FORMAL PROPERTIES OF PROBABILISTIC DEPENDENCIES AND THEIR GRAPHICAL REPRESENTATIONS

Judea Pearl
Thomas Verma

October 1986
CSD-860019

FORMAL PROPERTIES OF PROBABILISTIC DEPENDENCIES AND THEIR GRAPHICAL REPRESENTATIONS† *

(extended abstract)

Judea Pearl & Thomas Verma
Cognitive Systems Laboratory
Computer Science Department
University of California
Los Angeles, California 90024

1. INTRODUCTION

The primary focus of theoretical studies of database dependencies has, thus far, been on the formal characterization of dependencies in *categorical* relations, i.e., relations defined by categorical exclusion of subsets of tuples. Statistical databases are often of an entirely different nature: every tuple (representing some combination of events) may, in principle, be part of the database with some small, but non-zero, probability. This paper extends the scope of previous work by investigating the formal properties of probabilistic dependencies of the type "x is independent of y, given z." It is the existence of such dependencies that enables us to encode statistical databases in computationally manageable and conceptually meaningful representational schemes. Among such schemes, the most notable are graphical representations employing either undirected graphs (e.g., Markov fields [Darroch, 1980]) or directed, acyclic graphs (e.g., causal models [Blalock, 1971], inference networks [Duda, Hart and Nilsson, 1978]).

The formal basis underlying the correspondence between probabilistic dependencies and undirected graphs is described in [Pearl and Paz 1985], while this paper focuses on directed-acyclic graphs (dia-graphs). We define a criterion for detecting conditional independencies in dia-graphs and show that any minimal dia-graph constructed recursively from some probability distribution P never displays a conditional-independence relationship which is not in P. The paper then defines the class of dependencies capturable by dia-graphs and compares it to that capturable by undirected graphs. Allowing the introduction of auxiliary vertices, dia-graphs are shown to be more expressive than undirected graphs, i.e., they are capable of displaying a wider variety of probabilistic dependencies.

2. NOMENCLATURE AND OVERVIEW OF PREVIOUS WORK

2.1 Probabilistic Dependencies

Definition: Let $U = \{\alpha, \beta, \ldots\}$ be a finite set of discrete 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

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

i.e., 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

† *Topics:* Dependency Theory, Statistical Databases

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

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

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 amounts to relations that admit all tuples from the power set of U with varying weights (often negligibly small).

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,

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

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

$$I(x, \varnothing, y)_P \text{ iff } P(x,y) = P(x)\, P(y)$$

Theorem 1: [Pearl & Paz, 1985] 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	$I(x,z,y) \Longleftrightarrow I(y,z,x)$	(4.a)
Decomposition	$I(x,z,yw) \Rightarrow I(x,z,y)\ \&\ I(x,z,w)$	(4.b)
Intersection	$I(x,zw,y)\ \&\ I(x,zy,w) \Rightarrow I(x,z,yw)$	(4.c)
Weak Union	$I(x,z,yw) \Rightarrow I(x,zw,y)$	(4.d)
Contraction	$I(x,z \cup y, w)\ \&\ I(x,z,y) \Rightarrow I(x,z,yw)$	(4.e)

Remarks: The concatenation yw stands for the *conjunction* of events asserted by instantiating the set union $y \cup w$. Once I is defined on the set of disjoint triplets x, y, z it is also defined on the set of all triplets via $I(x,z,y) \Longleftrightarrow I(x-z, z, y)$

The intuitive interpretation of Eqs. (4.c) through (4.e) follows. (4.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. (4.d) states that learning an irrelevant fact (w) cannot help another irrelevant fact (y) become relevant to x. (4.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.

Axioms (4.a) through (4.e), with the exception of (4.c), are also satisfied by *Embedded Multi-Valued Dependencies* (EMVD) [Fagin, 1977] if $I(x,z,y)$ is interpreted as $z \rightarrow\rightarrow y$ with respect to the projection on the set $x \cup y \cup z$ of attributes. Axiom (4.c) is unique to probabilistic dependencies and stems from the assumption $P(\cdot) > 0$. It guarantees that every y has a unique minimal set z such that $z \rightarrow\rightarrow y$ (see Theorem 3 below), thus rendering these dependencies amenable to graphical representation. The question of completeness remains an open problem as it is not clear whether the restriction introduced by axiom (4.c) is sufficient to render EMVD axiomatizable [Sagiv & Walecka, 1982].

2.2 Graph Separation and Conditional Independence

Let $U = \{\alpha, \beta, \ldots\}$ be a finite set of elements (e.g., propositions, variables, etc.), let x, y and z stand for three disjoint subsets of elements in U and let M be a model which assigns truth values to the 3-place predicate $I(x, z, y)$. We next define the conditions under which every dependency $I(x, z, y)_M$ induced by M can be represented as a cutset condition $<x \mid z \mid y>_G$ in some graph G, where $<x \mid z \mid y>_G$ stands for "z is a vertex cutset of G, separating x from y."

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:

$$I(x, z, y)_M \Rightarrow <x \mid z \mid y>_G \quad (5)$$

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

$$I(x, z, y)_M \Leftarrow <x \mid z \mid y>_G \quad (6)$$

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.

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

(symmetry) $$I(x, z, y) \Longleftrightarrow I(y, z, x) \quad (7.a)$$

(decomposition) $$I(x, z, yw) \Rightarrow I(x, z, y) \;\&\; I(x, z, w) \quad (7.b)$$

(intersection) $$I(x, zw, y) \;\&\; I(x, zy, w) \Rightarrow I(x, z, yw) \quad (7.c)$$

(strong union) $$I(x, z, y) \Rightarrow I(x, zw, y) \;\; \forall \; w \subseteq U \quad (7.d)$$

(transitivity) $$I(x, z, y) \Rightarrow I(x, z, \gamma) \; or \; I(\gamma, z, y) \;\; \forall \;\; \gamma \notin x \cup z \cup y \quad (7.e)$$

Remark 1: (7.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. (7.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. (7.e) is the counter-positive 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.

Remark 2: (7.c) and (7.d) imply the converse of (7.b), which makes I completely defined by the set of triplets (x, z, y) in which x and y are individual elements of U. Note, also, that the union axiom (7.d) is unconditional and, therefore, stronger than (4.d), the one required for probabilistic dependencies. It provides a simple method of constructing the unique graph G_0, which is isomorphic to I -- 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.

2.3 Markov Networks

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 probability model P has a (unique) minimal I-map $G_0 = (U, E_0)$ produced by connecting *only* pairs (α, β) for which $I(\alpha - U - \alpha - \beta)_P$ is *FALSE*, *i.e.*,

$$(\alpha, \beta) \notin E_0 \quad \textit{iff} \quad I(\alpha, U - \alpha - \beta, \beta)_P \tag{8}$$

The proof uses only the symmetry and intersection properties of I.

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

$$I(\alpha, S, U - S - \alpha) \text{ and } \alpha \notin S \tag{9}$$

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 in $R_I^*(\alpha)$. $B_I(\alpha)$ is to be interpreted as a minimal set of variables that "shields" α from the influence of all other variables.

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 4 also makes use of the weak-union property (4.d).)

Corollary 1: The Markov net G_0 can be constructed by connecting each α to all members of its Markov boundary $B_I(\alpha)$.

Corollary 2: 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.,

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

(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.

3. BAYESIAN NETWORKS: Recursive Minimal Dia-graphs

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 arrow direction allows distinction between dependencies in various contexts. For instance, if the sound of a bell is functionally determined by the outcomes of two coins, we will use the network *coin* 1 $\rightarrow$ *bell* $\leftarrow$ *coin* 2, without connecting *coin* 1 to *coin* 2. This pattern of converging arrows should be interpreted as stating that the outcomes of the two coins are marginally independent but may become dependent upon knowing the outcome of the bell (or any external evidence bearing on that outcome).

Definition: Given a probability distribution $P(x_1 \ldots x_n)$ and an ordering d on the variables, the *Bayesian network* associated with the pair (P, d) is a directed acyclic graph ("dia-graph") constructed recursively by assigning to each node X_i a set of parent nodes $S_i \subseteq \{X_1 \ldots X_{i-1}\}$, which is the smallest set satisfying the condition

$$P(x_i \mid S_i) = P(x_i \mid x_{i-1}, \ldots, x_1) \tag{10}$$

In other words, S_i is the Markov boundary of X_i relative to the set $U_{(i)} = \{X_1, X_2, \ldots, X_i\}$ of variables. 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.

3.1 Dia-Graph Separation and Conditional Independence

In Markov nets, the correspondence between dependencies and the topology of the network was based on a simple graph separation criterion. In Bayes networks, it is based on a slightly more complex criterion of separation, 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, $X_1 \rightarrow X_2 \leftarrow X_3$

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 graphical criterion for testing conditional independence, in dia-graphs:

Theorem 5: [Verma, 1986] Let G_d be a dia-graph recursively 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 5 uses the contraction axiom (4.e).

One would normally expect that the introduction of directionality into the language of graphs would render them more expressive, capable of portraying a more refined set of dependencies, e.g., non-monotonic and non-transitive. Thus, it is natural to ask:

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 with edges (AB), (AC), (CD) and (BD) asserts the two independencies: $I(A, BC, D)$ 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. This limitation will always be encountered in nonchordal graphs (Tarjan & Yannakakis, 1984); 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. Despite these limitations, 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-graphs dependencies which clearly indicates where they differ from those of undirected graphs as well as probabilistic dependencies.

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 6: 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	$I(x, z, y) \iff I(y, z, x)$	(11.a)
Composition - Decomposition	$I(x, z, yw) \iff I(x, z, y) \,\&\, I(x, z, w)$	(11.b)
Intersection	$I(x, zw, y) \,\&\, I(x, zy, w) \Rightarrow I(x, z, yw)$	(11.c)
Weak Union	$I(x, z, yw) \Rightarrow I(x, zw, y)$	(11.d)
Contraction	$I(x, zy, w) \,\&\, I(x, z, y) \Rightarrow I(x, z, yw)$	(11.e)
Weak Transitivity	$I(x, z, y) \,\&\, I(x, z\gamma, y) \Rightarrow I(x, z, \gamma)$ *or* $I(\gamma, z, y)$	(11.f)
Chordality	$I(x, zw, y) \,\&\, I(z, xy, w) \Rightarrow I(x, z, y)$ *or* $I(x, w, y)$	(11.g)

Remarks: Axioms (11.a) and (11.c-e) are identical to those governing probabilistic dependencies (Eq. (4)). The left implication of (11.b) and the last two axioms, namely, composition, weak-transitivity and chordality, represent additional constraints over the system of Eq.(4). 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 (11.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 (11) to the axioms defining separation in undirected graphs (7), we note that (7) implies all axioms in (11) except chordality (11.g). In particular, weak-union is implied by strong union, composition and contraction are implied by (7.c) and (7.d) and, of course, weak transitivity is implied by transitivity (7.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 remains in effect when we associate independence with separation in dia-graphs but may be violated in some probability models which may allow for the co-occurrence of four conditions:
1. $I(x, \emptyset, y)_P$ 2. $I(x, \gamma, y)_P$ 3. $\neg I(x, \emptyset, \gamma)_P$ 4. $\neg I(y, \emptyset, \gamma)_P$

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. However, it can be shown that, if all variables in U are binary, then all probabilistic dependencies must be weakly transitive.

The purpose of the chordality axiom (11.f) is to exclude dependence models whose Markov nets are non-chordal since these cannot be completely captured by dia-graphs. Non-chordal graphs represent the one class of dependencies where undirected graphs exhibit expressiveness superior to that of dia-graphs. However, this superiority can be eliminated by the introduction of auxiliary variables.

Consider the diamond-shaped graph of Figure 1(a), which asserts the two independence relationships: $I(A, BC, D)$ and $I(B, AD, C)$.

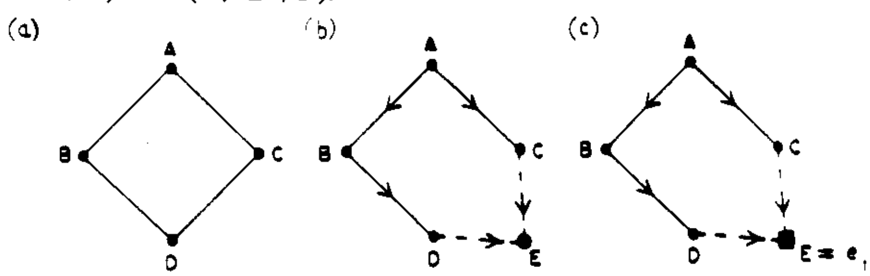

Figure 1

Introducing an auxiliary variable E in the manner shown in Figure 1(b) creates a dia-graph model on five variables whose dependencies are represented by the joint distribution function,

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

If we "clamp" the auxiliary variable E at some fixed value $E = e_1$, as in Figure 1(c), the dependency structure projected on A, B, C, D is identical to the original structure of Figure 1(a), i.e., $I(A, BC, D)$ and $I(B, AD, C)$.

In conclusion, we see that the introduction of auxiliary variables permits us to dispose of the chordality restriction of (11.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.

Another method of improving the expressive power of Bayesian networks (without introducing auxiliary nodes) is to permit both directed and undirected links but no directed cycles. Such hybrid-acyclic graphs (*ha*-graphs) can capture both non-chordal and non-weakly-transitive dependencies. Formal properties of dependencies expressible by *ha*-graphs, together with algorithms for their construction, will be presented in the full paper.

REFERENCES

H. M. Blalock, *Causal Models in the Social Sciences*, Macmillan, London, 1971.

J. N. Darroch, S. L. Lauritzen & T. P. Speed, "Markov Fields and Log-Linear Models for Contingency Tables," 1980.

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

R. Fagin, "Multivalued Dependencies and a New Form for Relational Databases," *ACM Transactions on Database Systems*, 2,3; September 1977, pp. 262-278.

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.

Y. Sagiv & S. F. Walecka, "Subset Dependencies and a Completeness Result for a Subclass of Embedded Mutltivalued Dependencies," *Journal of the Association for Computing Machinery*, Vol. 29, No. 1, January 1982, pp. 103-117.

R. E. 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.