The Semantics of Intermediate CPTs in Variable Elimination

Cory J. Butz
University of Regina, Canada
butz@cs.uregina.ca

Wen Yan
University of Regina, Canada
yanwe111@cs.uregina.ca

Abstract

Variable elimination (VE), a central component of Bayesian network inference, starts and ends with clear structure and semantics, yet all intermediate distributions, whether normalized or unnormalized, are denoted as potentials. In this paper, a condition is given stating when intermediate distributions are defined with respect to the joint distribution. Theoretical and practical advantages of these new semantics are given.

1 Introduction

A *Bayesian network* (BN) (Pearl, 1988; Cowell et al., 1999; Jensen and Nielsen, 2007; Kjaerulff and Madsen, 2008) consists of a *directed acyclic graph* (DAG) and a corresponding set of *conditional probability tables* (CPTs). The independencies encoded in a DAG on variable set U indicate that the product of CPTs is a *joint probability distribution* $p(U)$. BN reasoning centres around eliminating variables. *Variable elimination* (VE) (Zhang and Poole, 1996), an inference algorithm for answering a query $p(X|E = e)$, repeatedly calls the *sum-out* (SO) algorithm to remove variables. SO removes a variable v as a two-step process. First, the product of all distributions involving v is taken. Second, v is marginalized out from the obtained product.

Koller and Friedman (2009) state that it is interesting to consider the semantics of the distribution output by SO when evidence is not considered. They mention that SO outputs a CPT $\phi(X|Y)$, but not necessarily with respect to $p(U)$. Butz et al. (2010) have shown a stronger result, namely, that every multiplication and every addition operation during VE's execution yields a CPT, albeit perhaps not with respect to $p(U)$.

In this paper, we address the semantics of VE's intermediate CPTs by providing a condition stipulating when they are defined with respect to $p(U)$. Roughly speaking, $\phi(X|Y)$ is $p(X|Y)$ provided there exists a topological ordering of all variables in the BN in which those variables used to build $\phi(X|Y)$ appear consecutively. In such cases, we say an intermediate CPT $\phi(X|Y)$ has a "p-label" and denote $\phi(X|Y)$ as $p(X|Y)$. It is important to observe that $\phi(X|Y)$ can be normalized or unnormalized, as well as involve evidence variables or not. It is noted that there are two kinds of paths that violate our condition, which, respectively, have temporary and permanent influences on SO. This work helps reveal structure and semantics in probabilistic reasoning with BNs, a worthy goal according to Pearl (1988) and Shafer (1996). We will also mention a practical advantage of this new semantic knowledge using the latest optimization techniques that are being applied in join tree propagation.

This paper is organized as follows. Section 2 contains background knowledge. The CPT structure of SO is discussed in Section 3. In Section 4, we establish semantics of VE's CPT structure. We extend the semantics to involve evidence in Section 5. Section 6 contains theoretical and practical advantages. Conclusions are given in Section 7.

2 Background Knowledge

The following discussion draws mainly from Shafer (1996) and Olmsted (1983). Let $U = \{v_1, v_2, \ldots, v_n\}$ be a finite set of variables. Each v_i has a finite domain, denoted $dom(v_i)$. For a subset $X = \{v_i, \ldots, v_j\}$ of U, $dom(X)$ denotes the Cartesian product of the domains of the individual variables in X. Each element $x \in dom(X)$ is called a *configuration* of X.

A *potential* on $dom(X)$ is a function ψ such that $\psi(x) \geq 0$ for each $x \in dom(X)$, and at least one $\psi(x)$ is positive. A *joint probability distribution* on $dom(U)$ is a potential p on $dom(U)$ that sums to 1. A potential that sums to 1 is *normalized*; otherwise, it is *unnormalized*. We may write a set $\{v_1, v_2, \ldots, v_k\}$ as $v_1 v_2 \cdots v_k$ and use XY to denote $X \cup Y$. A *conditional probability table* (CPT) for X given disjoint Y, denoted $\phi(X|Y)$, is a potential on XY, satisfying the following condition: for each configuration $y \in dom(Y)$, $\sum_{x \in dom(X)} \phi(X = x \mid Y = y) = 1$. In writing $\phi(X|Y)$ with X and Y not disjoint, we always means $\phi(X|Y - X)$, and only configurations with non-zero probability are stored.

A discrete *Bayesian network* (BN) (Pearl, 1988) on $U = \{v_1, v_2, \ldots, v_n\}$ is a pair (B, C). B is a DAG with vertex set U. C is a set of CPTs $\{p(v_i|P_i) \mid i = 1, 2, \ldots, n\}$, where P_i denotes the parents (see below) of variable $v_i \in B$.

A *path* from v_1 to v_n is a sequence $v_1, v_2, \ldots, v_n$ with arcs (v_i, v_{i+1}), $i = 1, \ldots, n-1$ in B. With respect to a variable v_i, we define four sets: (i) the ancestors of v_i, denoted $A(v_i)$, are those variables having a path to v_i; (ii) the parents of v_i are those variables v_j such that arc (v_j, v_i) is in B; (iii) the descendants of v_i, denoted $D(v_i)$, are those variables to which v_i has a path; and, (iv) the children of v_i are those variables v_j such that arc (v_i, v_j) is in B. The ancestors of a set X of variables are defined as $A(X) = (\cup_{v_i \in X} A(v_i)) - X$. $D(X)$ is similarly defined. A *topological ordering* is an ordering $\prec$ of the variables in a BN B so that for every arc (v_i, v_j) in B, $v_i \prec v_j$. *Initial segments* of the ordering produce marginals of $p(U)$. A set W of variables in a DAG is an *initial segment* if the parents of each v_i in W are also in W.

Algorithm 1, called *sum-out* (SO), eliminates a single variable v from a set Φ of potentials, and returns the resulting set of potentials. The algorithm collect-relevant simply returns those potentials in Φ involving variable v.

Algorithm 1 sum-out(v,Φ)
begin
Ψ = collect-relevant(v,Φ)
ψ = the product of all potentials in Ψ
$\tau = \sum_v \psi$
Return $(\Phi - \Psi) \cup \{\tau\}$
end

Algorithm 2, called *variable elimination* (VE), computes $p(X \mid E = e)$ from a BN on U. VE calls SO to eliminate variables one by one. More specifically, in Algorithm 2, Φ is the set of CPTs in a BN, X is a list of query variables, E is a list of observed variables, e is the corresponding list of observed values, and σ is an elimination ordering for variables $U - XE$.

Algorithm 2 VE(Φ, X, E, e, σ)
begin
Set $E = e$ in all appropriate CPTs of Φ
While σ is not empty
  Remove the first variable v from σ
  Φ = sum-out(v, Φ)
$p(X, E = e)$ = the product of all $\phi \in \Phi$
$p(E = e) = \sum_X p(X, E = e)$
Return $p(X, E = e)/p(E = e)$
end

3 The CPT Structure of sum-out

Observe that VE starts and ends with clear structure and semantics, yet all intermediate distributions, whether normalized or unnormalized, are denoted as potentials. In their very comprehensive discussion, Koller and Friedman (2009) state that it is interesting to consider the semantics of the distribution constructed by summing out a variable from a BN not involving observed evidence. They point out that SO's marginalization step produces a CPT, but not necessarily with respect to the joint probability distribution $p(U)$.

Example 1. SO eliminates variable b from the BN in Figure 1 as:

$$\phi(c,e|a,d) \;=\; \sum_b p(b|a)\cdot p(c|b)\cdot p(e|b,d). \quad (1)$$

Thus, after marginalization, SO outputs a CPT.

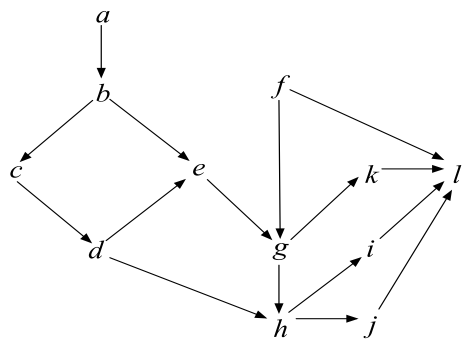

Figure 1: A Bayesian network.

Butz et al. (2010) have shown a stronger result, namely, every multiplication and every addition operation during VE's execution yields a CPT, albeit perhaps not with respect to $p(U)$.

Example 2. Each operation of (1) gives a CPT:

$$\begin{aligned} & \sum_b p(b|a)\cdot p(c|b)\cdot p(e|b,d) \\ =\; & \sum_b \phi(b,c|a)\cdot p(e|b,d) \quad (2) \\ =\; & \sum_b \phi(b,c,e|a,d) \quad (3) \\ =\; & \phi(c,e|a,d). \end{aligned}$$

Note that Shafer (1996) gives a condition under which each successive product of a sequence of CPTs always yields another CPT. In particular, for the case when multiplying just two CPTs $\phi(X_1|Y_1)$ and $\phi(X_2|Y_2)$, it is stated that

$$\phi(X_1X_2|Y_1Y_2) \;=\; \phi(X_1|Y_1)\cdot\phi(X_2|Y_2), \quad (4)$$

provided that X_2 is disjoint from X_1Y_1. The next example demonstrates, however, that (4) does not cover all possible cases encountered when applying SO on a BN.

Example 3. Consider eliminating variable c from the following BN:

$$p(a)\cdot p(b)\cdot p(c)\cdot p(d|a,b,c)\cdot p(e|c,d).$$

By Algorithm 1, we obtain

$$\sum_c p(c)\cdot p(d|a,b,c)\cdot p(e|c,d).$$

The optimal multiplication ordering, assuming binary variables and positive probabilities, is

$$\sum_c (p(c)\cdot p(e|c,d))\cdot p(d|a,b,c).$$

By (4), $p(c)\cdot p(e|c,d)$ is $\phi(c,e|d)$, giving

$$\sum_c \phi(c,e|d)\cdot p(d|a,b,c). \quad (5)$$

As (4) no longer applies, the product of the two CPTs in (5) must be denoted as a potential:

$$\sum_c \psi(a,b,c,d,e).$$

While (Butz et al., 2010) reveals the CPT structure of $\psi(a,b,c,d,e)$ as $\phi(c,d,e|a,b)$, the remaining unanswered question is semantic. Is $\phi(c,d,e|a,b)$ equal to $p(c,d,e|a,b)$?

4 Semantics Without Evidence

We first consider eliminating variables without observed evidence. It can easily be shown that SO's marginalization operation will always yield a CPT with the same kind of label as the CPT being marginalized. Thus, we focus on the semantics of multiplication.

Theorem 1. *Given a BN B on U and $X \subseteq U$. Then*

$$p(X|Y) \;=\; \prod_{v_i\in X} p(v_i|P_i),$$

if there is a topological ordering $\prec$ of B in which the variables in X appear consecutively, where $Y = (\cup_{v_i\in X} P_i) - X$.

Proof. Suppose there exists a topological ordering $\prec$ of the variables in B in which the variables in X appear consecutively. Let W be the set of all variables appearing in $\prec$ before any variable in X. By (Shafer, 1996), the variables in W and WX are both initial segments, meaning that

$$p(W) = \prod_{v_w\in W} p(v_w|P_w) \quad (6)$$

and

$$p(WX) = \prod_{v_w \in W} p(v_w|P_w) \cdot \prod_{v_x \in X} p(v_x|P_x). \quad (7)$$

By substitution of (6) into (7),

$$p(WX) \;=\; p(W) \cdot \prod_{v_x \in X} p(v_x|P_x).$$

By (Butz et al., 2010),

$$p(WX) \;=\; p(W) \cdot \phi(X|Y). \quad (8)$$

According to $\prec$, the variables Y must be contained in W, by (Shafer, 1996). Let $V = W - Y$, so $W = VY$. Then (8) can be rewritten as

$$p(VYX) \;=\; p(VY) \cdot \phi(X|Y).$$

Marginalizing away V yields

$$p(YX) \;=\; p(Y) \cdot \phi(X|Y).$$

By rearrangement, we obtain our desired result

$$p(X|Y) \;=\; \phi(X|Y).$$

□

Example 4. By Theorem 1, $\phi(b, c|a)$ in (2) is $p(b, c|a)$, since b and c can appear consecutively in a topological order $\prec$ of B in Figure 1. However, $\phi(b, c, e|a, d)$ in (3) is not guaranteed to be $p(b, c, e|a, d)$, since every topological order $\prec$ has d between c and e, i.e., $c \prec d \prec e$.

Our topological condition is sufficient but not necessary to ensure CPTs with p-labels.

Example 5. Consider eliminating b from the BN in Figure 1. Suppose the CPTs for $a, \ldots, e$ are defined such that their marginal has only one configuration with a non-zero probability, say, $p(a = 0, b = 0, c = 0, d = 0, e = 0)$ is 1. In this extreme case, it can be verified that

$$p(b, c, e|a, d) = p(b|a) \cdot p(c|b) \cdot p(e|b, d).$$

To ensure p-label CPTs for any BN instance B, we extend SO as *sum-out-as-p* (SOP), which calls *collect-topological* (CT) to collect any CPT needed to satisfy our topological ordering requirement in B.

Algorithm 3 sum-out-as-p(v,Φ)
begin
Ψ = collect-relevant(v, Φ)
Θ = collect-topological(Ψ, Φ)
$p(X|Y)$ is the product of all CPTs in Ψ and Θ
$p(X - v|Y) \;= \sum_v p(X|Y)$
Return $(\Phi - \Psi - \Theta) \cup \{p(X - v|Y)\}$
end

Algorithm 4 collect-topological(Ψ,Φ)
begin
X is the union of all X_i where $p(X_i|Y_i) \in \Psi$
Let Z be $A(X) \cap D(X)$
Let Ω be those $p(X_i|Y_i)$ in Φ with $X_i \cap Z \neq \emptyset$
Return Ω
end

Example 6. Consider how SOP eliminates h, b and f from Figure 1. For h, $\Psi = \{p(h|d, g)$, $p(i|h)$, $p(j|h)\}$. CT returns $\Omega = \emptyset$, as $X = hij$, $Z = A(hij) \cap D(hij) = \emptyset$. In SOP, $\Theta = \emptyset$, so

$$p(i, j|d, g) = \sum_h p(h|d, g) \cdot p(i|h) \cdot p(j|h).$$

For b, $\Psi = \{p(b|a),\ p(c|b),\ p(e|b, d)\}$ and, as $Z = A(bce) \cap D(bce) = d$, $\Theta = \{p(d|c)\}$. Thus,

$$p(c, d, e|a) = \sum_b p(b|a) \cdot p(c|b) \cdot p(d|c) \cdot p(e|b, d).$$

For f, $\Psi = \{p(f),\ p(g|e, f),\ p(l|f, k, i, j)\}$ and, since $Z = A(fgl) \cap D(fgl) = hijk$, we have $\Theta = \{p(i, j|d, g),\ p(k|g)\}$. Therefore,

$$\sum_f p(f)\ p(g|e, f)\ p(i, j|d, g)\ p(k|g)\ p(l|f, k, i, j)$$
$$= p(g, i, j, k, l\ |d, e).$$

Observe that every elimination in Example 6 yielded a p-label CPT. Also note that Z in the collect-topological algorithm can be quickly obtained from the transitive closure of a DAG, which can be found in $O(n^3)$ time (Cormen et al., 2009). It must be made clear that we are not advocating that SOP be considered as a new approach to inference. Instead, SOP can shed insight into the semantics of SO's intermediate CPTs. SO is ensured to yield a CPT with a p-label, if it collects the same CPTs as SOP does.

Example 7. Recall Example 6. When eliminating variable h, SO will yield a p-label, since SO and SOP collect the same CPTs. On the contrary, to eliminate variable b, SO can compute the following ϕ-label:

$$\phi(b, c, e|a, d) = p(b|a) \cdot p(c|b) \cdot p(e|b, d).$$

A p-label is not necessarily obtained here due to the fact that SOP also collects $\Theta = \{p(d|c)\}$.

More generally, every product taken in SO of two CPTs $\phi(X_1|Y_1)$ and $\phi(X_2|Y_2)$ will be $p(X_1X_2|Y_1Y_2)$, provided the topological requirement is met. Let us focus on eliminating a single variable from a BN. There are two kinds of paths warranting attention. The first only involves children of the variable being eliminated. When eliminating a variable v from a BN, the CPTs of v's children must be multiplied in an order consistent with some topological ordering of the DAG. The following example illustrates how this condition has a temporary influence within SO, meaning that intermediate CPTs can alternate between ϕ- and p-labels.

Example 8. Consider the elimination of variable b from the BN in Figure 2:

$$\begin{aligned}
& \sum_b p(b) \cdot p(e|b, d) \cdot p(d|a, b) \cdots && (9) \\
= & \sum_b \phi(b, e|d) \cdot p(d|a, b) \cdot p(g|b, f) \cdots && (10) \\
= & \sum_b p(b, d, e|a) \cdot p(g|b, f) \cdot p(f|b, c, e) \cdots && (11) \\
= & \sum_b \phi(b, d, e, g|a, f) \cdot p(f|b, c, e) \cdots && (12) \\
= & \sum_b p(b, d, e, f, g|a, c) \cdot p(i|b, g, h) \cdots && (13) \\
= & \sum_b \phi(b, d, e, f, g, i|a, c, h) \cdot p(j|b, i) && (14) \\
= & \sum_b \phi(b, d, e, f, g, i, j|a, c, h) && (15) \\
= & \ \phi(d, e, f, g, i, j|a, c, h). && (16)
\end{aligned}$$

Example 8 demonstrates how the intermediate CPTs can alternate between having and not having p-labels. A ϕ-label can be obtained when multiplying the CPTs for b and e in (10), since

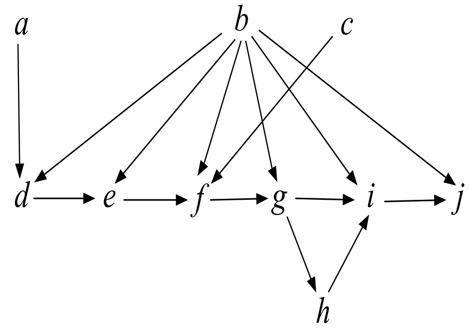

Figure 2: Illustrating the alternating pattern of intermediate CPTs with p-labels in Example 8.

there is a path from b to e going through d (so $b \prec d \prec e$). Since the CPT $p(d|a, b)$ for d has been collected by SO, a p-label can be subsequently re-obtained, as shown in (11). Similar remarks hold for multiplying this product with the CPT for g before that for f, as shown in (12) - (13).

The second kind of path, however, has a permanent influence on the semantics of SO's intermediate CPTs; it involves variables that are not children of the variable being eliminated. Recall Example 8 where variable b is being eliminated and consider (13) - (14). Once the CPT $p(i|b, g, h)$ is multiplied, all CPTs subsequently constructed by SO during the elimination of b can have a ϕ-label as in (14) - (16). The reason is that there is a path from b to i going through h (so $b \prec g \prec h \prec i$). However, the CPT $p(h|g)$ is not collected by SO as $p(h|g)$ does not involve b. Hence, the only way to ensure a subsequent p-label is to wait for $p(h|g)$ to be multiplied during a different call to SO, say to eliminate h.

Theorem 2. *Given a BN B on U, let $\phi(X|Y)$ be any CPT that VE computes by multiplication. Then $\phi(X|Y)$ is $p(X|Y)$, if there is a topological ordering $\prec$ of B in which the variables in DX appear consecutively, where D are those variables that were eliminated by SO in building $\phi(X|Y)$.*

The proof of Theorem 2 is similar to that of Theorem 1 and will be shown in a separate manuscript.

Example 9. Continuing from Example 8, consider the elimination of variable h.

$$\begin{aligned} & \sum_h p(h|g) \cdot \phi(d,e,f,g,i,j|a,c,h) \\ = & \sum_h p(d,e,f,g,h,i,j|a,c) \quad (17) \\ = & \; p(d,e,f,g,i,j|a,c). \end{aligned}$$

In Example 9, a p-label is obtained in (17) as there exists a topological ordering $\prec$ of the BN in Figure 2 where the variables in b,d,e,f,g,h,i,j appear consecutively.

5 Semantics With Evidence

Suppose we observe the values e of a set E of variables contained in U. Before a disjoint set D also contained in U is eliminated from the BN, those CPTs containing evidence variables are modified by multiplying them with evidence potentials. An *evidence potential*, denoted $1(E)$, assigns probability 1 to the single configuration e of E and probability 0 to all other configurations of E. If a BN CPT $p(v_i|P_i)$ contains at least one evidence variable, then 1_i denotes $1(E)$ restricted to those evidence variables appearing in $p(v_i|P_i)$. Hence, the product $p(v_i|P_i) \cdot 1_i$ keeps the configurations agreeing with $E = e$ while deleting the rest. By $1(\emptyset)$, we denote 1.

Intermediate potentials, constructed during inference by VE, can always have their conditional probabilities (Theorem 3) and semantics (Corollary 1) identified.

Theorem 3. *Given a BN B and evidence $E = e$, every VE distribution constructed by multiplication can be expressed as the product of a CPT and an evidence potential.*

Proof. Let the constructed distribution be $\phi_1 \cdot \phi_2$. We can always equivalently rewrite ϕ_1 as the marginalization of the product of the CPTs and evidence potentials used to build ϕ_1:

$$\phi_1 = \sum_{D_1} \prod_{i=1}^{j} (p(v_i|P_i) \cdot 1_i) \cdot \prod_{i=j+1}^{k} p(v_i|P_i),$$

where D_1 is the set of variables marginalized away by SO from the product of k CPTs in B, and where j evidence potentials were used. Since $E \cap D_1 = \emptyset$, we have

$$\begin{aligned} \phi_1 &= \prod_{i=1}^{j} 1_i \cdot \sum_{D_1} \prod_{i=1}^{j} p(v_i|P_i) \cdot \prod_{i=j+1}^{k} p(v_i|P_i) \\ &= \prod_{i=1}^{j} 1_i \cdot \sum_{D_1} \prod_{i=1}^{k} p(v_i|P_i). \end{aligned}$$

Similarly, for ϕ_2,

$$\begin{aligned} \phi_2 &= \sum_{D_2} \prod_{i=k+1}^{l} (p(v_i|P_i) \cdot 1_i) \cdot \prod_{i=l+1}^{m} p(v_i|P_i) \\ &= \prod_{i=k+1}^{l} 1_i \cdot \sum_{D_2} \prod_{i=k+1}^{l} p(v_i|P_i) \cdot \prod_{i=l+1}^{m} p(v_i|P_i) \\ &= \prod_{i=k+1}^{l} 1_i \cdot \sum_{D_2} \prod_{i=k+1}^{m} p(v_i|P_i). \end{aligned}$$

Thus, the product $\phi_1 \cdot \phi_2$ is

$$\prod_{i=1}^{j} 1_i \cdot \sum_{D_1} \prod_{i=1}^{k} p(v_i|P_i) \cdot \prod_{i=k+1}^{l} 1_i \cdot \sum_{D_2} \prod_{i=k+1}^{m} p(v_i|P_i).$$

By SO, D_2 and D_1 have no common variables with the CPTs $p(v_i|P_i)$, $i = 1, \ldots, k$ and $i = k+1, \ldots, m$, respectively. Therefore, we have:

$$\prod_{i=1}^{j} 1_i \cdot \prod_{i=k+1}^{l} 1_i \cdot \sum_{D_1 D_2} \prod_{i=1}^{k} p(v_i|P_i) \cdot \prod_{i=k+1}^{m} p(v_i|P_i).$$

Rearranging yields

$$\prod_{i=1}^{j} 1_i \cdot \prod_{i=k+1}^{l} 1_i \cdot \sum_{D_1 D_2} \prod_{i=1}^{m} p(v_i|P_i).$$

Let $\phi(X|Z) = \prod_{i=1}^{m} p(v_i|P_i)$. Then we have

$$\prod_{i=1}^{j} 1_i \cdot \prod_{i=k+1}^{l} 1_i \cdot \sum_{D_1 D_2} \phi(X|Z).$$

As $D_1 D_2 \subseteq X$ (Butz et al., 2010), let $W = X - D_1 D_2$. Thus, the multiplication $\phi_1 \cdot \phi_2$ is

$$1(E \cap WZ) \cdot \phi(W|Z),$$

where the product of the evidence potentials is $1(E \cap WZ)$. □

Corollary 1. *In the proof of Theorem 3, $\phi(W|Z)$ is $p(W|Z)$, provided there is a topological ordering $\prec$ of B in which the variables in DW appear consecutively, where D are those variables that were eliminated to build $\phi(W|Z)$.*

6 Advantages

We stress the improvement in clarity and suggest a direction of practical investigation.

Kjaerulff and Madsen (2008) suggest that in working with probabilistic networks it is convenient to denote distributions as potentials. Similarly, Koller and Friedman (2009) would denote the start of Example 8 as

$$\sum_b \psi_2(b) \cdot \psi_5(e,b,d) \cdot \psi_4(d,a,b) \cdots.$$

Observe that both the p-labels and the CPT structure have been destroyed even before the distributions in memory have been modified. It is then more meaningful to keep the CPT structure and semantics highlighted in (9) - (16).

Consider evidence $i = 1$ and $h = 0$ in the BN in Figure 3, which Koller and Friedman (2009) call non-trivial. All intermediate distributions are denoted as potentials in the computation of $p(j \mid i = 1, h = 0)$. However, it follows from Theorem 3 and Corollary 1 that structure and semantics can still be identified.

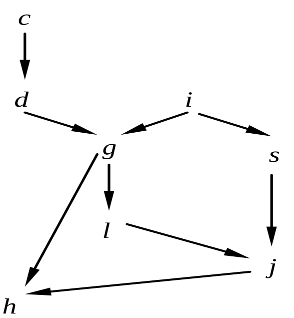

Figure 3: The DAG of a non-trivial BN.

Example 10. Computing $p(j \mid i = 1, h = 0)$ in the BN of Figure 3 involves, in part,

$$\sum_{c,d,g,s,l} p(c) \cdot p(d|c) \cdot p(i) \cdot p(s|i) \cdot p(g|d,i) \cdot p(l|g) \cdot p(j|l,s) \cdot p(h|g,j) \cdot 1(i=1) \cdot 1(h=0).$$

Eliminating variables c and d requires

$$\sum_d p(g|d,i) \cdot 1(i=1) \cdot \sum_c p(c) \cdot p(d|c) \quad (18)$$

$$= \sum_d p(g|d,i=1) \cdot \sum_c p(c,d) \quad (19)$$

$$= \sum_d p(g|d,i=1) \cdot p(d) \quad (20)$$

$$= \sum_d p(d,g|i=1)$$

$$= p(g|i=1).$$

Variable g can be eliminated as:

$$\sum_g p(g|i=1) \cdot p(l|g) \cdot p(h|g,j) \cdot 1(h=0)$$

$$\sum_g p(g|i=1) \cdot p(l|g) \cdot p(h=0|g,j)$$

$$= \sum_g p(g,l|i=1) \cdot p(h=0|g,j) \quad (21)$$

$$= \sum_g \phi(g,l,h=0|i=1,j) \quad (22)$$

$$= \phi(l,h=0|i=1,j). \quad (23)$$

The remainder of the example is omitted.

Example 10 shows that all intermediate distributions have structure and semantics, regardless of: the involvement of evidence potentials (18); the side or sides of the bar on which evidence appears (21), (22); marginalization operations (19); and p-labels (20) or ϕ-labels (23). Now let us turn to efficiency issues.

All previous join tree propagation algorithms either exclusively apply VE or *arc reversal* (AR) (Olmsted, 1983) at all join tree nodes (Madsen, 2010), or pick whether to apply VE or AR at each node (Butz et al., 2009a). A practical advantage of our new semantics is the ability to construct messages using both VE and AR at the *same* join tree node.

Example 11. Consider a join tree with three nodes $abcdefgh$, $fgij$ and fik. The CPTs assigned to $abcdefgh$ are $p(a)$, $p(b|a)$, $p(c|b)$, $p(d|c)$, $p(e|b,d)$, $p(f|e)$, $p(g|e)$, $p(h|a,b,c,d,e,f,g)$, while $fgij$ is provided $p(i|g)$ and $p(j|f,g,i)$, and fik is given $p(k|f,i)$. Butz et al.'s (2009b) message identification process indicates that $abcdefgh$ will pass $p(f)$ and $p(g|f)$ to node $fgij$, which, in turn, will pass $p(f)$ and $p(i|f)$ to fik. Madsen (2010) and Butz et al. (2009a) would apply AR at node $abcdefgh$, i.e., with some abuse of notation:

$$p(f) \cdot p(g|f) = \sum_{a,b,c,d,e}^{AR} p(a) \cdots p(f|e) \cdot p(g|e),$$

where h is removed as a barren variable. AR must be applied to remove the last variable e. However, by examining the semantics of VE, it can be verified that the elimination of variables a, b, c and d gives $p(e)$. Thus, apply VE to eliminate variables a, b, c and d, and then apply AR to eliminate variable e:

$$\sum_{e}^{AR} \sum_{a,b,c,d}^{VE} p(a) \cdot p(b|a) \cdot p(c|b) \cdot p(d|c) \cdots p(g|e)$$

$$= \sum_{e}^{AR} \sum_{b,c,d}^{VE} p(b) \cdot p(c|b) \cdot p(d|c) \cdot p(e|b,d) \cdots p(g|e)$$

$$= \sum_{e}^{AR} \sum_{c,d}^{VE} \phi(c,e|d) \cdot p(d|c) \cdot p(f|e) \cdot p(g|e)$$

$$= \sum_{e}^{AR} \sum_{d}^{VE} p(d,e) \cdot p(f|e) \cdot p(g|e)$$

$$= \sum_{e}^{AR} p(e) \cdot p(f|e) \cdot p(g|e).$$

It can be verified that applying AR at $abcdefgh$ requires more computation than applying the combination of VE and AR as shown above.

Empirical results will be reported separately.

7 Conclusions

Pearl (1988) emphasizes that probabilistic reasoning is not about numbers and is instead about the structure of reasoning. Our work here ascribes semantics to the intermediate CPTs of VE. This is the primary contribution of this paper. A practical advantage of these semantics was illustrated using the latest optimization techniques employed in join tree propagation and requires further study.

Intermediate CPTs constructed by VE could be labeled solely with p-labels, provided the label of each distribution is an expression rather than a single term. That is, label each CPT output by SO as a fraction, where the numerator is the p-label CPT output by SOP and the denominator is the factorization of "missing" CPTs in Θ. Thus, whereas SO can eliminate variable b in Example 7 as $\phi(c,e|a,d)$, it may be semantically more meaningful to take another step and label it $p(c,d,e|a)/p(d|c)$.

Acknowledgments

This research is supported by NSERC Discovery Grant 238880. The authors thank K. Williams for useful comments.

References

C.J. Butz, K. Konkel and P. Lingras. 2009a. Join tree propagation utilizing both arc reversal and variable elimination. In *Twenty Second International Florida Artificial Intelligence Research Society Conference*, pages 523–528.

C.J. Butz, H. Yao and S. Hua. 2009b. A join tree probability propagation architecture for semantic modeling, *Journal of Intelligent Information Systems*, 33(2):145-178.

C.J. Butz, W. Yan, P. Lingras and Y.Y. Yao. 2010. The CPT structure of variable elimination in discrete Bayesian networks. Advances in Intelligent Information Systems. SCI 265. Z.W. Ras and L.S. Tsay (Eds.). Springer, pages 245-257.

T.H. Cormen, C.E. Leiserson, R.L. Rivest and C. Stein. 2009. *Introduction to Algorithms*. MIT Press.

R.G. Cowell, A.P. Dawid, S.L. Lauritzen and D.J. Spiegelhalter. 1999. *Probabilistic Networks and Expert Systems*. Springer.

F.V. Jensen and T.D. Nielsen. 2007. *Bayesian Networks and Decision Graphs*. Springer.

U.B. Kjaerulff and A.L. Madsen. 2008. *Bayesian Networks and Influence Diagrams*. Springer.

D. Koller and N. Friedman. 2009. *Probabilistic Graphical Models: Principles And Techniques*. MIT Press.

A.L. Madsen. 2010. Improvements to message computation in Lazy propagation. *Int. J. Approx. Reason*, 51(5):499-514.

S. Olmsted. 1983. On representing and solving decision problems, Ph.D. Thesis, Department of Engineering Economic Systems, Stanford University, Stanford, California.

J. Pearl. 1988. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann.

G. Shafer. 1996. *Probabilistic Expert Systems*. SIAM.

N.L. Zhang and D. Poole. 1996. Exploiting causal independence in Bayesian network inference, *J. Artif. Intell. Res.*, (5):301-328.