BAYESIAN NETWORKS, CAUSAL INFERENCE AND KNOWLEDGE DISCOVERY *

Judea Pearl
Computer Science Departmen
University of California, Los Angeles, CA 90095
judea@cs.ucla.edu

March 5, 2001

Networks carrying probabilistic and causal information have a long and rich tradition, which began with the geneticist Sewall Wright (1921). Variants have appeared in many fields; within social science and economics, such models (usually linear) are known as Path Diagrams or Structural Equations Models (SEM), and in artificial intelligence, such models (usually nonlinear) are known as Bayesian networks. The capabilities for bidirectional inferences (e.g., prediction and diagnosis), quick debugging and reconfiguring, combined with a rigorous probabilistic foundation, led to the rapid emergence of Bayesian networks as the method of choice for uncertain reasoning in AI and expert systems, replacing earlier, *ad hoc* rule-based schemes [Pearl, 1988, Heckerman *et al.*, 1995, Jensen, 1996].

The nodes in a Bayesian network represent variables of interest (e.g., the temperature of a device, the gender of a patient, the price of a product, the occurrence of an event) and the links represent informational or causal dependencies among the variables. The dependencies are quantified by conditional probabilities for each node given its parents in the network. The network supports the computation of the probabilities of any subset of variables given

*Portions of this paper are based on (Pearl and Russell, 2001) evidence about any other subset.

Figure 1 illustrates a simple yet typical Bayesian network. It describes the causal relationships among the season of the year (X_1), whether it's raining (X_2), whether the sprinkler is on (X_3), whether the pavement is wet (X_4), and whether the pavement is slippery (X_5). Here, the absence of a direct link between X_1 and X_5, for example, captures our understanding that there is no direct influence of season on slipperiness—the influence is mediated by the wetness of the pavement. (If freezing is a possibility, then a direct link could be added.)

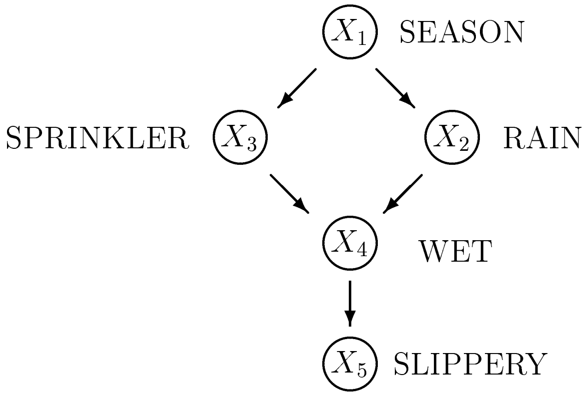

Figure 1: A Bayesian network representing causal influences among five variables.

Perhaps the most important aspect of Bayesian networks is that *they are direct representations of the world, not of reasoning processes.* The arrows in the diagram represent real causal connections and not the flow of information during reasoning (as in rule-based systems and neural networks). Inferences can be derived from Bayesian networks by propagating information in any direction. For example, if the sprinkler is on, then the pavement is probably wet (prediction); if someone slips on the pavement, that also provides evidence that it is wet (abduction). On the other hand, if we see that the pavement is wet, that makes it more likely that the sprinkler is on or that it is raining (abduction); but if we then observe that the sprinkler is on, that reduces the likelihood that it is raining (explaining away). It is this last form of reasoning, explaining away, that is especially difficult to model in rule-based systems and neural networks in any natural way.

Probabilistic interpretation.

Any complete probabilistic model of a domain must, either explicitly or implicitly, represent the *joint distribution*—the probability of every possible event as defined by the values of all the variables. There are exponentially many such events, yet Bayesian networks achieve compactness by factoring the joint distribution into local, conditional distributions for each variable given its parents. If x_i denotes some value of the variable X_i and pa_i denotes some set of values for X_i's parents, then $P(x_i|pa_i)$ denotes this conditional distribution. For example, $P(x_4|x_2, x_3)$ is the probability of wetness given the values of sprinkler and rain. The *global interpretation* of Bayesian networks specifies that the full joint distribution is given by the product

$$P(x_1, ..., x_n) = \prod_i P(x_i \mid pa_i) \tag{1}$$

In our example network, we have

$$P(x_1, x_2, x_3, x_4, x_5) = P(x_1)\ P(x_2|x_1)\ P(x_3|x_1)\ P(x_4|x_2, x_3)\ P(x_5|x_4) \tag{2}$$

Provided the number of parents of each node is bounded, it is easy to see that the number of parameters required grows only linearly with the size of the network, whereas the joint distribution itself grows exponentially. Further savings can be achieved using compact parametric representations—such as noisy-OR models, decision trees, or neural networks— for the local conditional distributions.

The validity of the factorization in Equation 1 rests on a set of *local* independence assumptions, asserting that each variable is independent of its nondescendants in the network given its parents. For example, the parents of X_4 in Figure 1 are X_2 and X_3 and they render X_4 independent of the remaining nondescendant, X_1. That is,

$$P(x_4|x_1, x_2, x_3) = P(x_4|x_2, x_3)$$

The collection of independence assertions formed in this way suffices to derive the factorization in Equation 1, and vice versa. These independencies are most useful in *constructing* Bayesian networks from human experts, because selecting as parents *all* the direct causes of a given variable invariably satisfies the local conditional independence conditions [Pearl, 2000, p.30] . These independencies also lead directly to a variety of algorithms for reasoning.

Evidential reasoning.

From the product specification in Equation 1, one can express the probability of any desired proposition in terms of the conditional probabilities specified in the network. For example, the probability that the sprinkler is on, given that the pavement is slippery, is

$$P(X_3 = on | X_5 = true) = \frac{P(X_3 = on, X_5 = true)}{P(X_5 = true)} \tag{3}$$

Both the numerator and denominator can be computed from Eq. (2) by summing over the rest of the variables. These summations can often be simplified in ways that reflect the structure of the network itself, and many algorithms have been developed, both exact and approximate, to perform these probabilistic calculations [Pearl, 1988, Lauritzen and Spiegelhalter, 1988; Zhang and Poole, 1996 Pearl, 1987, Jodan *et al* 1998]. The most appealing ones use distributed, message-passing schemes along the links of the original network [Pearl, 1988 (p. 235), Mackay *et al.*, 1998]

Causal networks.

All probabilistic models, no matter how refined and acccurate, describe a distribution over possible observed events—as in Eq. 1—but say nothing about what will happen if a certain *intervention* occurs. For example, what if I *turn the sprinkler on*? What effect does that have on the season, or on the connection between wetness and slipperiness? A *causal network* is a Bayesian network with the added property that the parents of each node are its direct causes—as in Figure 1. In such a network, the result of an intervention is obvious: the sprinkler node is set to $X_3 = on$ and the causal link between the season X_1 and the sprinkler X_3 is removed. All other causal links and conditional probabilities remain intact, so the new model is[1]

$$P(x_1, x_2, x_4, x_5) = P(x_1)\ P(x_2|x_1)\ P(x_4|x_2, X_3 = on)\ P(x_5|x_4)$$

Causal networks are defined, then, as oracles for interventions; the correct probability model after intervening to fix any node's value is given simply by deleting links from the node's

[1]Notice that this differs from *observing* that $X_3 = on$, which would result in a new model that included the term $P(X_3 = on|x_1)$. This mirrors the difference between seeing and doing: after observing that the sprinkler is on, we wish to infer that the season is dry, that it probably did not rain, and so on; an arbitrary decision to turn the sprinkler on should not result in any such beliefs.

parents. For example, $Fire \longrightarrow Smoke$ is a causal network whereas $Smoke \longrightarrow Fire$ is not, even though both networks are equally capable of representing any joint distribution on the two variables. Causal networks model the environment as a modular collection of stable mechanisms. These mechanisms may be reconfigured locally by interventions, with correspondingly local changes in the model. This, in turn, allows causal networks to be used very naturally for prediction by an agent that is considering various courses of action [Pearl, 1993b, 2000]

Causal structures and knowledge mining.

Many statistical routines are currently being developed under the enterprises of "knowledge mining" or "knowledge discovery," but none deserves this fancy title, because "knowledge" connotes stable relationships, invariant to local interventions and transportable across contexts – statistical routines are blind to considerations of stability. The general attitude is that statistical associations alone would be sufficient in prediction tasks that involve no manipulation.

This attitude is short sighted. First, black-box predictions are not as useful as those that are accompanied with causal understanding of the underlying processes. For example, when a statistical package predicts that customers who purchased product A are likely to purchase a product B in the future, the question always arises whether the association discovered is long-lived, and whether it is transportable across contexts. If one product is functionally supplementary to another, the association between the two demands is stable. If, on the other hand, demands for products A and B are correlated merely because the two were advertised simultaneously in the same medium, the association is short lived, and will disappear as soon as advertising strategies change.

Second, models are rarely used exclusively for passive predictions. Using an e-commerce example again, vendors constantly try new techniques of presentation, and new methods of capturing users' attention. These changes are the commercial analogue of scientific experimentation, and only causal models can capture the results of these experiments so as to predict response to future changes.

Finally, even purely predictive tasks can benefit from the modularity inherent in causal models. When some conditions in the environment undergo change, it is usually only a few causal mechanisms that are affected by the change; the rest remain unaltered. It is simpler and more effective, then, to reassess (judgmentally) or reestimate (statistically) the model parameters knowing that the corresponding change in the model is also local, involving just a few parameters, than to reestimate the entire model from scratch. In non-causal systems, such as neural nets or those based on regression equations, a local change in mechanism space would spread its effect over all model parameters, and that normally requires a major effort of re-estimation or re-training.

Where does the structure come from?

In many applications, users of statistical methods possess valuable theoretical and professional knowledge (say, that symptoms do not cause diseases) that permits one to combine causal and statistical information effectively – the human expert provides the qualitative causal structure (depicted by the diagram) and the data provides the basis for assessing the strengths of the causal connections. This symbiosis was in fact the motivating paradigm behind econometric modeling, before it went into hiding [2]. However, there have been two major (mental) barriers for implementing this symbiosis: (1) Investigators (especially statisticians) are reluctant to state causal information explicitly, because such information cannot be tested directly in nonexperimental data (2) Causal information, even when tested, cannot be expressed in the standard vocabulary of probability calculus. The second barrier, to my view, far outweighs the first, and the development of new mathematical tools for causation, both algebraic and graphical, now promises to reinstate causal modeling to its proper place in data interpretation and knowledge mining.

Learning in Bayesian networks.

Given a causal structure, the conditional probabilities $P(x_i|pa_i)$ can be updated continuously from observational data using gradient-based or EM methods [Lauritzen, 1995, Binder *et al*, 1997] as weights are adjusted in neural networks. When hidden variables are involved, some of the the conditional probabilities may not be identifiable, yet, even in such cases, many

[2]Most econometric texts in the past decade have refrained from defining what an economic model is, and those that attempt a definition, erroneously view models as compact representations of density functions (see Pearl, 2000, pp. 135-138)

causal quantities (e.g., total and direct effects) can still be assessed consistently from the data (Pearl, 2000; Chapters 3 and 4).

Causal discovery.

One of the most exciting prospects in recent years has been the possibility of using Bayesian networks to discover causal structures in raw statistical data [Pearl and Verma 1991, Spirtes *et al*, 1993, Pearl, 2000] previously considered impossible without controlled experiments. Consider, for example, the following *intransitive* pattern of dependencies among three events: A and B are dependent, B and C are dependent, yet A and C are independent. If you ask a person to supply an example of three such events, the example would invariably portray A and C as two independent causes and B as their common effect, namely, $A \rightarrow B \leftarrow C$. (For instance, A and C could be the outcomes of two fair coins, and B represents a bell that rings whenever either coin comes up heads.) Fitting this dependence pattern with a scenario in which B is the cause and A and C are the effects is mathematically feasible but very unnatural, because it must entail fine tuning of the probabilities involved; the desired dependence pattern will be destroyed as soon as the probabilities undergo a slight change.

Such thought experiments tell us that certain patterns of dependency, which are totally void of temporal information, are conceptually characteristic of certain causal directionalities and not others. When put together systematically, such patterns can be used to infer causal structures from raw data and to guarantee that any alternative structure compatible with the data must be less stable than the one(s) inferred; namely, slight fluctuations in parameters will enventually render that structure incompatible with the data. Using this mild assumption of stability, methods were developed for identifying genuine and spurious causes, with or without temporal information (Spirtes et al, 1993; Pearl, 2000; Chapter 2).

Alternative methods of identifying structure in data assign prior probabilities to the parameters of the network and use Bayesian updating to score the degree to which a given network fits the data [Cooper and Herskovitz, 1990, Heckerman *et al.*, 1994]. Likewise, one can trade off network complexity against degree of fit to the data [Friedman, 1998]. These methods have the advantage of operating well under small sample conditions, but encounter difficulties coping with hidden variables.

References

J. Binder, D. Koller, S. Russell, and K. Kanazawa. Adaptive probabilistic networks with hidden variables. *Machine Learning*, 29:213–244, 1997.

G.F. Cooper and E. Herskovits. A Bayesian method for constructing Bayesian belief networks from databases. *Proceedings of the Conference on Uncertainty in AI*, pages 86–94, 1990.

D. Heckerman, A. Mamdani, and M.P. Wellman (Guest Editors) Real-world applications of Bayesian networks. *Communications of the ACM*, 38(3):24–68, March 1995.

D. Heckerman, D. Geiger, and D. Chickering Learning Bayesian networks: The combination of knowledge and statistical data. *Uncertainty in Artificial Intelligence*, 10:293-301, 1994

F.V. Jensen. *An Introduction to Bayesian Networks.* Springer, New York, 1996.

N. Friedman. The Bayesian structural em algorithm. In G.F. Cooper and S. Moral, editors, *Uncertainty in Artificial Intelligence: Proceedings of the Fourteenth Conference*, pages 129–138, Madison, Wisconsin, 1998. Morgan Kaufmann.

S.L. Lauritzen. The EM algorithm for graphical association models with missing data. *Computational Statistics and Data Analysis*, 19:191–201, 1995.

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

J. Pearl. Evidential reasoning using stochastic simulation of causal models. *Artificial Intelligence*, 32(2):245–258, 1987.

J. Pearl. *Probabilistic Reasoning in Intelligent Systems.* Morgan Kaufmann, San Mateo, CA, 1988.

J. Pearl. Comment: Graphical models, causality and intervention. *Statistical Science*, (8):pages 266–9.

J. Pearl *Causality* Cambridge University Press, New York, NY, 2000.

J. Pearl, and S. Russell, Bayesian Networks, In M. Arbib (Ed.), *Handbook of Brain Theory and Neural Networks*, MIT Press, second edition, forthcoming, 2001.

J. Pearl and T. Verma. A theory of inferred causation. In J.A. Allen, R. Fikes, and E. Sandewall, editors, *Principles of Knowledge Representation and Reasoning: Proceedings of the Second International Conference*, pages 441–452. Morgan Kaufmann, San Mateo, CA, 1991.

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

P. Spirtes, C. Glymour, and R. Schienes. *Causation, Prediction, and Search.* Springer-Verlag, New York, 1993.

Related Links to Bayesian networks

UCLA: http://www.cs.ucla.edu/ judea/

UCLA: http://www.cs.ucla.edu/ darwiche/cs262a/

Stanford: http://www.stanford.edu/class/cs228/

http://www.cs.berkeley.edu/ murphyk/Bayes/bayes.html

http://www.cs.berkeley.edu/ murphyk/pomdp.html

Harvard: http://deas.harvard.edu/courses/cs281r/

Duke: http://www.stat.duke.edu/courses/Spring99/sta294/

UC Irvine: http://www.ics.uci.edu/ dechter/275B.html