In Michael A. Arbib (Ed.), *The Handbook of Brain Theory and Neural Networks*, Second Edition,
Cambridge, MA: The MIT Press, 157-160, 2003.

TECHNICAL REPORT
R-216

Bayesian Networks

Judea Pearl and Stuart Russell

Introduction

Probabilistic models based on directed acyclic graphs have a long and rich tradition, beginning with work by the geneticist Sewall Wright in the 1920s. Variants have appeared in many fields. Within statistics, such models are known as *directed graphical models*; within cognitive science and artificial intelligence (AI), they are known as *Bayesian networks*. The name honors the Reverend Thomas Bayes (1702–1761), whose rule for updating probabilities in light of new evidence is the foundation of the approach. The initial development of Bayesian networks in the late 1970s was motivated by the need to model the top-down (semantic) and bottom-up (perceptual) combination of evidence in reading. The capability for bidirectional inferences, 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; Shafer and Pearl, 1990; Jensen, 1996).

The nodes in a Bayesian network represent propositional variables of interest (e.g., the temperature of a device, the sex of a patient, a feature of an object, 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 evidence about any other subset.

Figure 1 illustrates a simple yet typical Bayesian network. It describes the causal relationships among five variables: the season of the year (X_1), whether it's raining or not (X_2), whether the sprinkler is on or off (X_3), whether the pavement is wet or dry (X_4), and whether the pavement is slippery or not (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;

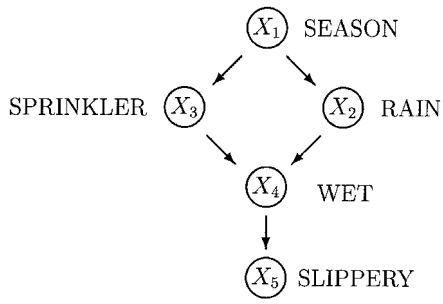

Figure 1. A Bayesian network representing causal influences among five variables. Each arc indicates a causal influence of the "parent" node on the "child" node.

the influence is mediated by the wetness of the pavement. (If freezing is a possibility, then a direct link could be added.)

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). Reasoning processes can operate on 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, or reasoning to a probable cause). 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, because it seems to require the propagation of information in two directions.

Probabilistic Semantics

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 $P(x_i \mid pa_i)$ denotes this conditional distribution. For example, $P(x_4 \mid x_2, x_3)$ is the probability of wetness given the values of sprinkler and rain. The *global semantics* of Bayesian networks specifies that the full joint distribution is given by the product

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

In our example network, we have

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

Provided that 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 conditional distributions. For example, in *sigmoid* networks (see Jordan, 1999), the conditional distribution associated with each variable is represented as a sigmoid function of a linear combination of the parent variables; in this way, the number of parameters required is proportional to, rather than exponential in, the number of parents.

There is also an entirely equivalent *local semantics* that asserts 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,

$$P(x_4 \mid x_1, x_2, x_3) = P(x_4 \mid x_2, x_3)$$

The collection of independence assertions formed in this way suffices to derive the global assertion in Equation 1, and vice versa. The local semantics is most useful in *constructing* Bayesian networks, because selecting as parents *all* the direct causes of a given variable invariably satisfies the local conditional independence conditions (Pearl, 2000, p. 30). The global semantics leads 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

$$P(X_3 = on \mid X_5 = true) = \frac{P(X_3 = on, X_5 = true)}{P(X_5 = true)}$$
$$= \frac{\sum_{x_1,x_2,x_4} P(x_1, x_2, X_3 = on, x_4, X_5 = true)}{\sum_{x_1,x_2,x_3,x_4} P(x_1, x_2, x_3, x_4, X_5 = true)}$$
$$= \frac{\sum_{x_1,x_2,x_4} P(x_1)P(x_2 \mid x_1)P(X_3 = on \mid x_1)P(x_4 \mid x_2, X_3 = on)P(X_5 = true \mid x_4)}{\sum_{x_1,x_2,x_3,x_4} P(x_1)P(x_2 \mid x_1)P(x_3 \mid x_1)P(x_4 \mid x_2, x_3)P(X_5 = true \mid x_4)}$$

These expressions can often be simplified in ways that reflect the structure of the network itself. The first algorithms proposed for probabilistic calculations in Bayesian networks used a local, distributed message-passing architecture, typical of many cognitive activities (Kim and Pearl, 1983). Initially this approach was limited to tree-structured networks, but it was later extended to general networks in Lauritzen and Spiegelhalter's (1988) method of join-tree propagation. A number of other exact methods have been developed and can be found in recent textbooks (Jensen, 1996; Jordan, 1999).

It is easy to show that reasoning in Bayesian networks subsumes the satisfiability problem in propositional logic and, hence, is NP-hard. Monte Carlo simulation methods can be used for approximate inference (Pearl, 1988), giving gradually improving estimates as sampling proceeds. (These methods use local message propagation on the original network structure, unlike join-tree methods.) Alternatively, variational methods provide bounds on the true probability (Jordan, 1999).

Uncertainty over Time

Entities that live in a changing environment must keep track of variables whose values change over time. Dynamic Bayesian networks, or DBNs, capture this process by representing multiple copies of the state variables, one for each time step (Dean and Kanazawa, 1989). A set of variables $\mathbf{X}_t$ denotes the world state at time t and a set of sensor variables $\mathbf{E}_t$ denotes the observations available at time t. The *sensor model* $P(\mathbf{E}_t \mid \mathbf{X}_t)$ is encoded in the conditional probability distributions for the observable variables, given the state variables. The *transition model* $P(\mathbf{X}_{t+1} \mid \mathbf{X}_t)$ relates the state at time t to the state at time $t + 1$. Keeping track of the world, known as *filtering*, means computing the current probability distribution over world states given all past observations, i.e., $P(\mathbf{X}_t \mid \mathbf{E}_1, \ldots, \mathbf{E}_t)$. Dynamic Bayesian networks include as special cases other temporal probability models, such as hidden Markov models (DBNs with a single discrete state variable) and Kalman filters (DBNs with continuous state and sensor variables and linear Gaussian transition and sensor models). For the general case, exact filtering is intractable, and a variety of approximation algorithms have been developed. The most popular and flexible of these is the family of *particle filtering* algorithms (see Doucet, de Freitas, and Gordan, 2001).

Learning in Bayesian Networks

The conditional probabilities $P(x_i \mid pa_i)$ can be updated continuously from observational data using gradient-based or Expectation-Maximization (EM) methods that use just local information derived from inference (Binder et al., 1997; Jordan, 1999), in much the same way as weights are adjusted in neural networks. It is also possible to learn the structure of the network, using methods that

trade off network complexity against degree of fit to the data (Friedman, 1998). As a substrate for learning, Bayesian networks have the advantage that it is relatively easy to encode prior knowledge in network form, either by fixing portions of the structure or by using prior distributions over the network parameters. Such prior knowledge can allow a system to learn accurate models from much less data than are required by *tabula rasa* approaches.

Causal Networks

Most probabilistic models, including general Bayesian networks, describe a distribution over possible observed events, as in Equation 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*, intuitively speaking, 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

$$P(x_1, x_2, x_3, x_4, x_5) = P(x_1)P(x_2 \mid x_1)P(x_4 \mid x_2, X_3 = on)P(x_5 \mid x_4)$$

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.

Causal networks are more properly defined, then, as Bayesian networks in which the correct probability model after intervening to fix any node's value is given simply by deleting links from the node's parents. For example, *fire* → *smoke* is a causal network, whereas *smoke* → *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 collection of stable component 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, 2000).

Functional Bayesian Networks

The networks discussed so far are capable of supporting reasoning about evidence and about actions. Additional refinement is necessary in order to process *counterfactual* information. For example, the probability that "the pavement would not have been slippery had the sprinkler been OFF, given that the sprinkler is in fact ON and that the pavement is in fact slippery" cannot be computed from the information provided in Figure 1 and Equation 1. Such counterfactual probabilities require a specification in the form of functional networks, where each conditional probability $P(x_i|pa_i)$ is replaced by a functional relationship $x_i = f_i(pa_i, \epsilon_i)$, where ϵ_i is a stochastic (unobserved) error term. When the functions f_i and the distributions of ϵ_i are known, all counterfactual statements can be assigned unique probabilities, using evidence propagation in a structure called a "twin network." When only partial knowledge about the functional form of f_i is available, bounds can be computed on the probabilities of counterfactual sentences (Pearl, 2000).

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, 2000)—a task 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 tossing two fair coins, and B could represent 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 render that structure incompatible with the data.

Plain Beliefs

In mundane decision making, beliefs are revised not by adjusting numerical probabilities but by tentatively accepting some sentences as "true for all practical purposes." Such sentences, called *plain beliefs*, exhibit both logical and probabilistic character. As in classical logic, they are propositional and deductively closed; as in probability, they are subject to retraction and can be held with varying degrees of strength. Bayesian networks can be adopted to model the dynamics of plain beliefs by replacing ordinary probabilities with nonstandard probabilities, that is, probabilities that are infinitesimally close to either zero or one (Goldszmidt and Pearl, 1996).

Discussion

Bayesian networks may be viewed as normative cognitive models of propositional reasoning under uncertainty. They handle noise and partial information using local, distributed algorithms for inference and learning. Unlike feedforward neural networks, they facilitate local representations in which nodes correspond to propositions of interest. Recent experiments suggest that they accurately capture the causal inferences made by both children and adults (Tenenbaum and Griffiths, 2001). Moreover, they capture patterns of reasoning, such as explaining away, that are not easily handled by any competing computational model. They appear to have many of the advantages of both the "symbolic" and the "subsymbolic" approaches to cognitive modeling, and are now an essential part of the foundations of computational neuroscience (Jordan and Sejnowski, 2001).

Two major questions arise when we postulate Bayesian networks as potential models of actual human cognition. First, does an architecture resembling that of Bayesian networks exist anywhere in the human brain? At the time of writing, no specific work has been done to design neurally plausible models that implement the required functionality, although no obvious obstacles exist. Second, how could Bayesian networks, which are purely propositional in their expressive power, handle the kinds of reasoning about individuals, relations, properties, and universals that pervade human thought? One plausible answer is that Bayesian networks containing propositions relevant to the current context are constantly being assembled, as needed, from a more permanent store of knowledge. For example, the network in Figure 1 may be assembled to help

explain why this particular pavement is slippery right now, and to decide whether this can be prevented. The background store of knowledge includes general models of pavements, sprinklers, slipping, rain, and so on; these must be accessed and supplied with instance data to construct the specific Bayesian network structure. The store of background knowledge must utilize some representation that combines the expressive power of first-order logical languages (such as semantic networks) with the ability to handle uncertain information. Substantial progress has been made on constructing systems of this kind (Koller and Pfeffer, 1998), but as yet no overall cognitive architecture has been proposed.

Road Maps: Artificial Intelligence; Learning in Artificial Networks
Related Reading: Bayesian Methods and Neural Networks; Decision Support Systems and Expert Systems; Graphical Models: Probabilistic Inference

References

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

Dean, T., and Kanazawa, K., 1989, A model for reasoning about persistence and causation, *Computat. Intell.*, 5:142–150.

Doucet, A., de Freitas, J., and Gordon, N., 2001, *Sequential Monte Carlo Methods in Practice*, Berlin: Springer-Verlag.

Friedman, N., 1998, The Bayesian structural EM algorithm, in *Uncertainty in Artificial Intelligence: Proceedings of the Fourteenth Conference* (G. F. Cooper and S. Moral, Eds.), San Mateo, CA: Morgan Kaufmann, pp. 129–138.

Goldszmidt, M., and Pearl, J., 1996, Qualitative probabilities for default reasoning, belief revision, and causal modeling, *Artif. Intell.*, 84:57–112.

Jensen, F. V., 1996, *An Introduction to Bayesian Networks*, New York: Springer-Verlag. ◆

Jordan, M. I., Ed., 1999, *Learning in Graphical Models*, Cambridge, MA: MIT Press. ◆

Jordan, M. I., and Sejnowski, T. J., Eds., 2001, *Graphical Models: Foundations of Neural Computation*, Cambridge, MA: MIT Press.

Kim, J. H., and Pearl, J., 1983, A computational model for combined causal and diagnostic reasoning in inference systems, in *Proceedings of the Eighth International Joint Conference on Artificial Intelligence* (*IJCAI-83*), San Mateo, CA: Morgan Kaufmann, pp. 190–193.

Koller, D., and Pfeffer, A., 1998, Probabilistic frame-based systems, in *Proceedings of the Fifteenth National Conference on Artificial Intelligence* (*AAAI-98*), Menlo Park, CA: AAAI Press, pp. 580–587.

Lauritzen, S. L., and Spiegelhalter, D. J., 1988, Local computations with probabilities on graphical structures and their application to expert systems (with discussion), *J. R. Statist. Soc.*, series B, 50:157–224.

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

Pearl, J., 2000, *Causality: Models, Reasoning, and Inference*, New York: Cambridge University Press. ◆

Shafer, G., and Pearl, J., Eds., 1990, *Readings in Uncertain Reasoning*, San Mateo, CA: Morgan Kaufmann.

Tenenbaum, J. B., and Griffiths, T. L., 2001, Structure learning in human causal induction, in *Advances in Neural Information Processing Systems 13*, Cambridge, MA: MIT Press.