Tailored-to-fit Bayesian Network Modeling of Expert Diagnostic Knowledge

Ruxandra Lupas Scheiterer, Dragan Obradovic and Volker Tresp

Siemens AG, Corporate Technology, Information and Communications

Abstract. This paper addresses issues in constructing a Bayesian network domain model for diagnostic purposes from expert knowledge. Diagnostic systems rely on suitable models of the domain, which describe causal relationships between problem classes and observed symptoms. Typically these models are obtained by analyzing process data or by interviewing domain experts. The domain models are usually built in the forward direction, i.e. by using the expert provided probabilities of symptoms given individual causes and neglecting the information in the backward direction, i.e. the knowledge about probabilities of problems given individual symptoms. In this paper we introduce a novel approach for the structured generation of a model that incorporates as closely as possible that subset of the unstructured multifaceted and possibly conflicting probabilistic information provided by the experts that they feel most confident in estimating.

1 Introduction

Bayesian (causal) networks are directed acyclic graphs with edges whose direction indicates causality and with parameters that capture the joint probability functions of the involved variables. Bayesian networks arose out of an attempt to add probabilities to expert systems, and this is still their most common use. Bayesian networks are well suited for the diagnostic examination of domains with many random variables, which are interrelated in non-transparent, non-deterministic relationships. They form a graphical representation of the domain variables and model their dependence and independence relations. Reference [1] gives an example of the application of troubleshooting via Bayesian networks in the domain of GSM cellular mobile systems, and [2], [3] of an application in the medical domain. Following a short review of Bayesian networks, we address issues of modeling a domain with such networks. The main limitation in building a Bayesian network based on expert knowledge is that the expert can provide only a subset of the marginal and conditional information needed to fully describe the joint probability of the problem and symptom variables. In addition, the expert provided information usually includes probability estimates in both directions, i.e from problems to symptoms and vice versa. Typically, Bayesian network models are built by exploiting the available information about the conditional probabilities in one direction only, usually the expert estimates of conditional probabilities of symptoms given individual problems, and then

making assumptions about the missing conditional probabilities describing the probabilities of symptoms given several problems simultaneously. In this paper we propose a novel way to match the model to that part of domain information that can be provided with ease and confidence by human experts. The herein proposed SFOBE (Smallest Forward-Backward Expert-based) approach enables the use of expert provided estimates of conditional probabilities in both directions, problem given symptoms and vice versa, in building Bayesian network based expert domain models. The use of the algorithm is illustrated on an example domain with four problems and two symptoms.

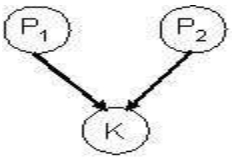
2 Forward and Backward Modeling Aspects

In this section we examine modeling aspects in constructing a Bayesian network for diagnosis assistance from expert provided information given in both directions, forward and backward. *Forward probabilities* are those needed to build a Bayesian network as detailed in appendix A, points 2 and 3, namely the marginal probabilities of root nodes and the conditional probabilities of children given parents. *Backward probabilities* are the marginal probabilities of leaf nodes and the conditional probabilities of parents given children. There are *two operating directions of a causal model:* forward, for model construction and simulation, and backward, for diagnosis. Our considerations are applicable to domains governed by cause and effect principles. In this paper we consider only binary variables.

Notation: To shorten formulas and derivations we abbreviate $P(R = 1) = P(R)$, $P(R = 0) = P(\overline{R})$. To avoid confusion we mark distributions by a "d" superscript, as in $P^d(K|P_1, P_2)$. The terms root-cause, cause and problem are used interchangeably to denote the parent in the examined causal hierarchy; and so are the terms effect, symptom or indicator, which denote the child.

2.1 Two Root-Causes

We first illustrate the various alternatives for modeling a causal dependence, and the associated degrees of freedom, on the simple case of a single binary symptom K that is causally dependent on two binary problems P_1 and P_2, as shown in Fig. 1. The table lists the conditional probabilities $P^d(K|P_1, P_2)$ for all 2^3 values of the three random variables. For example $P(K = 1|P_1 = 1, P_2 = 0) = p_1$ and $P(K = 1|P_1 = 0, P_2 = 1) = p_2$, while $P(K = 1|P_1 = 0, P_2 = 0) = l$ is the so-called "leak" probability, which is the probability that the effect is present even though none of the causes within the considered domain is present. As previously said, in order to completely determine a causal model, the following *forward* probabilities have to be specified: the probabilities of the root nodes, and the conditional probability tables of any child node given its parents. In this case this means specifying the probability distributions of the random variables P_1, P_2 and $K|P_1, P_2$, that is, the following 6 numbers, the other being determined by the requirement that the probability of disjoint mutually exhaustive events



$P^d(K\|P_1,P_2)$	0		1		P_1
	0	1	0	1	P_2
K=0	1-l	1-p_2	1-p_1	1-p	
K=1	l	p_2	p_1	p	

Fig. 1. Cause-to-effect model of simple domain with two problems and a single common symptom

sums to one:

$$\begin{array}{lll} P(P_1), & P(K|P_1,\overline{P_2}) = p_1, & P(K|P_1,P_2) = p \text{ (free parameter)} \\ P(P_2), & P(K|\overline{P_1},P_2) = p_2, & P(K|\overline{P_1},\overline{P_2}) = 0 \text{ or } l \text{ (leak)} \end{array}$$

The expert can readily estimate the values of the marginal probabilities $P(P_1)$, $P(P_2)$ which are the probabilities that problem P_i is present, the values p_1, p_2 that the symptom is present given that exactly one of the alternative causes is present, as well as the "leak" probability that the effect is due to a cause outside the modeled domain. But it is in general very difficult to obtain an estimate of the value of p, because two or more causes being present at the same time is such a rare event that the expert has no intuition about this estimate.

However in our experience the expert readily provides estimates of the values of the *backward* probabilities $P(K)$, $P(P_1|K)$ and $P(P_2|K)$. The first value is the probability that symptom K is present, and the other numbers are the probabilities that problem P_i is present given that symptom K is observed. These are estimates that the expert is quite comfortable with, since this is the direction of reasoning when doing troubleshooting. The expert detects a problem by the alarm state of an indicator, and then reasons backwards, thinking what problem is the most likely, given this finding.

When constructing the model from probability estimates provided by the expert or in the presence of incomplete data, it is desirable to use exactly those probabilities that the expert can specify with the highest confidence. These are:

$$\begin{array}{llll} P(P_1), & P(K|P_1,\overline{P_2}) = p_1, & P(K|\overline{P_1},\overline{P_2}) = l, & P(P_1|K) \\ P(P_2), & P(K|\overline{P_1},P_2) = p_2, & P(K), & P(P_2|K) \end{array}$$

that is, a mixture of forward and backward probabilities. Thinking of the freedom left in the model after specification of the forward probabilities (first 5) we see that *we would like to match 3 backward probabilities, having only one free parameter, p.* Thus one might ask for a causal model that exhibits probabilities with the "closest" approximation to the desired ones using distance measure $\mathcal{D}$, e.g. in the minimum mean square error sense, or in the Kullback-Leibler sense. There are several ways to do the proposed optimization of the model to the provided expert probabilities. We mention two:

1. Fix the forward probabilities, optimize only over the free parameter p, such that the backward probabilities are approximated as closely as possible. That is, define the cost function:

$$\begin{aligned} C_1 = \; & \mathcal{D}\left(P(K)_{(P(P_1),P(P_2),p_1,p_2,l,p)},\; P^t(K)\right) \\ + \; & \mathcal{D}\left(P(P_1|K)_{(P(P_1),P(P_2),p_1,p_2,l,p)},\; P^t(P_1|K)\right) \\ + \; & \mathcal{D}\left(P(P_2|K)_{(P(P_1),P(P_2),p_1,p_2,l,p)},\; P^t(P_2|K)\right) \end{aligned} \tag{1}$$

where in each parenthesis the second term is the target value specified by the expert, and the first term is the value resulting from fixing the functional arguments given in subscript parentheses.

2. Alternatively allow the forward probabilities to deviate from their specified values by a small amount ϵ. Or by ϵ_i, if a good reason for different "noises" along the different dimensions exists. That is, define

$$\begin{aligned} C_2 = C_1 \; & + \mathcal{D}\left(P(P_1), P^t(P_1)\right) + \mathcal{D}\left(P(P_2), P^t(P_2)\right) \\ & + \mathcal{D}\left(p_1, p_1^t\right) + \mathcal{D}\left(p_2,\; p_2^t\right) + \mathcal{D}\left(l,\; l^t\right) \end{aligned} \tag{2}$$

where in each summand the first term is the variable, and the second value is the target value specified by the expert. Both for $C = C_1$ and for $C = C_2$ the desired optimum of the cost function C over the variable space is the solution of the minimization:

$$\begin{array}{ll} \min\limits_{p \in [0,1]} & C \\ \text{subj.to} & \begin{array}{c} \mathcal{D}\left(P(P_1) - P^t(P_1)\right) \le \epsilon \\ \mathcal{D}\left(P(P_2) - P^t(P_2)\right) \le \epsilon \\ \mathcal{D}\left(p_1 - p_1^t\right) \le \epsilon \\ \mathcal{D}\left(p_2 - p_2^t\right) \le \epsilon \\ \mathcal{D}\left(l - l^t\right) \le \epsilon \end{array} \\ \text{s.t.} & \begin{array}{c} P(K) \in [0,1], \\ P(P_1|K) \in [0,1] \\ P(P_2|K) \in [0,1] \end{array} \end{array} \tag{3}$$

The minimization with cost function C_2 of Eq. (2) includes the one with cost function C_1 of Eq. (1) as a special case, as can be seen by setting $\epsilon = 0$ and will therefore result in a lower or equal minimum, i.e. a closer approximation, at the expense of increased computational effort.

2.2 Effect of a Leak from Outside the Domain

Concerning the "leak" mentioned before: If for a given symptom K the expert feels that even in the absence of all the modeled causes within the considered domain the probability of seeing the symptom present is nonzero, then this can be modeled so that the effect K results from an effect "KD" within the domain at hand and an effect L from outside this domain, as first introduced in [10]. Since "KD" and L are combined into K via an OR-junction, the following equation holds for the probabilities of K, KD and L:

$$P(\,\overline{K}\,) = P(\,\overline{KD}\,)P(\,\overline{L}\,) \tag{4}$$

From this equation we see that by the addition of a leak, $P(\overline{K})$ can not increase, hence $P(K)$ cannot decrease. Hence adding a leak to the model makes sense if

and only if the marginal probability $P(K)$ resulting from inputting the forward probabilities into the Bayesian network is *smaller* than the probability $P(K)$ specified by the expert or resulting from the data. If this is not the case, then the probability $P(K)$ in the Bayesian network has to be decreased. One possibility is the introduction of a so called "inhibitor" as described in [13]. Usually the leak "L" is modeled implicitly by inclusion of a corresponding entry "l" in the probability table of symptom K.

2.3 n Causes

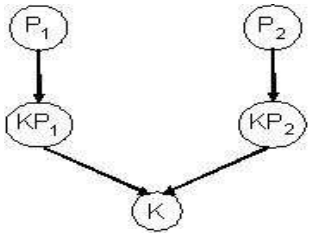
In the case of n problems the conditional probability table of the binary random variable $K|P_1, ..., P_n$ has 2^n entries that have to be specified. Out of these, according to our experience, the expert finds it feasible to specify $2n+2$ probabilities, namely n forward values of the form $P(K|P(i), \overline{P_1}, ..., \overline{P_{i-1}}, \overline{P_{i+1}}, ..., \overline{P_n})$, i.e. "probability that the symptom is present given that exactly one problem is present", the value of the leak $l = P(K|\overline{P_1}, ..., \overline{P_n})$, i.e. "probability that the symptom is present given that no problem is present", $P(K)$, probability of the symptom being present, and n backward probabilities of the form $P(P_i|K), i = 1, ..., n$, which are the probabilities that the problem is present given the symptom present. If several symptoms $K_1, \ldots, K_m$ are present, this has to be done in parallel for each symptom. For $n = 2$ as we saw the model is under-dimensioned, and can only approximate the expert estimates in for example a mean-square sense. Equality is given for 3 problems, $n = 3$, when $2^n = 2n + 2$, as after constructing the model there are exactly 4 free parameters to capture the 4 specified backward probabilities. For $n > 3$ the model is over-dimensioned compared to what an expert can reasonably specify.

2.4 Noisy-OR, or Reducing Complexity via Proxy Modeling

The classical way to reduce complexity in Bayesian networks is to model the n-way interaction as "Noisy-OR", first introduced by [9], or as "noisy" versions of AND, MAX, MIN, ADDER [11], SUM or ELENI [12]. For "Noisy-OR" this means associating with each cause an inhibitory mechanism that sometimes prevents the cause from producing the effect, and linking the single noisy causes with an OR function. This model has n unknowns in the $P(K|P_i)$ probability table, one for each inhibitory mechanism, reducing the original exponential assessment burden of 2^n, to n. Note that if a simple OR function would be used to link the causes, instead of a "Noisy-OR" function, there would be *no* unknown available to tune the model to the probabilities found in the domain at hand.

The way complexity is reduced in *proxy modeling* such as in the HealthMan project [2],[3], is by introduction of intermediate so-called Noisy-OR "proxy" nodes between problem and indicator, Fig. 2. The significance of the proxy nodes KP_i, "K due to P_i", is that they capture the event that the indicator K is due to precisely the problem P_i, which is seen from the probability table: $P(KP_i = 1|P_i = 0) = 0$, together with the following OR-junction of the proxies. That is, given that the P_i is not present, a possible $K = 1$ value cannot be due to P_i.

The following relations hold between the probabilities in the proxy model and



$P^d(KP_1\|P_1)$	0	1	P_1
KP_1=0	1	1-p_1'	
KP_1=1	0	p_1'	

$P^d(KP_2\|P_2)$	0	1	P_2
KP_2=0	1	1-p_2'	
KP_2=1	0	p_2'	

$P^d(K\|KP_1,KP_2)$	0		1		KP_1
(OR with leak)	0	1	0	1	KP_2
K=0	1-l'	0	0	0	
K=1	l'	1	1	1	

Fig. 2. Causal model of two problems with Noisy-OR proxy nodes between problem and indicator layers.

those specified by the expert:

1. The leaks in the two model are identical.
2. The p_1' and the p_2' in the proxy model are related to the p_1, p_2 of the direct model via: $p_1 = p_1' + l * (1 - p_1')$ $p_2 = p_2' + l * (1 - p_2')$.
3. In the absence of a leak, the p_1', p_2' in the proxy model are identical to the p_1, p_2 in the direct model.
4. In the proxy model there is no free parameter. $P(K|P_1, P_2)$ is not a free parameter, as in the direct model, but fixed by specification of p_1, p_2, l.

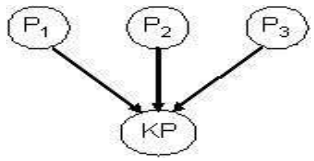
In our work we have found that this Noisy-OR proxy model is quite convenient to model the expert knowledge for a series of domains where troubleshooting or diagnosis has to be performed. However, after specifying the forward probabilities, the Noisy-OR proxy model has only one free parameter left. So this model has the limitation that it has no means to accommodate the expert-specified backward probabilities. Since it is important to reduce the complexity to an appropriate amount to be able to take into account, at least to a certain degree, all those probabilities that are easy to obtain, this paper proposes an enhanced model, which we call the SFOBE model: the Smallest Forward-Backward Expert-based model.

3 The SFOBE Model: the Smallest Forward-Backward Expert-based Model

The SFOBE model is the smallest model that has as least as many free parameters as there are expert-specified backward probabilities. Here "smallest"

means that there is no other model fulfilling this requirement that has fewer free parameters.

We propose a building block that is more complex than the Noisy-OR proxy building block and contains the same number of free parameters as there are given backward expert estimates that concern the involved problem nodes. Let a **three-cluster** denote three problem nodes with a common proxy and no leak, as shown in Fig.3. A general three-cluster would have 8 free parameters, but since



$P^d(KP\|P_1,P_2,P_3)$	0				1				P_1
	0		1		0		1		P_2
	0	1	0	1	0	1	0	1	P_3
$KP=0$	1	1-p_3'	1-p_2'	1-a	1-p_1'	1-b	1-c	1-d (or 0)	
$KP=1$	0	p_3'	p_2'	a	p_1'	b	c	d (or 1)	

Fig. 3. Definition of a three-cluster

a separate leak for each problem does not make sense, we can restrict attention to a three-cluster without leak, as shown in the probability table in Fig.3. So in a three-cluster if all the entering problem node values are zero we have a cluster node value of zero. Hence in a three-cluster there are 7 free parameters.

3.1 SFOBE Model Construction

With the three-cluster building-block, the SFOBE model that incorporates the $n+1$ forward and $n+1$ backward probabilities is built according to the following **steps**, which are detailed in section 3.2:

1. Build a model that for each symptom K lumps sets of 3 problems causally related to K into three-clusters and merge these into the node for symptom K via an OR-with-leak connection, as shown in Fig.4.
2. The root-node marginal probabilities are among the given forward probabilities. In the conditional probability tables in the top layer the given conditional forward probabilities together with the given leak determine the entries where exactly one problem is in state 1, leaving as only unknowns the entries where two or all three problems are in state 1. The OR-with-leak probability table is specified completely since the leak is known.
3. (a) Set $P(P_1=1,P_2=1,P_3=1)=1$. Compute the values of the remaining free parameters from the given backward probabilities $P(P_i|K)$. If the obtained parameter values are probabilities (i.e. $\in [0,1]$) the resulting model matches the given probabilities exactly. Done.

(b) Otherwise, set $P(P_1 = 1, P_2 = 1, P_3 = 1) = d$. Do a constrained optimization of the backward probabilities over the subspace $(a, b, c, d) \in [0, 1]^4$ under a chosen distance measure $\mathcal{D}$, to approximate as closely as possible the given backward probabilities.

Performance of this model has to be tested against its alternatives, a) the simple proxy model without explicit modeling of the backward probabilities, and b) the proxy model optimized via constrained optimization as shown in 2.1, generalized to n problems, to avoid overfitting the network to the prior knowledge. Incorporating mechanisms for learning from the incoming data provides both a check on the dependence of the quality of the prior knowledge, and a desirable adaptive component.

3.2 Procedure and Details

We now elaborate on the steps involved in constructing the SFOBE Model according to section 3.1. Proofs of assertions are found in the appendix.

Step 1

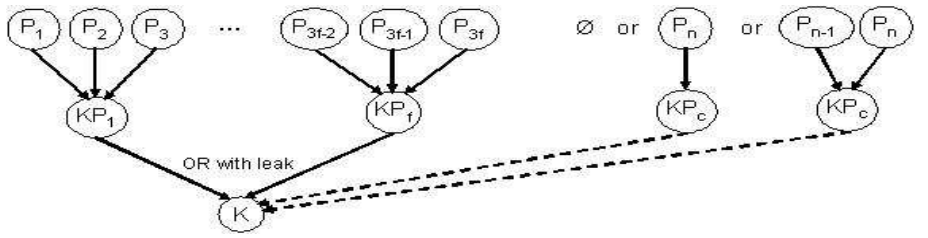
Let there be n binary problems identified as relevant to the considered symptom by the domain expert. According to the proposed model they are lumped into three-clusters as defined in Fig.3, which are then merged via an OR-with-leak connection into the observable domain symptom. Dividing n by 3 we can write $n = 3f + r$, with the rest $r \in \{0, 1, 2\}$. Depending on r, the division into three-clusters will either come out evenly, or there will be one or two left over problem nodes, as shown in Fig.4. There $f = \lfloor n/3 \rfloor$ and $c = \lceil n/3 \rceil$.[1] If n is not divisible by 3 the appendices E, F gives guidelines how to select the surplus problems. In choosing which problems to lump together the expert should be consulted whether there is any logical aggregation of problems into subdomains (such as for the symptom fever, "fever due to respiratory tract infections" and "fever due to abdominal problems"). If not, our current approach is to group problems arbitrarily.

Step 2

For $i = 1, \ldots, n$ let problem P_i be linked to the three-cluster node KP_k, with $k = \lceil i/3 \rceil$. We summarize the $3n + 2$ known probabilities:

$$\begin{array}{lll} \textbf{Given:} & P(P_i), & i = 1, \ldots, n \\ & p_i = P(K|P_i, \overline{P_j}, j \neq i), & i = 1, \ldots, n \\ & l \ = P(K|\overline{P_1}, ..., \overline{P_n}), & \\ & P(K), & \\ & P(P_i|K), & i = 1, ..., n. \end{array} \tag{5}$$

[1] Floor function $\lfloor x \rfloor$=largest integer not larger than x, ceiling function $\lceil x \rceil$=smallest integer not smaller than x.



$P^d(KP_c\|P_n)$	0	1	P_n
KP_c=0	1	1-$p_n^{'}$	
KP_c=1	0	$p_n^{'}$	

$P^d(KP_c\|P_{n-1},P_n)$	0		1		P_{n-1}
	0	1	0	1	P_n
$KP_c=0$	1	1-$p_n^{'}$	1-$p_{n-1}^{'}$	0	
$KP_c=1$	0	$p_n^{'}$	$p_{n-1}^{'}$	1	

$P^d(KP_k\|P_1,P_2,P_3)$	0				1				P_1
	0		1		0		1		P_2
	0	1	0	1	0	1	0	1	P_3
$KP_k=0$	1	1-$p_3^{'}$	1-$p_2^{'}$	1-a	1-$p_1^{'}$	1-b	1-c	1-d (or 0)	
$KP_k=1$	0	$p_3^{'}$	$p_2^{'}$	a	$p_1^{'}$	b	c	d (or 1)	

Fig. 4. SFOBE model. In the probability table of a three-cluster KP_k we may omit for simplicity the cluster index and call the three problems merging therein P_1, P_2, P_3.

The probabilities required to fill the minimum forward backward model are:

$$
\begin{array}{ll}
\textbf{Wanted:} & P(P_i), \qquad\qquad\qquad\qquad\qquad\quad i = 1,\ldots,n \\
& l^{'} = P(K|\overline{KP_k}, k = 1,\ldots,c) \\
& \text{For all clusters } KP_k: \qquad\qquad\quad k = 1,\ldots,f \\
& \quad p_1^{'} = P(KP_k|P_1,\overline{P_2},\overline{P_3}), \\
& \quad p_2^{'} = P(KP_k|P_2,\overline{P_1},\overline{P_3}), \\
& \quad p_3^{'} = P(KP_k|P_3,\overline{P_1},\overline{P_2}), \\
& \quad a \;\; = P(KP_k|\overline{P_1},P_2,P_3), \\
& \quad b \;\; = P(KP_k|\overline{P_2},P_1,P_3), \\
& \quad c \;\; = P(KP_k|\overline{P_3},P_1,P_2). \\
& \quad d \;\; = P(KP_k|P_1,P_2,P_3). \\
& \text{For } n \neq 3\lceil n/3\rceil \text{ need additionally: r=1: } p_n^{'},\ \text{r=2: } p_{n-1}^{'}, p_n^{'}.
\end{array}
\tag{6}
$$

Step 2a: The leaks in the two model are identical (for proof see appendix B).

$$\boxed{l^{'} = l} \tag{7}$$

Step 2b: Computation of the $p_i^{'}$, $i = 1,\ldots,n$ (for proof see appendix C):

$$\boxed{p_i^{'} = \frac{p_i - l}{1 - l}} \tag{8}$$

Step 3a: Exact backward step

As outlined we set $P(P_1 = 1, P_2 = 1, P_3 = 1) = 1$. In the probability tables of each three-cluster there are 3 remaining parameters, a, b, c. For each cluster independently, $j = 1, \ldots, f$, the values of a_j, b_j, c_j that effect exact matching of the backward probabilities are computed. The results may or may not be probabilities, as remains to be seen. We omit for simplicity the index j, and call the three problems merging into cluster KP: P_1, P_2, P_3. Let

$$\mathbf{x} = \begin{pmatrix} a \\ b \\ c \end{pmatrix} \tag{9}$$

Then $\mathbf{x}$ is the solution to the vector equation:

$$\boxed{(\mathbf{D_1}\mathbf{M_1} - \mathbf{D_2}\mathbf{M_2})\,\mathbf{x} = \mathbf{g} - \mathbf{D_1}\mathbf{v_1} + \mathbf{D_2}\mathbf{v_2}} \tag{10}$$

where the matrices $\mathbf{D_1}, \mathbf{D_2}, \mathbf{M_1}, \mathbf{M_2}$ and vectors $\mathbf{d}, \mathbf{v_1}, \mathbf{v_2}$ are given by:

$$\mathbf{D_1} = \begin{pmatrix} 1 - P(K|\overline{P_1}) & 0 & 0 \\ 0 & 1 - P(K|\overline{P_2}) & 0 \\ 0 & 0 & 1 - P(K|\overline{P_3}) \end{pmatrix} \tag{11}$$

$$\mathbf{D_2} = \begin{pmatrix} 1 - P(K|P_1) & 0 & 0 \\ 0 & 1 - P(K|P_2) & 0 \\ 0 & 0 & 1 - P(K|P_3) \end{pmatrix} \tag{12}$$

$$\mathbf{M_1} = \begin{pmatrix} 0 & P(\overline{P_2})P(P_3) & P(P_2)P(\overline{P_3}) \\ P(\overline{P_1})P(P_3) & 0 & P(P_1)P(\overline{P_3}) \\ P(\overline{P_1})P(P_2) & P(P_1)P(\overline{P_2}) & 0 \end{pmatrix} \tag{13}$$

$$\mathbf{M_2} = \begin{pmatrix} P(P_2)P(P_3) & 0 & 0 \\ 0 & P(P_1)P(P_3) & 0 \\ 0 & 0 & P(P_1)P(P_2) \end{pmatrix} \tag{14}$$

$$\mathbf{g} = \begin{pmatrix} P(K|P_1) - P(K|\overline{P_1}) \\ P(K|P_2) - P(K|\overline{P_2}) \\ P(K|P_3) - P(K|\overline{P_3}) \end{pmatrix} \tag{15}$$

$$\mathbf{v_1} = \begin{pmatrix} p_1^{'}P(\overline{P_2})P(\overline{P_3}) + P(P_2)P(P_3) \\ p_2^{'}P(\overline{P_1})P(\overline{P_3}) + P(P_1)P(P_3) \\ p_3^{'}P(\overline{P_1})P(\overline{P_2}) + P(P_1)P(P_2) \end{pmatrix} \tag{16}$$

$$\mathbf{v_2} = \begin{pmatrix} p_2^{'}P(P_2)P(\overline{P_3}) + p_3^{'}P(P_3)P(\overline{P_2}) \\ p_1^{'}P(P_1)P(\overline{P_3}) + p_3^{'}P(P_3)P(\overline{P_1}) \\ p_1^{'}P(P_1)P(\overline{P_2}) + p_2^{'}P(P_2)P(\overline{P_1}). \end{pmatrix} \tag{17}$$

All the probabilities on the right hand sides are given or easily derived: $P(\overline{P_i}) =$

$1 - P(P_i)$, from Bayes' formula,

$$P(K|P_i) = P(P_i|K)\frac{P(K)}{P(P_i)} \tag{18}$$

and, again with Bayes' formula,

$$P(K|\overline{P_i}) = P(\overline{P_i}|K)\frac{P(K)}{P(\overline{P_i})} = (1 - P(P_i|K))\frac{P(K)}{1 - P(P_i)}. \tag{19}$$

The proof is given in appendix D. The matrices $\mathbf{M_1}, \mathbf{M_2}$ and the vectors $\mathbf{v_1}, \mathbf{v_2}$ have a circular symmetry that ensures that they obey the requirement that all the rows of $(\mathbf{M_1 x} + \mathbf{v_1})P(P_i) + (\mathbf{M_2 x} + \mathbf{v_2})P(\overline{P_i})$ are equal, for $i = 1, 2, 3$. This results from the fact that, using (32), (33) for the last equality,

$$\begin{aligned} P(KP) &= P(KP, P_i) + P(KP, \overline{P_i}) = P(KP|P_i)P(P_i) + P(KP|\overline{P_i})P(\overline{P_i}) \\ &= \mathbf{e_i}^T(\mathbf{M_1 x} + \mathbf{v_1})P(P_i) + \mathbf{e_i}^T(\mathbf{M_2 x} + \mathbf{v_2})P(\overline{P_i}) \end{aligned} \tag{20}$$

independently of which unit vector $\mathbf{e_i}, i \in 1, 2, 3$ is chosen. The matrices $\mathbf{M_1}, \mathbf{M_2}$ and the vectors $\mathbf{v_1}, \mathbf{v_2}$ indeed fulfill this requirement, as can be easily verified.

So we have to solve, separately for each cluster, an inhomogeneous system of three equations with three unknowns, where all coefficients are known. The coefficient matrix $\mathbf{D_1M_1} - \mathbf{D_2M_2}$ is:

$$\mathbf{D_1M_1} - \mathbf{D_2M_2} = \tag{21}$$
$$\begin{pmatrix} -P(\overline{K}|P_1)P(P_2)P(P_3) & P(\overline{K}|\overline{P_1})P(\overline{P_2})P(P_3) & P(\overline{K}|\overline{P_1})P(P_2)P(\overline{P_3}) \\ P(\overline{K}|\overline{P_2})P(\overline{P_1})P(P_3) & -P(\overline{K}|P_2)P(P_1)P(P_3) & P(\overline{K}|\overline{P_2})P(P_1)P(\overline{P_3}) \\ P(\overline{K}|\overline{P_3})P(\overline{P_1})P(P_2) & P(\overline{K}|\overline{P_3})P(P_1)P(\overline{P_2}) & -P(\overline{K}|P_3)P(P_1)P(P_2) \end{pmatrix}$$

and its determinant equals:

$$\begin{aligned} &\det(\mathbf{D_1M_1} - \mathbf{D_2M_2}) = \\ &\quad = P(P_1)P(P_2)P(P_3)(1 - P(K))^2 \left(2P(\overline{K}) - \sum_{i=1}^{3} P(\overline{K}, P_i)\right) \end{aligned} \tag{22}$$

where each of the terms in the last parenthesis is computed as:

$$P(\overline{K}, P_i)) = P(P_i) - P(P_i|K)P(K). \tag{23}$$

If the determinant is nonzero, then the system matrix is full rank and the system has a unique solution. Looking at the r.h.s. of (22), the first four terms are

clearly nonzero, since a cause that is never true would not be modeled and a symptom that is always present isn't much of a symptom. Due to noise in the estimates there is a zero likelihood that the last term is zero. However, due to the necessity to obtain probabilities as solutions to (10), a very small determinant is undesirable as well. We know that $\forall i$, $P(\overline{K}, P_i) < P(\overline{K})$, in fact $P(\overline{K}, P_i) < \min(P(\overline{K}), P(P_i))$. So we expect $P(\overline{K}, P_i)$ to be much smaller than $P(\overline{K})$, resulting in a clearly positive determinant.

The problem is that in many cases the entries of the resulting solution vector $\mathbf{x}$ will not be probabilities. In that case one can do a constrained optimization over $[0,1]^3$, as outlined in (1).

Step 3b: Constrained optimization backward step

The constrained optimization is done as outlined in (1) to (3). The approach we have used in the numerical examples of the next section is to take the given forward estimates for granted and approximate the given backward estimates as closely as possible with the model. The general formulas for $P(K)$ and $P(P_i|K)$ are

$$P(\overline{K}) = 1 - P(K) = (1-l) \cdot \prod_{clusters} (1 - P(KP)) \tag{24}$$

and

$$P(P_i|K) = \frac{P(P_i)}{P(K)} \cdot \left(1 - (1 - P(K)) \frac{1 - P(KP|P_i)}{1 - P(KP)}\right), \tag{25}$$

where problem P_i belongs to cluster KP . For Noisy-OR $P(K)$ and $P(P_i|K)$ are completely determined by the forward estimates as:

Noisy-OR proxy:

$$\begin{aligned} P(K) &= 1 - (1-l) \cdot \prod_{i=1}^{n} \left(1 - p_i^{'} \cdot P(P_i)\right) \\ P(P_i|K) &= \frac{P(P_i)}{P(K)} \cdot \left(1 - P(\overline{K}) \frac{1 - p_i^{'}}{1 - p_i^{'} \cdot P(P_i)}\right) \end{aligned} \tag{26}$$

where the $p_i^{'}$ are as given in (8). This can be seen e.g. from the probability table in Fig.4 for the single left-over problem, which has a Noisy-OR connection. For three-clusters on the other hand the $P(P_i|K)$, for $P_i, i \in 1,2,3$ belonging to

cluster $KP_k, k \in 1, ..., \lceil n/3 \rceil$, and $P(K)$ are:

three-cluster:

$$
\begin{aligned}
P(K) &= 1 - (1-l) \cdot \prod_{k=1}^{\lceil n/3 \rceil} (1 - P(KP_k)) \\
k &\in 1, \ldots, \lfloor n/3 \rfloor : \\
P(KP_k) &= \mathbf{e_i}^T \left((\mathbf{M_1 x} + \mathbf{v_1}) P(P_i) + (\mathbf{M_2 x} + \mathbf{v_2}) P(\overline{P_i}) \right) \\
P(P_i|K) &= \tfrac{P(P_i)}{P(K)} \cdot \left(1 - P(\overline{K}) \tfrac{1 - \mathbf{e_i}^T (\mathbf{M_1 x} + \mathbf{v_1})}{1 - P(KP_k)} \right) \\
r &= 1 : \\
P(KP_{\lceil n/3 \rceil}) &= p_n^{'} \cdot P(P_n) \\
P(P_n|K) &= \tfrac{P(P_n)}{P(K)} \cdot \left(1 - P(\overline{K}) \tfrac{1 - p_n^{'}}{1 - p_n^{'} \cdot P(P_n)} \right) \\
r &= 2 : \\
P(KP_{\lceil n/3 \rceil}) &= p_n^{'} \cdot P(P_n) P(\overline{P_{n-1}}) + p_{n-1}^{'} \cdot P(\overline{P_n}) P(P_{n-1}) + \\
&\quad + P(P_n) P(P_{n-1}) \\
P(P_n|K) &= \tfrac{P(P_n)}{P(K)} \cdot \left(1 - P(\overline{K}) \tfrac{1 - p_n^{'} \cdot P(\overline{P_{n-1}}) - P(P_n) P(P_{n-1})}{1 - P(KP_{\lceil n/3 \rceil})} \right) \\
P(P_{n-1}|K) &= \tfrac{P(P_{n-1})}{P(K)} \cdot \left(1 - P(\overline{K}) \tfrac{1 - p_{n-1}^{'} \cdot P(\overline{P_n}) - P(P_n) P(P_{n-1})}{1 - P(KP_{\lceil n/3 \rceil})} \right)
\end{aligned}
\tag{27}
$$

where the $P(KP_k)$ were taken from (20), $\mathbf{e_i}, i \in 1, 2, 3$, denotes the unit vector and $\mathbf{x}$ is the vector of parameters a, b, c that we can vary within the unit cube in order to achieve a better approximation of the backward probabilities.

Setting $P(P_1 = 1, P_2 = 1, P_3 = 1) = d$: The above optimization can be enhanced by the one additional degree of freedom allowed by the 6p-model, namely by removing the fixed choice $P(P_1 = 1, P_2 = 1, P_3 = 1) = 1$, that was made in 3.1.3a, and allowing it to equal $d \in [0, 1]$. This way we gain the ability to match the backward probabilities more closely. The changes that need to be made before using (27) are to replace $\mathbf{M_1}$, $\mathbf{M_2}$, $\mathbf{v_1}$ and $\mathbf{x}$ by:

$$
\mathbf{M_1^d} = \begin{pmatrix} & P(P_2)P(P_3) \\ \mathbf{M_1} & P(P_1)P(P_3) \\ & P(P_1)P(P_2) \end{pmatrix} \qquad \mathbf{M_2^d} = \begin{pmatrix} & 0 \\ \mathbf{M_2} & 0 \\ & 0 \end{pmatrix} \tag{28}
$$

$$
\mathbf{v_1^d} = \begin{pmatrix} p_1^{'} P(\overline{P_2}) P(\overline{P_3}) \\ p_2^{'} P(\overline{P_1}) P(\overline{P_3}) \\ p_3^{'} P(\overline{P_1}) P(\overline{P_2}) \end{pmatrix} \qquad \mathbf{v_2^d} = \mathbf{v_2} \qquad \mathbf{x^d} = \begin{pmatrix} a \\ b \\ c \\ d \end{pmatrix} \tag{29}
$$

4 Numerical Example

In order to illustrate the approach two series of experiments were performed. For both experiments a Bayesian network with 4 problems and 2 symptoms was considered. In the first series of experiments, all the probabilities specified by the expert were randomly and independently drawn from a uniform distribution over $[0, 1]$, including the target (i.e. expert specified) probabilities in backward directions, $P(K_j)$ and $P(P_i|K_j)$. Although these probabilities are in reality highly dependent on each other, for illustration purposes we have generated them as independent random variables. The Noisy-OR proxy network has no freedom for further optimization once the forward probabilities are specified. On the other hand, the SFOBE network has some freedom to accommodate the specified probabilities in the backward direction to some degree. For each model the mismatch between the probabilities in the model and the target probabilities given by the expert is measured by different reasonable distance measures. We have calculated the Kullback-Leibler distance, the absolute value of the difference, and the square root of minimum mean square error. The statistics of the calculated distances after one thousand experiments is given in Tab. 1.

Model	Mean(KL)	Std(KL)	Mean(abs)	Std(abs)	Mean(sqrt)	Std(sqrt)
Noisy-OR proxy	0.37	0.18	0.48	0.11	0.18	0.03
SFOBE	0.27	0.16	0.41	0.11	0.16	0.03

Table 1. Distance of noisy-OR proxy model respectively SFOBE model to target for experiment 1, measured via Kullback-Leibler distance (KL), absolute value of difference (abs) and square root of minimum mean square error (sqrt).

As expected, the extra freedom of the SFOBE network has resulted in a smaller average error and a smaller or equal standard deviation of the error. Although the first series of experiments is equally fair to both models, it is not realistic, because the backward probabilities were generated as independent random variables, although they are not. For this reason the distances to the target probabilities are quite large in absolute value. A more realistic scenario is chosen for the second series of experiments. There all the probabilities in the *forward* direction necessary to fill the complete probability table of the network in Fig.5 were randomly and independently drawn from a uniform distribution over $[0, 1]$. The backward probabilities were then read off the network. Then the performance of the two models: Noisy-OR-proxy model of Fig.6 and SFOBE-model of Fig.7 were compared in the same way as for the first series of experiments. For each model the mismatch between the probabilities in the model and the target probabilities given by the expert was measured by different reasonable distance measures. We have calculated the Kullback-Leibler distance, the absolute value

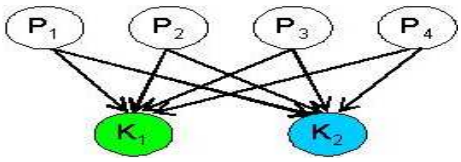


Fig. 5. Fully connected Bayesian network of domain with four root-causes and two symptoms. All variables are binary random variables with the states 1=yes and 0=no.

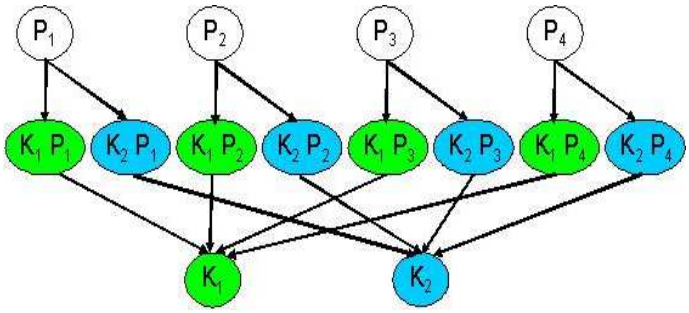


Fig. 6. Noisy-OR proxy model of Bayesian network with four root-causes and two symptoms. All variables are binary random variables with the states 1=yes and 0=no.

of the difference, and the square root of minimum mean square error. The statistics of the calculated distances after three thousand experiments is given in Tab. 2. Again the SFOBE model clearly outperforms the Noisy-OR proxy model in

Model	Mean(KL)	Std(KL)	Mean(abs)	Std(abs)	Mean(sqrt)	Std(sqrt)
Noisy-OR proxy	0.040	0.063	0.139	0.029	0.065	0.015
SFOBE	0.016	0.01	0.078	0.029	0.035	0.013

Table 2. Distance of noisy-OR proxy model respectively SFOBE model to target for experiment 2, measured via Kullback-Leibler distance (KL), absolute value of difference (abs) and square root of minimum mean square error (sqrt).

its ability to approximate that set of probabilities that the expert can estimate with confidence.

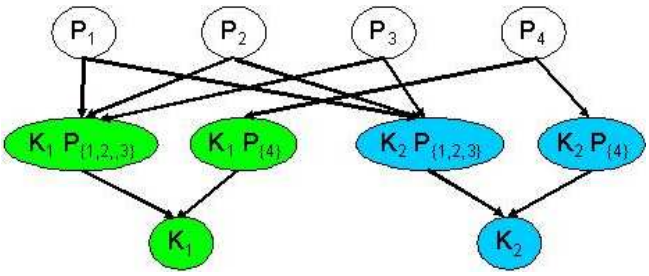


Fig. 7. SFOBE model of Bayesian network with four root-causes and two symptoms. All variables are binary random variables with the states 1=yes and 0=no.

5 Conclusion

This paper discussed the derivation of Bayesian network based domain models used for diagnostic purposes. These models are derived from the available human expert knowledge by interviewing domain experts. It was shown that a typical problem in expert interviewing is the presence of partial conditional information in both directions: from root-causes to symptoms and vice-versa, from symptoms to root-causes. This is because of the way the expert thinks about the domain when he is doing problem diagnosis. This paper has presented a rigorous mathematical analysis of the expert knowledge based design of Bayesian networks describing the dependencies between two groups of discrete binary random variables (root causes and symptoms) where the prior information characterizes both the forward and backward reasoning. A suitable parametric model was introduced which can accommodate information in both directions. The validity of the probability estimation algorithm is shown on a suitable example with four root-causes and two symbols. The obtained results might lead to the refinement of the information provided by the expert, which could render an improved model. The discussed examples have been binary, however the applications in [1], [2] and [3] deal with mixtures of binary and multivariate variables and the considerations presented here are amenable to multivariate extensions.

References

1. D. Obradovic, R. Lupas Scheiterer, *Troubleshooting in GSM Mobile Telecommunication Networks based on Domain Model and Sensory Information*, ICANN 2005.
2. J. Horn, T. Birkhölzer, O. Hogl, M. Pellegrino, R. Lupas Scheiterer, K.-U. Schmidt, V. Tresp, *Knowledge Acquisition and Automated Generation of Bayesian Networks*, Proc. AIME '01, Cascais, Portugal, July 2001, pp. 35-39.
3. J. Horn, T. Birkhölzer, O. Hogl, M. Pellegrino, R. Lupas Scheiterer, K.-U. Schmidt, V. Tresp, *Knowledge Acquisition and Automated Generation of Bayesian Networks for a Medical Dialogue and Advisory System*, S. Quaglini, P. Barahona, S. Andreassen, Eds, Artificial Intelligence in Medicine, Springer-Verlag, 2001, pp. 199-202.

4. F. V. Jensen, *An Introduction to Bayesian Networks*, UCL Press, 1996.
5. K. Murphy, *Software Packages for Graphical Models / Bayesian Networks*, last updated 31 October 2005 (status at paper submission), http://www.cs.ubc.ca/ murphyk/Bayes/bnsoft.html
6. galel@cs.huji.ac.il, *Bayesian Network Repository*, last modified: March 01, 2001 (status at paper submission), http://www.cs.huji.ac.il/labs/compbio/Repository/
7. J. Pearl, *Probabilistic Reasoning in Intelligent Systems*, Morgan-Kaufmann, 1988.
8. D. Heckerman, *A Tutorial on Learning With Bayesian Networks*, Technical Report, http://research.microsoft.com/h̃eckerman, March 1995.
9. J. Kim, J. Pearl, *A computational model for causal and diagnostic reasoning in inference engines*, Proceedings Eighth International Joint Conference on Artificial Intelligence, Karlsruhe, West Germany, pp. 190-193.
10. M. Henrion, *Some practical issues in constructing belief networks*, Proceedings of the Third Workshop on Uncertainty in Artificial Intelligence, Seattle, WA, pp. 132-139, Association for Uncertainty in Artificial Intelligence, Mountain View, CA. Also in Kanal, L., Levitt, T., and Lemmer, J., editors, Uncertainty in Artificial Intelligence 3, pp. 161-174. North-Holland, New York, 1989.
11. D. Heckerman, *Causal independence for knowledge acquisition and inference*, Proceedings of the 9th Conference on Uncertainty in Artificial Intelligence, Washington D.C., Morgan Kaufmann, San Mateo, Calif., 1993, pp. 122-127.
12. R. Lupas Scheiterer, *HealthMan Bayesian Network Description: Disease to Symptom Layers, Multi-Valued Syptoms, Multi-valued Diseases*, Siemens AG Internal Report, October 1999.
13. R. Lupas Scheiterer, *Bayesian Network Modeling Aspects Resulting from the HealthMan and GSM Troubleshooting Applications*, Siemens AG Internal Report, April 2003.

Appendix

A. Short Review of Bayesian Networks

The main ingredients of a causal network are, [4]:

1. Nodes, each with a finite state of mutually exclusive states, interconnected by directed arrows, to form a directed acyclic graph;
2. Prior probability distributions on all root nodes (i.e. nodes with no parents);
3. Conditional probability tables for each child node given its parents;
4. In some domains the directed links have a causal interpretation. Independently of a possible causal interpretation the property of conditional independence (d-separation in graphical model literature) holds, which says that:
   (a) For serial connections, A and C are conditionally independent given B (Fig. 8).
   (b) For diverging connections, B,C,...,N are conditionally independent given A (Fig. 9).

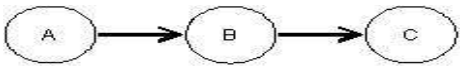


Fig. 8. Serial connection in causal model.

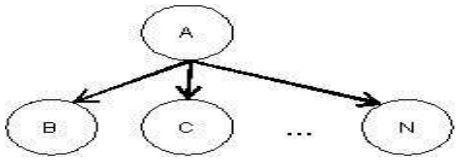


Fig. 9. Diverging connection in causal model.

(c) For converging connections, B,C,...,N, are conditionally independent if neither A nor one of its descendants have received evidence (Fig. 10). If however there is evidence present for A or one of its descendants, then its a-priori independent common ancestors become dependent. To see this, consider a binary domain where an effect can be explained by two different causes. Then if the effect is present our reasoning is that an increase in likelihood of one of these causes effects a decrease in likelihood of the other, a see-saw effect also called "explaining away".

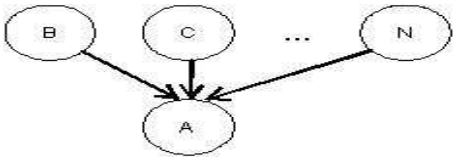


Fig. 10. Converging connection in causal model.

If possible, the dependence/independence relations asserted by the connections in the model should be validated by actual data.

With these ingredients, the important quantities in the domain can be modeled together with their dependence relations and with probability tables quantifying these dependencies. When evidence on the state of variables is received, the reasoning is performed by updating the various probabilities. There exists a large body of software packages for Bayesian networks, see e.g. [5]. A collection of different application domains and data sets can be found in [6]. The name "Bayesian network" comes from Bayes' formula for conditional probability, which is fundamental for the propagation of the evidence.

In summary, causal networks provide a rigorous and efficient framework for inference, i.e. for calculating the probability of non-observable variables given a set of observations of related observable variables. For further details we refer

to Pearl, e.g. [7], who is one of the founders of the field. A detailed theoretical discussion is found in [8] and the ten pages of references therein.

B. Proof of Step 2a:

$$\begin{aligned} l = P(K|\overline{P_1},\ldots,\overline{P_n}) &= P(K,\overline{KP_1},\ldots,\overline{KP_c}|\overline{P_1},\ldots,\overline{P_2}) \\ &= P(K|\overline{KP_1},\ldots,\overline{KP_c},\overline{P_1},\ldots,\overline{P_2})\cdot P(\overline{KP_1},\ldots,\overline{KP_c}|\overline{P_1},\ldots,\overline{P_2}) \\ &= P(K|\overline{KP_1},\ldots,\overline{KP_c})\cdot 1 = l^{'} \end{aligned} \tag{30}$$

We have used the independence properties of serial causal links and the fact that the KP_k are independent, given no evidence on K. The fact that the leaks are the same in the direct and the proxy model makes sense, because the leak node is a direct ancestor of the indicator node K, separate from the problem nodes, hence is not affected by the introduction of the proxies. □

C. Proof of Step 2b:

Let problem P_i belong to the cluster $KP_{C(i)}$, where we have chosen as a convenient cluster index $C(i)$ the set of indices of the problems connected to that cluster. For the n forward values specified by the expert of the form "probability that the symptom is present given that exactly one problem is present", the following equations hold:

$$\begin{aligned} p_i &= P(K|P_i,\overline{P_j},j\neq i) = \\ &= P(K,KP_{C(i)},\overline{KP_k},k\neq C(i)|P_i,\overline{P_j},j\neq i) + \\ &+ P(K,\overline{KP_{C(i)}},\overline{KP_k},k\neq C(i)|P_i,\overline{P_j},j\neq i) = \\ &= P(K|KP_{C(i)},\overline{KP_k},k\neq C(i),P_i,\overline{P_j},j\neq i)\cdot \\ &\qquad \cdot P(KP_{C(i)},\overline{KP_k},k\neq C(i)|P_i,\overline{P_j},j\neq i) + \\ &+ P(K|\overline{KP_{C(i)}},\overline{KP_k},k\neq C(i),P_i,\overline{P_j},j\neq i)\cdot \\ &\qquad \cdot P(\overline{KP_{C(i)}},\overline{KP_k},k\neq C(i)|P_i,\overline{P_j},j\neq i) = \\ &= P(K|KP_{C(i)},\overline{KP_k},k\neq C(i))\cdot P(KP_{C(i)}|P_i,\overline{P_j},j\neq i)\cdot \\ &\qquad \cdot P(\overline{KP_k},k\neq C(i)|P_i,\overline{P_j},j\neq i) + \\ &+ P(K|\overline{KP_{C(i)}},\overline{KP_k},k\neq I(i))\cdot P(\overline{KP_{C(i)}}|P_i,\overline{P_j},j\neq i)\cdot \\ &\qquad \cdot P(\overline{KP_k},k\neq C(i)|P_i,\overline{P_j},j\neq i) = \\ &= 1\cdot p_i^{'}\cdot 1 + l\cdot(1-p_i^{'})\cdot 1 = l + p_i^{'}\cdot(1-l) \end{aligned} \tag{31}$$

where the second equality holds since the event "$\overline{P_j},\forall j\neq i$" implies "$\overline{KP_k},k\neq C(i)$"; to obtain the one but last equality we have used for the 1st multiplicative term property 4a of Bayesian networks and for the 2nd multiplicative term property 4c; and for the last equality we have used the fact that $P(KP_{C(i)}|P_i,\overline{P_j},j\neq i) = P(KP_{C(i)}|P_i,\overline{P_j},j\in C(i),j\neq i)$ since $KP_{C(i)}$ and $\{P_j,j\notin C(i)\}$ are independent if no common descendent has received evidence.

Equation (8) follows. Remains to show that this assignment results in probabilities, i.e. the right hand side (r.h.s.) of (8) $\in (0,1)$. This is clearly true, as long as the leak is a small number not exceeding any probability p_i that the symptom is due to a single one of the modeled causes. Should the expert initially name a larger leak, contributors to this leak from outside the modeled domain need to be included in the model, say as L_1, and then the p_i are $P(K|P_i, \overline{P_j}, j \neq i, \overline{L_1})$, until the residue leak is indeed "negligible" compared to the modeled contributions. □

D. Proof of Step 3a:

From the large probability table in Fig.4 $P(KP|P_i)$ and $P(KP|\overline{P_i})$ are expressed in terms of the sought variables a, b, c as:

$$
\begin{aligned}
P(KP|P_1) &= p_1^{'}P(\overline{P_2})P(\overline{P_3}) + bP(\overline{P_2})P(P_3) + cP(P_2)P(\overline{P_3}) + P(P_2)P(P_3)\\
P(KP|P_2) &= p_2^{'}P(\overline{P_1})P(\overline{P_3}) + aP(\overline{P_1})P(P_3) + cP(P_1)P(\overline{P_3}) + P(P_1)P(P_3)\\
P(KP|P_3) &= p_3^{'}P(\overline{P_1})P(\overline{P_2}) + aP(\overline{P_1})P(P_2) + bP(P_1)P(\overline{P_2}) + P(P_1)P(P_2)\\
P(KP|\overline{P_1}) &= p_3^{'}P(P_3)P(\overline{P_2}) + p_2^{'}P(\overline{P_3})P(P_2) + aP(P_2)P(P_3)\\
P(KP|\overline{P_2}) &= p_1^{'}P(P_1)P(\overline{P_3}) + p_3^{'}P(\overline{P_1})P(P_3) + bP(P_1)P(P_3)\\
P(KP|\overline{P_3}) &= p_1^{'}P(P_1)P(\overline{P_2}) + p_2^{'}P(\overline{P_1})P(P_2) + cP(P_1)P(P_2)
\end{aligned}
$$

Or, compactly,

$$
\begin{pmatrix} P(KP|P_1)\\ P(KP|P_2)\\ P(KP|P_3)\end{pmatrix} = \mathbf{M_1}\cdot\begin{pmatrix} a\\ b\\ c\end{pmatrix} + \mathbf{v_1} \tag{32}
$$

and

$$
\begin{pmatrix} P(KP|\overline{P_1})\\ P(KP|\overline{P_2})\\ P(KP|\overline{P_3})\end{pmatrix} = \mathbf{M_2}\cdot\begin{pmatrix} a\\ b\\ c\end{pmatrix} + \mathbf{v_2}. \tag{33}
$$

with $\mathbf{M_1}, \mathbf{M_2}, \mathbf{v_1}, \mathbf{v_2}$ as given in (13), (14), (16), (17).

We need an equation containing $P(KP|P_i)$ and $P(KP|\overline{P_i})$ as the only unknowns. Clearly:

$$
\begin{aligned}
P(K|P_i) &= P(K, KP|P_i) + P(K, \overline{KP}|P_i) \qquad (34)\\
&= P(K|KP, P_i)\cdot P(KP|P_i) + P(K|\overline{KP}, P_i)\cdot P(\overline{KP}|P_i)\\
&= P(K|KP)\cdot P(KP|P_i) + P(K|\overline{KP})\cdot P(\overline{KP}|P_i)\\
&= 1\cdot P(KP|P_i) + (1 - P(KP|P_i))\cdot P(K|\overline{KP})
\end{aligned}
$$

and analogously

$$
P(K|\overline{P_i}) = P(KP|\overline{P_i}) + (1 - P(KP|\overline{P_i}))\cdot P(K|\overline{KP}) \tag{35}
$$

Therefore, equation the last term equal of both equations

$$
\frac{P(K|P_i) - P(KP|P_i)}{1 - P(KP|P_i)} = \frac{P(K|\overline{P_i}) - P(KP|\overline{P_i})}{1 - P(KP|\overline{P_i})} \tag{36}
$$

from which follows, after multiplying cross-wise and gathering terms

$$\boxed{P(KP|P_i)\cdot(1-P(K|\overline{P_i}))-P(KP|\overline{P_i})(1-P(K|P_i))=P(K|P_i)-P(K|\overline{P_i})} \tag{37}$$

As desired, this equation contains $P(KP|P_i)$ and $P(KP|\overline{P_i})$ as the only unknowns. Equivalently:

$$\mathbf{D_1}\cdot\begin{pmatrix}P(KP|P_1)\\P(KP|P_2)\\P(KP|P_3)\end{pmatrix}-\mathbf{D_2}\cdot\begin{pmatrix}P(KP|\overline{P_1})\\P(KP|\overline{P_2})\\P(KP|\overline{P_3})\end{pmatrix}=\mathbf{g} \tag{38}$$

where $\mathbf{D_1}, \mathbf{D_2}, \mathbf{g}$ are as given in (11), (12), (15). Inserting (32) and (33) into (38) results in (10). □

E. Case $r = 1$ (one surplus problem):

Equation (37) has to be obeyed also by a single surplus problem P_n left over after lumping groups of 3 problems together. From its probability table in Fig.4 we read that $P(KP|P_n) = p_n^{'}$ and $P(KP|\overline{P_n}) = 0$, hence:

$$p_n^{'}\cdot(1-P(K|\overline{P_n}))=P(K|P_n)-P(K|\overline{P_n}) \tag{39}$$

or, with (8),

$$\frac{p_n-l}{1-l}=\frac{P(K|P_n)-P(K|\overline{P_n})}{1-P(K|\overline{P_n})}=\frac{P(K|P_n)-P(K)}{1-P(P_n)-P(K)+P(P_n|K)P(K)} \tag{40}$$

where the last equality was obtained by using (19). Note that $p_n \neq P(K|P_n)$, but rather $P(K|P_n, \overline{P_j}, j \neq n)$. So, after using (18), we get:

$$p_n=l+(1-l)\cdot\frac{P(K)}{P(P_n)}\cdot\frac{P(P_n|K)-P(P_n)}{1-P(P_n)-P(K)+P(P_n|K)P(K)} \tag{41}$$

In the above equation all the quantities are given, being estimates provided by the expert. Therefore (41) is a consistency condition that has to be fulfilled by the given estimates for the problem that is chosen to be the surplus problem after clustering in groups of three. We have on purpose isolated on the l.h.s the estimate the expert feels least confident about, in our case the p_n, i.e. the probability that the symptom is present given that a single one of the possible causes is present. Then arguably the r.h.s. may constitute a better estimate. A reasonable way to achieve consistency is to compute for all the problems

$$\delta_i=|l+(1-l)\cdot\frac{P(K)}{P(P_i)}\cdot\frac{P(P_i|K)-P(P_i)}{1-P(P_i)-P(K)+P(P_i|K)P(K)}-p_i|, \tag{42}$$

then proceeding from the problem with smallest δ in ascending order ask the expert if he is comfortable with the replacement of his estimate of p_i by the r.h.s. of (41), then choose the first problem where he agrees as the surplus one.

Equivalently one could from the start ask the expert to provide a confidence indicator for each of his estimates - he will feel less confidant about cases he rarely sees - and replace the first value that has confidence under a threshold. Note that the consistency condition given here has to be obeyed by each problem in a Noisy-OR model.

F. Case $r = 2$ (two surplus problems):

Likewise, equation (37) has to be obeyed also by each of two surplus problems P_{n-1}, P_n left over after lumping groups of 3 problems together. From their probability tables in Fig.4 we read that $P(KP|P_n) = p_n^{'} \cdot P(\overline{P_{n-1}}) + P(P_{n-1})$ and $P(KP|\overline{P_n}) = p_{n-1}^{'} \cdot P(P_{n-1})$, hence:

$$\begin{pmatrix} P(\overline{P_{n-1}})(1 - P(K|\overline{P_n})) & -P(P_{n-1})(1 - P(K|P_n)) \\ -P(P_n)(1 - P(K|P_{n-1})) & P(\overline{P_n})(1 - P(K|\overline{P_{n-1}})) \end{pmatrix} \cdot \begin{pmatrix} p_n^{'} \\ p_{n-1}^{'} \end{pmatrix} = \tag{43}$$
$$= \begin{pmatrix} P(K|P_n) - P(P_{n-1}) - P(K|\overline{P_n})P(\overline{P_{n-1}})) \\ P(K|P_{n-1}) - P(P_n) - P(K|\overline{P_{n-1}})P(\overline{P_n})) \end{pmatrix}$$

The determinant of the square matrix can be shown to be

$$\begin{aligned} \det = P(\overline{K})[\ 1\ -P(K) - P(P_n) - P(P_{n-1}) + \\ + P(K)P(P_n|K) + P(K)P(P_{n-1}|K)] = \\ = P(\overline{K})(\ P\ (\overline{K}) - P(\overline{K}, P_n) - P(\overline{K}, P_{n-1})) \end{aligned} \tag{44}$$

hence is almost surely nonzero, and with high likelihood positive, if we expect the presence of a cause to more than halve the probability of absence of a relevant symptom. One could choose as the two surplus problems those whose p_n, p_{n-1} corresponding via (8) to the $p_n^{'}, p_{n-1}^{'}$ obtained from above are closest to the estimates thereof given by the expert. Note that these consistency conditions would have to be obeyed by each problem pair if we chose a model where we cluster problems together in pairs of two, instead of three, with the probability table as shown in Fig.4 for two problems.