Explaining Bayesian Networks using Argumentation*

Sjoerd T. Timmer[1], John-Jules Ch. Meyer[1],
Henry Prakken[1,2], Silja Renooij[1], and Bart Verheij[3]

[1] Utrecht University, Department of Information and Computing Sciences
[2] University of Groningen, Faculty of Law
[3] University of Groningen, Artificial Intelligence Institute

Abstract. Qualitative and quantitative systems to deal with uncertainty coexist. Bayesian networks are a well known tool in probabilistic reasoning. For non-statistical experts, however, Bayesian networks may be hard to interpret. Especially since the inner workings of Bayesian networks are complicated they may appear as black box models. Argumentation approaches, on the contrary, emphasise the derivation of results. Argumentation models, however, have notorious difficulty dealing with probabilities. In this paper we formalise a two-phase method to extract probabilistically supported arguments from a Bayesian network. First, from a BN we construct a *support graph*, and, second, given a set of observations we build arguments from that support graph. Such arguments can facilitate the correct interpretation and explanation of the evidence modelled in the Bayesian network.

Keywords: Bayesian networks, argumentation, reasoning, explanation, inference, uncertainty

1 Introduction

Reasoning about probabilities and statistics, and independence in particular, is a difficult task that easily leads to reasoning errors and miscommunication. For instance in the legal or medical domain the consequences of reasoning errors can be severe. Bayesian networks, which model probability distributions, have found a number of applications in these domains (see [9] for an overview). However, the interpretation of BNs is a difficult task, especially for domain experts who are not trained in probabilistic reasoning. Argumentation is a well studied topic in the field of artificial intelligence (see chapter 11 of [12] for an overview). Argumentation theory provides models that describe how conclusions can be justified. These models closely follow the same reasoning patters present in human reasoning. This makes argumentation an intuitive and versatile model

* This work is part of the research programme “Designing and Understanding Forensic Bayesian Networks with Arguments and Scenarios”, which is financed by the Netherlands Organisation for Scientific Research (NWO). See `http://www.ai.rug.nl/~verheij/nwofs/`.

for common sense reasoning tasks. This justifies a scientific interest in models of argumentation that incorporate probabilities. In this paper we formalise a new method to extract arguments from a BN, in which we first extract an intermediate support structure that guides the argument construction process. This results in numerically backed arguments based on probabilistic information modelled in a BN. We apply our method to a legal example but the approach does not depend on this domain and can also be applied to other fields where BNs are used.

In previous work [10] we introduced the notions of probabilistic rules and arguments and an algorithm to extract those from a BN. However, exhaustively enumerating every possible probabilistic rule and argument is computationally infeasible and also not necessary because many of the enumerated antecedents will never be met, and many arguments constructed in this way are superfluous because they argue for irrelevant conclusions. In a report [11] we proposed a new method that solves these issues. We proposed to split the process of argument generation into two phases: from the BN we construct a *support graph* at first, from which argument can be generated in a second phase. We introduced an algorithm for the first phase but the second phase has only been described informally. In the current paper we show a number of properties of the support graph formalisms and we fully formalise the argument generation phase.

In Section 2 we will present backgrounds on argumentation and BNs. In Section 3 we formally define and discuss support graphs. Using the notion of a support graph we present a translation to arguments in Section 4. One of the advantages of this method is that the support graph presents a dynamic model of evidence because when observations are added to the BN it does not need to be recomputed. Only the resulting argumentation changes.

2 Preliminaries

2.1 Argumentation

In argumentation theory, one possibility to deal with uncertainty is the use of defeasible inferences. A defeasible (as opposed to strict) rule can have exceptions. In a defeasible rule the antecedents do not conclusively imply the consequence but rather create a presumptive belief in it. Using (possibly defeasible) rules, arguments can be constructed. Figure 1, for instance, shows (on the left) an argument graph with a number of arguments connected by two rules. From a psychological report it is derived that the suspect had a motive and together with a DNA match this is reason to believe that the suspect committed the crime. Different formalisation of such systems exist [5,7,8,14]. In this paper we will construct an argumentation system where the rules follow from the BN. Since a BN captures probabilistic dependencies the inferences will be defeasible. Figure 1 also shows a possible counter-argument. Undercutting and rebutting attacks between arguments with defeasible rules have been distinguished [7]. A rebuttal attacks the conclusion of an argument, whereas an undercutter directly attacks the inference (as in this example). An undercutter exploits the fact that a rule is not strict by posing one of the exceptional circumstances under which it does not apply. Using rebuttals and undercutters, counter-arguments can be formulated. Arguments can be compared on their strengths to see which attacks succeed as defeats. Then Dung's theory of abstract argumentation [1] can be used to evaluate the acceptability status of arguments.

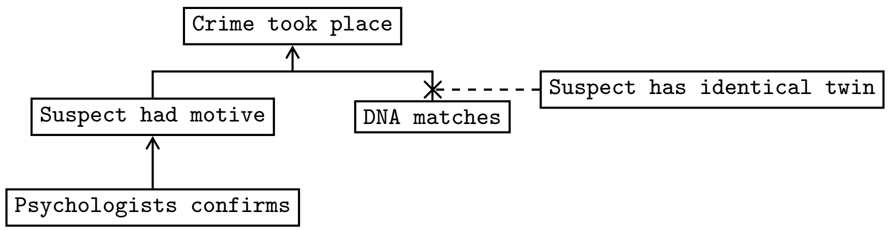

Fig. 1. An example of complex arguments and an undercutting counter-argument.

2.2 Bayesian networks

A Bayesian network (BN) contains a directed acyclic graph (DAG) in which nodes correspond to stochastic variables. Variables have a number of mutually exclusive and collectively exhaustive outcomes: upon observing the variable, exactly one of the outcomes will become true. Throughout this paper we will consider variables to be binary for simplicity.

Definition 1 (Bayesian network). *A Bayesian network is a pair $\langle G, P \rangle$ where G is a directed acyclic graph $(\mathbf{V}, \mathbf{E})$, with variables $\mathbf{V}$ as the set of nodes and edges $\mathbf{E}$, and P is a probability function which specifies for every variable V_i the probability of its outcomes conditioned on its parents $\mathrm{Par}(V_i)$ in the graph.*

We will use $\mathrm{Cld}(V_i)$ and $\mathrm{Par}(V_i)$ to denote the sets of children and parents respectively of a variable V_i in a graph. $\mathrm{Cld}(\mathbf{V}')$ (and $\mathrm{Par}(\mathbf{V}')$) will likewise denote the union of the children (and parents respectively) of variables in a set $\mathbf{V}' \subseteq \mathbf{V}$.

Given a BN, observations can be entered by instantiating variables; this update is then propagated through the network, which yields a posterior probability distribution on all other variables, conditioned on those observations. A BN models a joint probability distribution with independences among its variables implied by d-separation in the DAG [6].

Definition 2 (d-separation). *A* trail *in a DAG is a simple path in the underlying undirected graph. A variable is a* head-to-head *node with respect to a particular trail iff it has two incoming edges on that trail. A variable on a trail* blocks *that trail iff either (1) it is an unobserved head-to-head node without observed descendants, or (2) it is not a head-to-head on that trail and it is observed. A trail is* active *iff none of its variables are blocking it. Subsets of variables $\mathbf{V}_A$*

and $\mathbf{V}_B$ *are* d-separated *by a subset of variables* $\mathbf{V}_C$ *iff there are no active trails from any variable in* $\mathbf{V}_A$ *to any variable in* $\mathbf{V}_B$ *given observations for* $\mathbf{V}_C$*.*

If, in a given BN model, $\mathbf{V}_A$ and $\mathbf{V}_B$ are d-separated by $\mathbf{V}_C$, then $\mathbf{V}_A$ and $\mathbf{V}_B$ are probabilistically independent given $\mathbf{V}_C$. An example of a BN is shown in Figure 2. This example concerns a criminal case with five variables describing how the occurrence of the crime correlates with a psychological report and a DNA matching report. The variables `Motive` and `Twin` model the presence of a criminal motive and the existence of an identical twin. The latter can result in a false positive in a DNA matching test. In the following we will also require the notions of a Markov blanket and Markov equivalence [13].

Definition 3 (Markov blanket). *Given a BN graph, the* Markov blanket $\mathrm{MB}(V_i)$ *of a variable* V_i *is the set* $\mathrm{Cld}(V_i) \cup \mathrm{Par}(V_i) \cup \mathrm{Par}(\mathrm{Cld}(V_i))$*. I.e., the parents, children and parents of children of* V_i*.*

Definition 4 (Markov equivalence). *Given a BN graph, an* immorality *is a tuple* $\langle V_a, V_c, V_b \rangle$ *of variables such that there are directed edges* $V_a \rightarrow V_c$ *and* $V_b \rightarrow V_c$ *in the BN graph but no edges* $V_a \rightarrow V_b$ *or* $V_b \rightarrow V_a$*. Given two BN graphs, they are* Markov equivalent *if and only if they have the same underlying undirected graph, and they have the same set of immoralities.*

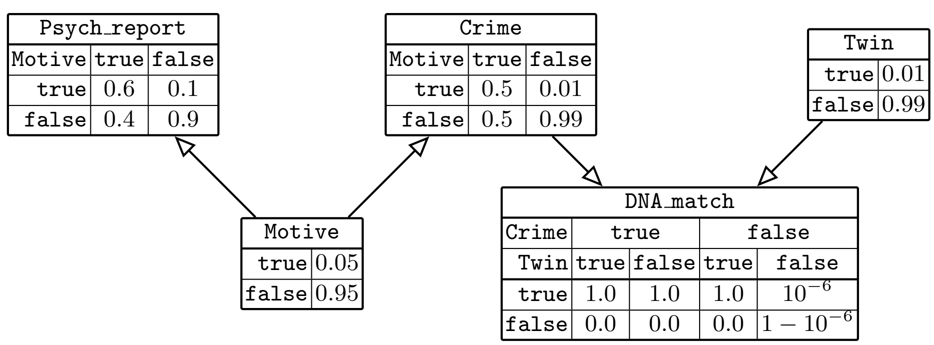

Fig. 2. A small BN concerning a criminal case. The conditional probability distributions are shown as tables inside the nodes of the graph.

3 Support graphs

We will split the construction of arguments for explaining a BN in two steps. We first construct a support graph from a BN, and subsequently establish arguments from the support graph. In this section we define the support graph and its construction.

Given a BN and a variable of interest $V^\star$, the support graph is a template for generating explanatory arguments. As such, it does not depend on observations of variables but rather models the possible structure of arguments for a particular variable of interest. This means that it can be used to construct an argument for any variable of our choice given any set of evidence, as we will show in the next section. When new evidence becomes available the same support graph can be reused. This means that the support graph should be able to capture the dynamics in d-separation caused by different observations. To enable this, each node in the support graph (which we will refer to as *support nodes* from here on) will be labelled with a *forbidden set* of variables $\mathcal{F}$. Moreover, since one BN variable can be represented more than once in a support graph, a function $\mathcal{V}$ is used to assign a variable to every support node. The support graph can now be constructed recursively. Initially a single support node $N^\star$ is created for which $\mathcal{V}(N^\star) = V^\star$ and $\mathcal{F}(N^\star) = \{V^\star\}$.

Definition 5 (Support graph). *Given a BN with graph* $G = (\mathbf{V}, \mathbf{E})$ *and a variable of interest* $V^\star$*, a* support graph *is a tuple* $\langle \mathcal{G}, \mathcal{V}, \mathcal{F} \rangle$ *where* $\mathcal{G}$ *is a directed graph* $(\mathbf{N}, \mathbf{L})$*, consisting of nodes* $\mathbf{N}$ *and edges* $\mathbf{L}$*,* $\mathcal{V} : \mathbf{N} \mapsto \mathbf{V}$ *assigns variables to nodes, and* $\mathcal{F} : \mathbf{N} \mapsto \mathcal{P}(\mathbf{V})$ *assigns sets of variables to each node, such that* G *is the smallest graph containing the node* $N^\star$ *(for which* $\mathcal{V}(N^\star) = V^\star$ *and* $\mathcal{F}(N^\star) = \{V^\star\}$*) closed under the following expansion operation:*

> *Whenever possible, a* supporter N_j *with variable* $\mathcal{V}(N_j) = V_j$ *is added as a parent to a node* N_i *(with* $V_i = \mathcal{V}(N_i)$*) iff* $V_j \in \mathrm{MB}(V_i) \setminus \mathcal{F}(N_i)$*. The forbidden set* $\mathcal{F}(N_j)$ *of the new support node is*
>
> – $\mathcal{F}(N_i) \cup \{V_j\}$ *if* V_j *is a parent of* V_i
> – $\mathcal{F}(N_i) \cup \{V_j\} \cup \{V_k \in \mathrm{Par}(V_j) | \langle V_i, V_j, V_k \rangle \text{is an immorality}\}$ *if* V_j *is a child of* V_i
> – $\mathcal{F}(N_j) \cup \{V_j\} \cup (\mathrm{Cld}(V_i) \cap \mathrm{Cld}(V_j))$ *otherwise*
>
> *If a support node with this forbidden set and the same* $\mathcal{V}(N_j)$ *already exists, that node is added as the parent of* N_i*, otherwise a supporting node* N_j *is created.*

To be able to represent d-separation correctly the *forbidden set* of variables assigns to every support node a set of variables that cannot be used in further support for that node. This forbidden set is inherited by supporters such that ancestors in the support graph cannot use variables from $\mathcal{F}$ either. Figure 3 shows the three cases of the forbidden set definition. The forbidden set of a new supporter N_i for variable V_i always includes the variable V_i itself which prevents circular reasoning. In a BN, parents of a common child often exhibit intercausal-interactions (such as explaining away) which means that the effect of one parent on the other is not the same as the combined effect from the parent to the child and then to the other parent. To support a variable V_i with one of its children and then support this child by a parent would incorrectly chain the inferences through a head-to-head node even though an intercausal-interaction is possible. Therefore we forbid the latter step by including any other parents that constitute immoralities in the second case. A reasoning step that uses the inference according to the intercausal-interaction is allowed by the third case.

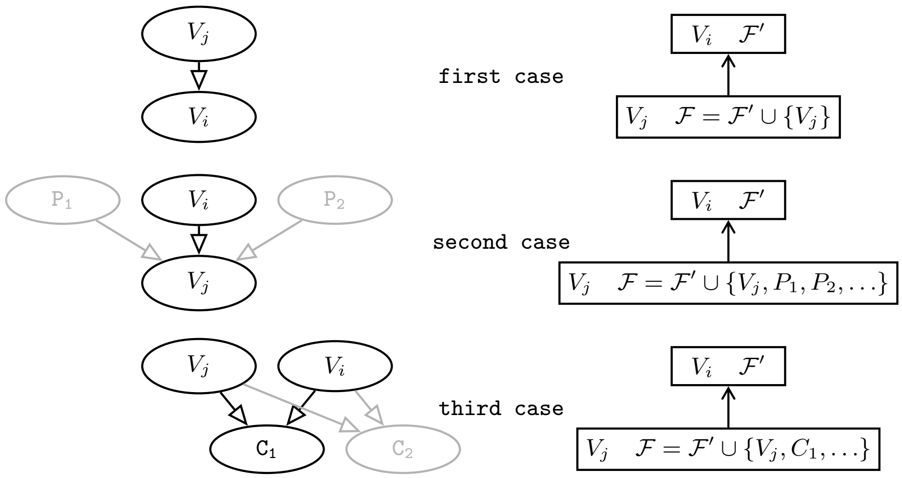

Fig. 3. Visual representation of the three cases in Definition 5. A support node for variable V_i can obtain support in three different ways from a variable V_j, depending on its graphical relation to V_i.

Now let us consider the example BN from Figure 2 and take `Crime` as the variable of interest. The initial support graph contains just one node with this variable and the forbidden set {`Crime`}. As can be seen in Figure 4, all of the three cases for $\mathcal{F}$ apply exactly once in this example. The `Crime` node can be supported by one parent (`Motive`), one child (`DNA_match`) and one parent of a child (`Twin`). In the first case the forbidden set leaves room to support the `Motive` node even further by adding a node for the `Psych_report` variable. This graph represents all possible dependencies in the BN model, where the actual dependencies will depend on the instantiation of evidence.

Property 1. Given a BN with $G = (\mathbf{V}, \mathbf{E})$, the constructed support graph contains $\mathcal{O}(|\mathbf{V}| * 2^{|\mathbf{V}|})$ nodes.

Proof (sketch). Variables can occur multiple times in the support graph but never with the same $\mathcal{F}$ sets (see the definition). This set contains subsets of other variables and therefore $2^{|\mathbf{V}|}$ is a strict upper bound on the number of times any variable can occur in the support graph. The total number of support nodes is therefore limited to $|\mathbf{V}| * 2^{|\mathbf{V}|}$. □

Property 2. In a given BN with a singly connected graph $G = (\mathbf{V}, \mathbf{E})$, every variable occurs exactly once in the support graph and the size of the support graph is $|\mathbf{V}|$.

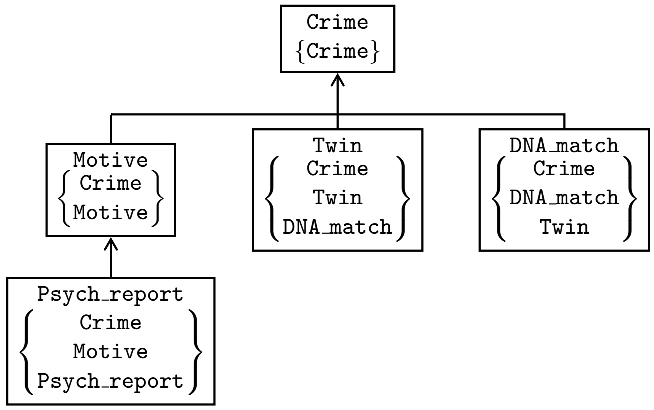

Fig. 4. The support graph corresponding to the example in Figure 2. For every node N_i we have shown the variable name $\mathcal{V}(N_i)$ togehter with the forbidden set $\mathcal{F}(N_i)$.

Proof (sketch). A variable can in theory occur multiple times in the support graph, but this only happens when the graph is loopy (multiply connected). □

Theorem 1. *Given two Markov equivalent BN graphs G and G', and a variable of interest $V^\star$, the two resulting support graphs are identical.*

Proof (sketch). Consider the BN graph G and the corresponding support graph. In a Markov equivalent graph G' an arbitrary number of edges may be reversed but not if this would create or remove immoralities. Following the three possible support steps we see that every supporter follows an edge from the skeleton (which stays the same) or an immorality (which also stays the same). What remains to be shown is that the forbidden sets will also be equal. Let us consider the three cases of the $\mathcal{F}$ update from Definition 5 (see also Figure 3). Suppose that in the support graph of G, N_i for variable V_i is supporting N_j for variable V_j:

- In the first case, reversal of the edge between V_i and V_j would change this to the second case in which variables V_k with an immorality $\langle V_i, V_j, V_k\rangle$ would be added to $\mathcal{F}$. However, since no immoralities are created those variables either do not exist, or the reversal is not allowed by the Markov equivalence.
- In the second case, reversal of any of the incoming edges of V_j is not allowed if V_j is involved in an immorality $\langle V_i, V_j, \_\rangle$. If that is the case, reversal is allowed and we end up in the first case but the forbidden set will be exactly the same.
- In the third case, there is no immorality between V_i and V_j through any of the shared children because if there were, a direct edge exists and either of the former cases would have taken precedence. None of these edges may therefore be reversed in G'.

□

What this theorem shows is that Markov equivalent models are mapped to the same support graph, which means that they will receive the same argumentative explanation. This takes one of the confusing aspects of BNs away, which is that the directions of edges do not have a clear intuitive interpretation.

4 Argument construction

In previous work we have already shown a method to identify arguments in a BN setting and how they can be enumerated exhaustively [10]. A disadvantage of the exhaustive enumeration of probabilistic rules and rule combinations is the combinatorial explosion of possibilities, even for realistically sized models. Using a support graph can reduce the number of arguments that need to be enumerated because only rules relevant to the conclusion of the argument are considered.

Definition 6 (Bayesian argument). *An argument A on the basis of a BN, a set of observations* $\mathbf{O}$*, and the corresponding support graph* $\langle \mathcal{G} = (\mathbf{N}, \mathbf{L}), \mathcal{V}, \mathcal{F} \rangle$*, is one of the following:*

- $\langle N, o \rangle$ *such that* $(\mathcal{V}(N) = o) \in \mathbf{O}$*, for which* $\mathrm{Obs}(A) = \{N = o\}$ *or*
- $\langle N_1, o_1 \rangle, \ldots, \langle N_n, o_n \rangle \Rightarrow \langle N, o \rangle$ *such that* $N_1, \ldots, N_n$ *are parents of* N *in the support graph,* $\langle N_1, o_1 \rangle$ *through* $\langle N_n, o_n \rangle$ *are arguments, and* o *is the most probable outcome of* $\mathcal{V}(N)$ *given the observations* $\mathrm{Obs}(A)$*, in which* $\mathrm{Obs}(A)$ *is the union of* $\mathrm{Obs}(B)$ *over subarguments* B*.*

In this definition $\langle N_1, o_1 \rangle$ *through* $\langle N_n, o_n \rangle$ *are the* immediate subarguments *of* $\langle N_1, o_1 \rangle, \ldots, \langle N_n, o_n \rangle \Rightarrow \langle N, o \rangle$*.*

Argument attack arises when two arguments assign outcomes to the same variable. We might be tempted to prefer the argument with the highest probability but that could lead to mistakes. For instance, when A, B and C collectively support a conclusion, situations can exist where the highest probability of that conclusion occurs when B is left out. It is, however, usually not acceptable to ignore evidence. The following definition meets this criterion:

Definition 7 (superseding). *An argument A supersedes another argument B iff* $\mathrm{Obs}(A) \supseteq \mathrm{Obs}(B)$*.*

Indeed, we prefer one argument over another iff it includes a superset of evidence. This resembles Pollock's concept of *subproperty defeat of the statistical syllogism* [7]. Superseding can be seen as a special case of undercutting, so attack and defeat follow naturally:

Definition 8 (Undercutting attack and defeat). *An argument A undercuts another argument B iff it supersedes B or one of the sub-arguments of B. An undercutting attack always succeeds and therefore A also defeats B.*

It can be shown that this instantiates a special case of the ASPIC+ [5] model of argumentation but a proof of that is omitted for brevity. In this special cases rebuttal and undermining are redundant due to the fact that for every rebuttal there is also an undercutter resolving the issue.

An interesting property of this approach is that conflicts between observations are resolved in the probabilistic setting within the argument and that the resolution is mirrored by the defeat relation of the extracted arguments, rather than decided by it. This means that the resulting argumentation system is rather simple which is ideal for a BN explanation method.

If we apply this system to the support graph from our example BN with the observations that `Psych_report=true` and `DNA_match=true`, we obtain (among others) the arguments shown in Figure 5. The argument on the right is in fact the formal version of the argument that we already showed in Figure 1. The undercutter from that figure was not extracted because no evidence for a twin was present in the set of observations.

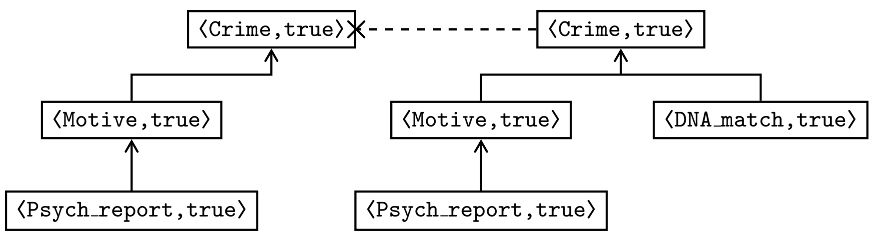

Fig. 5. Arguments resulting from our running example. The argument on the left is superseded by the one on the right. For readability we have only shown conclusions inside the nodes.

Property 3. Given a BN, a variable of interest, the resulting support graph and a set of observations, for every node in the support graph either no argument for this node exists at all, or exactly one of the arguments that exists supersedes all other arguments for the same node without itself being superseded.

Proof (sketch). Suppose no such un-superseded argument exists, then there must be two arguments A and B that supersede each other, i.e. $\mathrm{Obs}(A) \setminus \mathrm{Obs}(B) \neq \emptyset$ and $\mathrm{Obs}(B) \setminus \mathrm{Obs}(A) \neq \emptyset$. However, in that case an argument C combining the immediate subarguments of A and B also exists that strictly supersedes both A and B. □

Informally, the argument that includes all possible supporters that have ancestors in $\mathbf{O}$ will supersede any argument that includes fewer supporters. Since this holds for every node, there is in this argumentation system one unique tree in which every argument is supported by the maximal number of immediate sub-arguments given what is derivable from the evidence. Together with the fact that the outcome of the argument is based on the probability given the used observations, and that no d-separated paths are used in the argument this exactly mirrors the probabilistic reasoning.

5 Discussion

In this paper we formalised a two-phase argument extraction method. We have shown how support graphs help in the construction of arguments because they capture the argumentative structure that is present in a BN.

Many explanation methods for BNs (see e.g. [4,3]) focus on textual or visual systems. Other work on argument extraction includes that of Keppens [2], who focuses on Argument Diagrams. One advantage of structured argumentation is that counter-arguments can easily be modelled as well. Future research includes how arguments constructed from a BN can be combined with arguments from other sources, since often the available evidence is only partially probabilistic.

References

1. P. M. Dung. On the acceptability of arguments and its fundamental role in non-monotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, 77:321–357, 2005.
2. J. Keppens. Argument diagram extraction from evidential Bayesian Networks. *Artificial Intelligence & Law*, 20(2):109–143, 2012.
3. J. R. Koiter. Visualizing inference in Bayesian Networks. Master's thesis, Delft University of Technology, 2006.
4. C. Lacave and F. J. Díez. A review of explanation methods for Bayesian Networks. *Knowledge Engineering Review*, 17(2):107–127, 2002.
5. S. Modgil and H. Prakken. A general account of argumentation with preferences. *Artificial Intelligence*, 195:361–397, 2013.
6. J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, San Francisco, 1988.
7. J. L. Pollock. Justification and defeat. *Artificial Intelligence*, 67, 1994.
8. G. R. Simari and R. P. Loui. A mathematical treatment of defeasible reasoning and its implementation. *Artificial intelligence*, 53(2):125–157, 1992.
9. F. Taroni, C. Aitken, P. Garbolino, and A. Biedermann. *Bayesian Networks and Probabilistic Inference in Forensic Science*. John Wiley & Sons, Ltd, 2006.
10. S. T. Timmer, J.-J. C. Meyer, H. Prakken, S. Renooij, and B. Verheij. Extracting legal arguments from forensic Bayesian networks. In R. Hoekstra, editor, *Legal Knowledge and Information Systems. JURIX 2014: The Twenty-seventh Annual Conference*, volume 217, pages 71–80, 2014.
11. S. T. Timmer, J.-J. C. Meyer, H. Prakken, S. Renooij, and B. Verheij. A structure-guided approach to capturing Bayesian reasoning about legal evidence in argumentation. Technical report, Utrecht University, 2015. UU-CS-2015-003. Also submitted for publication.
12. F. H. van Eemeren, B. Garssen, E. C. W. Krabbe, A. F. S. Henkemans, B. Verheij, and J. H. M. Wagemans. *Handbook of Argumentation Theory*. Springer, Dordrecht, 2014.
13. T. Verma and J. Pearl. Equivalence and synthesis of causal models. In *Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence*, UAI '90, pages 255–270, New York, NY, USA, 1991. Elsevier Science Inc.
14. G. A. W. Vreeswijk. Abstract argumentation systems. *Artificial intelligence*, 90(1):225–279, 1997.