Soft Evidential Update for Probabilistic Multiagent Systems

Marco Valtorta*
Young-Gyun Kim
Department of Computer Science and Engineering
University of South Carolina
Columbia, SC 29208, U.S.A.

Jiří Vomlel
Department of Computer Science
University of Aalborg
Aalborg, Denmark

June 15, 2000

Abstract

We address the problem of updating a probability distribution represented by a Bayesian network upon presentation of soft evidence. Our motivation is the desire to let agents communicate with each other by exchanging beliefs, as in the Agent Encapsulated Bayesian Network (AEBN) model, and soft evidential update (under several different names) is a problem with a long history. We give methodological guidance to model soft evidence in the form of beliefs (marginals) on single and multiple variables, propositional logical formulae (arbitrary events in the universe of discourse), and even conditional distributions, by introducing *observation variables* and explaining their use. The extended networks with observation variables fully capture the independence structure of the model, even upon receipt of soft evidence. We provide two algorithms that extend the celebrated junction tree algorithm, process soft evidence, and have different efficiency characteristics. One of the extensions, the *big clique algorithm*, promises to be more time efficient at the cost of possible space penalties. The other extension requires only minimal modifications to the junction tree at the cost of possibly substantial time penalties. Our results open new avenues of application for graphical probabilistic models.

Keywords: Bayesian Networks, Graphical Probabilistic Models, Iterative Proportional Fitting Procedure (IPFP), Conditional IPFP (CIPFP), Jeffrey's Rule, Evidential Updating, Soft Evidence, Virtual Evidence.

*Corresponding author, mgv@cs.sc.edu

1 Introduction and Motivation

We address the problem of updating a probability distribution represented by a Bayesian net upon the presentation of soft evidence. We call this the problem of soft evidential update.

The motivation for this work is our desire to let agents that use probabilistic models (and especially Bayesian nets) communicate with each other by exchanging beliefs.

While this is not the focus of this paper, we need to describe briefly our agent model, which is called the Agent-Encapsulated Bayesian Network (AEBN) model, originally due to Bloemeke [Blo98]. Each agent in an AEBN model uses as its model of the world a single Bayesian network (which we also call an AEBN). The agents communicate via message passing. Each message is a distribution on variables shared between the individual networks.

The variables of each AEBN are divided into three groups: those about which other agents have better knowledge (*input set*), those that are only used within the agent (*local set*), and those of which the agent has the best knowledge, and which other agents may want (*output set*). The variables in the input set and the output set are shared, while those in the local set are not. An agent consumes (or subscribes to) zero or more variables in the input set and produces (or publishes) zero or more variables in the output set.

The mechanism for integrating the view of the other agents on a shared variable is to replace the agent's current belief in that variable with that of the communicating agent. When an agent receives a message from a publisher, it modifies the probabilities in its internal model, so that its local distribution either becomes consistent with the other agent's view or is inconsistent with it. In the rest of the paper, we assume the former case and provide neither a method for identification of inconsistency nor a method to deal with inconsistency, but note that Vomlel [Vom99] presents some methods for dealing with inconsistent evidence. Also cf. [GS93, ACS96].

Therefore, after updating using all evidence, we still require that all appropriate marginals of the updated distribution be equal to the evidence entered. The deservedly celebrated junction tree algorithm for probability update [LS88, SS90, Jen95, LJ97] was not designed to satisfy this requirement, and in fact it does not, as we will show in Section 4.1.

When a publisher makes a new observation, it sends a message to its subscribers. In turn, the subscribers adjust their internal view of the world and send their published values to their subscribers. Assuming that the graph of agent communication (which we simply call *agent graph* as in [Blo98]) is a DAG, equilibrium is reached, and a kind of global consistency is assured, because the belief in each shared variable is the same in every agent.

The restriction that one of the agents has oracular knowledge of a variable may seem excessive. However, it is permissible to have multiple views of a common variable. For example, in a multiagent system for interpretation, two agents may issue a report that corresponds to the same (unknown) physical quantity. Nothing prevents another agent from integrating the reports of these agents and effectively obtain a new (and possibly more accurate) view of the same underlying quantity. As another example, it is possible for a subscriber agent to model known reporting errors or biases of the publisher, as we will show in Section 2.2.

When the agent graph is not a tree, great care must be taken to deal with undirected cycles (loops). Such cycles lead to possible double counting of information, which is often known as the rumor problem. We do not address this important problem in this paper, but cf. [Blo98, BS99, BV]. It is also possible to consider directed cycles (and in particular, the tight cycles resulting from bi-directional communication in agent graphs), by appropriately sequencing messages between agents. We do not address this extension further in this paper, but the reader must be made aware of the very interesting and important work by Xiang on Multiply Sectioned Bayesian networks for related results [Xia96, XL00].

The rest of paper is organized as follows. In Section 2, we explain the notion of soft evidence and give some simple examples (in 2.1) and a more complex example that also illustrates the AEBN model (in 2.2). Section 3 is the major methodological part of our paper. We begin (in 3.1) by reviewing Jeffrey's rule and some of the problems that arise from its uncritical use, as presented by Pearl [Pea88, Pea90]. We then proceed by describing (in 3.2) our approach to Bayesian network modeling for soft evidential update, which is based on the notion of *observation variable*, is very general, and is simple to use. Section 4 is devoted to algorithmic issues. We begin by showing (in 4.1) that the junction tree algorithm does not support soft evidential update. We then present two modifications of the junction tree algorithm that support soft evidential update. The first modification (4.3) is iterative and does not require any additional space with respect to the original algorithm. The second modification is the *big clique algorithm* (4.4), which requires additional space but is more time efficient than the iterative algorithm. In Section 5, we concentrate on some advanced issues, including the treatment of conditional evidence and the detailed analysis of an example due to Pearl. Section 6 contains a summary and evaluation of our work and suggestions for future work.

2 Hard and Soft Evidence

We begin by defining what we mean by soft evidence. Evidence is a collection of findings on variables. A finding may be hard or soft. A *hard finding* specifies which value a variable is in. A *soft finding* specifies the probability distribution of a variable. *Hard evidence* is a collection of hard findings. *Soft evidence* is a collection of soft findings.

The definition of soft evidence could be generalized in three ways. Firstly, we may extend the definition of finding to allow conditional distributions. Secondly, we may allow joint (and possibly, conditional) distributions on a collection of variables. Thirdly, we may allow distributions on arbitrary events (equivalently, arbitrary logic formulae). These three extensions can be handled by the introduction of special *observation variables* as will be shown in Section 3.2.

2.1 Two Simple Examples

Here are two simple examples that illustrate the notion of soft evidence.

Suppose that the initial position and direction of an object are known precisely and that its acceleration is known to be zero. The position (P) of the object (which we will call a target) is then determined by its actual speed (S),

according to some probabilistic law. This simple situation may be modeled by the Bayesian network in Figure 1, together with appropriate conditional probability tables for $P(S)$ and $P(P|S)$.

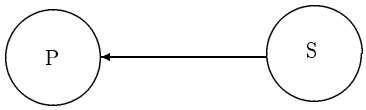

Figure 1: Position and speed.

Suppose that we cannot measure the speed precisely, but that we have uncertain information about it. We may think of E as the speed returned by a remote sensor. The sensor is embodied into a separate agent, whose task is to provide an estimate of speed. Since we live in the world of Bayesian networks, the estimate is a belief, i.e. a probability distribution over the possible values of speed.

We call this kind of evidence *soft evidence*. The new situation may be modeled by Figure 2, where the dashed line indicates that the evidence (E) about S is uncertain.

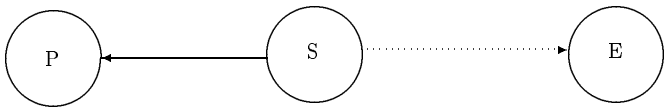

Figure 2: Position and speed with soft evidence about speed.

In our second example, position (P) depends on speed (S), again. We again have a separate sensor agent that provides us with soft evidence about the speed of the target. In contrast to the situation in the previous example, however, we now know that the state of the sensor (SS) that gives us the speed is affected by the weather (W). The situation is described in Figure 3.

Note that a finding (even when uncertain) about speed sets up a dependence between the weather and the target's position. In particular, if we knew the position and had evidence for the speed, we could tell the weather. Qualitatively, the argument goes like this: the position determines (although uncertainly) the speed of the target; whatever variability remains in the measured speed is explained by the weather (through the sensor model).

2.2 A More Complex Example

We extend the cow pregnancy network of [Jen95] to a three-agent system. One of the agents represents a farmer who needs to evaluate the probability that one of his or her cows is pregnant. The other agents represent a Urine Test (UT) expert and a Scanning Test (ST) expert, respectively.

The farmer subscribes to variable UT, which is published by the UT expert, and to variable ST, which is published by the ST expert, as indicated in Figure 4. While we require that the distribution of a variable remain the same across agents, nothing prevents the farmer from having a model of the experts and

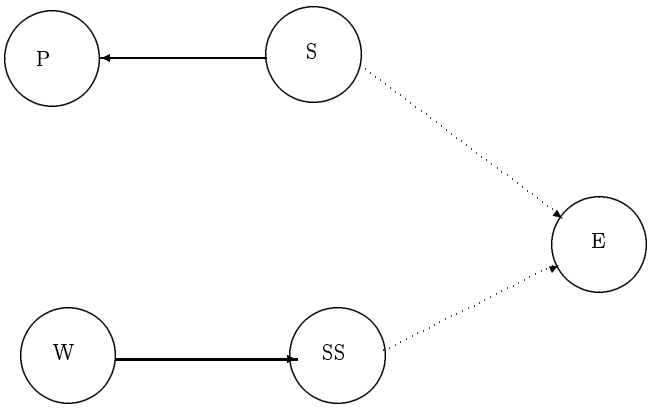

Figure 3: Speed with a sensor whose accuracy depends on the weather.

therefore somehow discounting their advice, as indicated in Figure 5, where a simple model of the reliability and sensitivity of the two experts is encoded in the link between the primed variables (UT' and ST'), which represent the farmer's view of the results of the Urine and Scanning tests, respectively, and the unprimed variables (UT and ST), which are the test results as communicated by the experts. The farmer possesses a more complicated model of the two experts in the situation described in Figure 6, where it is assumed that the farmer knows that the experts' advice is affected by some environmental factor E.

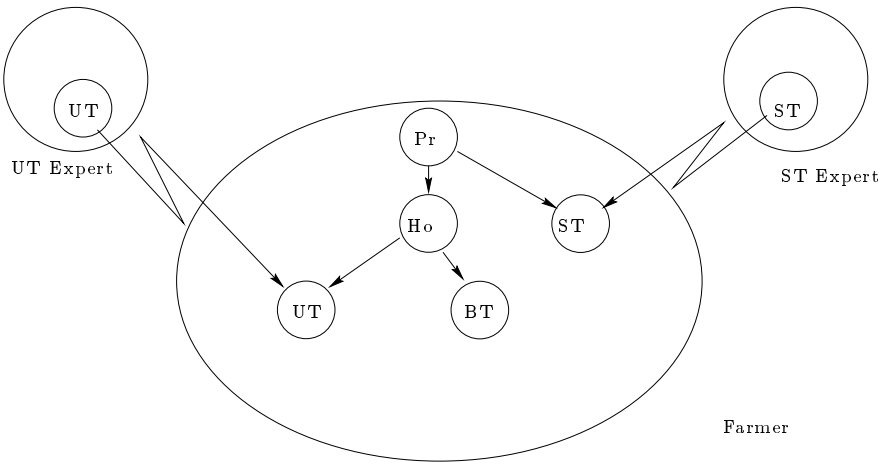

Figure 4: The agent graph of a three-agent system.

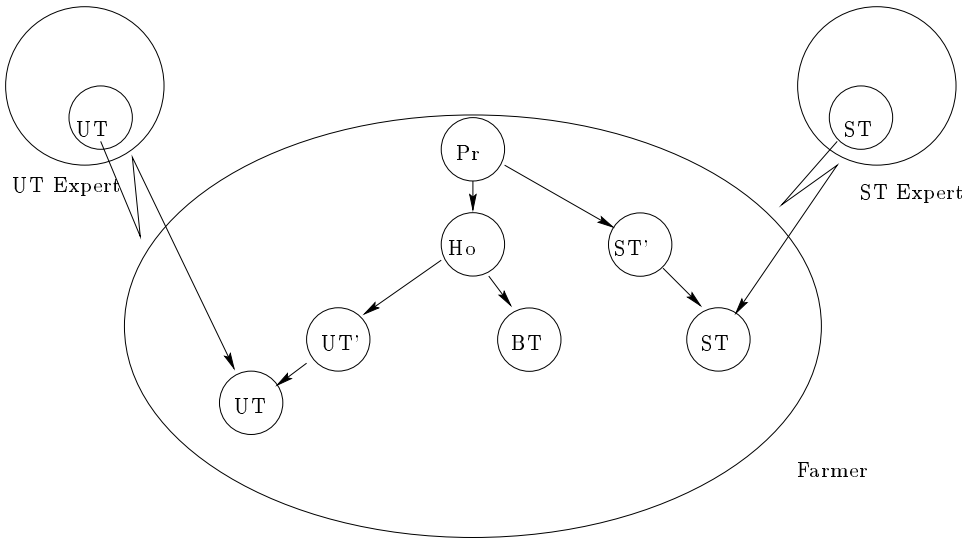

Figure 5: Agent graph of Figure 4 with simple models of the experts.

3 Modeling Issues

3.1 Jeffrey's Rule

Jeffrey's rule, also known as the rule of probability kinematics, provides a way to update a probability distribution from soft (uncertain, non-categorical, non-propositional) evidence. In this Section, we introduce the rule and some problems with its application, following the presentation by Pearl [Pea88, Pea90]. In the next subsection, we introduce a general modeling framework for soft evidence.

Jeffrey's rule can be written as: $Q(A) = \sum_i P(A|B_i) \cdot Q(B_i)$,
where $Q(B)$ is soft evidence, and $P(A|B)$ is the conditional probability of A given B *before evidence.*

Jeffrey's rule applies in situations in which $P(A|B)$ is invariant w.r.t. $P(B)$. This is not always the case, as shown in the following example, which is an extension by Pearl of an example by Jeffrey himself [Pea88].

The worth (W) of a piece of cloth depends on its color (C), as modeled by the Bayesian network structure of Figure 7, with $P(C)$ and $P(W|C)$. Before observing the cloth (whose color we know to be one of Green, Blue, and Violet), our belief is summarized by $P(Color) = (.3, .3, .4)$. We then observe the cloth by candlelight. The light is not good enough to allow us to distinguish the color precisely, but we revise our belief about the color of the cloth to $(.7, .25, .05)$. How should we revise the worth of the cloth?

Jeffrey's rule for this case is: $Q(W) = \sum P(W|c_i) \cdot Q(c_i)$. The rule applies only if $P(W|c_i)$ is not modified by the observation, which is clearly true in this example. Diaconis and Zabell comment that "this condition is an internal or psychological condition that must be checked or accepted at each stage. Mathematics has nothing to offer here" [DZ82, p.825]. There is, however, a

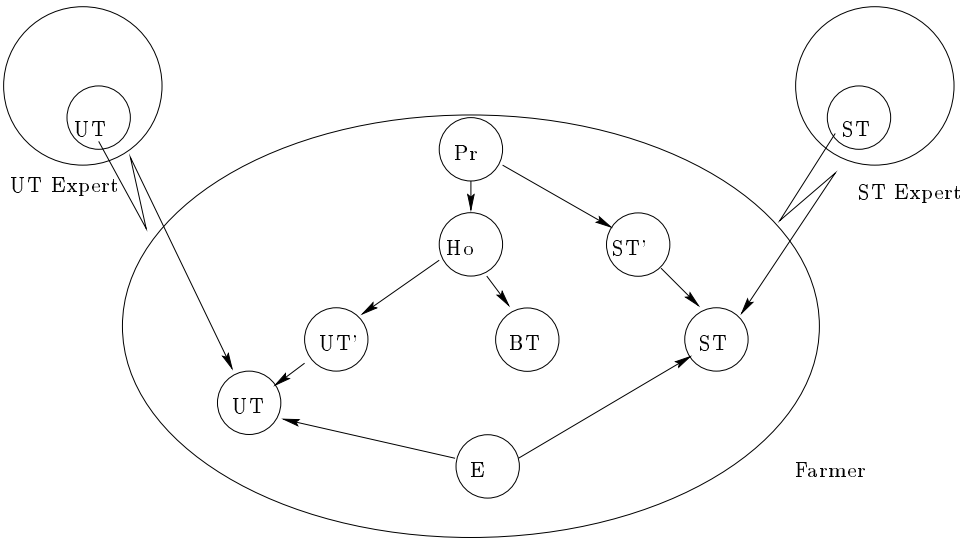

Figure 6: Agent graph of Figure 4 with complex models of the experts.

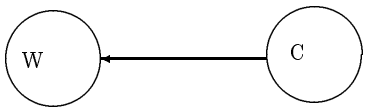

Figure 7: The worth of a cloth.

principled way to check whether the condition holds, which is based on the theory of Bayesian networks.

Augment the Bayesian net with a node that represents the evidence (E) and the appropriate edges. In the example, the evidence depends on the (true) color of the cloth, as shown in Figure 8. Now, check whether W is independent of E given C (using a d-separation algorithm). If this is the case, then $P(W|C) = P(W|C,E)$, and Jeffrey's rule may be applied. In our example, this is the case.

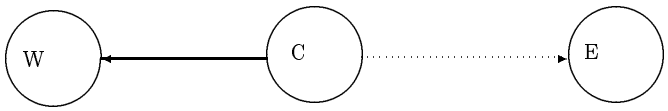

Figure 8: The worth of a cloth in the presence of soft evidence.

Let us now consider a situation in which Jeffrey's rule is not applicable. Suppose that we know the color of the cloth, and we want to update our belief in the type of candle (CA) used. In the absence of observations, the type of candle used does not affect the color of the cloth, and vice versa, so that:
$P(CA|C) = P(CA)$,
and the Bayesian network structure is given in Figure 9.

Figure 9: Candle type and cloth color.

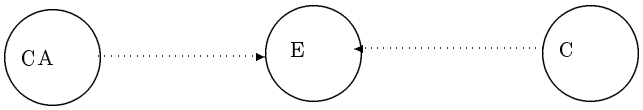

Figure 10: Candle type and cloth color in the presence of soft evidence.

The Bayesian network structure in the presence of soft evidence for Color is given in Figure 10. Since CA and C are not d-separated by E, $P(CA|C) \neq P(CA|C,E)$, and therefore Jeffrey's rule is not applicable.

Pearl [Pea88, Pea90] observes that the virtual evidence method (also cf. [Kyb87, Nea90]) can be viewed as formally equivalent to the likelihood ratio version of Jeffrey's rule. Suppose that we obtain evidence on a variable B, with values b_i. Letting $Q(b_i)$ be the evidence and $P(b_i)$ be the prior probability of B, the ratio $Q(B)/P(B)$ may be used in the virtual evidence method, resulting in an updated distribution whose marginal over B is the soft evidence $Q(B)$. We will exploit this result in Section 4.3.

3.2 A Unified Approach for Modeling Soft Evidence

Suppose that an agent-encapsulated Bayesian network needs to be updated in the light of information from another agent. There are two types of observed variables in a multiagent system:

1. One observed state of a random variable X is reported, but a publishing agent is unsure of its observation. E.g. a particular flip of a coin came up heads, but the agent is not sure since it has observed the coin from a long distance.

2. The probabilities that a random variable X takes certain states $P(X = x_i),\ i = 1, \ldots, n$ are reported. E.g. an agent will report $P(X = head) = 0.5$ and $P(X = tail) = 0.5$ in the case it believes that it has observed a fair coin.

There is an approach to update from non-crisp information that is often used for in Bayesian networks, e.g. in the Hugin program [AOJJ89], and which corresponds to the first type of variable. Pearl calls it *virtual evidence update* in [Pea88]. (Also cf. [Hec86, HH86, Kyb87, Pea90, Nea90].) Under certain conditions, which we discuss below, virtual evidence update may be an appropriate way to perform inter-agent communication. The second approach, for which we present a comprehensive analysis in this paper, corresponds to the second type of variable. We call it *soft evidential update*. Next, we will describe the basic differences between virtual evidence update and soft evidential update.

Virtual evidence update is a model of updating suitable for situations when agents obey the following property:

(V1) A publishing agent provides its belief in the true state of the observed variable. Its observation can be improved by something that the publishing agent can not observe. Variables in the subscribing agent model may have an influence on the true state of the variable observed by the publishing agent.

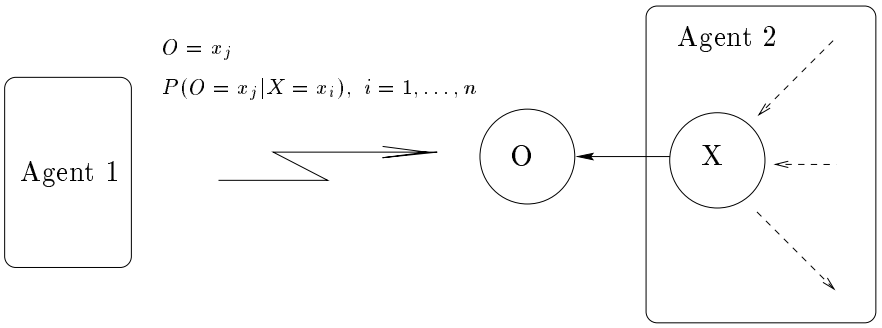

Figure 11: Virtual evidence update

A model of virtual evidence update is displayed on Figure 11. It consists of the following three steps:

1. A dummy node O is created. It has the same states as node X, and X is its only parent.

2. A publishing agent reports one observed state $x_j \in \{x_1, \ldots, x_n\}$ from n possible states of an observed variable X together with his reliability in the form of conditional probability distribution $P(O = x_j \mid X = x_i),\ i = 1, \ldots, n$ (Note that only values for the observed value j are required). This conditional probability distributions read as: "The probability of observing variable X being in state x_j if its true state is x_i."

3. Updating of the Bayesian network is performed. This corresponds to entering evidence $O = x_j$ with consequent propagation using the values of $P(x_j \mid X = x_i),\ i = 1, \ldots, n$. E.g. the junction tree algorithm [LS88, SS90, Jen95, LJ97] can be employed.

The approach for modeling soft evidence we propose is fundamentally different. The soft evidential update approach is suitable when the following assumption holds:

(S1) The publishing agent's belief on states of a variable cannot be improved by anything that is observed later, i.e. no variable in the model of a subscribing agent may have any influence on that publishing agent's observation.

The model of this approach is displayed on Figure 12. Note that, in contrast with virtual evidence update, the report of the publishing agent does not include

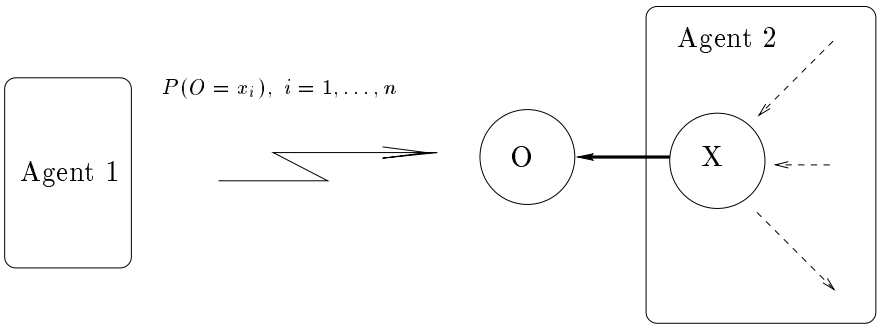

Figure 12: Soft evidential update

the conditional probability table $P(X \mid O)$ but consists only of a distribution on the observation variable.

This kind of situation exists in AEBNs, in which there is precisely one agent with specific authority for each variable that occurs in multiple agents. In this case, it is correct for the subscriber agents to treat their initial belief in that variable not as a prior probability to be updated upon receipt of additional information (as in the virtual evidence method), but as a guess to be *replaced* by the evidence provided by the authority. The soft evidential update approach more closely models the notion of observation as a property of the publishing agent and insures consistency across all agents that subscribe to the published variable, in the sense that the belief (marginal posterior probability) in a published variable is the same throughout the agent system.

The soft evidential update approach allows to combine the variable observed by the publishing agent and the subscribing agent's own opinion on the same variable, by the introduction of additional variables with the same domain. Dependence is modeled within the structure owned by the subscribing agent. By using this simple technique, it is reasonable to require that the probability distribution provided by the subscribing agent should not be changed (assumption S1) in a wider range of situations than it may superficially appear. In contrast with requiring that agents correspond by communicating either hard evidence or likelihood ratios, we can accommodate absolute beliefs, providing that the agent that uses them model them as such.

Next, it will be shown that using the soft evidential update approach several types of soft evidence can be handled in a unified manner. The update consists of the following three steps, which result in an extension of the original Bayesian network.

- First, we create a new node, which we call *observation node*, for each observation. Let O be a generic such node. Every state of node O corresponds to a possible outcome of an observation. We can handle any of the following types of soft evidence:

  (1) one-dimensional marginal probability distributions that are defined for values of a single variable,

  (2) higher-dimensional marginal probability distributions that are defined on Cartesian product of the values of two or more variables,

(3) conditional probability distributions,

(4) probability assignments to an arbitrary logical function, and

(5) probability assignments to a probabilistic function.

- Second, we add directed edges to node O from all nodes in the Bayesian network structure that have a direct influence on the observation, so that the set of parents of O in the extended network d-separates O from the rest of the network.
- Third, we model the dependence of the parents of O ($\mathrm{pa}(O)$) on O, by specifying the conditional probability table $P(O \mid \mathrm{pa}(O))$.

See Figure 13 where all the types of soft evidence listed above are displayed. Bold lines denote logical dependence while the others stand for a probabilistic dependence. We now address each of the five types of soft evidence individually.

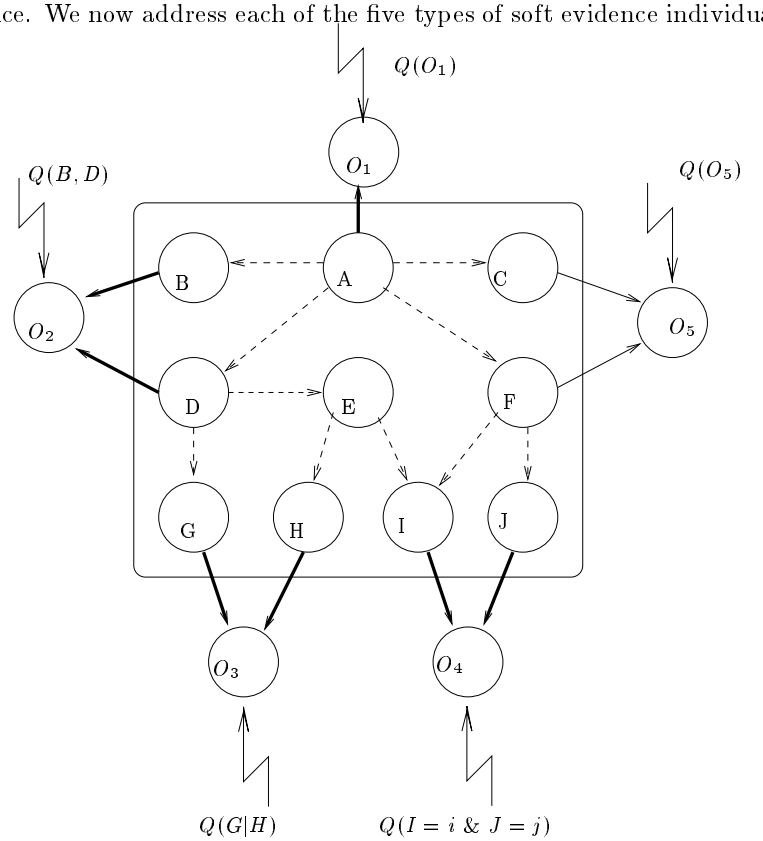

Figure 13: Example of an agent with five different types of soft evidence

In the first case, the observation corresponds to a one-dimensional marginal probability distribution that is defined for all values of a single variable. A child of a single node is created, such as O_1 (the child of A) in Figure 13. Child O is an exact copy of its parent $X \in pa(O)$. It has the same states as its parent X and is logically dependent on X, i.e.

$$P(O = o \mid X = x) = 1 \ \Leftrightarrow \ o = x$$

In the second case, the observation corresponds to a higher-dimensional marginal probability distribution that is defined on the Cartesian product

$$\mathbb{X}_{pa(O)} = \times_{i \in pa(O)} \mathbb{X}_i$$

of the set of values $\mathbb{X}_i$ of variables X_i, where $i \geq 2$. We create a single node, such as O_2 in Figure 13, that has all possible combinations of the values of its parents as its values, i.e. there exist a single state o of O for every $\vec{x} = (x_{i_j})_{i \in pa(O)} \in \mathbb{X}_{pa(O)}$. $X_{pa(O)} = \{X_i\}_{i \in pa(O)}$ denotes the multidimensional random variable taking values $\vec{x} \in \mathbb{X}_{pa(O)}$. Variable O is again logically dependent on its parents, i.e.

$$P(O = o \mid X_{pa(O)} = \vec{x}) = 1 \;\Leftrightarrow\; o \text{ corresponds to } \vec{x}$$

The third case is, with respect to modeling, equivalent to the second. The only difference is in the way the update of the probability distribution on the node (O_3 in Figure 13) is done (which is discussed in Section 4.2).

Note that in the previous three cases the parent nodes could be updated directly. The only reason for the introduction of node O_1 is that we want to handle evidence of all types listed above within a unified framework. In the second and third case, in addition to the previous reason, the introduction of observation nodes can be also understood as a modeling trick that allows us to update a distribution on a single node rather than on a set of variables. It is not necessary to add the observation nodes if we have a direct access to higher dimensional probability distributions in the computer representation of the Bayesian network.

In the fourth case, the observation is a logical function of some variables in the Bayesian network (i.e., some event), the probability table $P(O \mid \text{pa}(O))$, such as $P(O_4 \mid I, J)$ in Figure 13, models a logical dependence between the parent variables and therefore only contains zeroes and ones. Typically, the observation variable has as its values only two logical complements, i.e. values of O corresponds to a partition of $\mathbb{X}_{pa(O)}$. The possibility of updating on collections of subsets of $\mathbb{X}_{pa(O)}$ that are not partitions was discussed in [DZ82, Section 5.2]. For an example that involves a logical function of variables, see Section 5.

The fifth case is a generalization of the fourth one. In this case in order to update correctly using soft evidence, we need to specify fully the conditional probability table $P(O \mid \text{pa}(O))$, e.g. $P(O_5 \mid C, F)$ in Figure 13. Typically, the possible values of the observation variable O are the same as the possible values of one of its parents and the conditional probability distribution is used to model the influence of the other parents' values on the observation.

Figures 2 and 8 show situations in which the observation (E in both cases) is a child of a single node (S and C respectively). Figures 3 and 10 show situations in which the observation (E in both cases) is a child of two nodes. The two parent nodes are S and SS in the case of Figure 3 and C and CC in the case of Figure 10.

Consider now the case of the specification of the conditional probability for the speed sensor. To update correctly using soft evidence, we need to specify $P(E \mid S, SS)$. Note that such specification only involves the possible values of variables O, S, and SS, and does not require consideration of belief states. In fact, the possible values of O are the same as the possible values of S. In the case of the color of the cloth, we need to specify $P(E \mid C, CC)$. Again, only

the possible values of the three variables E, C, and CC need to be considered. Here, the possible values of E are the same as the possible values of C.

4 Algorithmic and Systems Issues

4.1 Inadequacy of the Junction Tree Algorithm

We show, by a simple example, that the junction tree algorithm does not treat soft evidence properly. For convenience, we refer to the so-called Hugin variant of the junction tree algorithm, as described, e.g., in [Jen95], and use similar terminology.

The skeleton of the argument is as follows. In the junction tree algorithm, messages are passed across separators from clique to clique. Exactly two messages are passed between two cliques (say, C_i and C_j), one in each direction, as shown in Figure 14. The first message is passed during the DistributeEvidence phase of the method (say, from C_i to C_j), the second during the CollectEvidence phase (say, from C_j to C_i). Suppose that clique C_i contains a node (say, V_i) for which we have a soft finding (say, $P(V_i)$), When C_i sends its message to C_j, $P(C_j)$ in modified (by calibration). After the probability of C_j has been fully updated (say, to $Q(C_j)$), C_j sends its message to C_i, and the probability of C_i is modified by calibration. In general, letting $Q(C_i)$ be the modified probability, we have that $\sum_{C_i \setminus \{V_i\}} Q(C_i) \neq P(V_i)$, which implies that the soft finding $P(V_i)$ is not treated as evidence.

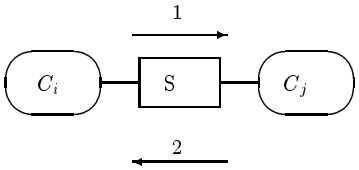

Figure 14: The soft finding $P(V_i)$ is not treated as evidence.

We now present the promised example that establishes our claim that the junction tree algorithm does not handle soft evidence properly. We use the "wet grass" Bayesian network, originally in [Pea88], as elaborated in [Jen95]. The network is given in Figure 15, with all variables binary (with values y and n, in that order), $P(R) = (0.2, 0.8)$, $P(S) = (0.1, 0.9)$, and the conditional probabilities in Tables 1 and 2.

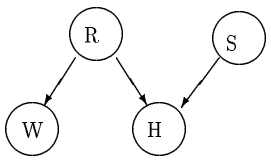

Figure 15: The wet grass Bayesian network structure.

	R	
W	y	n
y	1	0.2
n	0	0.8

Table 1: Table for $P(W|R)$ in the wet grass example.

	R	
S	y	n
y	(1,0)	(0.9,0.1)
n	(1,0)	(0,1)

Table 2: Table for $P(H|R,S)$ in the wet grass example. The first entry in each vector is for $H = y$.

The junction tree algorithm constructs the junction tree shown in Figure 16. Suppose that our hard evidence is that Holmes's lawn is certainly dry (i.e., $H = n$), while we have soft evidence that Watson's lawn, in the form $P(W) = (0.7, 0.3)$. If this soft evidence is absorbed in clique WR and the hard evidence is absorbed in clique HRS, propagation will consist of two messages through the separator R. After the messages are passed, P is updated to Q, and $Q(W) = (.4439, .5561) \neq (.7, .3) = P(W)$.

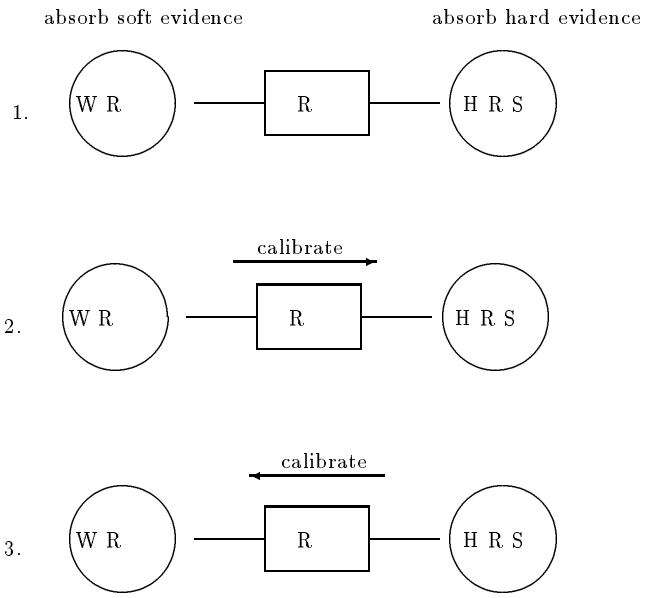

Figure 16: The junction tree algorithm and the wet grass network

We have shown that even if soft findings are absorbed correctly into cliques, the propagation method itself would not respect the evidence characteristics of

the findings.

The reader may wonder why examples such the one above do not also apply in the case of hard evidence. The reason is that it is not possible to enter new evidence that contradicts the evidence already entered, and therefore zero entries are treated in a special way by the junction tree algorithm: zeroes in probability tables remain zeroes after each message. (Cf. [Jen95, Lemma 4.1].)

4.2 Soft Evidential Update Method

We propose a universal approach (the *soft evidential update method*) for updating of a joint probability distribution $P(V)$, where V denotes the set of all variables in the model, in the light of soft evidence, which may correspond to any one of the five types discussed in Section 3.2.

(1) At first, one or more (say k) observation variables are defined as described in in Section 3.2. For each of the k observation variables, the values of probability distribution $Q(O_i = o_i), i = 1, \ldots, k$ are provided by the corresponding agent. The values correspond to the soft evidence, e.g. the beliefs that were assigned to the observed events. In the case (3) when a conditional probability distribution is provided by the agent, additional computation is necessary. The formulas are provided below (formulae 2 and 3); further details can be found in Section 5.

(2) The conditional probability distribution $P(V \setminus O \mid O)$ given the values of the observation variable is computed.

(3) The updated joint probability distribution is achieved by multiplication of conditional distribution $P(V \setminus O \mid O)$ with soft evidence stored in $Q(O)$,

$$Q(V) \quad = \quad P(V \setminus O \mid O) \cdot Q(O) \tag{1}$$

(4) If there is only one observation variable, then the formula above corresponds to Jeffrey's rule. If there are more observation variables $O_1, \ldots O_k$, we must update all of them and iterate until the difference between two subsequent cycles of the procedure is sufficiently small.

As we mentioned above, when a conditional probability distribution $Q(X_A \mid X_B)$ is provided by the agent, additional computation is necessary. $X_A = \{X_i\}_{i \in A}$ and $X_B = \{X_i\}_{i \in B}$ are (possibly) multidimensional random variables, where A, B denote two disjoint sets of variable indices. In the following formulae o denotes a particular value of variable O corresponding to the vector $\langle \vec{a}, \vec{b} \rangle$ concatenated from vectors $\vec{a}$ and $\vec{b}$, while $\vec{a}$ and $\vec{b}$ are particular values of the random variables X_A and X_B, respectively. There are two possibilities (discussed further in Section 5) as to how distribution $Q(O)$ can be computed. For every state o of the observation variable O the probability distribution $Q(O)$ is defined either by formula 2 or 3.

$$Q(O = o) \quad = \quad Q(X_A = \vec{a} \mid X_B = \vec{b}\,) \cdot P(X_B = \vec{b}\,) \tag{2}$$

$$Q(O = o) \quad = \quad Q(X_A = \vec{a} \mid X_B = \vec{b}\,) \cdot P(X_B = \vec{b}\,) \cdot c \cdot \exp\left(I\left(Q(X_A \mid X_B = \vec{b}\,) \parallel P(X_A \mid X_B = \vec{b}\,)\right)\right) \tag{3}$$

where c is a normalization constant given by:

$$c \quad = \quad \sum_{\vec{b}} \exp\left(I\left(Q(X_A \mid X_B = \vec{b}\,) \parallel P(X_A \mid X_B = \vec{b}\,) \right) \right)$$

and $I(P \parallel Q)$ denotes the well known *I-divergence* (also called Kullback-Leibler divergence or cross entropy) defined as

$$I(P \parallel Q) \quad = \quad \sum_{x} P(x) \log \frac{P(x)}{Q(x)} \tag{4}$$

The described soft evidential update method obeys several nice properties that are discussed in detail in Section 5. In the next two subsections, we present two approaches to an effective implementation of the soft evidential update method. Both approaches are based on the junction tree algorithm, which is modified to make it sound for soft evidential updates. The first approach requires no changes to the junction tree itself, but iteration on all variables is required. The second approach requires a different kind of triangulation, but iteration is localized to a single clique. The reader will recognize that the first approach is more space efficient, while the second approach is more time efficient.

4.3 Iterative Modification of the Junction Tree Algorithm

Three steps are involved.

1. Enter hard evidence and propagate normally. Call the resulting distribution Q. Q will be updated repeatedly by the following iterative process.

2. Iterate the following cycle (consisting of k applications of the pair of steps (2a) and (2b)) until convergence is achieved.

3. For each piece of soft evidence, $P(O_j), j = 1, \ldots, k$:

   (a) Absorb $P(O_j)$ in a clique C_{l_j} that contains O_j using the Iterative Proportional Fitting Procedure (IPFP) formula:

   $$Q(C_{l_j}) = \frac{P(C_{l_j}) \cdot P(O_j)}{Q(O_j)}$$

   (b) Propagate normally.

The convergence of this procedure follows from general results about the convergence of IPFP, as shown originally in [Csi75]. (Also, cf. [Jir91, Vom99].) We note that the procedure could, in principle, be implemented as a wrapper around the Hugin shell or other shells that support the virtual evidence method. Care must be taken, because the virtual evidence method requires that likelihoods be entered, rather than probabilities, and therefore soft evidence values must be divided by the appropriate prior probabilities at each cycle.

4.4 The Big Clique Algorithm

The *big clique algorithm* modifies the junction tree algorithm as follows:

1. Build a junction tree that includes all variables for which soft evidence is given in one clique, the *big clique* C_1. (These variables may appear in other cliques as well.)

2. Update $P(V)$ to a distribution $P^*(V)$ by executing the junction tree algorithm using only hard evidence. $P^*(V)$ is a distributed representation of $P(V|$ hard evidence), in the sense of the remark following Theorem 4.2 in [Jen95]: the product of all clique tables divided by the product of all separator tables is equal to $P(V|$ hard evidence).

3. Absorb all soft evidence in C_1 (with the algorithm described in Section 4.4.1).

4. Call the routine DistributeEvidence from C_1. This routine and the correctness of this step are presented in Section 4.4.2.

4.4.1 Absorption of Soft Evidence

We define absorption in the special big clique C_1 as the process by which the joint probability $P(C_1)$, is updated to conform to soft evidence on variables $A \subseteq C_1$, where $A = \{A_1, A_2, \ldots, A_k\}$. Let $Q(C_1)$ be the joint probability after absorption. Then $\forall i \ \sum_{C_1 \setminus A_i} Q(C_1) = P(A_i)$, where $P(A_i)$ is the soft evidence on $A_i, i = 1, \ldots, k$. Absorption of soft evidence in clique C_1 is carried out by using the Iterative Proportional Fitting Procedure (IPFP) and consists of cycles of k steps, one per finding. Each step corresponds to one soft finding. The appropriate formulae are:

$$\begin{aligned} Q_{(0)}(C_1) &= P(C_1) \\ Q_{(i)}(C_1) &= \frac{Q_{i-1}(C_1) \cdot P(A_j)}{Q_{i-1}(A_j)} \end{aligned}$$

where $j = (i-1) \bmod k + 1$.

For a simple example, suppose we have the clique $\{A, B\}$ with joint probability as given in Table 3 (all variables are binary). Suppose that soft evidence on variable B is available in the form of $P(B) = (.7, .3)$. We compute the updated joint probability $Q(B)$ in two steps. The result of the multiplication by $P(B)$ is in Table 4. The result of the division by $Q_{(0)} = (.59, .41)$ is in Table 5. Note that $\sum_{\{A,B\}\setminus\{B\}} Q(A, B) = P(B) = (0.7, 0.3)$, as claimed.

One step of IPFP is sufficient when there is only one soft finding. In general, however, several cycles may be necessary for IPFP to converge. See [Csi75, Hab74, Vom99] for the proof of convergence in the general discrete case and for bounds on the number of cycles in special cases.

4.4.2 Propagation of Soft Evidence

First, recall that Step 2 in the modified junction tree algorithm leads to a distributed representation of the posterior probability of all variables given all

	A	
B	y	n
y	.56	.03
n	.14	.27

Table 3: Table for $P(A, B)$.

	A	
B	y	n
y	.392	.021
n	.042	.081

Table 4: Table for $P(A, B, e)$.

hard findings. Second, observe that the product of the table for the special clique that has absorbed all soft evidence (as done in Step 3) multiplied by the tables for the other cliques and divided by the tables of the separators is a representation of the posterior probability of all variables given the soft evidence and the hard evidence. (Hard evidence had already been absorbed in Step 2).

We now need to restore consistency between the special clique and the other clique. To do so, we propagate from the clique that has absorbed soft evidence using the Hugin *DistributeEvidence* algorithm, which is described in [Jen95, Sect. 4.4.1]. This algorithm has three important properties: (1) it updates the probability tables of the other cliques while it maintains the invariant that the product of the clique tables divided by the separator tables is equal to the joint probability table for all variables in the Bayesian network; (2) it insures local consistency and (Theorem 4.5 in [Jen95]) global consistency; (3) it does not disturb hard findings, because it does not introduce new zeroes.

Finally, observe that the table for the clique that contains all variables for which we have soft findings is unchanged by a DistributeEvidence call that starts at that clique. Therefore, the result of propagation is to obtain a globally consistent distributed representation of the posterior in which all findings, hard and soft, hold.

We remark that the big clique algorithm could be simplified by removing the DistributeEvidence part from the second step. In other words, it is sufficient to carry out one CollectEvidence operation to the special clique (using only hard evidence) and one DistributeEvidence from the special clique (after absorbing soft evidence in the special clique). From this point on, we redefine the big clique algorithm to be this simplified version. Figures 17 and 18 illustrate the CollectEvidence and DistributeEvidence operations. Figure 19 illustrates the operation of the whole big clique algorithm on the lawn example. In this special case, the junction tree is the same as that constructed by the junction tree algorithm (cf. 16), but the order of operations is different. In particular, note how the absorption of soft evidence is delayed. The properties of the big clique algorithm are studied furter in Section 5.

	A	
B	y	n
y	.664	.036
n	.102	.198

Table 5: Table for $P(A, B, e)$ after normalization.

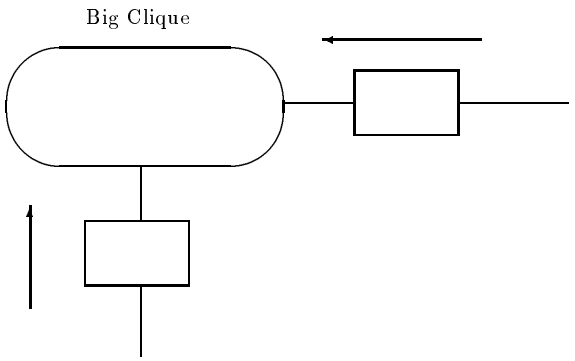

Figure 17: The big clique calls CollectEvidence.

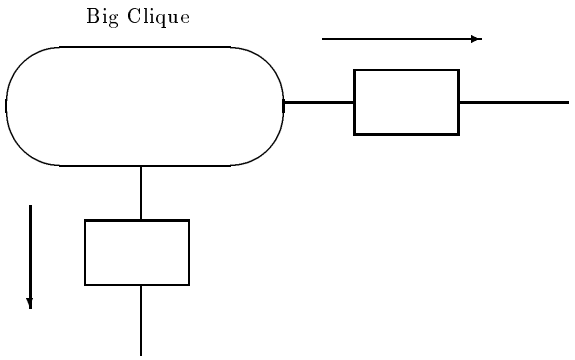

Figure 18: The big clique calls DistributeEvidence.

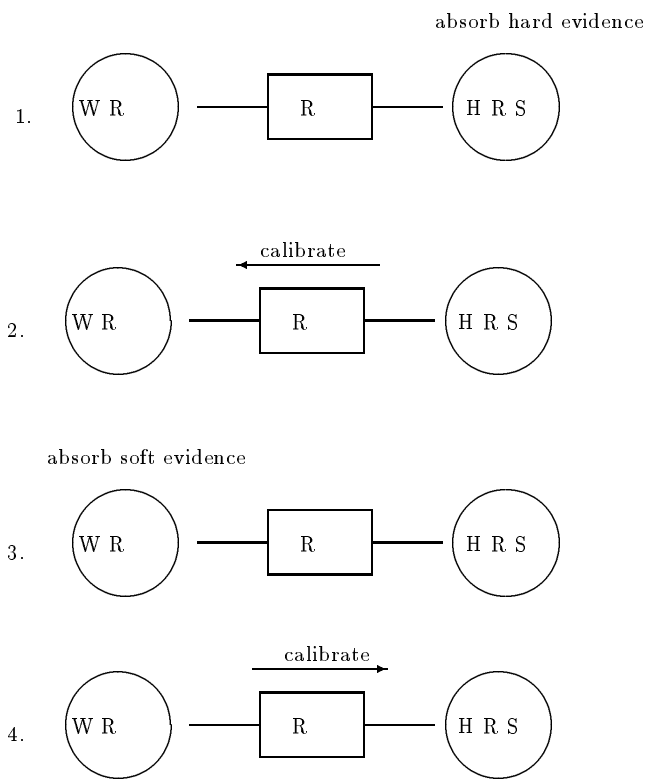

Figure 19: Operation of the big clique algorithm on the lawn example.

5 Properties of Soft Evidential Update Method

5.1 I-projections

Soft evidential updating can be formalized as a constrained optimization task. The goal is to find a probability distribution such that

(a) it satisfies all the constraints introduced by different types of soft evidence,

(b) it optimizes a chosen criterion among all distributions satisfying these constraints.

Cheeseman [Che83] proposed a method (based on Lagrange multipliers) that finds a joint probability distribution maximizing Shannon entropy over all distributions satisfying given constraints. However, the maximum entropy principle is not suitable for soft evidential update, since it does not take into account any information stored in the joint probability distribution that is going to be updated. We prefer a joint probability distribution that satisfies the constraints introduced by soft evidence and that is "close" to the joint probability distribution before updating.

I-divergence (defined by formula 4 in Section 4.2) may be used as a suitable distance criterion. (Also cf. [PV92].) Several arguments against application of this criterion to evidence updating were given in [Pea90, GH97]. These arguments are not against I-divergence *per se*, and we agree with their general tone: misuse of the I-divergence criterion leads to the dreaded "something for nothing" phenomenon, where inappropriate application of general formulae is substituted for the careful modeling of the domain of interest and, in particular, of its dependence structure. Moreover, in Section 5.4 we will show that some of Pearl's conclusions published in [Pea90] are erroneous and that his argument against the usage of I-divergence for conditional evidence updating vanishes if the optimization is performed with respect to the second argument of I-divergence. In this context, the notion of I-projections turns out to be useful.

The I_1-*projection* of a probability distribution P on a set of probability distributions $\mathcal{P}$ is a unique probability distribution that minimizes the I-divergence $I(R \parallel P)$ among all probability distributions $R \in \mathcal{P}$. The I_2-*projection* of P is a unique probability distribution minimizing $I(P \parallel R)$ among all probability distribution $R \in \mathcal{P}$.

5.2 Correspondence of Soft Evidential Update to IPFP

In the case of updating by a single marginal distribution, it does not matter whether the optimization is performed with respect to the first argument (I_1-projection) or the second argument (I_2-projection). There exists a single explicit formula for both I-projections in the case of updating based on single marginal probability distribution [KK66]: Formula 1 in Section 4.2.

In the case of several marginal probability distributions, the proposed soft evidential update method corresponds to the iterative proportional fitting procedure (IPFP) of Deming and Stephan [DS40]. Note that in Section 3 we have shown how updating based on either a probability assignment to an arbitrary logical function, or a probability assignment to a probabilistic function can be transformed to updating based on a marginal probability distribution. Therefore the correspondence of soft evidential update to IPFP holds in these cases

as well. Next, we will exploit some properties of IPFP to show that the computation of the resulting probability distribution can be decomposed.

5.2.1 Decomposability of Computation

The computation of a distribution that satisfies requirements (a) and (b) listed above can be decomposed as it is proposed for the big clique algorithm in Section 4.4. Recall that all variables $O_i, i = 1, 2, \ldots, n$ for which soft evidence is given are included in one clique C_1 (first step of big clique algorithm). The absorption of all soft evidence is performed locally in C_1 and then the absorbed evidence is distributed. The key that enables us to decompose the computation is the following lemma.

Lemma 1 *Let $P(V)$ be a given probability distribution and let C and B be two sets such that $C \cup B = V$ and that for the set of all observational variables $O_i, i = 1, 2, \ldots, n$, it holds that $\{O_1, O_2, \ldots, O_n\} \subseteq C$. Then $Q^\star$, the I_1-projection of probability distribution P on the set of all distributions having $Q(O_i), i = 1, 2, \ldots, n$ as their marginals can be computed as*

$$Q^\star(V) \quad = \quad Q^\star_C(C) \cdot P(B \setminus C \mid C)$$

where $Q^\star_C$ denotes the I_1-projection of marginal $P(C)$ of the original distribution P on the set of all distributions defined on C and having $Q(O_i), i = 1, 2, \ldots, n$ as their marginals.

For the proof see Appendix A.

If the set C corresponds to the big clique C_1 used in the method shown in Section 4.4.1, then $P(B \setminus C_1 \mid C_1) = P((B \setminus C_1 \mid C_1 \cap B)$ and the assertion of Lemma 1 can be written as $Q^\star(V) = Q^\star_{C_1}(C_1) \cdot P(B \setminus C_1 \mid C_1 \cap B)$, from which the following decomposition obtains:

$$Q^\star(V) = Q^\star_{C_1}(C_1) \cdot \frac{P((B \setminus C_1) \cup (C_1 \cap B))}{P(C_1 \cap B)} = Q^\star_{C_1}(C_1) \cdot \frac{P(B)}{P(C_1 \cap B)} \qquad (5)$$

By marginalizing $Q^\star(V)$ to $Q^\star(B)$, one obtains the calibration formula

$$Q^\star(B) \quad = \quad P(B) \cdot \frac{Q^\star_{C_1}(C_1 \cap B)}{P(C_1 \cap B)}$$

Therefore, we have obtained exactly the computation that is performed when the routine *DistributeEvidence* from C_1 is called (Section 4.4.2), thereby proving that the scheme of big clique algorithm proposed in Section 4.4 computes a distribution that both satisfies all the constraints introduced by soft evidence on $Q(O_i), i = 1, 2, \ldots, n$ (a) and minimizes I-divergence with respect to the original probability distribution P, among all probability distribution satisfying the given constraints (b).

5.2.2 Independence Causes Convergence within One Cycle

Next we will show that if all pairs of distinct observation variables are independent in the original distribution then they are independent in the updated distribution as well, and the soft evidential update method requires only one cycle to converge, i.e. only n steps are sufficient.

If all n observation variables are pairwise independent we write for $i, j = 1, \ldots, n,\ i \neq j : O_i \perp\!\!\!\perp O_j$. This independence in the original distribution can be equivalently written as: $P(O_1, \ldots O_n) = \prod_{i=1}^{n} P(O_i)$. Recall that using the big clique algorithm the evidence $Q(O_i), i = 1, \ldots n$ is absorbed so that we get a probability $Q^{\star}_{C_1}$, and it is sufficient to iterate on C_1 only (formula 5). We can use Lemma 1 with $C = \{O_1, \ldots, O_n\}$, perform the computations on C, and then extend the result to C_1 as given by formula 6:

$$Q^{\star}_{C_1}(C_1) = P(C_1 \setminus C \mid C) \cdot Q^{\star}_{C}(C) = \frac{P(C_1)}{\prod_{i=1}^{n} P(O_i)} \cdot \lim_{i \to \infty} Q_{(i)}(O_1, \ldots, O_n) \quad (6)$$

During the absorption of soft evidence (step 3 of the big clique algorithm) for every $i = 1, \ldots, n$ distribution $Q_{(i-1)}(O_1, \ldots, O_n)$ is updated by the corresponding soft evidence $Q(O_i)$. The starting distribution is $Q_{(0)}(O_1, \ldots, O_n) = P(O_1, \ldots, O_n) = \prod_{i=1}^{n} P(O_i)$. For $i = 1$ we can write

$$\begin{aligned} Q_{(1)}(O_1, \ldots, O_n) &= \frac{Q_{(0)}(O_1, \ldots, O_n) \cdot Q(O_1)}{Q_{(0)}(O_1)} = \frac{\prod_{j=1}^{n} P(O_j) \cdot Q(O_1)}{P(O_1)} \\ &= Q(O_1) \cdot \prod_{j=2}^{n} P(O_j) \end{aligned}$$

For $i = 2, \ldots, n$ we can similarly write

$$\begin{aligned} Q_{(i)}(O_1, \ldots, O_n) &= \frac{Q_{(i-1)}(O_1, \ldots, O_n) \cdot Q(O_i)}{Q_{(0)}(O_i)} = \\ &= \frac{\prod_{j=1}^{i-1} Q(O_j) \cdot \prod_{k=i}^{n} P(O_k) \cdot Q(O_i)}{P(O_i)} = \prod_{j=1}^{i} Q(O_j) \cdot \prod_{k=i+1}^{n} P(O_k) \end{aligned}$$

After performing n steps, we obtain the distribution

$$Q_{(n)}(O_1, \ldots, O_n) = \prod_{j=1}^{n} Q(O_j) \quad (7)$$

which has the property that $Q_{(n)}(O_j) = Q(O_j),\ j = 1, \ldots, n$, and therefore the procedure stops after n steps. It is obvious that $i, j = 1, \ldots, n,\ i \neq j : O_i \perp\!\!\!\perp O_j$ holds for $Q_{(n)}(O_1, \ldots, O_n)$ and thus consequently for $Q^{\star}_{C_1}(C_1)$ and $Q^{\star}(V)$ as well. If we know it in advance or have verified that all n observation variables are pairwise independent in the original probability distribution P, then we need not iterate at all since $Q^{\star}$ can be computed directly using formulae 5, 6, and 7.

The reader may now wonder whether it would be appropriate to ensure that the observed variables become independent after update, even when they are not independent in the original distribution. An immediate advantage of this position is that we could always update the big clique C_1 upon receipt of soft evidence on $\{O_1, \ldots, O_n\}$ in a single step by performing the following calibration operation, which is the operation done in the case of hard evidence:

$$Q(C_1) = P(C_1) \cdot \frac{Q(O_1) \cdot Q(O_2) \cdot \ldots \cdot Q(O_n)}{P(O_1, O_2, \ldots, O_n)}$$

In the hard evidence case, independence of the observed variables is necessary, because when $Q(O_1), Q(O_2), \ldots, Q(O_n)$ are hard findings, the joint distribution $Q(O_1, \ldots, O_n)$ necessarily is

$$Q(O_1) \cdot Q(O_2) \cdot \ldots \cdot Q(O_n),$$

since no other joint distribution has the required marginals. When the evidence consists of soft findings, however, the joint marginal distribution $Q(O_1, \ldots, O_n)$ is not uniquely specified by the observations. The following example shows that forcing independence of the observation variables in the subscribing agent may lead to incorrect results.

Imagine that we would like to have an agent that models the age and education levels of U.S. residents. We are only interested in two variables: age (X_1, with r possible age groups) and education (X_2, with s possible levels). Each person belongs to one of $r \times s$ groups. Our agent carried out a small survey and therefore constructed a joint probability distribution $P(X_1, X_2)$ that models the domain. Our agent now communicates with two other agents that provide it with accurate U.S. Census Bureau information concerning the age groups ($Q(X_1)$) and the education levels ($Q(X_2)$). By treating the two marginals as independent in the receiving agent, we would compute $Q_1(X_1, X_2) = Q(X_1) \cdot Q(X_2)$ and replace $P(X_1, X_2)$ by $Q_1(X_1, X_2)$. This is silly, of course, because the agent would lose all information from the small survey it performed. The agent should instead try to find another joint probability distribution $Q_2(X_1, X_2)$ such that

- The marginals from the U.S. Census Bureau agent are respected.
- $Q_2(X_1, X_2)$ is the distribution that is closest to the distribution $P(X_1, X_2)$. As a measure of distance among distribution, Kullback-Leibler divergence could be used.

As early as 1940, W.E. Deming and F.F. Stephen [DS40] proposed IPFP for solving such a problem. We propose that our agent does the same. One may argue that the task of this example is rather a *model learning* or *model revision* task than an *evidential update* task. In the context of AEBNs, it is more appropriate to consider this an evidential update task rather than a model revision task, because it may be necessary to carry it out as a routine activity in interagent communication. For example, consider a multiagent system in which there is an agent that is responsible for determining the nature of a coin. Another agent uses information about the coin for decision making. In the receiving agent, the initial marginal belief may be that the coin is ($P(Coin = head) = 1/2$), but evidence, which in this case is the published value of the coin, may come as $P(Coin = head) = 1/3$. The next time the receiving agent takes a decision, the evidence may have changed to $P(Coin = head) = 2/3$, and so on.

In AEBN systems, soft observations are messages from publishing agents and may be dependent when the AEBN graph is not a tree. The correct modeling of dependence in the receiving agent requires knowledge of the AEBN agent graph. A full treatment of this aspect is beyond the scope of this paper, but we mention Bloemeke's work on this topic [Blo98].

5.3 Soft Evidential Update and CIPFP

In the case of updating by a single conditional probability distribution, it does matter whether the optimization is performed with respect to the first or the second argument, since we get two different distributions. This fact was shown by Cramer in [Cra00]. We have already presented the explicit formulae for I_1-projection and I_2-projection in Section 4.2 (formulae 2 and 3).

In the case of several conditional or unconditional probability distributions the soft evidential update method corresponds to the conditional iterative proportional fitting procedure (CIPFP) of Bock [Boc89]. There are two versions of CIPFP depending on whether each step is an I_1-projection (CIPFP-1) or an I_2-projection (CIPFP-2). See [Cra00] for details. The properties of distributions $Q_1^\star$ and $Q_2^\star$, the limit distributions of CIPFP-1 and CIPFP-2, respectively, are summarized in the Table 6. These properties were proved by Cramer in [Cra00]. Note that there is an asymmetry, since $Q_2^\star$ generally minimizes neither the I_1-projection nor the I_2-projection, despite the fact that in every step the I_2-projection is computed. However, the important fact is that if there exist a distribution satisfying all required constraints then both versions of CIPFP converge to such a distribution.

$Q_i^\star$	minimizes *I_1-projection*	minimizes *I_2-projection*	satisfies constraints given by *soft evidence*
$i = 1$	yes	no	yes
$i = 2$	no	no	yes

Table 6: Properties of CIPFP

5.4 Tom and Mary Go to a Party

This is the example that was analyzed by Pearl in [Pea90]. We will use it to show how soft evidential updating can be performed using the approach we propose. We will also show that Pearl's conclusions concerning I-divergence were erroneous.

Suppose that we have an initial belief that Tom goes to the party $P(T = yes) = p$ and another initial belief that Mary goes to the party $P(M = yes) = q$. Initially, we suppose that these two events are independent. Therefore the probabilities corresponding to the beliefs given above uniquely define an initial joint probability distribution $P(T, M) = P(T) \cdot P(M)$. (See Table 7.)

We are going to present the example in a slightly generalized way with respect to the example analyzed by Pearl in [Pea90]. We allow the evidence to be soft, i.e. we assign a belief $v \in \langle 0, 1 \rangle$ to the information that we get. We do not assume that the initial probabilities of Tom and Mary going to the party are equal. However, if $v = 0$ and $p = q$, then the example corresponds to the one presented in [Pea90].

Imagine the following two situations:

	$M = yes$	$M = no$
$T = yes$	$p \cdot q$	$p \cdot (1-q)$
$T = no$	$(1-p) \cdot q$	$(1-p) \cdot (1-q)$

Table 7: Initial probability distribution $P(T, M)$

(1) We have got to know that *Tom and Mary will not be present at the party together.* In this case it seems reasonable to understand the given information as a probability assignment to the logical function $T = yes \ \& \ M = yes$. Since we may not be absolutely sure whether the information is right, we assign a probability value $v \in \langle 0, 1 \rangle$ to the event that *Tom and Mary will be present at the party together*, i.e. $Q(T = yes, M = yes) = v$. If we were sure that that event can not happen, this probability would be zero.

(2) We obtain a new piece of information: *Tom will not go to the party if Mary does.* This time the information should be rather understood as a conditional probability distribution. A probability value $v \in \langle 0, 1 \rangle$ is assigned to the conditional event $T = yes$ given $M = yes$, i.e. $Q(T = yes \mid M = yes) = v$.

(add 1) In the first case we want to update the initial probability distribution $P(T, M)$ so that we get a new probability distribution $Q(T, M)$ for which it holds that $Q(T = no, M = yes) = v$. As we have already mentioned above, in the case of updating by a marginal distribution, I_1-projection and I_2-projection are equivalent. The distribution that minimizes both criteria is given in Table 8.

	$M = yes$	$M = no$
$T = yes$	v	$\frac{p \cdot (1-q) \cdot (1-v)}{1-p \cdot q}$
$T = no$	$\frac{q \cdot (1-p) \cdot (1-v)}{1-p \cdot q}$	$\frac{(1-p) \cdot (1-q) \cdot (1-v)}{1-p \cdot q}$

Table 8: Updated probability distribution $Q(T, M)$ - case (1)

Since the observation is a logical function of the states of the Tom and Mary variables, using the soft evidential update approach, we first define the observation variable O taking two values o and $\neg o$, where

$$
\begin{aligned}
Q(O = o) \quad &= \quad Q(T = yes \ \& \ M = yes) \ = \ v \\
Q(O = \neg o) \ &= \\
&= \quad Q(T = yes \ \& \ M = no \ \vee \ T = no \ \& \ M = yes \ \vee \ T = no \ \& \ M = no) \\
&= \quad 1 - v
\end{aligned}
$$

Then we perform steps (2) and (3) of the soft evidential update method (Sec-

tion 4.2). They correspond to one step of IPFP, i.e.

$$Q(T,M) \quad = \quad \frac{P(T,M)}{P(O)} \cdot Q(O) \tag{8}$$

From Table 7 we may compute

$$\begin{aligned} P(O=o) &= p \cdot q \\ P(O=\neg o) &= p \cdot (1-q) + (1-p) \cdot q + (1-p) \cdot (1-q) = 1 - p \cdot q \end{aligned}$$

Thus using formula 8 we obtain the distribution that is equivalent to the I-projections already given in Table 8.

We repeat that the special case in which $v = 0$ and $p = q$ gives what Pearl [Pea90] calls the "Bayes conditionalization" solution. Conditioning in this case is done on the event "$M = yes$ & $T = yes$ is impossible", or $\neg(M = yes \ \& \ T = yes)$. Following the approach of Section 3, we chose to represent this event by introducing the observation variable O.

(add 2) In the second case we want to update the initial probability distribution $P(T,M)$ so that we get a new probability distribution $Q(T,M)$. In this case we require $Q(T = no \mid M = yes) = v$. Since we have no new information about the probability of Tom going to the party if Mary does not, the conditional probability $Q(T \mid M = no)$ should not change, i.e. $Q(T \mid M = no) = P(T \mid M = no)$. Therefore the conditional probability distribution $Q(T \mid M)$ is required to satisfy:

$$\begin{aligned} Q(T = yes \mid M = yes) &= v \\ Q(T = no \mid M = yes) &= 1 - v \\ Q(T = yes \mid M = no) &= \frac{p \cdot (1-q)}{(1-q)} \\ Q(T = no \mid M = no) &= \frac{(1-p) \cdot (1-q)}{(1-q)} \end{aligned}$$

Recall that in the case of updating by conditional probability distribution the I_1-projection and I_2-projection may differ. See Tables 9 and 10, where the I_1-projection and I_2-projection of P for the soft evidence $Q(T \mid M)$ are displayed.

	$M = yes$	$M = no$
$T = yes$	$\frac{v \cdot q \cdot (1-p)^{1-v} \cdot p^v}{q \cdot (1-p)^{1-v} \cdot p^v + (1-q) \cdot (1-v)^{1-v} \cdot v^v}$	$\frac{p \cdot (1-q) \cdot (1-v)^v \cdot v^v}{q \cdot (1-p)^{1-v} \cdot p^v + (1-q) \cdot (1-v)^{1-v} \cdot v^v}$
$T = no$	$\frac{(1-p) \cdot (1-q) \cdot (1-v)^v \cdot v^v}{q \cdot (1-p)^{1-v} \cdot p^v + (1-q) \cdot (1-v)^{1-v} \cdot v^v}$	$\frac{(1-p) \cdot (1-q)}{q \cdot (1-p)^{1-v} \cdot p^v + (1-q) \cdot (1-v)^{1-v} \cdot v^v}$

Table 9: Updated probability distribution $Q(T,M)$ - I_1-projection

Using the soft evidential update approach, we first define the observation variable O that takes four values

$$o_1 \quad = \quad (T = yes \ \& \ M = yes),$$

	$M = yes$	$M = no$
$T = yes$	$v \cdot q$	$p \cdot (1 - q)$
$T = no$	$(1 - v) \cdot q$	$(1 - p) \cdot (1 - q)$

Table 10: Updated probability distribution $Q(T, M)$ - I_2-projection

$$\begin{aligned} o_2 &= (T = no \ \& \ M = yes), \\ o_3 &= (T = yes \ \& \ M = no), \\ o_4 &= (T = no \ \& \ M = no). \end{aligned}$$

The probability distribution $Q(O)$ is computed using either the formula 2 or 3 of Section 4.2. In the case of I_2-projections, formula 2 gives for every o corresponding to $\langle t, m \rangle$, $t \in \{yes, no\}$, and $m \in \{yes, no\}$

$$Q(O = o) = P(M = m) \cdot Q(T = t \mid M = m)$$

Then we perform steps (2) and (3) of the soft evidential update method (Section 4.2), which for this example mean that $P(T, M)$ is simply replaced by $Q(T = t, M = m)$ defined by $Q(O = o)$, where o corresponds to $\langle t, m \rangle$. Observe that in this case the procedure corresponds to one step of CIPFP-2. Using the described approach we get the result equivalent to the I_2-projection of P given in Table 10.

Let us compute the marginal probability of Mary going to the party, $Q(M)$, for the case of I_1-projection.

$$\begin{aligned} Q(M = yes) &= \frac{q \cdot (1 - p)^{1-v} \cdot p^v}{q \cdot (1 - p)^{1-v} \cdot p^v + (1 - q) \cdot (1 - v)^{1-v} \cdot v^v} \\ Q(M = no) &= \frac{(1 - q) \cdot (1 - v)^v \cdot v^v}{q \cdot (1 - p)^{1-v} \cdot p^v + (1 - q) \cdot (1 - v)^{1-v} \cdot v^v} \end{aligned}$$

Next we will show that using the *minimum I-divergence* approach for any $0 \leq p, q, v \leq 1$ the probability of Mary going to the party is not increasing. This is just the opposite to what Pearl claims in his paper [Pea90]. It is well-known that the I-divergence of two probability distributions is always non-negative.

$$\begin{aligned} 0 &\leq I(Q(T \mid M = yes) \parallel P(T)) \\ 0 &\leq v \cdot \log \frac{v}{p} + (1 - v) \cdot \log \frac{1 - v}{1 - p} \end{aligned}$$

By exponentiation we obtain

$$1 \leq \left(\frac{v}{p}\right)^v \cdot \left(\frac{1 - v}{1 - p}\right)^{1-v}$$

Since $0 \leq q \leq 1$, we can multiply the previous inequality by $(1 - q)$ and obtain

$$\begin{aligned} (1 - q) \cdot (1 - p)^{1-v} \cdot p^v &\leq (1 - q) \cdot (1 - v)^{1-v} \cdot v^v \\ (1 - p)^{1-v} \cdot p^v &\leq (1 - q) \cdot (1 - v)^{1-v} \cdot v^v + q \cdot (1 - p)^{1-v} \cdot p^v \\ 1 &\geq \frac{(1 - p)^{1-v} \cdot p^v}{(1 - q) \cdot (1 - v)^{1-v} \cdot v^v + q \cdot (1 - p)^{1-v} \cdot p^v} \end{aligned}$$

By multiplication by q we obtain the following inequality

$$\begin{aligned} q &\geq \frac{q \cdot (1-p)^{1-v} \cdot p^v}{(1-q) \cdot (1-v)^{1-v} \cdot v^v + q \cdot (1-p)^{1-v} \cdot p^v} \\ P(M = yes) &\geq Q(M = yes) \end{aligned}$$

which proves that if the updated distribution Q is a I_1-projection then the value of $Q(M = yes)$ is less than or equal to the corresponding value before updating $P(M = yes)$.

For the distribution that is the result of the approach we propose (corresponding to I_2-projection) we get

$$\begin{aligned} Q(M = yes) &= v \cdot q + (1-v) \cdot q \;=\; q \\ Q(M = no) &= p \cdot (1-q) + (1-p) \cdot (1-q) = \; 1-q \end{aligned}$$

Observe that the marginal probability of Mary going to the party does not change, which corresponds to what seems intuitively correct in this example.

6 Conclusions and Future Work

Bayesian networks and related graphical probabilistic models (e.g., chain graphs, influence diagrams and decision networks) have established themselves as a useful tool for constructing intelligent systems. The application of graphical probabilistic models in multiagent systems clearly uncovers the limitations of simple conditioning as an update mechanism. In particular, it becomes necessary to allow update of the beliefs of an agent upon receipt of the beliefs of another agent (what we termed *soft evidence* in this paper), rather than on traditional, *hard* evidence.

We demonstrated that soft evidential update is possible in Bayesian networks without abandoning the basic commitments of graphical probabilistic modeling. In particular, we presented a principled way to extend Bayesian network structures with special *observation variables* in a way that, upon update from hard and soft evidence, the extended network structure correctly encodes the qualitative independence structure of the domain. Moreover, observation variables allow the correct processing of more general types of soft evidence than a collection of soft findings (single variable marginals).

We demonstrated that it is possible to extend the justly celebrated junction tree algorithm to handle soft evidence. The two extensions presented in this paper are based on a principled application of Jeffrey's rule and the Iterative Proportional Fitting Procedure (IPFP). IPFP was presented at one time as the "mechanical solution" [DZ82] to soft evidential updating. It was noted however that arbitrary and uncritical use of Jeffrey's rule and IPFP may lead to errors [GH97, Pea88, Pea90], and some authors have therefore expressed doubts about the possibility of a general method for soft evidential update. In this case, we believe that *in medio veritas* (the truth is in the middle) and hope that our paper will convince skeptics that soft evidential update in Bayesian networks is possible. Soft evidence should not be feared as a source of paradoxes and unextricable quandaries.

In order to make an impact in applications, the two extensions of the junction tree algorithm for soft evidential update described in Sections 4.3 and 4.4 should

be implemented and their performance should be empirically compared. We expect further work on Agent Encapsulated Bayesian Networks that will clarify the mechanism itself, exploiting Bloemeke's groundbreaking work [Blo98], will implement the necessary infrastructure for communication, and will extend the basic mechanism to decision-making and (limited) bi-directional communication.

We also expect work on the implementation and refinement of Conditional IPFP (CIPFP). The intriguing properties of CIPFP, some of which are briefly described in Section 5, warrant more theoretical work. Another line of theoretical investigation is the issue of sequential soft evidence update. While updating with hard findings (or virtual findings) may be done sequentially without changing the outcome, soft evidential update must be done concurrently. (Cf. [DZ82, BGHK97].) This is not a major stumbling block for applications, but it may have interesting theoretical ramifications.

The work presented in this paper has a methodological and practical bent. Accordingly, we used a rather plain style, with many illustrations of the concepts introduced, to try to reach a wide audience. The unified treatment of modeling for soft evidential update using observation variables should be especially useful for researchers and practitioners in Bayesian networks and open new avenues of application for graphical probabilistic models.

Acknowledgments

M.V.'s work has been supported in part by the U.S. Department of Defense through the DARPA Autonomous Negotiating Teams (ANTS) project (AO Number H359 and Contract Number F30602-99-2-0513). M.V. and Y.-G.K.'s work have been supported in part by the U.S. Office of Naval Research through project DAG (Grant Number N00014-97-1-0806). M.V. thanks in a special way Finn V. Jensen and his Decision Support Systems Group in the department of Computer Science at the University of Aalborg for providing the quiet and productive research environment that allowed the research in this paper to be completed during a sabbatical visit in March 2000. J.V. is grateful to his Ph.D. adviser Radim Jiroušek from the Laboratory of Intelligent Systems at the University of Economics in Prague for comments on several issues related to the paper. M.V. and J.V. thank Finn V. Jensen for comments on earlier versions of this paper.

References

[ACS96] Barry C. Arnold, Enrique Castillo, and Jose Maria Sarabia. Specification of distributions by combinations of marginals and conditional distributions. *Statistics and Probability Letters*, 26:153–157, 1996.

[AOJJ89] Stig K. Andersen, Kristian G. Olsen, Finn V. Jensen, and Frank Jensen. Hugin–a shell for building Bayesian belief universes for expert systems. In *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence*, pages 1080–1085, Detroit, 1989.

[BGHK97] Fahiem Bacchus, Adam J. Grove, Joseph Y. Halpern, and Daphne Koller. Probability update: Conditioning vs. cross-entropy. In *Proceedings of the Thirteenth Annual Conference on Uncertainty in Artificial Intelligence (UAI-97)*, pages 208–214, Providence, RI, August 1997.

[Blo98] Mark Bloemeke. *Agent Encapsulated Bayesian Networks.* PhD thesis, Department of Computer Science, University of South Carolina, 1998.

[Boc89] H. H. Bock. A conditional iterative proportional fitting (CIPFP) algorithm with applications in the statistical analysis of discrete spatial data. *Bull. ISI, Contributed papers of 47th Session in Paris*, pages 141–142, 1989.

[BS99] Mark Bloemeke and Wayne Smith. Agent encapsulated Bayesian networks—a solution to the rumor problem, Technical Report TR9908. Technical report, Department of Computer Science, University of South Carolina, 1999.

[BV] Mark Bloemeke and Marco Valtorta. Agent encapsulated Bayesian networks—a framework for probabilistic agent systems. in preparation.

[Che83] Peter Cheeseman. A method of computing generalized Bayesian probability values for expert systems. In *Proceedings of the Eighth International Joint Conference on Artificial Intelligence (IJCAI-83)*, pages 198–202, Karlsruhe, 1983.

[Cra00] Erhard Cramer. Probability measures with given marginals and conditionals: I-projections and conditional iterative proportional fitting. *Statistics and Decisions* (accepted for publication), 2000. Institut for Wirtschftmathematik und Statistik, Aachen University of Technology.

[Csi75] I. Csiszár. *I*-divergence geometry of probability distributions and minimization problems. *Ann. Prob.*, 3(1):146–158, 1975.

[DS40] W.E. Deming and F.F. Stephan. On a least square adjustment of a sampled frequency table when the expected marginal totals are known. *Annals of Mathematical Statistics*, 11:427, 1940.

[DZ82] Persi Diaconis and Sandi Zabell. Updating subjective probability. *Journal of the American Statistical Association*, 77:822–830, 1982.

[GH97] Adam J. Grove and Joseph Y. Halpern. Probability update: Conditioning vs. cross-entropy. In *Proceedings of the Thirteenth Annual Conference on Uncertainty in Artificial Intelligence (UAI-97)*, pages 208–214, Providence, RI, August 1997.

[GS93] Andrew Gelman and T.P. Speed. Characterizing a joint probability distribution by conditionals. *Journal of the Royal Statistical Society Series B (Methodological)*, 55:185–188, 1993.

[Hab74] S. Haberman. *The Analysis of Frequency Data.* The University of Chicago Press, Chicago and London, 1974.

[Hec86] David Heckerman. Probabilistic interpretation for MYCIN's certainty factors. In L. N. Kanal and J.F. Lemmer, editors, *Uncertainty in Artificial Intelligence*, pages 167–196. Kluwer Science Publishers, 1986.

[HH86] Eric Horvitz and David Heckerman. The inconsistent use of measures of certainty in artificial intelligence research. In L. N. Kanal and J.F. Lemmer, editors, *Uncertainty in Artificial Intelligence*, pages 137–151. Kluwer Science Publishers, 1986.

[Jen95] Finn V. Jensen. *An Introduction to Bayesian Networks.* Springer-Verlag, New York, NY, 1995.

[Jir91] R. Jiroušek. Solution of the marginal problem and decomposable distributions. *Kybernetika*, 27(5):403–412, 1991.

[KK66] S. Kullback and M. A. Khairat. A note on minimum discrimination information. *Ann. Math. Statist.*, 37:279–280, 1966.

[Kyb87] Henry E Kyburg, Jr. Bayesian and non-Bayesian evidential updating. *Artificial Intelligence*, 31:271–293, 1987.

[LJ97] S.L. Lauritzen and F.V. Jensen. Local computation with valuations from a commutative semigroup. In *Annals of Mathematics and Artificial Intelligence*, volume 21, pages 51–69. J.C. Baltzer A.G., 1997.

[LS88] Steffen L. Lauritzen and David Spiegelhalter. Local computations with probabilities on graphical structures and their application to expert systems (with discussion). *Journal of the Royal Statistical Society, Series B*, 50:157–224, 1988.

[Nea90] Richard E. Neapolitan. *Probabilistic Reasoning in Expert Systems: Theory and Algorithms.* John Wiley and Sons, New York, NY, 1990.

[Pea88] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* Morgan Kaufman, San Mateo, CA, 1988.

[Pea90] Judea Pearl. Jeffrey's rule, passage of experience, and *neo*-Bayesianism. In H.E. et al. Kyburg, Jr., editor, *Knowledge Representation and Defeasible Reasoning*, pages 245–265. Kluwer Academic Publishers, 1990.

[PV92] J.B. Paris and A. Vencovska. A method for updating that justifies minimum cross entropy. *International Journal of Approximate Reasoning*, 7:1–18, 1992.

[SS90] Glenn R. Shafer and Prakash P. Shenoy. Probability propagation. *Annals of Mathematics and Artificial Intelligence*, 2:327–352, 1990.

[Vom99] Jiri Vomlel. *Methods of Probabilistic Knowledge Integration.* PhD thesis, Department of Cybernetics, Faculty of Electrical Engineering, Czech Technical University, December 1999.

[Xia96] Y. Xiang. A probabilistic framework for cooperative multi-agent distributed interpretation and optimization of communication. *Artificial Intelligence*, 86:295–342, 1996.

[XL00] Y. Xiang and V. Lesser. Justifying multiply sectioned Bayesian networks. In *Proceedings of the Fourth International Conference on Multi-Agent Systems (ICMAS-2000)*, Boston, MA, July 2000. To appear; preprint available at the first author's web site.

A Proof of Lemma 1

Lemma 1 *Let $P(V)$ be a given probability distribution and let C and B be two sets such that $C \cup B = V$ and that for the set of all observational variables $O_i, i = 1, 2, \ldots, n$, it holds that $\{O_1, O_2, \ldots, O_n\} \subseteq C$. Then $Q^\star$, the I_1-projection of probability distribution P on the set of all distributions having $Q(O_i), i = 1, 2, \ldots, n$ as their marginals can be computed as*

$$Q^\star(V) \quad = \quad Q^\star_C(C) \cdot P(B \setminus C \mid C)$$

where $Q^\star_C$ denotes the I_1-projection of marginal $P(C)$ of the original distribution P on the set of all distributions defined on C and having $Q(O_i), i = 1, 2, \ldots, n$ as their marginals.

Proof. Let $\mathcal{P}$ denote the set of all probability distributions defined on variables from V. The I_1-projection of a probability distribution $P(V)$ on a set of probability distributions $\mathcal{P}$, having $Q(O_i), i = 1, 2, \ldots, n$, as their marginals, was defined to be the unique probability distribution

$$\begin{aligned} Q^\star &= \arg\min_{R \in \mathcal{P}} I(R \parallel P) \\ &= \arg\min_{R \in \mathcal{P}} \sum_x R(x) \log \frac{R(x)}{P(x)} \end{aligned}$$

For a particular combination of values $x = x^V$, we will write $x^V = \langle x^C, x^{B \setminus C} \rangle$, where for any set $S \subseteq V$ the symbol x^S denote particular values of the random variables $(X_i)_{i \in S}$. Using the definition of conditional probability we can write

$$\begin{aligned} Q^\star &= \arg\min_{R \in \mathcal{P}} \sum_x R(x) \log \frac{R(x^C) \cdot R(x^{B \setminus C} \mid x^C)}{P(x^C) \cdot P(x^{B \setminus C} \mid x^C)} \\ &= \arg\min_{R \in \mathcal{P}} \sum_x R(x) \log \frac{R(x^C)}{P(x^C)} + R(x) \log \frac{R(x^{B \setminus C} \mid x^C)}{P(x^{B \setminus C} \mid x^C)} \\ &= \arg\min_{R \in \mathcal{P}} \left(\begin{array}{l} \sum_{x^C} \left(\log \frac{R(x^C)}{P(x^C)} \cdot \sum_{x^{B \setminus C}} R(\langle x^C, x^{B \setminus C} \rangle) \right) + \\ \sum_{x^C} \left(R(x^C) \cdot \sum_{x^{B \setminus C}} R(x^{B \setminus C} \mid x^C) \log \frac{R(x^{B \setminus C} \mid x^C)}{P(x^{B \setminus C} \mid x^C)} \right) \end{array} \right) \end{aligned} \tag{9}$$

Observe that $\sum_{x^{B\setminus C}} R(\langle x^C, x^{B\setminus C}\rangle) = R(x^C)$ (i.e. $R(C)$ is the marginal of $R(V)$ on C), and therefore

$$\begin{aligned}\sum_{x^C}\left(\log\frac{R(x^C)}{P(x^C)}\cdot\sum_{x^{B\setminus C}} R(\langle x^C, x^{B\setminus C}\rangle)\right) &= \sum_{x^C}\left(\log\frac{R(x^C)}{P(x^C)}\cdot R(x^C)\right)\\ &= I\Big(R(C)\parallel P(C)\Big) \qquad (10)\end{aligned}$$

Furthermore , for every x^C,

$$\begin{aligned}&\sum_{x^{B\setminus C}} R(x^{B\setminus C}\mid x^C)\log\frac{R(x^{B\setminus C}\mid x^C)}{P(x^{B\setminus C}\mid x^C)}\\ &\qquad = I\Big(R(B\setminus C\mid x^C)\parallel P(B\setminus C\mid x^C)\Big) \qquad (11)\end{aligned}$$

Using formulae 10 and 11 we can rewrite formula 9 as

$$Q^\star = \arg\min_{R\in\mathcal{P}}\left(\begin{array}{l} I\Big(R(C)\parallel P(C)\Big) + \\ \sum_{x^C} R(x^C)\cdot I\Big(R(B\setminus C\mid x^C)\parallel P(B\setminus C\mid x^C)\Big)\end{array}\right)$$

Since the I-divergence of two probability distributions is always non-negative then all addends can be minimized independently. The first addend is minimized by the limit distribution $Q^\star_C$ of IPFP performed on the set C, i.e.

$$R(C) = Q^\star_C(C) \qquad (12)$$

The other addends are minimized (equal zero) if for every value of x^C it holds that

$$R(B\setminus C\mid x^C) = P(B\setminus C\mid x^C) \qquad (13)$$

The formulae 12 and 13 imply that the I_1-projection $Q^\star$ is equal to $R\in\mathcal{P}$ such that

$$R(V) = Q^\star_C(C)\cdot P(B\setminus C\mid C)$$

which proves the statement of the lemma. □