Knowledge Engineering for Bayesian Networks: How Common Are Noisy-MAX Distributions in Practice?

Adam Zagorecki [1] and **Marek Druzdzel** [2]

Abstract. One problem faced in knowledge engineering for Bayesian networks is the exponential growth of the number of parameters in their conditional probability tables (CPTs). The most common practical solution is application of the noisy-OR (or their generalization, the noisy-MAX) gates, which take advantage of independence of causal interactions and provide a logarithmic reduction of the number of parameters required to specify a CPT. In this paper, we propose an algorithm that fits a noisy-MAX distribution to an existing CPT and we apply it to search for noisy-MAX gates in three existing practical Bayesian networks. We show that noisy-MAX gate provides a surprisingly good fit for as many as 50% of CPTs in these networks. The importance of this finding is that it provides an empirical justification for the use of the noisy-MAX gate as a powerful knowledge engineering tool.

1 INTRODUCTION

Bayesian networks [1] provide a convenient and sound framework for encoding uncertain knowledge and for reasoning with it. A Bayesian network (BN) is essentially an acyclic directed graph encoding a factorization of a joint probability distribution. The structure of the graph represents the variables and independencies among them, while the probability distributions over the individual variables conditional on their direct predecessors (parents) represent individual components of the factorization.

When a node of a BN and all its parents are discrete, the conditional probability distributions are stored in *conditional probability tables* (CPTs) indexed by all possible combinations of states of the parents. This poses considerable difficulties to knowledge engineering, to learning BNs from data, and to inference algorithms for BNs. An ingenious practical solution to this problem has been the application of parametric conditional distributions, such as the noisy-OR. By taking advantage of *independence of causal interaction* (ICI), these gates offer a reduction of the number of parameters required to specify a conditional probability distribution from exponential to linear in the number of parents.

The two most widely applied ICI distributions are the binary noisy-OR model [2] and its extension to multi-valued variables, the noisy-MAX model [3, 4]. Noisy-OR and noisy-MAX gates have proven their worth in many real-life applications (e.g., [5, 6, 7]). Their foremost advantage is a small number of parameters that are sufficient to specify the entire CPT. This leads to a significant reduction of effort in knowledge elicitation from experts [3, 5], improves the quality of distributions learned from data [8], and reduces the spacial and temporal complexity of algorithms for Bayesian networks [9, 10].

Our research aims at better understanding of the applicability of the noisy-MAX gates in practical BN models. We achieve this by developing a technique for fitting noisy-MAX relationships to existing, fully specified CPTs. Having this technique, we can examine CPTs in existing practical BN models for whether they can be approximated by the noisy-MAX model.

We analyze CPTs in three existing sizeable Bayesian network models: ALARM [11], HAILFINDER [12] and HEPAR II [8]. We show that the noisy-MAX gate provides a surprisingly good fit for a significant percentage of distributions in these networks. We observed this in both, distributions elicited from experts and those learned from data and for two measures of distance between distributions. We test the robustness of this result by fitting the noisy-MAX distribution to randomly generated CPTs and observe that the fit in this case is poor. Obtaining a randomly generated CPT that can be reasonably approximated by a noisy-MAX gate is extremely unlikely, which leads us to the conclusion that our results are not a coincidence. We investigate the influence of the conversion to the noisy-MAX on the precision of posterior probability distributions.

We envision two applications of this technique. The first is a possible refocusing of knowledge engineering effort from obtaining an exponential number of numerical probabilities to a much smaller number of noisy-MAX parameters. While a parametric distribution may be only an approximation to a set of general conditional probability distributions, the precision that goes with the latter is often only theoretical. In practice, obtaining large numbers numerical probabilities is likely to lead to expert exhaustion and result in poor quality estimates. Focusing the expert's effort on a small number of parameters of a corresponding parametric distribution should lead to a better quality model. The second application of our technique is in approximate algorithms for Bayesian networks. Whenever the fit is good, a CPT can be replaced by an ICI gate, leading to potentially significant savings in computation [9].

The remainder of this paper is organized as follows. Section 2 introduces the noisy-OR and the noisy-MAX gates. Section 3 proposes our algorithm for fitting noisy-MAX parameters to an arbitrary CPT. Section 4 presents the empirical results of the experiments testing the goodness of fit for several existing practical networks. In Section 5, we propose an explanation of the observed results.

2 NOISY-OR AND NOISY-MAX

We will represent random variables by upper-case letters (e.g., X) and their values by indexed lower-case letters, (e.g., x_1). We use

[1] Defence Academy of the UK, Cranfield University, UK, email: zagorecki@gmail.com
[2] University of Pittsburgh, School of Information Sciences, Decision Systems Laboratory, USA, email: marek@sis.pitt.edu

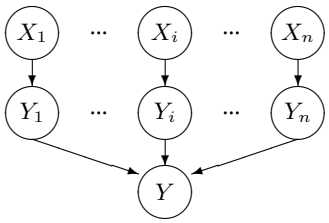

Figure 1. Direct modelling of noisy-OR

$Rng(X)$ to denote the range (the set of possible values) of a variable X. We assume that all variables are discrete. Let there be n binary nodes $X_1, \ldots, X_n$, each with values from $Rng(X_i) = \{x_i, \overline{x}_i\}$. Let the variables X_i be the parents of an effect variable Y that assumes values y and $\overline{y}$.

A useful concept for modeling the noisy-OR by means of the deterministic OR, is that of the *inhibitor* nodes [1, 10], which model individual probabilistic relations between each cause and the effect individually. The general model including inhibitor nodes is shown in Fig. 1. The CPT of Y defines how those individual effects combine to produce Y. For the noisy-OR model, the CPT of node Y is equivalent to a deterministic OR. The *inhibitor* nodes Y_i introduce *noise* — the probabilistic effect of X_i on Y. The CPT of each Y_i is of the form: $\Pr(y_i|x_i) = p_i$ where $p_i \in [0, 1]$, and $\Pr(y_i|\overline{x}_i) = 0$. The noisy-MAX [3, 4] is basically an extension of the noisy-OR model to multi-valued variables. The noisy-MAX assumes that the variable Y has n_y states and that these states are ordered. The inhibitor nodes Y_i take values from the same domain as Y and their states follow the same ordering. Every parent variable X_i has n_i values. We use q_{ijk} to denote the element of CPT of inhibitor node Y_i that corresponds to the j-th value of parent node X_i and k-th value of Y_i: $\forall_{ijk}\ q_{ijk} = \Pr(y_i^k|x_i^j)$. Probabilities q_{ijk} are noisy-MAX parameters. The inhibitor variables Y_i have the same range as Y and their CPTs are constrained in the following way:

$$q_{ijk} = \begin{cases} 1 & \text{if } j = 1, k = 1 \\ 0 & \text{if } j = 1, k \neq 1 \\ p \in [0 \ldots 1] & \text{if } j \neq 1\ . \end{cases}$$

The CPT of Y is a deterministic MAX defined by the ordering relation over states of Y. It is a common practice to add a *leak* term to the noisy-OR model. The leak is an auxiliary cause that serves the purpose of modeling the influence of causes that are not explicitly included in the model. In our experiments, we always assume that the noisy-MAX model includes the leak probability.

3 CONVERTING CPT INTO NOISY-MAX GATE

In this section, we propose an algorithm that fits a noisy-MAX distribution to an arbitrary CPT. In other words, it identifies the set of noisy-MAX parameters that produces a CPT that is the *closest* to the original CPT. Let C_Y be the CPT of a node Y that has n parent variables $X_1, \ldots, X_n$. Each variable X_i can take one of n_{X_i} possible values. We use $\mathbf{p}_i$ to denote i-th combination of the parents of Y and $\mathbf{P}$ to denote the set of all combinations of parents values, $\mathbf{P} = \{\mathbf{p}_1, \ldots, \mathbf{p}_m\}$, where m is the product of the numbers of possible values of the X_is, i.e., $m = \prod_{i=1}^{n} n_{X_i}$. There exist several measures of similarity of two probability distributions of which two are commonly used: Euclidean distance and Kullback-Leibler (KL) divergence. The main difference between the two is that the Euclidean distance is based on absolute differences and, hence, is insensitive to large relative differences in very small probabilities, which is its major drawback. The KL divergence addresses this problem, but on the other hand, the problem with KL divergence is that for cases when the estimated probability is zero and the goal probability is non-zero, it is infinity. In our experiments, in such cases we replaced zero with a constant close to zero, which is a common practice.

The problem of defining a distance between two CPTs is somewhat more complicated, because a CPT is a set of probability distributions. The easiest approach is to define distance between two CPTs as a sum of distances of corresponding probability distributions in both CPTs. However, in practice not all distributions in the CPT are equally important. This is because typically some of the configurations of parents' states are far more likely than the others.

Definition 1 (Euclidean distance between two CPTs) *The distance D_E between two CPTs, $\Pr_A(Y|\mathbf{P})$ and $\Pr_B(Y|\mathbf{P})$ is a weighted sum of Euclidean distances between their corresponding probability distributions:*

$$D_E(\underset{A}{\Pr}(Y|\mathbf{P}), \underset{B}{\Pr}(Y|\mathbf{P})) = \\ = \sum_{i=1}^{m} w_{\mathbf{p}_i} \sum_{j=1}^{n_Y} \left(\underset{A}{\Pr}(y_j|\mathbf{p}_i) - \underset{B}{\Pr}(y_j|\mathbf{p}_i) \right)^2 . \qquad (1)$$

Definition 2 (KL divergence between CPTs) *The divergence D_{KL} between the goal CPT $\Pr_A(Y|\mathbf{P})$ and its approximation $\Pr_B(Y|\mathbf{P})$ is a weighted sum of KL divergences between their corresponding probability distributions:*

$$D_{KL}(\underset{A}{\Pr}(Y|\mathbf{P}), \underset{B}{\Pr}(Y|\mathbf{P})) = \\ \sum_{i=1}^{m} w_{\mathbf{p}_i} \sum_{j=1}^{n_Y} \underset{A}{\Pr}(y_j|\mathbf{p}_i) \ln \frac{\Pr_A(y_j|\mathbf{p}_i)}{\Pr_B(y_j|\mathbf{p}_i)} , \qquad (2)$$

In both definitions $w_{\mathbf{p}_i}$ is a weighting constant for each distribution in the CPT and $w_{\mathbf{p}_i} = P(\mathbf{p}_i)$.

Definition 3 (MAX-based CPT) *A MAX-based CPT $\Pr_q(Y|\boldsymbol{P})$ is a CPT constructed from a set of noisy-MAX parameters $\boldsymbol{q}$.*

Our goal is to find for a given fully specified CPT $\Pr_{cpt}(Y|\mathbf{P})$, such set of the noisy-MAX parameters $\mathbf{q}$ that minimizes the Euclidean distance between the original CPT and the MAX-based CPT $\Pr_q(Y|\mathbf{p}_i)$. For simplicity, we will use θ_{ij} to denote the element of CPT that corresponds to the i-th element of $\mathbf{P}$ and j-th state of Y. When this parameter is given in a fully defined CPT, we use upper index θ_{ij}^{cpt}, and when the parameter was obtained from the MAX-based CPT, we use upper index θ_{ij}^{max}. We can now rewrite Eq. 1 as:

$$\sum_{i=1}^{m} w_{\mathbf{p}_i} \sum_{j=1}^{n_Y} \left(\theta_{ij}^{cpt} - \theta_{ij}^{max} \right)^2 .$$

Because further part of our discussion relies heavily on cumulative probabilities, we introduce cumulative probability distributions based on the parameters θ_{ij} and q_{ij}. We define Θ_{ij} as:

$$\Theta_{ij} = \begin{cases} \sum_{k=1}^{j} \theta_{ik} & \text{if } j \neq 0 \\ 0 & \text{if } j = 0\ , \end{cases}$$

which constructs a cumulative probability distribution function for $\Pr(Y|\mathbf{p}_i)$. It is easy to notice, that $\theta_{ij} = \Theta_{ij} - \Theta_{i(j-1)}$. The next step is to express the MAX-based CPT parameters θ_{ij}^{max} in terms of the noisy-MAX parameters. In similar manner, we define the cumulative probability distribution of noisy-MAX parameters as:

$$Q_{ijk} = \begin{cases} \sum_{l=1}^{k} q_{ijl} & \text{if } j \neq 0 \\ 0 & \text{if } j = 0\,. \end{cases}$$

Pradhan et al. [6] proposed an algorithm exploiting cumulative probability distributions for efficient calculation of the MAX-based CPT that computes parameters of the MAX-based CPT as follows:

$$\Theta_{ij}^{max} = \prod_{x_p^r \in \mathbf{p}_i} Q_{prj}\,. \tag{3}$$

The product in Eq. 3 is taken over all elements of the cumulative distributions of noisy-MAX parameters, such that the values of a parent node X_i belong to a combination of parent states in CPT. Eq. 4 shows how to compute the element θ_{ij}^{max} from the noisy-MAX parameters:

$$\begin{aligned} \theta_{ij}^{max} &= \Theta_{ij}^{max} - \Theta_{i(j-1)}^{max} \\ &= \prod_{x_p^r \in \mathbf{p}_i} Q_{prj} - \prod_{x_p^r \in \mathbf{p}_i} Q_{pr(j-1)} \\ &= \prod_{x_p^r \in \mathbf{p}_i} \sum_{k=1}^{j} q_{prk} - \prod_{x_p^r \in \mathbf{p}_i} \sum_{k=1}^{j-1} q_{prk}\,. \end{aligned} \tag{4}$$

However, parameters θ_{ij}^{max} have to obey the axioms of probability, which means that we have only $n_Y - 1$ independent terms and not n_Y, as the notation suggests. Hence, we can express θ_{ij}^{max} in the following way:

$$\theta_{ij}^{max} = \begin{cases} \prod_{x_p^r \in \mathbf{p}_i} \sum_{k=1}^{j} q_{prk} - \prod_{x_p^r \in \mathbf{p}_i} \sum_{k=1}^{j-1} q_{prk} & \text{if } j \neq n_Y \\ 1 - \prod_{x_p^r \in \mathbf{p}_i} \sum_{k=1}^{n_Y - 1} q_{prk} & \text{if } j = n_Y\,. \end{cases}$$

Theorem 1 *The distance D_E between an arbitrary CPT $\Pr_{cpt}(Y|\boldsymbol{P})$ and a MAX-based CPT $\Pr_q(Y|\boldsymbol{P})$ of noisy-MAX parameters $\boldsymbol{q}$ as a function $\boldsymbol{q}$ has exactly one minimum.*

Proof 1 *We prove that for each noisy-MAX parameter $q \in \boldsymbol{q}$, the first derivative of D_E has exactly one zero point. We will subsequently show that the second derivative is always positive, which indicates that D_E has exactly one minimum and proves the theorem. The first derivative of D_E over q is*

$$\begin{aligned} \frac{\partial}{\partial q} \sum_{i=1}^{m} w_{\mathbf{p}_i} \sum_{j=1}^{n_Y - 1} \left(\theta_{ij}^{cpt} - \prod_{x_p^r \in \boldsymbol{p}_i} \sum_{k=1}^{j} q_{prk} + \prod_{x_p^r \in \boldsymbol{p}_i} \sum_{k=1}^{j-1} q_{prk} \right)^2 \\ + \frac{\partial}{\partial q} w_{\mathbf{p}_i} \sum_{i=1}^{m} \left(- \sum_{j=1}^{n_Y - 1} \theta_{ij}^{cpt} + \prod_{x_p^r \in \boldsymbol{p}_i} \sum_{k=1}^{n_Y - 1} q_{prk} \right)^2 . \end{aligned}$$

Each of the three products contains at most one term q and, hence, the expression takes the following form:

$$\frac{\partial}{\partial q} \sum_{i,j} \left(A_{ij} + B_{ij} q \right)^2 , \tag{5}$$

where A_{ij} and B_{ij} are constants. At least some of the terms B_{ij} have to be non-zero (because external sum in Eq. 5 runs over all elements of the CPT). The derivative

$$\frac{\partial}{\partial q} \sum_{i,j} \left(A_{ij} + B_{ij} q \right)^2 = 2 \sum_{i,j} \left(A_{ij} B_{ij} \right) + 2q \sum_{i,j} B_{ij}^2$$

is a non-trivial linear function of q. The second order derivative is equal to $2 \sum_{i,j} B_{ij}^2$ and always takes positive values. Therefore, there exist exactly one local minimum of the original function.

In our approach, for a given CPT we try to identify the set of noisy-MAX parameters that minimizes the distances D_E or D_{KL}.

The problem amounts to finding the minimum of the distance which is a multidimensional function of the noisy-MAX parameters. We proved that for the Euclidean distance, there exists exactly one local minimum. Therefore, any mathematical optimization method ensuring convergence to a single minimum can be used. In case of KL divergence, we have no guarantee that there exists exactly one local minimum. But for the purpose of our experiments, this is a conservative assumption. To perform our experiments we implemented a simple gradient descent algorithm that takes a CPT as an input and produces noisy-MAX parameters and a measure of fit as an output.

4 HOW COMMON ARE NOISY-MAX GATES IN REAL MODELS

We decided to test several sizeable real world models in which probabilities were specified by an expert, learned from data, or combination of both. Three models that contained sufficiently large CPTs which were defined without parametric distributions were available to us: ALARM [12], HAILFINDER [11] and HEPAR II [8]. We verified by contacting the authors of these models that none of the CPTs in these networks were specified using the noisy-OR/MAX assumption. For each of the networks, we first identified all nodes that had at least two parents and then we applied our conversion algorithm to these nodes. HEPAR contains 31 such nodes, while ALARM and HAILFINDER contain 17 and 19 such nodes respectively. We tried to fit the noisy-MAX model to each of these nodes using both D_E and D_{KL} measures.

We used two criteria to measure the goodness of fit between a CPT and its MAX-based equivalent: (1) *Average*, the average Euclidean distance (with square root, not weighted by probabilities of parents instantiations) between the two corresponding parameters and (2) *Max*, the maximal absolute value of difference between two corresponding parameters, which is an indicator of the single worst parameter fit for a given CPT.

Fig. 2 shows the results for the three tested networks for Euclidean and KL measures respectively. The figures show the distance for all networks on one plot. The nodes in each of the networks are sorted according to the corresponding distance (*Average* or *MAX*) and the scale is converted to percentages. We can see for the *MAX* distance that for roughly 50% of the variables in two of the networks the greatest difference between two corresponding values in the compared CPTs was less than 0.1.

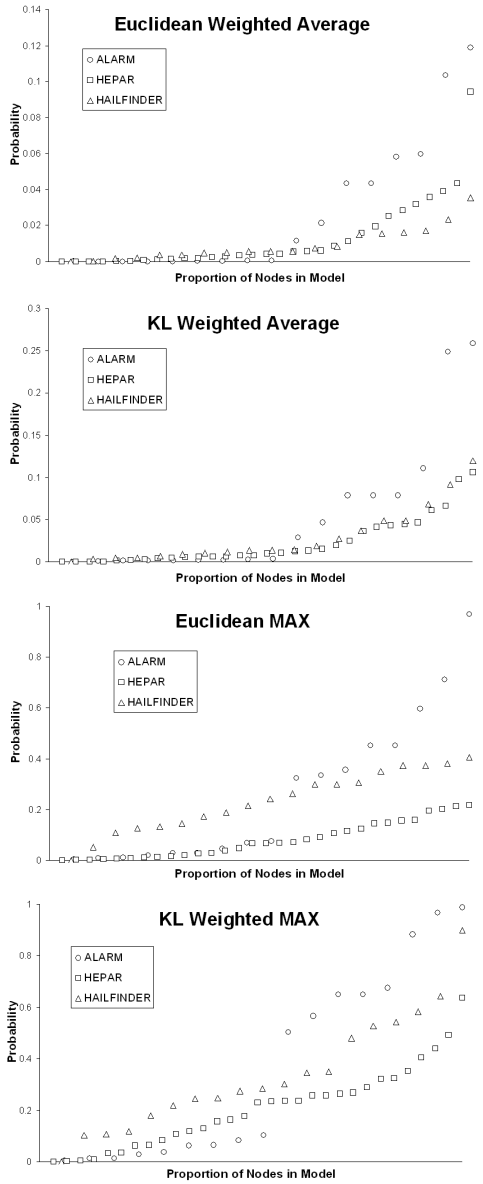

Figure 2. The *Average* and *MAX* distance for the nodes of the three analyzed networks obtained using weighted distances.

One possible explanation of our findings is that the noisy-MAX model is likely to fit well any randomly selected CPT. We decided to verify this by generating CPTs for binary nodes, with 2-5 parents (10,000 CPTs for every number of parents, for a total of 40,000 CPTs), whose parameters were sampled from the uniform distribution. Fig. 3 shows the results. On the X-axis there are generated CPTs sorted according to their fit to the noisy-OR using average and MAX measures. Except for the cases with two parents, the results are qualitatively different from the results we obtained using real-life models. They clearly indicate that approximating a randomly generated CPT by the noisy-OR is highly improbable. Additionally, these results can provide empirical grounds for interpretation of values of distance measures.

The small difference in the conditional probabilities does not necessarily imply that differences in the posterior probabilities will be of

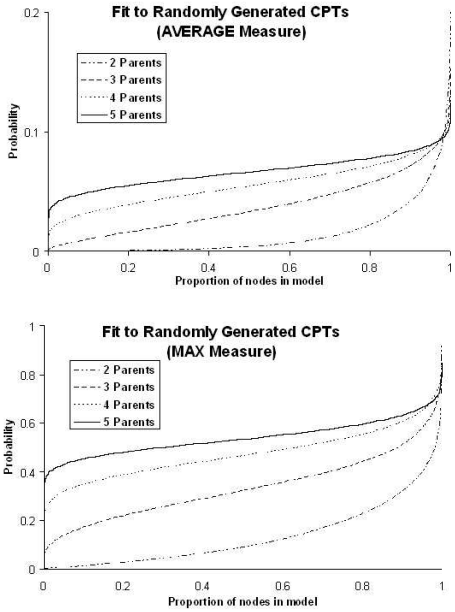

Figure 3. The *Average* and *MAX* distances for randomly generated CPTs.

a similar magnitude. We decided to test the accuracy of models with some nodes converted into the noisy-MAX. For each of the tested networks we converted one by one the selected nodes into noisy-MAX gates using Euclidean and KL measures. In this way, after each node had been converted, the new model was created. For each such model we generated random evidence using logic sampling for 10% of the nodes in the network and calculated the posterior probability distributions over the remaining nodes. We compared these posterior probabilities to those obtained in the original model, which we treated as the gold standard. We repeated the procedure described above 1000 times for each of the three models.

We decided to perform fit of the noisy-MAX model to CPTs without taking into account the probabilities of parent instantiations. One may argue that to answer the question if a local distribution fits the noisy-MAX should not take into account the propor distribution of parent variables. It is equivalent to assuming that the constants $w_{\mathbf{p}_i}$ in Eqs. 1 and 2 are always equal to 1. We decided to report these results together with the weighted distances. In the sequel, we will refer to uniformly weighted distributions as *simple* and the weighted according to probabilities of parent combinations as *weighted*.

Fig. 4 shows the results of tests for accuracy of posterior probabilities for the three networks. On the X-axis there are nodes sorted by the goodness of fit. On the Y-axis there is absolute average maximal error between posterior probabilities for 1,000 trials. We observe a consistent tendency that the accuracy of the posterior probabilities is poorer for nodes that have worse fit to the noisy-MAX. From these results one can conclude that the weighted KL divergence is superior to other distances, when it comes to CPTs which are good fits to the noisy-MAX (these at the left hand side of X-axis). The nodes on the right hand side are usually not of interest, because they represent nodes that are not good fit anyway. The nature of maximal error can explain the weighted distances' worse performance in case of HEPAR II network. The HEPAR II network has many small probabilities, and these lead to orders of magnitude differences in probabilities of parents' instantiations. The weighted distances work well on the

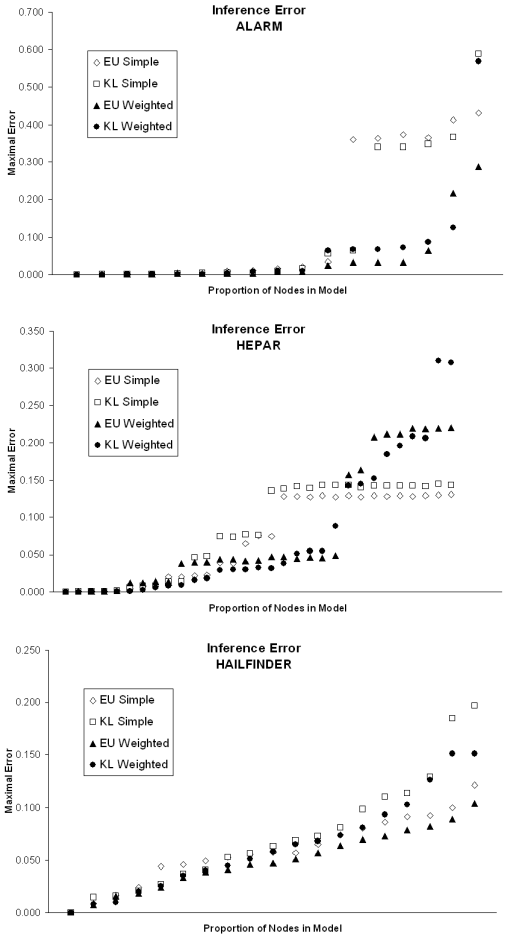

Figure 4. Accuracy of the posterior probabilities for the three networks. Evidence sampled from the posterior distribution.

average, but their performance can be worse for less likely scenarios and this additionally can be amplified by using the maximal measure, which captures the worst case scenario.

In the later version of the HEPAR II model domain expert's knowledge was used to identify variables that are candidates for the noisy-MAX distributions [8]. We had an access to this later version of the model and we could compare results obtained by our algorithm to the variables indicated by the expert. The surprising part is that the expert did not identify variables that were the best fit – she listed only 4 of the 10 variables with the best fit. Apparently the experts strategy was to convert as much as possible variables with large CPTs to the noisy-MAX, even though they were not a good fit. After some closer investigation, we concluded that these nodes are indeed a good fit to the noisy-MAX, but they required relatively complex manipulations on the order of states to fit the noisy-MAX model which was possibly beyond expert's modeling skills.

5 DISCUSSION

In this paper, we introduced two measures of distance between two CPTs – one based on the Euclidean distance and one based on the KL divergence. We proved that Euclidean distance between any CPT and a MAX-based CPT, as a function of the noisy-MAX parameters of the latter, has exactly one minimum. We applied this result to an algorithm that given a CPT finds a noisy-MAX distribution that provides the best fit to it. Subsequently, we analyzed CPTs in three existing Bayesian network models using both measures. Our experimental results suggest that the noisy-MAX gate may provide a surprisingly good fit for as many as 50% of CPTs in practical networks. We demonstrated, that this result can not be observed in randomly generated CPTs. We tested the influence of accuracy of the approximation of a CPTs by noisy-MAX gates on the accuracy of posterior probabilities showing that models with some nodes converted to the noisy-MAX provide good approximation of the original models. Our results provide strong empirical support for the practical value of the noisy-MAX models. We showed that the relation defined by the noisy-MAX often approximates interactions in the modeled domain reasonably well. It seems to us, based on this result, that the independence of causal interactions is fairly common in real-world distributions.

ACKNOWLEDGEMENTS

This research was supported by the Air Force Office of Scientific Research under grant F49620-03-1-0187.

REFERENCES

[1] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Mateo, CA: Morgan Kaufmann Publishers, Inc., 1988.

[2] Y. Peng and J. A. Reggia, "Plausibility of diagnostic hypotheses," in *Proceedings of the 5th National Conference on AI (AAAI–86)*, Philadelphia, 1986, pp. 140–145.

[3] M. Henrion, "Some practical issues in constructing belief networks," in *Uncertainty in Artificial Intelligence 3*, L. Kanal, T. Levitt, and J. Lemmer, Eds. New York, N. Y.: Elsevier Science Publishing Company, Inc., 1989, pp. 161–173.

[4] F. J. Díez, "Parameter adjustement in Bayes networks. The generalized noisy OR–gate," in *Proceedings of the 9th Conference on Uncertainty in Artificial Intelligence*. Washington D.C.: Morgan Kaufmann, San Mateo, CA, 1993, pp. 99–105.

[5] F. J. Díez, J. Mira, E. Iturralde, and S. Zubillaga, "DIAVAL, a Bayesian expert system for echocardiography," *Artificial Intelligence in Medicine*, vol. 10, pp. 59–73, 1997.

[6] M. Pradhan, G. Provan, B. Middleton, and M. Henrion, "Knowledge engineering for large belief networks," in *Proceedings of the Tenth Annual Conference on Uncertainty in Artificial Intelligence (UAI–94)*. San Francisco, CA: Morgan Kaufmann Publishers, 1994, pp. 484–490.

[7] M. Shwe, B. Middleton, D. Heckerman, M. Henrion, E. Horvitz, H. Lehmann, and G. Cooper, "Probabilistic diagnosis using a reformulation of the INTERNIST–1/QMR knowledge base: I. The probabilistic model and inference algorithms," *Methods of Information in Medicine*, vol. 30, no. 4, pp. 241–255, 1991.

[8] A. Oniśko, M. J. Druzdzel, and H. Wasyluk, "Learning Bayesian network parameters from small data sets: Application of noisy-OR gates," *International Journal of Approximate Reasoning*, vol. 27, no. 2, pp. 165–182, 2001.

[9] F. J. Díez and S. F. Galán, "Efficient computation for the noisy-MAX," *International Journal of Intelligent Systems*, vol. 18, no. 2, pp. 165–177, 2004.

[10] D. Heckerman and J. S. Breese, "A new look at causal independence," in *Proceedings of the Tenth Annual Conference on Uncertainty in Artificial Intelligence (UAI–94)*. San Francisco, CA: Morgan Kaufmann Publishers, 1994, pp. 286–292.

[11] I. Beinlich, H. Suermondt, R. Chavez, and G. Cooper, "The ALARM monitoring system: A case study with two probabilistic inference techniques for belief networks," in *Proceedings of the Second European Conference on Artificial Intelligence in Medical Care*, London, 1989, pp. 247–256.

[12] B. Abramson, J. Brown, W. Edwards, A. Murphy, and R. Winkler, "Hailfinder: A Bayesian system for forecasting severe weather," in *International Journal of Forecasting*, Amsterdam, 1996, pp. 57–71.