Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence (2008)

A General Framework for Generating Multivariate Explanations in Bayesian Networks

Changhe Yuan
Department of Computer Science and Engineering
Mississippi State University
Mississippi State, MS 39762
cyuan@cse.msstate.edu

Tsai-Ching Lu
HRL Laboratories, LLC
Malibu, CA, 90265
tlu@hrl.com

Abstract

Many existing explanation methods in Bayesian networks, such as Maximum a Posteriori (MAP) assignment and Most Probable Explanation (MPE), generate complete assignments for target variables. A priori, the set of target variables is often large, but only a few of them may be most relevant in explaining given evidence. Generating explanations with all the target variables is hence not always desirable. This paper addresses the problem by proposing a new framework called *Most Relevant Explanation* (MRE), which aims to automatically identify the most relevant target variables. We will also discuss in detail a specific instance of the framework that uses *generalized Bayes factor* as the relevance measure. Finally we will propose an approximate algorithm based on Reversible Jump MCMC and simulated annealing to solve MRE. Empirical results show that the new approach typically finds much more concise explanations than existing methods.

Introduction

Bayesian networks offer compact and intuitive graphical representations of uncertain relations among random variables in a domain and provide a foundation for many diagnostic expert systems. However, these systems typically focus on disambiguating single-fault diagnostic hypotheses because it is hard to generate "just right" multiple-fault hypotheses that only contain the most relevant faults. Explanation methods in Bayesian networks, such as Maximum a Posteriori (MAP) assignment and Most Probable Explanation (MPE), can be applied in multiple-fault diagnosis. Both MAP and MPE find complete assignments to a set of target variables as the best explanation for given evidence. A priori, the set of target variables is often large; a real-world diagnostic model may have tens or even hundreds of them. Even the optimal solution by MAP or MPE may have an extremely low probability, say in the order of 10^{-6}, given that so many variables are involved. It is hard to make any decisions based on such explanations. Furthermore, it is observed that usually only a few of target variables may be most relevant in explaining specific observations. For example, there are many possible diseases in any medical domain, but a patient can have at most a few diseases at one time, as long as he or she does not delay treatments for too long. Other diseases should not be included in following diagnosis or treatment. Therefore, it is desirable to find diagnostic hypotheses only for those relevant diseases. The question is how to identify them.

This work addresses the problem by proposing a new framework called *Most Relevant Explanation* (MRE), which aims to automatically identify the most relevant target variables by searching for a partial assignment of the target variables that maximizes a chosen relevant measure. We will also discuss a specific instance of the framework that uses *generalized Bayes factor* as the relevance measure, based on recent research in Bayesian confirmation theory. Finally, we will propose an approximate algorithm based on Reversible Jump MCMC and simulated annealing to solve MRE. Our results show that the new approach can typically find much more concise explanations than existing methods.

A Running Example

The proposed work is best motivated using a simple example. Consider the circuit in Figure 1(a) adapted from (Poole & Provan 1991). Gates A, B, C and D are faulty if they are closed. The prior probabilities that the gates close independently are $0.016, 0.1, 0.15$ and 0.1 respectively. The circuit can be modeled with a diagnostic Bayesian network as in Figure 1(b). Nodes A, B, C and D correspond to the gates in the circuit and each has two states: "defective" and "ok". Other nodes are input or output nodes and take states "current" or "noCurr". Uncertainty is introduced to the model such that an output node is in state "current" with certain probability if its parent gate, when exists, is "defective" and any of its other parents is in state "current". Otherwise, it is in "noCurr" state with probability 1.0. For example, node $output\ of\ B$ takes state "current" with probability 0.99 if parent gate B is in state "defective" and parent $Input$ is in state "current".

Suppose we observe that current flows through the circuit, which means that nodes $Input$ and $Total\ Output$ in the Bayesian network are both in the state "current". The task is to diagnose the system and find the best fault hypothesis. Based on our knowledge of the model, we can easily identify the following three *minimal* scenarios that most likely lead to the observation:

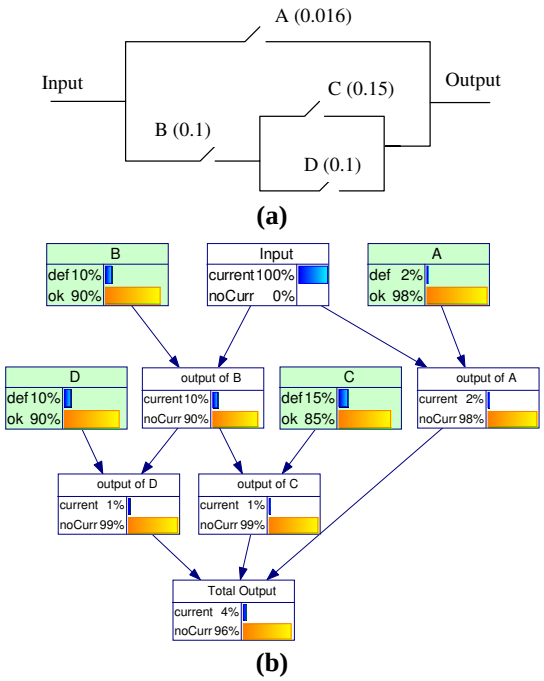

Figure 1: **(a)** A digital circuit, **(b)** A diagnostic Bayesian network

- A is defective;
- B and C are defective;
- B and D are defective.

These scenarios are minimal in the sense that any scenario with fewer faults than any of them will not produce the observed symptom.

Related Work

Many existing diagnostic approaches based on Bayesian networks make simplifying assumptions for computational convenience. For example, it is often assumed that the target variables are mutually exclusive and collectively exhaustive, and there is conditional independence of evidence given any hypothesis (Jensen & Liang 1994; Kalagnanam & Henrion 1988). Therefore, we only need to consider the singleton hypotheses. However, single-fault hypotheses may be *underspecified* and are unable to fully explain given evidence. For the circuit example, the posterior probabilities of A, B, C, and D failing independently are $0.391, 0.649, 0.446$, and 0.301 respectively. Therefore, $\neg B$ is the best single-fault hypothesis (We use a variable and its negation to stand for its "ok" and "defective" states respectively). However, B alone does not fully explain the evidence. C or D has to be involved.

For a domain in which faults are interdependent, multiple-fault hypotheses are often more natural for explaining given observations. However, existing methods often produce hypotheses that are *overspecified*. MAP finds a configuration of the target variables that maximizes the joint posterior probability given partial or complete evidence on the other variables. For the circuit example, MAP will find the multiple-fault hypothesis $(A \wedge \neg B \wedge \neg C \wedge D)$ as the best explanation. However, when B and C are faulty, A and D are not really necessary for explaining the observation. MPE finds hypotheses with even more variables. Several other approaches make use of the conditional independence relations encoded in Bayesian networks to identify the best multiple-fault hypotheses (Shimony 1993). They will find the same explanation as MAP because all target variables are dependent given the evidence. Yet several other approaches measure the quality of explanation candidates using likelihood of evidence (Chajewska & Halpern 1997; Gärdenfors 1988). Unfortunately they will choose $(\neg A \wedge \neg B \wedge \neg C \wedge \neg D)$ as the best explanation, because the likelihood of the evidence given that all targets fail is almost 1.0.

There have been efforts trying to find or define more reasonable explanations. Henrion and Druzdzel (1991) assume that a system has a set of pre-specified scenarios as potential explanations and look for the scenario with the highest posterior probability. Flores et al. propose to build an explanation tree by incrementally adding the most informative variable remaining while maintaining the probability of each explanation above certain threshold (Flores, Gamez, & Moral 2005). Similar problems have also been studied in the area of model-based diagnosis (de Kleer, Mackworth, & Reiter 1992).

The above discussion shows that many existing methods generate multiple-fault hypotheses that are either too complex (overspecified) or too simple (underspecified). They fail to find explanations that focus on the most relevant target variables.

Most Relevant Explanation

This paper addresses the aforementioned problems by proposing a new framework for explanation in Bayesian networks that aims to automatically identify the most relevant faults. Our goal is to find a concise explanation that best accounts for given evidence. We formally define *explanation* in our setting as follows.

Definition 1 *Given a set of target variables* $\mathbf{X}$ *in a Bayesian network and evidence* $\mathbf{e}$ *on the remaining variables, an* explanation $\mathbf{x_{1:k}}$ *for the evidence is a partial instantiation of the target variables, i.e.,* $\mathbf{X_{1:k}} \subseteq \mathbf{X}$ *and* $\mathbf{X_{1:k}} \neq \emptyset$.

To be able to compare all possible explanations, we have to put them on an equal footing using an appropriate relevance measure. We will discuss the selection of the measure in the next section. For now, let us denote it as $\mathcal{F}(\mathbf{x_{1:k}}, \mathbf{e})$. We now introduce the following new formulation for explanation in Bayesian networks.

Definition 2 *Let* $\mathbf{X}$ *be a set of target variables, and* $\mathbf{e}$ *be the evidence on the remaining variables in a Bayesian network.* Most Relevant Explanation *is the problem of finding an explanation* $\mathbf{x_{1:k}}$ *that maximizes a relevance function* $\mathcal{F}(\mathbf{x_{1:k}}, \mathbf{e})$ *measuring the degree of the partial assignment* $\mathbf{x_{1:k}}$ *for explaining* $\mathbf{e}$*, i.e.,*

$$MRE(\mathbf{X}, \mathbf{e}) \triangleq \arg\max_{\mathbf{x_{1:k}}, \mathbf{X_{1:k}} \subseteq \mathbf{X}, \mathbf{X_{1:k}} \neq \emptyset} \mathcal{F}(\mathbf{x_{1:k}}, \mathbf{e}) \,. \quad (1)$$

Therefore, MRE traverses the trans-dimensional space containing all the partial assignments of $\mathbf{X}$ and finds the partial assignment that maximizes the relevance measure $\mathcal{F}$. The proposed framework provides a formulation fundamentally different from MAP or MPE. While MAP and MPE generate explanations for all given target variables, MRE finds an explanation with only a subset of the variables; The subset can be as large as all target variables, or as small as singletons. Therefore, MRE selects the best explanation from all solution candidates of existing approaches.

Selecting Relevance Measures

MRE is a general framework that can be instantiated with different relevant measures for different decision-making purposes. Selecting an appropriate relevance measure is crucial for MRE to pick out the most relevant target variables. *Posterior probability* is the first metric that one may think of, but it is not viable because probability biases strongly towards solutions with fewer variables; the final solution will inevitably be the singleton explanation with the highest probability.

Bayesian confirmation theory provides a basis for selecting the measure. The theory says that evidence $\mathbf{e}$ favors hypothesis $\mathbf{x_{1:k1}}$ over hypothesis $\mathbf{x_{1:k2}}$, according to Bayesian relevance measure r, if and only if $r(\mathbf{x_{1:k1}}, \mathbf{e}) > r(\mathbf{x_{1:k2}}, \mathbf{e})$. In other words, $\mathbf{e}$ provides more evidence for $\mathbf{x_{1:k1}}$ than for $\mathbf{x_{1:k2}}$. Note that $\mathbf{x_{1:k1}}$ may or may not have a higher probability than $\mathbf{x_{1:k2}}$. Although there is still debate over the choice of r, recent studies show that *generalized Bayes factor* (GBF), defined as $\Pr(\mathbf{e}|\mathbf{x_{1:k1}})/\Pr(\mathbf{e}|\overline{\mathbf{x_{1:k1}}})$ for hypothesis $\mathbf{x_{1:k1}}$ and evidence $\mathbf{e}$, satisfies many theoretical properties and is an excellent measure for representing the degree of evidential support (Fitelson 2001). More importantly, GBF clearly has the capability to rank explanations from different dimensional spaces.

Likelihood $P(\mathbf{e}|\mathbf{x_{1:k}})$ is another measure that can put all the partial assignments on a same ground. The following theorem shows that GBF provides more discrimination power than likelihood.

Theorem 1 *For an explanation $\mathbf{x_{1:k}}$ with a fixed likelihood $P(\mathbf{e}|\mathbf{x_{1:k}})$ greater than or equal $P(\mathbf{x_{1:k}})$, $GBF(\mathbf{x_{1:k}}, \mathbf{e})$ increases monotonically as the prior probability $P(\mathbf{x_{1:k}})$ increases.*

Therefore, GBF takes into account the relative magnitude of the probabilities such that the explanations that cannot be distinguished by likelihood can be ranked using GBF. Since lower-dimensional explanations typically have higher probabilities, GBF has the intrinsic capability to penalize more complex explanations. In fact, we have the following theorem, which further confirms the property.

Theorem 2 *For any explanation $\mathbf{x_{1:k}}$ with $P(\mathbf{e}|\mathbf{x_{1:k}}) \geq P(\mathbf{x_{1:k}})$, adding any state y of a variable Y that is independent from variables in $\mathbf{x_{1:k}}$ and $\mathbf{e}$ to the explanation decreases its GBF, i.e.,*

$$GBF(\mathbf{x_{1:k}}, \mathbf{e}) \geq GBF(\mathbf{x_{1:k}} \cup \{y\}, \mathbf{e}). \quad (2)$$

The above theorem can be further relaxed to accommodate cases where $\mathbf{e}$ decreases the probability of y.

Running Example Revisited

For the circuit example, the top hypothesis according to GBF is:

$$GBF(\neg B, \neg C) = 42.62\ . \quad (3)$$

$(\neg B, \neg C)$ is a better explanation than both $(\neg A)$ (39.44) and $(\neg B, \neg D)$ (35.88) because both its prior and posterior probabilities are relatively high in comparison; Their posterior probability are 0.394, 0.391, and 0.266 respectively. More importantly, MRE simultaneously identifies the set of target variables and their states that best explain the evidence. Results also suggest that GBF automatically penalizes higher-dimensional explanations and gets rid of variables are less relevant or irrelevant. For example,

$$GBF(\neg B, \neg C) > GBF(\neg B, \neg C, A)\ \&\ GBF(\neg B, \neg C, D) > GBF(\neg B, \neg C, A, D)\ .$$

Reversible Jump MCMC for Solving MRE

We want to develop algorithms for solving MRE problems. Theoretical results have shown that the decision problem of MAP is NP^{PP}-complete (Park 2002). Note that MRE has a much larger search space than MAP and is believed to be even harder. Therefore, we focus on developing approximate methods for MRE in this paper. In particular, due to the need to find solution in a trans-dimensional space, we propose an MRE algorithm based on Reversible Jump MCMC algorithm (Green 1995) and simulated annealing.

Reversible Jump MCMC

The reversible jump MCMC algorithm (Green 1995) allows sampling the space $\chi = \cup_{k \in \mathcal{K}}(k \times \mathcal{R}^{n_k})$ with invariant distribution π. To implement RJMCMC, we need reversible moves to traverse the whole space χ. This is accomplished by a deterministic, differentiable, invertible dimension matching function $f_{n \to m}$ that transforms the parameters

$$f_{n \to m}(\theta_n, u_n) = (\theta_m, u_m), \quad (4)$$

where u_n and u_m are random quantities used to ensure dimension matching between the communication spaces, i.e.,

$$d(\theta_n) + d(u_n) = d(\theta_m) + d(u_m). \quad (5)$$

The inverse transform $f_{m \to n} = f^{-1}_{n \to m}$ gives the move to the other direction. If $q_{n \to m}(\theta_n, u_n)$ is the probability density for the proposed move and $q(n, m)$ is the probability for the move $n \to m$, the acceptance probability can be written as

$$\alpha_{n \to m} = min\{1, \frac{\pi(\theta_m, u_m) q(m, n) q_{m \to n}(\theta_m, u_m)}{\pi(\theta_n, u_n) q(n, m) q_{n \to m}(\theta_n, u_n)} \mathcal{J}_{f_{n \to m}}\}, \quad (6)$$

where $\mathcal{J}$ is the Jacobian of the transformation from (θ_n, u_n) to (θ_m, u_m), i.e.,

$$\mathcal{J}_{f_{n \to m}} = |\frac{\partial(\theta_m, u_m)}{\partial(\theta_n, u_n)}|. \quad (7)$$

We propose to integrate reversible jump MCMC with simulated annealing to solve MRE. Simulated annealing is a

technique (Kirkpatrick, Gelatt, & Vecchi 1983) that finds the global maximum of a distribution $p(x)$ by simulating a non-homogeneous Markov chain whose invariant function is no longer $p(x)$, but

$$p_i(x) \propto p^{1/T_i}(x) \ , \tag{8}$$

where $\lim_{i\to\infty} T_i = 0$. Under weak regularity assumption, $p^{\infty}(x)$ is a probability density that concentrates itself on the modes of $p(x)$. Andrieu et al. has applied a similar scheme to optimizing radial basis function (RBF) networks (Andrieu, de Freitas, & Doucet 2000).

Trans-dimensional Model

We use the simple Bayesian graphical model in Figure 2 to represent the MRE problem. Node K stands for the dimensionality of the explanation. It takes value 1 through k_{max}, where k_{max} is the total number of target variables. Node X represents the current explanation and contains a set of states of the selected target variables. The shaded node e stands for the observed evidence. Therefore, the full trans-dimensional model can be expressed as:

$$P(K, X, e) = P(K)P(X|K)P(e|X). \tag{9}$$

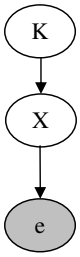

Figure 2: The trans-dimensional model for MRE problem.

This Bayesian model allows unknowns K and X be regarded as drawn from a hierarchical prior parameterized as follows.

Prior $P(K)$ can be parameterized to penalize more complex explanations as in model selection. However, as shown in previous discussion, generalized Bayes factor seems to have the intrinsic capability to penalize more complex explanations. So we use the following prior $P(K) \propto \sum_{c_k \in C(k_{max},k)} \prod_{X_i \in c_k} |X_i|$, where $C(.,.)$ are all the possible ways of selecting k variables. In other words, $P(K)$ is proportional to the number of explanations in each subspace with K target variables.

Next we consider the parametrization of $P(X|K)$. Intuitively, this conditional relation models the variable and state selection process. A unique aspect of MRE is that there is a fixed set of labeled variables $X_1, X_2, ..., X_{k_{max}}$ to choose from. Also, there are different number of joint states for any given k variables. Therefore, we set $P(X|K)$ to be uniform over all possible explanations for each K. The above parameterizations of $P(K)$ and $P(X|K)$ ensure that a uniform prior is specified over all partial assignments of the target variables.

Finally, $P(e|X)$ is not the conditional probability in the Bayesian network but a likelihood measure in a "calibrated" posterior distribution and represents the evidence that the candidate explanation conforms to our objective to maximize GBF. Hence, we define the likelihood measure as

$$P(e|X) \propto GBF(X, e). \tag{10}$$

The trans-dimensional model thus explicitly defines a joint dimensional space of K and X that MRE needs to explore.

Reversible Jump Moves for MRE

Since MRE needs to move across subspaces of different dimensions, we select the following types of moves:

- **Birth move:** Add a variable randomly selected from the available possibilities to the current explanation;
- **Death move:** Delete a randomly selected variable in the explanation;
- **Update move:** Update the states of existing variables in the explanation.

At each iteration, one of the candidate moves, birth, death, and update, is randomly chosen. Let the probabilities for choosing these moves be b_k, d_k, u_k respectively, such that $b_k + d_k + u_k = 1$. Typically, we let $d_k = 1/3, b_k = 1/3$, and $u_k = 1/3$. However, care has to be taken to ensure that k lies between 1 and k_{max}. When $k = 1$, death move is prohibited. Similarly when $k = k_{max}$, birth move is prohibited. We discuss the details of each move in the following.

First, the simplest move is the update move. Since this move does not change the dimension of the model, it reduces to the standard Gibbs sampling transition kernel. Given K, we can easily calculate the full conditionals of each individual variable X_i and update it using Gibbs sampling according to (Yuan, Lu, & Druzdzel 2004).

Second is the birth move. Since the $f_{n\to m}$ function represents an identity mapping, $\mathcal{J}$ is equal to 1. Assume the current model has k variables, the more concrete form of the acceptance probability of birth move used in this paper is

$$\alpha_{birth} = min\{1, \frac{P(k+1, x_{1:k+1}, e) d_{k+1} \frac{1}{k+1}}{P(k, x_{1:k}, e) b_k \frac{1}{k_{max}-k} \frac{1}{|X_{k+1}|}}\}. \tag{11}$$

The proposal distributions are defined as follows. $\frac{1}{k_{max}-k}$ is the probability for selecting a variable from possible candidates. Suppose X_{k+1} is the variable selected, $\frac{1}{|X_{k+1}|}$ is the probability for selecting an arbitrary state uniformly. $\frac{1}{k+1}$ is the probability of proposing the opposite death move. Clearly, the reversibility of the moves is achieved. Furthermore, the dimensions are carefully maintained. There are k variables in the original model, plus two random quantities for proposing the birth move. In the proposed model, there are $k+1$ variables, and one random quantity is needed to propose the opposite death move. The detailed steps of birth move is as follows.

Birth Move:

1. Randomly choose a variable not in the current explanation;

2. Randomly choose a state for the variable;
3. Accept the move with probability defined in Eqn. 11;
4. If the move is accepted, update the state to be $(k+1, x_{1:k+1})$;
5. Otherwise, stay in the current explanation.

The death move is defined analogously. For the same reason, the Jacobian $\mathcal{J}$ is also equal to 1 for the death move. The acceptance probability of death move is

$$\alpha_{death} = min\{1, \frac{P(k-1, x_{1:k-1}, e) b_{k-1} \frac{1}{k_{max}-k+1} \frac{1}{|X_k|}}{P(k, x_{1:k}, e) d_k \frac{1}{k}}\}. \quad (12)$$

Similarly to the birth move, $\frac{1}{k}$ is the probability for selecting a variable to kill. Suppose X_k is the variable selected, $\frac{1}{k_{max}-k+1}\frac{1}{|X_k|}$ defines the probability of proposing the opposite birth move. $\frac{1}{k_{max}-k+1}$ is the probability of proposing the variable, and $\frac{1}{|X_k|}$ is the probability for selecting the state. Reversibility and dimensions are also maintained. The detailed steps of death move is as follows.

Death Move:

1. Randomly choose a variable already in the current explanation to kill;
2. Accept the move with probability defined in Eqn. 12;
3. If the move is accepted, update the state to be $(k-1, x_{1:k-1})$;
4. Otherwise, stay in the current explanation.

It is possible to design more informative moves to bias towards better candidate solutions. We choose these move designs for their simplicity.

Dealing with Extreme Values

Since we define $P(e|X)$ in the trans-dimensional model in proportional to $GBF(x, e)$, we may encounter numerical difficulty when $P(X)$ and $P(X|e)$ take extreme values 0 and 1. First, if $P(X) = 0$, $P(X|e)$ will be 0 also. Such explanations are obviously not MRE solutions, and we can avoid proposing such explanations in the algorithm. For the same reason, we should also avoid explanations with $P(X) \neq 0$ and $P(X|e) = 0$. The most tricky situation is when $P(X|e) = 1$, for which GBF is undefined. The fact that an explanation has posterior probability 1, we believe, warrants it being an MRE solution. We also note that the following equivalence holds:

$$P(x_1, x_2, ..., x_n) = 1 \Leftrightarrow P(x_i) = 1, \forall i. \quad (13)$$

Therefore, we can use a preprocessing step to find all the target variables and states with marginal posterior probability equal to 1 and take their conjunction as the MRE solution. Unless we are interested in finding multiple answers, we can stop the algorithm at this point and output the solution.

Empirical Evaluation

We tested the RJMCMC algorithm on a set of benchmark models listed in Table 1, including Alarm, Circuit, Hepar, Munin, and Win95pts. We chose these several models because we have the diagnostic versions of these networks, where each node has been classified into three categories: *target*, *observation*, and *auxiliary*. For each test case we randomly selected states for all observation nodes as evidence and took all the target nodes as target variables. In our experiments, we used simple geometric annealing schedule for simulated annealing. Computing GBF is tractable for the models we tested. Otherwise, we may have to resort to approximate inference methods.

Case Study: Alarm

We randomly generated a test case on Alarm network and reported the convergence of both K, dimension of explanation, and GBF, generalized Bayes factor, in Figure 3.

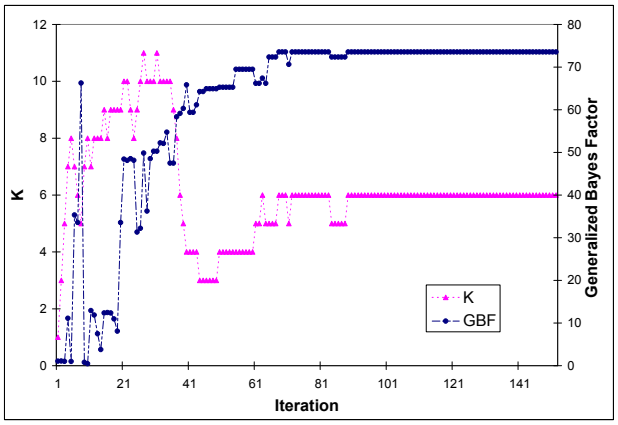

Figure 3: Convergence of K and GBF on a random test case on Alarm network.

We note that the generalized Bayes factor increased stochastically with the number of iterations of simulated annealing and eventually converged to a maximum, which is confirmed to be the global maximum by the exhaustive search. We also observe that the RJMCMC algorithm typically does not need many iterations to converge to a maximum in our experiments.

Results on Benchmark Models

We randomly generated 50 cases for the set of benchmark models and reported the average number of variables in the final explanations in Table 1. For comparison purpose, we also reported the results by using likelihood as the relevance measure. The results of the exhaustive search for the models with fewer than 12 target nodes are also included. Model Win95pts has 35 target variables and is too difficult for the exhaustive search to solve.

The results clearly show that generalized Bayes factor typically returns explanations that are much more concise

Net	N	A-GBF	A-L	S-GBF	S-L
Alarm	12	5.3/42	11.82/43	4.4	10.6
Circuit	4	1.14/50	3.24/50	1.14	3.24
Hepar	9	2.64/44	8.18/46	2.56	8.04
Munin	4	2/50	3.2/50	2	3.2
Win95pts	35	12.46	18.04	-	-

Table 1: The average number of variables in explanations generated by A-GBF (RJMCMC with GBF), A-L (RJMCMC with likelihood), S-GBF (exhaustive search with GBF), and S-L (exhaustive search with likelihood). The number of cases solved optimally out of 50 cases by RJMCMC are also reported when exact answers were available.

than likelihood. Also, the RJMCMC algorithm finds optimal solutions for most of the cases in a reasonable number of iterations.

Concluding Remarks

In this paper, we proposed a new framework called Most Relevant Explanation (MRE) for finding explanations for given evidence in Bayesian networks. A specific instance of the framework using generalized Bayes factor (GBF) as the relevance measure was discussed in detail. We also developed an approximate algorithm based on Reversible Jump MCMC and simulated annealing that can solve MRE efficiently and accurately on a set of benchmark networks. Our results show that the new approach can typically generate much more concise explanations than existing approaches. The results also show that GBF can factor in both prior and posterior probabilities in ranking the explanations and has the intrinsic capability to penalize more complex explanations.

Evaluating the proposed framework on real-world multiple-fault test cases remains an important future task. Currently such test cases are extremely rare. Furthermore, the proposed RJMCMC algorithm can be generalized to deal with Bayesian networks with continuous variables. We also plan to study the theoretical complexity of MRE and investigate other relevance measures for different decision making settings. The proposed methodologies are applicable to a broad range of real-world problems, including but not restricted to diagnosis.

Acknowledgements

We thank Brian Milch and the anonymous reviewers for their insightful comments that have led to improvements in the paper. All experimental data have been obtained using SMILE, a Bayesian inference engine developed at the Decision Systems Laboratory and available at `http://genie.sis.pitt.edu`.

References

Andrieu, C.; de Freitas, N.; and Doucet, A. 2000. Reversible jump MCMC simulated annealing for neural networks. In *Proceedings of the 16th Annual Conference on Uncertainty in Artificial Intelligence (UAI-00)*, 11–18. San Francisco, CA: Morgan Kaufmann.

Chajewska, U., and Halpern, J. Y. 1997. Defining explanation in probabilistic systems. In *Proceedings of the Thirteenth Annual Conference on Uncertainty in Artificial Intelligence (UAI–97)*, 62–71. San Francisco, CA: Morgan Kaufmann Publishers.

de Kleer, J.; Mackworth, A.; and Reiter, R. 1992. Characterizing diagnosis and systems. *Artificial Intelligence* 56:197–222.

Fitelson, B. 2001. *Studies in Bayesian Confirmation Theory*. Ph.D. Dissertation, University of Wisconsin, Madison, Philosophy Department.

Flores, J.; Gamez, J. A.; and Moral, S. 2005. Abductive inference in bayesian networks: finding a partition of the explanation space. In *Eighth European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty, ECSQARU'05*, 63–75. Springer Verlag.

Gärdenfors, P. 1988. *Knowledge in Flux: Modeling the Dynamics of Epistemic States*. MIT Press.

Green, P. 1995. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrica* 82:711–732.

Henrion, M., and Druzdzel, M. J. 1991. Qualitative propagation and scenario-based schemes for explaining probabilistic reasoning. In Bonissone, P.; Henrion, M.; Kanal, L.; and Lemmer, J., eds., *Uncertainty in Artificial Intelligence 6*. New York, N. Y.: Elsevier Science Publishing Company, Inc. 17–32.

Jensen, F. V., and Liang, J. 1994. drHugin: A system for value of information in Bayesian networks. In *Proceedings of the 1994 Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, 178–183.

Kalagnanam, J., and Henrion, M. 1988. A comparison of decision analysis and expert rules for sequential diagnosis. In *Proceedings of the 4th Annual Conference on Uncertainty in Artificial Intelligence (UAI-88)*, 253–270. New York, NY: Elsevier Science.

Kirkpatrick, S.; Gelatt, C. D.; and Vecchi, M. P. 1983. Optimization by simulated annealing. *Science* (4598):671–680.

Park, J. D. 2002. MAP complexity results and approximation methods. In *Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence (UAI–02)*, 388–396.

Poole, D., and Provan, G. M. 1991. What is the most likely diagnosis? In Bonissone, P.; Henrion, M.; Kanal, L.; and Lemmer, J., eds., *Uncertainty in Artificial Intelligence 6*. New York, N. Y.: Elsevier Science Publishing Company, Inc. 89–105.

Shimony, S. E. 1993. The role of relevance in explanation I: Irrelevance as statistical independence. *International Journal of Approximate Reasoning* 8(4):281–324.

Yuan, C.; Lu, T.; and Druzdzel, M. J. 2004. Annealed MAP. In *Proceedings of the 20th Annual Conference on Uncertainty in Artificial Intelligence (UAI–04)*, 628–635.