Reducing online contaminant monitoring uncertainty using a Bayesian belief network

W. J. Dawsey[1], B. S. Minsker[2], and V. L. VanBlaricum[3]

[1]Graduate Research Asst., Dept of Civil & Environmental Engineering, University of Illinois-Urbana, 205. N. Mathews Ave, Urbana, IL 61801; PH (217) 333-6979; dawsey@uiuc.edu

[2]Associate Professor, Dept of Civil & Environmental Engineering, University of Illinois-Urbana, 205. N. Mathews Ave, Urbana, IL 61801; PH (217) 333-9017; minsker@uiuc.edu

[3]General Engineer, U. S. Army Corps of Engineers U. S. Army Engineer Research & Development Center, Construction Engineering Research Laboratory (ERDC-CERL), P. O. Box 9005, Champaign, IL 61826-9005, Phone 217-373-6771, Vicki.L.VanBlaricum@erdc.usace.army.mil

Abstract

There is a great deal of uncertainty in real time characterization of water distribution system contamination events. Much of this uncertainty is due to the lack of targeted sensors which makes it necessary to use surrogate water quality parameters to indirectly measure the presence of a contaminant. A positive sensor detection can often be validated by pieces of evidence observed in a distribution system. This paper illustrates how Bayesian belief networks can be used to represent distribution system contamination scenarios. A framework was developed that integrated sensor data with other validating evidence of a contamination event. This framework was used to express causality between the events and observed evidence that comprise contamination scenarios.

Introduction

Drinking water utilities face many challenges in recognizing, characterizing, and responding to a contamination event. Chief among these is coping with the daunting number of dangerous chemical, biological, and radiological substances that may be accidentally or purposefully introduced into a drinking water system. Furthermore, real-time analytical tools for characterizing many of these contaminants in-situ do not currently exist, or are prohibitively expensive (ASCE, 2004). As a result, current online contaminant monitoring system design makes use of surrogate water quality measures such as total organic carbon (TOC), turbidity, pH, chlorine concentration and others which may be correlated to "fingerprint" contamination events (ASCE, 2004). However, when considered within the context of an entire water collection,

treatment, and distribution system, it is apparent that there are numerous other system events that could obscure or validate such a contaminant detection. Related pieces of information such as physical security alarms, distribution system model topology, and surrogate contaminant measurements need to be tied-in to expert knowledge of potential contamination scenarios. A Bayesian belief network (BBN) is a useful framework for representing the causal relationships between events and observations that comprise a contamination scenario.

Bayesian methodologies, such as BBNs used in this project, have been used to combine diverse data inputs for numerous applications including battlefield strategy, optical recognition, fault detection, advanced driver assistance systems, and sensor network data fusion (Sanzotta and Sherrill, 1997; Fox et al., 2003; Liu and Zhang, 2002; Coue et al., 2003; Karlsson et al., 2002). Bayesian networks have been used to fuse data from multiple sensor networks in many applications (Bonci et al., 2002; Beckerman, 1992; Brown et al., 1995). In the energy domain, Bayesian modeling methods were used to analyze the distribution of the failure rate at nuclear power plants (Chu, 1995). Schlumberger et al. (2002) utilized coupled dynamic models with a BBN to assess the voltage stability limits for part of France's subsystem and to identify more efficient rules of operation.

Substantial research in the Environmental Engineering domain has utilized Bayesian approaches for a number of applications, but few studies have utilized BBNs specifically and none have involved applications to water supply protection as illustrated here. In groundwater remediation, BBNs coupling an expert knowledge base with process models have been used to evaluate the potential of naturally occurring reductive dechlorination at sites contaminated with TCE. (Siber et al., 1999). Marcot (2001) combined expert knowledge with ecological data within a BBN to model the causal relationships between planning decisions and impacts on at-risk wildlife species habitats. Stow et al. (2003) compared a BBN approach with two deterministic models for predicting the effect of nitrogen loading on estuarine chlorophyll a concentrations.

Research has been conducted in the military domain to apply Bayesian networks and decision trees to support battlefield decision-making. The CoRaven system described by Jones et al. (1998) and Hayes et al. (2000) utilized a BBN structure to make inferences from data observations to a commander's information requirements. Franzen (1999) modeled the decision structure of battle damage assessment (BDA) within a BBN. Das et al. (2002) presented an approach to battlefield situation assessment based on the real-time combination of small Bayesian network components to form a BBN for a specific high-level scenario. Therrien (2002) used a BBN to model human and environmental parameters influencing risk assessment and stress in combat scenarios. Information sources included observations, training, orders, and report. Our research utilizes a BBN to integrate sensors and other relevant data to better characterize a system for real-time response. This task is similar to military and other studies which utilized Bayesian approaches to combine diverse sources of data and intelligence in a battlefield setting or other response scenario.

Bayesian Belief Networks

Bayes theorem states that:

$$P(h \mid D) = \frac{P(D \mid h)P(h)}{P(D)}$$

where:

P(h) = prior probability of h with no knowledge of observation D,
P(D) = prior probability of D with no knowledge of h,
P($h|D$) = posterior probability of h after observation D, and
P($D|h$) = probability of observation D given that h is true.

The Bayesian prior probability h is updated to a posterior probability P($h|D$) that reflects an observation D. Conceptually, this updating process mimics the reasoning of people presented with new information about uncertain phenomena. For a classic illustration of this idea, consider a person stranded on a desert island with amnesia. This person observes that the sun sets and wishes to determine the probability that it will rise again using Bayes theorem. The castaway puts a black rock and a white rock into a bag to represent the chances that the sun will rise again or not. Each day that the sun rises, he puts another white rock into the bag. After one day, the probability of sunrise increases from (1 white rock/2 total rocks) = 0.5 to (2 white rocks/3 total rocks) = 0.67. After another observed sunrise on the second day the posterior probability increases to 0.75. The white rocks will continue to accumulate until the probability that the sun will not rise is negligible.

A BBN represents the conditional independence assumptions among a set of variables, thus specifying the joint probability distribution. The BBN is typically presented as a directed acyclic graph with nodes representing variables and arcs representing assertions of conditional independence. A node is conditionally independent of its non-descendents. In Figure 1, variable **d** can be said to be conditionally independent of variable **c,** given **a** and **b**. The joint probability for the assignment of any set of values (x_1,…,x_n) to the set of variables (X_1,…,X_n) in a BBN can be determined by:

$$P(x_1,...,x_n) = \prod_{i=1}^{n} P(x_i \mid Parents(X_i))$$

where, Parents(X_i) is the set of values for preceding nodes in the network. The joint probability of any set of variables can be inferred from observed values or distributions for any subset of remaining variables. There are a number of exact and approximate algorithms that have been developed to infer posterior probabilities for BBNs (Jenson, 1996). This study utilized a generalized variable elimination approach in which posteriors are derived from the marginal probabilities for a set of variables in the BBN. Further details of this algorithm are provided by Cozman (2000).

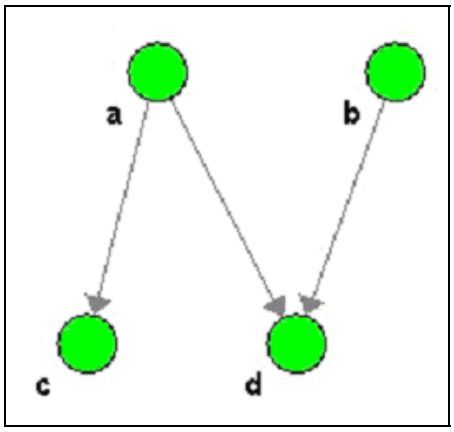

Figure 1: Bayesian belief network

Distribution System Model

The distribution system model used in this study is for a hypothetical campus-type facility that was created for research, development, and demonstration purposes only. The relevant water infrastructure components were extracted from GIS data for the model facility. There was no calibration performed since there were no observed data for the hypothetical system. The demands were estimated based on design guidelines for per capita consumption in the assortment of office, residential, commercial, and other sectors that were present in the model facility. Sensor locations were identified based on a qualitative inspection of the model flow patterns in extended period simulations performed with EPANET (Rossman, 2000). An attempt was made to maximize the sensor's upstream coverage, however, no rigorous mathematical optimization was performed. Others have conducted research into algorithms to determine the optimal placement of sensors (Lee and Deininger, 1992), which is a topic beyond the scope of this study. Each sensor was assumed to be capable of identifying contaminants from the region of the distribution system upstream of the sensor's location. Figures 2, 3, and 4 show a schematic view of the distribution system model and the coverage region of sensors 1, 2, and 3, respectively. The upstream region for sensor 3 contains that of both sensor 1 and sensor 2. Sensors were assumed to provide a simple yes/no indication of the presence of a contaminant with an initial estimated false positive and false negative rate. Sensors that provide continuous values for concentration or other measure could be converted to a discrete state using a threshold value, statistical test, learning algorithm (e.g. artificial neural network), clustering, or other method (ASCE, 2004).

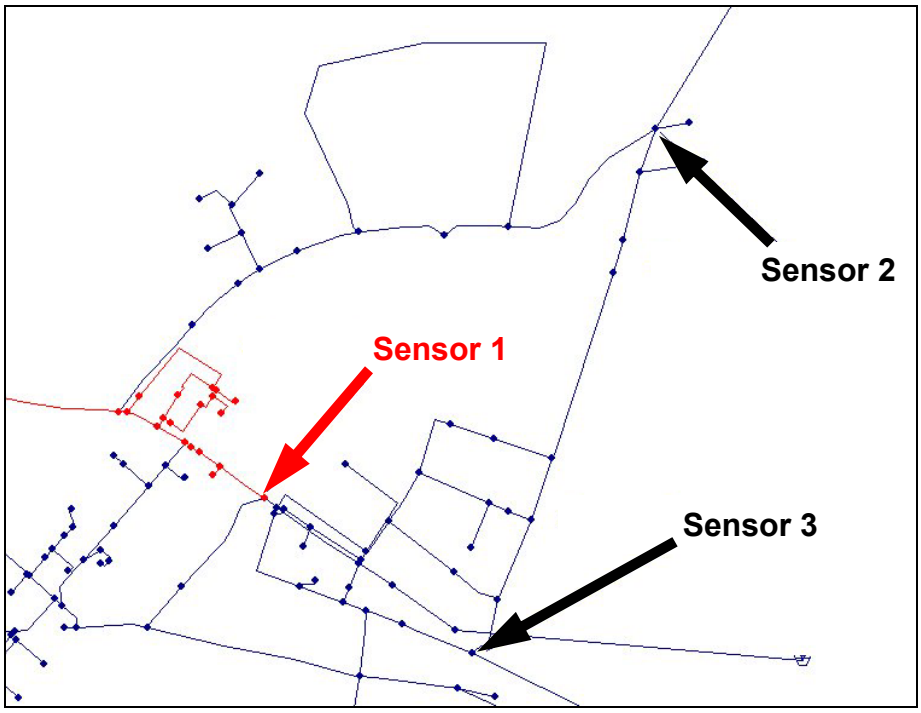

Figure 2: Distribution system model showing coverage region of sensor 1.

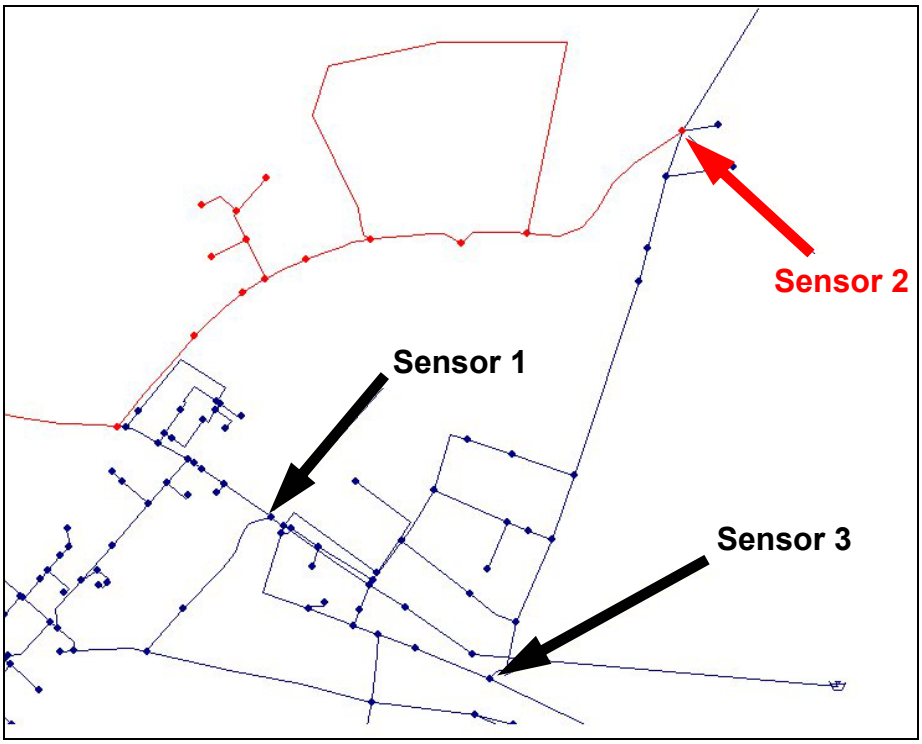

Figure 3: Distribution system model showing coverage region of sensor 2.

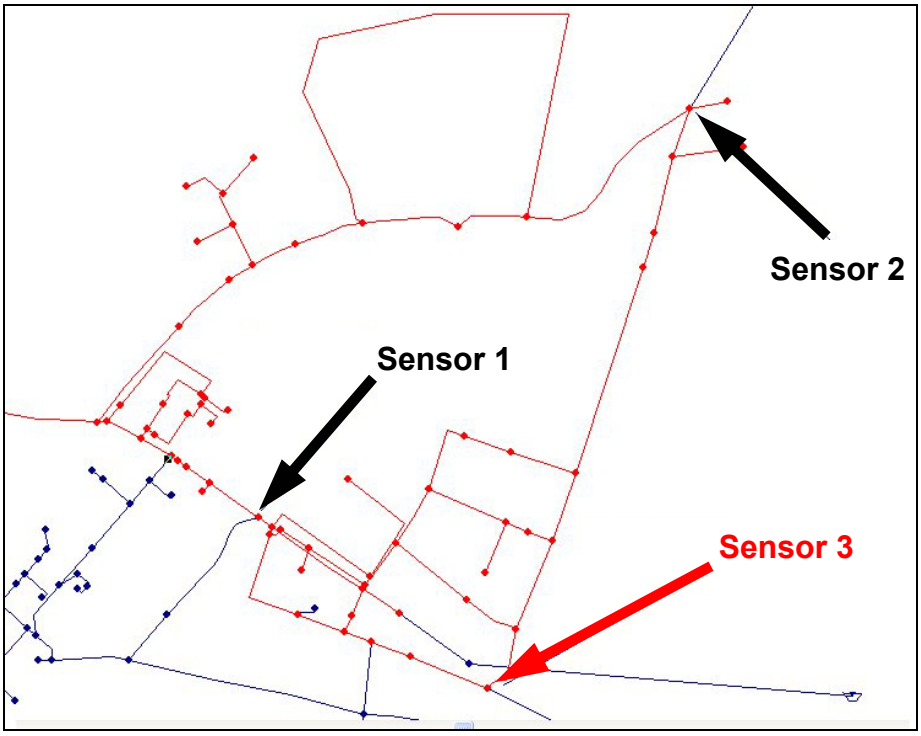

Figure 4: Distribution system model showing coverage region of sensor 3.

A BBN was generated to represent the joint probabilities of a few selected contamination scenarios (Figure 5). The BBN includes nodes representing observable evidence such as physical security alarms, sensors, and operational logs. Other nodes represent events that were not directly observable such as the release of a contaminant or a change in system operation. These are simplifications and abstractions of the relevant components that might exist in an actual distribution system. Additional observable evidence such as specific threat intelligence could easily be incorporated into the BBN.

The initial false positive and false negative rates for each sensor and other observable nodes were estimated, as were the prior probabilities of a contaminant release and other event nodes. The BBN can be recursively updated during the implementation period to reflect further refinement of some of these initial estimates. However, it may be difficult to estimate the probability of very low occurrence events such as an intentional contaminant release. For these cases it may be necessary to make an estimate of the prior probability based upon expert judgment.

This BBN was used as a framework for expressing the complex causal relationships and conditional probabilities that comprise contamination scenarios. The scenarios themselves are the product of collaboration among people that may be involved in vulnerability assessment such as utility managers, engineers, and others with detailed knowledge of a water system. Prior probabilities of infrequent events simply reflect these expert's judgments.

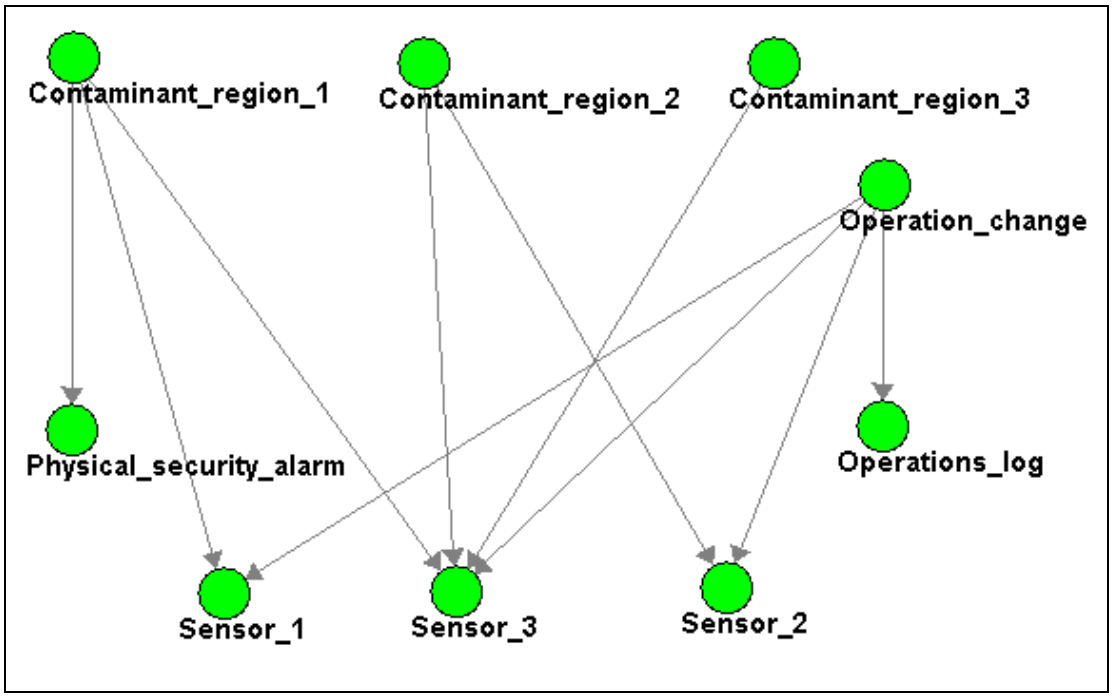

Figure 5: Bayesian belief network for distribution system contamination

Each node of the BBN utilizes a table containing conditional probabilities of discrete Boolean states given the state of that node's parents, or $P(x_i \mid Parents(X_i))$. For top level nodes, this table is simple the prior probability of that event, $P(x_i)$ and the prior probability of *not* that event, $P(\neg x_i)$. Table 1 shows the matrix of probabilities estimated for S*ensor_1* and its parent nodes, *Contaminant_region_1* and *Operation_change*. *Contaminant_region_1* refers to the introduction of a contaminant into the region that is covered by *Sensor_1*. *Operation_change* refers to actions such as system flushing, booster pump activation, valve maintenance, or others that might cause false sensor detections. This node could be subdivided and refined during implementation to reflect further knowledge of system operational characteristics. In Table 1, *Sensor_1* refers to a positive sensor detection, and ¬*Sensor_1* refers to a negative sensor detection. The number of combinations of parent node states is $2^{(i)}$, where i = number of parent nodes. The false positive and false negative rates for the sensor were estimated to be 0.05 and 0.03, respectively. The prior probability of contaminant release was arbitrarily estimated to be 0.0001, and the prior probability of a change in operation was 0.1.

Table 1: Matrix of prior probabilities for sensor 1 and its parent nodes

	Cont. Region 1 Oper. Change	**Cont. Region 1 ¬Oper. Change**	**¬Cont. Region 1 Oper. Change**	**¬Cont. Region 1 ¬Oper. Change**
Sensor 1	0.985	0.97	0.5	0.05
¬Sensor 1	0.015	0.03	0.5	0.95

Results and Discussion

The BBN was used to explore hypothetical combinations of sensor detections and other evidence that might occur in a contamination event. The posterior probability that a contaminant had been introduced was inferred from changes to observable nodes in the BBN. When a node was 'observed', its value became fixed, and the probability of that node's parent(s) was updated to reflect the new observation. For example, when *Operations_log* was observed to be 'true', then the prior probability of O*peration_change* was updated from 0.1 to a posterior value of 0.917 representing the relative certainty that the cause of the log entry was in fact a change in system operation. When *Sensor_3* was observed to be 'true', then the posterior probability of *Operation_change* was 0.525 and the probabilities of a contaminant in region 1, 2, or 3 each increased to only 0.00102. This result is somewhat counterintuitive, since it seems to indicate that S*ensor_3* has little value for contaminant detection. It illustrates the statistical impact of relatively high false positive rates and the importance of prior probabilities in Bayes theorem. Even an apparently low false positive rate of 0.03 is in this case drastically higher than the prior probability of a contaminant release (0.0001). In such imbalanced cases, a positive detection by a sensor is much more likely to be caused by something other than contamination. This effect would occur regardless of the prior probability as long as it is relatively insignificant in comparison to the false positive rate.

Figure 6 shows a detection scenario in which other evidence validated the positive result at *Sensor_3*. In this case, the probability of *Contaminant_region_1* increased from 0.0001 to 0.0627. Again, the result appears to be counterintuitive since the probability remains relatively low despite overwhelming evidence. However, the probability increased by a factor of 627 from its prior value. When the prior probability for *Contaminant_region_1* was set to 0.001, the resulting posterior value was 0.401. This demonstrates the high sensitivity of posteriors to different estimates of prior probability. The factor by which the two different priors increased, however, was similar. It may, therefore, be more informative to consider the *change* in probability for very low frequency events rather than the absolute value.

The spatial and temporal characteristics of the sensor data and distribution system model were not directly expressed in the BBN. This approach does not, for example, provide information regarding the exact location within an upstream region that a contaminant was released. The travel time between sensors or time between physical security alarms and sensor detections was also not accounted for in this approach. The timing of these observations would be relevant to the probability that they are evidence of the same causal event. This approach is also limited by the ability of experts to imagine contamination scenarios. The possibility would always exist that a terrorist could attack a water system in an unpredictable way that would not be characterized accurately by the BBN. During implementation, observed evidence would likely be augmented by data from field testing kits that could be deployed to a region of the distribution system. The evidence from these tests could be integrated into the BBN simply by adding additional observation nodes.

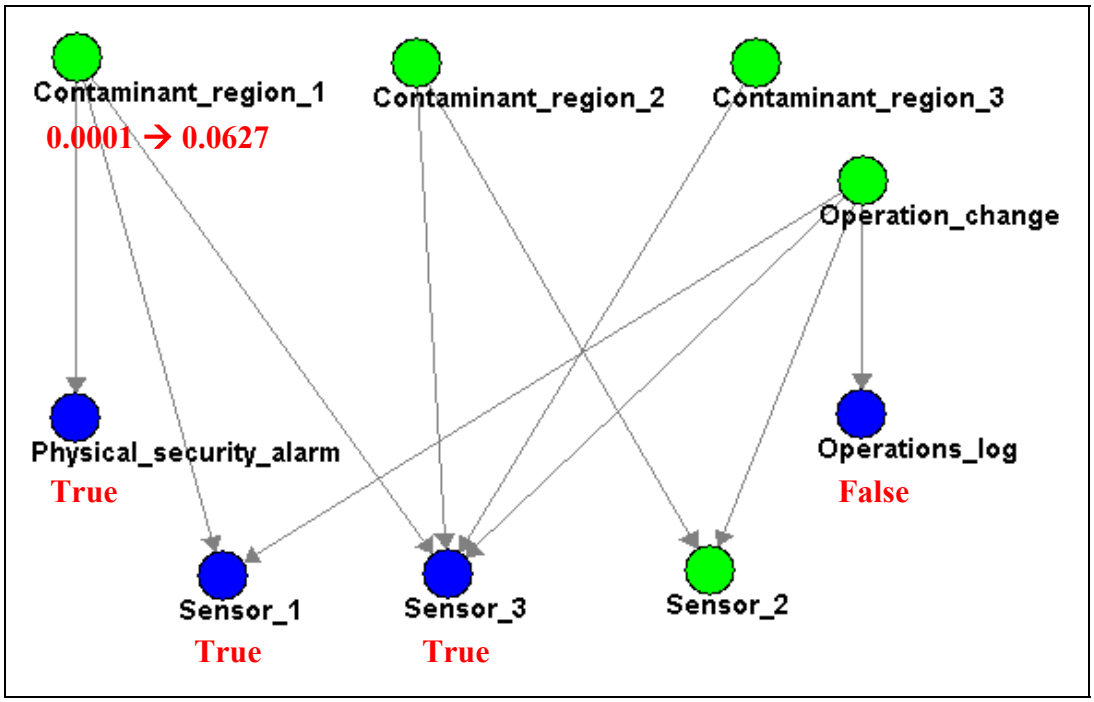

Figure 6: Updated posterior probability for contamination in region 1 reflecting system observations.

Conclusions

This study demonstrated that BBNs are capable of expressing complex causal relationships among the events and observations that comprise contamination scenarios in water distribution systems. These scenarios can be better understood when explicitly visualized in a graphical probabilistic model such as a BBN. This approach has the potential to be incorporated into both security planning and real time response to contaminant detection.

Acknowledgements

This research was supported by the Office of Naval Research grant number N00014-04-1-0437 through the Technology Research, Education and Commercialization Center (TRECC). The authors would like to acknowledge Tim Perkins of the Army Corps of Engineers Construction Engineering Research Laboratory (CERL) for his efforts in creating the distribution system model used in this study. We would also like to acknowledge Fabio Cozman, author of the Bayesian belief network modeling software JavaBayes (http://www-2.cs.cmu.edu/~javabayes/).

References

American Society of Civil Engineers (ASCE), Interim Voluntary Guidelines for Designing an Online Contaminant Monitoring System, Reston VA, 2004.

Beckerman, M., A Bayes-maximum entropy method for multi-sensor data fusion, *Proceedings - IEEE International Conference on Robotics and Automation*, v 2, pp. 1668-1674, 1992.

Bonci, A., G. Di Francesco, and S. Longhi, A Bayesian approach to the hough transform for video and ultrasonic data fusion in mobile robot navigation, *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics*, v 3, pp. 350-355, 2002.

Brown, C., M. Marengoni, G. Kardaras, Bayes nets for selective perception and data fusion, *Proceedings of SPIE - The International Society for Optical Engineering*, v 2368, pp. 117-127, 1995.

Chu, T. L., Estimation of initiating event distribution at nuclear power plants by Bayesian procedure, *ISSAT international conference on reliability and quality in design (2nd),*Orlando, Florida, 1995.

Coue, C., T. Fraichard, P. Bessikre and E. Mazer, Using Bayesian programming for multi-sensor mu1ti-targe t tracking in automotive applications, *Proceedings of the 2003 IEEE International Conference on Robotics & Automation*, IEEE, pp. 2104 – 2109, 2003

Cozman, F. G., Generalizing variable elimination in Bayesian networks, *Workshop on Probabilistic Reasoning in Artificial Intelligence*, Atibaia, Brazil, 2000
http://www-2.cs.cmu.edu/~javabayes/Home/

Das, S., R. Grey, and P. Gonsalves, Situation assessment via Bayesian belief networks, *Proceedings of the Fifth International Conference Information Fusion*, pp. 664-671, v. 1, 2002.

Fox, D., J. Hightower, L. Liao, D. Schulz, and G. Borriello, Bayesian Filtering for Location Estimation, *Pervasive Computing*, Pervasive Computing, IEEE, pp: 24- 33, v. 2(3), 2003.

Franzen, D. W., Bayesian Decision Model for Battle Damage Assessment, Master's thesis, Air Force Institute of Technology, Air University, Air Education and Training Command, 1999.

Hayes, C., R. Penner, H. Ergan, L. Lu, N. Tu, P. Jones, P. Asaro, R. Bargar, O. Chernyshenko, I. Choi, N. Danner, O. Mengshoel, J. Sniezek, D. Wilkins, CoRaven: model-based design of a cognitive tool for real-time intelligence monitoring and

analysis, *Systems, Man, and Cybernetics 2000 IEEE International Conference,*pp. 1117-1122, v. 2, 2000.

Jenson, F. V., *An Introduction to Bayesian Networks*, New York, Springer Verlag, 1996.

Jones, P. M., C. C. Hayes, D. C. Wilkins, R. Bargar, J. Sniezek, P. Asaro, O. Mengshoel, D. Kessler, M. Lucenti, I. Choi, N. Tu, and MAJ. J. Schlabach, CoRAVEN: modeling and design of a multimedia intelligent infrastructure for collaborative intelligence analysis, *Systems, Man, and Cybernetics 1998 IEEE International Conference,* pp. 914-919, v. 1, 1998.

Karlsson, B., J. Jan-Ove, P. Wide, A fusion toolbox for sensor data fusion in industrial recycling, *IEEE Transactions on Instrumentation and Measurement*, v 51(1), PP. 144-149, 2002.

Lee, B. H. and R. A. Deininger, Optimal Locations of Monitoring Stations in a Water Distribution System, *Journal of Environmental Engineering*, v. 118:4.

Liu, E., and D. Zhang, Diagnosis of component failures in the space shuttle main engines using Bayesian Belief Networks: a feasibility study, *Proceedings of the 14th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'02)*, pp. 181-188, 2002.

Marcot, B. G., Using Bayesian belief networks to evaluate fish and wildlife population viability under land management alternatives from an environmental impact statement, *Forest Ecology and Management*, v 153 (1-3), n 1-3, pp. 29-42, 2001.

Rossman, L. A., EPANET2 Users Manual, United States Environmental Protection Agency, 2000. Available at http://www.epa.gov/ORD/NRMRL/wswrd/epanet.html

Sanzotta, M. A. and E. T. Sherrill, Approximation Probability of Detection in the Janus Model, Technical report, United States Military Academy, West Point, New York, 1997.

Schlumberger, Y., J. Pompee, and M. De Pasquale, Updating operating rules against voltage collapse using new probabilistic techniques, *Transmission and Distribution Conference and Exhibition 2002: Asia Pacific*, pp 1139-1144, IEE/PES, v. 2, 2002.

Siber, N. A., M. Pantazidou, and M. J. Small, Expert system methodology for evaluating reductive dechlorination at TCE sites, *Environmental Science and Technology*, v 33(17), pp. 3012-3020, 1999.

Stow, C. A., C. Roessler, M. E. Borsuk, J. D. Bowen, and K. H. Reckhow, Comparison of Estuarine water quality models for total maximum daily load

development in Neuse River Estuary, *Journal of Water Resources Planning and Management*, v 129 (4), pp. 307-314, 2003.

Therrien, S. S., Bayesian Model to Incorporate Human Factors in Commanders' Decision Making, Master's thesis, USNAVY Postgraduate School, Monterey, California, 2002.