In *Fundamenta Informaticae, 30(3-4):241-254, 1997*
Special issue on Knowledge Representation and Machine Learning

Five Useful Properties of Probabilistic Knowledge Representations From the Point of View of Intelligent Systems

Marek J. Druzdzel
University of Pittsburgh
Department of Information Science
and Intelligent Systems Program
Pittsburgh, PA 15260, U.S.A.
marek@lis.pitt.edu

Abstract

Although probabilistic knowledge representations and probabilistic reasoning have by now secured their position in artificial intelligence, it is not uncommon to encounter misunderstanding of their foundations and lack of appreciation for their strengths. This paper describes five properties of probabilistic knowledge representations that are particularly useful in intelligent systems research. **(1)** Directed probabilistic graphs capture essential qualitative properties of a domain, along with its causal structure. **(2)** Concepts such as relevance and conflicting evidence have a natural, formally sound meaning in probabilistic models. **(3)** Probabilistic schemes support sound reasoning at a variety of levels ranging from purely quantitative to purely qualitative levels. **(4)** The role of probability theory in reasoning under uncertainty can be compared to the role of first order logic in reasoning under certainty. Probabilistic knowledge representations provide insight into the foundations of logic-based schemes, showing their difficulties in highly uncertain domains. Finally, **(5)** probabilistic knowledge representations support automatic generation of understandable explanations of inference for the sake of user interfaces to intelligent systems.

1 Introduction

Reasoning within such disciplines as engineering, science, management, or medicine is usually based on formal, mathematical methods employing probabilistic treatment of uncertainty. While heuristic methods and ad-hoc reasoning schemes may in many domains perform well, most engineers will be reluctant to rely on them whenever the cost of making an error is high. To give an extreme example, few people would choose to fly airplanes built using heuristic principles over airplanes built using the laws of aerodynamics enhanced with probabilistic reliability analysis. The attractiveness of probability theory lies in its soundness and its guarantees concerning long-term performance. Similarly to the first order logic in deterministic reasoning, probability theory can be viewed as a gold standard for rationality in reasoning under uncertainty. Following its axioms protects from some elementary inconsistencies. Their violation, on the other hand, can be demonstrated to lead to sure losses [19]. Application of probabilistic methods in intelligent systems makes these systems philosophically distinct from those based on the mainstream artificial intelligence methods. Rather than imitating humans, they support human reasoning by a normative theory of decision making. A useful analogy is that of an electronic calculator: the calculator aids people's limited capacity for mental arithmetics rather than imitating it. The distrust for human capabilities for reasoning under uncertainty has a substantial empirical support [17].

This paper argues that probability theory has much to offer to builders of intelligent systems. The five sections to follow describe each one property of probabilistic knowledge representations that is particularly useful in intelligent systems research. They summarize in an accessible and informal way the most important research results that support the thesis, giving pointers to original papers for those readers who are interested in details and in a formal exposition.

2 Foundations of Probabilistic Knowledge Representation

As outlined carefully by Leonard Savage in his influential book on the foundations of Bayesian probability theory and decision theory [19], probabilistic reasoning is always confined to a well defined set of uncertain variables, which Savage refers to as "small world." A probabilistic model consists of an explicit specification of these variables and the information about the probability distribution over all possible combinations of their values,[1] known as the joint probability distribution. It is a fundamental assumption of the Bayesian approach that the joint probability distribution exists and if needed can be elicited from a human expert.[2] If there are n propositional variables in a model, there are 2^n states of the model and, effectively, the joint probability distribution consists of 2^n numbers. It is seldom the case that all these numbers have to be elicited and stored in a model. By factorizing the joint probability distribution and exploring the independences existing in the domain, one can reduce it to a product of a small number of probabilities. If, for example, a model consists of three variables x, y, and z, we can specify the joint probability distribution $\Pr(xyz|\mathcal{S})$, where $\mathcal{S}$ is the state of available information. Such a specification of the joint probability distribution can be rewritten as a product of conditional probability distributions of each of the variables. This is called factorization of the joint probability distribution. Two possible factorizations of three variables x, y, and z are

$$\begin{aligned} \Pr(xyz|\mathcal{S}) &= \Pr(x|yz\mathcal{S}) \Pr(y|z\mathcal{S}) \Pr(z|\mathcal{S}) \\ \Pr(xyz|\mathcal{S}) &= \Pr(z|xy\mathcal{S}) \Pr(y|x\mathcal{S}) \Pr(x|\mathcal{S}) \,. \end{aligned} \tag{1}$$

Simple combinatorics shows that a joint probability distribution of n variables can be factorized in $n!$ possible ways, so there are 6 possible factorizations of $\Pr(xyz|\mathcal{S})$.

[1] A combination of outcomes of all variables, i.e., an element of the Cartesian product of sets of outcomes of all individual model's variables, can be succinctly defined as a state. Many terms have been used to describe states of a model: extension, instantiation, possible world, scenario, etc. Throughout this paper, I will attempt to use the term *state of a model* or briefly *state* whenever possible.

[2] It is not necessary, however, to specify it numerically in order to perform useful reasoning — in fact a specification of the constraints on this joint probability distribution and reasoning in terms of these constraints leads to schemes of less specificity and even purely qualitative schemes, as will be shown in Section 5.

Knowledge of conditional independences among variables allows for simplifications in the factorized formulas. For example, if we know that x is conditionally independent of y given z, i.e.,

$$\Pr(xy|z\mathcal{S}) = \Pr(x|z\mathcal{S})\Pr(y|z\mathcal{S}) \,,$$

we have by Bayes theorem that

$$\Pr(x|yz\mathcal{S}) = \Pr(x|z\mathcal{S}) \,.$$

This allows for a simplification in the factorized formula (1)

$$\Pr(xyz|\mathcal{S}) = \Pr(x|yz\mathcal{S})\Pr(y|z\mathcal{S})\Pr(z|\mathcal{S}) = \Pr(x|z\mathcal{S})\Pr(y|z\mathcal{S})\Pr(z|\mathcal{S}) \qquad (2)$$

The above generalizes easily to conditional independence involving sets of variables. Simplifications in the factorized form of a joint probability distribution lead to representational savings. If x, y, and z are propositional, the conditional distribution $\Pr(x|yz\mathcal{S})$ can be specified by a $2 \times 2 \times 2$ probability matrix. The simplified form $\Pr(x|z\mathcal{S})$, exploring independence between x and y conditional on z, can be specified by a 2×2 matrix. The factorization of joint probability distribution and explicit use of conditional independences in the factorized form underlie the idea of *Bayesian belief networks* (BBNs) [18]. Nodes in a BBN represent random variables. Lack of a directed arc between a node a and a node b means that variables a and b are independent conditional on some subset Ψ of other variables in the model (Ψ can also be empty). Figure 1 shows Bayesian belief networks for factorizations (1) and (2) (left and right graph respectively). Lack of a direct arc between x and y in the right graph expresses conditional independence of x and y given z.

Figure 1: Bayesian belief networks for factorizations (1) and (2) (left and right graph respectively).

Figure 2 shows an example of a BBN modeling various causes of low level of car engine oil. There are many independences represented explicitly in this graph. And so, *loose bolt* and *crack in the gasket* are independent. They become dependent conditional on *oil leak* or any of its descendants. *Worn piston rings* is independent on *clean exhaust* conditional on *excessive oil consumption*. The graphical model, such as the one in Figure 2, is usually supplemented by its numerical properties, expressed by matrices of conditional probabilities stored in each of the nodes. With each of the 12 variables in this model being propositional, the complete joint probability distribution contains $2^{12} = 4096$ numbers. Explicit information about independences included in the model allows for specifying it by only 54 numbers (or 27, if we take into account that for every propositional variable x, $\Pr(\overline{x}) = 1 - \Pr(x)$). A popular approximation of the interaction between a node and its direct predecessors in a BBN is the Noisy–OR gate [18]. In Noisy–OR gates, each of the arcs is described by a single number expressing the causal strength of the interaction between the parent and the child. If there are

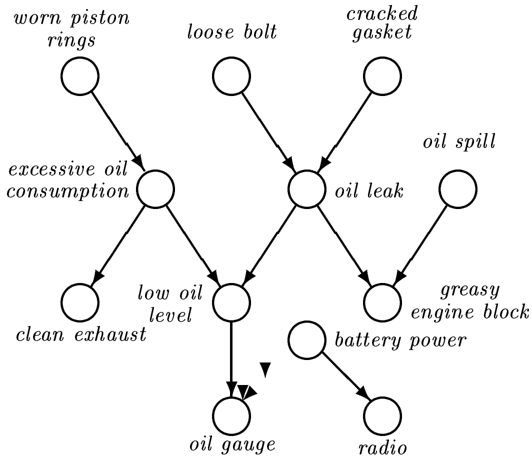

Figure 2: Example of a Bayesian belief network

other, unmodeled causes of a, we need one additional number, known as *leak probability*, denoting the causal strength of all unmodeled causes of a. If each of the interactions in our model is approximated by a leaky Noisy–OR gate, 23 numbers suffice to specify the entire joint probability distribution.

Both, the structure and the numerical probability distributions in a BBN are elicited from a human expert and are a reflection of the expert's subjective view of a real world system. Scientific knowledge about the system, both in terms of the structure and frequency data, if available, can be easily incorporated in the model. It is apparent from the above example that BBNs offer a compact representation of joint probability distributions and are capable of practical representation of large models. BBNs can be easily extended with decision and value variables for modeling decision problems. Such amended graphs are known as *influence diagrams* [20].

3 Probability, Causality, and Action

It seems to be an accepted view in psychology that humans attempt to achieve a coherent interpretation of the events that they observe by organizing their knowledge in schemas consisting of cause-effect relations. This holds for both scientific and everyday reasoning. Scarcity of references to causality in most statistics textbooks and the disclaimers that usually surround the term "causation" create the impression that causality forms a negative and unnecessary ballast on human mind that cannot be reconciled with the probabilistic approach. In fact, causality and probability are closely related. While probabilistic relations indeed do not imply causality, causality normally implies a pattern of probabilistic interdependencies. A generally accepted necessary condition for causality is statistical dependence. For a to be considered a cause of b in a context $\mathcal{S}$, it is necessary that $\Pr(b|a\mathcal{S}) \neq \Pr(b|\overline{a}\mathcal{S})$, i.e., the presence of a must have impact on the probability of b.

Directed graphs readily combine the symmetric view of probabilistic dependence with the asymmetry of causality. A directed graph can be given causal interpretation and can be viewed as a structural model of the underlying domain. Simon and I [10] tied the work on structural equations models in econometrics to probabilistic models and formulated the semantic conditions under which a directed probabilistic graph is causal. We have shown that a node and all its direct predecessors in a graph play a role that is equivalent to that of a structural equation. Structural equations in econometric are equations describing unique mechanisms acting in the system [22]. For example, in a simple physical system such as a pendulum, one of the mechanisms might be described by the equation $f = mg$, where m is the mass of the pendulum, g is Earth's gravitational constant, and f the force with which Earth acts on the pendulum. Mechanisms are identifiable by underlying physical, chemical, social, or other laws, physical adjacency, connection, or interaction. As we have shown, one can view each node in a probabilistic graph along with its direct predecessors as a qualitative specification of a mechanism acting in a system equipped with its approximate numerical description.

There are two important reasons for interest in causality in the context of intelligent systems. The first is that models that include causal information are natural and in general easier to construct and modify than models that are not causal [14, 21]. Such models are also easier for the system to explain and for their users to comprehend [2, 25]. The theoretical link between structural equations models and directed probabilistic graphs shows how prior theoretical knowledge about a domain, captured in structural equations, can aid construction of BBNs. If we happen to know the mechanism tying a group of variables, we can make these variables adjacent in the constructed graph. Existing theoretical knowledge, if incorporated at the model building stage, can aid human experts, make model building easier, and, finally, improve the quality of constructed models.

The second reason for interest in causality is that autonomous intelligent planning systems should be able to predict the effects of their actions. For this, the model that they base their reasoning on, i.e., their picture of the world, needs to be causal. Spirtes et al. [23] show in what they call the *manipulation theorem*, that it is straightforward to predict the effect of manipulating a variable in a probabilistic causal graph. The probability distribution over the manipulated graph can be obtained by modifying the conditional distributions of the manipulated variables. Imposing a value on a variable x through an external intervention, in particular, amounts to removing all arcs in the graph that point at x. And so, manipulation of the variable *greasy engine block* (for example, by washing the engine) will have no effect on any other variable in the model of Figure 2. On the other hand, manipulation of the variable *low oil level* (for example, by adding oil) will impact the indication of the *oil gauge*, but not variables *excessive oil consumption*, *oil leak*, or any of the other variables in the graph.

4 Relevance in Probabilistic Models

Typically, an intelligent system includes a large body of domain knowledge that is essential for its reasoning. An important problem that such a system faces is identifying those parts of the domain knowledge that are relevant for the query that it is addressing. "Small worlds" modeled by probabilistic systems may include hundreds or thousands of variables. Each of the variables of a probabilistic model may be relevant for some types of reasoning within this domain, but rarely will all of them participate in reasoning related to a single query. Too much information may unnecessarily degrade the system's overall performance. Focusing on the most relevant part of the model is also crucial in explanation: too many marginally relevant facts will have a confounding effect on most users. It is important, therefore, to identify a subset of the "small world" including only those elements of the domain model that are directly relevant to a particular problem. Suermondt and I [11] recently summarized methods that can be used for such reduction in probabilistic models. Each of these methods is fairly well understood theoretically and has been practically implemented. While I would like to direct interested readers to our paper for a comprehensive treatment of the issue of relevance in probabilistic models, I will give a flavor of these methods below.

One possible way of reducing the size of the model is instantiating evidence variables to their observed values. The observed evidence may be causally sufficient to imply the values of other, as yet unobserved nodes (e.g., if a patient is male, it implies that he is not pregnant). Similarly, observed evidence may imply other nodes that are causally necessary for that evidence to occur (e.g., observing that the *radio* works might in our simple model imply *battery power*). Each instantiation reduces the number of uncertain variables and, hence, reduces the computational complexity of inference. Further, instantiations can lead to additional reductions, as they may screen off other variables by making them independent of the variables of interest (discussed below).

Parts of the model that are probabilistically independent from a node of interest t given the observed evidence are clearly not relevant to reasoning about t. Geiger et al. [12] show a computationally efficient way of identifying nodes that are probabilistically independent from a set of nodes of interest given a set of observations by exploring independences implied by the structural properties of the graph. They base their algorithm on a condition known as d-separation, binding probabilistic independence to the structure of the graph. Reduction achieved by means of d-separation can be significant. For example, observing *excessive oil consumption*, makes each of the variables in the example graph independent of *worn piston rings*. If this is the variable of interest, almost the whole graph can be reduced.

Further reduction of the graph can be performed by removing nodes that are not computationally relevant to the nodes of interest given the evidence, known as *barren nodes* [20]. Barren nodes are uninstantiated child-less nodes in the graph. They depend on the evidence, but do not contribute to the change in probability of the target node and are, therefore, computationally irrelevant. If the presence of *low oil level* is unknown, then the probability distribution of *low oil level* is not necessary for computing the belief in *clean exhaust*, *excessive oil consumption*, *oil leak*, and ancestors of the latter two.

A probabilistic graph is not always capable of representing independences explicitly [18]. The d-separation criterion assumes, for example, that an instantiated node makes its predecessors probabilistically dependent. One reflection of this phenomenon is a common pattern of reasoning known as "explaining away." For example, given *low oil level*, observing *oil leak* makes *excessive oil consumption* less likely. Noisy–OR gates, for example, violate this principle: predecessors of a Noisy-OR gate remain conditionally independent when the common effect has been observed absent (e.g., when the *oil level* has been observed normal, *oil leak* and *excessive oil consumption* remain independent [9]. A careful study of the probability distribution matrices in a graph may reveal additional independences and further opportunities for reduction. Procedures for this examination follow straightforwardly from the probabilistic definition of independence.

For some applications, such as user interfaces, there is another class of variables that can be reduced. This class consists of those predecessor nodes that do not take active part in propagation of belief from the evidence to the target, called *nuisance nodes*. A nuisance node, given evidence e and variable of interest t, is a node that is computationally related to t given e but is not part of any active trail from e to t. The idea here is that only the active trails from e to the t are relevant for explaining the impact of e on t. If we are interested in the relevance of *worn piston rings* to *low oil* *level*, then *oil leak* and all its ancestors fall into the category of nuisance nodes and can be removed.

The above methods do not alter the quantitative properties of the underlying graph (removal of nodes has no effect on the probability distribution over the remaining nodes) and are, therefore, exact. In addition, for a collection of evidence nodes e and a node of interest t, there will usually be nodes in the BBN that are only marginally relevant for computing the posterior probability distribution of t. Identifying the nodes that have non-zero but small impact on the probability of t and pruning them can lead to a further simplification of the graph with only a slight loss of precision of the conclusions. To identify such nodes, one needs a suitable metric for measuring changes to the distribution of t, as well as a threshold beyond which changes are unacceptable. Such metrics can be derived solely from the probabilities (e.g., cross entropy), or from decision and utility models involving the distribution of t. In INSITE, a system that generates explanations of BBN inference, Suermondt [24] found cross entropy to be the most practical measure. Use of such a metric and threshold allows us to discriminate between more and less influential evidence nodes, and to identify nodes and arcs in the BBN that might, for practical purposes, be omitted from computations and from explanations of the results.

Relevance in probabilistic models has a natural interpretation and probability theory supplies effective tools that aid in determining what is at any given point most crucial for the inference. The common denominator of the above methods is that they are theoretically sound and quite intuitive. They are exact or, as it is the case with the last method, they come with an apparatus for controlling the degree of approximation, preserving correctness of the reduced model.

5 Qualitative Probabilistic Reasoning

Probabilistic reasoning schemes are often criticized for the undue precision they require to represent uncertain knowledge in the form of numerical probabilities. In fact, such criticism is misplaced since probabilistic reasoning does not need to be conducted with a full numerical specification of the joint probability distribution over a model's variables. Useful conclusions can be drawn from merely constraints on the joint probability distributions. Most of relevance reasoning, described in the previous section, is purely qualitative and based only on the structure of the directed probabilistic graph. Another instance of qualitative probabilistic reasoning can be obtained by amending reasoning about relevance with reasoning about its sign.

Wellman introduced a qualitative abstraction of BBNs, known as *qualitative probabilistic networks* (QPNs)[26]. QPNs share the structure with BBNs, but instead of numerical probability distributions, they represent the signs of interactions among variables in the model. A proposition a has a positive *influence* on a proposition b, if observing a to be true makes b more probable. QPNs generalize straightforwardly to multivalued and continuous variables. QPNs can replace or supplement quantitative Bayesian belief networks where numerical probabilities are either not available or not necessary for the questions of interest. An expert may express his or her uncertain knowledge of a domain directly in the form of a QPN. This requires significantly less effort than a full numerical specification of a BBN. Alternatively, if we already possess a numerical BBN, then it is straightforward to identify the qualitative relations inherent in it, based on the formal probabilistic definitions of the properties. QPNs are useful for structuring planning problems and identification of dominating alternatives in decision problems [26]. Another application of QPNs is in model building — the process of probability elicitation can be based on a combination of qualitative and quantitative information [5]. Figure 3 shows a QPN for the example of Figure 2.

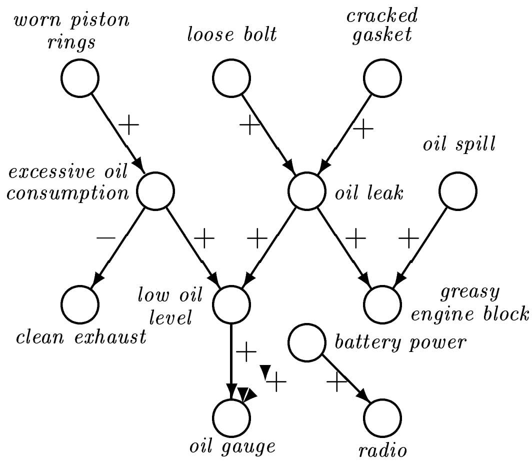

Figure 3: Example of a qualitative probabilistic network

Henrion and I [8] proposed an efficient algorithm for reasoning in QPNs, called *qualitative belief propagation*. Qualitative belief propagation traces the effect of an observation e on other graph variables by propagating the sign of change from e through the entire graph. Every node t in the graph is given a label that characterizes the sign of impact of e on t. Figure 4 gives an example of how the algorithm works in practice. Suppose that we have previously observed *low oil level* and we want to know the effect of observing blue exhaust (i.e., *clean exhaust* to be false) on other variables in the model. We set the signs of each of the nodes to 0 and start by sending a negative sign to *clean exhaust*, which is our evidence node. *Clean exhaust* determines that its parent, node *excessive oil consumption*, needs updating, as the sign product of $(-)$ and the sign of the link $(-)$ is $(+)$ and is different from the current value at the node (0). After receiving

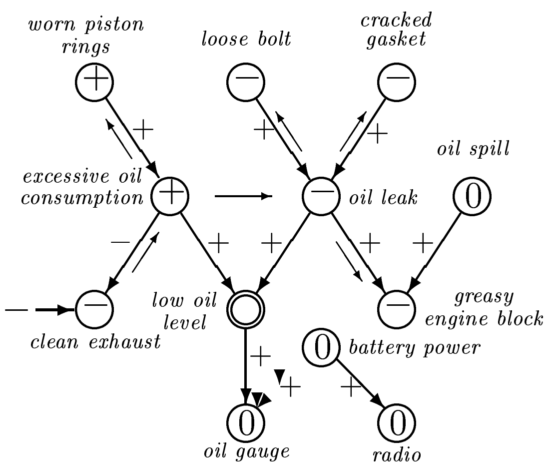

Figure 4: Example of qualitative belief propagation.

this message, *excessive oil consumption* sends a positive message to *worn piston rings*. Given that the node *low oil level* has been observed, *excessive oil consumption* will also send a negative intercausal message to *oil leak* (this is an instance of "explaining away" captured by a condition called *product synergy* [9, 13, 16, 27]). No messages are passed to *oil gauge*, as it is d–separated from the rest of the graph by *low oil level. Oil leak* sends negative messages to *loose bolt*, *cracked gasket*, and *greasy engine block. Oil spill* is d–separated from *oil leak* and will not receive any messages. The final sign in each node (marked in Figure 4) expresses the sign of change in probability caused by observing the evidence (in this case, blue exhaust). Once the propagation is completed, one can easily read off the labeled graph exactly how the evidence propagates through the model, including all intermediate nodes through which the evidence impacts a target variable.

If the signs of impact of two pieces of evidence e_1 and e_2 on a node t are different, we are dealing with conflicting evidence. We speak about conflicting evidence also when an evidence variable e impacts t positively through one path and negatively through another. The labels placed on each node in the graph by the qualitative belief propagation algorithm allows a computer program, in case of sign-ambiguity, to reflect about the model at a meta level and find the reason for ambiguity, for example, which paths are in conflict. Hence, it can suggest ways in which the least additional specificity could resolve the ambiguity.

6 From Probability to Logics

One way of looking at models of uncertain domains is that they describe a set of possible states of the world. This view is explicated by the logic-based approaches to reasoning under uncertainty — at any given point various extensions of the current body of facts are possible, one of which, although unidentified, is assumed to be true. Since the number of possible extensions of the facts is exponential in the number of uncertain variables in the model, it seems to be intuitively appealing, and for sufficiently large domains practically necessary, to limit the number of extensions considered. Several artificial intelligence schemes for reasoning under uncertainty, such as case-based or script-based reasoning, abduction, or non-monotonic logics, seem to be following this path. This very approach may be taken by humans in reasoning under uncertainty. I have investigated theoretically the conditions under which this approach is justifiable [3]. It turns out that for a wide range of models, one can expect a small number of states to cover most of the probability space. I have demonstrated that the probabilities of individual states of the model can be expected to be drawn from lognormal distributions. The probability mass carried by the individual states follows also lognormal distribution, but it is usually strongly shifted towards higher probability values and cut off at the point $p = 1.0$. The asymmetry in individual prior and conditional probability distributions determines the variance in the distribution of probabilities of single states (probabilities of states are spread over many orders of magnitude) and also determines the magnitude of the shift towards the higher values of probabilities. For sufficiently asymmetric distributions (i.e., for distributions describing well known systems, where there is not too much uncertainty), a small fraction of states can be expected to cover a large portion of the total probability space with the remaining states having practically negligible probability. In the limit, when there is no uncertainty, one single state covers the entire probability space.

Intuitively, the more we know about a domain, the more asymmetry individual conditional probabilities will show. When the domain and its mechanisms are well known, probability distributions tend to be extreme. This implies a small number of very likely states of the model. When an environment is less familiar, the probability distributions tend to be less extreme, the shift in contribution function is small and none of the states is very likely. Figure 5 shows theoretically derived probability density functions for two models consisting of ten binary variables, in which individual conditional probability distributions were 0.2 and 0.8 (left diagram) and 0.1 and 0.9 (right diagram). The ordinate is in decimal logarithmic scale — the lognormal distributions found in practical models tend to span over many orders of magnitude and are extremely skewed, making them unreadable in linear scale. Note that the distribution pictured in the right

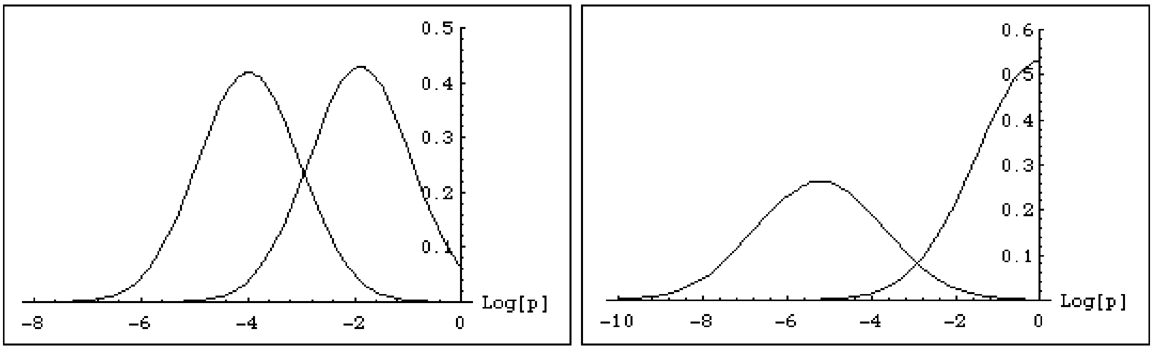

Figure 5: Theoretically derived distributions for identical conditional probability distributions for 10 binary variables with probabilities of outcomes equal to 0.2 and 0.8 (left diagram) and 0.1 and 0.9 (right diagram).

diagram is for a system with more symmetry in the distribution, i.e., a system that we know less about. In this case, the shift towards higher probabilities is small, most states will have low probabilities and, hence, no very likely states will be observed. In the spirit of a demonstration device similar to those proposed by Gauss or Kapteyn to show a mechanism by which a distribution is generated, I performed several simulation studies in which I randomly generated belief networks models and subsequently studied the distribution of states in these networks. These studies corroborated the theoretical findings. A stronger support for this analysis comes from studying the properties of a real model. The most realistic model with a full numerical specification that was available to me was ALARM, a medical diagnostic model of monitoring anesthesia patients in intensive care units [1]. With its 38 random variables, each having two or three outcomes, ALARM has a computationally prohibitive number of states. I selected, therefore, several self-contained subsets of ALARM consisting of 7 to 13 variables, and analyzed the distribution of probabilities of all states within those subsets. Figure 6 shows the result of one of such run, identical with the results of all other runs with respect to the form of the observed distribution. It is apparent that the histogram of states appears to be for normally distributed variables, which, given that the ordinate is in logarithmic scale, supports the theoretically expected lognormality of the actual distribution. The histogram also indicates a small contribution of its tail to the total probability mass. The subset studied contained 13 variables, resulting in 525,312 states. The probabilities of these states were spread over 22 orders of magnitude. Only the most likely states, spread over the first five orders of magnitude, provided meaningful contribution to the total probability mass. Of all states, there was one state with probability 0.52, 10 states with probabilities in the range $(0.01, 0.1)$ and the total probability of 0.23, and 48 states with probabilities in the range $(0.001, 0.01)$ and the total probability of 0.16. The most likely state covered 0.52 of the total probability space,

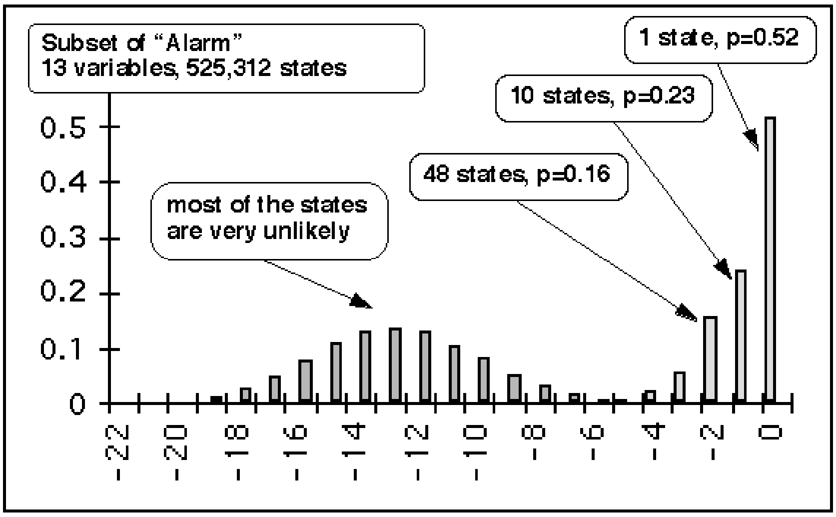

Figure 6: Histograms of the probabilities of various states (the bell-shaped curve) and their contribution to the total probability mass (the peak on the right side) for a subset of 13 variables in the ALARM model.

the 11 most likely states covered 0.75 of the total probability space, and the 59 most likely states (out of the total of 525,312) covered 0.91 of the total probability space.

The above result gives some insight into the logic-based schemes for reasoning under uncertainty, showing when and why they will work and when they will not perform too well. In the domains that are well known, there will be usually a small number of very likely states and these states can be modeled in logic. In the domains that contain much uncertainty, logic-based approaches will fail: there will be many plausible states and commitment to any of them is unreasonable.

7 Human Interfaces to Probabilistic Systems

Decision analysis, which is the art and science of applying decision theory to aid decision making in the real world, has developed a considerable body of knowledge in model building, including elicitation of the model structure and elicitation of the probability distribution over its variables. These methods have been under a continuous scrutiny of psychologists working in the domain of behavioral decision theory and have proven to cope reasonably well with the dangers related to human judgmental biases. The approach taken by decision analysis is compatible with that of intelligent systems. The goal of decision analysis is to provide insight into a decision. This insight, consisting of the analysis of all relevant factors, their uncertainty, and criticality of some assumptions, is even more important than the actual recommendation.

Probability theory is known to model well certain patterns of human plausible reasoning, such as mixing predictive and diagnostic inference, discounting correlated sources of evidence, or intercausal reasoning [13]. BBNs offer several advantages for automatic generation of explanation of reasoning to the users of intelligent systems. As they encode the structure of the domain along with its numerical properties, this structure can be analyzed at different levels of precision. The ability to derive lower levels of specification and, therefore, changing the precision of the representation makes probabilistic models suitable for both computation and explanation. Soundness of the reasoning procedure makes it easier to improve the system, as explanations based on a less precise abstraction of the model provide an approximate, but correct picture of the model. Possible disagreement between the system and its user can always be reduced to a disagreement over the model. This differs from the approach taken by some alternative schemes for reasoning under uncertainty, where simplicity of reasoning is often achieved by making simplifying, often arbitrary assumptions (such as independence assumptions embedded in Dempster–Shafer theory and possibility theory) [28]. Ultimately, it is hard to determine in these schemes whether possibly counterintuitive or wrong advice is the result of errors in the model or errors introduced by the reasoning algorithm.

Qualitative belief propagation, presented in Section 5, appears to be easy to follow for people and it can be used for generation of verbal explanations of probabilistic reasoning. The individual signs along with the signs of influences can be translated into natural language sentences describing paths of change from the evidence to the variable of interest. Explanation of each step involves reference to a usually familiar causal

```
Qualitative influence of greasy engine block on worn piston rings:
Greasy engine block is evidence for oil leak.  Oil leak and excessive oil
consumption can each cause low oil level.  Oil leak explains low oil level
and so is evidence against excessive oil consumption.  Decreased likelihood
of excessive oil consumption is evidence against worn piston rings.
Therefore, greasy engine block is evidence against worn piston rings.
```

Figure 7: Example of qualitative explanations.

or diagnostic interaction of variables. In general, explanations based on qualitative reasoning are easier to understand than explanations using numerical probabilities. So even where a quantified BBN is available, it may often be clearer to reduce it to the qualitative form, and base explanations on purely qualitative reasoning. An example of a qualitative belief propagation-based explanation is given in Figure 7. More details on generation of verbal explanations of reasoning based on qualitative belief propagation can be found in [4, 7].

Another method for generating explanations is based on the observation that in most models there is usually a small number of very likely states (this was discussed in Section 6). If there is a small number of very likely states, most likely states of the model can be identified and presented to the user. This is the essence of *scenario-based explanations* [6, 16]. An example of a scenario-based explanation is given in Figure 8.

8 Conclusion

I have described five properties of probabilistic knowledge representations that are useful, if not crucial, for intelligent systems research. Probability theory is based on sound qualitative foundations that allow for capturing the essential properties of a domain, along with its causal structure. Directed probabilistic graphs model explicitly independences and tie probability with causality, allowing for a concise and insightful representation of uncertain domains. Probabilistic knowledge representations and reasoning do not need to be quantitative — there is a whole spectrum of possible levels of specifying models, ranging from independence or relevance to full numeric specification. The amount of specificity in a model can be made dependent on available information and a reasoning agent can dynamically move between different levels of specification to do the most with the least possible effort. Concepts such as relevance and conflicting evidence

```
The observed low oil level can be caused by excessive oil consumption or by
oil leak.

Scenarios supporting excessive oil consumption are:
1.  There is no oil leak, excessive oil consumption causes low oil level
(p=0.35).
2.  Cracked gasket causes oil leak, excessive oil consumption and oil leak
cause low oil level (p=0.15).
3.  Other, less likely scenarios (p=0.05).

Scenarios disproving excessive oil consumption are:
1.  Cracked gasket causes oil leak, there is no excessive oil consumption,
oil leak causes low oil level (p=0.36).
2.  Loose bolt causes oil leak, there is no excessive oil consumption, oil
leak causes low oil level (p=0.04).
3.  Other, less likely scenarios (p=0.05).

Therefore, excessive oil consumption is more likely than not (p=0.55).
```

Figure 8: Example of scenario-based explanations.

have in probabilistic representations a natural, formally sound meaning. Finally, probabilistic knowledge representations directly support user interfaces. Their structural properties make it possible to refer to the causal structure of the domain. Full numerical specification of a domain, if available, allows for manipulating with the level of precision for the sake of simplification. The view that probability theory is a numerical scheme, difficult to comprehend for humans, requiring a prohibitive number of expert judgments, and demanding high computational power seems, as I have argued, to be misplaced.

Acknowledgments

Partial support for this work was provided by ARPA's Computer Aided Education and Training Initiative under grant N66001–95–C–8367. Anonymous reviewers provided useful remarks.

References

[1] I.A. Beinlich, H.J. Suermondt, R.M. Chavez, and G.F. Cooper. The ALARM monitoring system: A case study with two probabilistic inference techniques for belief networks. In *Proceedings of the Second European Conference on Artificial Intelligence in Medical Care*, pages 247–256, London, 1989.

[2] William J. Clancey. Use of MYCIN's rules for tutoring. In Bruce G. Buchanan and Edward H. Shortliffe, editors. *Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project*, chapter 26, pages 464–489. Addison-Wesley, Reading, MA, 1984.

[3] Marek J. Druzdzel. Some properties of joint probability distributions. In *Proceedings of the Tenth Annual Conference on Uncertainty in Artificial Intelligence (UAI–94)*, pages 187–194, Seattle, WA, July 1994.

[4] Marek J. Druzdzel. Qualitative Verbal Explanations in Bayesian Belief Networks. To appear in *Artificial Intelligence and Simulation of Behaviour Quarterly*, special issue on belief networks, 1996.

[5] Marek J. Druzdzel and Linda C. van der Gaag. Elicitation of Probabilities for Belief Networks: Combining Qualitative and Quantitative Information. In *Proceedings of the Eleventh Annual Conference on Uncertainty in Artificial Intelligence (UAI–95)*, pages 141–148, Montreal, Canada, August 1995.

[6] Marek J. Druzdzel and Max Henrion. Using scenarios to explain probabilistic inference. In *Working notes of the AAAI–90 Workshop on Explanation*, pages 133–141, Boston, MA, July 1990. American Association for Artificial Intelligence.

[7] Marek J. Druzdzel and Max Henrion. Belief propagation in qualitative probabilistic networks. In N. Piera Carrete and M.G. Singh, editors, *Qualitative Reasoning and Decision Technologies*, pages 451–460. CIMNE, Barcelona, 1993.

[8] Marek J. Druzdzel and Max Henrion. Efficient reasoning in qualitative probabilistic networks. In *Proceedings of the 11th National Conference on Artificial Intelligence (AAAI–93)*, pages 548–553, Washington, D.C., July 1993.

[9] Marek J. Druzdzel and Max Henrion. Intercausal reasoning with uninstantiated ancestor nodes. In *Proceedings of the Ninth Annual Conference on Uncertainty in Artificial Intelligence (UAI–93)*, pages 317–325, Washington, D.C., July 1993.

[10] Marek J. Druzdzel and Herbert A. Simon. Causality in Bayesian belief networks. In *Proceedings of the Ninth Annual Conference on Uncertainty in Artificial Intelligence (UAI–93)*, pages 3–11, Washington, D.C., July 1993.

[11] Marek J. Druzdzel and Henri J. Suermondt. Relevance in probabilistic models: "backyards" in a "small world". In *Working notes of the AAAI 1994 Fall Symposium Series: Relevance*, pages 60–63, New Orleans, Louisiana, November 1994. (An extended version of this paper is currently under review.)

[12] Dan Geiger, Thomas S. Verma, and Judea Pearl. Identifying independence in Bayesian networks. *Networks*, 20(5):507–534, August 1990.

[13] Max Henrion. Uncertainty in artificial intelligence: Is probability epistemologically and heuristically adequate? In Jeryl Mumpower, Ortwin Renn, Lawrence D. Phillips, and V.R.R. Uppuluri, editors, *Expert Judgment and Expert Systems, Proceedings of the NATO Advanced Research Workshop on Expert Judgment and Expert Systems*, pages 105–129, Porto, Portugal, August 1986. Berlin, Germany: Springer Verlag.

[14] Max Henrion. Some practical issues in constructing belief networks. In L.N. Kanal, T.S. Levitt, and J.F. Lemmer, editors, *Uncertainty in Artificial Intelligence 3*, pages 161–173. Elsevier Science Publishers B.V. (North Holland), 1989.

[15] Max Henrion and Marek J. Druzdzel. Qualitative and linguistic explanation of probabilistic reasoning in belief networks. In *Proceedings of the Third International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU)*, pages 225–227, Paris, France, July 1990.

[16] Max Henrion and Marek J. Druzdzel. Qualitative propagation and scenario-based approaches to explanation of probabilistic reasoning. In P.P. Bonissone, M. Henrion, L.N. Kanal, and J.F. Lemmer, editors, *Uncertainty in Artificial Intelligence 6*, pages 17–32. Elsevier Science Publishers B.V. (North Holland), 1991.

[17] Daniel Kahneman, Paul Slovic, and Amos Tversky, editors. *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge University Press, Cambridge, 1982.

[18] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers, Inc., San Mateo, CA, 1988.

[19] Leonard J. Savage. *The Foundations of Statistics (Second Revised Edition)*. Dover Publications, New York, NY, 1972.

[20] Ross D. Shachter. Evaluating influence diagrams. *Operations Research*, 34(6):871–882, November–December 1986.

[21] Ross D. Shachter and David E. Heckerman. A backwards view for assessment. In J.F. Lemmer and L.N. Kanal, editors, *Uncertainty in Artificial Intelligence 2*, pages 317–324. Elsevier Science Publishers B.V. (North Holland), 1988.

[22] Herbert A. Simon. Causal ordering and identifiability. In William C. Hood and Tjalling C. Koopmans, editors, *Studies in Econometric Method. Cowles Commission for Research in Economics. Monograph No. 14*, chapter III, pages 49–74. John Wiley & Sons, Inc., New York, NY, 1953.

[23] Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*. Springer Verlag, 1993.

[24] Henri Jacques Suermondt. *Explanation in Bayesian Belief Networks*. PhD thesis, Department of Computer Science and Medicine, Stanford University, Stanford, CA, March 1992.

[25] William R. Swartout. XPLAIN: a system for creating and explaining expert consulting programs. *Artificial Intelligence*, 21(3):285–325, September 1983.

[26] Michael P. Wellman. Fundamental concepts of qualitative probabilistic networks. *Artificial Intelligence*, 44(3):257–303, August 1990.

[27] Michael P. Wellman and Max Henrion. Explaining "explaining away". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(3):287–292, March 1993.

[28] Ben P. Wise and Max Henrion. A framework for comparing uncertain inference systems to probability. In L.N. Kanal and J.F. Lemmer, editors, *Uncertainty in Artificial Intelligence*, pages 69–83. Elsevier Science Publishers B.V. (North Holland), 1986.