EURO J Decis Process (2013) 1:205–231
DOI 10.1007/s40070-013-0001-x

ORIGINAL ARTICLE

Structuring and analyzing competing hypotheses with Bayesian networks for intelligence analysis

Christopher W. Karvetski · Kenneth C. Olson · Donald T. Gantz · Glenn A. Cross

Received: 30 July 2012 / Accepted: 5 March 2013 / Published online: 30 April 2013

Abstract Intelligence analysis often tackles questions shrouded by deep uncertainty, such as those that deal with chemical and biological terrorism or nuclear weapon detection. In dealing with such questions, the task falls on intelligence analysts to assemble collected items of information and determine the consistency of the body of reporting with a set of conflicting hypotheses. One popular procedure within the Intelligence Community for distinguishing a hypothesis that is "least inconsistent" with evidence is analysis of competing hypotheses (ACH). Although ACH aims at reducing confirmation bias, as typically implemented, it can fall short in diagramming the relationships between hypotheses and items of evidence, determining where assumptions fit into the modeling framework, and providing a suitable model for "what-if" sensitivity analysis. This paper describes a facilitated process that uses Bayesian networks to (1) provide a clear probabilistic characterization of the uncertainty associated with competing hypotheses, and (2) prioritize information gathering among the remaining unknowns. We illustrate the process using the 1984 Rajneeshee bioterror attack in The Dalles, Oregon, USA.

Keywords Competing hypotheses · Facilitated modeling · Bayesian network · Intelligence analysis · Rajneeshee bioterror attack

Mathematics Subject Classification 62 · 91

C. W. Karvetski (✉) · K. C. Olson · D. T. Gantz
Department of Applied Information Technology,
George Mason University, Fairfax, VA, USA
e-mail: ckarvets@gmu.edu

G. A. Cross
Federal Bureau of Investigation, Washington, DC, USA

Introduction

Overview of research problem

Within the Intelligence Community and other organizations, situations of uncertainty require that hypothesis development and comparison involve analysts who often have different perspectives on a problem. These differences can result from varying levels of training, experience, organizational biases, or access to information. Bringing together analysts allows them to compare background knowledge and observations from multiple sources to form and deliver inferences about competing hypotheses. Consider for example the case of interagency experts assessing the current status of a country's weapons of mass destruction (WMD) program or assessing attribution of a biological or chemical terrorism event. These assessment settings are highly prone to human and group judgment biases,[1] and require creative, cooperative thinking and aggregation of opinions among experts. The assessment could be better informed by a facilitated modeling session in which the assessment model is socially constructed.

Facilitated modeling differs from expert modeling. In facilitated modeling, a set of experts works with facilitators within a general modeling structure to cooperatively build a complete model. In expert modeling, the domain experts provide limited inputs to the modeling experts (Franco and Montibeller 2010). Expert modeling approaches for dealing with the example problem of WMD detection include statistical regression analysis using past economic, political, and other data (Jo and Gartzke 2007; Singh and Way 2004) and systems dynamic modeling with semi-Markov processes linked with Bayesian networks (Caswell and Pate-Cornell 2011). McLaughlin and Pate-Cornell (2005) use a dynamic Bayesian approach for modeling the likelihood of hypotheses concerning Iraq's nuclear program as evidence becomes available. Caswell et al. (2011) use decision analysis methods to model the acquisition of nuclear weapons in Iran. Parnell et al. (2010) use decision analysis for studying bioterror risks with intelligent adversaries.

Expert models are necessary in intelligence analysis, as statistical and other theoretically sophisticated models can inform judgments of analysts, but each analyst must still cooperate with other domain experts to develop an assessment for senior policy makers. In this context, an assessment benefits from facilitated modeling, which is a process by which the formal model is collaboratively developed face-to-face with a group, with or without computer software support (Franco and Montibeller 2010; Eden and Radford 1990). The value of the model is directly proportional to both the quality of the inputs provided by the analysts and the level of participation among analysts.

One well-known structured and facilitated approach in the Intelligence Community for comparing hypotheses is analysis of competing hypotheses (ACH; Heuer 1999). While the level of structured facilitation can vary (e.g., with or without software), the general steps of ACH are designed to allow consideration

[1] The example of the erroneous Intelligence Community judgment on the status of Iraq's WMD programs in the runup to the 2003 US invasion is illustrative of this point (Whitney 2005).

of items of evidence across a set of hypotheses and enable a final assessment of hypotheses in terms of feasibility. The steps include identifying all possible hypotheses, identifying all items of evidence, weighing reliability and relevance for each item, preparing a two-dimensional matrix that catalogs the consistency of evidence with each hypothesis, drawing conclusions about future analyses, and analyzing the influence of uncertainties and future data collection with the aim of reporting to policy makers.

Analysis of competing hypotheses has advantages and disadvantages. Among the advantages is that ACH is easy to implement in a facilitated setting, as the inputs are limited, and ACH software provides a rudimentary display of how evidence relates to a set of mutually exclusive and exhaustive hypotheses (Palo Alto Research Center 2006). This matrix can help filter through large sets of evidence and hypotheses. There is some empirical support that ACH reduces confirmation bias (Lehner et al. 2008), the tendency to interpret evidence in favor of the hypothesis perceived as most likely a priori (Lord et al. 1979).

The deficiencies of ACH are documented (see e.g., van Gelder 2008). The inputs of ACH are not well defined, and the process itself does not safeguard against logic errors related to uncertainty and uncertainty propagation. ACH does not deliver a defensible measure of uncertainty among non-discreditable hypotheses, it cannot consider the confluence of evidence with regard to a hypothesis, and it does not arrive at a usable model for a meaningful sensitivity and what-if analysis. There are purported ways to work around some of the shortcomings, such as treating each meaningful confluence of evidence as an individual piece of evidence, but these come at the cost of simplicity and tractability.

Previous efforts have tried to remedy some of the deficiencies of ACH by using statistical models or logic calculi (Duncan and Wilson 2008; Pope and Josang 2005), but these approaches are difficult to apply within a group of analysts in a facilitated setting. Valtorta et al. (2005) are the first to suggest and argue for the pairing of ACH with Bayesian networks because Bayesian networks offer many features that can remedy the aforementioned shortcomings of ACH.

Bayesian networks provide a visual of the relationships between a set of evidence and a set of hypotheses in the form of a directed graph consisting of arcs and nodes. Each node is a random variable with two or more states. The arcs imply conditional probabilistic dependence between nodes. Uncertainty concerning node states is included in the form of conditional probabilities that form a joint probability distribution over the states (Jensen 2001; Pearl 1988).

Importantly, once the nodes, node states, and arcs are described, there is available research to elicit and aggregate probabilities across analysts through simple aggregation procedures (Clemen and Winkler 1993, 1999). Clemen et al., (2000) find that subjects are able to forecast conditional dependence quite well with just a simple seven-point relationship scale and subjects improve with some training.

Before aggregating probabilities, much consideration is still needed for assembling the Bayesian network from nodes, states, and arcs in a facilitated setting. The facilitated setting constrains the modeling time and requires a repeatable and structured process for constructing the network that ensures the assessments are consistent with how analysts reason about items of evidence.

Organization of paper

In this paper, we describe a structured process for which a group of analysts form a set of evidence and a set of hypotheses and build a Bayesian network with the help of a facilitator. The second section describes the Rajneeshee bioterror case that is used throughout this paper to demonstrate the process. The third section describes in more detail ACH and literature on related methods for this effort. We further elaborate in the third section the reasons why Bayesian networks provide the proper framework for assessing the uncertainty of hypotheses. The fourth section describes step-by-step the primary analysis process with direct application to the 1984 Rajneeshee bioterror attack on The Dalles, Oregon, USA. The fifth section provides discussion and conclusions of the process and the case, and also highlights future and ongoing work.

Rajneeshee bioterror case

Overview of events

This section describes the case that is used throughout this paper to illustrate the process. In September, 1984, in the town of The Dalles, Oregon, the Centers for Disease Control (CDC) was tasked to investigate a *Salmonella* outbreak involving at least eight different restaurants (Deisler 2002; Carus 2000). In total, 751 cases were reported in three waves that lasted several weeks. The reported symptoms of the cases were generally severe fevers, diarrhea, and other discomforts, with some cases requiring hospitalization. The common thread among the infected was that nearly all had eaten at the salad bar of one of the multiple restaurants, and the infected also included workers of each of the restaurants.

The *New York Times* reported in October of 1984 that while "a single, absolute answer to the outbreak's cause might never be found […] the leading hypothesis is ill food handlers […] may have contaminated the raw salad bar items" (Staff 1984). This explanation was similar to the initial conclusion of the CDC. However, some residents were not convinced of this leading hypothesis for the contamination. James Weaver, a congressman from Oregon, believed the Rajneeshee followers of the Indian mystic, known at that time as Bhagwan Shree Rajneesh, were responsible for the outbreak. Weaver presented his argument to the US House of Representatives describing the extreme rarity of an outbreak of such magnitude and its inconsistency with the ill food-handlers hypothesis (Weaver and James 1985). Congressman Weaver concluded that the outbreak was the result of an attack by the Rajneeshees. A formal, expert statistical analysis of the case is also presented by Torok et al. (1997).

Congressman Weaver was correct in his assertion. The Rajneeshees were hoping to expand their community's land holdings, but were receiving pushback from officials in The Dalles. Several Rajneeshees were running for county-level public office, and other followers were trying to infect local voters in The Dalles to prevent

the residents from voting against their candidate in the county elections. The attack remains the largest bioterror attack in US history.

Use of case within paper

The *Salmonella* contamination case of The Dalles is used herein to demonstrate the process described in this paper. The purpose is not to present a retrospective analysis of the investigation by the CDC, but rather to compare the process of this paper with ACH on a real case and to simulate how a group of analysts would find the process of this paper useful. This case is presented as if at the onset of the investigation, the time when information is initially limited and additional information should be sought. This additional information is used to show the value of the model for considering unknowns, assessing the confluence of evidence, and performing sensitivity analysis.

The items of evidence that are available at this point in the case are summarized in Table 1. The set consists of 11 items related to the prior history leading up the contamination, as well as observations that are revealed with investigation. This set of evidence is consistent with how Heuer (1999) broadly defines the set of evidence for ACH as "all the factors that have an impact on […] judgments about the hypotheses."

The evidence for the case includes the case-specific observations that patrons who ate the same food at the banquet rooms as served in the salad bars did not become ill, and the outbreak was confined to just The Dalles.

Table 1 The items of evidence for the Rajneeshee bioterror case

Evidence	Description
No other towns report outbreak	The outbreak was restricted to just restaurants within The Dalles
No prior cases in The Dalles in last 2 years	There are no reported *Salmonella* cases in the past 2 years in The Dalles
Leading causes of *Salmonella* are meat contamination and farm runoff	*Salmonella* is typically caused in the kitchen by improper food handling or at the farm by sewage runoff
Banquet patrons not sick	Restaurant patrons who ate the same food at banquet rooms within the restaurant were not sick
Multiple items contaminated at salad bars	The salad bar item contaminated varied across restaurants
Contamination not easily passed between other foods, people	Contamination is not easily passed; food and people must have contact with contaminated food
At least eight salad bars contaminated	Confirmed cases were tied to at least 8 different restaurants in The Dalles, all with salad bars
No reason to suspect attack	At the time, there was no history of successful bioterrorism
Outbreak in multiple waves	The outbreaks seemed to occur in three separate waves
Same salad bar suppliers (?)	It is believed that restaurants had different salad bar suppliers
Suppliers distribute elsewhere (?)	It is believed that the suppliers, including farms, distributed to other restaurants outside The Dalles

Prior information and assumptions are included, like the item describing a typical cause of a *Salmonella* outbreak as poor food handling or runoff near farms. Also appearing in the set of evidence are items that are case-relevant, potential observables, which can be helpful for determining the diagnostic ability of case-specific, already observed items of evidence when the confluence of these two types of evidence is considered. These items include whether or not the restaurants had the same food suppliers for items on the salad bars and whether or not the farms that provided produce to the contaminated restaurants distributed outside of The Dalles. These items are described with a question mark, as they are uncertain.

The nine hypotheses include permutations of who was responsible, whether the outbreak was an accident or attack, and where the initial contamination took place. The hypotheses consider that the contamination occurred either at the farm, the kitchens, or the salad bars, that the contamination was the result of worker or non-worker actions, and that the contamination was an accident or a large-scale attack on the restaurants' patrons.

Facilitated modeling

Background

In risk analyses done within and across agencies and organizations, the modeling phase needs to support and engage multiple analysts rather than a single analyst. For example, Phillips (2007) describes decision conferencing as decision analysis with a group of analysts or stakeholders that build a model in real time with the help of a facilitator. With decision conferencing, the goal is selecting a "best" alternative course of action. The modeling requires the participation of the key analysts, impartial facilitation, real-time modeling with continuous output, and iteration. The goal of decision conferencing is to work toward the stopping criterion of a "requisite" model (Phillips 1984), or a model that is sufficient to generate all insights and capture the intuition of all analysts. More generally, with facilitated modeling, the purpose is to enable participants to work together much more effectively in resolving the issues of concern that brought them together (Franco and Montibeller 2010). Facilitated modeling is concerned with outcomes related to both the group process and the model.

Facilitated modeling is an important domain of risk analysis (Karvetski and Lambert 2012). Aven (2010) describes the purpose of a risk assessment is to (1) obtain an objective description of the unknowns, or (2) to obtain a scientific judgment about the unknowns from a qualified group of experts. This first aim fails to provide usable information in many situations, thus requiring the collective judgment of analysts. The treatment of uncertainty concerning hypotheses can range from simple verbal description and detection to a complete probabilistic characterization (Pate-Cornell 1996).

ACH and facilitated modeling

While ACH catalogs the consistency of a set of hypotheses with a body of evidence, ACH does not include a probabilistic characterization of the uncertainty. With ACH, a judgment tool that shows continuous output is often used (Palo Alto Research Center 2006), and the evidence is listed down the left column, with mutually exclusive hypotheses listed across the top.

Table 2 shows an example ACH implementation setup with the Rajneeshee case, with the 11 pieces of evidence and 9 hypotheses. Evidence is scored across each hypothesis (across each row) to evaluate the question "are the evidence and the hypothesis consistent with each other?" Matrix entries can consist of *Inconsistent* (I), *Very Inconsistent* (II), *Not Applicable* (NA), *Consistent* (C), or *Very Consistent* (CC). Evidence should be judged on its "diagnosticity", or its ability to distinguish which hypotheses are true or false (Heuer 1999).

Weighted inconsistency scores can be assigned to the hypotheses by summing the inconsistencies of each hypothesis over items of evidence, and weighting the summands by credibility of the evidence and the relevance of the evidence (Palo Alto Research Center 2006). Consistency of evidence with hypotheses is not included in the sum, as the goal is to disprove hypotheses. Credibility refers to the reliability of the evidence and relevance refers to the degree of timeliness of the information and the degree to which the evidence should be considered among the other pieces of evidence (Heuer 1999).

In Table 2, 11 credibility ratings are needed, 11 relevance ratings are needed, and 99 consistency ratings are needed to facilitate ACH for this example. A significant number of ratings result in *Not Applicable* (NA), and the ratings for some hypotheses' columns are the same, for example those with accident at the salad bar, because the evidence does not differentiate between worker or non-worker source.

The final step is to draw conclusions about future analyses, such as whether or not a hypothesis can be discredited and rejected given the inconsistency of the available evidence. If more than one hypothesis remains, the analysts need to determine how to disprove the remaining hypotheses so that one or none are left. It is important to analyze the sensitivity of conclusions to validity or interpretation of evidence and assumptions, and determine if deception is in play. With ACH, this might simply amount to changing consistency scores, and one or two consistency scores are not likely to affect the judgment. Reported conclusions include a qualitative description of feasibility for all hypotheses, with future needs for observations.

Advantages and disadvantages of ACH

In an experiment conducted over email, Lehner et al. (2008) tested whether ACH can reduce confirmation bias, which is characterized by the search for and interpretation of information to confirm a favored hypothesis, while discarding or misinterpreting evidence against other hypotheses. Confirmation bias can appear in three manifestations: (1) evidence supporting one hypothesis is incorrectly interpreted to support a favored hypothesis; (2) neutral evidence is interpreted to

Table 2 A typical ACH implementation for the Rajneeshee bioterror case

Evidence	Cred.	Rel.	Hypotheses								
			Accident, farm	Sabotage, farm	Accident, kitchen, workers	Sabotage, kitchen, workers	Sabotage, kitchen, non-workers	Accident, salad bars, workers	Accident, salad bars, non-workers	Sabotage, salad bars, workers	Sabotage, salad bars, non-workers
No other towns report outbreak	Med.	High	NA	NA	NA	NA	NA	NA	NA	NA	NA
No prior cases in The Dalles in last 2 years	Med.	Med.	NA	NA	NA	NA	NA	NA	NA	NA	NA
Leading causes of *Salmonella* are meat contamination and farm runoff	High	High	C	NA	C	NA	NA	NA	NA	NA	NA
Banquet patrons not sick	High	High	I	I	I	CC	C	I	I	CC	CC
Multiple items contaminated at salad bars	High	High	I	I	I	C	C	C	C	CC	CC
Contamination not easily passed between other foods, people	Med.	Med.	I	I	I	C	C	C	C	CC	CC
At least eight salad bars contaminated	High	High	C	C	II	C	C	II	II	CC	CC
No reason to suspect attack	High	High	C	I	C	I	I	C	C	I	I
Outbreak in multiple waves	Med.	Med.	C	C	I	C	C	II	II	C	C
Same salad bar suppliers (?)	Low	High	II	II	NA	NA	NA	NA	NA	NA	NA
Suppliers distribute elsewhere (?)	Low	High	II	II	NA	NA	NA	NA	NA	NA	NA

support a favored hypothesis; or (3) evidence supporting a favored hypothesis is correctly interpreted, but given more weight for drawing conclusions of likelihood than evidence supporting other hypotheses. They found that ACH is able to minimize the confirmation bias among subjects without backgrounds as intelligence analysts. The most common form of the confirmation bias is the greater weighting of evidence that supports a favored hypothesis. While these results do favor the implementation of ACH, the experiment was conducted over email and not in a facilitated setting.

In practice, ACH can fall short in sequencing the conversation among analysts, and suffers from poorly defined and implemented steps that produce questionable output. ACH does not describe where the key assumptions should be situated or elicited in the model. In particular, items of evidence can increase the *diagnosticity* of another item of evidence, or nullify the impact of another item of evidence, without any reporting of this effect. The two-dimensional matrix fails to display the connectedness of evidence, assumptions, and hypotheses. The lack of a visual map implies a significant cognitive burden on analysts, as they must continually recall and piece together the evidence, assumptions, and hypotheses as they build the model (Conklin 2006).

The measures of consistency, relevance, and credibility are poorly defined and elicited unreliably. This allows for highly subjective and unique interpretations among analysts. For example, the consistency measure should answer a well-defined question such as, "Given hypothesis H, how likely are we to see evidence e?" rather than the question "How consistent are hypothesis H and evidence e?" Emphasizing the direction of the question can clear up confusion between interpretations. Additionally, when the evidence is checked against all hypotheses, there are many cases when the relationship is insignificant, resulting in many *Not Applicable* entries.

While an analyst might prefer verbal scales for eliciting probability inputs, the verbal scale should be based on an underlying quantitative scale that can remedy confusion of interpretation (Renooij and Witteman 1999). Without a precisely defined scale for these inputs, the facilitator cannot know if differences among the inputs from two analysts should be attributed to linguistic imprecision or true differences in the likelihood estimation. Having a probability scale is useful when the inputs are to be combined across analysts and when the final likelihood assessments concerning hypotheses are interpreted.

Pope and Josang (2005) explain that the type of evidence should affect our reasoning process and how we analyze the consistency of evidence with hypotheses. Evidence that could cause or that precedes a hypothesis should be analyzed using probabilistic deductive reasoning while evidence that could result or proceed from a hypothesis should be analyzed using probabilistic abductive reasoning. The complement of evidence e^c should be included in both cases.

Probabilistic deductive and abductive reasoning can be translated to intuitive conditional probabilities. If we consider the example question of whether it has rained (H) or not (H^c), our knowledge that a low pressure system (e_1) preceded H or H^c would be causal evidence, and we should use $P(H|e_1)$ and deductively ask, "Given the low pressure system, how likely are we to see rain?" along with, "Given

NOT a low pressure system, how likely are we to see rain?" The evidence that our lawn is wet (e_2) is derivative evidence, and we should use $P(e_2|H)$ and abductively ask, "Given it has rained, how likely are we to see a wet lawn?" along with $P(e_2|H^c)$, "Given it HAS NOT rained, how likely are we to see a wet lawn?" The latter question forces the consideration of what else might cause a wet lawn. Research suggests that this strategy of considering more alternatives is a way to combat many cognitive biases (e.g., Lord et al. 1984).

If the assessments are made probabilistically, Bayes' Theorem gives a way to make logical inferences, ensuring that all necessary components for inference are elicited (Kaplan 1997; Schum 1987; Zlotnick 1972). Causal evidence can be used when setting a base-rate or prior probability of a hypothesis $P(H)$. Bayes theorem can use abductive reasoning on derivative evidence described by $P(e|H)$ and the assessment $P(e|H^c)$ to make a reversed inference on the hypothesis as

$$P(H|e) = \frac{P(e|H)P(H)}{P(e)} = \frac{P(e|H)P(H)}{P(e|H)P(H) + P(e|H^c)P(H^c)}$$

For this assessment, we should consider the states e^c and H^c and ask about $P(e^c|H)$, $P(e^c|H^c)$, and $P(H^c)$, to ensure coherence as $P(e|H) + P(e^c|H) = 1$, $P(e|H^c) + P(e^c|H^c) = 1$, and $P(H) + P(H^c) = 1$. Even in simple cases, coherence should not always be assumed (Mandel 2005). Any assessment of $P(H|e)$ that does include all of the above components needed for Bayes' Theorem is incomplete and prone to error.

Previous enhancements to ACH

While few objections are raised with respect to the philosophy behind ACH, researchers have tried various enhancements to ACH. Duncan and Wilson (2008) implement ACH using a multinomial Dirichlet Bayesian statistical model, where evidence is weighted using the analogy to prior sample size. This implementation results in inferences such as interval estimates and Bayes factors. The interpretation of the modeling inputs can be ambiguous, such as "prior sample size" for an item of evidence, and "relative importance of evidence", and the focus of the paper is on the posterior inference, rather than the process of engaging multiple analysts with the goal of communication and reduction in biases. Pope and Josang (2005) describe how ACH can be implemented using subjective logic, noting that the questions of ACH should be worded precisely to eliminate biases caused by causal interpretation.

A straightforward evolution of the ACH two-way matrix is a bi-partite graph, where each consistent/inconsistent rating represents and arc between an item of evidence and a hypothesis. Valtorta et al. (2005) claim this simple extension does not permit dependency between hypotheses, dependencies between evidence, and does not enable the ability to model "context" for hypotheses. However, their recommendations for implementation lack description, they do not use a real case demonstration, and there is no mention of how variability among types of evidence should be included in the Bayesian network, or how such an approach would be implemented with a group of analysts.

Process and application

Overview of process

In this section, we are motivated by the initial work of Valtorta et al. (2005) to supplement ACH with Bayesian networks. In practice, the structured process of this paper must be understandable to analysts and feasible through cooperation in a facilitated setting. Analysts must be able to communicate their interpretation of output from the process to policy makers along with an underlying rationale and confidence level. The goal is to have the analysts with the help of facilitators construct the entirety of a model using a general framework, therefore owning and better understanding the model. The facilitators use software to provide a visual representation of analysts' ideas and guide the analysts to discuss those ideas.

The graph of a Bayesian network offers an intuitive visualization, while the probabilistic reasoning system ensures that assessments are made logically, as Bayesian networks use Bayes' Theorem for including evidence (Pearl 1988). Bayesian networks therefore safeguard against making the reasoning errors described previously. Bayesian probability stresses degree of belief, and probabilities can be elicited for one-time events (De Finetti 1990). This property of Bayesian networks along with graphing supports multiple forms of natural reasoning.

Together, the features of Bayesian networks assure that inferences are made with consideration to all relationships between modeling elements and that uncertainty and assumptions are carried throughout the entirety of the reasoning process. The Senate Select Committee on Intelligence (United States Senate Select Committee on Intelligence 2004) criticized the prewar assessments of WMDs in Iraq for the tendency of analysts to only consider uncertainty at each separate stage of reasoning rather than over the whole chain of reasoning. Heuer (1999) was *not* unaware of this problem, but he offered limited advice on the subject for ACH users. However, a Bayesian network effectively allows "localized assessments" to be tied together to deliver a literal "big-picture" assessment. If one analyst believes state A is dependent on state B, and another analyst believes state B is dependent on state C, the network will show through their shared beliefs how state A is dependent on state C.

The exercise presented in this section describes the modeling efforts of the authors only, and therefore is a hypothetical example of how a group of analysts would perform on the historical case. We describe the testing of the process in the next section, and use that experience to outline the role of the facilitator at each step within the exercise presented. While the process is described as a sequence of steps within this section, the actual process with analysts is not a set of discrete steps, but one with fluid transitions and often requiring revisiting of past steps. The facilitators have to gauge when it is time to push the analysts forward.

Like all renditions of ACH, the process begins with a partially formed set of hypotheses X, where each hypothesis of X is a descriptive story of what has occurred, is occurring, or could occur, and a partially formed set of items of evidence Y, where each item of evidence of Y is a piece of information that might be relevant for discerning the likelihood among hypotheses. Key features of the process

are (1) the use of multiple nodes to build new hypotheses and describe all hypotheses and combinations thereof; and (2) the identification and classification of items of evidence of Y into subsets of background information and assumptions, observed items of evidence that specifically fit the case at hand, and variables that are uncertain but observable. Each subset is paired with a different type of input of a Bayesian network.

Defining hypothesis nodes

The first step for this method in construction of a Bayesian network is the collection of the differing dimensions across the hypotheses of X, which generally include who, what, when, where, why, and how, as well as other distinguishing dimensions. Analysts are needed to identify these dimensions, with some dimensions possibly having very little evidence to distinguish among the uncertain states. For the Bayesian network, each hypothesis of X is described using a set of hypothesis nodes, denoted $H = \{H_1,\ldots, H_n\}$. Each node H_j describes one dimension of the hypotheses (e.g., the "who" dimension), and each node H_j has mutually exclusive and exhaustive states $\{H_{j_1},\ldots, H_{jn_j}\}$, with $n_j \geq 2$ for all j. In the Bayesian network, each hypothesis of X is then an element of $H_1 \times \ldots \times H_n$.

For example, with the Rajneeshee case, the *H_Who* node describes the party responsible for the *Salmonella* appearing in the patrons' food, and the states of this node are *workers* and *non-workers.* Likewise the *H_Where* node consists of the states *SaladBar*, *Kitchen,* or *Farm*, and describes where the contamination *first* took place. Finally, the *why* or intent of the *Salmonella* outbreak is described in the *H_Why* node, with states *attack* or *accident.* Importantly, every initial hypothesis is visible in the Bayesian network. Questions about the biological agent and timing of the contamination could also be asked, but these three nodes are sufficient to express hypotheses such as an accidental contamination in the kitchen by workers or attack at the produce farm by non-workers.

Defining hypotheses using multiple hypothesis nodes structures the hypotheses to elicit the prior probability among hypotheses in a way that considers all dimensions and the causal interactions within hypotheses. This way also might help identify hypotheses that were previously unconsidered interactions between the hypothesis nodes, possibly representing the convergence of partially formed hypotheses put forth by multiple analysts. With the Rajneeshee example, there are 12 permutations of the three hypothesis nodes, whereas before with ACH there were only nine. Nevertheless, there is a practical tradeoff between generating valuable, previously unconsidered hypotheses and avoiding unnecessary hypotheses. However, within the Bayesian network, this is not a theoretical problem. If a hypothesis combination is initially deemed considerably less plausible than the others, it can be assigned a sufficiently small prior probability.

When making assessments, as we will see, analysts can focus on a key dimension of one or more hypotheses and one or more items of evidence without considering other unrelated dimensions of a hypothesis. This can eliminate redundant elicitations.

The job of the facilitators at this stage is to prompt the analysts to consider the major intelligence questions that need to be addressed, regardless of a preconceived lack of evidence. Depending on the breadth of the analyses, the facilitators should also try to limit the number of states in each hypothesis node to capture the essence of what needs to be answered for each particular dimension. The nodes can be projected onto a screen so that the analysts begin to see that they are building a model.

Defining evidence nodes

Next a general set of evidence is collected from analysts. At the onset, the set of evidence Y is in line with Heuer's definition of evidence (Heuer 1999), as any observation, fact, or assumption that the analysts deem relevant. This set might be sparse, as the level of creativity required to collect evidence and hypotheses should not be understated. Various brainstorming methods can aid with the construction of each of these sets. The evidence should then be gathered considering the conditions that might have led to the hypotheses, conditions that could have shaped intent or lack of intent, and evidence that could be the result of each hypothesis. The list of evidence should be assembled with the following questions in mind for each hypothesis:

- Given a hypothesis, what evidence would we expect to find that we have found?
- Given a hypothesis, what evidence would we expect to find that we have not found, or have not tried to find?
- Given a hypothesis, what evidence would we not expect to find that we have found?
- Given a hypothesis, what evidence would we not expect to find that we have not found, or have not tried to find?

Argument mapping might be useful here in the process to consider relationships between pieces of evidence and missing items of evidence along with assumptions (van Gelder 2008). Analysts should consider all items of evidence together (rather than independently) and can construct an argument map of how the observational evidence could be consistent with a hypothesis, what additional assumptions are needed, and what prior knowledge can support each hypothesis. This exercise serves as a way to stimulate creative thinking. An example of argument mapping for one hypothesis is shown in Fig. 1.

There are two objectives when forming the evidence nodes for a Bayesian network. First, evidence nodes should prompt analysts to consider how rare a true observation actually is; for example, $P(e|H)$ has to join $P(e^c|H)$. Second, evidence nodes should only consist of items of evidence that are truly observations. Part of the difficulty in incorporating a set of evidence into a Bayesian network is the variability in types of evidence. This set of evidence in Table 1 consists of four "prior" or background items of evidence or assumptions, five truly observed items of evidence specific to the case, and two items that are uncertain but possibly observable.

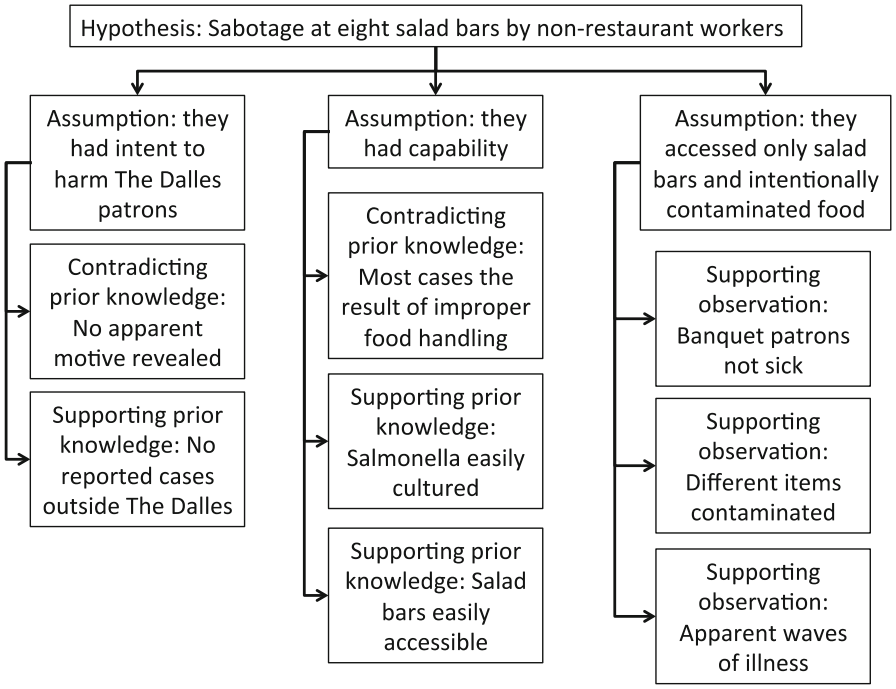

Fig. 1 An example of an argument map that identifies what observations and what prior knowledge support and contradict a hypothesis

For example, the evidence of *Leading cause meat/runoff* describes that in previous documented cases of *Salmonella* poisoning, the source has typically been improper handling of meat in the kitchen or runoff at the farms. This type of evidence is different than the item of evidence *E_Banquet patrons not sick*, describing that those patrons who ate at the banquets of some of the restaurants did not become sick. The former piece of evidence is something that should inform a prior probability judgment among hypotheses, but the *diagnosticity* of this item of evidence can be minimized by the observation that banquet patrons were not sick. In other words, the observed items update the prior probabilities. Keeping with the example, if contamination occurred from improper handling in the kitchen or runoff at the farms, contaminated food would have been served at both the salad bar and the banquet. Only the patrons who ate food from the salad bar were sick, suggesting that food at the banquet was not contaminated, and further suggesting that food at the salad bar was not contaminated in the typical manner.

When considered in confluence, the uncertain, but observable items can help in judging the likelihood of seeing the already observed items, and will help with a what-if analysis. For example, the information that no other towns were infected has the ability to distinguish where the contamination took place or did not take place, depending on whether or not these farms distribute to other places. If farms distribute outside of The Dalles, it is less likely that the contamination took place at the farms because restaurant patrons of other towns would likely show symptoms.

With this in mind, the set of evidence as defined by Heuer is partitioned into a set of prior background knowledge and assumptions B, a set of case-relevant observed evidence E, and a set of observables that are still uncertain Q, with $Y = B \cup E \cup Q$, and B, E, and Q mutually exclusive. Heuer's definition of evidence does not differentiate between these types of evidence. The separation is important for considering consistency, as prior knowledge will help differentiate among hypotheses before considering case-specific observations, which typically result from a hypothesis (Pope and Josang 2005).

The prior background knowledge and assumptions of B are not included as multistate nodes in the Bayesian network, but appear as single-state "nodes" only to remind the analysts as they set prior probabilities among the hypothesis nodes of H. The set of case-relevant observed evidence of E is next used to form the nodes for the evidence. For each element of E, we form a separate node, with the observation itself being the first labeled state of the node, and then define at least one more state to make sure the states of each node are mutually exclusive and exhaustive.

In the Rajneeshee case, the food sources of contamination were diverse, and included vegetables, prepared salads, sauces, salad dressings, and pastas. While this observation is described as the state *DifferentItems*, we pair this state with the *SameItem* state. Eventually this node is switched to *DifferentItems*, but pairing the two states together allows for some comparison in determining the rarity of the observation.

A final set of nodes represents the elements of Q and includes states of the world that are not observed with certainty, but could be observed. The elements of Q could also include assumptions or conjectures of the analysts that can be supported or contradicted by evidence. The observables of Q should be collected with the goal of increasing *diagnosticity* among the hypotheses; that is, contradicting or refuting a subset of hypotheses states. For example, at an early stage, we might want to determine if the restaurants all had the same salad bar food suppliers. These nodes also contain at least two states. After the probabilities are elicited, the value of obtaining this information can be viewed during the what-if analysis by switching between states after probabilities are assessed.

The overall role of the facilitator at this step is to encourage the analysts to think hard about what intelligence has been collected or what could be collected. Initially, the facilitators could use whiteboards to archive the discussed items of evidence. The facilitators should also begin to label the evidence nodes and clarify what the states of each node are, as in many cases it is not obvious. Furthermore, the facilitator should search for the fewest possible states that will accommodate the ideas of analysts for each node.

Defining arcs

After the hypothesis nodes of H are defined, along with evidence nodes, the arcs are elicited to be consistent with how analysts reason most comfortably. Causality is a useful way to elicit arcs for a Bayesian network (Walshe and Burgman 2010; Nadkarni and Shenoy 2001, 2004), even though the arcs imply conditional probability, which does not require a causal relationship. The first set of arcs defined

is between the hypothesis nodes of *H*, as these arcs will be used for setting the probabilities among hypotheses. In the Rajneeshee case, these arcs might be thought of in a chronological order. This way of viewing the arcs can provide a natural way of setting arcs when multiple possibilities exist.

For example, given we assume who is the responsible party, we can make better inference on whether the event was an accident or not, and where the original contamination took place. Considering who and why the contamination took place might give us better insight into where the contamination took place. Therefore, arcs emanate from *H_Who* to *H_Why* and *H_Where*, while another arc emanates from *H_Why* to *H_Where*. These arcs are shown between the hypothesis nodes of Fig. 2.

The next arcs to be added are between nodes of *H* and nodes of *E*. An arc in this situation represents the conditional dependence between a hypothesis node and a case-specific item of evidence. Given that the arcs represent a causal influence, they will typically emanate from the nodes of *H* to the nodes of *E*.

Sometimes a hypothesis node might not have a clear causal impact on an element of *E*; thus an arc is not needed. Having multiple nodes for each hypothesis can reduce the number of judgments when compared with the full ACH. For example, knowing that the reported incidents came in multiple waves can help us differentiate between whether or not the contamination was the result of an accident or attack, and where the contamination took place, but will not help us directly distinguish whether it was the workers or not. Thus, arcs emanate from the *H_Why* and *H_Where* nodes to the *E_MultipleWaves* node, but there is no arc between *H_Who* and *E_MultipleWaves*. *E_MultipleWaves* communicates with *H_Who* through *H_Why* in a *serial connection* (Jensen 2001).

The next set of arcs is to be added between uncertain observable nodes of *Q* and the hypothesis nodes of *H* and/or the evidence nodes of *E*. As described before, the

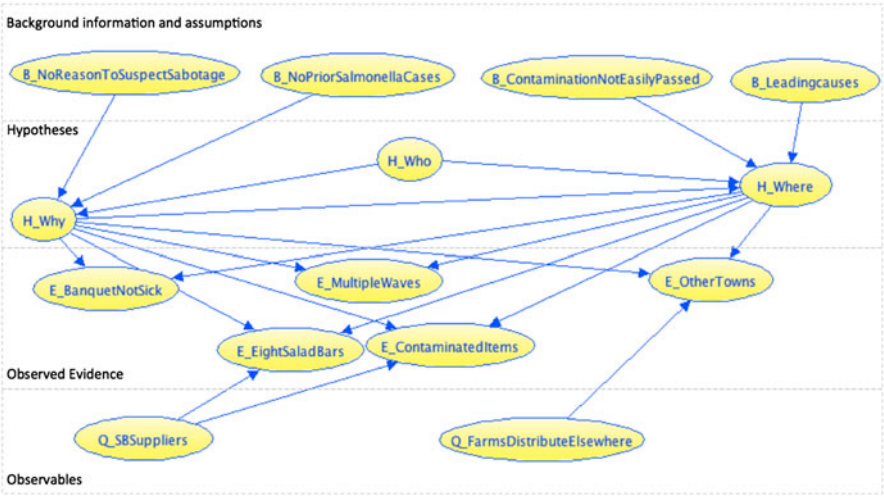

Fig. 2 The final Bayesian network graph with the hypotheses described by three nodes, and the set of evidence broken down into the subsets of background information and assumptions, observed evidence, and uncertain observables

information that no other towns were infected can help distinguish where the contamination took place, depending whether or not these farms distribute to other places. Therefore, there is a *converging connection* of *H_Where* and *Q_FarmsDistributeElsewhere* to *E_OtherTowns*. When *E_OtherTowns* is set to the state *OnlyTheDallesCont*, switching between the states of *Q_FarmsDistributeElsewhere* shows how the confluence of these observations can affect the ability to distinguish between states of *H_Where* (Jensen 2001).

With the nodes defined at this stage, the facilitators can provide the model that consists only of nodes, and coach the analysts as they draw in arcs either with software or with pen and paper. Once collected, the arcs can be discussed, combined, and inserted. The idea of causality might need to be described throughout, and the task of the facilitators is to ensure that the final model represents casual relationships where possible. Naturally, the inclusion of arcs will not always be a linear process described herein, as additional arcs can emerge or become unnecessary during the probability elicitation stage, when analysts explore how they form their probability estimates and how these estimates differ. At least one iteration in the process is necessary. For this case demonstration, Fig. 2 shows the final network structure of the Bayesian network, which includes the arcs from the nodes of *B* for visual purposes only.

Eliciting probabilities

Throughout the process of building the network, the analysts are tasked with providing probabilities that are specified by the network. It is important that these probabilities are elicited in a manner that is comfortable with the analysts, and the elicitation should incorporate qualitative descriptions and visual aids in addition to quantitative estimates (Renooij and Witteman 1999). The scale for eliciting probabilities can be a seven-point verbal and quantitative scale (Witteman and Renooij 2002; Renooij 2001) or a five-point scale (Kent 1964). The scale allows each analyst to visualize, quantify, and communicate uncertainty in a comfortable way.

Empirically, Clemen et al. (2000) find that the subjects are able to forecast conditional dependence quite well with just a seven-point probability scale. Winkler and Clemen (2004) find forecasts of the dependence parameters improve when each participant uses multiple methods (or questions) and when the forecasts are aggregated across multiple participants. In general, the forecasts improve at a greater rate when more participant forecasts are combined than when more methods are used, but the marginal improvement declines in either case.

The first set of probabilities to be elicited is the set of unconditional and conditional probabilities for the nodes of *H*. The analysts should be continuously exposed to assumptions and prior background knowledge of *B* when making these probability judgments. Importantly, more assumptions may emerge at this stage as individual differences in reasoning processes are discovered. These probabilities are shown in Fig. 3.

For example, the *H_Who* prior probabilities are set to 0.9 for the workers, and 0.1 for the non-workers. The prior probability for the *H_Why* node is then set

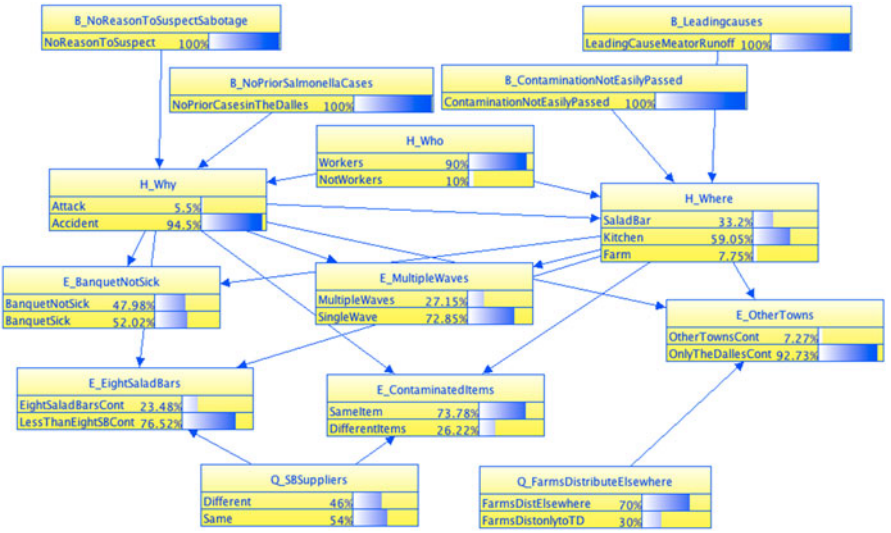

Fig. 3 The Bayesian network with the probabilities described, including the probability of observed evidence given the prior probabilities of the hypothesis nodes and the conditional probabilities

conditionally on the states of the *H_Who* node. With the conditional probabilities of Table 2, we have

$$\begin{aligned} P(\text{Attack}) =& P(\text{Attack}|\text{Workers})P(\text{Workers}) + P(\text{Attack}|\text{NotWorkers})P(\text{NotWorkers}) \\ =& (0.05)(0.9) + (0.1)(0.1) = 0.055. \end{aligned}$$

Then, the conditional probabilities are elicited for the arcs that emanate from the nodes of H and point to the nodes of E and possibly Q. These probabilities are shown in Fig. 3 as well, using the conditional probabilities assessed in Table 4.

The degree to which the probability elicitation is a group exercise depends on time factors, as well as social factors. The simplest form of elicitation involves group discussion of each probability. Conversely, the probability elicitation procedure can be done in multiple steps to allow analysts to explain their reasoning (Burgman et al. 2011). Estimates can be elicited from each analyst independently, and then the analysts can be presented with how the estimates differed. The analysts can discuss why their estimates differed, and then each analyst can resubmit an estimate. If necessary, new nodes, states, or arcs can be added to the network if key assumptions are uncovered. However, there is again a practical tradeoff, as this two-stage approach is time-consuming when dozens of elicitations are needed.

Another role of the facilitator is to identify cases when not all probabilities need to be elicited. With more than three arcs entering a node, there are at least 8 probabilities that need to be accounted for. Nevertheless, sometimes the majority of the probabilities are the same for all cases besides one particular case (Heckerman 1990), or the elicitation can be simplified with *noisy-or* assumptions (Pearl 1988). With this in mind, it is the job of the facilitator to elicit only the non-redundant

Table 3 Probability tables to set prior probabilities for the hypothesis nodes

H_Who	--
Workers	0.9
NotWorkers	0.1

	H_Who	
H_Why	Workers	NotWorkers
Attack	0.05	0.1
Accident	0.95	0.9

	H_Who, H_Why			
	Workers		NotWorkers	
H_Where	Attack	Accident	Attack	Accident
SaladBar	0.3	0.3	0.8	0.6
Kitchen	0.65	0.65	0.1	0.05
Farm	0.05	0.05	0.1	0.35

probabilities. For a method to shrink the number of probability elicitations for a large conditional probability table, see Wang and Druzdzel (2000).

After final probability estimates are elicited from analysts, they can be aggregated into a single estimate, and large remaining discrepancies can be noted for sensitivity analysis (Clemen and Winkler 1993, 1999). Figure 3 displays an example of the Bayesian network now quantified with the marginal probabilities that are calculated using the conditional probabilities of Tables 3 and 4. We note the prior probability on the hypothesis variables describes that an accident by the workers in the kitchen is the most likely scenario.

For aggregating probabilities, a traditional approach to aggregation is calculating the mean. Another approach is to take the median, as this approach reduces the influence of a non-compliant analyst. The median aggregation has been shown to outperform the average, is robust to outliers, and minimizes gaming or disruptive behavior with the assumption of single-peaked judgments[2] (Scholz and Hansmann 2007; Schummer and Vohar 2007).

Consider for example three analysts where one analyst thinks an event will occur with probability 0.7, while other analysts propose estimates 0.2 and 0.3. The mean is 0.4, but the first analyst can sway the mean estimate by proposing a probability of 1 rather than 0.7. This new mean estimate would be 0.5 rather than 0.4, whereas the median is 0.3 in both cases. The only way the first analyst can sway the median probability in this case is to report something below 0.3, which moves the median estimate further from 0.7, and thus, by assumption, is judged as worse by that analyst. When an analyst's true estimate is below the median, a similar argument shows he or she can only change the estimate by reporting something greater than the median. Finally, when an analyst's estimate is the median, it is best that the analyst report his or her true estimate. In general, the median rule is a *strategy-proof*

[2] Given an expert has a true probability estimate of p^*, single-peakedness implies that for p^1 and p^2, if $p^* \leq p^1 < p^2$ or $p^2 < p^1 \leq p^*$, then the analyst truthfully views p^1 as a better estimate to p^2 (Schummer and Vohra 2007).

Table 4 Conditional probability tables to set probabilities among observed and uncertain items of evidence

	H_Where, H_Why					
	SaladBar		Kitchen		Farm	
E_BanquetNotSick	Attack	Accident	Attack	Accident	Attack	Accident
BanquetNotSick	0.9	0.9	0.3	0.3	0.05	0.05
BanquetSick	0.1	0.1	0.7	0.7	0.95	0.95

Q_Suppliers	--
Different	0.46
Same	0.54

	H_Where, H_Why					
	SaladBar		Kitchen		Farm	
E_MultipleWaves	Attack	Accident	Attack	Accident	Attack	Accident
MultipleWaves	0.5	0.05	0.5	0.3	0.8	0.8
SingleWaves	0.5	0.95	0.5	0.7	0.2	0.2

Q_FarmsDistributeElsewhere	--
FarmsDistElsewhere	0.7
FarmsDistonlytoTD	0.3

	H_Where, H_Why, Q_FarmsDistributeElsewhere											
	FarmsDistributeElsewhere						FarmsDistonlytoTD					
	SaladBar		Kitchen		Farm		SaladBar		Kitchen		Farm	
E_OtherTowns	Attack	Accident	Attack	Accident	Attack	Accident	Attack	Accident	Attack	Accident	Attack	Accident
OtherTownsCont	0.05	0.02	0.05	0.02	0.95	0.95	0.05	0.02	0.05	0.02	0.05	0.05
OnlyTheDallesCont	0.95	0.98	0.95	0.98	0.05	0.05	0.95	0.98	0.95	0.98	0.95	0.95

	H_Where, H_Why, Q_SBSuppliers											
	Different						Same					
	SaladBar		Kitchen		Farm		SaladBar		Kitchen		Farm	
E_EightSaladBars	Attack	Accident	Attack	Accident	Attack	Accident	Attack	Accident	Attack	Accident	Attack	Accident
EightSaladBarsCont	0.8	0.02	0.5	0.02	0.7	0.7	0.8	0.8	0.5	0.02	0.7	0.7
LessThanEightSBCont	0.2	0.98	0.5	0.98	0.3	0.3	0.2	0.2	0.5	0.98	0.3	0.3

	H_Where, H_Why, Q_SBSuppliers											
	Different						Same					
	SaladBar		Kitchen		Farm		SaladBar		Kitchen		Farm	
E_ContaminatedItems	Attack	Accident	Attack	Accident	Attack	Accident	Attack	Accident	Attack	Accident	Attack	Accident
SameItem	0.2	0.2	0.8	0.8	0.9	0.9	0.5	0.95	0.8	0.8	0.9	0.9
DifferentItems	0.8	0.8	0.2	0.2	0.1	0.1	0.5	0.05	0.2	0.2	0.1	0.1

rule, or one that incentivizes an analyst to report a "true estimate" (Schummer and Vohar 2007). Other procedures should be used to ensure that the estimates are calibrated and coherent.

Making inferences

With the Bayesian network populated with probabilities, inferences can be made as to the level of uncertainty concerning key hypothesis states, and sensitivity analysis can show how analyst disagreement affects the likelihood of key variables. A first step is to switch all case-specific observation nodes of *E* to the state of the observation. Software will propagate probability through the network using Bayes' Theorem to provide posterior estimates for the hypothesis nodes of *H* and the nodes of *Q*.[3] This activity is shown in Fig. 4. The next step is to use the nodes of *Q* in a what-if analysis, to determine how much more uncertainty concerning the nodes of *H* can be removed by finding out more information.

The first form of the what-if sensitivity analysis is shown in Fig. 5 and involves switching the *Q* nodes. Here switching the *Q_SBSupplier* and *Q_FarmsDistributeElsewhere* nodes to *Different* suppliers and *FarmsDistElsewhere*, which are the most likely states of the respective nodes, further implies the occurrence of an attack at the salad bar, but the change to the probabilities of the *H_Why* and *H_Where* hypothesis nodes is not large. This combination of new states of the *Q* nodes loads

[3] Existing software can help with these inferences (see Eleye-Datubo et al. 2006 for a case study with Hugin software). UnBBayes is publicly available software for potential use (available at http://sourceforge.net/projects/unbbayes/).

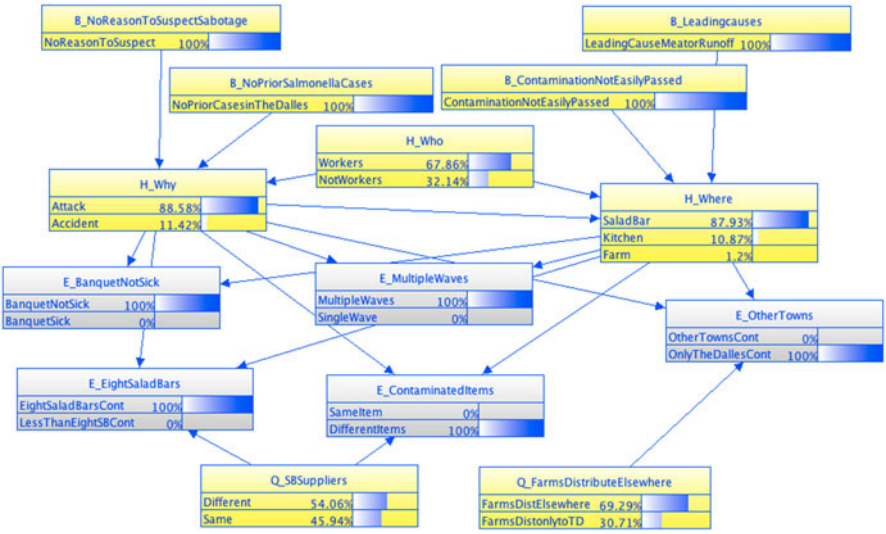

Fig. 4 The posterior probabilities among the hypothesis nodes and uncertain observable nodes after the observations are entered into the Bayesian network

more probability onto *Attack* and *SaladBar*, and thus demonstrates how the new information can confirm the a posteriori most likely hypotheses.

Analysts can also use to the model to understand how one or more of the opposite findings could change the posterior probabilities on the hypothesis nodes. By switching the *Q_SBSupplier* and *Q_FarmsDistributeElsewhere* nodes to *Same* suppliers and *FarmsDistOnlytoTD*, we find the posterior probability on *H_Why* goes from 88.58 % on *Attack* to 81.36 %, the posterior probability on *H_Who* goes from 67.86 % on *Workers* to 69.05 %, and the posterior probability on *H_Where* goes from 87.93 % on *SaladBar*, 10.87 % on *Kitchen*, and 1.2 % on *Farm*, to 83.50 % on *SaladBar*, 12.43 % on *Kitchen*, and 4.07 % on *Farm* (not shown). Analysts can then consider these changes, and report the expected effects of collecting information for these nodes.

A second form of sensitivity analysis would look at how probability differences could affect the conclusions. For example, considering Fig. 4, where we obtain the posterior probabilities for the hypothesis nodes, we consider how analyst disagreement resulting from differing assumptions of background knowledge nodes changes the resulting posterior probabilities. Assuming a more conservative analyst sets the prior probability distribution on *H_Who* now at 50 % on *Workers*, and 50 % on *NotWorkers,* we include the five observations as in Fig. 4 for the *E* nodes. We find that the posterior probability on *H_Why* goes from 88.58 % on *Attack* to 92.98 %, the posterior probability on *H_Who* goes from 67.86 % on *Workers* to 19.00 %, and the posterior probability on *H_Where* goes from 87.93 % on *SaladBar*, 10.87 % on *Kitchen*, and 1.2 % on *Farm*, to 94.88 % on *SaladBar*, 3.66 % on *Kitchen*, and 1.46 % on *Farm* (not shown). For prioritizing this type of

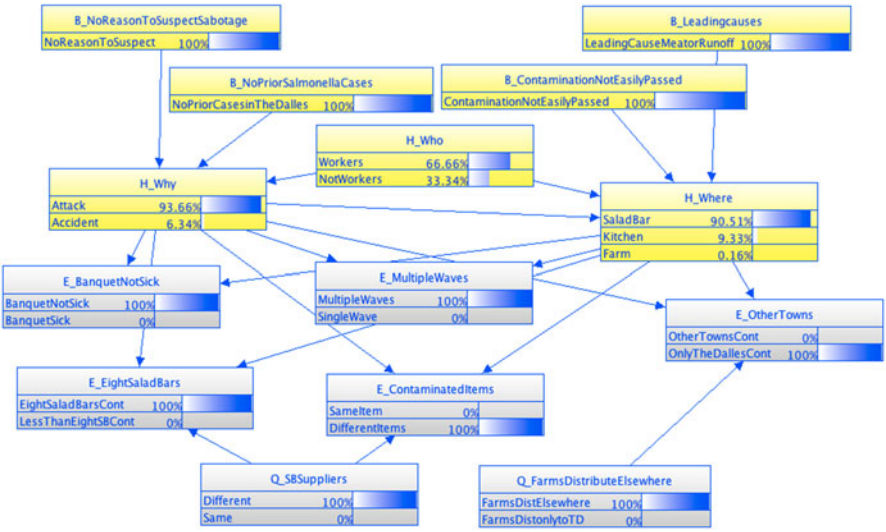

Fig. 5 The posterior probabilities among the hypothesis nodes after the observations are entered into the Bayesian network and the uncertain observation nodes are set to a certain state

sensitivity analysis on the probabilities, the facilitators should look for the probability assessments with the greatest ranges across the analysts.

This sensitivity analysis on the *H_Who* prior probability distribution shows the high degree of sensitivity of the posterior distribution to the prior probability for this particular node. This should prompt a new type of search and what-if analysis. In particular, by looking at the network in Figs. 3, 4, 5 and 6, we see that the *H_Who* node is the only hypothesis node without evidence flowing out of the node. The analysts' task is then to identify intelligence that should be collected to make direct inferences on the *H_Who* node, such as questioning the workers, or looking for workers who worked at multiple restaurants. In fact, in the original case, it was eventually uncovered that the workers got sick and that there was not overlap in the workforces of the multiple restaurants (Weaver and James 1985).

Multiple what-if analyses, such as those presented above, will allow analysts to make logical, well-structured recommendations to policy makers. Strategic insight emerges when analysts consider gathering intelligence that would greatly change the probability of a hypothesis being correct and resolve debate. In total, a "what-if" analysis can help by (1) identifying the lynchpin assumptions that could change the characterization of uncertainty among hypotheses, thereby forcing analysts to further validate key assumptions, particularly those about inaccurate information and forms of deception (Stech and Elasser 2004); (2) showing how disagreement among analysts can change the characterization of uncertainty among hypotheses; and (3) identifying instances of surprise when a node of *E* is not switched to the supposed observation, to determine whether the observation is expected, given the remaining body of observed evidence.

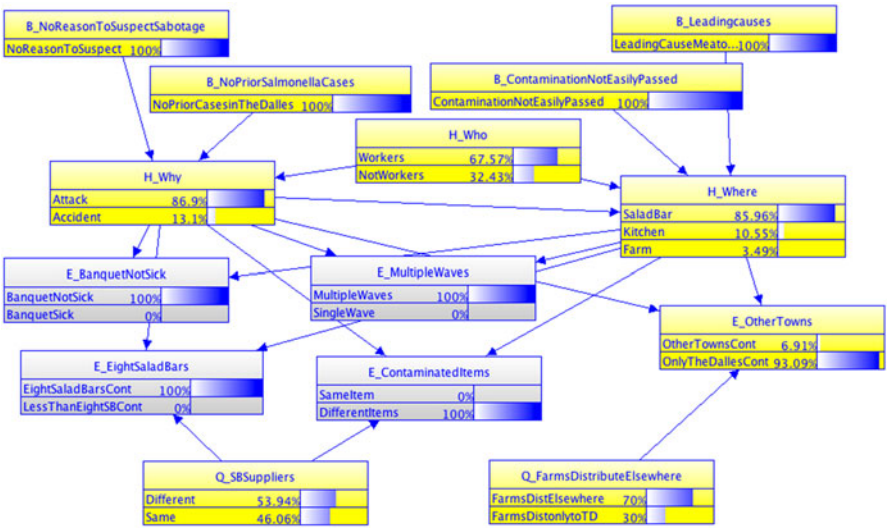

Fig. 6 Understanding the agreement of the observed items of evidence. By switching all nodes of *E* but *E_OtherTowns*, we see the observation of *OnlyTheDallesCont* is the most likely outcome

This third form of sensitivity analysis can determine the consistency of a body of observed evidence. Figure 6 is based on Fig. 4, but all of the nodes of *E* are switched to their observed states, except for *E_OtherTowns*. This allows for an understanding of the agreement of the observed items of evidence. We see the observation of *OnlyTheDallesCont* is the most likely outcome by a large margin, which confirms that the observation that "only restaurants within The Dalles were contaminated" is not a suspicious item of evidence.

Conclusion

Discussion of process and application

Many people in the Intelligence Community spend their time trying to assess events with political implications. Intelligence analysts have a variety of approaches for the task some which are interactive (e.g., Nemeth et al. 2001), argument-based (e.g., Pherson 2008), or statistical (e.g., Sticha et al. 2005). The process of intelligence analysis presented in this paper builds on all the above, as the process uniquely places simple statistical modeling at the discretion of a group of analysts. The analysts construct the entirety of a model using a general framework, therefore "owning" the model (Phillips 2007). The facilitators interact with software as they guide the analysts' interactions, making sure they understand each step.

The process for forming a Bayesian network relies on a natural taxonomy of variables, which pairs well with the inputs for a Bayesian network. In the Rajneeshee case, the characterization of evidence types and the confluence of

evidence are valuable for the evolving assessment. A picture develops of the probabilistic reasoning about the intelligence problem.

One major difference between this method and others that build on ACH (Duncan and Wilson 2008; Valtorta et al. 2005; Pope and Josang 2005) is that this method is implemented within a group of analysts. The method is a literal revision process in which the group's Bayesian network serves as a visual representation of a model to aid communication and inter-analyst understanding. The process takes advantage of disagreement during sensitivity analysis so that consensus is not forced upon analysts, but a shared understanding develops of how inferences might hinge on a few contested relationships between hypotheses and evidence. It therefore provides order for analysts to consider and challenge each other's ideas, thereby encouraging them to see again their own ideas in a new way. In this sense, many of the benefits of dialogue mapping are shared with this process (Conklin 2006).

The power of constructing a shared Bayesian network comes from the ability to consider localized assessments and then use existing software to produce a "big picture" assessment. Bayesian networks allow for interactions and stimulate cognition and conversation in a way that a two-dimensional matrix is not capable of. Although a Bayesian network is a more sophisticated model than ACH, it can be less tedious by eliminating repeated elicitations after partitioning hypotheses into multiple dimensions and focusing on local relationships between variables. With ACH, 121 inputs were needed to define the model in Table 2, whereas 118 conditional probabilities were needed in Tables 3 and 4 to define the Bayesian network.

With respect to the entire intelligence progression, the probabilities inferred about the states of hypothesis nodes and findings of the sensitivity analysis are an output of comprehensive analysis, all of which can be communicated with policy makers (Keisler and Noonan 2012). The output of the process can be used as inputs for later policy-making analyses. This is important for documenting why decisions are made when policy makers are interrogated. The proposed process helps to create intelligence assessments that are more transparent to outside reviewers.

Summary of future research

Implicit in the process for group judgment is an assumption of the lack of gaming or disruptive behavior. For example, the analysts should all be seeking to uncover the truth, and not just paint a picture that is best for an individual analyst or agency. If gaming behavior is anticipated, the process should further strive to be *strategy-proof*, implying that the decision process incentivizes each stakeholder to truthfully report his or her honest preferences and other inputs (Schummer and Vohra 2007). Another important area of future research involves testing the process on new, challenging cases with a variety of group participants.

While this paper has motivated and presented a process for using Bayesian networks for hypotheses analysis, it offers only a hypothetical case demonstration of the value of the process. Structured analytic techniques often lack empirical support for their value (Marrin 2007), but we are rigorously testing the proposed process in an experimental setting. Initial testing with a group of analysts is favorable, but

future work is needed to formally evaluate the process with a variety of cases. In particular, we are developing an action-researching approach for evaluating process effectiveness and output effectiveness, which describe the quality of the group analysis process and the modeling output, and these measures of effectiveness are evaluated with user surveys, measured outcomes, and facilitator reflection and generalization (Montibeller 2007; Schilling et al. 2007). For example, the surveys compare the new process with ACH, but also with an "ideal" process on criteria related to ability to articulate independent ideas, exchange of information across analysts, transparency and comprehensibility of process, rationality of process, structuring of group interaction, and ability to generate creative thinking and insights.

Acknowledgments We would like to thank IC Postdoctoral Research Fellowship Program for providing funding for this effort. Opinions expressed in this paper are those of the authors and do not necessarily represent the views or positions of the Federal Bureau of Investigation.

References

Aven T (2010) On the need for restricting the probabilistic analysis in risk assessment to variability. Risk Anal 30(3):354–360

Burgman MA, McBride M, Ashton R, Speirs-Bridge A, Flander L, Wintle B, Fidler F, Rumpff L, Twardy C (2011) Expert status and performance. PLoS ONE 6(7):1–7

Carus, W.S. 2000. The Rajneeshees. In (eds.) J.B. Tucker. *Toxic Terror*. MIT Press. Cambridge, MA

Caswell DJ, Pate-Cornell ME (2011) Probabilistic analysis of a country's program to acquire nuclear weapons. Mil Oper Res 16(1):5–20

Caswell DJ, Howard RA, Pate-Cornell ME (2011) Analysis of national strategies to counter a country's nuclear weapons program. Decis Anal 8(1):30–45

Clemen RT, Winkler RL (1993) Aggregating point estimates: a flexible modeling approach. Manage Sci 39(4):501–515

Clemen RT, Winkler RL (1999) Combining probability distributions from experts in risk analysis. Risk Anal 19(2):187–203

Clemen RT, Fischer GW, Winkler RL (2000) Assessing dependence: some experimental results. Manage Sci 46:1100–1115

Conklin J (2006) Dialogue mapping: building shared understanding of wicked problems. Wiley, West Sussex

De Finetti B (1990) Theory of probability. A critical introductory treatment. Wiley, New York

Deisler PF (2002) A perspective: risk analysis as a tool for reducing the risks of terrorism. Risk Anal 22(3):405–413

Duncan KA, Wilson JL (2008) A multinomial-Dirichlet model for analysis of competing hypotheses. Risk Anal 28(6):1699–1709

Eden C, Radford J (1990) Tackling strategic problems: the role of group decision support. Sage, London

Eleye-Datubo AG, Wall A, Saajedi A, Wang J (2006) Enabling a powerful maritime and offshore decision-support solution through Bayesian network technique. Risk Anal 26(3):695–721

Franco LA, Montibeller G (2010) Facilitated modeling in operational research. Eur J Oper Res 205:489–500

Heckerman D (1990) Probabilistic similarity networks, Ph.D. thesis. Program in Medical Information Sciences, Stanford University, Stanford. Report STAN-CS-90-1316

Heuer RJ (1999) Psychology of intelligence analysis. United States Government Printing

Jensen FV (2001) Bayesian networks and decision graphs. Springer, New York 268p

Jo D, Gartzke E (2007) Determinants of nuclear weapons proliferation. J Conflict Resolut 51(1):167–195

Kaplan S (1997) The words of risk analysis. Risk Anal 17(4):407–417

Karvetski CW, Lambert JH (2012) Evaluating deep uncertainties in strategic priority-setting with an application to facility energy investments. Syst Eng 15(4):483–493

Keisler JM, Noonan PS (2012) Communicating analytic results: a tutorial for decision consultants. Decis Anal 9(3):274–292
Kent S (1964) Words of estimative probability. Stud Intell 8:49–65
Lehner PE, Adelman L, Cheikes BA, Brown MJ (2008) Confirmation bias in complex analyses. IEEE Trans Syst Man Cybern A Syst Hum 38(3):584–592
Lord CG, Ross L, Lepper MR (1979) Biased assimilation and attitude polarization: the effects of prior theories on subsequently considered evidence. J Perspect Soc Psychol 37:2098–2109
Lord CG, Lepper MR, Preston E (1984) Considering the opposite: a corrective strategy for social judgment. J Perspect Soc Psychol 47:1231–1243
Mandel DR (2005) Are risk assessments of a terrorist attack coherent? J Exp Psychol Appl 11(4):277–288
Marrin S (2007) Intelligence analysis: structured methods or intuition? Am Intell J 25:1–10
McLaughlin J, Pate-Cornell ME (2005) A Bayesian approach to Iraq's nuclear program intelligence analysis: a hypothetical illustration. 2005 international conference on intelligence analysis. https://analysis.mitre.org/proceedings/Final_Papers_Files/85_Camera_Ready_Paper.pdf
Montibeller G (2007) Action-researching MCDA interventions. In: Shaw D (ed) Key-note papers, 49th British operational research conference. The OR Society
Nadkarni S, Shenoy PP (2001) A Bayesian network approach to making inferences in causal maps. Eur J Oper Res 128:479–498
Nadkarni S, Shenoy PP (2004) A causal mapping approach to constructing Bayesian networks. Decis Support Syst 38:259–281
Nemeth C, Brown K, Rogers J (2001) Devil's advocate vs. authentic dissent: stimulating quantity and quality. Eur J Soc Psychol 31:707–720
Palo Alto Research Center (PARC) (2006) ACH: Version 2.0.3
Parnell GS, Smith CM, Moxley FI (2010) Intelligent adversary risk analysis: a bioterrorism risk management model. Risk Anal 30(1):32–48
Pate-Cornell ME (1996) Uncertainties in risk analysis: six levels of treatment. Reliab Eng Syst Saf 54:95–111
Pearl J (1988) Probabilistic reasoning in intelligent systems. Morgan Kaufmann, San Francisco
Pherson R, Schwarts A, Manak E (2008) Anticipating rare events: the role of ACH and other structured analytic techniques. In: Anticipating rare events: a scientific perspective on problems, pitfalls, and potential solutions. Office of the Secretary of Defense/DDR&E/RTTO, Washington, DC
Phillips LD (1984) A theory or requisite decision models. Acta Psychol 56:29–48
Phillips LD (2007) Decision conferencing. In: Edwards W, Miles RF, von Winterfeldt D (eds) Advances in decision analysis. Cambridge University Press, Cambridge
Pope S, Josang A (2005) Analysis of competing hypotheses using subjective logic. In: 10th CCRTS: the future of command and control, pp 1–30
Renooij S (2001) Probability elicitation for belief networks: issues to consider. Knowl Eng Rev 16(3):255–269
Renooij S, Witteman CLM (1999) Talking probabilities: communicating probabilistic information with words and numbers. Int J Approx Reason 22:169–194
Schilling MS, Oeser N, Schaub C (2007) How effective are decision analyses? Assessing decision process and group alignment effects. Decis Anal 4(4):227–242
Scholz RW, Hansmann R (2007) Combining experts' risk judgments of technology performance of phytoremediation: self-confidence ratings, averaging procedures, and formative consensus building. Risk Anal 27(1):225–240
Schum DA (1987) Evidence and inference for the intelligence analyst, vol II. University Press of America, New York
Schummer J, Vohra RV (2007) Mechanism design without money. In: Nisam N, Roughgarden T, Tardos E, Vazirani VV(eds) Algorithmic game theory. Springer, New York, pp 317–332
Singh S, Way CR (2004) The correlates of nuclear proliferation. J Conflict Resolut 48(6):859–885
Staff (1984) Ill handlers suspected in Oregon food poisonings. The New York Times. http://www.nytimes.com/1984/10/21/us/ill-handlers-suspected-in-oregon-food-poisonings.html. Retrieved 15 Sep 2011
Stech FJ, Elasser C (2004) Midway revisited: deception by analysis of competing hypothesis. MITRE Corporation [online]. http://www.mitre.org/work/tech_papers/tech_papers_04/stech_deception/stech_deception.pdf

Sticha P, Buede D, Rees RL (2005) APOLLO: an analytical tool for predicting a subject's decision making. Presented at the international conference on intelligence analysis methods and tools, McLean

Torok TJ, Tauxe RV, Wise RP, Livengood JR, Sokolow R, Mauvais S, Birkness KA, Skeels MR, Horan JM, Foster LR (1997) A large community outbreak of salmonellosis caused by intentional contamination of restaurant salad bars. J Am Med Assoc 278(5):389–395

United States Senate Select Committee on Intelligence (2004) Report on the U.S. Intelligence Community's Prewar Intelligence Assessments on Iraq, One Hundred Eighth Congress, Second Session. U.S. Government Printing Office, Washington, DC

Valtorta M, Dang J, Goradia H, Huang J, Huhns M (2005) Extending Heuer's analysis of competing hypotheses method to support complex decision analysis. International conference on intelligence analysis, Washington, DC

van Gelder T (2008) Can we do better than ACH? Australian Institute of Professional Intelligence Officers News, 55

Walshe T, Burgman M (2010) A framework for assessing and managing risks posed by emerging diseases. Risk Anal 30(2):236–249

Wang H, Druzdzel M (2000) User interface tools for navigation in conditional probability tables and elicitation of probabilities in Bayesian networks. In: Proceedings of the sixteenth conference on uncertainty in artificial intelligence. Morgan Kaufmann, San Francisco

Weaver J (1985) The town that was poisoned (PDF). Congressional Record 131(3–4). United States Government Printing Office, Washington, DC, 99th United States Congress, 1st session, pp 4185–4189. Transcription at WikiSource. http://commons.wikimedia.org/wiki/File:1985_Feb_28_Congressman_Weaver_THE_TOWN_THAT_WAS_POISONED.pdf. Retrieved 15 Sep 2011

Whitney CR (2005) The WMD mirage. Public Affairs, New York

Winkler RL, Clemen RT (2004) Multiple experts vs. multiple methods: combining correlation assessments. Decis Anal 1(3):167–176

Witteman C, Renooij S (2002) Evaluation of a verbal-numerical probability scale. Int J Approx Reason 33:117–131

Zlotnick J (1972) Bayes' theorem for intelligence analysis. Stud Intell 16(2). https://www.cia.gov/library/center-for-the-study-of-intelligence/kent-csi/vol16no2/pdf/v16i2a03p.pdf