An Experimental Procedure for Evaluating User-Centered Methods for Rapid Bayesian Network Construction

Michael Farry, Jonathan Pfautz & Zach Cox
Charles River Analytics, Inc.
625 Mount Auburn St.
Cambridge, MA 02138

Ann Bisantz & Richard Stone
Industrial and Systems Engineering
University at Buffalo
Amherst NY 14260

Emilie Roth
Roth Cognitive Engineering
Brookline, MA

Abstract

Bayesian networks (BNs) are excellent tools for reasoning about uncertainty and capturing detailed domain knowledge. However, the complexity of BN structures can pose a challenge to domain experts without a background in artificial intelligence or probability when they construct or analyze BN models. Several canonical models have been developed to reduce the complexity of BN structures, but there is little research on the accessibility and usability of these canonical models, their associated user interfaces, and the contents of the models, including their probabilistic relationships. In this paper, we present an experimental procedure to evaluate our novel Causal Influence Model structure by measuring users' ability to construct new models from scratch, and their ability to comprehend previously constructed models. [Results of our experiment will be presented at the workshop.]

1. INTRODUCTION AND MOTIVATION

A Bayesian network (BN) (Jensen, 2001; Pearl, 1988) is a probabilistic model used to reason under uncertainty. Successful efforts in applying Bayesian modeling to a variety of domains (e.g., computer vision (Rimey & Brown, 1994), social networks (Koelle et al., 2006), human cognition (Guarino et al., 2006; Glymour, 2001), and disease detection (Pang et al., 2004)) have inspired knowledge engineers to use BNs to capture domain knowledge from experts. However, expressing an expert's domain knowledge in a BN is cumbersome due to the complex, tedious, and mathematical nature of conditional probability table (CPT) construction. Adding states and parents to a node quickly results in an exponential explosion in the number of CPT entries required (Pfautz et al., 2007). Canonical models such as Noisy-OR (Henrion, 1989; Pearl, 1988), Noisy-MAX (Diez & Galan, 2003; Diez, 1993; Henrion, 1989), Qualitative Probabilistic Networks (Wellman, 1990) and Influence Networks (IN) (Jensen, 1996; Rosen & Smith, 1996a; Rosen & Smith, 1996b) have been developed to mitigate this problem. In response to some issues raised by those models, and to simplify the Bayesian modeling process through novel user interface techniques, we developed a new canonical model, the Causal Influence Model (CIM) (Cox & Pfautz, 2007; Pfautz et al., 2007). The CIM paradigm was inspired by anecdotal evidence gained by developing systems for domain experts interacting with BNs and by an analysis of other canonical models to determine the constraints that limit their generalizability and applicability.

There have been few user-centered evaluation efforts to assess how (and if) canonical models help domain experts elicit their knowledge and understanding of models presented to them, or how graphical interfaces and their features and properties impact the way people create, interpret, reason with, or base actions on Bayesian networks. The purpose of our study is to provide baseline information on how people construct and describe CIMs presented and created within a graphical user interface.

1.1 BACKGROUND

A *canonical model* (Diez & Druzdzel, 2001) is a modeling pattern that allows probabilistic relationships between nodes to be specified by a reduced set of parameters (i.e., without completing every cell in a CPT). By assuming that the reduced parameters can still accurately represent the domain being modeled, users can quickly build a complex BN that would otherwise take a large amount of time. Most canonical models achieve their reduced parameters by assuming the independent effects of parents. This assumption allows a linear number of parameters to quantify an entire CPT; in the best-case scenario, only a single parameter per parent is needed. Canonical models can also serve as a "front-end" tool for the initial model-building effort, since the CPTs can always be refined by hand or with data at a later time. Some of the simplified patterns followed by canonical models have been motivated by the process followed when eliciting key factors and probabilistic relationships from domain experts (O'Hagan et al., 2006; Hastie & Dawes, 2001).

A review of canonical models sheds light on the advantages and drawbacks of each model. The Influence Network (IN) model can only be used with Boolean nodes. It assumes that the child node has a baseline probability of occurring independently of any parent effects and that each parent independently influences the child to be more or less likely to be true. Since a single baseline probability for the child and a single change in probability for each parent are simple parameters for users to specify, the IN represents a powerful mechanism for capturing domain knowledge. However, since only Boolean nodes are allowed in the IN model, model flexibility is significantly reduced. BNs commonly contain nodes that represent concepts other than the occurrence or non-occurrence of events, and INs cannot be used to simplify these BNs without considerably re-architecting the model.

The Noisy-OR model is also used only with Boolean nodes and assumes that a true state in any parent can cause the child to be true independently of the other parents, with some uncertainty. Similar to INs, the main drawback of the Noisy-OR is its limitation to only Boolean nodes. The Noisy-MAX model generalizes the Noisy-OR and allows ordinal nodes at the expense of increasing the complexity of parameters. Although Noisy-MAX does work with ordinal nodes, it cannot be used with more general discrete nodes that do not have ordered states. These nodes, referred to as categorical nodes, have an arbitrary number of unordered states and usually represent the category or type of something. Qualitative Probabilistic Networks (QPNs) allow for the construction of purely qualitative relationships between nodes in a network, to abstract from the highly quantitative and numerical nature of typical Bayesian models. QPNs consider the "signs" inherent in probabilistic relationships between nodes, and consider the additive synergies between nodes to capture more complicated probabilistic relationships between them (i.e., if A and B both have a positive influence on node C, their influences may be synergistic in nature: if A and B are both true, their cumulative influence upon C may be greater than just the sum of their individual influences.) QPNs allow for more qualitative model elicitation and may therefore be appropriate for interactions with non-technical experts, but they are limited in their ability to provide hard, numerical estimates of the likelihood of events.

The Causal Influence Model (CIM) is a canonical model that retains the desirable properties of the IN while providing solutions to its problems. The CIM assumes that each node is discrete and has an arbitrary number of states with arbitrary meaning. Each node has a baseline probability distribution, independent of any parent effects. Each parent independently influences these baseline probabilities to be more or less likely. The CIM also introduces simplifications that govern the generation of conditional probability relationships, enabling Boolean, ordinal, and categorical nodes to be included. A full description of the mathematical formulas that govern CIMs, including formulas to translate CIM link strengths into conditional probability tables, is provided in (Cox et al., 2007).

Studies have been conducted to analyze and mitigate complexities that arise in the construction of Bayesian models as a result of knowledge elicitation (Onisko, Druzdzel, & Wasyluk, 2001), but no studies to date have assessed the accessibility and usability of various canonical models and associated user interfaces when provided directly to domain experts. The following study investigates how users interpret and create CIMs within a particular user interface.

2. METHOD

2.1 PARTICIPANTS

Up to twenty participants are recruited from the university community to perform the study. After providing informed consent, participants are given the Ishihara Test for color blindness. Participants who pass this screening continue with the study.

2.2 EXPERIMENTAL SYSTEM

We have developed an CIM-enabled version of our BNet.Builder product to allow us to experiment with graphical interfaces for Bayesian network modeling (Pfautz et al., 2007). Using a simple point-and-click interface, users can create, label, connect, and move nodes in the model. Users can also create and modify causal links to represent positive or negative influences between nodes and the strength of those relationships. Users can also post or remove evidence to any node and view the effects of posted evidence on the belief states of other nodes. Link strengths are converted using CPTs based on algorithms provided in (Cox et al., 2007; Pfautz et al., 2007). The positivity or negativity of a causal link and the link strength are represented visually by the color and thickness of the link, respectively.

To simplify model construction for this particular experiment, the CIM interface has been constrained so that all nodes are Boolean; initial beliefs are set to 0.5 for all nodes and cannot be changed directly by the user (but can change based on evidence or link strengths); and only "hard" evidence can be posted (e.g., evidence that the node was either fully true, or fully false). This represents a set of simplifications we have found useful in other work, particularly among users less familiar with Bayesian modeling techniques. Our main goal in this study is to determine whether participants can reason about previously constructed CIMs and construct models to match a given situation. Since these are specific, novel, and fundamental questions with little previous research behind them, we have started with a simple case. The inclusion of additional node types, in particular, is useful for future work in comparing CIMs to other canonical models such as INs, Noisy-OR, and Noisy-MAX.

2.3 EXPERIMENTAL TASKS

Participants will be asked to provide descriptions of and answer questions about a series of CIMs shown in the BNet.Builder interface. In the first task, participants will be shown a model and asked questions about the structure and nature of relationships in the model (specifically, questions asking them to describe elements of the model, and questions related to abductive and deductive reasoning using the model). For instance, given the following example model (Figure 1), participants would be asked:

- *Description:* This picture shows a model of part of a car. Describe what causes headlights to be dim, or not dim.
- *Abductive Reasoning:* If the headlights are dim, what does that mean about the other parts of the car?
- *Deductive Reasoning:* The alternator is working. What does that suggest about the headlights? The battery is old. What does that suggest about the headlights? What if the battery is new and the alternator is failing?

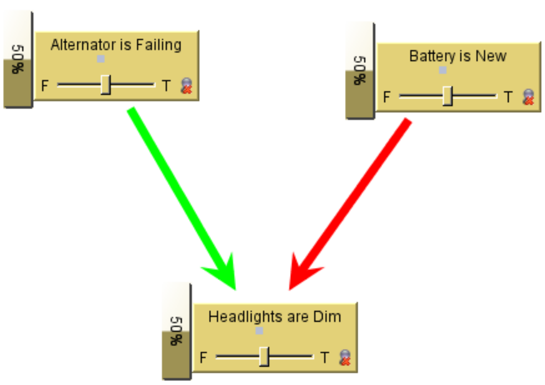

Figure 1. Example model used in the experiment. The green link represents positive influence, while the red link represents negative influence within our CIM-enabled interface.

In the second task, participants can manipulate the causal links and post evidence to see how changing the strength and directionality of the links between the nodes, and evidence about the state of the nodes, affects beliefs about whether the nodes are true or false. They will respond to similar sets of questions as provided in the first task. Finally, in the third task, participants will be asked to construct models from scratch using the interface based on several different vignettes, such as the following:

> *The headlight system on a car is dependent on two components: a battery, which stores energy to power the lights, and an alternator, which converts mechanical energy from the car's engine into stored energy in the battery. When the car is running, the alternator "recharges" the battery. This process only works if the alternator is working, and the battery is new.*

Four models/vignettes have been constructed for each task (a total of 12). Each model has the following relationships: 1 child/1 parent, 2 children/1 parent, 1 child/2 parents, 2 children/2 parents. In all cases, all children are linked to all parents. Also, in all but the 1 child/1 parent case, one parent-child link is negative. This simplification provides the basis for the initial study. We expect to expand upon this simple representation with later empirical work.

2.4 INDEPENDENT VARIABLE

Two stimuli sets are created based on the 12 models. Either the nodes in the models (or phrases in the vignette) are phrased positively, or they include at least one node that uses negative phrasing (e.g., "battery is not new"). This difference allows us to investigate how semantic properties of the model or situation affect task performance. This condition has been inspired by our experience in domain expert interaction with CIM modeling interfaces, where we observed the articulation of variable names as a source of common confusion. The use of negatives in the variable name (e.g., "not raining") or logical antonyms (e.g., "happy" and "sad") tends to lead to later confusion in expressing causal relationships (e.g., "if it is not not-raining, then it is unlikely that Rakesh will not bring his umbrella"). By including this specific independent variable, we will be able to assess which specific patterns of reasoning are most difficult for users. Participants are randomly assigned to one of the two stimuli sets (up to 10 participants per condition). This sample size is consistent with those used in usability type tests, and will allow us to analyze verbal protocols of participants to look for patterns across conditions.

2.5 DEPENDENT MEASURES AND ANALYSIS

Throughout all three tasks, participants are asked to "talk aloud" while performing the task to describe how they are thinking about or creating the models. Screen capture software is used to record participants' interaction with and construction of models. Participants are also fitted with a view point eye tracker (lightweight glasses that have an attached camera that tracks the corneal movements of the participant's eye to assess gaze relative to the computer screen they are working on). The eye tracking system is used to record aspects of gaze position and dwell time at a screen location. Time to complete the tasks is also being recorded.

Data from the audio, eye track, and screen capture processes is combined to create a "process trace" of each participant's behavior describing and creating CIMs (Woods, 1993). Verbalizations and actions are coded and analyzed (Bainbridge & Sanderson, 1995; Sanderson & Fisher, 1994; Woods, 1993) to identify the correctness and completeness of the descriptions and answers provided by participants in the first task, the processes with which participants constructed the models in the second task, and the form and content of the models produced in the third task.

3. ANTICIPATED RESULTS AND DISCUSSION

The purpose of this study is to provide baseline information regarding how people construct and describe CIM models presented and created within the BNet.Builder interface. There is continued interest in simplifying the manner in which domain expertise is elicited, and the creation and presentation of Bayesian network models through direct manipulation and visualization. However, information on how these tools are used by practitioners, how they affect the models that people produce, and how they affect the way that people interpret models or predict outcomes is missing. We anticipate that users will have more difficulty explaining and constructing models with more parent-child connections. We also anticipate users having more difficulty explaining and constructing models when there are more nodes with negative causal links because of the increase in complexity of the models.

In this study, we intend to measure reasoning patterns involving negative quantities that give users the most trouble. We anticipate that users will have the most difficulty interpreting and creating models when nodes are presented with "negatively phrased" labels (e.g., assessing the influence of a node labeled "battery is not new" on a node labeled "headlights are dim"). If this is the case, it suggests a need for developers of CIMs (and BNs in general) to encourage users to employ certain modeling patterns, possibly by constraining the description of nodes. These constraints, in turn, can be accomplished through prior training or interface wizards, or through intelligent, automatic processing of user entries, and provision of suggested alternatives (e.g., pop-up suggestions). These interventions could be tested in further studies.

The primary contribution of this paper will be process- and product-oriented descriptions of how this graphical tool is used to interpret and create CIMs. Future research could compare how models created within the CIM framework compare to those using more traditional BN structures, from the point of view of the user. This study used simple Bayesian models, with constrained parameters and interaction capabilities, and used only Boolean nodes. Future studies, guided by these initial findings, can be conducted using more complex models, a greater variety of node types (e.g., categorical, ordinal), and allow subjects greater flexibility in manipulating CPTs and posting evidence. Other issues for investigation include measuring and mitigating user tendencies to confuse "evidence" and "belief" (both as terms, and in the values these terms represent), measuring tendencies to disregard parental independence when constructing CIMs, and further observation of user reaction to non-intuitive but correct behavior (e.g., becoming confused when particular variables appear overly sensitive or insensitive to posted evidence.)

The CIM interface provides a user-friendly way to express causal influences between nodes, vastly decreasing the number of parameters needed to construct causal models and providing the capability for a much broader base of users to perform Bayesian modeling. Within the experimental interface, participants express relative degrees of influence over a range of 11 steps (from positive to negative 5, with a neutral intermediate value). Additional studies are necessary to clarify the appropriate level of granularity of influence assignment (e.g., 3 steps? 11 steps? 51 steps?) as well whether other methods of assigning strengths across sets of links (e.g., normalized strengths, rank ordered strengths) have merit. Finally, detailed studies with real-world models, situations, and domain experts are required.

Acknowledgements

We would like to thank David Koelle, Geoffrey Catto, Joseph Campolongo, Sam Mahoney, Sean Guarino, and Eric Carlson for their contributions in the development of the CIM and identifying hypotheses to investigate. We also extend our deepest gratitude to Greg Zacharias for his continued funding and support of our work with Bayesian networks.

References

Bainbridge, L., & Sanderson, P. (2005). Verbal protocol analysis. In J. R. Wilson & E. N. Corlett (Eds.), Evaluation of Human Work (pp. 159 - 184). Boca Raton: Taylor and Francis.

Cox, Z. & Pfautz, J. (2007). *Causal Influence Models: A Method for Simplifying Construction of Bayesian Networks.* (Rep. No. R-BN07-01). Cambridge, MA: Charles River Analytics Inc.

Diez, F. J. (1993). Parameter Adjustment in Bayes Networks: The Generalized Noisy OR-Gate. In *Proceedings of the 9th Conference of Uncertainty in Artificial Intelligence,* (pp. 99-105). San Mateo, CA: Morgan Kaufmann.

Diez, F. J. & Druzdzel, M. J. (2001). Fundamentals of Canonical Models. In *Proceedings of Ponencia Congreso: IX Conferencia De La Asociacion Espanola Para La Inteligencia Artificial (CAEPIA-TTIA 2001),* (pp. 1125-1134).

Diez, F. J. & Galan, S. F. (2003). An Efficient Factorization for the Noisy MAX. *International Journal o Intelligent Systems,* 18165-177.

Glymour, C. (2001). *The Mind's Arrows: Bayes Nets and Graphical Causal Models in Psychology*. Cambridge, MA: The MIT Press.

Guarino, S., Pfautz, J., Cox, Z., & Roth, E. (2006). Modeling Human Reasoning About Meta-Information. In *Proceedings of 4th Bayesian Modeling Applications Workshop at the 22nd Annual Conference on Uncertainty in AI: UAI '06*. Cambridge, Massachusetts.

Hastie, R. & Dawes, R. M. (2001). *Rational Choice in an Uncertain World: The Psychology of Judgment and Decision-Making*. London, UK: Sage Publications.

Henrion, M. (1989). Some Practical Issues in Constructing Belief Networks. In L. Kanal, T. Levitt, & J. Lemmer (Eds.), *Uncertainty in Artificial Intelligence 3* (pp. 161-173). North Holland: Elsevier Science Publishers.

Jensen, F. V. (1996). *An Introduction to Bayesian Networks*. London: University College London Press.

Jensen, F. V. (2001). *Bayesian Networks and Decision Graphs*. New York: Springer-Verlag.

Koelle, D., Pfautz, J., Farry, M., Cox, Z., Catto, G., & Campolongo, J. (2006). Applications of Bayesian Belief Networks in Social Network Analysis. In *Proceedings of 4th Bayesian Modeling Applications Workshop at the 22nd Annual Conference on Uncertainty in AI: UAI '06*. Cambridge, Massachusetts.

Kraaijeveld, P., Druzdzel, M., Onisko, A., & Wasyluk, H. (2005). GeNIeRate: An Interactive Generator of Diagnostic Bayesian Network Models. In *Proceedings of Working Notes of the 16th International Workshop on Principles of Diagnosis (DX-05),* (pp. 175-180).

O'Hagan, A., Buck, C., Daneshkhah, A., Eiser, R., Garthwaite, P., Jenkinson, D. et al. (2006). *Uncertain Judgements: Eliciting Experts' Probabilities*. New York: Wiley & Sons.

Onisko, A., Druzdzel, M., & Wasyluk, H. (2001). Learning Bayesian Network Parameters From Small Data Sets: Application of Noisy-OR Gates. *International Journal of Approximate Reasoning,* 27(2), 165-182.

Pang, B., Zhang, D., Li, N., & Wang, K. (2004). Computerized Tongue Diagnosis Based on Bayesian Networks. *IEEE Transactions on Biomedical Engineering,* 51(10), 1803-1810.

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

Pfautz, J., Cox, Z., Koelle, D., Catto, G., Campolongo, J., & Roth, E. (2007). User-Centered Methods for Rapid Creation and Validation of Bayesian Networks. In *Proceedings of 5th Bayesian Applications Workshop at Uncertainty in Artificial Intelligence (UAI '07)*. Vancouver, British Columbia.

Rimey, R. & Brown, C. (1994). Control of Selective Perception Using Bayes Nets and Decision Theory. *International Journal of Computer Vision,* 12(2-3), 173-207.

Rosen, J. & Smith, W. (1996a). Influence Net Modeling With Causal Strengths: An Evolutionary Approach. In *Proceedings of Command and Control Research and Technology Symposium*.

Rosen, J. A. & Smith, W. L. (1996b). Influencing Global Situations: A Collaborative Approach. *US Air Force Air Chronicles*.

Sanderson, P. M., & Fisher, C. (1994). Exploratory sequential data analysis. *Human Computer Interaction*, 9(3), 251 - 317.

Van der Gagg, L.C., Geenen, P.L., & Tabachneck-Schijf, H.J.M. (2006). Verifying Monotonicity of Bayesian Networks with Domain Experts. In *Proceedings of 4th Bayesian Modeling Applications Workshop at the 22nd Annual Conference on Uncertainty in AI: UAI '06*. Cambridge, Massachusetts.

Woods, D. D. (1993). Process tracing methods for the study of cognition outside of the experimental psychology laboratory. In G. A. Klein, J. Orasanu, R. Calderwood & C. E. Zsambok (Eds.), *Decision-making in action: Models and Methods* (pp. 228 - 251). Norwood NJ: Ablex Publishers.

Wellman, M. P. (1990). Fundamental Concepts of Qualitative Probabilistic Networks. *Artificial Intelligence,* 44(3), 257-303.