In *Artificial Intelligence and Simulation of Behaviour Quarterly,*
Special issue on Bayesian belief networks, 94:43-54, 1996

Qualitative Verbal Explanations in Bayesian Belief Networks

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

Abstract

Application of Bayesian belief networks in systems that interact directly with human users, such as decision support systems, requires effective user interfaces. The principal task of such interfaces is bridging the gap between probabilistic models and human intuitive approaches to modeling uncertainty. We describe several methods for automatic generation of qualitative verbal explanations in systems based on Bayesian belief networks. We show simple techniques for explaining the structure of a belief network model and the interactions among its variables. We also present a technique for generating qualitative explanations of reasoning.

Keywords: Explanation, Bayesian belief networks, qualitative probabilistic networks

1 Introduction

The purpose of computing is insight, not numbers.
Richard Wesley Hamming

As the increasing number of successful applications in such domains as diagnosis, planning, learning, vision, and natural language processing demonstrates, Bayesian belief networks have secured their position as effective and practical representations of knowledge for reasoning under uncertainty. In many of these applications, systems are fairly autonomous and their most important characteristic is the ultimate reasoning performance. Other applications, such as those in decision support systems, involve interaction with human users. Because in such cases the system fulfills merely an advisory role, the ultimate decision is made by the user. It is, therefore, crucial that users be able to understand the underlying probabilistic model, its assumptions, and its recommendations. In addition to improving the overall acceptability of the system [22], effective user interfaces have a significant impact on the ultimate quality of decisions [14]. The insight gained during the interaction is even more important than the actual recommendation. In addition, good user interfaces facilitate understanding of a model's assumptions and limitations, allowing to discover its shortcomings and potential improvements.

Even though probability theory satisfactorily models most types of uncertain reasoning, it seems to be, as a large body of empirical evidence indicates, remote from the cognitive representation of uncertainty [12]. The magnitude of probabilistic influence and sometimes even its direction are often counterintuitive. One possible way to avoid this problem is to forego probabilistic methods when building systems that interact with human users and to try to imitate the reasoning of human experts. This underlies much expert systems work. Presumably, if the system's reasoning imitates a human expert's reasoning, it should be relatively easy to produce explanations comprehensible to the user. However, the literature demonstrating that probability theory is not a good model of human reasoning also demonstrates that human reasoning under uncertainty is liable to systematic biases and inconsistencies. Indeed, this is part of the argument for using normatively based decision aids to help improve on unaided human intuition. The descriptive approach carries the danger that, along with imitating humans, we may imitate their cognitive biases. This poses a dilemma: must we choose between the soundness and the intuitiveness of inference methods? Must we compromise the performance for understandability?

The challenge to build effective interfaces to probability theory has been accepted by several researchers. One of the first approaches to explanation of probabilistic reasoning was based on the concept of evidence weight (e.g., [10, 18, 20]), which is the logarithm of the likelihood ratio describing the influence of the finding on the variable of interest. Weights of evidence combine additively and provide a simple foundation for graphical explanations of numerical results. Other approaches concentrated on verbal (e.g., [8, 16, 19]) or graphical (e.g., [3]) methods for explaining single-step Bayesian updating. Madigan et al. [15] proposed a collection of graphical methods for explaining inference in graphical models. Of verbal explanations, two types addressed large models (as opposed to local interactions): scenario-based approach [4, 11] and belief propagation-based approach [11, 21].

In this paper, we concentrate on qualitative verbal explanations. We distinguish two components of explanations in the context of probabilistic systems: explanation of assumptions and explanation of reasoning. *Explanation of assumptions* focuses on communicating the system's model of the domain. An important role of this component of explanation is addressing the differences between this model and the user's beliefs about the domain. This is essential for both improving users' insight into the domain and finding possible errors in the model. *Explanation of reasoning* focuses on describing how conclusions are being extracted from the assumptions coded in the original model and the observed evidence. The main purpose here is to explain probabilistic belief updating, i.e., explaining the posterior probability that an outcome of a variable has assumed in light of the available evidence.

The complexity of generation of verbal explanations cannot be underestimated. For an explanation to be effective and create an efficient human–computer collaboration that is the basis for successful use, its form and content must carefully match users' competence, knowledge, and styles of reasoning. The language used in explanations needs to resemble the language used by a human expert explaining his or her advice. The goal of this paper is far more modest: it shows the technical foundations for automatic generation of explanations and provides a vehicle for building explanations that are normatively sound, yet comprehensible. These foundations can be extended to address all aspects of the interaction and used for building complete user interfaces.

The remainder of this paper consists of three parts. Section 2 covers three useful tools that will be used in generating qualitative verbal explanations: qualitative probabilistic networks, methods for focusing explanations on the most relevant parts of the graphical model, and translation of numerical probabilities into verbal phrases. Section 3 discusses explanation of assumptions, i.e., explanation of probabilistic models. Section 4 discusses explanation of reasoning, i.e., explanations of probabilistic updating.

2 The Basics

2.1 Qualitative Probabilistic Networks

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 unfounded as 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 mere constraints on the joint probability distributions. Many forms of probabilistic reasoning, such as reasoning about structure, including independence, relevance, and conflicting evidence, is often purely qualitative and based only on the structure of the directed probabilistic graph. Commonly used qualitative abstraction of Bayesian belief networks (BBNs) are Qualitative Probabilistic Networks (QPNs) [24]. 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 (denoted by $S^+(a, b)$), if observing a to be true makes b more probable. Two direct predecessors of c, a and b exhibit a negative *product synergy* with respect to a value c_0 of c (denoted by $X^-(\{a, b\}, c_0)$) if given c_0 observing a make b less likely. The practical implication of a negative product synergy is that it forms a sufficient condition for a common pattern of reasoning known as *explaining away*. Explaining away is when given an observed effect and increase in probability of one cause, other causes of that effect become less likely. The qualitative properties are captured in formal definitions expressed in terms of sets of constraints on the joint probability distribution [6, 24]. QPNs generalize in a straightforward manner to multivalued and continuous variables. QPNs can replace or supplement quantitative 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 a straightforward process to identify the qualitative relations inherent in it, based on the formal probabilistic definitions of the properties.

Figure 1 shows an example of a QPN that captures the interaction of various variables related to car engine oil. All variables in the example are binary. *Worn piston rings* can cause *excessive oil consumption*, which, in turn, can result in observing *blue exhaust* and *low oil level*. *Low oil level* can be also caused by an *oil leak* (e.g., through a *loose bolt* or a *cracked gasket*). *Oil leak* and possible *oil spills* during adding or replacing oil can give the *engine block* a *greasy* look. *Low oil level* is indicated by an *oil gauge*, but this and the *car radio* will work correctly only if the *battery power* is on. All relations in this model are uncertain. Links in a QPN are labeled by signs of the qualitative influences, while each pair of links coming into a node is described by the signs of the synergies between them (not pictured). There is a negative product synergy between *excessive oil consumption* and *oil leak* with respect to *low oil level*. Even though they can be assumed to be probabilistically independent, once we know that the oil level is low, this independence vanishes. Upon obtaining additional evidence for *oil leak* (e.g., observing *greasy engine block*), we find the likelihood of *excessive oil consumption* diminished — *oil leak* "explains away" *excessive oil consumption*.

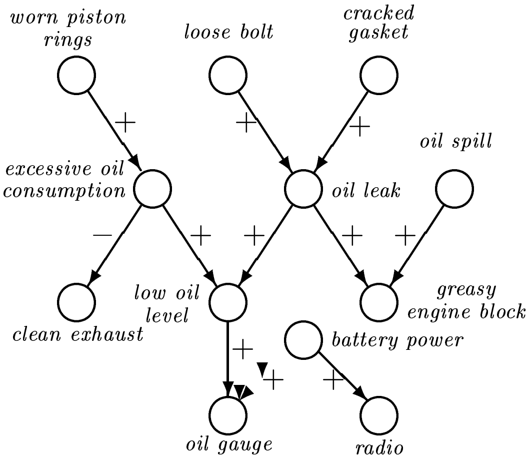

Figure 1: An example of a qualitative probabilistic network

2.2 Relevance for Explanations

It is not unusual for probabilistic models to include hundreds or thousands of variables. Each of these variables may be relevant for some types of reasoning, but rarely will all of them participate in reasoning related to a single query. Focusing on the most relevant part of the model is crucial in communicating its results: too many irrelevant facts will have a confounding effect on most users. It is important, therefore, to identify a subset of the model including only those elements of the domain model that are directly relevant to a particular problem. Druzdzel and Suermondt [7] recently summarized various 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. We provide below only the flavor of these methods, referring the interested readers to [7] for details and references to original sources.

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.

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. These can be identified using 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.

A probabilistic graph is not always capable of representing independences explicitly [17]. The d-separation criterion assumes, for example, that an instantiated head-to-head node makes its predecessors probabilistically dependent. This is not the case, for example, for a common type of interaction known as Noisy–OR gate, when the common effect has been observed to be absent [6]. A careful study of the probability distribution matrices in a graph may reveal similar circumstances and further opportunities for reduction. Procedures for this examination follow in a straightforward manner from the probabilistic definition of independence.

There is another class of variables that can be reduced when the focus of reasoning is explanation. This class consists of those predecessor nodes that do not take an 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. 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 reduced.

The above methods do not alter the quantitative properties of the underlying graph 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 nodes that have non-zero but small impact on t and pruning them can lead to a further simplification of the graph with only a slight impact on the precision of the conclusions.

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.

2.3 Verbal Expressions of Uncertainty

When the time comes to explain the relations among uncertain events, the magnitude of their associated probabilities must somehow be conveyed. One appealing approach to render numerical probabilities more digestible is to translate them into verbal phrases, often the preferred means of communication between people.

There are several empirical studies[1] that report the mapping between absolute numerical probabilities (both point probabilities and probability ranges) and verbal expressions of uncertainty. In general, these studies have found a considerable within-subject consistency in the usage of the verbal phrases, although this is very sensitive to the context in which the expressions are used. Between-subject consistency has been observed to be much lower, although the ordering of the expressions according to their intended or perceived numerical equivalents seems to persist across subjects and contexts (with exception of such ambiguous phrases as “possible” or “probable”).

In the implementation of the explanations that are the subject of this paper, we include translation tables between numerical ranges and verbal expressions of uncertainty. An example of such a table, a compilation of the least ambiguous expressions found in the literature, is given in Figure 1. Other conversion tables between numerical and verbal probabilities were derived from empirical studies. Each probability interval has two

Probability Range	Adjective	Adverb
0.0	*impossible*	*never*
0.0 — 0.1	*very unlikely*	*very rarely*
0.1 — 0.25	*unlikely*	*rarely*
0.25 — 0.4	*fairly unlikely*	*fairly rarely*
0.4 — 0.5	*less likely than not*	*less often than not*
0.5	*as likely as not*	*as often as not*
0.5 — 0.6	*more likely than not*	*more often than not*
0.6 — 0.75	*fairly likely*	*fairly often*
0.75 — 0.9	*likely*	*commonly*
0.9 — 1.0	*very likely*	*very commonly*
1.0	*certain*	*always*

Table 1: Sample mapping from probability ranges to verbal expressions of uncertainty.

[1] See [23] or the monograph by Krause and Clark [13] for some pointers to the empirical literature on verbal expressions of uncertainty.

phrases associated with it: one in adjectival form and one in adverbial form. In addition to simple correspondence between verbal and numerical probabilities, we implemented tables for verbal expressions of relative likelihood (such as “more likely than” or “a great deal more likely than”) proposed by Elsaesser and Henrion [9]. To address the variability of phrases to different contexts, the data structure for each node in the belief network includes a pointer to a translation table that reflects the best the context of the model with respect to the variable. Verbal expressions of uncertainty used in the subsequent examples will all be obtained automatically from numerical values using the translation tables described in Table 1.

Due to variations between people and contexts, there is some inevitable vagueness inherent to most verbal expressions. Some users may find a verbal expression attractive, while others my find the same verbal expression disturbing. An effective interface to a quantitative system should include a variety of formats, including both numerical and verbal expressions, and should allow users to indicate their preferred mode of communication. One option, used in the examples throughout this paper, is using both verbal and numerical expressions of uncertainty.

3 Explanation of the Model

A graphical belief network is in itself a clear representation of the structure of the model. Directed arcs depict dependences between the variables they connect. This information can be expressed verbally as follows:

```
Low oil level depends directly on excessive oil consumption and
oil leak.
```

As far as the structure of the domain is concerned, the graphical form seems to be superior to the verbal description, although the former may be supplemented by the latter in cases where the graphical model is very large. Also, verbal explanations of the structure of the network may aid model debugging: if a verbal description sounds counterintuitive, it may mean that some part of the model contains a structural error. Such an error may sometimes be easier to catch when studying a verbal description of the model.

Verbal explanations of the model structure may be valuable whenever the explanation involves both the dependences and their magnitude, as the latter is not reflected in the directed graph. We might explain the prior probability of the nodes in the network of Figure 1 as follows:

```
Cracked gasket is unlikely (p=0.15).
Loose bolt is very unlikely (p=0.08).
```

Simple influences along with their magnitude can be explained as follows:

```
Cracked gasket more likely than not (p=0.55) causes oil leak.
Worn piston rings commonly (p=0.8) cause excessive oil
consumption.
```

Comparisons between likelihood of various outcomes can be made using the verbal expressions of relative likelihood [9], such as the following:

```
Cracked gasket is more likely than loose bolt (0.15/0.08).
Excessive oil consumption is a great deal more likely than oil
leak (0.05/0.002).
```

An explanation system should be able to distinguish between different types of links in the graph in order to be able to support, for example, cause-effect relations (*worn piston rings* cause *excessive oil consumption*), supposed causes (smoking is believed to cause lung cancer), risk factors (lack of exercise is a risk factor for heart disease), relations like IS-A (cat is an animal), HAS-A (cat has a tail), etc. This information can be easily collected and included in the data structure at the model-building stage. In the simple implementation described in this paper, we did not make this distinction, but rather concentrated on explaining causal relations. As earlier experiences with explanations (e.g., [2]) indicated, causal domain models are easier for the system to explain and for their users to comprehend.

Observed evidence usually changes the structure of the model. By the properties of conditional independence, parts of the model may become independent and other parts may become conditionally dependent. And so, even though *oil leak* and *oil spill* are independent a-priori, they become dependent after *greasy engine block* has been observed. Similarly, *blue exhaust* is probabilistically related to *low oil level*: observing *blue exhaust* makes us alert about possible *excessive oil consumption* and, therefore, possible *low oil level*. Once we know for sure that *excessive oil consumption* is true, however, observation of *blue exhaust* does not bring any new information to our expectation of *low oil level*. Although knowledge of independence and dependence of variables is very basic and believed to be cognitively robust, an explanation system should be prepared for restating them in those cases where a user has doubts about the independence or dependence between two variables. An explanation may clear the doubts or lead to the discovery of an error in the model. Conditional dependence of *oil leak* and *oil spill* can be explained as follows:

```
Oil leak and oil spill are possible causes of greasy engine
block.  Given the observation of greasy engine block, they are
dependent.
```

Conditional independence of *blue exhaust* and *low oil level* can be explained as below:

```
Excessive oil consumption causes blue exhaust.  Excessive oil
consumption causes low oil level.  Given excessive oil
consumption, blue exhaust and low oil level are independent.
```

4 Explanation of Reasoning

Explanation of reasoning consists of explaining how the system reached its conclusions from the assumptions encoded in the model and from the observed evidence. In probabilistic systems, the main form of reasoning is Bayesian updating, i.e., changing the belief in light of new evidence. This is the main focus of the explanations described in this section.

In the explanations, we will focus on explaining the qualitative effect of observing the state of a single evidence variable on the probability distribution for a single target variable. If there are more than one evidence variable, we can explain their impact on the target variable successively, adding each piece of evidence one at a time, concluding with a summary of the joint effect. Similarly, if we are interested in more than one target, we can generate an explanation for the effect on each target individually.

4.1 Qualitative Belief Propagation

Druzdzel and Henrion [5] proposed an efficient algorithm for reasoning in QPNs, called *qualitative belief propagation.* The idea of the algorithm is based on studies of verbal protocols of human subjects solving simple problems involving reasoning under uncertainty. 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 2 gives an example of how the algorithm works in practice. Suppose that we

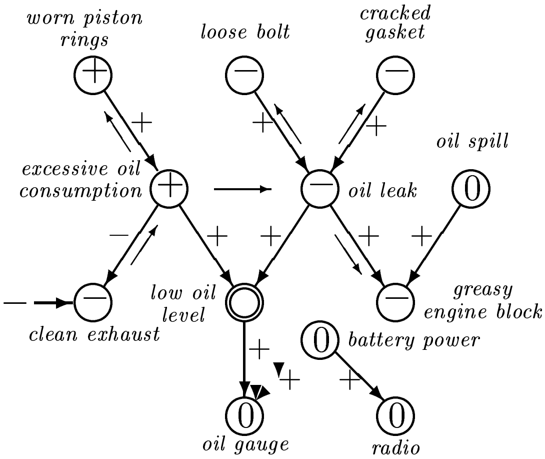

Figure 2: Example of qualitative belief propagation.

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 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* (the sign is determined by the positive sign of *excessive oil consumption* and the negative sign of product synergy between *excessive oil consumption* and *oil leak*). 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 2) 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.

4.2 Basic Steps in Generation of Explanations

Generation of a qualitative belief propagation-based explanation of reasoning in a quantitative belief network involves three steps (see Figure 3). First, given a query that involves a single evidence node e and a single target node t, we extract from the original belief network the qualitative influence graph, which is a graph consisting of all trails between e and t. This can be easily accomplished using the methods outlined in Section 2.2. The algorithm for qualitative sign propagation is then invoked to mark all nodes in the qualitative influence graph with the sign of belief change. In effect, each node in this graph, including t, will be labeled with the sign of impact of e on its probability. A qualitative verbal explanation is generated as a list of elementary qualitative inferences along with the summary conclusion.

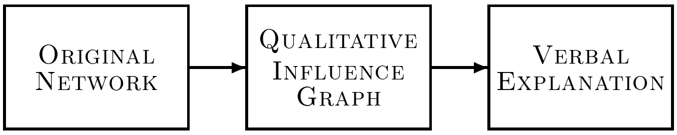

Figure 3: Three steps in generation of qualitative belief propagation-based explanations.

It is possible that the qualitative belief propagation algorithm will be unable to resolve all signs on the active trails from the evidence to the target. In such case, the explanation scheme can rely on a quantitative algorithm to determine these signs. An additional benefit of applying a quantitative algorithm is that it will produce the magnitude of change (i.e., the difference between the prior and the posterior probability of each node) in addition to its sign. This can be used to prune less relevant nodes and links before entering the text generation stage.

4.3 Elementary Qualitative Inferences

Explanations based on qualitative belief propagation mimic, in a sense, the qualitative belief propagation algorithm outlined in Section 4.1. There are three types of elementary explanations corresponding to the three types of qualitative inference: predictive, diagnostic, and intercausal inference. These three are presented schematically in Figure 4.

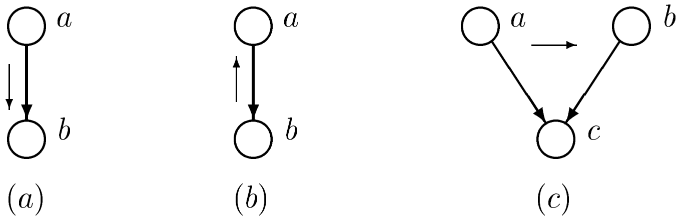

Figure 4: Three elementary inferences in explanations based on qualitative belief propagation: (a) predictive, (b) diagnostic, and (c) intercausal.

Predictive (or causal) inference (Figure 4-a) is in the direction of the arc in the belief network. The fact that increased belief in a causes an increased belief in b is expressed verbally as:

`Increased probability of` a `makes` b `more likely`

or simply

a `can cause` b`.`

Diagnostic inference (Figure 4-b) is in the reverse direction. The fact that increased belief in b results in an increased belief in a is expressed verbally as:

`Increased probability of` b `makes` a `more likely`

or simply

b `is evidence for` a`.`

Intercausal inference (Figure 4-c) gives the qualitative impact of evidence for one variable a on another variable b when both influence a third variable c, about which we have independent evidence (for example, c has been observed). The fact that, given evidence for c, increased belief in a results in an decreased belief in b (explaining away) is expressed verbally as:

a `and` b `can each cause` c`.`
a `explains` c `and so is evidence against` b`.`

Selection of the appropriate phrases that describe the propagation in the most natural and accurate way is heuristic in nature and involves careful experimentation to discover what linguistic forms people find most natural and comprehensible. There are several variants of the qualitative belief propagation-based explanations possible, three of the most obvious are symmetric, causal, and evidential. Symmetric explanations avoid any mention of asymmetry between the variables and speak purely in terms of influences, e.g., "a makes b more likely." Causal explanations follow the causal directions of influences and use causal language, e.g., "a causes b." In cases where the direction of the arrow is evidential, the explanation takes a passive form, such as "a can be caused by b." Evidential explanations use purely evidential terms, e.g., "a is evidence for b." So far, the most naturally sounding explanations have been either symmetric explanations (this may be explained by the fact that by being neutral and symmetric, although not completely natural, they never sound wrong) or combinations of the three types. This suggests that the optimal type of explanation might be data dependent and should perhaps be elicitated at the model-building stage on a per-influence basis.

4.4 Example

Suppose that *low oil level* has been observed and we are interested in the impact of seeing *greasy engine block* on the likelihood of *excessive oil consumption*. The first step is to reduce the model to the qualitative influence graph, shown in Figure 5. Subsequently, the qualitative belief propagation algorithm is invoked to mark the nodes in the graph with the sign of observing *greasy engine block* (a positive sign applied to *greasy engine block*). The signs are marked on the graph. Finally, the elementary qualitative inferences made in the algorithm are translated into verbal form, yielding the following explanation of impact of observing *greasy engine block* on the possibility of *worn piston rings*, given that *low oil level* has been observed earlier.

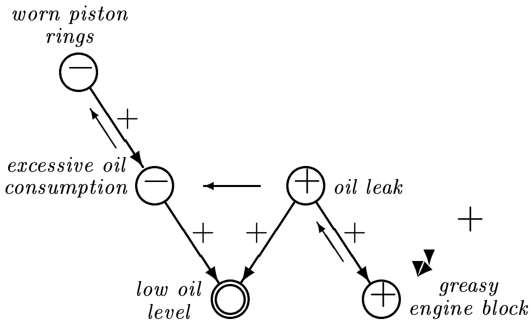

Figure 5: Example explanation: Qualitative influence graph.

```
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.
```

4.5 Coping With Multiple Connections

In case of multiply connected graphs, the explanation system is faced with the necessity to serialize the multiple paths of influence from the evidence to the target variable. The question of how this should be done is empirical in nature and needs to be extensively tested with human users. In the simple introspective solution that we applied, explanations are progressively generated for all nodes located between the evidence and the node of interest. At each point, the node is selected that (1) has at least two incoming paths, and (2) is at the closest distance to the evidence node in terms of the number of links on the shortest active trail from the evidence. The sign of this node is explained in terms of the signs coming from the different parallel trails from the previously explained nodes (that includes the evidence node). Explanation proceeds in stages, each of which is concentrated on the nodes with multiple incoming or outgoing arcs.

Suermondt [21] provides a thorough quantitative treatment of several issues that are complementary to the qualitative belief propagation-based explanations. He concentrates his explanation on the concept of *chains of reasoning*, which are direct active paths in the graph between the evidence node e and the node of interest t. The multiply connected graph between e and t is broken into parts at places where all parallel paths run through one single node (called a *knot*). Multiply connected segments of the graph between knots are explained using a heuristic technique that selects as the next node for explanation the one that is a member of the highest number of active trails.

5 Conclusion

We have presented several methods for automatic generation of qualitative verbal explanations in systems based on Bayesian belief networks. We have shown simple techniques for explaining the structure of a belief network model and the interactions among its variables. We also presented a technique for generating qualitative explanations of reasoning.

Bayesian belief networks offer several advantages for building user interfaces and, in particular, automatic generation of explanation of reasoning. Their structural properties make it possible to encode and to refer to the causal structure of the domain. Concepts such as relevance and conflicting evidence have in probabilistic representations a natural, formally sound meaning. Availability of the numerical specification of the model allows us to study the model 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. In qualitative belief propagation-based explanations, we decreased the precision to signs and used these to explain the significance of evidence for the variables in question. 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) [25]. 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.

The long-term goal of the research described in this paper is to provide a user interface to a quantitative probabilistic reasoner consisting of a variety of explanation schemes and communication methods. The idea behind this variety of schemes is to avoid dogmatic commitment to one single scheme, leaving the choice to the user. Any specific explanation will be too concise for some users, too verbose for others, and confusing for yet others. Users are normally the best judges of their preferences, abilities, and interests, so leaving the choice of style and form to them may work the best. It is important to provide a variety of schemes and many parameters that are adjustable to individual tastes.

6 Acknowledgments

While I am solely responsible for possible shortcomings of this paper, I feel indebted to Max Henrion for numerous interactions that had a large impact of the direction and scope of this work. Support for this work was provided by the National Science Foundation under grant IRI–8807061 to Carnegie Mellon University and by ARPA's Computer Aided Education and Training Initiative under grant N66001–95–C–8367 to University of Pittsburgh.

References

[1] Bruce G. Buchanan and Edward H. Shortliffe, editors. *Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project.* Addison-Wesley, Reading, MA, 1984.

[2] William J. Clancey. Use of MYCIN's rules for tutoring. In Buchanan and Shortliffe [1], chapter 26, pages 464–489.

[3] William G. Cole. Understanding Bayesian reasoning via graphical displays. In *SIGCHI Convention*, Austin, TX, April 1989.

[4] 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, 1990. American Association for Artificial Intelligence.

[5] 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., 1993.

[6] 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., 1993.

[7] 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, LA (An extended version of this paper is currently under review.), 1994.

[8] Christopher Elsaesser. Explanation of probabilistic inference. In L.N. Kanal, T.S. Levitt, and J.F. Lemmer, editors, *Uncertainty in Artificial Intelligence 3*, pages 387–400. Elsevier Science Publishers B.V. (North Holland), 1989.

[9] Christopher Elsaesser and Max Henrion. Verbal expressions for probability updates: How much more probable is "much more probable"? In M. Henrion, R.D. Shachter, L.N. Kanal, and J.F. Lemmer, editors, *Uncertainty in Artificial Intelligence 5*, pages 319–328. Elsevier Science Publishers B.V. (North Holland), 1990.

[10] David E. Heckerman, Eric J. Horvitz, and Bharat N. Nathwani. Toward normative expert systems: Part i. the Pathfinder project. *Methods of Information in Medicine*, 31:90–105, 1992.

[11] 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.

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

[13] Paul Krause and Dominic Clark. *Representing Uncertain Knowledge: An Artificial Intelligence Approach.* Kluwer Academic Publishers, Dordrecht, The Netherlands, 1993.

[14] Paul E. Lehner, Theresa M. Mullin, and Marvin S. Cohen. A probability analysis of the usefulness of decision aids. In M. Henrion, R.D. Shachter, L.N. Kanal, and J.F. Lemmer, editors, *Uncertainty in Artificial Intelligence 5*, pages 427–436. Elsevier Science Publishers B.V. (North Holland), 1990.

[15] David Madigan, Krzysztof Mosurski, and Russell G. Almond. Explanation in belief networks. (under review), 1994.

[16] Steven W. Norton. An explanation mechanism for Bayesian inferencing systems. In John F. Lemmer and Laveen N. Kanal, editors, *Uncertainty in Artificial Intelligence 2*, pages 165–174. Elsevier Science Publishers B.V. (North Holland), 1988.

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

[18] James A. Reggia and Barry T. Perricone. Answer justification in medical decision support systems based on Bayesian classification. *Computers in Biology and Medicine*, 15(4):161–167, 1985.

[19] Peter Sember and Ingrid Zukerman. Strategies for generating micro explanations for Bayesian belief networks. In *Proceedings of the 5th Workshop on Uncertainty in Artificial Intelligence*, pages 295–302, Windsor, Ontario, August 1989.

[20] David J. Spiegelhalter and Robin P. Knill-Jones. Statistical and knowledge-based approaches to clinical decision-support systems, with an application in gastroenterology. *Journal of the Royal Statistical Society*, 147, Part 1:35–77, 1984.

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

[22] Randy L. Teach and Edward H. Shortliffe. An analysis of physicians' attitudes. In Buchanan and Shortliffe [1], chapter 34, pages 635–652.

[23] Thomas S. Wallsten and David V. Budescu. A review of human linguistic probability processing: General principles and empirical evidence. *The Knowledge Engineering Review*, 10(1):43–62, March 1995.

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

[25] 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.