To appear in M. Arbib (Ed.), *Handbook of Brain Theory and Neural Networks*, MIT Press, 2001.

TECHNICAL REPORT
R-277
November 2000

BAYESIAN NETWORKS

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

Stuart Russell
Computer Science Department
University of California, Berkeley, CA 94720
russell@cs.berkeley.edu

Probabilistic models based on directed acyclic graphs (DAGs) have a long and rich tradition, which began with the geneticist Sewall Wright (1921). Variants have apeared in many fields; within cognitive science and artificial intelligence, such models are known as Bayesian networks. Their initial development 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, Heckerman *et al.*, 1995, Jensen, 1996].

The nodes in a Bayesian network represent propositional variables of interest (e.g., the temperature of a device, the gender 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 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.)

Perhaps the most important aspect of a 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

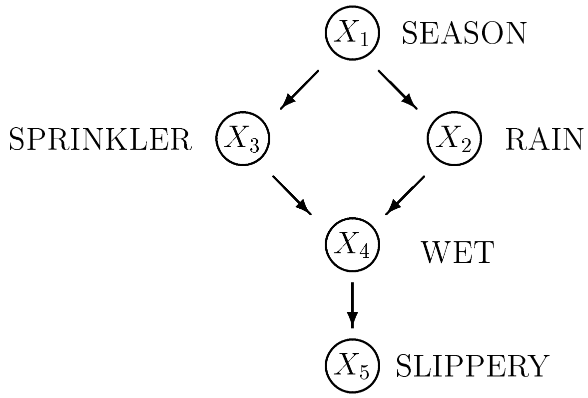

Figure 1: A Bayesian network representing causal influences among five variables.

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 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 $P(x_i|pa_i)$ 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 semantics* 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 conditional distributions.

There is also an entirely equivalent *local semantics*, which 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,

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

The collection of independence assertions formed in this way suffices to derive the global assertion in Equation 1, and vice versa. The local semantics is most useful in *constructing*

Bayesian networks, because selecting as parents the direct causes of a given variable automatically satisfies the local conditional independence conditions. 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

$$
\begin{aligned}
P(X_3 = on | X_5 = true) &= \frac{P(X_3 = on, X_5 = true)}{P(X_5 = true)} \\
&= \frac{\sum_{x_1, x_2, x_4} P(x_1, x_2, X_3 = on, x_4, X_5 = true)}{\sum_{x_1, x_2, x_3, x_4} P(x_1, x_2, x_3, x_4, X_5 = true)} \\
&= \frac{\sum_{x_1, x_2, x_4} P(x_1)\ P(x_2|x_1)\ P(X_3 = on|x_1)\ P(x_4|x_2, X_3 = on)\ P(X_5 = true|x_4)}{\sum_{x_1, x_2, x_3, x_4} P(x_1)\ P(x_2|x_1)\ P(x_3|x_1)\ P(x_4|x_2, x_3)\ P(X_5 = true|x_4)}
\end{aligned}
$$

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 [Pearl, 1982, Kim and Pearl, 1983]. Initially, this approach was limited to tree-structured networks, but was later extended to general networks in Lauritzen and Spiegelhalter's (1988) method of join-tree propagation. Other exact methods include cycle-cutset conditioning [Pearl, 1988] and variable elimination [Zhang and Poole, 1996].

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, 1987], 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 [Jordan *et al.*, 1998] provide bounds on the true probability.

Uncertainty over time. Entities that live in a changing environment must keep track of variables whose values change over time. *Dynamic* Bayesian networks [Dean and Kanazawa, 1989] capture this process by representing multiple copies of the state variables, one for each time step. 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* $P(\mathbf{E}_t|\mathbf{X}_t)$ is encoded in the conditional probability distributions for the observable variables, given the state variables. The *transition model* $P(\mathbf{X}_{t+1}|\mathbf{X}_t)$ relates the state at time t to the state at time $t+1$. Keeping track of the world means computing the current probability distribution over world states given all past observations, i.e., $P(\mathbf{X}_t|\mathbf{E}_1, \ldots, \mathbf{E}_t)$. Dynamic Bayesian networks are strictly more expressive than other temporal probability models such as hidden Markov models and Kalman filters.

Learning in Bayesian networks. The conditional probabilities $P(x_i|pa_i)$ can be updated continuously from observational data using gradient-based or EM methods that use just local information derived from inference [Lauritzen, 1995, Binder *et al.*, 1997]—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].

Causal networks. Most probabilistic models, including general Bayesian networks, 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*, 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[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 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, $Fire \longrightarrow Smoke$ 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 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, 1996].

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 Eq. 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 know, 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. [Balke and Pearl, 1995, 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 and Verma, 1991, Spirtes *et al.*, 1993, 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 two fair coins, and B represents a bell that rings whenever either coin comes up heads.)

[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.

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 to varying degrees of entrenchment. Bayesian networks can be adopted to model the dynamics of plain beliefs by replacing ordinary probabilities with non-standard probabilities, that is, probabilities that are infinitesimally close to either zero or one [Goldszmidt and Pearl, 1996].

Models of cognition. 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 [Tenenbaum and Griffiths, 2001] suggest that they capture accurately the causal inferences made by both children and adults. 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 modelling.

Two major questions arise when we postulate Bayesian networks as potential models of actual human cognotion. 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 pervades 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 [Halpern, 1990, Koller and Pfeffer, 1998], but as yet no overall cognitive architecture has been proposed.

References

[Balke and Pearl, 1995] A. Balke and J. Pearl. Counterfactuals and policy analysis in structural models. In P. Besnard and S. Hanks, editors, *Uncertainty in Artificial Intelligence 11*, pages 11–18. Morgan Kaufmann, San Francisco, 1995.

[Binder *et al.*, 1997] John Binder, Daphne Koller, Stuart Russell, and Keiji Kanazawa. Adaptive probabilistic networks with hidden variables. *Machine Learning*, 29:213–244, 1997.

[Dean and Kanazawa, 1989] Thomas Dean and Keiji Kanazawa. A model for reasoning about persistence and causation. *Computational Intelligence*, 5(3):142–150, 1989.

[Friedman, 1998] 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.

[Goldszmidt and Pearl, 1996] M. Goldszmidt and J. Pearl. Qualitative probabilities for default reasoning, belief revision, and causal modeling. *Artificial Intelligence*, 84(1–2):57–112, July 1996.

[Halpern, 1990] J. Y. Halpern. An analysis of first-order logics of probability. *Artificial Intelligence, volume =*, 1990.

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

[Jensen, 1996] Finn V. Jensen. *An Introduction to Bayesian Networks*. Springer, New York, 1996.

[Jordan *et al.*, 1998] M. I Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models. In M.I. Jordan, editor, *Learning in Graphical Models*. Kluwer, Dordrecht, The Netherlands, 1998.

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

[Koller and Pfeffer, 1998] D. Koller and A. Pfeffer. Probabilistic frame-based systems. In *Proceedings of the Fifteenth National Conference on Artificial Intelligence (AAAI-98)*, Madison, Wisconsin, 1998. AAAI Press.

[Lauritzen and Spiegelhalter, 1988] 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.

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

[Pearl and Verma, 1991] 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.

[Pearl, 1982] J. Pearl. Reverend Bayes on inference engines: A distributed hierarchical approach. In *Proceedings AAAI National Conference on AI*, pages 133–136, Pittsburgh, PA, 1982.

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

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

[Pearl, 1996] J. Pearl. Causation, action, and counterfactuals. In Y. Shoham, editor, *Theoretical Aspects of Rationality and Knowledge, Proceedings of the Sixth Conference*, pages 51–73. Morgan Kaufmann, San Francisco, CA, 1996.

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

[Shafer and Pearl, 1990] G. Shafer and J. (Eds.) Pearl. *Readings in Uncertain Reasoning*. Morgan Kaufmann, San Mateo, CA, 1990.

[Spirtes *et al.*, 1993] P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. Springer-Verlag, New York, 1993.

[Tenenbaum and Griffiths, 2001] J. B. Tenenbaum and T. L. Griffiths. Structure learning in human causal induction. In *Advances in Neural Information Processing Systems*, volume 13, Denver, Colorado, 2001. MIT Press.

[Wright, 1921] S. Wright. Correlation and causation. *Journal of Agricultural Research*, 20:557–585, 1921.

[Zhang and Poole, 1996] N.L. Zhang and D. Poole. Exploiting causal independence in Bayesian network inference. *Journal of Artificial Intelligence Research*, 5:301–328, 1996.