An Empirical Study of Probability Elicitation under Noisy-OR Assumption

Adam Zagorecki and **Marek Druzdzel**
Decision Systems Laboratory
School of Information Science
University of Pittsburgh, Pittsburgh, PA 15260
{adamz,marek}@sis.pitt.edu

Abstract

Bayesian network is a popular modeling tool for uncertain domains that provides a compact representation of a joint probability distribution among a set of variables. Even though Bayesian networks significantly reduce the number of probabilities required to specify probabilistic relationships in the domain, the number of parameters required to quantify large models is still a serious bottleneck. Further reduction of parameters in a model is usually achieved by utilization of parametric probability distributions such as noisy-OR gates. In this paper report the results of an empirical study that suggests that under the assumption, that the underlying modeled distribution follows the noisy-OR assumptions, human experts provide parameters with better accuracy using elicitation of noisy-OR parameters than when eliciting conditional probability tables directly. It also seems that of the two alternative noisy-OR parameterizations due to Henrion and Díez the latter results in better elicitation accuracy.

Introduction

Bayesian network formalism (Pearl 1988) provides a convenient and powerful tool for modeling uncertain domains. A Bayesian network (BN) is a compact representation of the joint probability distribution over a set of variables. A BN consists of two parts: the qualitative part, that represents relationships among variables by the means of an acyclic directed graph, and the quantitative part that defines local probability distributions for each variable in the model. We restrict our discussion to models that consist exclusively of discrete variables, as the noisy-OR model is inherently discrete.

The qualitative part of a Bayesian network is an acyclic directed graph, in which each node represents a variable. The arcs between nodes encode statistical relationships among the variables. The graph structure encodes independencies among the variables, which leads to a significant reduction of the number of probabilities required to specify the joint probability distribution. The quantitative part of a BN specifies the joint probability distribution by means of local probability distributions associated with every node. In case of nodes that do not have direct predecessors, there is a single prior probability distribution; in case of nodes that have predecessors, there is a set of conditional probability distributions, each corresponding to one combination of the parents states. BNs provide a framework for efficient and sound inference algorithms, that permit diagnostic and predictive reasoning.

In terms of model building, BNs provide a great dose of flexibility – models can be obtained from data using theoretically sound learning techniques, built completely from experts knowledge, or a combination of both. Typically, the graphical part of a model is obtained from an expert. The graphical part can be also learned from data, enhanced with some constraints imposed by the expert. Parameters are learned from data using machine learning techniques, elicited from human expert or both. Typically, it is useful, if not necessary, to elicit numerical probabilities from experts. Therefore, need for good knowledge elicitation schemes for Bayesian networks has been recognized and is currently a field of active research (Druzdzel & van der Gaag 2000).

Even though BNs significantly reduce the number of parameters required to specify a joint probability distribution, the number of parameters in a model remains still one of the major bottlenecks of this framework. This problem arises, when a model contains nodes with multiple parents. Every node that has parent nodes has a conditional probability table (CPT) associated with it. The CPT consists of a set of discrete probability distributions indexed by all possible combinations of parents states. For example, if all variables in a model are binary and a variable X has 3 parents, the CPT that corresponds to X will have $2^3 = 8$ probability distributions (because there are 3 parents of X with 2 states each). Hence, the number of parameters in a CPT grows exponentially in the number of the parent nodes, easily reaching practically prohibitively large values. One way of reducing this number is to assume some functional relationship that defines the interaction among parents on the child node. The most accepted and widely applied solution to this problem is the noisy-OR model (Pearl 1988). The noisy-OR model gives the interaction between the parents and the child a causal interpretation and assumes that all causes (parents) are independent of each other in terms of their abilities to influence the effect variable (child). Given this assumption the noisy-OR model provides a logarithmic reduction in the number of parameters relatively to CPT, which virtually makes construction of large models for real-life problems feasible (Pradhan *et al.* 1994). The noisy-MAX model is an extension of the noisy-OR for the multi-valued variables. In another paper we investigate, how often the noisy-OR/MAX models can provide a good approximation of the CPTs that were acquired without the noisy-OR/MAX assumption. It turned out that for the models under study up to 50% of CPTs could be reasonably approximated by the noisy-OR/MAX.

In this paper, we describe an empirical comparison of accuracy in probability elicitation for fully specified CPT and the noisy-OR model. We believe that our study is important, because it empirically compares the two widely applied frameworks, showing that under assumption that the underling phenomenon can be reasonably modeled by a noisy-OR gate, the direct elicitation of noisy-OR parameters yields significantly better accuracy than direct elicitation of CPT. We also investigate which of two noisy-OR parameterizations proposed in the literature leads to better accuracy. We limit our study to the case of three parent variables and a single effect under assumption that all the variables are binary. We believe that this setting is actually most favorable for elicitation of parameters for the full conditional probability table. In cases when the number of parents is larger, the advantage of noisy-OR becomes more apparent.

In the following sections, we first describe the noisy-OR model, its assumptions, important equations and we explain the difference between the two parameterizations. Then we present the design of the experiment and discuss its results. Finally, we draw conclusions and discuss limitations of this study.

Noisy-OR

The noisy-OR gate is a member of the family of models often referred to as independence of causal influences (ICI) (Heckerman & Breese 1994; Srinivas 1993; Díez & Druzdzel 2003). The name of the family reflects the basic assumption behind these models, which states that the influence of each modeled cause on the effect is independent of other causes. The word *noisy* in the name indicates that the causal interaction is not deterministic, in the sense that any cause can produce the effect with some probability. The second part of the name – *OR* originates from the functional relation, which determines the manner in which independent causes combine to produce their common effect.

One can think about the noisy-OR model as a probabilistic extension of the deterministic OR model. Similarly to the deterministic OR model, the noisy-OR model assumes that the presence of each cause X_i is sufficient to produce the presence of the effect Y and that its ability to produce that effect is independent of the presence of other causes. However, the presence of a cause X_i in the noisy-OR model does not guarantee that the effect Y will occur.

For modeling the noisy-OR by means of the deterministic OR, we can introduce the concept of inhibitor nodes (Pearl 1988; Heckerman & Breese 1994), which capture probabilistic relations between each cause and the effect individually. The general model is shown in Figure 1. Causal independence is represented here by the inhibitor nodes Y_i, whose CPTs encode individual effect of corresponding X_i

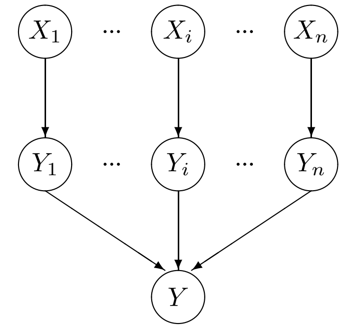

Figure 1: Direct modeling of noisy-OR

on Y. The CPT of Y defines how the individual effects of X_is interact to produce Y. For the noisy-OR, the CPT of node Y is equivalent to a deterministic OR. The concept of *inhibitor* nodes Y_i is introduced to represent *noise* — the probabilistic effect of X_is on Y. The CPT of Y_i is of the form:

$$\begin{aligned}\Pr(y_i|x_i) &= p_i\\ \Pr(y_i|\overline{x}_i) &= 0\,.\end{aligned}$$

The value of p_i represents the probability that the presence of X_i will cause the presence of Y and is sometimes referred to as *causal strength*. The noisy-OR model assumes that the absence of X_i never produces presence of Y. The above assumptions allow us to specify the entire CPT for Y by means of only n parameters p_i. Of course, such CPT is constrained and it is not possible to represent an arbitrary CPT by means of the noisy-OR distribution.

Leaky Noisy-OR

In practice, it is fairly impossible to capture all possible causes that can produce the effect in the model. Therefore, in practical models, the situation when absence of all *modeled* causes guarantees absence of the effect almost never happens. To address this weakness of the noisy-OR model, Henrion (1989) introduced the concept of *leak* or *background probability* which allows for modeling combined influence of such unmodeled factors.

The leaky noisy-OR model can formally be represented by the noisy-OR. The concept of leak is intended to help domain experts to simplify knowledge engineering process. The leak can be conceptualized as an additional cause X_0 and its corresponding inhibitor node Y_0. We assume that X_0 is always present and the probability that it will cause Y_0 is $\Pr(y_0|x_0) = p_0$, where $0 < p_0 < 1$. Alternatively, we can conceptualize the leak as a binary variable X_0 with prior probability $\Pr(x_0) = p_0$, whose presence always produces presence of Y.

The interpretation of p_0 is following: p_0 is the probability that the effect Y will be produced by the unmodeled causes in absence of all the modeled causes. We should note, that there is an underlying assumption that the unmodeled causes produce the effect independently of the modeled causes X_is.

Parameterizations of Leaky Noisy-OR

Introduction of the leak to the noisy-OR model has an interesting consequence from the point of view of model parameterization. For the noisy-OR model, parameter p_i of the inhibitor Y_i is equal to probability $P(y|\overline{x}_1, \ldots, x_i, \ldots, \overline{x}_n)$. For the noisy-OR it is not necessary to talk explicitly about the causal strength of one cause on the effect. Instead we can ask our expert for the probability of presence of Y under the condition that only one of the modeled causes is present and all the other causes are absent. Please note that this allows us for estimating the noisy-OR parameters directly from a database.

However, this is not true anymore for the leaky noisy-OR gate. This fact is attributed to the assumption that leak is unobserved. For convenience of further discussion, let us denote $p_i = \Pr(y_i|x_i)$ and $q_i = P(y|\overline{x}_1, \ldots, x_i, \ldots, \overline{x}_n)$. It is easy to note, that for the leaky noisy-OR the relation between p_i and q_i is the following:

$$q_i = 1 - \frac{1 - p_i}{1 - p_0} \, . \tag{1}$$

The leaky noisy-OR model can be specified using either causal strengths p_i or the conditional probabilities q_i. Parameterization based on p_i was introduced by Henrion (1989), and we will refer to it as Henrion's parameterization. The alternative parameterization based on q_i was introduced by Díez (1993), we will refer to it as Díez's parameterization. Although both specifications are formally equivalent in the sense of specifying the same mathematical model (Equation 1 defines the transition between the two parameterizations), they differ significantly with respect to knowledge engineering.

Henrion's definition based on conditional probabilities of Y seems to be more appropriate for learning from data – in this setting one can calculate probabilities $P(y|\overline{x}_1, \ldots, x_i, \ldots, \overline{x}_n)$ directly from a database. When parameters p_i are obtained from a human expert, the question asked of the expert is: *What is the probability that Y is present when X_i is present and all other causes of Y that we are considering in the model are absent.*

Díez's parameterization seems be more applicable to the situation, when the human expert has some knowledge of the causal relations between X_is and Y. This means that the expert needs to know explicit parameters for the inhibitor nodes. The question asked of the expert in case of Díez's parameterization: *What is the probability that X_i produces Y?* or *What is the probability that Y is present when X_i is present and all other causes of Y (including those not modeled explicitly) are absent?*

Experimental Design

The goal of our experiment was to compare the accuracy of direct elicitation of CPTs and elicitation of the noisy-OR parameters using two alternative parameterizations, under the assumption that the modeled mechanism follows the noisy-OR distribution.

Our research design is based on an objective method of evaluating probability elicitation methods proposed in (Wang, Dash, & Druzdzel 2002). This approach assumes that expert's knowledge is a database of records which have been observed in the course of gaining the expertise. To reduce the influence of factors other than actually observed instances (like, for example, textbook, witness reports) and to ensure a similar level of expertise among our subjects, we defined an artificial domain, which was reasonably distinct from any real-life domain. We first trained our subjects so that they got some degree of expertise. Subsequently we asked them to provide numerical values of probabilities using the three elicitation methods under study. We compared the obtained probabilities to the probabilities that the subjects were actually exposed to.

We trained our subjects in modeling an artificial domain involving four variables: three causes and a single effect and then we asked them to specify numerical parameters of interaction among the causes and the effect. Providing an artificial domain had the purpose of ensuring that all subjects were equally familiar with the domain, by making the domain totally independent of any real-life domain that they might have had prior knowledge of.

Subjects

The subjects were 44 graduate students enrolled in the course *Decision Analysis and Decision Support Systems* at the University of Pittsburgh. The experiment was performed in the final weeks of the course, which ensured that subjects were sufficiently familiar with Bayesian networks in general and conditional probabilities in particular. The subjects were volunteers who received partial course credit for their participation in the experiment.

Design and Procedure

The subjects were first asked to read the following instructions that introduced them to an artificial domain that we defined for the purpose of this study. In addition, we had full knowledge over what the subjects have actually seen and should have learned about the domain.

> *Imagine that you are a scientist, who discovers a new type of extraterrestrial rock on Arizona desert. The rock has an extraordinary property of producing anti-gravity and can float in the air for short periods of time. However, the problem is, that it is unclear to you what actually causes the rock to float. In a preliminary study, you discovered that there are three factors that can help the rock to levitate. These three factors are: light, X-rays, and high air temperature.*
>
> *Now your task is to investigate, to what degree, each of these factors can produce anti-gravity force in the rock. You have a piece of this rock in a special apparatus, in which you can expose the rock to (1) high intensity halogen light, (2) high dose of X-rays and (3) rise the temperature of the rock to 1000K.*
>
> *You have 160 trials, in each trial you can set any of those three factors to state present or absent. For example, you can expose the rock to light and X-ray while temperature is low.*
>
> *Be aware of the following facts:*

- *Anti-gravity in the rock appears sometimes spontaneously, without any of these three factors present. Make sure to investigate this as well.*
- *You can expect that anti-gravity property of the rock is dependent on all these three factors. Make sure to test interactions among them.*

Additionally, subjects were presented with a Bayesian network for the domain shown in Figure 2 and told, that at the end of experiment they will be asked to answer some questions about the conditional probabilities of the node *Anti-gravity*. The subjects had unlimited time to perform the 160 trials.

In our experiment, the interaction between the node *Anti-gravity* and its parents was assumed to follow the noisy-OR distribution. The subjects were not aware of the fact that the underlying distribution was the noisy-OR and throughout the whole experiment, special caution was exercised not to cue the subjects to this fact.

We decided not to introduce noise to the distribution from which sampling is drown, because we believe it would not strongly extend external validity of the study, and it is highly disputable what value of noise and in which form would be appropriate. We would like to note that the subjects actually observed distributions that were not precisely the noisy-OR. This was caused by a relatively small sample of data observed by the subject during the learning phase.

In order to ensure that our results would not be an artifact of some unfortunate choice of initial parameters, each subject was assigned a unique underlying noisy-OR distribution for the node *Anti-gravity*. To ensure that the probabilities fell in range of modal probabilities, each model had the noisy-OR parameters sampled from uniform distribution ranging from 0.2 to 0.9. To ensure significant difference between Henrion and Díez parameters, the leak values should be significantly greater than zero (otherwise both parameterizations are virtually equivalent). We sampled them from a uniform distribution ranging from 0.2 to 0.5.[1]

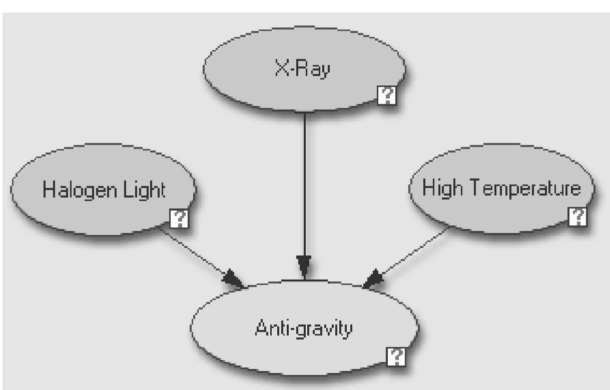

Figure 2: BN used in the experiment

In each of the 160 trials, the subjects were asked to assign values to the independent variables (Halogen Light, X-Ray, and High Temperature) and submit their values to perform the 'experiment' (Figure 3). Subsequently, the screen appeared showing the result – a levitating rock or rock on the ground. An example of the screen that the subject could see is presented in Figure 4.

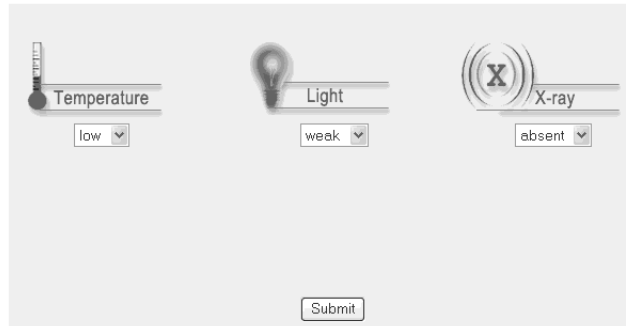

Figure 3: Screen snapshot for setting the three factors

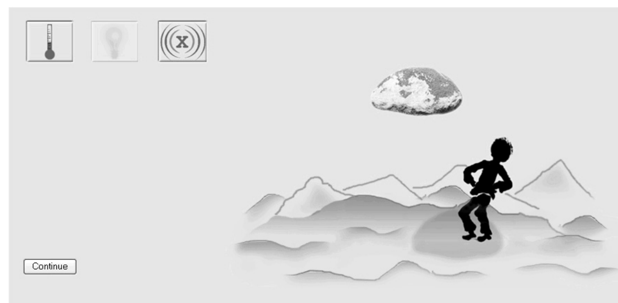

Figure 4: Screen snapshot of the result of a single trial

At the end of the experiment, the subjects were asked to answer questions on conditional probability distribution of the node *Anti-gravity*. We applied a within-subject design. Each subject was asked to express his or her judgement of probabilities by answering three separate sets of questions, which asked for parameters required to define the conditional probability distribution using

1. a complete CPT with 8 parameters,
2. a noisy-OR gate with 4 parameters using Díez's parameterization, and
3. a noisy-OR gate with 4 parameters using Henrion's parameterization.

To compensate for the possible carry-over effects, we counter-balanced the order of the above questions across the subjects. Additionally, we disallowed the subjects to see previously answered questions for the other sets.

We chose a within-subject as opposed to between-subject design due to an anticipated small number of subjects and the possible resulting lack of statistical power. We also feared that learning different distributions for each of tested parameterizations, as suggested by one of the reviewers, would result in an unacceptable time burden on the subjects.

[1] All values are given using Díez parameters.

Results

We decided to remove records of three subjects from further analysis, as we judged these to be outliers. Two of these subjects very likely reversed their probabilities and in places where one would expect large values they entered small values and vice versa. The third subject did not explore all combinations of parent values, making it impossible to compare the elicited probabilities with the cases actually observed by the subject. Therefore, the final number of data records was 41.

We did not measure the individual times for performing the tasks. For most of the subjects, the whole experiment took between 20 and 30 minutes, including probability elicitation part.

As a measure of accuracy, we used the Euclidean distance between the elicited parameters and the probabilities actually seen by the subject. The Euclidean distance is one of the measures used to compare probability distributions. The other commonly used measure is the Kullback-Leibler measure, which is sensitive to extreme values of probabilities. The reason why we selected Euclidean distance is the following. In our study we do not really deal with extreme probabilities and even if the value is close to 1, the subjects preferred entering parameters with accuracy of 0.01. Comparing parameters with this accuracy to accurate probabilities (those presented to the subject) would result in unwanted penalty in case of the Kullback-Leibler measure. The table below shows the mean Euclidean distance between the observed and the elicited probabilities for each of the three methods:

Method	Distance
CPT	0.0800
Henrion	0.0796
Díez	0.0663

For each pair of elicitation methods we performed one-tailed, paired t-test to compare their accuracy. Results suggest that Díez's parameterization performed significantly better than CPT and Henrion's parameterizations (respectively with $p < 0.0008$ and $p < 0.0001$). The difference between Henrion's parameterization and CPT is not statistically significant ($p \approx 0.46$).

We observed consistent tendency among the subjects to underestimate parameters. The average difference per parameter was -0.11 for Henrion's parameters and CPT and -0.05 for Díez's parameters with the individual errors distribution was bell-shaped and slightly asymmetrical. The medians were correspondingly: -0.07, -0.08, and -0.02 respectively.

We tested whether the sampled distributions follow the noisy-OR assumption and whether this had any influence on the accuracy of the elicitation. Figure 5 shows that the sampling distributions followed fairly well the original noisy-OR distributions and no clear relationship between sampling error and the quality of elicitation was observed. This might suggest that for distributions that are further away from the noisy-OR, elicitation error under the noisy-OR assumption might be also smaller than one for direct CPT elicitation.

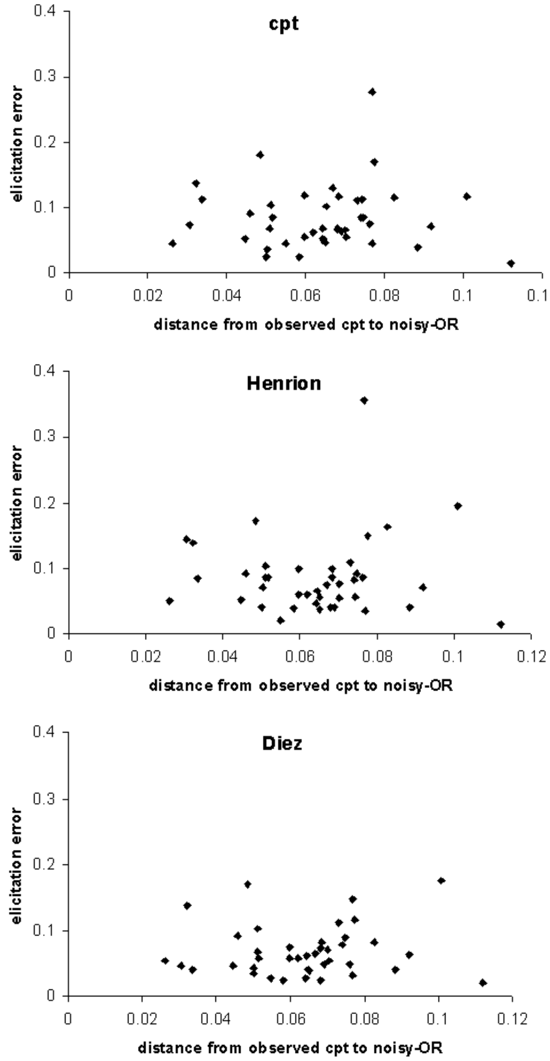

Figure 5: Elicitation error as a function of the distance from observed CPT to noisy-OR

Discussion

We believe that these results are interesting for several reasons. First of all, we show that if an observed distribution follows noisy-OR assumptions, the elicitation of the noisy-OR parameters does not yield worse accuracy than direct elicitation of CPT, even when the number of parameters in the CPT is still manageable.

In our approach, we had a simple model with three binary parent variables and a binary effect variable. We believe, that this setting is favorable for applying CPT framework. When the number of parents increases, the noisy-OR framework will offer significant advantage, as it will require logarithmically fewer parameters. The exponential growth of the number of parameters required to specify full CPT works strongly against this framework for models with larger number of parents.

In our experiments, expert's domain knowledge comes exclusively from observation. It is impossible for a subject to understand the mechanisms of interaction between causes, because such mechanisms are fictional. In light of this fact, it is surprising that Díez's parameters, which assume understanding of causal mechanisms, perform better than Henrion's parameters. The latter are more suitable in situations where one has a set of observations without understanding of relationships between them. One possible rival hypothesis is the following: subjects were unable to provide Díez's parameterization (because it requires separating leak from the cause, which was challenging), so they provided numbers suitable for Henrion's parameterization. In fact, roughly 50% of subjects seemed to act this way. This, in conjunction with the observed tendency to underestimate probabilities, might have led to the situation, where these two contradicting tendencies might have cancelled out leading to more precise results.

Acknowledgements

This research was supported by the Air Force Office of Scientific Research under grants F49620–00–1–0112 and F49620-03-1-0187. All experimental data have been obtained using SMILE, a Bayesian inference engine developed at Decision Systems Laboratory and available at `http://www.sis.pitt.edu/~genie`. We would like to thank all students in the Decision Analysis and Decision Support Systems course at the University of Pittsburgh, Spring 2003, who agreed to take part in this experiment.

References

Díez, F. J., and Druzdzel, M. J. 2003. Canonical probabilistic models for knowledge engineering. Forthcoming.

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

Druzdzel, M. J., and van der Gaag, L. 2000. Building probabilistic networks: Where do the numbers come from? guest editor's introduction. In *Transactions on Knowledge and Data Engineering*. 12(4):481–486.

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

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

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

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

Srinivas, S. 1993. A generalization of the noisy-OR model. In *Proceedings of the Ninth Annual Conference on Uncertainty in Artificial Intelligence (UAI–93)*, 208–215. San Francisco, CA: Morgan Kaufmann Publishers.

Wang, H.; Dash, D. H.; and Druzdzel, M. J. 2002. A method for evaluating elicitation schemes for probabilistic models. In *Transactions on Systems, Man, and Cybernetics-Part B: Cybernetics*. 32(1):38–43.