Approximate Explanation of Reasoning in Bayesian Networks*

Wim Wiegerinck
SNN, Radboud University Nijmegen
Geert Grooteplein 21, 6525 EZ, Nijmegen, The Netherlands

Abstract

We address the question of explaining the computation of the posterior probability of a node in a Bayesian network. In polytrees, this probability can be explained by decomposing it into a product of causal terms (from the ancestors) and diagnostic terms (from the descendants). In models with loops, this decomposition is in general not valid. To proceed, we propose a scheme to approximate the posterior of a loopy model locally by a polytree. In this approximating model, the node probability can again be decomposed into causal and diagnostic terms. This decomposition can then be used as a rough explanation of reasoning for a user. The method is illustrated by numerical examples.

1 Introduction

One of the advantages of Bayesian networks (Pearl, 1988; Jensen, 1996; Castillo et al., 1997) is their model transparency. Reasoning in a Bayesian network, on the other hand, is often a complex computational process. For a human user, it is often difficult to oversee this process numerically. To help the human to understand what is going on in a Bayesian network a large number of explanation tools have been proposed, see (Lacave and Díez, 2002) for an overview. Here one can distinguish between several types of explanation. The most important ones are explanation of evidence and explanation of reasoning. An example of explanation of evidence is medical diagnosis. There the aim is the determination of factors (diseases, disorders) that caused –or explain– the observed anomalies in patient findings, (symptoms, test results outside the normal range etc.).

In this paper, we focus on explanation of reasoning. Explanation of reasoning is the explanation of how the numerical value of the posterior probability of a given node (the focal node) is achieved, and what are the factors that led to this probability. Understanding of reasoning in an expert system can enhance the acceptance of the system by human users. In addition, it can be helpful in building and debugging such systems. So, transparency of reasoning can be of importance in the model choice. For example, a nice property of a naive Bayes model is that the posterior probability of the parent node (which is typically the focal node) can be completely understood in terms of its prior and the likelihoods of the evidence at the child nodes. In polytrees (Pearl, 1988), a similar decomposition is possible for every node in the model (Sember and Zukerman, 1989). However in more general models, this factorization fails. For an exact explanation of reasoning, one should in principle consider the junction tree, or resort to conditioning arguments. In complex networks, both approaches will yield cumbersome, incomprehensible expressions. In this paper we aim to simplify this explanation by defining an approximate polytree around the focal node. The interactions of the focal node with its neighbors are copied from the original model. The neighboring nodes in the approximate model have adaptive biases which are optimized to the original model (with evidence). The approximate polytree can again be decomposed, thus giving an approximate explanation of reasoning.

This paper is organized as follow. In the remainder of this section we review Bayesian net-

*Workshop Probabilistic Graphical Models 2004 (PGM '04)

works. In the next section, we consider the case of polytrees, and review how posterior probabilities factorize into causal and diagnostic terms. In section 3, we propose the approximate model, and the optimization criteria. In section 4, we present some numerical results on toy models and we end in section 5 with discussion.

1.1 Bayesian networks

A Bayesian network is a probabilistic model P on a finite directed acyclic graph (DAG). For each node i in the graph, there is a random variables X_i which can assume the states x_i, together with a conditional probability distribution $p(x_i|x_{\mathsf{pa}(i)})$, where $\mathsf{pa}(i)$ are the parents of i in the DAG. In this paper, the number of states that X_i can assume is finite. The joint distribution of the Bayesian network is

$$P(x) = P(x_1, \ldots, x_n) = \prod_{i=1}^{n} p(x_i|x_{\mathsf{pa}(i)}) \quad (1)$$

Since a Bayesian network is a probabilistic model, one can compute marginal distributions, and conditional distributions by applying the standard rules of probability calculus. In general networks, these computations involve summations over many states. To keep these operations tractable, one often needs to order these summations, e.g. by applying a clustering algorithm such as the junction tree algorithm (Jensen, 1996; Castillo et al., 1997). In this paper, we are particularly interested in the the posterior distribution $P(x_f|e)$, which is the conditional distribution of the focal node x_f given the evidence e. The expression for this conditional distribution is

$$P(x_f|e) = \frac{\sum_{x\setminus x_f} \prod_k p(x_k|x_{\mathsf{pa}(k)})}{Z} \quad (2)$$

where it is understood that the summation is over the states that are compatible with the evidence. Z is a normalization factor, which is equal to the evidence $P(e)$.

2 The posterior distribution in polytrees

A polytree is a Bayesian network with a singly connected DAG. The message propagation algorithm by Pearl (Pearl, 1988) immediately implies that the posterior distribution in polytrees completely factorize into contributions from each of the parents and each of the children. We will rederive this result shortly in a slightly different way, by first considering the posterior on the *Markov blanket* (Pearl, 1988).

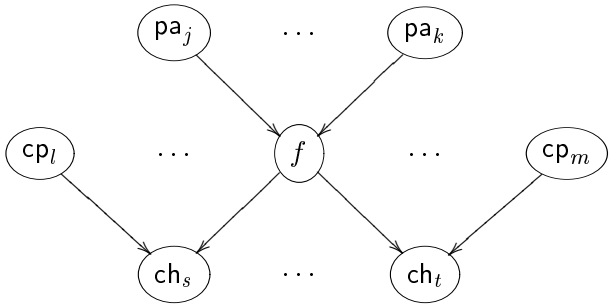

Figure 1: Markov blanket of node f in a polytree. pa are the parents of node f, ch are the children of node f, and cp are the co-parents of the children of node f.

2.1 Markov blanket

The Markov blanket $\mathsf{mb}(i)$ of node i consists, its parents $\mathsf{pa}(i)$, its children $\mathsf{ch}(i)$ and its coparents (the other parents of its children), $\mathsf{cp}(i)$. In this paper, we also consider the extended Markov blanket $\mathsf{emb}(i)$ which is the Markov blanket together with the node i itself, $\mathsf{emb}(i) = \mathsf{mb}(i) \cup i$. For notational convenience, the i dependence is suppressed if the node in question is the focal node f (i.e. $\mathsf{emb} \equiv \mathsf{emb}(f)$, etc.). See figure 1.

In general, the posterior probability distribution of the extended Markov blanket of the focal node f is given by

$$\begin{aligned} P(x_{\mathsf{emb}}|e) &= \frac{1}{Z} \sum_{x\setminus x_{\mathsf{emb}}} \prod_k p(x_k|x_{\mathsf{pa}(k)}) \\ &= \frac{1}{Z} p(x_f|x_{\mathsf{pa}}) \prod_{j\in\mathsf{ch}} p(x_j|x_{\mathsf{pa}(j)}) \\ &\quad \times \psi(x_{\mathsf{mb}}) \end{aligned} \quad (3)$$

Again the summation is over states that are compatible with the evidence. Z is a normalizing constant, and

$$\psi(x_{\mathsf{mb}}) = \sum_{x\setminus x_{\mathsf{emb}}} \prod_{i\setminus\{f,\mathsf{ch}\}} p(x_i|x_{\mathsf{pa}(i)}) \quad (4)$$

(again with summation compatible with the evidence). If we define the 'reduced model' $P_{\setminus f}$ as the original model P from which f and all the incoming links to the children $k \in \mathsf{ch}$ are removed,

$$P_{\setminus f}(x_{\setminus f}) \propto \prod_{i \setminus \{f,\mathsf{ch}\}} p(x_i|x_{\mathsf{pa}(i)}) \tag{5}$$

then (4) shows that the 'potential' ψ is proportional to the conditional probability on mb in the reduced model $P_{\setminus f}$

$$\psi(x_{\mathsf{mb}}) \propto P_{\setminus f}(x_{\mathsf{mb}}|e) \tag{6}$$

If P is a polytree, the graph of $P_{\setminus f}$ consists of disconnected subgraphs, and we can factorize ψ into potentials $\psi_j(x_j)$ $(\propto P_{\setminus f}(x_j|e))$ for each of the parents j and $\psi_k(x_{k,\mathsf{pa}(k)\setminus f})$ $(\propto P_{\setminus f}(x_{k,\mathsf{pa}(k)\setminus f}|e))$ for each of the children k, i.e.,

$$\psi(x_{\mathsf{mb}}) = \prod_{j \in \mathsf{pa}} \psi_j(x_j) \prod_{k \in \mathsf{ch}} \psi_k(x_{k,\mathsf{pa}(k)\setminus f}) \tag{7}$$

Substitution of (7) into (3) shows that the posterior distribution on the extended Markov blanket can be expressed as

$$P_{\mathsf{polytree}}(x_{\mathsf{emb}}|e) = \frac{1}{Z} p(x_f|x_{\mathsf{pa}}) \prod_{j \in \mathsf{pa}} \psi_j(x_j) \times \prod_{k \in \mathsf{ch}} p(x_k|x_{\mathsf{pa}(k)}) \psi_k(x_{k,\mathsf{pa}(k)\setminus f}) \tag{8}$$

Defining $\pi_j(x_j) = \psi_j(x_j)$ for the parents and

$$\lambda_k(x_f) = \sum_{x_{k,\mathsf{pa}(k)\setminus f}} p(x_k|x_{\mathsf{pa}(k)}) \psi_k(x_{k,\mathsf{pa}(k)\setminus f}) \tag{9}$$

for the children, we can express the posterior distribution of node f in polytrees as

$$P_{\mathsf{polytree}}(x_f|e) \propto \sum_{x_{\mathsf{pa}}} p(x_f|x_{\mathsf{pa}}) \prod_{j \in \mathsf{pa}} \pi_j(x_j) \times \prod_{k \in \mathsf{ch}} \lambda_k(x_f) \tag{10}$$

By summing over the parents we obtain Pearl's expression for the posterior distribution in polytrees,

$$P_{\mathsf{polytree}}(x_f|e) \propto \pi(x_f) \prod_{k \in \mathsf{ch}} \lambda_k(x_f) \tag{11}$$

with

$$\pi(x_f) = \sum_{x_{\mathsf{pa}}} p(x_f|x_{\mathsf{pa}}) \prod_{j \in \mathsf{pa}} \pi_j(x_j) \tag{12}$$

In polytrees, we can explain the posterior distribution (in the form (10)) in terms of causal support $\pi_j(x_j)$ contributed by each of the parents and diagnostic support $\lambda_k(x_k)$ contributed by each of the children (Sember and Zukerman, 1989).

In models that are not polytrees, the posterior of the focal node may still be explained in this way as long as the extended Markov blanket at the focal node factorizes as in (8). In other cases, the posterior distribution cannot be explained in terms of individual causal and diagnostic supports. In some cases one can cluster parent nodes into non-overlapping parent super-nodes γ and children nodes into non-overlapping children super-nodes κ such that the posterior assumes the form

$$P_{\mathsf{clustertree}}(x_f|e) \propto \sum_{x_{\mathsf{pa}}} p(x_f|x_{\mathsf{pa}}) \prod_{\gamma \subset \mathsf{pa}} \pi_\gamma(x_\gamma) \times \prod_{\kappa \subset \mathsf{ch}} \lambda_\kappa(x_f) \tag{13}$$

with

$$\lambda_\kappa(x_f) = \sum_{x_{\kappa,\mathsf{pa}(\kappa)\setminus f}} \prod_{k \in \kappa} p(x_k|x_{\mathsf{pa}(k)}) \psi_\kappa(x_{\kappa,\mathsf{pa}(\kappa)\setminus f}) \tag{14}$$

where $\mathsf{pa}(\kappa)$ is the union of all parents of children node $k \in \kappa$. In cases where this is not possible (e.g., where parents and children of f are not separated by f), one can try to present the expressions for the posterior in some other way. This approach is expected to give cumbersome expressions, which will be difficult to comprehend by a user. Therefore we do not pursue this direction. Our approach is still to try to give an explanation in terms of individual causal and diagnostic supports. This explanation is not exact, but in some way approximate.

3 Approximate explanation

3.1 Desiderata

The goal of this paper is to approximately explain the posterior distribution of a focal node in

terms of causal support from each of the parents and diagnostic support from each of the children. This goal is not exactly defined. Therefore we define some desiderata for our procedure.

- If an exact explanation is possible, then the procedure should reproduce this explanation.
- The explanation should reflect the interactions of the focal node with its direct neighbors.
- The reconstruction of the focal node probability from the causal and diagnostic terms should be equal to the exact posterior of the focal node.
- If the probability can be factorized into cluster factors, the method should respect this factorization. In particular, if a cluster contains only a single node (parent or child of the focal node), the corresponding causal or diagnostic term should be recovered.

3.2 Approximating Markov blanket posterior

Our approach is to explain the posterior of the focal node by first approximating the posterior of the extended Markov blanket, and then do the explaining in this approximating model. Our approximating extended Markov blanket posterior will be of the form

$$Q(\mathbf{x}_{\mathsf{emb}}) = \frac{1}{Z} p(x_f|x_{\mathsf{pa}}) \prod_{j \in \mathsf{pa}} \phi_j(x_j) \times \prod_{k \in \mathsf{ch}} p(x_k|x_{\mathsf{pa}(k)}) \phi_k(x_{k,\mathsf{pa}(k)\setminus f}) \quad (15)$$

in which the tables $p(x_f|x_{\mathsf{pa}})$ and $p(x_k|x_{\mathsf{pa}(k)})$ are copied from the original model P and kept fixed. The potentials ϕ are to be optimized. The factor Z is for normalization.

Note that the model Q is of the same functional form as (8). However, since parents and co-parents of f may intersect, or different children may share some other parents other than f, in other words, if the extended Markov blanket of f contains loops, this functional form is not sufficient for the model to be a local polytree. To remedy this, we insert cloned nodes in the original model where necessary. For example if child k and child l share the parent i, we insert clones i_k and i_l of i and make them parents of node k and l respectively. The node i is disconnected from k and l, and connected as a parent to its clones $c = i_k$ and $c = i_l$ with probability table

$$p_{\mathsf{clone}}(x_c|x_k) = \begin{cases} 1 & \text{if } x_c = x_i \\ 0 & \text{otherwise} \end{cases} \quad (16)$$

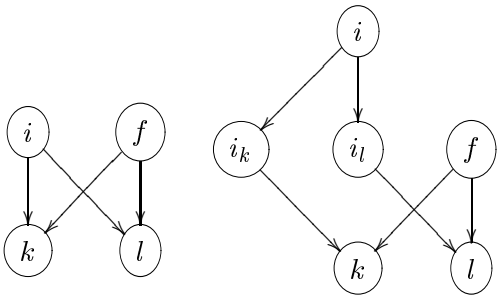

Figure 2: Left: Example a network such that the extended Markov blanket of focal node x is loopy. Right: Inclusion of cloned nodes i_k, i_l that clones i (i.e. $p(x_{i_k}|x_i) = 1$ if $x_{i_k} = x_i$ and 0 otherwise) remove loops from the extended Markov blanket of focal node f without changing the model.

This procedure is illustrated in figure 2. Since cloned nodes are hard-coupled to their parent, it will not have any effect on the probability of the original nodes. However, the inclusion of cloned nodes guarantees that the approximating extended Markov blanket posterior (15) is indeed a polytree as desired.

3.3 Optimizing the approximating model

It is tempting to optimize the parameters ϕ by directly minimizing the *Kullback-Leibler divergence* (Whittaker, 1990) between $P(x_{\mathsf{emb}}|e)$ and $Q(x_{\mathsf{emb}})$,

$$\mathsf{KL}(P(X_{\mathsf{emb}}|e)||Q(X_{\mathsf{emb}})) = \sum_{\{x_{\mathsf{emb}}\}} P(x_{\mathsf{emb}}|e) \log \frac{Q(x_{\mathsf{emb}})}{P(x_{\mathsf{emb}}|e)} \quad (17)$$

This procedure would lead to an approximation in which the node probabilities $Q(x_i)$ equal the node probabilities $P(x_i|e)$ for all the nodes except for, unfortunately, the focal node f [1]. It is desirable to have at least the focal node right, and therefore, we adapt this procedure by minimizing KL under the constraint that the focal node has the right probability. Since there is no guarantee (at least not known to us) that there is a solution which satisfies the constraint, we implement the constraint as a penalty term and minimize $E(Q)$,

$$\begin{aligned}\beta E(Q) &= \mathsf{KL}(P(X_{\mathsf{emb}}|e)||Q(X_{\mathsf{emb}})) \\ &+ \beta\mathsf{KL}(P(X_f|e)||Q(X_f)) \quad (18)\end{aligned}$$

in the limit $\beta \to \infty$.

However, even this minimization may yield undesired effects. For instance in a model with only one parent of f, and in which ancestors of f are separated from the descendants of f (so the causal support $\pi(x)$ is well defined), the outcome of the procedure may yield a causal support that is not equal to the true causal support. Such a deviation in causal support can be the result of a compensation of the misfit in x_f due to fitting the node probabilities of the children. To remedy this compensation effect that mix causal and diagnostic support, we will first disentangle them before we optimize the model parameters. After disentangling, we will optimize the parameters for causal and diagnostic support separately, which guarantee that they will not be mixed.

3.3.1 Step 1: Disentangling causal and diagnostic support

We disentangle causal and diagnostic information as follows: We clone the focal node f by g in the original model, disconnect f from its other children and connect these to g. Next we decouple f and g. Since g is now a root, it needs a prior $r(x_g)$. The table values of this prior is to be choosen. Below, we propose a procedure for choosing these values. The resulting distribution of the disentangled model R_r is

$$R_r(x) = p(x_f|x_{\mathsf{pa}})r(x_g)\prod_{i\neq f,g} p(x_i|x_{\mathsf{pa}(i)}) \quad (19)$$

The subindex r is to indicate that the model dependents on the choice of r. The procedure is sketched in figure 3.

If P would be a model where ancestors and descendants are separated by the focal node, – for instance if P is a polytree – then it is clear that the total causal contribution is $\pi_r(x_f) = R_r(x_f|e)$ and the total diagnostic contribution is proportional to $\hat{\lambda}_r(x_g) = R_r(x_g|e)/r(x_g)$, regardless the table values of r. Defining $\lambda_r(x_f)$ as the table for x_f with the same table values as $\hat{\lambda}_r(x_g)$, the posterior is proportional to the product of both terms,

$$P(x_f|e) \propto \pi_r(x_f)\lambda_r(x_f) \quad (20)$$

If there are other connections from ancestors to descendants, the choice of r matters and gives different results (the higher r, the more will be explained by g and less by ancestors of f). In addition, (20) will not hold in general. Since one of the desiderata is that (20) does hold, we consider the product of causal and diagnostic terms

$$pr_r(x_f) \propto \pi_r(x_f)\lambda_r(x_f) \quad (21)$$

and tune r such that pr_r matches the exact posterior as close as possible. Using the KL to measure the mismatch, the optimal r_{opt} is

$$r_{\mathsf{opt}} = \operatorname*{argmin}_r \mathsf{KL}(P(X_f|e)||pr_r(X_f)) \quad (22)$$

and our final disentangled model is $R \equiv R_{r_{\mathsf{opt}}}$.

3.3.2 Step 2: Optimizing the causal terms

Now that we have disentangled the total causal from the total diagnostic support, we can further factorize the total causal support $\pi(x_{f,\mathsf{pa}}) = R(x_{f,\mathsf{pa}}|e)$. First we try to find non-overlapping clusters γ such that

$$R(x_{\mathsf{pa}}|e) = \prod_\gamma R(x_\gamma|e) \quad (23)$$

[1] This can be verified by computing the gradient of (17) with respect to ϕ_j and ϕ_k (where $j \in \mathsf{pa}$ and $k \in \mathsf{ch}$) and setting it to zero . The resulting stationary equations are $Q(x_j) = P(x_j|e)$ and $Q(x_{k,\mathsf{pa}(k)\setminus f}) = P(x_{k,\mathsf{pa}(k)\setminus f}|e)$.

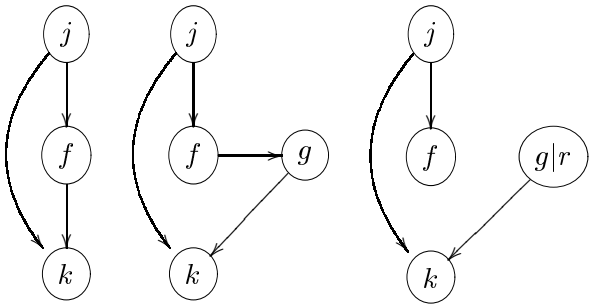

Figure 3: Disentangling causal from diagnostic support. Left: original model P with focal node f. Middle: original model with focal node f and its clone g. Right: disentangling model R in which f and g are disconnected. The probability of x_g depends on the table values of r

If each cluster contains exactly one parent, we are done. Otherwise we aim for each cluster γ to find a product distribution $\prod_{j\in\gamma}\phi_j(x_j)$ such that

$$\sum_{x_\gamma}\pi(X_f|x_{\mathsf{pa}})\prod_{j\in\gamma}\phi_j(x_j)=\pi(X_f|x_{\mathsf{pa}_\setminus\gamma}) \quad (24)$$

for each state of the remaining parents. In such a solution, the causal support from cluster γ is optimally preserved. The solution is the minimum of the set of objective functions

$$\begin{aligned} &E(\{\phi_j\}_{j\in\gamma}|x_{\mathsf{pa}\setminus\gamma}) \\ &\quad = \quad \mathsf{KL}(\sum_{x_\gamma}\pi(X_f|x_{\mathsf{pa}})\pi(x_\gamma)\ldots \\ &\qquad \ldots||\sum_{x_\gamma}\pi(X_f|x_{\mathsf{pa}})\prod_{j\in\gamma}\phi_j(x_j))) \end{aligned} \quad (25)$$

for each state of the remaining parents. To combine this into one cost function, we weight these by the probability of the states,

$$E(\{\phi_j\}_{j\in\gamma})=\sum_{x_{\mathsf{pa}\setminus\gamma}}\pi(x_{\mathsf{pa}\setminus\gamma})E(\{\phi_j\}_{j\in\gamma}|x_{\mathsf{pa}\setminus\gamma}) \quad (26)$$

This optimization problem is equivalent to

$$\begin{aligned} E(\{\phi_j\}_{j\in\gamma}) \;&=\; \mathsf{KL}(\pi(X_{f,\mathsf{pa}\setminus\gamma})\ldots \\ &\quad \ldots||\sum_{x_\gamma}\pi(X_{f,\mathsf{pa}\setminus\gamma})\prod_{j\in\gamma}\phi_j(x_j))) \end{aligned} \quad (27)$$

which can be viewed as a maximizing likelihood problem with missing values and thus can be solved by the *Expectation-Maximization* (E-M) algorithm (Dempster et al., 1977). The resulting causal support is

$$\pi_Q(x_f)=\sum_{x_{\mathsf{pa}}}p(x_f|x_{\mathsf{pa}})\prod_j\phi_j(x_j) \quad (28)$$

3.3.3 Step 3: Optimizing the diagnostic terms

Finally, we have to factorize the diagnostic support. Again we start with searching for non-overlapping clusters $\kappa\subset\mathsf{ch}\cap\mathsf{cp}$ such that

$$R(x_{\mathsf{ch},\mathsf{cp}}|x_g,e)=\prod_\kappa R(x_\kappa|x_g,e_\kappa) \quad (29)$$

and we fit for each cluster the approximating distribution

$$\begin{aligned} Q_\kappa(x_{g,\kappa}) \;&\equiv\; \frac{1}{Z}\pi_Q(x_g)\prod_{k\in\mathsf{ch}\cap\kappa}p(x_k|x_{\mathsf{pa}(k)}) \\ &\quad\times\; \phi_k(x_{k,\mathsf{pa}(k)\setminus g}) \end{aligned} \quad (30)$$

parameterized by the potentials ϕ_k to the disentangled distribution R_{π_Q} on the set $\kappa\cup g$ by minimization of

$$\begin{aligned} \beta E(Q_\kappa) \;&=\; \mathsf{KL}(R_{\pi_Q}(X_{g,\kappa}|e)||Q_\kappa(X_{g,\kappa})) \\ &\quad+\; \beta\mathsf{KL}(R_{\pi_Q}(X_g|e)||Q_\kappa(X_g)) \end{aligned} \quad (31)$$

in the limit $\beta\to\infty$.

Again this KL minimization problem is closely related to log-likelihood maximization problems with missing values, and the E-M algorithm may be invoked for the optimization. In practice, this minimization procedure involves a cooling scheme, where one starts with minimizing with small β. The optimized ϕ's are then used as starting point in the next optimization with larger β. A two step approximation which still seems to work (at least for small problems) is obtained by setting the first β equal to 0 and the second one to ∞.

After optimization, we have potentials for the parents ϕ_j from minimizing (27) and potentials for the children (and co-parents) ϕ_k from minimizing (31). With (15), these are recombined to an approximating polytree on the extended Markov blanket. By summing co-parents, and then over children, approximating diagnostic

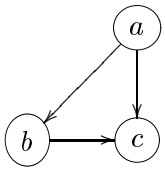

Figure 4: Network structure in the numerical experiments.

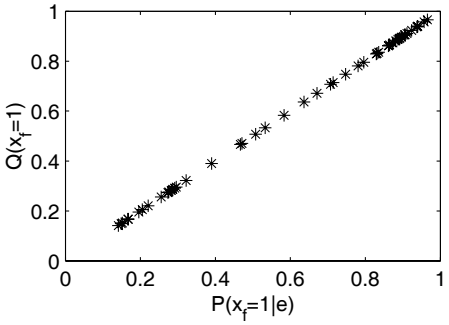

Figure 5: Approximating posterior $Q(x_f)$ against true posterior $P(x_f|e)$

support terms λ_k can be computed. The final result can be written in the form (10), and presented in some way to the user.

4 Numerical examples

In this section, we illustrate the method by some results on toy problems. The toy model is a fully connected model $P(a)P(b|a)P(c|a,b)$ as in figure 4. States are binary $(0,1)$. In the first problem, we check whether in this small network the exact posterior distributions of the focal node are recovered by the approximating distribution of the focal node. We do this by defining random tables, and randomly take node a or b as the focal node x_f. Node c is clamped to 1. In all these networks the posterior $P(x_f|c=1)$ is compared to $Q(x_f)$ achieved by the method. The results are plotted in figure 5. These show that, at least in these toy models, the exact posterior is always recovered.

In the second experiment, we take a as the focal node, and node c is again clamped to 1. We take $p(a) = 0.5$ (using notation $p(a) = p(a=1)$ and $p(\bar{a}) = p(a=0)$). Furthermore, we take $p(c|a,b) = 0.99$, $p(c|\bar{a},b) = p(c|a,\bar{b}) = 0.099$ and $p(c|\bar{a},\bar{b}) = 0.0099$. The probability of b is parameterized as $p(b|a) = p(\bar{b}|\bar{a}) = u$.

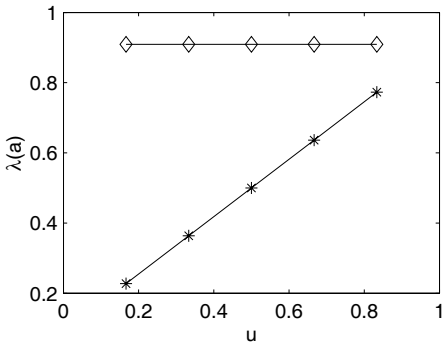

Figure 6: Diagnostic support $\lambda_b(a)$ (stars) and $\lambda_c(a)$ (diamonds) as a function of the interaction strength u between a and b. See text for more model details.

In figure 6, the outcomes of the normalized diagnostic support $\lambda_b(a)$ and $\lambda_c(a)$ are plotted as function of the interaction strength u. (normalization of the λ's means that we multiplied both components by a constant such that $\lambda(\bar{a}) = 1 - \lambda(a)$). We see that $\lambda_c(a)$ is constant, satisfying $\lambda_c(a) = 10\lambda_c(\bar{a})$. This is a natural outcome since the state a is 10 times as likely as the state $\bar{a}$ when the state c is clamped, regardless of the probability of b. Furthermore, we see that with increasing interaction term u, the support via b increases, as expected. Note that with $u = 0.5$, there is no interaction and there should also be no support ($\lambda_b(a) = \lambda_b(\bar{a}) = 0.5$) as is indeed the case. With u smaller than 0.5, there is 'negative' support. In this case a and b are anti-correlated, and there is a significant probability that the evidence is explained by $\bar{a}, b$.

5 Discussion

We presented a method for an approximate explanation of the posterior distribution of a focal node in terms of causal support from each of its parents and diagnostic support from each of the children. The method aims to make a decomposition that looks similar to the exact decomposition in polytrees. The method consists of fitting a local polytree to posterior distribution of the extended Markov blanket around the

focal node.

We feel that there is need for such a decomposition. It is our experience that in particular users with superficial but non-zero acquaintance with probabilistic models – typically this acquaintance is limited to the naive Bayes model (one parent connected to a number of children) – tend to try to understand the posterior in terms of influences of neighboring nodes. An automated method to decompose the posterior according some criteria could be advantageous to a rough and more ad-hoc estimation by hand.

The notion of approximate explanation is not well defined. In our paper, we made several choices for this approximation. Some of them may be better founded than others, and certainly more discussion is needed on the desiderata of an approximate method. To find out whether the method is helpfull, and to find out what is further needed for a practical method, field tests with implementations in real-world expert systems with human users/modelers need to be performed.

Finally, we would like to remark that the method is very different from the loopy belief propagation algorithm which recently received much attention (Murphy et al., 1999). This latter algorithm is a global algorithm for approximate inference in models for which the posterior distribution is intractable for exact computation. Our method on the contrary, requires the exact posterior distribution as an input to make a local fit by a polytree.

Acknowledgments

This research is supported by the Dutch Technology Foundation STW. I would like to thank the anonymous reviewers for useful comments.

References

Castillo, E., Gutierrez, J. M., and Hadi, A. S. (1997). *Expert Systems and Probabilistic Network Models.* Springer.

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum-likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, B39:1–38.

Jensen, F. (1996). *An Introduction to Bayesian networks.* UCL Press.

Lacave, C. and Díez, F. (2002). A review of explanation methods for bayesian networks. *The Knowledge Engineering Review*, 17(2):107–127.

Murphy, K. P., Weiss, Y., and Jordan, M. I. (1999). Loopy belief propagation for approximate inference: An empirical study. In *Proceedings of Uncertainty in AI*, pages 467–475.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent systems: Networks of Plausible Inference.* Morgan Kaufmann Publishers, Inc.

Sember, P. and Zukerman, I. (1989). Strategies for generating micro explanations for Bayesian belief networks. In *Proceedings of the 5th workshop on Uncertainty in Artificial Intelligence*, pages 295–302.

Whittaker, J. (1990). *Graphical Models in Applied Multivariate Analysis.* Wiley, New York.