d-Separation: Strong Completeness of Semantics in Bayesian Network Inference

Cory J. Butz[1], Wen Yan[1], and Anders L. Madsen[2,3]

[1] Department of Computer Science, University of Regina, Canada
{butz,yanwe111}@cs.uregina.ca
[2] HUGIN EXPERT A/S, Aalborg, Denmark
anders@hugin.com
[3] Aalborg University, Department of Computer Science, Denmark

Abstract. It is known that d-separation can determine the minimum amount of information needed to process a query during exact inference in discrete Bayesian networks. Unfortunately, no practical method is known for determining the semantics of the intermediate factors constructed during inference. Instead, all inference algorithms are relegated to denoting the inference process in terms of potentials. In this theoretical paper, we give an algorithm, called *Semantics in Inference* (SI), that uses d-separation to denote the semantics of every potential constructed during inference. We show that SI possesses four salient features: polynomial time complexity, soundness, completeness, and strong completeness. SI provides a better understanding of the theoretical foundation of Bayesian networks and can be used for improved clarity, as shown via an examination of Bayesian network literature.

1 Introduction

In [12], Pearl advocated the restoration of probabilistic methods in artificial intelligence systems and explored the possibility of representing and manipulating probabilistic knowledge in graphical forms, latter called Bayesian networks. When recounting the development of Bayesian networks, Pearl [14] states that perhaps [12] made its greatest immediate impact through the notion of *d-separation*. As a method for deciding which conditional independence relations are implied by the directed acyclic graph of a Bayesian network, d-separation provides the semantics needed for defining and characterizing Bayesian networks. Observe that Pearl emphasizes the importance of d-separation with respect to Bayesian network modeling. With respect to inference, Pearl only states that d-separation can determine the minimum information needed for answering a query posed to a Bayesian network. No claim has ever been made that d-separation can also provide semantics during Bayesian network inference.

Koller and Friedman [8] state that it is interesting to consider the semantics of the potentials constructed during inference. They mention that sometimes the probabilities are defined with respect to the joint distribution, but not at other times. As no practical algorithm exists for deciding the semantics of inference, all inference algorithms denote the intermediate factors constructed during inference

O. Zaïane and S. Zilles (Eds.): Canadian AI 2013, LNAI 7884, pp. 13–24, 2013.

as potentials. Potentials have no constraints [8] meaning they do not have clear physical interpretation [4].

In this theoretical paper, we present *Semantics in Inference* (SI), an algorithm for denoting semantics during exact inference in discrete Bayesian networks. SI works by introducing the notion of evidence normal form to organize how each potential was constructed. SI then decides semantics of the potential by performing one d-separation test. Formal properties of the SI algorithm are obtained, namely, polynomial time complexity, soundness, completeness, and strong completeness. SI can be utilized for clarity of exposition in Bayesian network literature, since the semantics of potentials can now be articulated.

2 Inference

Here we consider only discrete Bayesian networks. $U = \{v_1, v_2, \ldots, v_n\}$ is a finite set of random variables and each $v_i \in U$ can take a value from a finite domain, $dom(v_i)$. Given $X \subseteq U$, $dom(X)$ is the Cartesian product of $dom(v_i)$, $v_i \in X$. A *potential* on $dom(X)$ is a function ψ on $dom(X)$ such that $\psi(x) \geq 0$ for each $x \in dom(X)$, and at least one $\psi(x)$ is positive. For brevity, we refer to ψ as a mapping on X rather than $dom(X)$. A potential p on U that sums to 1 is called a *joint probability distribution* on U, denoted $p(U)$. A *conditional probability table* (CPT) for X given disjoint Y, denoted $\psi(X|Y)$, is a potential on XY that sums to 1, for each configuration $y \in dom(Y)$. The *unity-potential* $1(v_i)$ for v_i is a function 1 mapping every element of $dom(v_i)$ to one. The unity-potential for a non-empty set $X = \{v_1, v_2, \ldots, v_k\}$ of variables, denoted $1(X)$, is defined as $1(X) = 1(v_1) \cdot 1(v_2) \cdots 1(v_k)$. For simplified notation, we may write $\{v_1, v_2, \ldots, v_k\}$ as $v_1, v_2, \ldots, v_k$.

A *Bayesian network* [13] is a pair (B, C). B denotes a directed acyclic graph with vertex set U and C is a set of *conditional probability tables* (CPTs) $\{p(v_i|P(v_i)) \mid i = 1, 2, \ldots, n\}$, where $P(v_i)$ denotes the parents (immediate predecessors) of $v_i \in B$. The product of CPTs in C is a joint probability distribution $p(U)$. For example, the directed acyclic graph in Figure 1 is called the *extended student Bayesian network* (ESBN) [8]. We give CPTs in Table 1, where only binary variables are used in examples, and probabilities not shown can be obtained by definition. By the above,

$$p(U) \ = \ p(c) \cdot p(d|c) \cdot p(i) \cdot p(g|d, i) \cdots p(h|g, j). \tag{1}$$

We say X and Z are *conditionally independent* [16] given Y in $p(U)$, denoted $I_p(X, Y, Z)$, if given any $x \in dom(X)$, $y \in dom(Y)$, for all $z \in dom(Z)$: $p(x|y, z) = p(x|y)$, whenever $p(y, z) > 0$, where $X, Y, Z \subseteq U$.

Pearl [12] gave a method, called d-separation, for determining those independencies encoded in a directed acyclic graph. The following is the definition of d-separation based on [8]. In a Bayesian network B, a trail (an undirected path) $v_1, v_2, \ldots, v_n$ is *active* given Y, if: (i) whenever we have a v-structure $v_{i-1} \rightarrow v_i \leftarrow v_{i+1}$, then v_i or one of its descendants are in Y; (ii) no other node along the trail is in Y. Note that if v_1 or v_n are in Y the trail is not active.

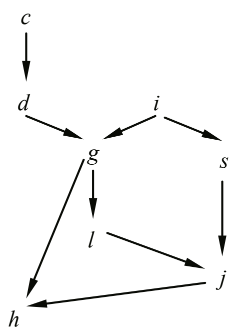

Fig. 1. The directed acyclic graph of ESBN

Table 1. CPTs for the ESBN in Figure 1

c	$p(c)$
0	0.20

i	$p(i)$
0	0.75

c	d	$p(d\|c)$
0	0	0.40
1	0	0.70

d	i	g	$p(g\|d,i)$
0	0	0	0.90
0	1	0	0.20
1	0	0	0.50
1	1	0	0.40

g	l	$p(l\|g)$
0	0	0.30
1	0	0.60

g	j	h	$p(h\|g,j)$
0	0	0	0.25
0	1	0	0.65
1	0	0	0.50
1	1	0	0.85

s	l	j	$p(j\|s,l)$
0	0	0	0.10
0	1	0	0.60
1	0	0	0.45
1	1	0	0.50

i	s	$p(s\|i)$
0	0	0.40
1	0	0.80

We say that X and Z are *d-separated* given Y in B, denoted $I_B(X,Y,Z)$, if there is no active trail between any variable $v \in X$ and $v' \in Z$ given Y.

In inference, $p(X|E=e)$ is the most common query type, which are useful for many reasoning patterns, including explanation, prediction, intercausal reasoning, and many more [8]. Here, X and E are disjoint subsets of U, and E is observed taking value e. We describe a basic algorithm for computing $p(X|E=e)$, called *variable elimination* (VE), first put forth in [17]. We do not consider alternative approaches to inference such as conditioning [6] and join tree propagation [1,2,10]. Inference involves the elimination of variables. Algorithm 1, called *sum-out* (SO), eliminates a single variable v from a set Φ of *potentials* [8], and returns the resulting set of potentials. The algorithm *collect-relevant* simply returns those potentials in Φ involving variable v.

Algorithm 1. SO(v,Φ)
Ψ = collect-relevant(v,Φ)
ψ = the product of all potentials in Ψ
$\tau = \sum_v \psi$
return $(\Phi - \Psi) \cup \{\tau\}$

SO uses Lemma 1, which means that potentials not involving the variable being eliminated can be ignored.

Lemma 1. [15] *If ψ_1 is a potential on W and ψ_2 is a potential on Z, then the marginalization of $\psi_1 \cdot \psi_2$ onto W is the same as ψ_1 multiplied with the marginalization of ψ_2 onto $W \cap Z$, where $W, Z \subseteq U$.*

The *evidence potential* for $E = e$, denoted $1(E = e)$, assigns probability 1 to the single value e of E and probability 0 to all other values of E. Hence, for a variable v observed taking value λ and $v \in \{v_i\} \cup P(v_i)$, the product $p(v_i|P(v_i)) \cdot 1(v = \lambda)$ keeps only those configurations agreeing with $v = \lambda$.

Algorithm 2, taken from [8], computes $p(X|E = e)$ from a discrete Bayesian network B. VE calls SO to eliminate variables one by one. More specifically, in Algorithm 2, Φ is the set C of CPTs for B, 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 $U - XE$, where XE denotes $X \cup E$.

Algorithm 2. VE(Φ, X, E, e, σ)
Multiply evidence potentials with appropriate CPTs
While σ is not empty
  Remove the first variable v from σ
  $\Phi \ = \ \text{sum-out}(v, \Phi)$
$p(X, E = e)$ = the product of all potentials $\psi \in \Phi$
return $p(X, E = e) / \sum_X p(X, E = e)$

As in [8], suppose the observed evidence for the ESBN is $i = 1$ and $h = 0$ and the query is $p(j|h = 0, i = 1)$. The weighted-min-fill algorithm [8] can yield $\sigma = (c, d, l, s, g)$. VE first incorporates the evidence:

$$\begin{aligned} \psi(i=1) &= p(i) \cdot 1(i=1), \\ \psi(d,g,i=1) &= p(g|d,i) \cdot 1(i=1), \\ \psi(i=1,s) &= p(s|i) \cdot 1(i=1), \\ \psi(g,h=0,j) &= p(h|g,j) \cdot 1(h=0). \end{aligned}$$

To eliminate c, the SO algorithm computes

$$\psi(d) \ = \ \sum_c p(c) \cdot p(d|c).$$

SO computes the following to eliminate d

$$\psi(g, i=1) \ = \ \sum_d \psi(d) \cdot \psi(d, g, i=1).$$

To eliminate l,

$$\psi(g, j, s) \ = \ \sum_l p(l|g) \cdot p(j|l, s).$$

SO computes the following when eliminating s,

$$\psi(g, i = 1, j) \;=\; \sum_{s} \psi(i = 1, s) \cdot \psi(g, j, s). \tag{2}$$

For g, SO can compute:

$$\begin{aligned} & \sum_{g} \psi(g, i = 1, j) \cdot \psi(g, i = 1) \cdot \psi(g, h = 0, j) \\ =\; & \sum_{g} \psi(g, i = 1, j) \cdot \psi(g, h = 0, i = 1, j) \\ =\; & \psi(h = 0, i = 1, j). \end{aligned} \tag{3}$$

Next, VE multiplies all remaining potentials as

$$p(h = 0, i = 1, j) \;=\; \psi(i = 1) \cdot \psi(h = 0, i = 1, j).$$

Finally, VE answers the query by

$$p(j|h = 0, i = 1) \;=\; \frac{p(h = 0, i = 1, j)}{\sum_{j} p(h = 0, i = 1, j)}.$$

3 Understanding Semantics

We review the current limited understanding of semantics in inference.

Kjaerulff and Madsen [7] suggest that in working with probabilistic networks it is convenient to denote distributions as potentials. In fact, the use of potentials is built into the standard inference algorithms (see the SO and VE algorithms, for instance). For example, suppose query $p(j)$ is posed to the ESBN [8]. Even without evidence being considered, the initial step of VE is to regard CPTs as potentials, i.e., $p(U)$ is factorized as

$$p(U) \;=\; \psi(c) \cdot \psi(c, d) \cdot \psi(i) \cdots \psi(g, h, j). \tag{4}$$

By comparing (1) and (4), it is clear that semantics are destroyed even before the CPTs in computer memory are modified. The notation used for potentials does not convey the semantic meaning of the probabilities comprising the potential.

Darwiche [6] ascribes meaning during inference by representing each potential by what we will call evidence expanded form, except that products involving evidence potentials are taken. Let ψ be any potential constructed by VE. The *evidence expanded form* of ψ, denoted $F(\psi)$, is the unique expression defining how ψ was built using the multiplication and marginalization operators on the Bayesian network CPTs together with any appropriate evidence potentials.

For example, consider potential $\psi(g, i = 1, j)$ in (2). $F(\psi(g, i = 1, j))$, the evidence expanded form, can be easily obtained in a recursive manner as follows:

$$\begin{aligned} & \sum_s \psi(i=1,s) \cdot \psi(g,j,s) \\ = & \sum_s \psi(i=1,s) \cdot (\sum_l (p(l|g) \cdot p(j|l,s))) \\ = & \sum_s ((p(s|i) \cdot 1(i=1)) \cdot (\sum_l (p(l|g) \cdot p(j|l,s)))). \end{aligned} \quad (5)$$

Henceforth, parentheses are understood and may not be shown. Unfortunately, the expanded form by itself does not directly articulate semantics.

By semantics, we mean that a CPT $\psi(X|Y)$ constructed by VE's manipulation of Bayesian network CPTs is not necessarily equal to the CPT $p(X|Y)$ obtained from the defined joint probability distribution $p(U)$. For instance, it can be verified that in the ESBN,

$$p(h|g,j) \cdot \sum_d p(g|d,i) \cdot \sum_c p(c) \cdot p(d|c) \quad (6)$$

produces the CPT $\psi(g, h|i, j)$ in Table 2 (left). In contrast, the CPT $p(g, h|i, j)$ built from the joint distribution $p(U)$ in (1) is shown in Table 2 (right).

Table 2. (left) CPT $\psi(g, h|i, j)$ built by (6). (right) CPT $p(g, h|i, j)$ built from $p(U)$ in (1).

i	j	g	h	$\psi(g,h\|i,j)$	i	j	g	h	$p(g,h\|i,j)$
0	0	0	0	0.1890	0	0	0	0	0.1960
0	0	0	1	0.5670	0	0	0	1	0.5880
0	0	1	0	0.1220	0	0	1	0	0.1080
0	1	0	0	0.4914	0	1	0	0	0.4762
0	1	0	1	0.2646	0	1	0	1	0.2564
0	1	1	0	0.2074	0	1	1	0	0.2272
1	0	0	0	0.0680	1	0	0	0	0.0846
1	0	0	1	0.2040	1	0	0	1	0.2537
1	0	1	0	0.3640	1	0	1	0	0.3309
1	1	0	0	0.1768	1	1	0	0	0.1518
1	1	0	1	0.0952	1	1	0	1	0.0817
1	1	1	0	0.6188	1	1	1	0	0.6515

Semantics in inference are not well understood. In their comprehensive and highly recommended text, Koller and Friedman [8] consider the semantics of potential $\psi(b, c, d)$ built by eliminating variable a from the Bayesian network B in Figure 2 (left):

$$\psi(b,c,d) \;=\; \sum_a p(a) \cdot p(b|a) \cdot p(d|a,c). \quad (7)$$

Koller and Friedman [8] incorrectly state

$$p(b, d|c) \neq \psi(b, c, d). \quad (8)$$

While this claim is almost always true, there are a few exceptions to refute it. For one counter-example, eliminating variable a using the CPTs in Table 3 yields:

$$p(b, d|c) = \psi(b, c, d). \quad (9)$$

Koller and Friedman [8] also state it must necessarily be the case that

$$p'(b, d|c) = \psi(b, c, d), \quad (10)$$

where $p'(U)$ is defined by a *different* Bayesian network B' - the one given in Figure 2 (right). Our objective is to stipulate semantics in the *current* Bayesian network B - the one on which inference is being conducted.

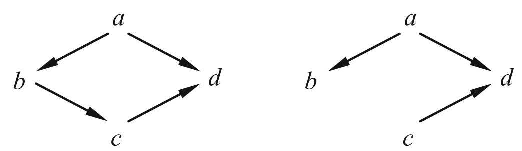

Fig. 2. Bayesian networks B (left) and B' (right)

Table 3. Exceptional CPTs for B in Figure 2 (left)

a	$p(a)$	a	b	$p(b\|a)$	b	c	$p(c\|b)$	a	c	d	$p(d\|a,c)$
0	0.2	0	0	0.4	0	0	0.5	0	0	0	0.5
		1	0	0.9	1	0	0.5	0	1	0	0.5
								1	0	0	0.5
								1	1	0	0.5

4 CPT Structure

It is instructive to review that, when evidence is not considered, each potential built by VE is a CPT.

A *topological ordering* [8] is an ordering $\prec$ of the variables in a Bayesian network B so that for every arc (v_i, v_j) in B, v_i precedes v_j in $\prec$. For example, $c \prec d \prec i \prec g \prec s \prec l \prec j \prec h$ is a topological ordering of the directed acyclic graph in Figure 1, but $d \prec c \prec i \prec g \prec h \prec l \prec j \prec s$ is not.

Recall this feature of Bayesian networks,

$$p(U) = \prod_{v_i \in U} p(v_i|P(v_i)).$$

This can be established by showing

$$1 = \sum_U \prod_{v_i \in U} p(v_i|P(v_i)).$$

More generally, we have the following two lemmas.

Lemma 2. [3] *Consider a Bayesian network (B, C) on U. Given any non-empty subset X of U, $\prod_{v_i \in X} p(v_i|P(v_i))$ is a CPT $\psi(X|P(X))$, where $P(X) = (\cup_{v_i \in X} P(v_i)) - X$.*

Lemma 3. [3] *When evidence is not considered, each potential constructed by VE is a CPT.*

Lemma 3 can be seen as first applying Lemma 1 on the evidence expanded form of a potential built by VE, keeping in mind $E = \emptyset$, and then applying Lemma 2.

For example, consider the potential ψ built by (6), which is already in evidence expanded form. By applying Lemma 1,

$$\sum_d \sum_c p(h|g, j) \cdot p(g|d, i) \cdot p(c) \cdot p(d|c). \tag{11}$$

By Lemma 2,

$$\psi(g, h|i, j) \;=\; \sum_d \sum_c \psi(c, d, g, h|i, j),$$

Thus, the potential ψ built by (6) is, in fact, a CPT $\psi(g, h|i, j)$, in Table 2 (left).

5 Denoting Semantics

The evidence expanded form $F(\psi)$ of any potential ψ constructed by VE is in evidence normal form, if $F(\psi)$ is written as

$$\gamma \cdot N,$$

where γ is the product of 1 and all evidence potentials in $F(\psi)$, and N is the same factorization as $F(\psi)$ except without products involving evidence potentials.

Recall $\psi(g, h = 0, i = 1, j)$ in (3). The evidence expanded form $F(\psi)$ is

$$p(h|g, j) \cdot 1(h = 0) \cdot \sum_d p(g|d, i) \cdot 1(i = 1) \cdot \sum_c p(c) \cdot p(d|c), \tag{12}$$

and the evidence normal form $\gamma \cdot N$ is

$$1(h = 0, i = 1) \;\cdot\; p(h|g, j) \cdot \sum_d p(g|d, i) \cdot \sum_c p(c) \cdot p(d|c), \tag{13}$$

namely, $\gamma = 1(h = 0, i = 1)$ and N is (6).

Lemma 4. *The evidence expanded form $F(\psi)$ of any potential ψ constructed by VE always can be equivalently written in normal form, i.e., $F(\psi) = \gamma \cdot N$.*

Proof. Since evidence variables are never marginalized in VE, the claim follows from Lemma 1.

Observe that, by Lemma 3, N in evidence normal form is a CPT. We may denote evidence normal form $\gamma \cdot N$ simply as N with evidence γ understood, since γ only serves to select configurations of N agreeing with the evidence. We now turn to denoting semantics.

To understand when $N = p(X|Y)$ in evidence normal form, some terminology is needed. 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}) in B, $i = 1, \ldots, n-1$. With respect to a variable v_i, we define three sets: (i) the ancestors of v_i, denoted $A(v_i)$, are those variables having a path to v_i; (ii) the descendants of v_i, denoted $D(v_i)$, are those variables to which v_i has a path; and, (iii) 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 $X \subseteq U$ are defined as $A(X) = (\cup_{v_i \in X} A(v_i)) - X$. The descendants $D(X)$ are defined similarly. $I_B(X, Y, Z)$ means an independence statement $I(X, Y, Z)$ [13] holds in B by d-separation, where $X, Y, Z \subseteq U$.

We now give the *Semantics in Inference* (SI) algorithm, which uses d-separation to denote the semantics of any potential ψ built by VE on B. Each potential ψ constructed by VE is represented in evidence normal form $\psi(X|Y)$. If the semantics of B ensure the $\psi(X|Y) = p(X|Y)$, then ψ is denoted as $p_B(X|Y)$; otherwise, it is denoted as $\phi_B(X|Y)$. S is the set of variables marginalized in $F(\psi)$. $A(XS)$ and $D(XS)$ are computed from the *transitive closure*, denoted T, of B [5].

Algorithm 3. SI(ψ)
Compute the evidence expanded form $F(\psi)$ of ψ
Compute the evidence normal form $\gamma \cdot N$ of $F(\psi)$
Compute the CPT structure $\psi(X|Y)$ of N
Compute $Z = A(XS) \cap D(XS)$
Compute $X_1 = X \cap P(Z)$
if $I_B(X_1, \emptyset, Y)$ holds in B by d-separation
  return $p_B(X|Y)$
else
  return $\phi_B(X|Y)$

Recall $\psi(g, i = 1, j)$ in (2). The evidence expanded form is (6). Its evidence normal form $\gamma \cdot N$ is $\gamma = 1(i = 1)$ and $N = \psi(j|g, i)$. Now $X = \{j\}$, $Y = \{g, i\}$ and $S = \{l, s\}$. By the transitive closure T of the ESBN, $A(XS) = \{c, d, i, g\}$ and $D(XS) = \{h\}$. Hence, $Z = \emptyset$, $P(Z) = \emptyset$, and $X_1 = \emptyset$. Trivially, $I_B(X_1, \emptyset, Y)$ holds. Thus, SI denotes $\psi(g, i = 1, j)$ in (2) as $p_B(j|g, i = 1)$.

Now consider $\psi(g, h = 0, i = 1, j)$ in (3). The evidence expanded form is (12). The evidence normal form $\gamma \cdot N$ is (13). Here $N = \psi(g, h|i, j)$, as seen in (6). With $X = \{g, h\}$, $Y = \{i, j\}$ and $S = \{c, d\}$, from T on the ESBN we have $A(\{c, d, g, h\}) = \{i, j, l, s\}$ and $D(\{c, d, g, h\}) = \{j, l\}$. Thus, $Z = \{j, l\}$, giving $P(Z) = \{g, s\}$ and $X_1 = \{g\}$. Now, $I_B(X_1, \emptyset, Y)$ does not hold. Thereby, SI denotes $\psi(g, h = 0, i = 1, j)$ in (3) as $\phi_B(g, h = 0|i = 1, j)$.

In this example, there is a path from $XS = \{c, d, g, h\}$ to XS through $Z = \{j, l\}$, starting at $X_1 = \{g\}$. With $X_1 = \{g\}$ and $Y = \{i, j\}$, we focus on $I_B(g, \emptyset, ij)$. Note that when deciding semantics of $\psi(X|Y)$, the independence to be tested is $I_B(X_1, \emptyset, Y)$ and not $I_B(XS, Y, A(XSY))$. In Figure 2 (left), $I_B(abd, c, \emptyset)$ holds, but $p(b, d|c) \neq \psi(b, d|c)$ in (8) is possible.

6 Theoretical Foundation

We present four salient features of SI. Only the proofs of time complexity and strong completeness are shown due to space considerations.

Theorem 1. *Let ψ be any potential built by VE during exact inference in a discrete Bayesian network with n variables. Then the time complexity of the SI algorithm to determine the semantics of ψ is $O(n^3)$.*

Proof. As ψ may require $n-1$ multiplications and n marginalizations, computing $F(\psi)$ takes $2n$ steps. The normal form $\gamma \cdot N$ can be decided in linear time, as can the CPT structure $\psi(X|Y)$ of N. The transitive closure T of the directed acyclic graph can be computed in $O(n^3)$ [5]. Let XS be a set of k variables, $1 \leq k \leq n$. Then $A(XS)$ and $D(XS)$ each can be computed in $O(k \cdot n)$. Now Z and X_1 each can be computed in $O(n^2)$. Testing $I_B(X_1, \emptyset, Y)$ is linear in the size of B [6]. Thus, the semantics of ψ can be determined by SI in $O(n^3)$.

Theorem 2. *In a Bayesian network (B, C) defining a joint distribution $p(U)$, suppose VE computes a potential ψ whose evidence normal form is $\gamma \cdot N$. If SI denotes the semantics of N as $p_B(X|Y)$, then $N = p(X|Y)$.*

Theorem 2 guarantees that if SI denotes the semantics of a VE potential ψ as $\gamma \cdot p_B(X|Y)$, then

$$\psi \ = \ \gamma \cdot p(X|Y).$$

Recall potential $\psi(g, i = 1, j)$ in (2). As illustrated in Table 4, Theorem 2 ensures that $\psi(g, i = 1, j)$ is equal to $p(j|g, i = 1)$, since SI denotes it as $p_B(j|g, i = 1)$.

Table 4. Potential $\psi(g, i = 1, j)$ in (2) is $p(j|g, i = 1)$

$\psi(g, i = 1, j) \ = \ p(j|g, i = 1) \ =$

i	g	j	$p_B(j\|g, i = 1)$
1	0	0	0.457
1	0	1	0.543
1	1	0	0.334
1	1	1	0.666

With respect to inference, the question of completeness is this. Can SI determine the semantics of every VE potential defined with respect to the joint distribution? The answer is no.

Theorem 3. *In a Bayesian network B on U, suppose VE computes a potential ψ whose evidence normal form is $\gamma \cdot N$. If SI denotes the semantics of N as $\phi_B(X|Y)$, there exists a set C of CPTs for B defining a joint distribution $p(U)$ such that $N \neq p(X|Y)$.*

Theorem 3 states that whenever SI indicates that a potential is not defined with respect to the joint distribution, then this is true for at least one set of CPTs for the given Bayesian network. Recall once again $\psi(g, h = 0, i = 1, j)$ in (3), which SI denotes as $\phi_B(g, h = 0, l|i = 1, j)$. With respect to $p(U)$ defined by the CPTs in Table 1, we have

$$\psi(g, h = 0, i = 1, j) \;\neq\; p(g, h = 0|i = 1, j).$$

However, Theorem 3 can be made significantly stronger.

Lemma 5. [11] *Except for a measure zero set in the space of all joint distributions $p(U)$ defined by all discrete Bayesian networks (B, C), the independencies satisfied by $p(U)$ are precisely those satisfied by d-separation in B.*[1]

Lemma 5 says that for nearly all choices C of CPTs for a Bayesian network B defining $p(U)$, d-separation perfectly characterizes the independencies in $p(U)$, i.e., for $X, Y, Z \subseteq U$,

$$I_p(X, Y, Z) \iff I_B(X, Y, Z).$$

Theorem 4. *Except for a measure zero set in the space of all joint distributions $p(U)$ defined by all discrete Bayesian networks (B, C), for any potential ψ built by VE,*

$$\psi \;=\; \gamma \cdot p(X|Y) \iff \text{SI denotes } \psi \text{ as } p_B(X|Y),$$

where $\gamma \cdot N$ is the evidence normal form of ψ.

Proof. ($\Rightarrow$) Suppose VE constructs a ψ whose evidence normal form is $\gamma \cdot N$ and whose semantics are defined with respect to $p(U)$. By contraposition, suppose SI denotes N as $\phi_B(X|Y)$. By SI, $I_B(X_1, \emptyset, Y)$ does not hold. Then, by Lemma 5, $I_p(X_1, \emptyset, Y)$ does not hold in essentially all possible $p(U)$ defined over B. It follows that for each such $p(U)$,

$$\gamma \cdot p(X|Y) \;\neq\; \gamma \cdot N.$$

A contradiction to our initial assumption. Therefore, SI correctly denotes the potential ψ as $p_B(X|Y)$.

($\Leftarrow$) Follows directly from Theorem 2.

Let B be any Bayesian network. Theorem 4 states that for nearly all choices C of CPTs for B, the SI algorithm correctly denotes the semantics of potentials constructed by VE during exact inference on B.

[1] A set has measure zero if it is infinitesimally small relative to the overall space [8].

7 Conclusion

We extend d-separation's role from determining the minimum amount of information needed to answer a query $p(X|E = e)$ [12] to also giving the semantics of the potentials constructed when answering $p(X|E = e)$. Our results contribute to a deeper understanding of Bayesian networks, since semantics of VE's intermediate factors are now articulated with respect to the joint distribution. The main result (Theorem 4) showed that our SI algorithm correctly denotes the semantics of inference in nearly all Bayesian networks. Future work will include applying the results here to differential semantics in Bayesian networks [6,9].

References

1. Butz, C.J., Hua, S., Konkel, K., Yao, H.: Join Tree Propagation with Prioritized Messages. Networks 55(4), 350–359 (2010)
2. Butz, C.J., Konkel, K., Lingras, P.: Join Tree Propagation Utilizing both Arc Reversal and Variable Elimination. Int. J. Approx. Reasoning 52(7), 948–959 (2011)
3. Butz, C.J., Yan, W., Lingras, P., Yao, Y.Y.: The CPT Structure of Variable Elimination in Discrete Bayesian Networks. In: Ras, Z.W., Tsay, L.S. (eds.) Advances in Intelligent Information Systems. SCI, vol. 265, pp. 245–257. Springer, Heidelberg (2010)
4. Castillo, E., Gutiérrez, J., Hadi, A.: Expert Systems and Probabilistic Network Models. Springer, New York (1997)
5. Cormen, T.H., Leiserson, C.E., Rivest, R.L., Stein, C.: Introduction to Algorithms. MIT Press, Cambridge (2009)
6. Darwiche, A.: Modeling and Reasoning with Bayesian Networks. Cambridge University Press, New York (2009)
7. Kjaerulff, U.B., Madsen, A.L.: Bayesian Networks and Influence Diagrams. Springer, New York (2008)
8. Koller, D., Friedman, N.: Probabilistic Graphical Models: Principles and Techniques. MIT Press, Cambridge (2009)
9. Madsen, A.L.: A Differential Semantics of Lazy AR Propagation. In: 21st Conference on Uncertainty in Artificial Intelligence, pp. 364–371. Morgan Kaufmann, San Mateo (2005)
10. Madsen, A.L.: Improvements to Message Computation in Lazy Propagation. Int. J. Approximate Reasoning 51(5), 499–514 (2010)
11. Meek, C.: Strong Completeness and Faithfulness in Bayesian Networks. In: 11th Conference on Uncertainty in Artificial Intelligence, pp. 411–418. Morgan Kaufmann, San Mateo (1995)
12. Pearl, J.: Fusion, Propagation and Structuring in Belief Networks. Artif. Intell. 29, 241–288 (1986)
13. Pearl, J.: Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann, San Francisco (1988)
14. Pearl, J.: Belief Networks Revisited. Artif. Intell. 59, 49–56 (1993)
15. Shafer, G.: Probabilistic Expert Systems. SIAM, Philadelphia (1996)
16. Wong, S.K.M., Butz, C.J., Wu, D.: On the Implication Problem for Probabilistic Conditional Independency. IEEE Trans. Syst. Man Cybern. A 30(6), 785–805 (2000)
17. Zhang, N.L., Poole, D.: A Simple Approach to Bayesian Network Computations. In: 7th Canadian Conference on Artificial Intelligence, pp. 171–178. Springer, New York (1994)