Belief Networks Revisited *

Judea Pearl
Cognitive Systems Laboratory
Computer Science Department
University of California, Los Angeles, CA 90024
judea@cs.ucla.edu

January 6, 1994

1 Introduction

The article "Fusion Propagation and Structuring in Belief Networks" (hereafter *Fusion*) was the culmination of a series of papers (e.g., [15][12][16]) in which I advocated the restoration of probabilistic methods in AI systems and explored the possibility of representing and manipulating probabilistic knowledge in graphical forms, later called *belief networks* (also known as *Bayesian networks* and *causal diagrams*). In recent years, belief networks have become a tool of great versatility and power and are now considered the most common representation scheme for probabilistic knowledge. They have been used to aid diagnosis of medical patients and malfunctioning systems, to understand stories, to interpret pictures, to perform filtering, smoothing and prediction, to facilitate planning in uncertain environments, and to study causation, nonmonotonicity, action, change, and attention. [1]

The following is a brief personal account of the development of belief networks, both before and after the publication of *Fusion*, although space

*This work was supported in part by NSF grant IRI-9157936 and by State of California MICRO grants 91-124 and 91-125.

[1]Some of these applications are described in a recent tutorial article by Charniak [1].

permits but a sketchy account of the wealth of recent developments in this area. [2]

2 Origins

The idea of studying distributed probabilistic computations on graphical models began brewing in my mind in the late 1970s, after I read Rumelhart's paper on reading comprehension [23]. In this paper, Rumelhart presented compelling evidence that text comprehension must be a distributed process that combines both top-down and bottom-up inferences. Strangely, this dual mode of inference, so characteristic of Bayesian analysis, did not match the capabilities of either the "certainty factors" calculus or the inference networks of PROSPECTOR – the two major contenders for uncertainty management in the 1970s. I thus began to explore the possibility of achieving distributed computation in a "pure" Bayesian framework, so as not to compromise its basic capacity to combine bi-directional inferences (i.e., predictive and abductive). Not caring much about generality at that point, I picked the simplest structure I could think of (i.e., a tree) and tried to see if anything useful can be computed by assigning each variable a simple processor, forced to communicate only with its neighbors. This gave rise to the tree-propagation algorithm reported in [15] and, a year later, the Kim-Pearl algorithm [12], which supported not only bi-directional inferences but also intercausal interactions, such as "explaining-away." These two algorithms were described in Section 2 of *Fusion*.

In the course of developing these algorithms, it became clear that *conditional independence* is the most fundamental relation behind the organization of probabilistic knowledge and the most crucial factor facilitating distributed computations. I therefore decided to investigate systematically how directed and undirected graphs could be used as a language for encoding, decoding, and reasoning with such independencies. At about the same time, Howard and Matheson were studying the properties of influence diagrams [10] and were asking similar questions about graphs and dependencies, albeit from a somewhat different perspective: the links in the diagrams were treated

[2] A more complete account and an updated bibliography are provided in the revised second printing of my book *Probabilistic Reasoning in Intelligence Systems* [18].

as traces of the information that a person finds convenient to consult while assessing probabilities.

The inconclusive results of Howard and Matheson's report jolted me into trying a different approach, in which the links are designated specifically to *causal* associations. However, having found no satisfactory definition of causality in the literature, I decided to search for one myself by concentrating on the fundamental mathematical relationships that may exist between probabilities and directed acyclic graphs. I began by asking how a directed acyclic graph (dag) can be extracted from a given probability distribution, whether the extracted dag is unique, what kind of distributions can be specified by a given dag, how we can read off the independencies that are embedded in the dag, and whether they match those associated with causal organizations. This line of inquiry resulted in Section 1 of *Fusion*, in which the construction, consistency, and completeness of belief networks were demonstrated and the *d*-separation criterion was presented. Eventually, this inquiry developed into the axiomatic theory of *graphoids* [20][18][5], in which directed and undirected graphs are treated as approximate representations of abstract mathematical objects, called *dependency models*, and are interpreted and manipulated by the logic of conditional independence. [3]

3 Motivations and Speculations

Fusion was motivated by a busy mixture of observations and speculations, some of which I recall quite vividly:

1. The failure of rule-based systems to exhibit certain plausible patterns of reasoning is symptomatic of fundamental limitations, and these limitations can be overcome only by grounding automated reasoning in some safe and friendly calculus of uncertainty.

2. The consistent agreement between plausible reasoning and probability calculus could not be coincidental, but strongly suggests that human intuition invokes some crude form of probabilistic computation.

[3] *Fusion* has been criticized for "substituting mathematics for clarity" (e.g., R. E. Barlow, in [14], page 117). In my judgment, it was precisely this conversion of networks and diagrams to mathematically defined objects that led to their current acceptance in practical reasoning systems.

3. In light of the speed and effectiveness of human reasoning, the computational difficulties that plagued earlier probabilistic systems could not be very fundamental and should be overcome by making the right choice of simplifying assumptions.

4. No reasoning can take place unless our knowledge embodies many (conditional) independence assumptions, and graphical forms are the only plausible way in which these assumptions could be represented.

5. If a graphical knowledge representation could be found, then it should be possible to use the links as message-passing channels, and we could then update beliefs by parallel distributed computations, reminiscent of neural architectures.

6. If belief updating could be achieved by such distributed mechanisms, then the update would be easier to explain, since the flow of information would transverse conceptually meaningful paths.

7. If distributed updating were feasible, then probabilistic inference would be as easy to program and execute (even on a serial machine) as rule-based systems, since no timing information, hence only simple control mechanisms, would be required.

In hindsight, some of these speculations were rather naive. For example, fully distributed updating turned out to be feasible only in singly connected networks, and some conditional independence relationships were shown to defy graphical representation altogether. Nevertheless, many of these speculations have survived the test of time, as the following section reflects.

4 The Main Contributions

The key contribution of *Fusion* was the formulation and demonstration of some of the basic properties and capabilities of belief networks:

1. Graphical methods make it easy to maintain consistency and completeness in probabilistic knowledge bases. They also define modular procedures of knowledge acquisition that reduce significantly the number of

probability assessments required, [4] and they guard the model builder from assigning numerical values that lead to unintended dependencies.

2. Independencies can be dealt with explicitly. They can be articulated by an expert, encoded graphically, read off the network, and reasoned about, yet they forever remain robust to numerical imprecision. Every conditional independency embedded in the network can be recognized in linear time (using the *d*-separation rule).

3. Graphical representations uncover opportunities for efficient computation. Distributed updating is feasible in knowledge structures rich enough to exhibit intercausal interactions (e.g., "explaining away"), and, when extended by clustering or conditioning, tree-propagation algorithms are capable of updating networks of arbitrary topology.

4. The combination of predictive and abductive inferences has resolved many problems encountered by first generation expert systems and has rendered belief networks a viable model for cognitive functions requiring both top-down and bottom-up inferences.

5. Causal utterances such as "X is a direct cause of Y" were given a probabilistic interpretation as distinctive patterns of conditional independence relationships that can be verified empirically. "Hidden causes" were given operational definition and, under certain conditions, were shown to be identifiable by efficient algorithms.

5 Recent Progress

d-separation and Graphoids. In retrospect, perhaps *Fusion* made its greatest immediate impact through the introduction of the *d*-separation criterion. d-*separation* (the "*d*" denoting "directional") is a simple graphical test for deciding which conditional-independence relations are implied by a given network's topology. It provides, therefore, the semantics needed for defining and characterizing belief networks. Technically, the *d*-separation criterion has facilitated immediate solutions to three practical problems (see[19], Section

[4] A further reduction has been achieved by Heckerman's *similarity networks* [8].

4.4.): [5] (1) How to characterize precisely the set of graphical transformations (e.g., arc reversals, node removals, node collapsing) that can legitimately be performed on a network, (2) how to test whether one network is entailed by or is equivalent to another, and (3) how to delineate the minimum information needed for answering a given query. On the conceptual side, by identifying the independencies embedded in directed acyclic graphs, the d-separation criterion has also identified special patterns of independencies that are characteristics of causal organizations. These patterns have since been used to define causation and to uncover causal relationships in data [21] (see also last paragraph in this section).

Verma [28] has proved the soundness of the d-separation criterion using the semi-graphoid axioms [20], thus rendering the criterion valid for a wide class of informational dependencies, including probabilistic, graphical, correlational, and database dependencies. Geiger [5] has shown that the criterion cannot be improved; namely, d-separation reveals *all* the independencies that can be inferred from the information provided by the network builder. A more comprehensive separation criterion, applicable to networks containing deterministic nodes, was developed by Geiger, Verma, and Pearl [6] and has been shown to be testable in time proportional to the number of edges in the network.

The relation of conditional independence has received an axiomatic characterization using the theory of graphoids [20] (see also [18], Chapter 3), which provides symbolic machinery for deciding whether one independency follows from others and whether we can capture such independencies by graphs. Representations using undirected graphs (also known as Markov fields) are discussed in [18], Chapter 3, and [5]; representations using multigraphs and annotated graphs have been developed by Geva and Paz ([7].

Network Updating Techniques. Since the publication of *Fusion*, many techniques for updating belief networks have been developed and refined. Among the most popular are Shachter's method of node elimination [24], Lauritzen and Spiegelhalter's method of graph-triangulation and clique-tree propagation [13], and the method of loop-cut conditioning (*Fusion* Section 2.4. While the task of computing probabilities in general networks is

[5] Specific aspects of these problems (e.g., arc reversals [24]) had been worked out in the literature on influence diagrams, but the general problems remained unsettled until quite recently [25].

NP-hard [22] [2], the complexity of the first two methods is exponential in the size of the largest clique found in some triangulation of the network. The third method might yield a higher complexity in some networks, but it is convenient in networks with a few long loops. It is fortunate that these complexities may be estimated prior to actual processing, because, when the estimates exceed reasonable bounds, we can switch to an approximation method such as stochastic simulation [9][17]. Statistical techniques have also been developed for systematic updating of the conditional probabilities annotating the network so as to achieve a better match with past empirical data [26]. The preprocessing method of tree-decomposition with hidden variables (*Fusion*, Section 3) is still not well developed.

Causal Discovery. One of the most exciting prospects in recent years has been the possibility of using belief networks to discover causal relationships in raw statistical data. Technically, the probabilistic semantics that belief networks attribute to the links and their orientations has rendered this prospect feasible, and several systems have been developed for this purpose.

Pearl and Verma [21] have developed a probabilistic account of causation based on minimal-model semantics, [6] This theory provides criteria for identifying genuine and spurious causes, with and without temporal information, and yields algorithms for recovering causal networks with hidden variables from statistical data. A fast algorithm for recovering sparse networks is described by Spirtes and Glymour [27], and Bayesian methods of computing the "probability that X is a cause for Y" were developed by Cooper and Herskovits [3]. Causal reorganization of categorical databases is studied in [4]. In addition to their likely impact on the practice of building knowledge systems, these developments also promise finally to give causation a purely empirical semantics – the illusive goal of many philosophers and statisticians since the time of Hume.

6 Regrets and Near Misses

One regrettable step in *Fusion* was my betting on what turned out to be the less attractive way of extending tree-propagation to multiply connected net-

[6]In this semantics, a variable X is said to have a causal influence on a variable Y if there is a directed path from X to Y in all minimal causal networks (dags) consistent with the data.

works. I speculated that the loop-cut conditioning method would be more efficient than the one I labeled "compounding," that is, forming clusters of compound variables that are tree structured and applying the tree-propagation algorithm to the resulting tree. Lauritzen and Spiegelhalter [13], and later Jensen et al. [11], have perfected this tree-clustering method to the point that it is now the most widely used algorithm in practical applications. The popularity of the tree-clustering method stems from its inheriting the distributed character, and hence robustness and versatility of the basic tree-propagation algorithm, as described in *Fusion* (Section 2.1).[7] Thus, my regrets are somewhat mitigated by the realization that concentrating all my initial efforts on trees and polytrees did yield some useful insights.

Finally, to the many readers intrigued by the lengthy review process for *Fusion* (*Received January 1982; revised version received February 1986.*): Yes, it indeed took four years to get the article accepted, but the reviewers were not at fault. The article simply got lost (literally!) twice, which was not entirely without virtue; each time the editor asked me to replace a lost copy, I would seize the opportunity and send an improved version. I hope the final outcome was worth the wait.

References

[1] E. Charniak, Bayesian networks without tears, *AI Magazine* 12 (4) (1991) 50-63.

[2] G.F. Cooper, Computational complexity of probabilistic inference using Bayesian belief networks (research note), *Artificial Intelligence* 42 (2) (1990) 393-405.

[3] G. F. Cooper and E. Herskovits, A Bayesian method for constructing Bayesian belief networks from databases, in: *Proceedings of the Conference on Uncertainty in AI* (Morgan Kaufmann, San Mateo, CA) (1990) 86-94.

[7] Structurally, the two algorithms are essentially the same; the one described in *Fusion* propagates messages in a tree of singletons, whereas Lauritzen and Spiegelhalter's algorithm propagates these messages in a tree of compound variables known as a *join tree* ([18], pp. 111-113) or *junction tree* [11].

[4] R. Dechter and J. Pearl, Directed constraint networks: A relational framework for casual modeling, in: *Proceedings of the 12th International Joint Conference of Artificial Intelligence (IJCAI-91)* Sydney, Australia (Morgan Kaufmann, San Mateo, CA) (1991) 1164-1170.

[5] D. Geiger, Graphoids: A qualitative framework for probabilistic inference, Ph.D. dissertation, University of California, Los Angeles, CA (1990).

[6] D. Geiger, T.S. Verma and J. Pearl, Identifying independence in Bayesian networks, *Networks* 20 (5) (1990) 507-534.

[7] R.Y. Geva and A. Paz, Towards complete representation of graphoids in graphs, in: *Proceedings of the 15th International Workshop on Graph Theoretic Concepts in Computer Sciences*, Rodluc (Springer-Verlag, New York) (1989) 41-62.

[8] D. Heckerman, Probabilistic similarity networks, *Networks* 20 (5) (year) 607-636. Also MIT Press (1991).

[9] M. Henrion, Propagation of uncertainty by probabilistic logic sampling in Bayes' networks, in: J.F. Lemmer and L.N. Kanal, eds., *Uncertainty in Artificial Intelligence 2* (Elsevier Science Publishers/North-Holland, Amsterdam, Netherlands, 1988) 149-164.

[10] R.A. Howard and J.E. Matheson, Influence diagrams, in: R.A. Howard and J.E. Matheson, eds., *The Principles and Applications of Decision Analysis Vol. 2* (Strategic Decisions Group, Menlo Park, CA, 1984) 721-762.

[11] F.V. Jensen, K.G. Olsen and S.K. Andersen, An algebra of Bayesian belief universes for knowledge-based systems, *Networks* 20 (5) (1990) 637-660.

[12] J.H. Kim and J. Pearl, A computational model for combined causal and diagnostic reasoning in inference systems, in: *Proceedings IJCAI-83*, Karlsruhe, Germany (1983) 190-193.

[13] S.L. Lauritzen and D.J. Spiegelhalter, Local computations with probabilities on graphical structures and their application to expert systems (with discussion), *Journal Royal Statistical Society*, Series B 50 (2) (1988) 157-224.

[14] R.M. Oliver and J.Q. Smith, eds., *Influence Diagrams, Belief Nets and Decision Analysis* (John Wiley, Rexdale, Ontario, Canada, 1990).

[15] J. Pearl, Reverend Bayes on inference engines: A distributed hierarchical approach, in: *Proceedings AAAI National Conference on AI*, Pittsburgh, PA (1982) 133-136.

[16] J. Pearl, How to do with probabilities what people say you can't, in: *Proceedings 2nd IEEE Conference on AI Applications*, Miami, FL (1985) 6-12.

[17] J. Pearl, Evidential reasoning using stochastic simulation of causal models, *Artificial Intelligence* 32 (2) (1987) 245-258.

[18] J. Pearl, *Probabilistic Reasoning in Intelligent Systems*, Morgan Kaufmann, Palo Alto, CA, 1988). Revised second printing (1991).

[19] J. Pearl, D. Geiger and T.S. Verma, The logic of influence diagrams, in: [14] 67-88.

[20] J. Pearl and A. Paz, 1989. On the logic of representing dependencies by graphs, in: *Proceedings 1986 Canadian AI Conference*, Montreal, Ontario, Canada (1986) 94-98.

[21] J. Pearl and T. Verma, A theory of inferred causation, in: J.A. Allen, R. Fikes and E. Sandewall, eds., *Principles of Knowledge Representation and Reasoning: Proceeding of the Second International Conference* Morgan Kaufmann, San Mateo, CA) (1991) 441-452.

[22] A. Rosenthal, A computer scientist looks at reliability computations, in: Barlow et al., eds., *Reliability and Fault Tree Analysis* (SIAM, Philadelphia, 1975) 133-152.

[23] D.E. Rumelhart, Toward an interactive model of reading, Tech. Report #CHIP-56, University of California, La Jolla, CA (1976).

[24] R.D. Shachter, Evaluating influence diagrams, *Operations Research* 34 (6) (1986) 871-882.

[25] J.Q. Smith, Influence diagrams for statistical modeling, *Annals of Statistics* 17 (2) (1989) 564-572.

[26] D.J. Spiegelhalter and S.L. Lauritzen, Sequential updating of conditional probabilities on directed graphical structures, *Networks* 20 (5) (1990) 579-605.

[27] P. Spirtes and C. Glymour, An algorithm for fast recovery of sparse causal graphs, *Social Science Computer Review* 9 (1) (1991) 62-72.

[28] T. Verma, Causal networks: semantics and expressiveness, Tech. Report #R-65, Cognitive Systems Laboratory, University of California, Los Angeles, CA (1986). Also in: *Proceedings of the 4th Workshop on Uncertainty in Artificial Intelligence*, Minneapolis, MN (Advanced Decision Systems, Mountain View, CA) (1988) 352-359.