Evaluating uncertainty representation and reasoning in HLF systems

Paulo Cesar G. Costa, Rommel N. Carvalho, Kathryn B. Laskey, Cheol Young Park
Center of Excellence in C4I / The Sensor Fusion Lab
George Mason University, MS 4B5
Fairfax, VA 22030-4444 U.S.A.
Email: pcosta@c4i.gmu.edu, rommel.carvalho@gmail.com, klaskey@c4i.gmu.edu, ypark@c4i.gmu.edu

Abstract—High-level fusion of hard and soft information from diverse sensor types still depends heavily on human cognition. This results in a scalability conundrum that current technologies are incapable of solving. Although there is widespread acknowledgement that an HLF framework must support automated knowledge representation and reasoning with uncertainty, there is no consensus on the most appropriate technology to satisfy this requirement. Further, the debate among proponents of the various approaches is laden with miscommunication and ill-supported assumptions, which inhibits advancement of HLF research as a whole. A clearly defined, scientifically rigorous evaluation framework is needed to help information fusion researchers assess the suitability of various approaches and tools to their applications. This paper describes requirements for such a framework and describes a use case in HLF evaluation.

Keywords: Uncertainty reasoning, high-level fusion, evaluation framework, simulation, Bayesian theory, Dempster-Shafer theory, fuzzy theory..

I. INTRODUCTION

Events such as the recent attempted car bombing in Times Square underscore the challenge faced by today's information systems. As in previous such incidents (*e.g.*, the failed suicide bomber onboard an American Airlines flight), many evidence items pointed to a clear threat, yet delays in "connecting the dots" put innocent people in danger. Forensic analyses have the advantage of focusing on a single individual. Real-time prediction requires sifting through a myriad of soft and hard information items, to weed out a handful of malign actors attempting to blend in among millions of innocent people.

Higher-level fusion (HLF) technologies require the use of expressive representational frameworks capable of explicitly representing semantics of the domain. A key requirement for these HLF frameworks is to provide principled support for representation and reasoning with uncertainty, which is ubiquitous to all applications involving knowledge exchange.

High-level fusion of hard and soft information from diverse sensor types still depends heavily on human cognition. This results in a scalability conundrum that current technologies are incapable of solving. A major roadblock to successful automated reasoning with hard and soft information is the lack of a fundamental theory of HLF, backed by a consistent mathematical framework and supporting algorithms. Although there is widespread agreement that an HLF framework must support automated knowledge representation and reasoning

with uncertainty, there is no consensus on the most appropriate technology to satisfy this requirement. Further, the debate on the appropriateness of prominent approaches is laden with miscommunication and ill-supported assumptions, greatly jeopardizing attempts by the community to converge on a fundamental mathematical theory of HLF that: (1) supports representation of semantics and pragmatics, (2) provides a solid mathematical foundation underlying fusion algorithms, and (3) supports scalability of products such as common and user-defined operational pictures.

We argue that a rigorous, scientifically principled evaluation framework[1] is needed to help information fusion researchers to assess which tools are most suitable for use in their applications. Given the broad spectrum of uncertainty reasoning approaches, in this preliminary work we restrict attention to the Bayesian and Dempster-Shafer (D-S) theories, two of the most widely advocated approaches to uncertainty management for HLF. The next section provides some background information on the current state of affairs in terms of uncertainty reasoning techniques to HLF. Section 3 elaborates on the most important aspects we found to be essential when evaluating uncertainty in HLF systems. Then, Section 4 presents our current work in evaluating the PROGNOS system as a case study for a broader framework.

II. UNCERTAINTY IN HIGH-LEVEL FUSION

Military situations are inherently uncertain, and the available data are inevitably noisy and incomplete. The ability to represent and reason with uncertainty is fundamental to fusion systems at all levels. At lower levels of the JDL hierarchy, probabilistic methods have become standard. The Kalman filter and its many variants have their mathematical basis in the theory of stochastic processes. Sensors for detecting various kinds of objects are typically characterized by receiver operating curves that allow false positive and false negative probabilities to be traded off as detection thresholds are adjusted. Classification systems are characterized by confusion matrices, which represent misclassification probabilities for various entity/event types.

---

[1]In this work, we use the term *evaluation framework* to denote the set of requirements, use cases, and evaluation criteria that collectively form a comprehensive, unbiased means to evaluate how well a given approach addresses the needs of HLF applications.

As HLF became a more prominent concern within the community, new demands for complex knowledge representation have sparked a debate over the most appropriate formalism for representing and reasoning with uncertainty. While many different approaches have been proposed for high-level fusion of hard and soft information, the Bayesian and D-S approaches are two of the most prominent.

In addition to its unquestioned applicability to many problems of sensor-level fusion, probability theory has enjoyed practical success in a wide array of applications that involve soft information (*e.g.*, Mcculloh *et al.* in [1] apply probability theory to social network data). Probability theory provides a mathematically coherent framework for combining physical, logical, empirical, and subjective information to perform evidential reasoning under uncertainty. Bayesian inference unifies learning and inference as a foundational theory of belief dynamics. Citing these advantages, some have argued for probability theory as a privileged, uniquely justifiable approach to uncertainty management (*e.g.*, [2]).

Others argue that probability has serious limitations for problems characterized by ignorance or incompleteness of evidence. Dempster-Shafer (D-S) theory has been put forward as a superior approach for such situations (*cf.* [3], [4]). There is a substantial literature, spanning several decades, comparing these approaches (*e.g.* [2], [5]–[9]), but there is still no consensus on their suitability for uncertainty representation and reasoning in general, and for HLF applications in particular. In short, it is clear that the debate on this subject has failed to produce definitive answers to many of the open questions within the HLF community.

Tasks at higher levels of the JDL fusion framework, such as the Level 3 task of predicting threat behavior, require reasoning about complex situations in which entities of different types are related to each other in diverse ways. This is particularly true in asymmetric warfare where the threats are elusive, secretive, and decentralized entities that often appear unconnected and their stealthy behavior is very difficult to predict. Automated methods for reasoning about such complex situations require expressive representation languages that can represent and reason with uncertainty. Recent years have seen rapid advances in the expressive power of probabilistic languages (*e.g.*, [10]–[13]). These expressive languages provide much-enhanced ability to represent and reason about complex situations. Similarly, expressive languages based on fuzzy logic have been developed (*e.g.*, [14], [15]), and Dempster-Shafer belief functions have been applied to merging beliefs about uncertain statements in an expressive language (*e.g.*, [16]).

As these technologies have become more mature, there has been increasing interest in their potential applicability to in many fields of research in which uncertainty plays a major role. Starting in the eighties in the field of artificial intelligence (*e.g.*, [2], [5]–[7]), there has been intense discussion of the suitability of different techniques to address the requirements of various applications. Despite decades of debate, the matter is far from being settled. Unfortunately, at times the debate has provoked misconception and wasted effort as much as it has provided enlightenment. This state of affairs inhibits advancement of the community as a whole. A credible, unbiased framework is needed to support HLF researchers in finding the most appropriate methods and techniques to apply to their problems. This paper is intended to provide the basis for such a framework.

III. EVALUATING PROBABILISTIC MODELS

The continuing stalemate in the long-standing debate on the most appropriate uncertainty management formalism for HLF applications can be traced to the lack of a comprehensive evaluation framework to support meaningful comparisons. More specifically, much of the debate has been at a philosophical level, focused on artificially simple problems. There has been little testing that explores behavior of the different formalisms on complex problems exhibiting characteristics fundamental to today's high-level fusion systems.

As an example of the lack of effective comparison, a common criticism of probability theory is its purported inability to deal with novelty and surprise. Situations outside those explicitly incorporated in a probabilistic model, it is said, are assigned probability zero and treated as impossible, thus preventing a probabilistic model from adjusting to surprising outcomes. A common response to this criticism is that it is true of classical probability in which the space of possibilities is given a priori. In more sophisticated approaches, the sample space is flexible and expandable to accommodate surprises. In such approaches, probabilistic models are constructed from modular components. In response to surprises, one should look for a set of models similar in some way to the surprising situation, extract relevant modules, and combine into a new model that accounts for the surprise [17]. Both the criticism and its associated rebuttal have merits, and the matter cannot be settled through philosophical debate alone. A step towards a definitive solution must involve a well-defined requirement (*i.e.*, the ability to adapt to surprises), use cases in which the requirement is testable, and a set of evaluation criteria that comprehensively and fairly evaluates the degree to which a solution meets the requirement. In the context of the above example, questions to be addressed by our research include whether such adaptive approaches can adequately deal with surprise, in what HLF situations this is important, and how to measure the efficacy of a given approach.

To summarize, there is no shortage of comparisons between major approaches for representing and reasoning with uncertainty. However, from the standpoint of the HLF community there are enough questions left unanswered to make it difficult for a researcher or practitioner to choose an approach best suited to his/her problem. This situation not only causes a waste of resources in various steps of the research and development processes but also creates an environment of myths, half-truths, and misconceptions that causes segmentation in the field and jeopardizes the advancement of the community as a whole. This paper aims to be a first step in changing this state of affairs. If this debate has a conclusion, then

reaching it must be though an approach that leverages the existing body of research on technologies for uncertainty representation and reasoning, the high level of interest on the subject within the HLF community, and on expertise in the fundamentals of uncertainty management, to produce a comprehensive, unbiased evaluation framework. A successful outcome will not only provide important insights into uncertainty management for HLF, but will also produce a valuable tool for the community to support future comparisons. The following subsections convey some ideas for building such a framework.

A. The Evaluation Design Process

The high-level overview of the approach for HLF evaluation we propose involves:

1) Performing an in-depth analysis of the major requirements for representing and reasoning with uncertainty from the HLF perspective;
2) Developing a set of use cases with enough complexity to cover the identified requirements;
3) Defining a comprehensive set of criteria to evaluate how well a given methodology addresses the representational and reasoning needs of each use case; and
4) Conducting an evaluation of the major uncertainty management approaches that could be used to address the use cases.

Further, we propose to leverage the path followed by the W3C group [18] and start this effort with the definition of an ontology for HLF systems, which would provide support to the steps defined above.

A key methodological component of this process is the design of use cases that exemplify complex situations involving difficult aspects of uncertainty representation and reasoning, especially those that any HLF system must be capable of handling. The use cases must be derived from, and support identification of, requirements that any HLF ought to address. The evaluation plan must also include development of metrics to evaluate performance of the uncertainty management methods on the use case scenarios.

Given the multidisciplinary nature and broad spectrum of skills needed to perform this process, the embedded analysis must address not only the fundamental issues of representing and reasoning with uncertainty, but also pragmatic topics germane to knowledge exchange problems. These latter topics include consistency, accuracy and scalability. Although the core subject of the above process is to establish the requirements that will guide the development of an evaluation framework for uncertainty representation and reasoning in HLF systems, an important part is to perform an analysis of the major uncertainty representation approaches. This step should provide a means to further ensure a solution that encompasses the required level of complexity needed to arrive at meaningful conclusions, while keeping the evaluation criteria as fair and unbiased as possible. Subsections III-B, III-C, and III-D explore the main points for evaluating an "uncertainty-friendly"

HLF system according to our experience with the case study depicted in Section IV.

B. Expressive Power

A key attribute for a HLF system to operate with uncertainty is its ability to capture knowledge patterns and statistical regularities of the situations it will encounter during operations. This stored domain knowledge is the basis for the inferential process, which supports essential capabilities such as predictive analysis and situational awareness. Therefore, the knowledge representation technology employed by a given HLF method is not only a major aspect supporting its feature set, but also an obvious candidate parameter for evaluating its suitability to meet operational requirements. Vintage information fusion systems have usually relied on proprietary database schemas and other rudimentary means of knowledge representation, which were well suited to store information about domain entities, their status, and other parameters of interest. These were adequate for JDL levels 0 and 1, but as the trend towards higher levels of the JDL spectrum became clear their limitations have become increasingly apparent.

Ontologies are the current paradigm for knowledge representation in HLF systems, providing a means to capture domain knowledge not only about about entities and their respective status, but also domain semantics (concepts, relationships, etc). Further, explicit representation of formal semantics such as Description Logics [19] enables automated reasoning, another requirement for current HLF systems. However, as we have noted elsewhere [20], ontologies lack standardized support for uncertainty representation and reasoning, a liability for systems that must operate in environments plagued with incomplete, ambiguous, and dissonant knowledge. Clearly, most HLF applications fall into this category, and uncertainty management is vital to their success.

The lack of a standard, principled representation of uncertainty was already recognized by the W3C, which has created an incubator group with the following assignment (*ipsis literis* from [18]):

1) To identify and describe situations on the scale of the World Wide Web for which uncertainty reasoning would significantly increase the potential for extracting useful information;
2) To identify methodologies that can be applied to these situations and the fundamentals of a standardized representation that could serve as the basis for information exchange necessary for these methodologies to be effectively used.

This realization of the need to a principled means to represent and reason with uncertainty is hardly surprising, considering the strong parallels between the demands for information exchange found in the Semantic Web and those needed for HLF systems. Even more interesting than this common realization is the charter's explicit determination that "The scope does not include recommending a single methodology but to investigate whether standard representations of uncertainty can be identified that will support requirements across a wide spectrum of

reasoning approaches." [18]. This resonates both in spirit and intention to the approach being proposed in this paper, since it is within the best interests of the HLF community to better understand its requirements for uncertainty representation and reasoning than to engage in acrimonious debate over which method is superior.

C. Solid Mathematical Foundation

Ontology-based systems usually have their logical basis in some variation of classical logic. HLF systems, as argued above, operate in environments fraught with uncertainty and thus must be capable of performing automated reasoning with incomplete knowledge. Designing evaluation criteria for the mathematical foundation of a given approach might seem a straightforward task, given the various computational tools available to support it. However, there is a hidden danger that is often overlooked in most evaluations or debates on the consistency of a model's mathematical support, namely its degree of plausibility and realism. In other words, there is too often a tendency to evaluate whether the math behind a model or an approach is consistent, while spending little time in verifying whether its equations and associated outcome properly reflect the phenomena being modeled. To ensure the latter is addressed, it is important to devise an evaluation plan that draws on expert knowledge (*e.g.*, SME participation) to assess how well the system's behavior matches the domain being modeled.

D. Performance and Scalability

Performing multi-source data merging from a huge body of data demands not only that knowledge be aggregated consistently, but also that inference is scalable. Performing plausible reasoning in support of complex decisions typically requires examining numerous entities with many dependent indicators. These include occurrence of related activities, pointers to a common plan of action, similar spatio-temporal evolution within a situation map, etc. This can lead to exponential growth in the number of hypotheses, to the point at which considering all of them explicitly would be intractable.

HLF systems apply various techniques to address this issue, such as approximation algorithms, hypothesis management, pre-compiled libraries, and others. It is a challenge to design a well balanced combination of these techniques in a way that preserves the accuracy of the model (*cf.* III-C) while ensuring that the system can handle the computational demands of the reasoning process. Although evaluating this aspect seems to be relatively easy to do in an objective way, there are potential traps in the evaluation design. Different uncertainty approaches will have distinct computational costs and will react differently to each combination of optimization techniques, which implies the possibility that an evaluation plan could overemphasize aspects in which a given technique excels over the others. The key for an efficient evaluation is to ensure that all aspects being analyzed closely reflect the average demand expected for the system's daily usage, which stresses the importance of having use cases that represent such usage in a comprehensive

and balanced fashion. Care also must be taken to represent rare but high-impact use cases, which might be overlooked in an "average case" analysis, but for which good performance is essential.

The case study presented in Section IV below depicts the process used to evaluate a HLF system. Although all the aspects cited above were taken into consideration, this case study should not be considered as a model for an evaluation framework or even a benchmark against it others should be compared. Instead, the issues raised there and the respective approaches used to address it should be considered as an illustration to facilitate the discussion on the subject.

IV. CASE STUDY: EVALUATING PROGNOS

PROGNOS [21], [22] (**PR**obabilistic **O**ntolo**G**ies for **N**et-centric **O**peration **S**ystems) is a naval predictive situational awareness system devised to work within the context of U.S. Navy's FORCENet. The system uses the UnBBayes-MEBN framework [23], which implements a MEBN reasoner capable of saving MTheories in PR-OWL format.
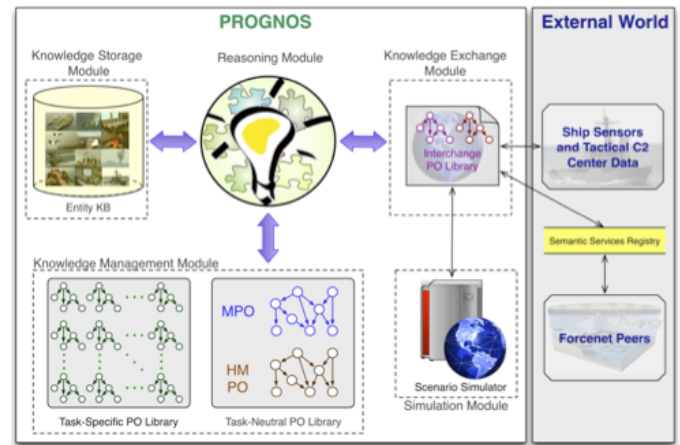

Figure 1. PROGNOS system architecture.

Figure 1 shows the high level architecture of the PROGNOS system, which was designed with five independent modules to provide a scalable, easily maintainable system. For more information on each of these modules, see [22].

According to this architecture, each FORCEnet platform (*e.g.*, a ship) would have its own system that receives information from the platform's sensors and from its FORCEnet peers. It is assumed that these inputs provide a fairly precise tactical view in which the geographical position of the entities surrounding the platform is known and well discriminated. The platform is also a peer in FORCEnet and exchanges data and knowledge as services with its peers.

One of the major challenges in systems like PROGNOS is evaluating the situational awareness and prediction generated by its probabilistic model.

The literature in Machine Learning, Artificial Intelligence, and Data Mining usually work with real data by separating it into training and testing sets. However, in systems that try to

predict rare events, such as terrorist attacks, either there is not enough data or the data available is classified. Therefore, in these cases, there is not sufficient data to be used for testing.

To overcome this limitation, a common approach is to create different use cases or scenarios manually. This use case generation process is discussed in Section IV-A. However, this is a tedious process and usually not statistically sufficient to confidently assess how good these probabilistic models are. In Subsection IV-B we present a framework that can be used for automatically creating different and random, yet consistent, scenarios to provide sufficient statistical data for testing. It is to be stressed, however, that this testing is only as good as the use cases incorporated into the testing process, and there is no substitute for real-world evaluation.

We will use the probabilistic model from Carvalho *et al.* [24] to illustrate the manual and automatic testing solutions presented in Section IV-A and Section IV-B, respectively.

A. Creating scenarios manually

In the first iteration of the probabilistic model presented in Carvalho *et al* [24] the main goal is to identify if a ship is of interest, *i.e.*, if the ship seems to be suspicious in any way. The assumption in this model is that a ship is of interest if and only if there is at least one terrorist crew member.

To test the model created for this iteration and to show that it provides consistent results, we created six different scenarios by increasing not only the complexity of the generated model, but also the probability that a ship, `ship1`, is of interest. These increases are due to new evidence that is available in every new scenario, which supports the hypothesis that `ship1` is of interest. For more information on the types of evidence that support this hypothesis, see [24].

In scenario 1, the only information available is that `person1` is a crew member of `ship1` and that `person1` is related to at least one terrorist. Figure 2 shows that there is a 70.03% probability of `ship1` being of interest, which is consistent with the fact that one of its crew members might be a terrorist.

In scenario 2, besides having the information available from scenario 1, it is also known that `ship1` met `ship2`. Now the probability of `ship1` being of interest has increased to 89.41%, which is consistent with the new supporting evidence that `ship1` met `ship2`.

In scenario 3, besides having the information available from scenario 2, it is also known that `ship1` has an unusual route. Now the probability of `ship1` being of interest has increased to 97.19%, which is consistent with the new supporting evidence that `ship1` is not going to its destination using a normal route.

In scenario 4, besides having the information available from scenario 3, it is also known that `navyShip` has detected an ECM. Figure 3 shows the probability of `ship1` being of interest has increased to 99.97%, which is consistent with the new supporting evidence that `ship1` is probably the ship that deployed the ECM. It is important to notice that there are only two ships that could deploy the ECM in this scenario, which
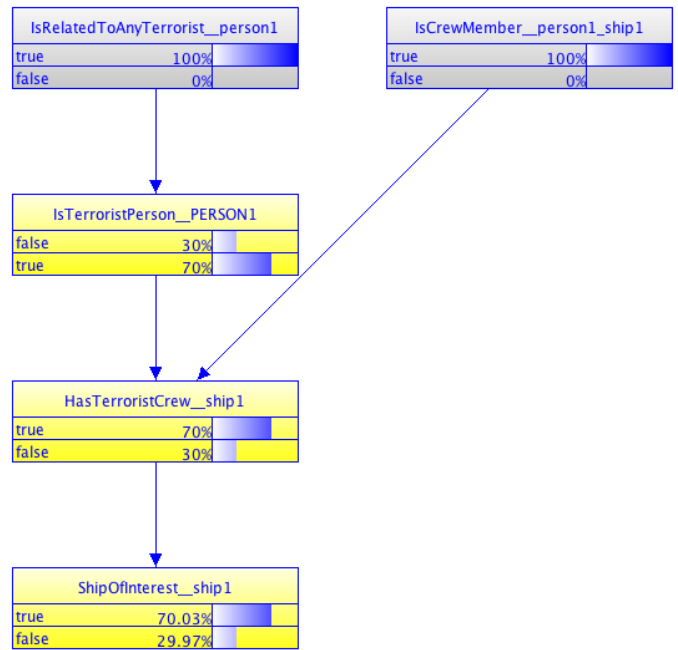

Figure 2. SSBN generated for scenario 1.

are the ships within range of `navyShip`'s radar (`ship1` and `ship2`). From the other evidence that supports the fact that `ship1` is most likely a ship of interest, it becomes more likely that `ship1` is the one that deployed the ECM. That is why the probability that `ship2` having deployed the ECM is so low (due to explaining away).

In scenario 5, besides having the information available from scenario 4, it is also known that `ship1` has neither a responsive radio nor a responsive AIS. Figure 4 shows that the probability of `ship1` being of interest is still high, though a bit smaller then in scenario 4, 99.88%. This is probably due to the fact that having these non-responsive electronics could be related to the fact that the electronics are not working (non-evidence RV). If, by any chance, the electronics were not working, this would make the probability of `ship1` having an evasive behavior decrease, due to explaining away. However, it is reasonable to assume that if it was known that the electronics were working, then the probability of `ship1` having an evasive behavior would increase in comparison to scenario 4, which then would make the probability of `ship1` being of interest increase even more when compared to scenario 4. In fact, if we enter this evidence, scenario 6, the probability of `ship1` being of interest increases to 100%.

At the end of every iteration of the probabilistic model presented in Carvalho *et al.* [24], a set of scenarios like the ones presented above was defined to test whether the posterior probabilities seemed to make sense with respect to what is expected by the subject matter experts (SMEs) who provided expertise to support the project.

The following iterations of the probabilistic model presented in Carvalho *et al.* [24] provides clarification on the reasons behind declaring a ship as being of interest and

IsCrewMember__person1_ship1
false	0%
true	100%

IsRelatedToAnyTerrorist__person1
false	0%
true	100%

IsCrewMember__PERSON1_SHIP2
false	99.28%
true	0.72%

IsTerroristPerson__PERSON1
false	0.07%
true	99.93%

HasTerroristCrew__ship1
true	99.93%
false	0.07%

HasTerroristCrew__SHIP2
true	0.71%
false	99.29%

ShipOfInterest__ship1
true	99.97%
false	0.03%

ShipOfInterest__SHIP2
true	0.86%
false	99.14%

AreMeeting__ship1_NAVYSHIP
false	24.86%
true	75.14%

AreMeeting__ship1_SHIP2
false	6.72%
true	93.28%

EvasiveBehavior__SHIP1
true	99.66%
false	0.34%

EvasiveBehavior__SHIP2
true	0.87%
false	99.13%

UnusualRoute__ship1
true	97.5%
false	2.5%

AreMeetingReport__ship1_ship2
false	0%
true	100%

DeployedECM__SHIP1
false	0.5%
true	99.5%

DeployedECM__SHIP2
false	99.08%
true	0.92%

UnusualRouteReport__ship1
false	0%
true	100%

DetectedECM__navyShip
false	0%
true	100%

Figure 3.   SSBN generated for scenario 4.

detects an individual crew member's terrorist affiliation given his close relations, group associations, communications, and background influences.

To test this final probabilistic model, we created 4 major scenarios:

1) a possible bomb plan using fishing ship;
2) a possible bomb plan using merchant ship;
3) a possible exchange illicit cargo using fishing ship;
4) a possible exchange illicit cargo using merchant ship.

For each of these major scenarios we created 5 variations:

1) "sure" positive;
2) "looks" positive;
3) unsure;
4) "looks" negative;
5) "sure" negative.

All 20 different scenarios were analysed by the SMEs and were evaluated as reasonable results (what was expected).

B. Creating scenarios automatically

Besides being a tedious process, there are a few problems with the manual creation of scenarios as presented in Subsection IV-A. In the first set of scenarios created for the first iteration, it is clear that the test designers just tested how well the model behaves when all the evidence is in favor of the hypotheses being tested. However, how will the model behave if we receive evidence both in favor of and against the hypotheses being tested? Is it still a good model in these cases?

In fact, this is a problem that the last set of scenarios presented in Subsection IV-A addresses. This, the fact that some evidence favors the hypotheses and some does not, is why there are scenarios where the expected result is "looks" positive/negative or unsure. However, even twenty different scenarios is not enough considering the amount of information that is used as evidence in the final model. Let's clarify by presenting the numbers. In the final model there are more than 20 evidence nodes with at least 2 states each (some have more than 10 states). This gives more than $2^{20} = 1,048,576$ different configurations of evidence. In other words, while we tried to cover different types of scenarios, 20 is still an extremely small number compared with the possible configurations. However, it is unreasonable to think a human being will be able to generate and analyze more than one million different scenarios. For this reason, we created a framework for simulating different scenarios automatically.

There are three basic steps in our simulation framework:

1) Create entities and generate some basic static ground truth for them (*e.g.*, create ships, define their type, and their plan);
2) Generate dynamic data for entities based on their static ground truth data (*e.g.*, if the ship is a fishing ship and it has a normal behavior, it will go from its origin port
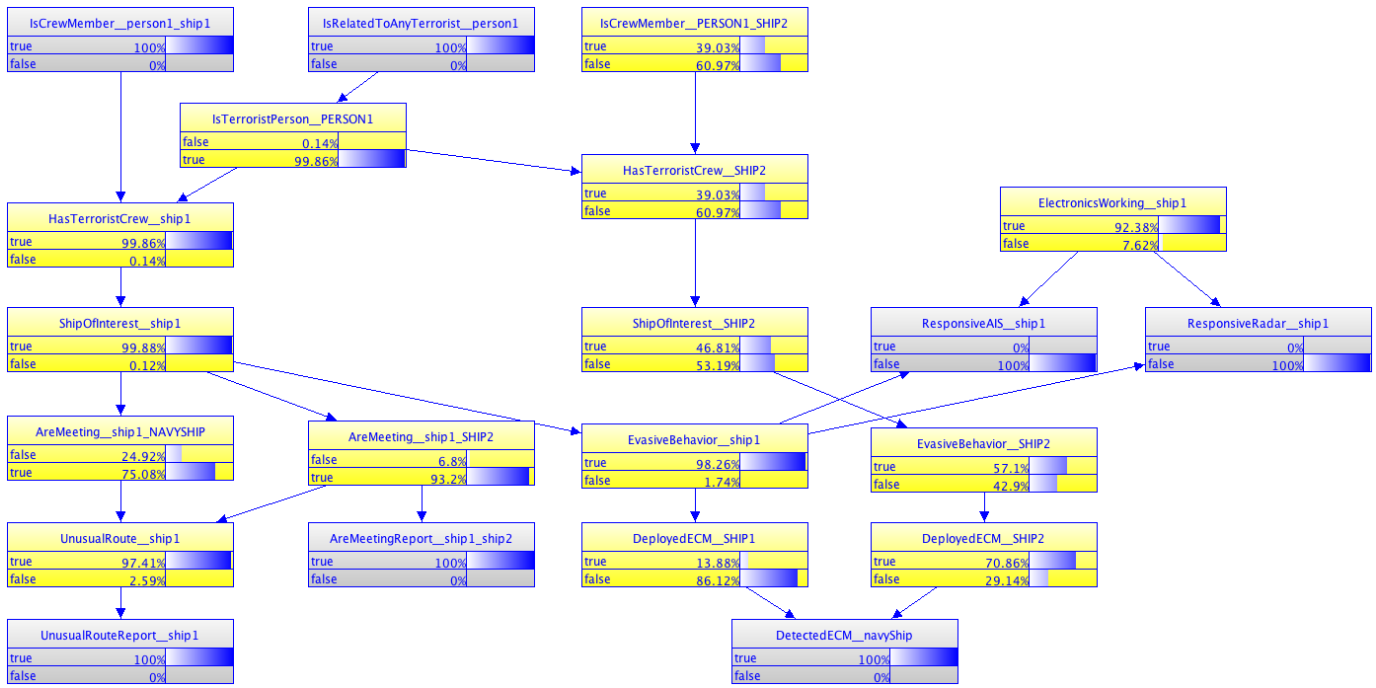
Figure 4. SSBN generated for scenario 5.

to its fishing area and after some time it will go to its destination port);

3) Simulate reports for different agencies. Each agency has a probability of generating a correct report (*e.g.*, saying a person is from Egypt when he is actually from Egypt), an incorrect report (*e.g.*, saying a person is not from Egypt when he is in fact from Egypt), and no report at all (*e.g.*, not being able to say where a person is from). The idea is that different agencies are expected to be more accurate in certain types of information than others (*e.g.*, the Navy is expected to have more accurate data on a ship's position than the FBI).

The information generated in the first two steps are considered the ground truth, while the reports generated in the third step is given as input to the probabilistic model, like the one developed for PROGNOS [24]. The probabilistic model can then use these information as evidence to provide situational awareness and prediction after the reasoning process through its posterior probabilities. Once we know what the model "thinks" is more reasonable (*e.g.*, if a ship is of interest), we can ask the simulation for the correct information, *i.e.*, the ground truth with respect to the hypotheses being tested (*e.g.*, if the ship is indeed of interest). We can then evaluate if the model provided a correct result. Since this process is automatic, we can run this evaluation process as many times as we need to and finally compute some metrics (*e.g.*, confusion matrix) to evaluate how well our model performs.

In the case of the PROGNOS evaluation, the subject matter experts who evaluated the use cases also supported the domain knowledge engineering effort. A more rigorous evaluation process would make use of independent subject matter experts who had not been involved in the system design process. These independent evaluators would develop use cases for evaluation and rate the quality of the system's output.
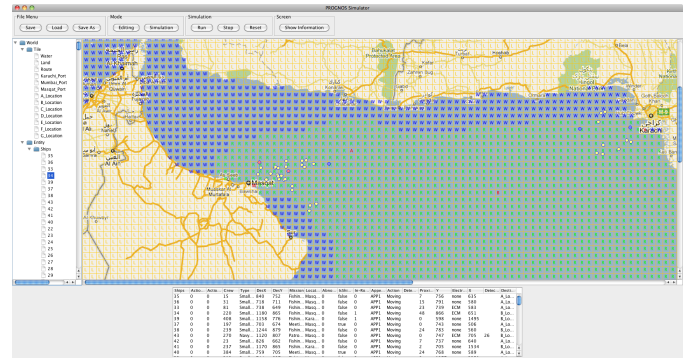

Figure 5. Simulation editor.

However, to be able to compute the three steps described above, we need to define some basic characteristics of the simulation. For instance, what is the geographical region considered, which cells correspond to land and which correspond to land, where are the ports of interest, what are the usual routes between areas of interest, where are the common fishing areas, etc. Figure 5 presents the simulation editor used to define this information.

V. DISCUSSION

This paper focused on the problem of evaluating uncertainty representation and reasoning approaches for HLF systems. The main aspects involved were to establish features required

of any quantitative uncertainty representation for exchanging soft and hard information in a net-centric environment; to develop a set of use cases involving information exchange and fusion requiring sophisticated reasoning and inference under uncertainty; and to define evaluation criteria supporting an unbiased comparison among different approaches applied to the use cases. The process we have described was an abstraction of our own experience with the HLF evaluation case study presented in Section IV. However, the most important message is the need to establish a commonly agreed understanding of the fundamental aspects of uncertainty representation and reasoning that a HLF system must encompass. This can be achieved through an unbiased, in-depth analysis of the system requirements and the ability of each approach to meet the requirements.

REFERENCES

[1] I. A. McCulloh, K. M. Carley, and M. Webb, "Social network monitoring of Al-Qaeda," *Network Science*, vol. 1, no. 1, pp. 25–30, 2007.

[2] P. Cheeseman, "In defense of probability," in *Proceedings of the 9th international joint conference on Artificial intelligence - Volume 2*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1985, pp. 1002–1009. [Online]. Available: http://portal.acm.org/citation.cfm?id=1623611.1623677

[3] V. Sarma and S. Raju, "Multisensor data fusion and decision support for airborne target identification," *Systems, Man and Cybernetics, IEEE Transactions on*, vol. 21, no. 5, pp. 1224–1230, 1991.

[4] R. Murphy, "Dempster-Shafer theory for sensor fusion in autonomous mobile robots," *Robotics and Automation, IEEE Transactions on*, vol. 14, no. 2, pp. 197–206, 1998.

[5] J. Y. Halpern, "Let many flowers bloom: a response to an inquiry into computer understanding," *Computational Intelligence*, vol. 6, no. 3, pp. 184–188, 1990. [Online]. Available: http://portal.acm.org.mutex.gmu.edu/citation.cfm?id=95957

[6] J. Pearl, "Reasoning with belief functions: an analysis of compatibility," *International Journal of Approximate Reasoning*, vol. 4, no. 5-6, pp. 363–389, 1990.

[7] G. M. Provan, "The validity of Dempster-Shafer belief functions," *Int. J. Approx. Reasoning*, vol. 6, pp. 389–399, May 1992. [Online]. Available: http://portal.acm.org/citation.cfm?id=149073.149083

[8] K. Laskey, "Belief in belief functions: An examination of shafer's canonical examples." North-Holland, 1989. [Online]. Available: http://mars.gmu.edu:8080/dspace/handle/1920/1738

[9] D. M. Buede and P. Girardi, "A target identification comparison of bayesian and Dempster-Shafer multisensor fusion," *IEEE Transactions on Systems, Man, and Cybernetics—Part A: Systems and Humans*, vol. 27, no. 5, pp. 569–577, Sep. 1997.

[10] K. B. Laskey, "MEBN: a language for first-order bayesian knowledge bases," *Artif. Intell.*, vol. 172, no. 2-3, pp. 140–178, 2008. [Online]. Available: http://portal.acm.org/citation.cfm?id=1327646

[11] L. Getoor and B. Taskar, *Introduction to Statistical Relational Learning*. The MIT Press, 2007. [Online]. Available: http://portal.acm.org/citation.cfm?id=1296231

[12] P. Domingos, D. Lowd, S. Kok, H. Poon, M. Richardson, and P. Singla, "Just add weights: Markov logic for the semantic web," in *Uncertainty Reasoning for the Semantic Web I*, 2008, pp. 1–25. [Online]. Available: http://dx.doi.org/10.1007/978-3-540-89765-1_1

[13] D. Heckerman, C. Meek, and D. Koller, "Probabilistic models for relational data," Microsoft Research, Redmond, WA, USA, Technical Report MSR-TR-2004-30, 2004.

[14] T. Lukasiewicz, "Probabilistic description logic programs," *Int. J. Approx. Reasoning*, vol. 45, no. 2, pp. 288–307, 2007. [Online]. Available: http://portal.acm.org/citation.cfm?id=1265854

[15] U. Straccia, "A fuzzy description logic for the semantic web," in *Fuzzy logic and the Semantic Web: capturing intelligence*. Elsevier, 2005, pp. 167—181. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.60.2720

[16] M. Nickles and R. Cobos, "An approach to description logic with support for propositional attitudes and belief fusion," in *Uncertainty Reasoning for the Semantic Web I*, P. C. Costa, C. D'Amato, N. Fanizzi, K. B. Laskey, K. J. Laskey, T. Lukasiewicz, M. Nickles, and M. Pool, Eds. Berlin, Heidelberg: Springer-Verlag, 2008, pp. 124–142. [Online]. Available: http://dx.doi.org/10.1007/978-3-540-89765-1_8

[17] R. Haberlin, P. C. G. Costa, and K. B. Laskey, "Hypothesis management in support of inferential reasoning," in *Proceedings of the Fifteenth International Command and Control Research and Technology Symposium*. Santa Monica, CA, USA: CCRP Publications, Jun. 2010.

[18] K. Laskey and K. Laskey, "Uncertainty reasoning for the world wide web: Report on the URW3-XG incubator group," W3C, URW3-XG, 2008. [Online]. Available: http://ite.gmu.edu/~klaskey/papers/URW3_URSW08.pdf

[19] F. Baader, D. Calvanese, D. McGuinness, D. Nardi, and P. Patel-Schneider, *The Description Logic Handbook: Theory, Implementation and Applications*. Cambridge University Press, Mar. 2003.

[20] P. C. G. Costa, "Bayesian semantics for the semantic web," PhD, George Mason University, Jul. 2005, brazilian Air Force. [Online]. Available: http://digilib.gmu.edu:8080/xmlui/handle/1920/455

[21] P. C. G. Costa, K. B. Laskey, and K. Chang, "PROGNOS: applying probabilistic ontologies to distributed predictive situation assessment in naval operations," in *Proceedings of the Fourteenth International Command and Control Research and Technology Conference (ICCRTS 2009)*, Washington, D.C., USA, Jun. 2009. [Online]. Available: http://c4i.gmu.edu/~pcosta/pc_publications.html#2009iccrts

[22] P. Costa, K. Chang, K. Laskey, and R. N. Carvalho, "A multi-disciplinary approach to high level fusion in predictive situational awareness," in *Proceedings of the 12th International Conference on Information Fusion*, Seattle, Washington, USA, Jul. 2009, pp. 248–255.

[23] R. N. Carvalho, "Plausible reasoning in the semantic web using Multi-Entity bayesian networks - MEBN," M.Sc., University of Brasilia, Feb. 2008. [Online]. Available: http://hdl.handle.net/123456789/159

[24] R. N. Carvalho, R. Haberlin, P. C. G. Costa, K. B. Laskey, and K. Chang, "Modeling a probabilistic ontology for maritime domain awareness," in *Proceedings of the 14th International Conference on Information Fusion*, Chicago, US, Jul. 2011.