ON THE LOGIC OF PROBABILISTIC DEPENDENCIES

Judea Pearl

Cognitive Systems Laboratory, Computer Science Dept., UCLA, Los Angeles, CA. 90024

ABSTRACT

This paper uncovers the axiomatic basis for the probabilistic relation "x is independent of y, given z" and offers it as a formal definition of informational dependency. Given an initial set of such independence relationships, the axioms established permits us to infer new independencies by non-numeric, logical manipulations. Additionally, the paper legitimizes the use of inference networks to represent probabilistic dependencies by establishing a clear correspondence between the two relational structures. Given an arbitrary probabilistic model, P, we demonstrate a construction of a unique edge-minimum graph G such that each time we observe a vertex x separated from y by a subset S of vertices, we can be guaranteed that variables x and y are independent in P, given the values of the variables in S.

1. INTRODUCTION

Any system that reasons about knowledge and beliefs must make use of information about dependencies and relevancies. 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, information not available from z. 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. Similarly, knowing the color of x's car normally tells me nothing about the color of Y's. However, if X were to tell me that he almost mistook Y's car for his own, the two pieces of information become relevant to each other -- 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 this type of reasoning?

In probability theory, the notion of relevance is given precise quantitative underpinning using the device of *conditional independence*. A variable x is said to be independent of y given the information z if

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

However, it is rather unreasonable to expect people or machines to resort to numerical verification of equalities in order to extract relevance information. The ease and conviction with which people detect relevance relationships strongly suggest that such information is readily available from the organizational structure of human memory, not from numerical values assigned to its components. Accordingly, it would be interesting to explore how assertions about relevance can be inferred qualitatively from various models of memory and, in particular, whether the logic of such assertions coincides with that of probabilistic dependencies.

Since models of human memory are normally portrayed in terms of semantic networks of concepts and relations [Woods 1975], a natural question to ask is whether the notion of probabilistic dependency can be captured by a network representation, in the sense that all dependencies and independencies in a given probabilistic model could be detected from the topological properties of some network. For a given probability distribution P and any three variables x, y and 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, the topology of networks which display the underlying dependencies is not explicitly given by the numeric representation of probabilities.

This paper accomplishes two tasks. First, it 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. Second, the paper legitimizes the use of networks to represent probabilistic dependencies by establishing a clear correspondence between the two relational structures. Given an arbitrary probabilistic model, P, we demonstrate a construction of a unique edge-minimum graph G such that each time we observe a vertex x separated from y by a subset S of vertices, we can be guaranteed that variables x and y are independent in P, given the values of the variables in S. This correspondence provides a semantic for the topology of propositional inference networks like those used in expert systems [Duda, Hart and Nilsson 1976].

*This work was supported in part by the National Science Foundation, Grant #DSR 83-13875.

2. AN AXIOMATIC BASIS FOR PROBABILISTIC DEPENDENCIES

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

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

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

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

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

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,

$$I(x, z, y)_P \ iff\ P(x, y \mid z) = P(x \mid z)\, P(y \mid z) \quad (3)$$

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

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

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

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

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

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

$$I(x, z, y) \Longrightarrow I(x, (z, f(y)), y) \quad (5.a)$$

$$I(x, z, y) \Longrightarrow I(f(x, z), z, y) \quad (5.b)$$

The proof of these properties can be derived by elementary means from the definition (1). 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." The next set of properties constitute 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:

$$I(x, z, y) \Longleftrightarrow I(y, z, x)$$
Symmetry $\quad (6.a)$

$$I(x, z, y \cup w) \Longrightarrow I(x, z, y) \ \&\ I(x, z, w)$$
Decomposition $\quad (6.b)$

$$I(x, z \cup w, y) \ \&\ I(x, z \cup y, w) \Longrightarrow I(x, z, y \cup w)$$
Exchange $\quad (6.c)$

$$I(x, z, y \cup w) \Longrightarrow I(x, z \cup w, y)$$
Expansion $\quad (6.d)$

$$I(x, z \cup y, w) \ \&\ I(x, z, y) \Longrightarrow I(x, z, y \cup w)$$
Contraction $\quad (6.e)$

For technical convenience we shall 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. (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 expansion 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 proof of Theorem 1 can be derived by elementary means from the definition (3) and from the basic axioms of probability theory. The exchange 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. 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 $\varepsilon \to 0$. 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.

A graphical interpretation for properties (6.a) through (6.e) can be obtained by envisioning a graph with a set of vertices U and associating the relationship $I(A,B,C)$ with the statement "B intervenes between A and C" or, in other words, "the removal of a set B of nodes would render the nodes in A disconnected from those in C." The validity of (6.c) through (6.e) is clearly depicted by the chain x-z-y-w.

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

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

Although we have not been able to establish a general proof of completeness, we were not able to find any violating example, i.e., we could not find another general property of conditional independence which is not implied by (6.a) through (6.e).

3. A GRAPHICAL REPRESENTATION FOR PROBABILISTIC DEPENDENCIES

Let G be an undirected graph and let $<x \mid S \mid y>_G$ stand for the assertion that removing a subset S of nodes from G would render nodes x and y disconnected. Ideally, we would like to display independence between variables by the lack of connectivity between their corresponding nodes in some graph G. Likewise, we would like to require that finding $<x \mid S \mid y>_G$ should correspond to conditional independence between x and y given S, namely, $<x \mid S \mid y>_G \Longrightarrow I(x\ , S\ , y)_P$ and, conversely, $I(x\ , S\ , y)_P \Longrightarrow <x \mid S \mid y>_G$. 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 might be incompatible; there might exist no way to display *all* the dependencies and independencies embodied in P by vertex separation in a graph.

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

$$I(x\ , S\ , y)_P \Longrightarrow <x \mid S \mid y>_G \qquad (7)$$

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

$$I(x\ , S\ ,y)_P \Longleftarrow <x \mid S \mid y>_G \qquad (8)$$

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 graph are trivial D-maps, while complete graphs are trivial I-maps.

Given an arbitrary graph G, the theory of *Markov Fields* [Lauritzen 1982] tells us how to construct a probabilistic model P for which G is both a D-map and an I-map. We now ask whether the converse construction is possible.

Lemma: There are probability distributions for which no graph can be both a D-map and an I-map.

Proof: Graph separation always satisfies $<x \mid S_1 \mid y>_G \Longrightarrow <x \mid S_1 \cup S_2 \mid y>_G$ for any two subsets S_1 and S_2 of vertices. Some P's, however, may induce both $I(x\ , S_1\ , y)_P$ and *NOT* $I(x, S_1 \cup S_2\ , y)_P$. Such P's cannot have a graph representation which is both an I-map and a D-map because D-mapness forces G to display S_1 as a cutset separating x and y, while I-mapness prevents $S_1 \cup S_2$ from separating x and y. No graph can satisfy these two requirements simultaneously. Q.E.D.

A simple example illustrating the conditions of the proof 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 are mutually independent, i.e., $S_1=\emptyset$. However, if we notice the bell (S_2), then learning the outcome of one coin should change our opinion about the other coin. The only I-map for this example is a complete graph on the three variables involved. It is obviously not a D-map because it fails to display the basic independence of the coin outcomes.

Being unable to provide a graphical description for *all* independencies, we settle for the following compromise: we will consider only I-maps but will insist that the graphs in those maps capture as many of P's independencies as possible, i.e., they should contain no superfluous edges.

Definition: A graph G is a *minimal I-map of P* if no edge of G can be deleted with destroying its I-mapness. We call such a graph a *Markov-Net* of P.

Theorem 2: Every P has a (unique) minimal I-map $G_O = (U, E_O)$ produced by connecting *only* pairs (α,β) for which $I(\alpha\ , U-\alpha-\beta\ , \beta)_P$ *is FALSE.* (8)

i.e.,

$$(\alpha,\beta) \notin E_O \quad \textit{iff}\ I(\alpha\ , U-\alpha-\beta\ , \beta)_P \qquad (9)$$

The proof is given in [Pearl and Paz 1985] and uses only the symmetry and exchange properties of I.

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

$$I(\alpha, S, U-S-\alpha) \text{ and } \alpha \notin S \qquad (10)$$

Let $R_I^*(\alpha)$ stand for the set of all relevance spheres 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 is 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 (10).

Theorem 3: [Pearl and Paz 1985] Every variable $\alpha \in U$ has a unique *relevance boundary* $B_I(\alpha)$. $B_I(\alpha)$ coincides with the set of vertices $B_{G_0}(\alpha)$ adjacent to α in the Markov net G_0. The proof of Theorem 3 makes use of the expansion property (6.d).

Corollary 1: The set of relevance 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 relevance 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 $B_{G_O}(\alpha)$ guarantees that many independence relationships can be validated by tests for graph separation at the knowledge level itself [Pearl 1985].

Thus we see that the major graphical properties of probabilistic independencies are consequences of the exchange and expansion 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.

Illustration 1 (abstract)

To illustrate the role of these axioms, consider a set of four integers $U = \{(1, 2, 3, 4)\}$, and let I be the set of twelve triplets listed below:

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

It is easy to see that I satisfies (6.a)-(6.d) and thus it has a unique minimal I-map G_0, shown in Figure 1.

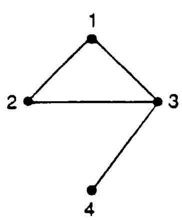

Figure 1: The Minimal I-Map, G_o, of I

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 I 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 exchange axiom (6.c). Indeed, if we try to construct G_O by the usual criterion of edge deletion, the graph in Figure 1 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 2, and the edge-minimum graph is clearly not unique.

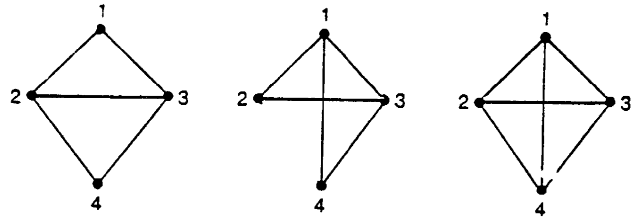

Figure 2: The Three I-Maps of I '

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 expansion axiom (6.d). Since no triple of the form $(\alpha, U-\alpha-\beta, \beta)$ appears in I'', the only I-map for this list is the complete graph. However, 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) of probabilistic dependencies. Therefore, there is no probabilistic model capable of inducing this set of independence relationships unless we also add the triplet (1, 2, 3) to I.

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 to a meeting (see Introduction, 1st paragraph). 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 Person-1's watch;
2) the time shown on Person-2's watch;
3) the correct current time;
4) the time A will show up at the meeting place;
5) the agreed time for starting the meeting;
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_O 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 values of all other variables in the system will render our belief in α sensitive to the value of β. We know, for example, that 7 will depend on 4, no matter what values are assumed by all the other variables and, on that basis, we may connect node 7 to node 4 and, proceeding in that fashion through all pairs of variables, the graph of Figure 3 may be constructed.

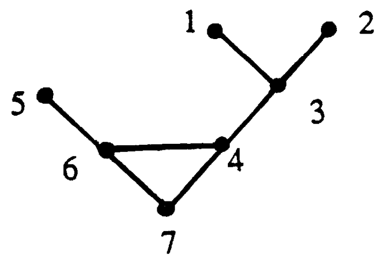

Figure 3

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 commonsense task, for instance, to decide that, once we know the current time (3), no other variable may affect what we expect to read on Person-1's watch (1). Similarly, once we know the current time (3) and that we are not about to be late (7), we still must know when our partner will actually show up (6) before we can estimate our arrival time (4) independent of the agreed time (5). On the basis of these considerations, we may connect 1 to 3; 4 to 6, 7 and 3; and so on. After finding the immediate neighbors of any six variables in the system, the graph G_O will emerge, identical to that of Figure 3.

Once established, G_O can be used as an inference instrument. For example, the fact that knowing 4 renders 2 independent of 5 (i.e., $I(2,4,5)$) can be inferred from the fact that 4 is a cutset in G_O, separating 2 from 5. 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 only ascertain that the relevance boundaries identified by the knowledge provider (e.g., the expert) form a neighbor system.

4. CONCLUSIONS

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), exchange (6.c), expansion (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_O 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, in 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_O, then we can correctly deduce independence relationships by testing whether the set of currently known propositions constitutes a cutset in G_O.

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 permit 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) and *belief nets* (Pearl, 1985) are finding a wider application in reasoning systems. A systematic axiomatization of these graphical representations is currently under way.

REFERENCES

Doyle, J. (1979), "A Truth Maintenance System," *Artificial Intelligence*, vol. 12, no. 3.

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

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

Pearl, J. (1985), "Fusion, Propagation and Structuring in Belief Networks," UCLA *CSD Technical Report* #850022, Los Angeles, CA., June 1985; to be published in *Artificial Intelligence*, fall 1986.

Pearl, J., and Paz, A., (1985), "Graphoids: A Graph-based Logic for Reasoning about Relevance Relations," UCLA *CSD Technical Report* #850038, Los Angeles, CA., December 1985.

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