the underlying distribution and is, therefore, order-invariant. Thus, if we succeed in finding an ordering d in which a given conditional independence relationship becomes graphically transparent, that relationship remains valid even though it may not induce a graph-separation pattern in networks corresponding to other orderings. This permits the use of Bayes networks for identifying, by inspection, a *Markov blanket* for any given node, namely, a set S of variables that renders a given variable independent of all variables not in S. The intersection of the Markov blankets induced by all possible orderings gives, of course, the Markov boundaries. The d-separation criterion for Bayes networks guarantees that the union of the following three types of neighbors is sufficient for forming a Markov blanket: direct parents, direct successors and all direct parents of the latter. Thus, if the network consists of a single path (traditionally called a Markov chain), the Markov blanket of any non-terminal node consists of its two immediate neighbors while, in trees, the Markov blanket consists of the (unique) father and the immediate successors. In Figure 1, however, the Markov blanket of X_3 is $\{X_1, X_5, X_2\}$.

4.3 How Expressive are Dia-Graphs?

One would normally expect that the introduction of directionality into the language of graphs would render them more expressive, capable of portraying a greater number of conditional independencies. We saw, indeed, that the d-separation criterion permits us to display both nonmonotonic and nontransitive dependencies that were excluded from the Markov net vocabulary. Thus, it is natural to ask how the expressive power of dia-graphs compares with that of undirected graphs and numerical representations of probability. This brings up two questions:

1. Are all dependencies representable by Markov nets also representable by a Bayesian net?

2. How well can Bayesian nets represent the type of dependencies induced by probabilistic models?

The answer to the first question is, clearly, negative. For instance, the dependency structure of a diamond-shaped Markov net (e.g., Figure 1) with edges (AB), (AC), (CD) and (BD) asserts the two independence relationships: $I(A, BC, D)$ and $I(B, AD, C)$. No Bayesian net 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 the latter but not the former. In view of property (4) of Theorem (7), it is clear that this difficulty will always be encountered in nonchordal graphs. No matter how we direct the arrows, there will always be a pair of non-adjacent parents sharing a common child, a configuration which yields independence in Markov nets but dependence in Bayes nets.

The inability of dia-graphs to display some common probabilistic dependencies is also obvious. It is hampered by the failure of every graphical representation to distinguish connectivity between sets from connectivity among their elements. In other words, in graphs (directed as well as undirected) 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, pairwise independence does not imply joint independence (see Eq.(6.b)) as demonstrated in the coins-and-bell example. When the coins are both fair, all three pairs of variables are mutually independent; yet, every variable is dependent (deterministically) on the other two.

Despite these shortcomings, we will see that the dia-graph representation is far more flexible than its undirected graph counterpart and, in addition, captures the great majority of probabilistic independencies, especially those which are conceptually meaningful. To this end, we offer an axiomatic characterization of dia-graph dependencies, which clearly indicates where they differ from those of undirect graphs (10) as well as probabilistic dependencies (6).

Definition: A dependency model M is said to be a *dia-graph isomorph* if there is a dia-graph G_d which is a perfect map of M relative to d-separation, i.e.,

$$I(x, z, y)_M \iff \langle x \mid z \mid y \rangle_{G_d}$$

Theorem 10: A necessary condition for a dependency model M to be a dia-graph isomorph is that $I(x, z, y)_M$ satisfies the following independent axioms (the subscript M dropped for clarity):

Symmetry

$$I(x,z,y) \iff I(y,z,x) \tag{34.a}$$

Composition - Decomposition

$$I(x,z,y \cup w) \iff I(x,z,y) \;\&\; I(x,z,w) \tag{34.b}$$

Intersection

$$I(x,z \cup w,y) \;\&\; I(x,z \cup y,w) \Rightarrow I(x,z,y \cup w) \tag{34.c}$$

Weak Union

$$I(x,z,y \cup w) \Rightarrow I(x,z \cup w,y) \tag{34.d}$$

Contraction

$$I(x,z \cup y,w) \;\&\; I(x,z,y) \Rightarrow I(x,z,y \cup w) \tag{34.e}$$

Weak Transitivity

$$I(x,z,y) \;\&\; I(x,z \cup \gamma,y) \Rightarrow I(x,z,\gamma) \;\; or \;\; I(\gamma,z,y) \tag{34.f}$$

Chordality

$$I(x,z \cup w,y) \;\&\; I(z,x \cup y,w) \Rightarrow I(x,z,y) \;\; or \;\; I(x,w,y) \tag{34.g}$$

Remarks: Axioms (34.a) and (34.c-e) are identical to those governing probabilistic dependencies (Eq. (6)). The left implication of (34.b) and the last two axioms, namely, composition, weak-transitivity and chordality, represent additional constraints over the system of Eq.(6). Thus, every dependency model which is a dia-graph isomorph also has a probabilistic representation but not vice-versa. The composition axiom (left implication of (34.b)) asserts that separation between sets is completely defined in terms of separation between singletons. Therefore, there will be no loss of generality in treating the first and third arguments of each triplet as individual elements of U.

Comparing (34) to the axioms defining separation in undirected graphs (10), we note that (10) implies all axioms in (34) except chordality (34.g). In particular, weak-union is implied by strong union, composition and contraction are implied by (10.c) and (10.d) and, of course, weak transitivity is implied by transitivity (10.e).

Weak transitivity asserts that, if two variables, x and y, are both unconditionally independent and conditionally independent given a third variable γ, then it is impossible for both x and y to be dependent on γ. This restriction, which may be violated in some probability models, remains in effect when we associate independence with separation in dia-graphs. Indeed, if both x and y are d-connected to γ in some dia-graph, then there must be an unblocked path from x to γ and an unblocked path from y to γ. These two form at least one path from x to y via γ. Now, if that path traverses γ along converging arrows, it should get unblocked by instantiating γ; so, x and y could not be d-separated given γ. Conversely, if the arrows meeting at γ are non-converging, x and y could not be d-separated by any set not containing γ (i.e., γ uninstantiated).

Probability theory, on the other hand, does not insist on weak transitivity, as it allows for the co-occurrence of the following four conditions:

1. $I(x,\emptyset,y)_P$ 2. $I(x,\gamma,y)_P$ 3. $\neg I(x,\emptyset,\gamma)_P$ 4. $\neg I(y,\emptyset,\gamma)_P$

For example, if both x and y are binary variables $x, y \in \{true, false\}$ and γ is a ternary variable $\gamma \in \{1, 2, 3\}$, we may have x dependent on γ via:

$$P(x = true \mid \gamma) = (1/2,\ 1/4,\ 1/8),$$

and γ dependent on y via:

$$P(\gamma \mid y = true) = (1/3,\ 1/3,\ 1/3)$$

$$P(\gamma \mid y = false) = (1/2,\ 1/2,\ 0).$$

Yet, x and y are mutually independent both conditionally (given γ) and unconditionally.

Thus, although dia-graphs seem better capable of displaying non-transitive dependencies than undirected graphs, even they require some weak form of transitivity and fall short of capturing totally non-transitive probabilistic dependencies. It can be shown, however, that if all variables in U are binary, then all probabilistic dependencies must be weakly transitive.

The purpose of the chordality axiom (34.f) is to exclude dependence models whose Markov nets are non-chordal (such as the one in Figure 1) since these cannot be completely captured by dia-graphs. Axiom (34.f), in essence, insists on either adding the appropriate chords to any long cycle (length ≥ 4), thus falsifying the antecedent of (34.f), or nullifying some of its links, thus satisfying the consequent part of (34.f).

Non-chordal graphs represent the one class of dependencies where undirected graphs exhibit expressiveness superior to that of dia-graphs. As we shall see next, this superiority can be eliminated by the introduction of auxiliary variables.

Consider the diamond-shaped graph of Figure 8(a), which asserts the two independence relationships: $I(A, BC, D)$ and $I(B, AD, C)$. Introducing an auxiliary variable E in the manner shown in Figure 8(b) creates a dia-graph 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 at some fixed value $E = e_1$, as in Figure 8(c). The dependency structure induced by the clamped dia-graph on A, B, C, D is identical to the original structure of Figure 8(a). Indeed, applying the d-separation criterion to Figure 8(c) recreates exactly 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

$$P(A, B, C, D \mid E = e_1) = \frac{P(A, B, C, D, e_1)}{P(e_1)}$$

$$= K\ P(e_1 \mid D, C)\, P(D \mid B)\, P(C \mid A)\, P(B \mid A)\, P(A)$$

$$= g_1(D, C)\, g_2(D, B)\, g_3(A, C)\, g_4(A, B) \quad ,$$

and, using the analysis of Section 4.1, we see that this distribution is equivalent to the one portrayed by Figure 8(a).

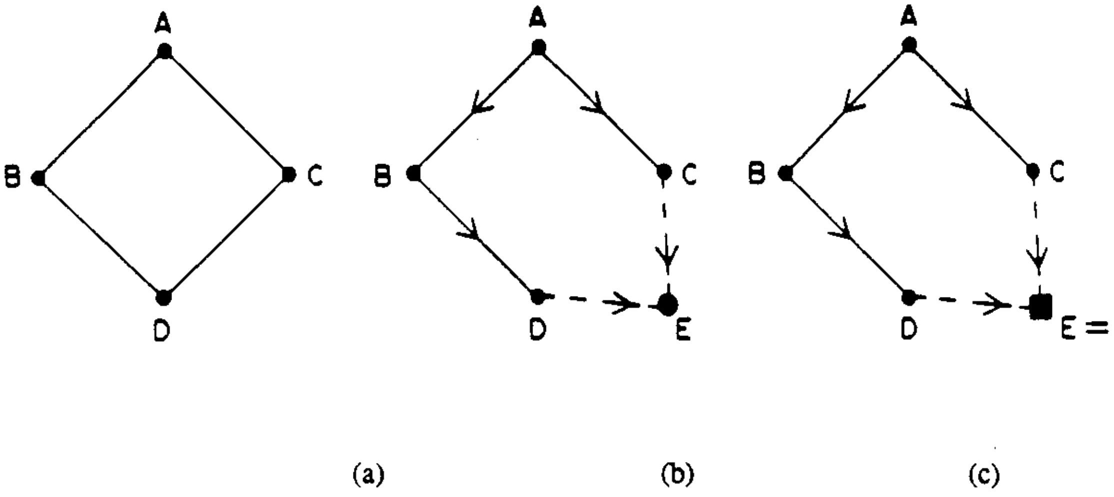


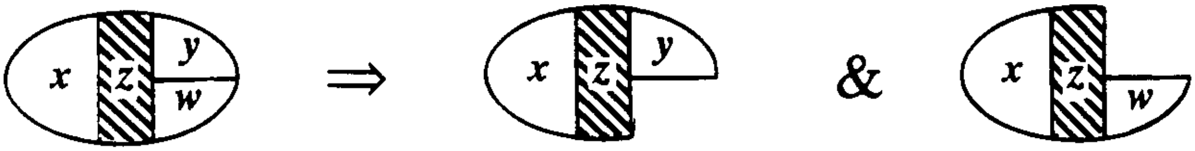
Figure 8

In conclusion, we see that the introduction of auxiliary variables permits us to dispose of the chordality restriction of (34.f) and renders the dia-graph representation superior to that of undirected graphs; that is, every dependency model expressible by the latter is also expressible by the former.

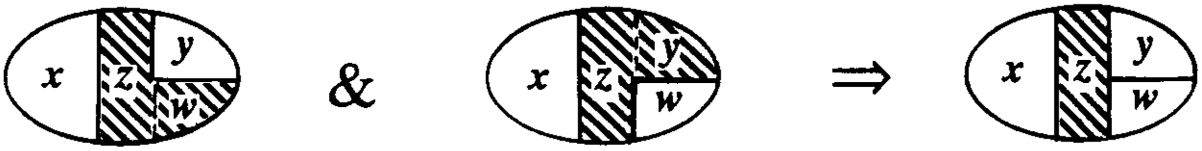
APPENDIX I

DESCRIPTIVE SCHEMATICS FOR AXIOMS 6(a)-6(e)

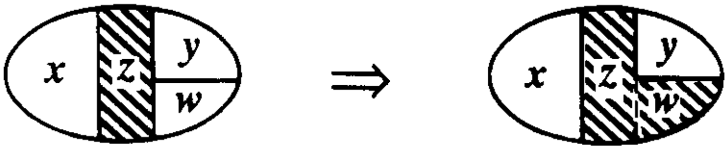
Decomposition



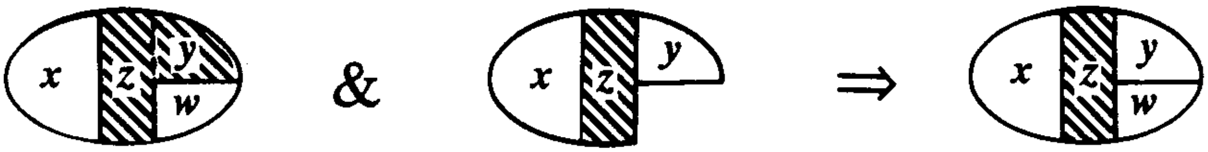
Intersection



Weak Union



Contraction



D. J. Spiegelhalter, ''Probabilistic Reasoning in Predictive Expert Systems,'' Kanal, L. N. & Lemmer, J., (Eds.), *Uncertainty in Artificial Intelligence,* North-Holland, Amsterdam, 1986.

Tarjan & M. Yannakakis, ''Simple Linear-Time Algorithms to Test Chordality of Graphs, Test Acyclicity of Hypergraphs and Selectively Reduce Acyclic Hypergraphs,'' *SIAM J. Computing,* Vol. 13, 1984, pp. 566-579.

T. S. Verma, ''Causal Networks: Semantics and Expressiveness,'' UCLA Cognitive Systems Laboratory *Technical Report R-65,* Los Angeles, California, in preparation, 1986.

W. A. Woods, ''What's in a Link? Foundations for Semantic Networks,'' Bobrow and Collins (Eds.), *Representation and Understanding,* Academic Press, Inc. 1975.

REFERENCES

Jon Doyle, ''A Truth Maintenance System,'' *Artificial Intelligence,* Vol. 12, No. (3), 1979, pp. 231-272.

R. O. Duda, P. E. Hart & N. J. Nilsson, ''Subjective Bayesian Methods for Rule-Based Inference Systems,'' *Proceedings, 1976 National Computer Conference (AFIPS Conference Proceedings),* 45, 1075-1082, 1976.

S. Geman & D. Geman, ''Stochastic Relaxations, Gibbs Distributions and the Bayesian Restoration of Images,'' *IEEE Trans. on Pattern Analysis and Machine Intelligence,* PAMI-6, November 1984, pp. 721-742.

S. A. Goldman & R. L. Rivest, ''Making Maximum Entropy Computations Easier by Adding Extra Constraints,'' *Proceedings, 6th Annual Workshop on Maximum Entropy and Bayesian Methods in Applied Statistics,* August, 1986.

R. A. Howard & J. E. Matheson, *Reading on the Principles and Applications of Decision Analysis,* Strategic Decisions Group, Menlo Park, California, 1984.

V. Isham, ''An Introduction to Spatial Point Processes and Markov Random Fields,'' *International Statistical Review,* Vol. 49, 1981, pp. 21-43.

S. L. Lauritzen, *Lectures on Contingency Tables,* 2nd Ed., University of Aalborg Press, Aalborg, Denmark, 1982.

U. Montanari, ''Networks of Constraints,'' *Information Science,* Vol. 7, 1974, pp. 95-132.

J. Pearl, ''Distributed Revision of Belief Commitment in Multi-Hypotheses Interpretation,'' UCLA Computer Science Department *Technical Report CSD-860045 (R-64),* June 1986; also, *2nd AAAI Workshop on Uncertainty in Artificial Intelligence,* Philadelphia, PA., August 1986(a).

J. Pearl, ''Fusion, Propagation and Structuring in Belief Networks,'' UCLA Computer Science Department *Technical Report 850022 (R-42);* also, *Artificial Intelligence,* Vol. 29, No. 3, September 1986(b), pp. 241-288.

J. Pearl & A. Paz, ''On the Logic of Representing Dependencies by Graphs,'' UCLA Computer Science Department *Technical Report CSD-860047 (R-56);* also, *Proceedings, 1986 Canadian AI Conference,* Montreal, May 1986 pp. 94-98.

J. Pearl & A. Paz, ''GRAPHOIDS: a Graph-based Logic for Reasoning about Relevance Relations,'' UCLA Computer Science Department *Technical Report 850038 (R-53),* October 1985; also, *Proceedings, ECAI-86,* Brighton, U.K., June 1986.

R. Schank, ''Conceptual Dependency: a Theory of Natural Language Understanding,'' *Cognitive Psychology,* Vol. 3, No. (4), 1972.