The Knowledge Engineering Review, Vol. 17:2, 107–127. © 2002, Cambridge University Press
DOI: 10.1017/S026988890200019X Printed in the United Kingdom

A review of explanation methods for Bayesian networks*

CARMEN LACAVE[1] and FRANCISCO J DÍEZ[2]

[1] *Dept. Computer Science, University of Castilla-La Mancha, Paseo de la Universidad, s/n 13071 Ciudad Real, Spain; e-mail: clacave@inf-cr.uclm.es*
[2] *Dept. Artificial Intelligence, UNED, Senda del Rey, 9, 28040 Madrid, Spain; e-mail: fjdiez@dia.uned.es*

Abstract

One of the key factors for the acceptance of expert systems in real-world domains is the ability to explain their reasoning (Buchanan & Shortliffe, 1984; Henrion & Druzdzel, 1990). This paper describes the basic properties that characterise explanation methods and reviews the methods developed to date for explanation in Bayesian networks.

1 Introduction

Expert systems originated in the 1970s as computer programs capable of imitating human experts and even substituting them when necessary. One of the essential qualities of real experts is their ability to communicate their knowledge and explain their reasoning. This ability is especially important in the case of expert systems, not only for tracing performance during the construction and evaluation of the system, but also for justifying their results when the system is deployed in an operating environment. In fact, an experiment performed at the MYCIN project showed that physicians are very reluctant to accept the advice of a machine if they do not understand how it was obtained (Teach & Shortliffe, 1984).

In the decades that followed, i.e. the 1980s and 1990s, the main goal of artificial intelligence shifted from *imitating* natural intelligence to *supporting* human beings in a synergistic way. In fact, Clancey (1993) points to the notes to authors' in the *Knowledge Acquisition* journal: "The key issue is not *artificial* intelligence, but how to extend *natural* intelligence through knowledge-based systems." In this context, explanation in expert systems becomes even more important, because human-computer collaboration requires mutual understanding: machine models must take into account human cognitive processes and at the same time must make artificial reasoning understandable to human users. It is surprising, however, that the amount of research devoted to this subject has been relatively small compared to other areas of artificial intelligence. There are only isolated pieces of work, seldom used in real-world applications; in fact, the explanation capability of most of today's expert systems and packages is even poorer than that of MYCIN, which is often regarded as the first expert system.

There was another shift in the 1980s, due to the needs of uncertain reasoning in virtually all applications of expert systems, especially in medicine. It was shown that MYCIN's model of certainty factors was inconsistent and might lead to wrong results. The model of PROSPECTOR was no better, and Dempster-Shafer theory, besides lacking a universally accepted interpretation, was impractical for real-world problems – at least until graphical methods came on the scene. Only fuzzy-set methods seemed to be useful for building expert systems in uncertain domains, but due to the lack of a clear

* This work has been partially supported by the Spanish CICYT under project TIC-97-1135-C04. We thank Marek Druzdzel and the anonymous reviewers of the *Knowledge Engineering Review* for their comments on this paper.

https://doi.org/10.1017/S026988890200019X Published online by Cambridge University Press

semantics and of a consistent methodology, many researchers considered those methods a heterogeneous set of ad hoc solutions without a firm theoretical basis. In the same decade, advances in Bayesian networks (Castillo, 1997; Jensen, 1996; Pearl, 1988) and influence diagrams (Howard, 1984; Shachter, 1986b) showed theoretically and empirically that it was feasible to build probabilistic expert systems without introducing unrealistic assumptions of independence. The advantages of these models are that they have a clear interpretation, easily combine subjectively estimated probabilities and statistical data and can be justified on theoretical grounds. In contrast, the main disadvantage is that their reasoning method follows a normative approach and, consequently, the explanation of inference is more difficult than in the methods that try to imitate human reasoning. Therefore the need for explanation methods is even more important in Bayesian networks than in heuristic expert systems.

The aim of this paper is to analyse what has been done to date and what remains to be done in the field of explanation in Bayesian networks. (In a future work we will review explanation methods for heuristic expert systems.) Therefore, after introducing very briefly Bayesian networks in Section 2.1 and explanation in Section 2.2, we study the fundamental properties of explanation in Section 3. In the light of such properties, we review the methods of explanation for Bayesian networks proposed in the literature (Section 4), and conclude by pointing out possible lines for future research (Section 5).

2 Preliminaries

2.1 *Bayesian networks*

A Bayesian network consists of an acyclic directed graph (ADG),[1] whose nodes represent random variables, together with a conditional probability distribution for each node X_i given its parents,[2] $P(x_i|pa(x_i))$. The conditional probability for a node without parents is just its prior probability $P(x_i|\varnothing) = P(x_i)$. These probabilities can be obtained from statistical data (for instance, from a database), from the literature on the specific domain or by the judgement of human experts.

The joint probability represented by a Bayesian network is

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

This distribution satisfies the *d-separation* property (Pearl, 1988) and its equivalent, the Markov property (Jensen1996; Neapolitan1990), which states that a node is independent of its non-descendants in the graph given its parents. Roughly speaking this property implies that a link $X \rightarrow Y$ in a Bayesian network represents a probabilistic dependence between X and Y, while the lack of a link represents a probabilistic independence.

A *finding* is a piece of information that states with certainty the value of a random variable; a finding may be, for example, that the patient is a male; other findings might be that he is 54 years old, that he has a fever, that he does not usually have headaches, etc. The set of findings is called *evidence*, **e**. Probabilistic reasoning consists of computing the posterior probability of the unobserved variables given the evidence; for instance, $P(x_i|\mathbf{e})$ or $P(x_i, x_j, x_k|\mathbf{e})$. This process is usually called *evidence propagation*, and is based, more or less explicitly, in the application of Bayes theorem.

2.2 *Defining explanation*

From the origins of philosophy and science, many researchers of both disciplines have tried to determine the meaning of explanation, connecting it with comprehension and description. According

[1] A cycle is a closed directed path $X_i \rightarrow X_j \rightarrow X_k \rightarrow \ldots \rightarrow X_i$. A restriction of Bayesian networks is that their graphs cannot contain cycles.

[2] A node X_i is a parent of X_j if and only if there is a directed link $X_i \rightarrow X_j$. It is usual to represent a configuration of the parents of X_i by $pa(x_i)$.

https://doi.org/10.1017/S026988890200019X Published online by Cambridge University Press

to the *Concise Oxford Dictionary of Current English*, an explanation is "a statement or circumstance that explains something | declaration made with a view to mutual understanding or reconciliation" For the *Webster's New World Dictionary*, some meanings of the term *explain* are "to make clear, plain or understandable | to give the meaning or interpretation of; expound | to account for; state reasons for; to give an explanation".

After these definitions, we can conclude that explaining consists of **exposing something** in such a way that it is

understandable for the receiver of the explanation, which implies that he/she improves his/her knowledge about the object of the explanation; and
satisfactory as far as it covers the receiver's expectations.

In the field of expert systems, there are different concepts of explanation. For instance, the rule-based expert system MYCIN (Buchanan & Shortliffe, 1984), the first artificial intelligent program able to explain its reasoning, could show *how* it had obtained some conclusion (i.e. which rules it had applied to deduce a certain proposition) or *why* it was requesting additional information from the user (i.e. which rule it was trying to trigger). For some authors, such as Pearl (1988), the best explanation is the most probable assignment of values to a set of variables – the process of obtaining this kind of explanation is also called *abduction* (Charniak & Shimony, 1994; Gámez, 1998; Santos, 1991; Shimony, 1991). Other systems try to offer a simple but comprehensible report about the domain and the reasoning (Carolis *et al*., 1996; Chandrasekaran *et al*., 1989; Horvitz *et al*., 1986; Langlotz *et al*., 1988; Reggia & Perricone, 1985; Strat, 1987; Swartout, 1983; Wick & Slagle, 1989, Wick & Thompson, 1992). Finally, in the most sophisticated methods, explanation constitutes an 'intelligent' dialogue with the system user through natural language by way of interactive methods (Carenini *et al*., 1995; Cawsey, 1991; Cawsey, 1993; Cawsey, 1994; Cawsey *et al*., 1993; Fiedler, 2001; Moore, 1994).

3 Properties of explanation

In this section we define ten properties of explanation, which constitute the framework for the analysis of the methods studied in Section 4. Some of these properties are taken from Wick (1989), and we have classified them in three categories (see Table 1):

content: **what** to explain;
communication: **how** the system interacts with the user;

Table 1 Properties of explanation methods.

Category	Property	Options
Content	Focus	evidence/model/reasoning
	Purpose	description/comprehension
	Level	micro/macro
	Causality	causal/non-causal
Communication	User-system interaction	menu/predefined questions /natural language dialogue
	Display of explanations	text/graphics/multimedia
	Expressions of probability	numeric/linguistic/both
	User's knowledge about the domain	no model/scale /dynamic model
Adaptation	User's knowledge about the reasoning method	no model/scale /dynamic model
	Level of detail	fixed/threshold/auto

https://doi.org/10.1017/S026988890200019X Published online by Cambridge University Press

adaptation: to **whom** the explanation is addressed.

Please note that these categories correspond to the definition of explanation given in Section 2.2: exposing something (*content*) in such a way that it is understandable (*communication*) to the receiver of the explanation, which implies that he/she improves his/her knowledge about the object of the explanation, and it is satisfactory as far as it covers the receiver's expectations (*adaptation*). We briefly describe them in the following subsections.

3.1 Content

One of the most important aspects of explanation is related to *what* is going to be explained, i.e. what an explanation must include to be understandable to the user. This question has a difficult answer because it depends on several issues such us the focus of explanation, the purpose, the level of detail and causality, which we are going to analyse.

3.1.1 Focus of explanation

In the case of Bayesian networks (and indeed for any expert system), there are three basic issues that can (and must) be explained: the knowledge base, the reasoning process performed by the system to obtain (or not) a conclusion and the evidence propagated, if any. For this reason, we can classify explanation methods in three groups, differing according to the focus taken on explanation:

- Explanation of **evidence** consists of determining which values of the unobserved variables justify the available evidence. This process is usually called *abduction*, and it is based on the (usually implicit) assumption that there is a causal model – see Section 3.1.4. In this context, an *explanation* is a configuration of the unobserved variables, and the goal of the inference process is to obtain the most probable explanation (MPE) or the k most probable explanations. In general, the variables that take the value "present" or "positive" in the MPE are considered to be the causes that *explain* the evidence. The purpose of this kind of explanation is basically to offer a diagnosis for a set of observed anomalies. For instance, in medical expert systems, an explanation consists of determining the disease or diseases that explain the evidence: symptoms, signs, test results, etc.
- Explanation of the **model** is sometimes called *static* (Henrion & Druzdzel, 1990) and consists of displaying (verbally, graphically or in frames) the information contained in the knowledge base. One of its objectives is to assist human experts in the *construction* of expert systems; for instance, Elvira's explanation facility has been used to debug HEPAR-II, a medical Bayesian network (Lacave *et al*., 2001). The other objective of explaining the model is to offer a novice user some knowledge about the domain for *instructional purposes*.
- Explanation of **reasoning** is sometimes known as *dynamic explanation* (Henrion & Druzdzel, 1990) and may provide three kinds of justification:
  - The **results obtained** by the system and the **reasoning process** that produced them. During the *evaluation* of the expert system, some of the test cases are properly solved by the system and some of them are not. In the cases properly solved, the evaluators (knowledge engineers and human experts) can check that all the steps on the way to a solution were correct (an incorrect argument might lead to a correct solution by accident). In the case that the expert system makes a mistake, the explanation of the reasoning process is an invaluable tool for isolating and correcting the wrong pieces of information in the knowledge base. During the *deployment* of an expert system, the explanation capability is very useful for convincing the user of the correctness of the results; in particular, physicians are very reluctant to accept the advice of a machine if they can not understand how it was obtained (Teach & Shortliffe, 1984). In the case of *tutoring systems*, the explanation capability is essential for improving the student's skills.
  - The **results not obtained** by the system, despite the user's expectations. It may also be of interest, especially for educational systems, to explain why the system did not produce a

https://doi.org/10.1017/S026988890200019X Published online by Cambridge University Press

certain conclusion expected by the user; in particular, which findings oppose such a conclusion and which findings would be necessary to support it.

Hypothetical reasoning, i.e. what results the system would have returned if one or more given variables had taken on different values from those observed.

3.1.2 Purpose

There are two different goals an explanation capability might try to accomplish, which leads to two types of explanation:

Description This kind of explanation consists in showing the underlying *knowledge base* (in the case of model explanation), or providing further details on the *conclusions* or displaying *intermediate results* (in the case of reasoning explanation).

Comprehension In this case, the explanation tries to make the user understand the implications of the model or the conclusions of the system and/or the relation between them. For instance, how each finding affects the conclusion, individually or in conjunction with other findings.

3.1.3 Levels of explanation

Sember and Zukerman (1990) proposed another classification of explanation methods for Bayesian networks, thought it might also be applied to other types of expert system:

Micro level A micro-level explanation consists of a detailed justification of the variations produced in a particular node as a consequence of the variations in its neighbours. In the case of a rule-base expert system, the micro level would consist of analysing the variables contained in a certain rule, or the rules containing a certain variable.

Macro level This level of explanation analyses the main lines of reasoning (the paths in the Bayesian network) that lead from the evidence to a certain conclusion. A macro-level explanation for a traditional expert system would consist of one or several chains of rules.

3.1.4 Causality

From a mathematical point of view, a Bayesian network is just a model for representing probabilistic dependencies and independencies; in this case, a link, considered by itself, has no meaning. However, when a Bayesian network is built as a model of a real world system, a link $A \rightarrow B$ is *causal* when A is a cause of B, i.e. when there is a mechanism by which the value taken on by A influences the value of B. A Bayesian network is said to be causal when all of its links are causal. The reasons for using causal models in artificial intelligence, especially in probabilistic expert systems, are the following:

Human beings tend to interpret events in terms of cause-effect relations (Kahneman *et al.*, 1982; Pennington & Hastie, 1988). Therefore causal models are easier to construct and modify (Henrion, 1989; Pennington & Hastie, 1988), and also more easily understood by users (Druzdzel, 1993; Suermondt, 1992).

The identification of *invariant*[3] causal relationships in a domain allows the prediction of effects of both spontaneous causes (usually corresponding to random variables) and actions (sometimes called manipulations or interventions) (Pearl, 1999).

Causality and probability are closely related, because causality normally implies a pattern of probabilistic interdependencies, which provides clues about causality. In fact, a necessary condition for establishing the presence of causality is statistical correlation (Druzdzel, 1993).

In the same line, the axiomatic properties of Bayesian networks (*d*-separation and the Markov property) correspond to probabilistic dependencies and independencies that appear in causal domains (Díez, 1999).

[3] "Invariant" means that when one mechanism is subjected to changes, the others remain intact.

https://doi.org/10.1017/S026988890200019X Published online by Cambridge University Press

There exist canonical probabilistic models (noisy OR, noisy MAX, noisy AND, etc.) based on the interpretation of the parents of a node as causes or conditions for that node and on the assumption of independence of causal interactions, although they might also be used as mere probabilistic models. These models reduce the number of parameters of the network and simplify the acquisition of knowledge, and some of them lead to a more efficient computation (Díez & Druzdzel, 2002).

Causal Bayesian networks support certain qualitative reasoning patterns, which can be identified in order to explain the results of inference (Druzdzel, 1993; Druzdzel, 1996; Henrion & Druzdzel, 1990). Some of these patterns are specific of certain canonical models; for instance, *explaining away* is a phenomenon typical of the noisy OR

Finally, the concept of explanation is very closely tied to the notion of causation; in fact, one of the modalities of scientific explanation consists of finding the causes of the observed facts.

In summary, some of the explanation methods for Bayesian networks are specifically designed for causal networks, or even for specific canonical models, while other methods are general, in the sense that they do not assume a causal interpretation of the network.

3.2 Communication

A crucial aspect of explanation is the way in which it is offered to the user. This depends, basically, on the methods of interaction between the user and the system and the manner in which it is presented. Of particular interest in the case of Bayesian networks is also the way in which probabilities are expressed.

3.2.1 User–system interaction

In some systems, users can request explanations by selecting some variables or options from a certain menu. In other systems, users can pose questions; for instance, in MYCIN the available questions were "how [did the system arrive at a conclusion]" and "why [is the system asking for this piece of information]". Finally, other systems try to offer a natural language dialogue by analysing and building a model of the conversation, in order to "understand" the context of each question and answer.

On the other hand, some systems only allow the user to ask for an explanation after the program has presented its conclusions, while other systems allow the user to interrupt the process of inference and request an explanation.

3.2.2 Display of explanations

We can distinguish how explanations are presented to the user in the following ways:

Verbally, by using text and numbers.

Graphically, in two forms:

by using bar diagrams, pie charts, plots or any other graphical tools that represent the relations between values or variables, the evolution of the probabilities associated with the variables during the evidence propagation, etc.

by using the Bayesian network's own graph, so that changes in the probability distributions are indicated by adding text, numbers or signs, or by colouring the nodes and/or the links.

Multimedia, integrating the above, plus text, numbers and graphs, in a hypertext and multimedia environment that combines interactive explanations with images, video and sound.

3.2.3 Expression of probability

Expressions of probability may be *numerical* or *quantitative*, such as 0.98 or 72.6%, and *linguistic* or *qualitative*, such as "seldom", "very likely" or "almost sure". Similarly, when comparing two probabilities it is possible to have quantitative expressions, such as "the odds between A and B are 10^{-3}", or qualitative expressions, such as "B is much more probable than A". There is a significant amount of research on the assignment of linguistic expressions of probability to numeric values and

https://doi.org/10.1017/S026988890200019X Published online by Cambridge University Press

vice versa – see Druzdzel (1989) and Druzdzel (1993) for a review, and also Section 4.3.1 in this paper.

3.3 Adaptation

Explanation always means explaining something to *somebody*. Therefore one of the key features of an effective explanation is the ability to address each user's specific needs and expectations, which essentially depends on the knowledge he/she has. There are three issues to be considered independently.

3.3.1 User's knowledge about the domain

Some explanation methods do not take into account the variability of domain knowledge between different users. Therefore, explanations are generated for a hypothetical user having a certain knowledge: some explanation methods are intended for beginners, others assume the user is an expert.

Other methods rely on a **static user model**, i.e. they do not consider that the user's knowledge increases as he/she interacts with the system. In many cases, the user model often resorts to using a **scale** with two or three categories: "novice, advanced, expert". In theory, an expert system could have the capability of classifying users automatically, but in practice, the user generally has to classify himself/herself in this scale at the beginning of his/her interaction with the system.

Finally, other explanation methods possess a **dynamic model** for each user, which explicitly represents his/her knowledge and changes as he/she proves to have learned new concepts or relations during his/her operation with the expert system. This model allows the explanation module to generate the explanation adapted to the knowledge that the user has at each moment.

3.3.2 User's knowledge about the reasoning method

Analogously, it would be possible for an explanation method to take into account the user's knowledge about the reasoning method employed by the expert system. In the case of a Bayesian network, the explanation generated for a user that is familiar with the concepts of prevalence, prior/posterior odds and likelihood ratios should be very different from the explanation generated for a user who has never heard about them. As in the case of knowledge about the domain, the possibilities rank from not considering the variations in user's knowledge about the reasoning method to having a dynamic model that explicitly represents the user's knowledge at each moment.

3.3.3 Level of detail

The level of detail of an explanation is closely related to the user's knowledge about the domain and about the reasoning method, but can also be established independently. For instance, it is possible for an expert system that does take into account the user's knowledge to assign an importance factor to each item (each rule or each variable, for instance) and an importance threshold, so that the explanation module only displays those items above the threshold. By lowering the threshold, the user can increase the level of detail, and vice versa. Thus it is possible to offer different levels of detail without having a user model. Obviously, instead of having a fixed importance factor for each item, it would be desirable to dynamically adjust the importance factors as a function of the user's domain knowledge, the available evidence and interaction with the system. Similarly, the aspects of explanation related to the reasoning method should also be offered at different levels of detail.

4 Explanation methods for Bayesian networks

In this section we review the techniques developed for generating explanation in Bayesian networks, in the light of the ten properties discussed above. We divide the methods into three groups, according

https://doi.org/10.1017/S026988890200019X Published online by Cambridge University Press

to the focus of explanation (see Section 3.1.1): explanation of evidence, explanation of the model and explanation of reasoning.

4.1 Explanation of evidence: abduction

We first analyse the methods of abduction, which focus on explaining the evidence. As mentioned in Section 3.1.1, in this context an explanation **w** is an assignment of values to all the variables in a certain subset **W** of the variables of the network. Since the values of observed variables are known with certainty, only unobserved variables are the object of scrutiny in abductive methods. The goal of abduction is to find the most probable explanation (MPE), i.e. the configuration **w** with the maximum a-posteriori probability $P(\mathbf{w}|\mathbf{e})$, where **e** is the available evidence. Some methods are able to find the k most probable explanations. When **W** includes all the unobserved variables, the process is known as *total abduction*; otherwise, it is called *partial abduction*.

In principle, the goal of the methods we review in this section is just to find the MPEs, without trying to justify why they are more probable than the others, i.e. the purpose of these methods is **description**, not comprehension. The distinction between micro and macro **level** makes no sense in this case, because there is no analysis of the reasoning process. The interpretation of a configuration (an assignment of values) as an explanation implicitly assumes that there is a **causal** model, so that the variables that take on the value "present" in the MPE are those that *explain* the observed anomalies. However, since these methods are purely mathematical, they can be applied to any network, independently of whether it is causal or not.

With respect to user–system communication, research on abduction in Bayesian networks usually limits itself to finding the MPEs, without paying attention to interaction with and adaptation to the user. For this reason, most of the criteria for analyzing explanation methods (Section 3) do not apply to abduction. At most we can mention that the probability of each explanation is given **numerically**; no attempt has been made to express qualitatively the probabilities.

We discuss in chronological order the most relevant works on abduction. Please note that all these methods only work for discrete variables.

4.1.1 Pearl's π-λ propagation

The first work on explanation in Bayesian networks (Pearl, 1988, Chapter 5) was a method of total abduction based on the property that, for each value x of a given variable X in **W**, there is a best explanation for the rest of the variables, $\mathbf{W}\backslash X$. This property is similar to the principle of optimality in dynamic programming. Therefore the value of X in the MPE can be obtained by finding the best explanation for $\mathbf{W}\backslash X$ and then choosing the best value of X. Since the computations are local for each node, the process can be designed as an exchange of π-λ messages among nodes. The algorithm has linear complexity for the case of polytrees and exponential complexity for networks with loops. However, the search for the MPE may also be very complicated for polytrees, since even variables that are not interesting for the hypothesis must be taken into account. A shortcoming of this method is that it can only find the two MPEs ($k=2$) (Neapolitan, 1990). As a consequence, this method is only adequate when the two MPEs are much more probable than the rest of the explanations.

4.1.2 Linear restrictions system

Santos (1991) proposed a method of total abduction that transforms a Bayesian network into an equivalent linear restriction system. A *restriction system* $L(W)$ is a tuple (Γ, I, ψ) where Γ is a set of variables, I is a finite set of inequalities defined over the variables of Γ, and ψ is a function from $\Gamma \times \{true, false\}$ to $\mathbb{R}$. The restrictions guarantee that each variable takes only one value and that the probability of a configuration of a set of variables is calculated by the corresponding set of conditioned probabilities. The construction of the restriction system is performed in linear time with respect to the number of variables of the Bayesian network. The search of all solutions involves solving a sequence of restriction systems in which each system is derived from the previous one. One of the disadvantages of this method is that, like Pearl's, it assigns values to all variables, including those that are not interesting for the hypothesis.

https://doi.org/10.1017/S026988890200019X Published online by Cambridge University Press

4.1.3 Irrelevance in partial abduction

Shimony (1991) and Suermondt (1992) (see Section 4.4.1) addressed the problem of obtaining explanations consisting of only relevant variables. In the case of partial abduction, Shimony proposed three definitions of irrelevance:

statistical independence: a variable X should not be part of the explanation $\mathbf{w}$ if it does not affect the probability of the evidence: $P(\mathbf{e} \mid \mathbf{w}, x) = P(\mathbf{e} \mid \mathbf{w})$; it is said that X is irrelevant;
δ-independence: it is less restrictive than the former and it indicates that a fact X is irrelevant if given facts (W) are independent of it with a tolerance of δ: $|P(\mathbf{e} \mid \mathbf{w}, x) - P(\mathbf{e} \mid \mathbf{w})| \leq \delta$, and
quasi-independence, which means that a parent node is relevant only if its contribution to the probability of the node it explains is greater than its own prior probability.

The algorithm for finding relevant explanations relies on the fact that the ancestors of a node V that have not been observed must remain without assigning them any value if they do not affect the probability of V (because as we have mentioned they cannot explain V). Furthermore, if a node V is not an ancestor of node N, then V cannot explain N because it is not a possible cause of N. Please note again the assumption of a causal network in this method.

4.1.4 Graphs of weighted Boolean functions

Charniak and Shimony (1994) proved that the problem of abduction in Bayesian networks is equivalent to finding the minimum cost assignment for the variables of a certain *Weighted Boolean Function Acyclic Directed Graph* (WBFADG) that results from a transformation of the Bayesian network together with evidence $\mathbf{e}$. The WBFADG is solved by a best-first algorithm that, applied successively, produces the enumeration of the assignments in increasing order of cost, i.e. the explanations in decreasing order of probability. Two disadvantages of this method are its computational and conceptual complexities.

4.1.5 Approximate partial abduction

Finally, Gámez (1998) has made a detailed study about abduction, both total and partial. Among his most relevant contributions are:

He has studied how to adapt clustering algorithms (clique tree propagation) for partial abduction.
He has developed new approximate algorithms that can be used in those cases in which exact algorithms are inefficient. Those algorithms are based on *genetic algorithms* and *stochastic annealing* (de Campos *et al*., 1999a),
He has proposed some criteria to simplify explanations (de Campos *et al*., 1999b). Those are:
Normality The simplification is based on showing the user only those values that are not the *usual* value of the variable. For example, in medicine, the usual value for the variable `Meningitis` is `absent`.
Probabilistic independence, by analysing the graph that represents the Bayesian network. He has developed an algorithm for detecting some kinds of independency.
Relevance This implies that the omitted information is "almost" irrelevant (in the sense of statistical independence) for the observed facts.
He has proved that the criteria based on relevance and independence produce better results in those cases in which it is not easy to decide which is the *usual* value of a variable.

4.2 Explanation of the model

In Section 3.1.1 we mentioned the advantages that the explanation of the model, also known as *static explanation*, presents in the phase of building the Bayesian network and also with educational purposes. We classify the methods according to the way they present their explanations.

https://doi.org/10.1017/S026988890200019X Published online by Cambridge University Press

4.2.2 Graphical display of the network

The most direct and intuitive way of showing the information embodied in a Bayesian network is to display the corresponding graph: each node is represented by an oval containing the name (or a short description) of the associated variable, and links are drawn as arrows. Some of the earlier tools that permitted the graphical edition and visualisation of Bayesian networks are Analytica, IDEAL (Srinivas & Breese, 1990), DAVID (Shachter, 1986a), HUGIN (Andersen *et al*., 1990) and PATHFINDER (Heckerman, 1991). However, when the size of the network grows large, it becomes more and more difficult to read the graph. For this reason, other packages, such as Analytica and GeNIe (Druzdzel, 1999), offer the possibility of defining submodels. When contracted, a submodel is represented by a special type of node; when expanded, it displays all the nodes and the links it contains.[4]

In any case, this kind of explanation is only a **description** of the model, and it is not intended to improve **comprehension**. The interaction with the user is done by means of **menus**, which give access to the properties of nodes, links and conditional probability tables. The parameters of the model appear in **numerical** form. In the tools we have examined, there is no possibility of **adaptation** to the user, except for minor formatting options.

4.2.2 Verbal description of the network

Druzdzel (Druzdzel, 1993; Henrion & Druzdzel, 1990) proposed a method for translating the qualitative and quantitative information of a Bayesian network into linguistic expressions. There are patterns for indicating the prior probability of a node, such as "`Cold is very unlikely (p=0.08)`" and for comparing probabilities: "`Cold is slightly less likely than [having a] cat (0.08/0.10)`".

In this method, **causal relations** play an important role. In particular, there are explanation patterns for describing the leaky noisy OR, a model that assumes the independence of causal interactions: "`Cold very commonly (p=0.9) causes sneezing. Allergy very commonly (p=0.9) causes sneezing. Cold does not affect the tendency of allergy to cause sneezing, and vice versa. There are also other unlikely (p=0.1) causes of sneezing`." Relations of conditional dependence or independence can be expressed by sentences like this: "`Given sneezing, cold and allergy are independent`."

The purpose of this explanation model lies between **description** and **comprehension**. Explanations are offered at the **micro level**. User–system **interaction** is not described in the references. The presentation of explanation is in the form of **text** containing **linguistic** and **numeric** expressions of probability.

Other systems, like B2 (see Section 4.4.1) and Elvira (see Section 4.2.4) are also capable of offering verbal explanations of the model.

4.2.3 Menu-driven navigation

Díez (1994), in his expert system DIAVAL, developed a method for explaining both the model and the reasoning. His method distinguishes several types of link and node corresponding to the different types of **causal** influence: in the noisy OR gate, the parents of a node are considered *causes* in the strict sense, while in the general model parents are regarded as *factors* that influence the child variable.

User-system interaction is based on a system of windows and **menus**, which offer different options:

For each *node*, the user can view its prevalence (prior probability), posterior probability, list of causes or factors, conditional probability table (for nodes with parents), list of effects (children) and definition.

For each *parameter*, which is a node with an associated measure (generally from echocardiography), the user can see measured value, intervals, pathophysiological meaning, and formula (for those parameters calculated from others).

[4] The URL's for HUGIN, Analytica, GeNIe and other Bayesian network tools can be found at `http://www.ia.uned.es/~fjdiez/bayes/#software`

https://doi.org/10.1017/S026988890200019X Published online by Cambridge University Press

For the *links* that make part of a noisy OR the options are cause, effect, sensitivity and specificity (only for binary variables), efficiency (the probability that the cause produces the effect) and a text that explains the associated causal mechanism.

By visiting the parents (causes) and children (effects) of nodes, the user can navigate across the network. Explanation fits into the **micro level**, because explanation focuses on one node or link at each moment. Explanations are displayed as **text** inside different windows, and probability is only expressed **numerically**.

4.2.4 Static explanation in Elvira

Elvira[5] is an environment for the editing and evaluation of Bayesian networks and influence diagrams, developed as a research project among several Spanish universities. Following Druzdzel's proposal (Druzdzel, 1993) (see Section 4.2.2), Elvira offers **verbal** explanations at the **micro level**. The main difference with his method is that in Elvira the nodes are classified into several categories, such as *symptom* or *disease*. This information is used to generate model explanations. An example of an explanation could be: "`The disease X has the following symptoms: s1, s2, ...`". It is also capable of offering **graphical** explanations, such us nodes expansion, as well as verbal information about specific nodes or links. In a similar way to DIAVAL (Díez 1994), it also allows navigation across the network. The interaction with the user is made by way of **windows** and **menus**.

Another useful option consists of automatically colouring the links of the network, in order to offer qualitative insight about the conditional probability tables (Lacave *et al*., 2001). More specifically, given two ordinal discrete variables A and C such that there is a link $A \rightarrow C$, this link is said to be **positive** if higher values of A lead to higher values of C for any configuration of B, where B represents the set of the other parents of C. The definitions of **negative link** and **null link** are analogous. When the influence is neither positive, nor negative, nor null, then it is said to be **unknown** (Wellman, 1990). Typical orderings of values of a variable are $+a > \neg a$, *present* > *absent*, *severe* > *moderate* > *mild* > *absent*, *positive* > *negative*, etc. If A and C are binary variables, the above definition implies that link $A \rightarrow C$ is positive if and only if $P(+c \mid +a, b) > P(+c \mid \neg a, b)$ for each value b of B. If variable A represents a cause or a risk factor for C, or C is a test that detects A, then influence $A \rightarrow C$ is in general positive. In causal networks, most of the links are positive.

4.2.5 Assisted construction of medical Bayesian networks

We conclude this section on static explanation by mentioning the system MEDICUS, developed by Folckers *et al*. (Folckers *et al*., 1996; Shcroder, 1996) as a tool for the construction of explanation models in which the knowledge is complex and uncertain, as in medicine. The main contribution is that it is a shell in which the user can develop a Bayesian network for representing a certain domain, independently of his/her **level of knowledge** about probability. This is done by means of **micro-level** explanations generated by a linguistic editor of the model, together with a **graphic** editor of the Bayesian network. The explanations are expressed both **quantitatively** and **qualitatively** and are presented **verbally** and **graphically**.

4.3 Explanation of reasoning at the micro level

In this section we describe the methods used to explain the reasoning at the micro level in Bayesian networks. They are characterised by the fact that at each moment they focus only on one variable, usually called *focal hypothesis*, for generating the explanations.

4.3.1 Verbal description of variations in probability

Experimental research has shown that human beings understand linguistic expressions of probability better than numerical data. For this reason, Elsaesser (1990) proposed a method for generating

[5] Information about Elvira can be found at `http://www.ia.uned.es/~elvira`

https://doi.org/10.1017/S026988890200019X Published online by Cambridge University Press

linguistic explanations by means of a template based on Polya's *shaded inductive patterns*, which are a set of heuristic rules for describing changes in probabilities.

For example, one of those rules, compatible with Bayes formula, is: "*If* $A \rightarrow B$ *and B is true*, *then the existence of A is more credible*." The template is filled with natural language terms that represent probabilistic data and their variations. The set of expressions denoting probability vary between those that reflect high probabilities, like *almost certain* (for values between 0.99 and 0.91) or *highly probable* (for the interval 0.90 to 0.82) and to those that designate small values, like *improbable* (for the range 0.18 to 0.09) or *highly improbable* (for the range 0.08 to 0.01). There are also expressions for defining changes from probability p_1 to probability p_2; for instance, when $p_2/p_1 \geq 5$, the expression is "a great deal more likely"; when $2.5 < p_2/p_1 \leq 5$, the expression is "much more likely, " etc.

Later, an experiment by Elsaesser and Henrion (1989) proved that the linguistic expressions assigned by people to different pairs (p_1, p_2) depend on the difference $p_2 - p_1$ more than on the ratio p_2/p_1 or the odds ratio $(p_1/(1-p_1))/(p_2/(1-p_2))$.

Clearly, the translation of numerical probabilities into linguistic expressions makes **no** assumption on **causality**. In this method, intended for users that are not familiar with probability theory, there is no explicit **user model**.

4.3.2 Graphical display of probabilities

There are several software tools for processing Bayesian networks that offer the possibility of showing graphically the variations in probability of certain variables, such as HUGIN, Netica or Elvira. The basic idea common to all of them is to show the variations of probability by plotting a bar proportional to the probability of each state of a node, together with its numerical value. The purpose of this kind of explanation facility is the **description** of the reasoning process and it is useful both for **causal** and **non-causal** Bayesian networks. Probabilities are expressed **quantitatively** and **qualitatively**.

However, Elvira differs from other tools in that it is able to simultaneously store and display several evidence cases (Lacave *et al*., 2000; Lacave *et al*., 2001). The user is allowed to navigate across the set of evidence cases, saving them in files, generating new cases, expanding or contracting the selected nodes, modifying the inference options, etc. This facility permits the user to observe the variations in the probability of each variable due to the sequential introduction of new findings, and to perform "what-if" analysis by introducing different hypothetical findings, even for the same variable. Moreover, when evaluating the impact of evidence, the user can know if the probability of a certain node has increased or decreased with respect to the previous case or to another fixed case, selected by him/her, because nodes are coloured depending on the changes on its probabilities.

The main objective of Elvira's graphical explanation facility is the **comprehension** of the reasoning process and it is useful both for **causal** and **non-causal** Bayesian networks. The interaction with the user is performed by **windows** and **menus**. Probabilities are expressed **quantitatively** and **qualitatively**, by means of coloured bars. There is a rudimentary **adaptation** capability consisting of controlling the expansion and importance thresholds.

4.3.3 Explanation of local updates in polytrees

In a polytree (a network without loops), the posterior probability for variable B can be obtained by

$$Bel(b) = \alpha\lambda(b)\pi(b) \tag{2}$$

(Pearl, 1988) where $\lambda(b)$ is the support from the effects (the children) of B and $\pi(b)$ is the causal support. Sember and Zukerman (1990) developed an explanation method for justifying the value of $Bel(b)$ in terms of $\pi(b)$ and $\lambda(b)$. The user's expectations are featured by the changes produced in π(b), since it represents causal information. If $\pi(b)$ increases, decreases or does not change, then the reader expects $Bel(b)$ to increase, decrease or remain unchanged, respectively. When the expectation is met, their method generates an explanation like this: "*The belief in B has augmented due to an increase in its causal support*." In those cases when the user's expectation is not met, the explanation consists of identifying which values of $\pi(b)$ and $\lambda(b)$ caused the deviation.

https://doi.org/10.1017/S026988890200019X Published online by Cambridge University Press

The purpose of this method is **comprehension** and it assumes a **causal** Bayesian network. Explanations are presented as **text** and variations in the probability are expressed **linguistically** ("has augmented/has decreased"). Sember and Zukerman do not describe the user–system **interaction** nor any possibility of **adaptation**. The explanations require the user to be familiar with the reasoning method.

This method is direct and intuitive in the case of binary variables, but becomes more complicated for multivalued variables. The main limitation of this method is that it only works for polytrees, and therefore can seldom be applied to real-world problems, whose models almost always have loops.

4.3.4 Analysis of local variations of probability in DIAVAL

As we said in Section 4.2.3, the expert system DIAVAL (Díez, 1994) had a method for explaining reasoning which allows the user to select each diagnosis and open a menu with different options, including the possibility of visiting the parents and children of the nodes, as described in Section 4.2.3. This way the user can investigate which neighbours of a node have increased or decreased their probability. Additionally, if a node Y is the child of a noisy OR/MAX gate, the probability that each parent X_i has produced Y can be computed as $P(+x_i \mid \mathbf{e}) \times c_i$, where c_i is a parameter of the OR gate that represents the probability that X_i, when being present, produces Y.

The purpose of this explanation method for **causal** Bayesian networks is mainly to **describe** the results of inference and includes specific explanation patterns for the noisy OR/MAX gates. User–system **interaction** is implemented by a system of windows and menus. Probabilities are expressed **numerically**. There is no user model and no adaptation.

4.3.5 Analysis of the impact of evidence on a variable

One of the first systems that included some kind of explanation was PATHFINDER (Heckerman, 1991). The method basically consists of discriminating between two diseases or two groups of diseases, by showing how each value v_i of a certain variable V affects its probability distribution. To do this, the system draws a graphic bar proportional to the likelihood ratio for each value v_i of the selected variable in favour of one disease D_1 relative to the other D_1 given the evidence e. The measurement unit used is the *evidence weight* proposed by Good (1977).

$$\log\left(\frac{P(v_i \mid D_1, e)}{P(v_i \mid D_2, e)}\right) \tag{3}$$

The purpose of PATHFINDER explanations is the **description** of reasoning and can be applied to both causal and non-causal networks. The presentation of the explanation is **graphical** and the interaction with the user is led by way of menus, windows and dialogue boxes. Probabilities are expressed by **numbers** and there is no adaptation. The user must have some knowledge of probability theory.

Another method was Suermondt's INSITE (Suermondt, 1992; Suermondt & Cooper, 1993), whose main objective is to identify the findings that influence the posterior probability of a certain hypothesis, as well as the paths through which the evidence flows, which we will describe in Section 4.4.1. The influence of evidence $\mathbf{e}$ on a certain variable D is measured by a cost-function. Suermondt (1992) analyses different cost functions and concludes that the most suitable function is cross-entropy,

$$H(P(D \mid \mathbf{e}); P(D)) = \sum_i \left[p(d_i \mid \mathbf{e}) \log\left(\frac{p(d_i \mid \mathbf{e})}{p(d_i)}\right) \right], \tag{4}$$

where d_i represents the possible values of D. The influence of individual findings or subsets of findings is determined by a *sensitivity analysis* which computes the cost of omission of such findings.

The purpose of INSITE is the **comprehension** of reasoning and it can be applied to both causal and non-causal networks. Interaction with the user is led by way of menus, windows, dialogue boxes and

https://doi.org/10.1017/S026988890200019X Published online by Cambridge University Press

buttons, and by selecting nodes or arcs in the network display. Explanations are presented as a combination of **graphics** and **text**. In the latter case, the probabilities are expressed both by **numbers** and **linguistic expressions**.

4.4 Explanation of reasoning at the macro level

The methods analysed below are characterised by generating macro-level explanations, which typically implies that they attempt to explain the main paths in the Bayesian network through which the evidence flows.

4.4.1 Quantitative analysis of reasoning chains

NESTOR, developed by Greg Cooper (1984), is one of the first Bayesian networks. It included a facility for explanation which offered two possibilities: to *compare* how two diagnostic hypotheses account for the evidence, and to *critique* a hypothesis with respect to all other possible diagnoses. In both cases, it could generate a **verbal** explanation for describing **qualitatively** the chains among the evidence and a given hypothesis, which basically consisted of translating into English the causal links of each chain. On the other hand, in the case of the "compare" command, the user could visualise how each finding affected the relative probability of two hypotheses selected by the user. In the case of the "critique" command, the user could visualize the relative probability of the hypothesis with respect to all other hypothesis. The explanation method was defined only for **causal** networks. Explanations were displayed both **graphically** and **verbally**. Probabilities were expressed **numerically**. The interaction with the user was led by way of commands (predefined questions). It had no adaptation capability and no user model.

Graphical display of the weight of evidence In the same line, Madigan *et al*. (1996) propose a **graphical method** that consists of showing how the evidence is propagated through a causal Bayesian network. The impact of the evidence on a binary node H is measured, as in PATHFINDER, by computing Good's *evidence weight*

$$W(H:\mathbf{e}) = \log\left(\frac{P(\mathbf{e}\mid +h)}{P(\mathbf{e}\mid \neg h)}\right) \tag{5}$$

and it is displayed on the network's own graph by varying the colour and thickness of nodes and links, by drawing different kind of links, etc.

The method also provides macro explanations for two types of question. The first one is, "What is the relative importance of each finding f about the variable of interest H?" This is answered by visualizing $W(H{:}f)$ in a graphical evidence balance sheet, in which all findings are shown together with their importance for the hypothesis. Please note that the importance of each finding depends on the order in which the user introduces the evidence.

The second type of question is, "Why does a concrete node have so much influence on a given variable?" The answer consists of showing the relevant paths, in a way similar to Suermondt's *chains of reasoning*.

The purpose of this method is the **comprehension** of the reasoning process. Although Madigan *et al*. speak of **causal** Bayesian networks, their method also works in non-causal networks. User–system interaction is **menu-driven** and probabilities are **qualitatively** codified by means of colours and thicknesses. No adaptation capability has been developed for this method.

INSITE Suermondt's method (see Section 4.3.5) also looks for relevant *chains of reasoning*, i.e. paths from **e** to D that are computationally related to D given evidence **e**. He analyses the *strength*[6] of each of them and whether it conflicts or not with the inference result. The chains are presented graphically

[6] The strength of a chain is viewed as the importance of the chain in obtaining the inference result.

https://doi.org/10.1017/S026988890200019X Published online by Cambridge University Press

to the user by shading the conflicting ones and highlighting the consistent ones. Chains can also be described verbally. Moreover, INSITE defines a method for detecting conflicts among different variables.

As we said, the purpose of INSITE is the **comprehension** of reasoning and the method can be applied to both causal and non-causal networks. Explanations are presented as combinations of **graphics** and **text**, expressing probabilities by means of **numbers** and **linguistic expressions**. The method has no user model but the **level of detail** can be adapted to the needs of a given user.

One interesting aspect of INSITE is its independence of the algorithm of evidence propagation. In contrast, its main shortcoming is the computational complexity, which grows exponentially with the number of nodes and arcs.

BANTER Based on Suermondt's INSITE method, Haddawy, Jacobson and Kahn (Haddawy *et al*., 1994a; Haddawy *et al*., 1994b; Haddawy *et al*., 1997) have developed BANTER, a tool for decision-support and training. The tool is aimed at medicine, although it works on any network consisting of hypotheses, observations and diagnostic tests. Given evidence, BANTER can offer the probability of a hypothesis or select the most informative test for confirming or discarding a hypothetical diagnosis. When used as an educational tool, BANTER randomly generates scenarios and asks the user to make a diagnosis or select a test.

BANTER can also generate verbal explanations (by using a modification of Suermondt's method) by identifying the most influential pieces of evidence and by selecting the strongest[7] and shortest paths between the evidence and the hypothesis.

Both methods share the purpose of **comprehension** of the reasoning process. They differ in that BANTER is designed for **causal** Bayesian networks and only generates **verbal** explanations, while INSITE can also display graphic explanations. In BANTER, the expressions of probability are only **quantitative**. None of them offers any adaptation to the user. As BANTER is especially conceived for novice or inexperienced users, it does not require any knowledge of Bayesian networks, but only some familiarity with the application domain and an elementary understanding of probability.

B2 However, there are some problems when BANTER is used as an educational tool: it is not capable of making hypothetical reasoning (see Section 3.1.1) and it sometimes offers irrelevant information for the conclusion, which can be misleading for the user. On the other hand, verbal explanations are difficult to understand. In order to solve these deficiencies, McRoy *et al*. (McRoy *et al*., 1996; McRoy *et al*., 1997a; McRoy *et al*., 1997b) have developed B2 as an extension of BANTER. It generates both **graphical** and **verbal** explanations in natural language, in a more consistent way than BANTER does. It generates an explanation of the **reasoning**. This is done by the representation of a discourse model by way of a propositional semantic network. This allows the system to generate only relevant information, since it is capable of doing some reasoning about the content and the structure of its interaction with the user.

Therefore, the objective is **comprehension** of the **reasoning**. It also offers the capacity for doing **hypothetical reasoning**. As in BANTER, the probabilities are expressed **quantitatively** and there is no user **adaptation**. User-system **interaction** is performed by windows and questions are expressed in **natural language**.

Probability of evidence Dittmer and Jensen (1997) focus on analysis tools for generating macro explanations in Bayesian networks, but instead of analysing the paths in the network, they compute the probability of evidence in order to detect conflicts among data, and perform a *sensitivity analysis* to determine how changes in the evidence affect the conclusion. They have applied them to the BOBLO Bayesian network for determining if the pedigree assigned to the cattle is correct. The main objective of this method is **description** and it is useful both for causal and non-causal models. We have found no reference in relation to user–system interaction or the possibility of adaptation.

[7] The strength of a path C given evidence e is defined as $Strength(C) = \min |(P(n) - P(n|e)|$, $\forall$ node $n \in C$.

https://doi.org/10.1017/S026988890200019X Published online by Cambridge University Press

4.4.2 Qualitative explanations

Druzdzel and Henrion (Druzdzel, 1993; Druzdzel, 1996; Druzdzel & Henrion, 1993a; Druzdzel & Henrion, 1993b; Henrion & Druzdzel, 1990) also proposed another explanation method similar to the previous one but it is based on the *qualitative* analysis of the reasoning chains. It consists of transforming a causal Bayesian network into a qualitative probabilistic network (QPN) (Wellman, 1990a; Wellman, 1990b), in which the relation between two adjacent nodes is denoted as positive (+), negative (−), null (0) or unknown (?); there are also relations that involve more than two nodes, such as additive (Wellman, 1990a; Wellman, 1990b) or multiplicative synergies (Druzdzel & Henrion, 1993a; Druzdzel & Henrion, 1993b). The main advantage of QPNs is that they simplify the construction of models, because they do not require the elicitation of numerical parameters; as a consequence, their main disadvantage is lack of precision in the results, especially because very often the combination of "positive" and "negative" influences leads to "unknown" influences. The motivation for this explanation method is that people usually reason and explain their reasoning in terms of qualitative relations.

The qualitative propagation algorithm proposed by Druzdzel and Henrion (1993b) consists of exchanging messages among neighbour nodes. Evidence nodes are marked with + or −. Unobserved nodes are initially marked with 0. The sign of the message (+, −, 0 or ?) from X_i to X_j is determined by the product of the sign of X_i and the sign of the link.

The objective of explanation is to determine the qualitative impact that each finding *f* has produced on a certain variable of interest *V* and to find out the active paths from *F* to *V*. There are three kinds of elementary explanations corresponding to the three kinds of elementary qualitative inferences:

predictive inference goes from causes to effects, i.e. in the direction of the links; the explanation for this kind of inference relative to a link $A \rightarrow B$ is of the type "*A* may cause *B*".

diagnostic inference goes from effects to causes, i.e. in the opposite direction of the arcs; the explanation may be "*B* is evidence for *A*".

intercausal inference analyses the qualitative impact of the evidence for a variable *A* over another variable *B* when both have an influence on a third variable *C*, about which there is independent evidence (for example, *C* has been observed); the explanation generated in this case is of the kind: "*A* and *B* may each cause *C*; as *A* explains *C*, there is no evidence for *B*"

If there are several findings and several target variables, the process will be repeated several times. Similarly, if there are several active paths from finding *f* to target variable *V*, explanation is sequenced, since both human reasoning and natural language are essentially sequential.

The purpose of this method is the **comprehension** of the reasoning process. It assumes that the network is **causal**. Explanations are displayed as **text**, which includes **linguistic** expressions of probability. As in the case of scenarios, the authors do not describe user–system **interaction** or any possibility of **adaptation**. In principle, these qualitative explanations do not require the user to be familiar with probabilistic reasoning, although knowledge on the propagation method may help him/her to understand the explanations.

Trade-offs resolution It is also worthwhile to mention the work done by Renooij *et al*. (Renooij & van der Gaag, 1999; Renooij *et al*., 2000) in order to avoid the ambiguous results that may appear when doing inference in a qualitative network. They have designed a formalism for trade-off resolution without making use of numerical information (Renooij & van der Gaag, 1999). It is called *enhanced qualitative networks* and is based on distinguishing between strong and weak influences, in contrast with qualitative networks which do not make this distinction. The way they do this is by associating a relative strength with influences. The algorithm for sign-propagation in this kind of network is a generalisation of that for qualitative networks, although its main difference is that strong influences dominate over conflicting weak influences.

Together with Parsons and Green (in Renooij *et al*., 2000), they have also designed another algorithm for calculating informative results for the node of interest, in order to solve its ambiguous sign after doing inference. It is based on focusing on the subnetwork formed by the chains between the

https://doi.org/10.1017/S026988890200019X Published online by Cambridge University Press

observed node and the node of interest. Then, an informative result is constructed for the *pivot node*,[8] because its sign determines the sign of the node of interest. The ambiguity at the pivot node is solved in terms of the relative strengths of the influences between them as well as the signs of the node's resolvers.

4.4.3 Scenario-based explanation

According to Druzdzel and Henrion (Druzdzel, 1993; Druzdzel & Henrion, 1990; Henrion & Druzdzel, 1990), a *scenario* is an assignment of values to variables that are relevant to a certain conclusion, ordered in such a way that they form a coherent story – a causal story, if possible – compatible with the evidence. An example of a scenario could be: `Age 5, Fever high, Exanthema present`. The use of scenarios is founded on psychological studies (Kahneman *et al.*, 1982; Pennington & Hastie, 1988) showing that humans tend to interpret and explain processes by weighing up the most credible stories that include the hypothesis being demonstrated.

Although a scenario may contain all the nodes in the network, it is more reasonable to include only those nodes that are *relevant* for a certain task (Druzdzel & Suermondt, 1994; Lin & Druzdzel, 1997; Ludwig, 1998; Subramanian *et al.*, 1997). If there is a certain focal hypothesis H selected by the user, the relevant nodes are those that affect the posterior probability of H given the observed evidence **e**. Otherwise, the relevant nodes are all those whose probabilities depend on **e**. After selecting the relevant nodes, the scenarios are generated by some of the methods of partial abduction – see Section 4.1. The explanation consists of showing the evidence, the most probable scenarios compatible with the hypothesis and those incompatible with the hypothesis, and a comparison of the probabilities of the most probable scenarios.

The purpose of scenario-based explanation is the **comprehension** of the reasoning process, even though the propagation algorithm is not based on scenarios. The method is mainly designed for **causal** Bayesian networks. Explanations are presented as a **text** that describes each scenario in natural language. The probabilities of scenarios are given **numerically**. Druzdzel and Henrion do not describe user–system **interaction** or any possibility of **adaptation**. In principle, these explanations do not require the user to be familiar with probabilistic reasoning, although knowledge of the methods involved will certainly help him/her to understand the explanations.

4.5 Other research

Finally, we describe other methods related to the generation of explanations for Bayesian networks.

4.5.1 Selection of diagnoses

Most of the Bayesian network tools limit the process of inference to displaying the posterior probabilities of the values of each variable. This is clearly insufficient for practical expert systems, especially when the network contains a great number of nodes. For this reason, the explanation capability of the PATHFINDER expert system (Heckerman, 1991) (see Section 4.3.5) basically consists of showing only how the evidence affected all the possible diseases. For this purpose, an ordered list with all the possible diseases and their associated probabilities is displayed. These probabilities define the order of the list. The main goal of this method is **description**. The presentation of the explanation is **verbal** and interaction with the user is led by way of menus, windows and dialogue boxes. Probabilities are expressed by **numbers** and there is no user adaptation.

Similarly, DIAVAL (Díez, 1994) also implemented a method for selecting the most probable and relevant diagnoses. Relevance in DIAVAL is a measure of the importance from the medical point of view: for instance, a disease like mitral stenosis is more "relevant" than an intermediate pathophysiological variable, like left atrium hypertension; during the construction of the Bayesian network, each node V is subjectively assigned two relevance factors: the positive (negative) relevance

[8] The pivot node is a node that separates the part of the relevant network that contains the trade-offs from the part that does not.

https://doi.org/10.1017/S026988890200019X Published online by Cambridge University Press

factor applies when the most probable value is $+v$ ($\neg v$). The system only displays the diagnoses whose posterior probability and relevance exceed both the *certainty threshold* and the *relevance threshold*.

The purpose of this method is to **describe** the results of inference although it offers some rudimentary assistance for the **comprehension** of reasoning. The method assumes that the model is **causal**. **Interaction** between the user and the system is based on windows and menus. Probabilities are expressed **numerically**. There is no user model, and the only adaptation capability is that the user can control the **level of detail** in the selection of diagnoses by modifying the relevance and certainty thresholds.

4.5.2 Nice argument generator

Among the ongoing projects, the current research of Zukerman, McConachy and Corb (McConachy *et al*., 1998; Zukerman *et al*., 1998a; Zukerman *et al*., 1998b) stands out. Their method focuses on the use of Bayesian networks to generate arguments expressed in natural language. In their latest investigations, they propose the utilisation of Bayesian networks as the basis of the Nice Argument Generator (NAG) system, which serves to analyse and generate 'good enough' reasoning to convince the user. So, given a goal proposition introduced by the user, the context in which it is made and a grade of credibility, the system generates the arguments to justify it; moreover, it analyses the reasoning provided by the user and prepares other arguments to refute it when necessary. Both the user model and the normative model are represented by Bayesian networks. The new arguments are added to the normative model when they are created. Since both networks can be too complex, there is a mechanism for focusing on the relevant subnetworks, over which reasoning will be performed. The intersection of both network structures defines the reasoning graph. In order to analyse it, a propagation over both networks is made, using the probabilistic information associated with each model.

5 Conclusions

Explanation of reasoning is one of the key factors in the success of expert systems. As we discussed in Section 3.1.1, explanation capability is crucial for debugging a model, for convincing the user that the results are correct and for educational purposes.

There are two main approaches for building expert systems. The *heuristic approach* tries to mimic the reasoning process of human experts, generally by using rules and structured objects. In contrast, the *normative approach* is based on probability and decision theory; in practice, it amounts to using Bayesian networks for diagnostic expert systems and influence diagrams for decision-support systems. There are both theoretical and empirical studies indicating that the normative approach leads to more accurate and robust expert systems, but on the other hand the explanation capability is even more necessary, because normative reasoning methods are more foreign to human beings than heuristic methods.

In neither of the approaches are there explanation methods satisfactory for the end-user. In fact, there are only isolated proposals, partial solutions insufficient to constitute a standard method suitable for all the expert systems that use similar reasoning techniques.

In this paper (Section 3) we have discussed the basic properties of explanation methods and summarised them in Table 1. In the light of these, we analysed explanation methods to date for Bayesian networks by considering three kinds of explanation: explanation of evidence (Section 4.1), explanation of the model (Section 4.2) and explanation of the reasoning (Sections 4.3 and 4.4). The first kind correspond to the philosophic concept of explanation as a search for the causes of a certain phenomenon. The second and third kinds of explanation correspond to the concept of explanation used in artificial intelligence: the processes that help the user understand the model and the performance, respectively, of an expert system.

If we come back to Section 3 and look at Table 1 as a checklist of the features that an explanation capability should offer, we realise that all the methods currently available suffer from serious limitations. A significant part of the research limits itself to theoretical models that have not even been implemented in prototype applications. And in the methods implemented so far, interaction with the

https://doi.org/10.1017/S026988890200019X Published online by Cambridge University Press

user amounts, in the best cases, to displaying a menu with a few options. No explanation method for Bayesian networks has even tried to offer the possibility of a dialogue with the user.

In the same way, no explanation method for Bayesian networks takes into account the user's knowledge, not even by using the distinction between novice and experienced users. (In heuristic expert systems, there are several explanation methods that used either a novice-advanced–experienced scale or a dynamic model that explicitly represents the user's knowledge at each moment.)

The conclusion of this paper is that there is much research to be done in the area of explanation in Bayesian networks. We trust that this paper may stimulate some researchers to build on the current methods (by remedying the shortcomings of one of them or by combining different methods) or to explore new approaches aimed at the problems thus far unaddressed. In our opinion, the most promising lines are the study of causal methods – in particular, canonical models – and the application of user models, both for the domain knowledge and knowledge about the reasoning method.

References

Andersen, S, Olesen, K, Jensen, F and Jensen, F, 1990, "HUGIN: a shell for building belief universes for expert systems" in *Readings in Uncertainty* 332–337.

Buchanan, B and Shortliffe, E (eds), 1984, *Rule-Based Expert Systems*: *The MYCIN Experiments of the Stanford Heuristic Programming Project* Addison-Wesley.

Carenini, G, Mittal, V and Moore, J, 1994, "Generating patient-specific interactive natural language explanations" in *Proceedings of* 18*th Symposium on Computer Applications in Medical Care* (SCAMC 94).

Carolis, BD, Rosis, FD, Grasso, F, Rossiello, A, Berry, D and Gillie, T, 1996, "Generating recipient-centered explanations about drug prescription" *Artificial Intelligence in Medicine* **8** 123–145.

Castillo, E, Gutiérrez, JM and Hadi, AS, 1997, *Expert Systems and Probabilistic Network Models* Springer Verlag, New York.

Cawsey, A, 1991, "Generating interactive explanations" in *Proceedings of the Ninth National Conference on Artificial Intelligence* (AAAI-91) 86–91.

Cawsey, A, 1993, "User-modelling in interactive explanations" *Journal of User Modelling and User Adapted Interaction* **3**(3) 221–247.

Cawsey, A, 1994, "Developing an explanation component for a knowledge-based system: discussion" *Expert Systems with Applications* **8**(4) 527–531.

Cawsey, A, Galliers, J, Logan, B, Reece, S and Jones, KS, 1993, "Revising beliefs and intentions: a unified framework for agent interaction" in *Proceedings of the Conference of the Society for Artificial Intelligence and Simulation of Behaviour* 130–139.

Chandrasekaran, B, Tanner, M and Josephson, J, 1989, "Explaining control strategies in problem solving" *IEEE Expert* **4** 9–24.

Charniak, E and Shimony, S, 1994, "Cost-based abduction and MAP explanation" *Artificial Intelligence* **66**(2) 345–374.

Clancey, WJ, 1993, "Notes on 'heuristic classification'" *Artificial Intelligence* **59** 191–196.

Cooper, G, 1984, "NESTOR: a computer-based medical diagnostic aid that integrates causal and diagnostic knowledge" Ph.D. thesis, Stanford University, Stanford, CA.

De Campos, L M, Gámez, J and Moral, S, 1999a, "Partial abductive inference in Bayesian belief networks using a genetic algorithm" *Pattern Recognition Letters* **20** 1211–1217.

De Campos, LM, Moral, S and Gámez, J, 1999b, "Simplifying explanations in Bayesian belief networks" Technical Report DECSAI-99–02–01, Dept. Ciencias de la Computación e Inteligencia Artificial, Universidad de Granada, Spain.

Díez, F, 1994, "Sistema experto bayesiano para ecocardiografía" Ph.D. thesis, UNED, Dept. Informática y Automática.

Díez, FJ and Druzdzel, M, 2002, "Canonical probabilistic models. Part I: Knowledge engineering" Technical Report, Decision System Laboratory, University of Pittsburgh. In preparation.

Díez, FJ, 1999, "From causal graphs to Bayesian networks" in *Proceedings of the Workshop on Conditional Independence Structures and Graphical Models*.

Dittmer, S and Jensen, F, 1997, "Tools for explanation in Bayesian networks with application to an agricultural problem" in *Proceedings of the First European Conference for Information Technology in Agriculture*.

Druzdzel, M and Henrion, M, 1990, "Using scenarios to explain probabilistic inference" in *Working Notes of the AAAI*-90 *Workshop on Explanation* 133–141.

Druzdzel, M and Henrion, M, 1993a, "Belief propagation in qualitative probabilistic networks" in NP Carrete and M Singh (eds) *Qualitative Reasoning and Decision Technologies* CIMNE.

https://doi.org/10.1017/S026988890200019X Published online by Cambridge University Press

Druzdzel, M and Henrion, M, 1993b, "Efficient reasoning in qualitative probabilistic networks" in *Proceedings of the* 11*th National Conference on Artificial Intelligence* 548–553.

Druzdzel, M and Suermondt, H, 1994, "Relevance in probabilistic models: backyards in a small world" in *Working Notes of the AAAI*-94 *Workshop on Fall Symposium Series*: *Relevance* 60–63.

Druzdzel, M, 1989, "Verbal uncertainty expressions: literature review" Technical Report CMU-EPP-1990–03–02, Dept. Engineering and Public Policy, Carnegie Mellon University.

Druzdzel, M, 1993, "Probabilistic reasoning in decision support systems: from computation to common sense" Ph.D. thesis, Department of Engineering and Public Policy, Carnegie Mellon University.

Druzdzel, M, 1996, "Qualitative verbal explanations in Bayesian belief networks" *Artificial Intelligence and Simulation of Behaviour Quarterly* **94** 43–54.

Druzdzel, M, 1999, "GeNIe: a development environment for graphical decision-analytic models" in *Proceedings of the 1999 Annual Symposium of the American Medical Informatics Association* (AMIA-1999) 1206.

Elsaesser, C and Henrion, M, 1989, "Explanation of probabilistic inference" in L Kanal, T Levitt and J Lemmer (eds) *Uncertainty in Artificial Intelligence* Volume 3 Elsevier Science Publishers.

Elsaesser, C, 1990, "Verbal expressions for probability updates. How much more probable is "much more probable"?" in L Kanal, T Levitt and J Lemmer (eds) *Uncertainty in Artificial Intelligence* Volume 5 Elsevier Science Publishers.

Fiedler, A, 2001, "Dialog-driven adaptation of explanations of proofs" in *Proceedings of the Seventeenth International Joint Conference of Artificial Intelligence* 1295–1300.

Folckers, J, Möbus, C, Schröder, O and Thole, H, 1996, "An intelligent problem solving environment for designing explanation models and for diagnostic reasoning in probabilistic domains" in *Intelligent Tutoring Systems*, 353–362.

Gámez, J, 1998, "Inferencia abductiva en redes causales". Ph.D. thesis, Dept. Ciencias de la Computación e Inteligencia Artificial, Universidad de Granada, Spain.

Good, I, 1977, "Explicativity: a mathematical theory of explanation with statistical applications" in *Proceedings of the Royal Statistical Society*, 354, 303–330.

Haddawy, P, Jacobson, J and Kahn, C, 1994a, "An educational tool for high-level interaction with Bayesian networks" in *Proceedings of the 6th IEEE International Conference on Tools with Artificial Intelligence*.

Haddawy, P, Jacobson, J and Kahn, C, 1994b, "Generating explanations and tutorial problems from Bayesian networks" in *Proceedings of the 18th Annual Symposium on Computer Applications in Medical Care*, 770–774.

Haddawy, P, Jacobson, J and Kahn, C, 1997, "BANTER: a Bayesian network tutoring shell" *Artifical Intelligence in Medicine*, **10** 177–200.

Heckerman, DE, 1991, *Probabilistic Similarity Networks* MIT Press.

Henrion, M and Druzdzel, M, 1990, "Qualitative propagation and scenario-based approaches to explanation of probabilistic reasoning" in *Proceedings of the 6th Conference on Uncertainty in Artificial Intelligence*, 17–32.

Henrion, M, 1989, "Some practical issues in constructing belief networks" in *Uncertainty in Artificial Intelligence*, Volume 3 Elsevier Science Publishers, Amsterdam.

Horvitz, E, Heckerman, D, Nathwani, B and Fagan, L, 1986, "The use of a heuristic problem-solving hierarchy to facilitte the explanation of hypothesis-directed reasoning" in *MEDINFO 86* 27–31. Elsevier Science Publishers.

Howard, RA and Matheson, JE, 1984, "Influence diagrams" in RA Howard and JE Matheson (eds) *Readings on the Principles and Applications of Decision Analysis* Strategic Decisions Group.

Jensen, F, 1996, *An Introduction to Bayesian Networks* UCL Press.

Kahneman, D, Slovic, P and Tversky, A, (eds), 1982, *Judgement under uncertainty: Heuristics and biases*. Cambridge University Press.

Lacave, C, Atienza, R and Díez, FJ, 2000, "Graphical explanations in Bayesian networks" in *Proceedings of the First International Symposium on Medical Data Analysis* (ISMDA 2000) 122–129.

Lacave, C, Oniśko, A and Díez, FJ, 2001, "Debugging medical Bayesian networks with Elvira's explanation capability" in *Proceedings of the Workshop on Bayesian Models in Medicine, Eighth European Conference on Artificial Intelligence in Medicine* (AIME-2001) 47–52.

Langlotz, C, Shortliffe, E and Fagan, L, 1988, "A methodology for generating computer-based explanations of decision-theoretic advice" *Medical Decision Making*, **8**(4) 290–303.

Lin, Y and Druzdzel, M, 1997, "Computational advantages of relevance reasoning in Bayesian belief networks" in *Proceedings of the 13th Conference on Uncertainty in Artificial Intelligence* 342–350.

Ludwig, K, 1998, "Functionalism, causation and causal relevance" *Psyche: An Interdisciplinary Journal of Research and Consciousness*, **4**(3).

McConachy, R, Korb, K and Zukerman, I, 1998, "Deciding what not to say: an attentional-probabilistic approach to argument presentation" in *Proceedings of the Twentieth Annual Meeting of the Cognitive Science Society* 669–674.

https://doi.org/10.1017/S026988890200019X Published online by Cambridge University Press

McRoy, S, Haller, S and Ali, S, 1997a, "B2: an interactive tool for explaining Bayesian reasoning in natural language" in *Conference Proceedings of Energy-Week 97*, Volume V, 153–159, Houston, TX Special Session on Computers in Engineering.

McRoy, S, Haller, S and Ali, S, 1997b, "Uniform knowledge representation for language processing in the B2 system" *Journal of Natural Language Engineering* **3**(2).

McRoy, S, Liu-Perez, A, Helwig, J and Haller, S, 1996, "B2: A tutoring shell for Bayesian networks that supports natural language interaction" in *Working notes of the AAAI 96 Spring Symposium on Artificial Intelligence*.

Madigan, D, Mosurski, K and Almond, R, 1997, "Graphical explanations in belief networks" *Journal of Computational and Graphic Statistics*, **6**(2) 160–181.

Moore, J, 1994, *Participating in Explanatory Dialogues: Interpreting and Responding to Questions in Context* MIT Press.

Neapolitan, R, 1990, *Probabilistic Reasoning in Expert Systems: Theory and Algorithms* Wiley-Interscience.

Pearl, J, 1988, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference* Morgan Kaufmann Publishers.

Pearl, J, 1999, "Reasoning with cause and effect" Technical report, Department of Computer Science. University of California, Los Angeles, California. `http://bayes.cs.ucla.edu/jphome.html`.

Pennington, N and Hastie, R, 1988, "Explanation-based decision making: effects of memory structure on judgment" *Journal of Experimental Psychology: Learning, Memory and Cognition* **14**(3) 521–533.

Reggia, J and Perricone, B, 1985, "Answer justification in medical decision support systems based on Bayesian classification" *Computers in Biology and Medicine* **15**(4) 161–167.

Renooij, S and van der Gaag, L, 1999, "Enhancing QPNs for trade-off resolution" in *Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence* 559–566.

Renooij, S, van der Gaag, L, Parsons, S and Green, S, 2000, "Pivotal prunning of trade-offs in QPNs" in *Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence* 515–522.

Santos, E, 1991, "On the generation of alternative explanations with implications for belief revision" in *Proceedings of the 7th Conference on Uncertainty in Artificial Intelligence* 339–347.

Schröder, O, Mobus, C, Folckers, J and Thole, H, 1996, "Supporting the construction of explanation models and diagnostic reasoning in probabilistic domains" Technical report, OFFIS Institute, Oldenburg, Germany.

Sember, P and Zukerman, I, 1990, "Strategies for generating micro explanations for Bayesian belief networks" in *Uncertainty in Artificial Intelligence* Volume 5 295–302 Elsevier Science Publishers.

Shachter, R, 1986a, "DAVID: Influence diagram processing system for the Macintosh" in *Proceedings of the Workshop on Uncertainty in Artificial Intelligence* 243–248.

Shachter, R, 1986b, "Evaluating influence diagrams" *Operations Research* **34**(6) 871–882.

Shimony, S, 1991, "A probabilistic framework for explanation" Ph.D. thesis, Department of Computer Science, Brown University, Technical Report CS-91–57.

Srinivas, S and Breese, J, 1990, "IDEAL: a software package for analysis of influence diagrams" in *Proceedings of the 6th Workshop on Uncertainty in Artificial Intelligence* 212–219.

Strat, T, 1987, "The generation of explanations within evidential reasoning systems" in *Proceedings of the 10th International Joint Conference on Artificial Intelligence* 1097–1104.

Subramanian, D, Greiner, R and Pearl, J, 1997, "The relevance of relevance" *Artificial Intelligence* **97**(1–2) 1–5.

Suermondt, H and Cooper, G, 1993, "An evaluation of explanations of probabilistic inference" *Computers and Biomedical Research* **26** 242–254.

Suermondt, H, 1992, "Explanation in Bayesian belief networks" Ph.D. thesis, Department of Computer Science, Stanford University, Stanford, CA STAN-CS–92–1417.

Swartout, W, 1983, "XPLAIN: a system for creating and explaining expert consulting programs" *Artificial Intelligence* **21** 285–325.

Teach, RL and Shortliffe, EH, 1984, "An analysis of physician's attitudes" in BG Buchanan and EH Shortliffe (eds) *Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project* Addison-Wesley, 635–652..

Wellman, M, 1990a, "Fundamental concepts of qualitative probabilistic networks" *Artificial Intelligence* **44** 257–303.

Wellman, M, 1990b, "Graphical inference in qualitative probabilistic networks" *Networks*, **20** 687–701.

Wick, M and Slagle, J, 1989, "An explanation facility for today's expert systems" *IEEE Expert* **4** 24–36.

Wick, M and Thompson, W, 1992, "Reconstructive expert system explanation" *Artifical Intelligence* **54** 33–70.

Wick, M, 1989, "The 1988 Workshop on Explanation" *Artifical Intelligence Magazine* **10**(3) 22–26.

Zukerman, I, Korb, K and McConachy, R, 1998a, "Perambulations on the way to an architecture for a nice argument generator" in *Notes of the ECAI-96 Workshop on Gaps and Bridges: New Directions in Planning and Natural Language Generation* 31–36.

Zukerman, I, McConachy, R and Korb, R, 1998b, "Bayesian reasoning in an abductive mechanism for argument generation and analysis" in *Proceedings of the Fifteenth National Conference on Artificial Intelligence* 833–838.

https://doi.org/10.1017/S026988890200019X Published online by Cambridge University Press

https://doi.org/10.1017/S026988890200019X Published online by Cambridge University Press