ELSEVIER

Artificial Intelligence 82 (1996) 45-74

Artificial Intelligence

Knowledge representation and inference in similarity networks and Bayesian multinets

Dan Geiger[a,*], David Heckerman[b]

[a] Department of Computer Science, Technion Israel Institute of Technology, Haifa 32000, Israel
[b] Microsoft Corporation, One Microsoft Way, Redmond, WA 98052-6399, USA

Received February 1993

Abstract

We examine two representation schemes for uncertain knowledge: the similarity network (Heckerman, 1991) and the Bayesian multinet. These schemes are extensions of the Bayesian network model in that they represent asymmetric independence assertions. We explicate the notion of relevance upon which similarity networks are based and present an efficient inference algorithm that works under the assumption that every event has a nonzero probability. Another inference algorithm is developed that works under no restriction albeit less efficiently. We show that similarity networks are not *inferentially complete*—namely—not every query can be answered. Nonetheless, we show that a similarity network can always answer any query of the form: "What is the posterior probability of an hypothesis given evidence?" We call this property *diagnostic completeness*. Finally, we describe a generalization of similarity networks that can encode more types of asymmetric conditional independence assertions than can ordinary similarity networks.

1. Introduction

Traditional probabilistic approaches to knowledge acquisition and inference for diagnostic, classification, and pattern-recognition systems face a critical choice: either specify precise relationships between all relevant variables or make uniform independence assumptions throughout the model. The first choice is computationally infeasible except in very small domains, whereas the second choice is rarely justified and often yields inaccurate conclusions. *Bayesian networks* offer a compromise between the two extremes by encoding independence when possible and dependence when necessary.

[*] Corresponding author. E-mail: dang@cs.technion.ac.il.

SSDI 0004-3702(95)00014-3

They allow a wide spectrum of independence assertions to be considered by the model builder, so that a practical balance can be established between computational needs and the accuracy of conclusions.

Although Bayesian networks considerably extend traditional approaches, they are not sufficiently expressive to encode every independence assertion that may facilitate knowledge acquisition or speed up inference. One deficiency is their inability to represent naturally *asymmetric independence* assertions. Such assertions state that variables are independent for some but not necessarily for all of their values.

The *similarity network* is a probabilistic representation that addresses this deficiency for problems of diagnosis. The representation employs multiple Bayesian networks, and thereby allows the representation of asymmetric independence assertions. In practice, the representation has proved to be extremely useful, facilitating the construction of expert systems for the diagnosis of breast, lymph-node, intestine, ovary, skin, soft-tissue, testis, and thymus pathology [16,27], sleep disorders [28], eye diseases [13], and efficiency problems in gas turbines that generate electricity [2].

Heckerman [14] introduces the similarity network representation. In his work, he describes the representation from the perspective of the user, emphasizing the benefits of the representation for knowledge acquisition. The development does not concentrate on issues of probabilistic inference. In particular, Heckerman describes how a similarity network can be converted to a Bayesian network, and proposes that probabilistic inference be performed using the Bayesian network rather than using the similarity network. The disadvantage of this approach is that, in the process of generating a Bayesian network from a similarity network, one encodes asymmetric independence in the numbers rather than in the topology of the Bayesian network. Consequently, these asymmetric assertions are not available to the inference algorithm to speed up computations. In addition, Heckerman's developments are limited to models where (1) there exists an ordering over all variables that is consistent with every Bayesian network within the similarity network, and (2) no relationship among variables is deterministic. We overcome these constrains and also discuss more fully the situation when diagnostic hypotheses are not mutually exclusive or when the hypothesis variable is not a root node.

Moreover, in this paper, we offer several enhancements to the similarity network representation. We present an efficient inference algorithm that works under the assumption that every event has a nonzero probability. Another inference algorithm is developed that works under no restriction albeit less efficiently. In the processes of developing the later algorithm we introduce another representation of asymmetric independence called a *Bayesian multinet*, describe an algorithm that converts a similarity network into a Bayesian multinet, and show how to perform inference using the Bayesian multinet obtained.

We show that similarity networks are not *inferentially complete*—namely—not every query can be answered. Nonetheless, we show that every similarity network can answer any query of the form: "What is the posterior probability of an hypothesis given evidence?" We call this property *diagnostic completeness*. Finally, we describe a generalization of similarity networks that can encode more types of asymmetric conditional independence assertions than can ordinary similarity networks.

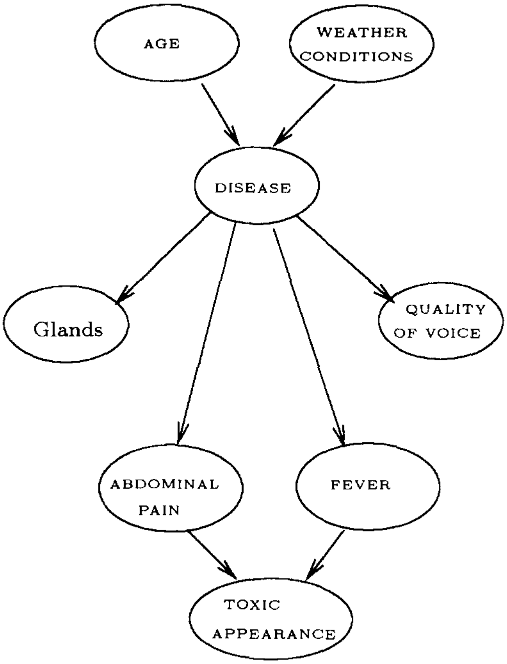

Fig. 1. A Bayesian network for the diagnosis of sore throat.

2. Bayesian networks: an overview[1]

2.1. Informal description

The Bayesian network paradigm was introduced to the AI community by Pearl [31, 32]. It is best explained via a simple example:

> Age and weather influence whether a child gets a sore throat. There are five mutually exclusive and exhaustive types of a sore throat: *viral pharyngitis*, *tonsillar cellulitis*, *mononucleosis*, *strep throat*, and *peritinsillar abscess*. Several symptoms are associated with a sore throat, including fever, toxic appearance, abdominal pain, swollen glands, and voice quality. Most symptoms occur independently of each other in patients having a sore throat, except toxic appearance, which depends upon having fever or abdominal pain.[2]

A Bayesian network representing this description is shown in Fig. 1. The network is constructed from cause-and-effect relationships by placing links from each cause to its direct consequences. For example, *fever* and *pain* are causes for *toxic appearance*, and *disease* is their common cause.

[1] This section is based on Geiger [8].

[2] A modified example of Heckerman [14].

Each node represents a variable having a finite set of values. Continuous variables such as *age* and *fever* are made discrete. For example, the values of *age* can be partitioned into: *infant*, *toddler*, and *school-age child*, and the values of *fever* can be partitioned into: *normal*, *moderately elevated*, and *markedly elevated*. (Work on continuous variables without discretization can be found in [10,32,36].)

Each variable u is associated with a conditional distribution $P(u \mid \pi(u))$, where $\pi(u)$ is the set of parents of u in the network. For example, $P(fever \mid disease)$ is specified by fifteen numbers: one for each value combination of the variables *disease* and *fever*. When u has no parents, that is, when $\pi(u) = \emptyset$, then u is associated with the marginal distribution $P(u)$. For example, $P(age)$ is specified by three numbers depending upon the relative proportion of infants, toddlers, and school-age children among the intended patients. Such distributions are associated with each node in the network.

From the chaining rule of probability, we know that

$$P(u_1,\ldots,u_n) = \prod_i P(u_i \mid u_1,\ldots,u_{i-1}). \tag{1}$$

If P satisfies

$$P(u_i \mid \pi(u_i)) = P(u_i \mid u_1,\ldots,u_{i-1}) \tag{2}$$

for $i = 1,\ldots,n$, then

$$P(u_1,\ldots,u_n) = \prod_i P(u_i \mid \pi(u_i)). \tag{3}$$

Each of the n equalities in Eq. (2) corresponds to an *independence assertion* stating that, given $\pi(u_i)$, u_i is independent of $\{u_1,\ldots,u_{i-1}\} \setminus \pi(u_i)$. (The symbol $\setminus$ stands for set difference.) The structure of a Bayesian network represents these independence assertions, as well as all those assertions implied by them. Thus, according to Eq. (3), a Bayesian network and its associated distributions $P(u_i \mid \pi(u_i))$ determine the joint distribution over $u_1,\ldots,u_n$.

According to Eq. (2), the network of Fig. 1 represents the assertion that *age* and *weather conditions* are independent—that is, $P(age) = P(age \mid weather)$. This assertion appears reasonable. Nonetheless, if this assertion—or any other independence assertion encoded in the network—does not accurately reflect the beliefs of the constructor of the network, then additional nodes or links are drawn until a sufficiently accurate model is realized. For example, one may argue that the life span on the North pole is generally shorter than that in California where weather conditions are more benign. This dependency between age and weather conditions could be modeled by adding a new node called climate and making it the parent of both weather conditions and age.

Another independence assertion implied by the network of Fig. 1 is that toxic appearance is independent of disease, given the values for fever and abdominal pain. That is,

$$\begin{aligned} &P(\textit{toxic appearance} \mid \textit{fever, pain, disease}) \\ &\quad = P(\textit{toxic appearance} \mid \textit{fever, pain}). \end{aligned}$$

This assertion reflects the view that fever and pain are the only intervening mechanisms by which a disease related to a sore throat causes toxic appearance. If there are other intervening mechanisms beside pain and fever, then these mechanisms can be represented in the network by either adding new nodes or by adding a direct link between disease and toxic appearance that summarizes the effect of these mechanisms, while keeping them implicit. There are other assertions of independence implied by this graph. All of these assertions can be read directly from the graph (see Section 2.3).

The network and the probability distribution that result from this judgemental process provide a model of a domain as conceived by an expert. Philosophical justifications for the use of probabilities to represent expert's judgments are given in [7,24,34]. In addition, Bayesian networks can be constructed directly from data or from a combination of expert knowledge and data [3,5,6,15,25,33,37].

2.2. *Notations and basic definitions*

Before we state the definition of Bayesian networks, we provide some notational conventions. Let $\{u_1, \ldots, u_n\}$ be a finite set of variables each having a finite set of values, P be a probability distribution having the Cartesian product of these sets of values as its sample space, and $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n$ be arbitrary values for $u_1, \ldots, u_n$, respectively. Throughout this article, we often say that P is a probability distribution over U, keeping the remaining details implicit.

We use capital letters from the end of the alphabet (e.g., X, Y, Z) to denote sets of variables and the respective bold characters to denote their values. For example, if $X = \{u_1, u_2\}$, then $\boldsymbol{X} = \{\boldsymbol{u}_1, \boldsymbol{u}_2\}$ is a value of X. We use the notation $P(\boldsymbol{X} \mid \boldsymbol{Y})$ to denote the conditional probability $P(X = \boldsymbol{X} \mid Y = \boldsymbol{Y})$, and $P(X \mid Y) = P(X)$ to denote $P(\boldsymbol{X} \mid \boldsymbol{Y}) = P(\boldsymbol{X})$ for all values of X and Y such that $P(\boldsymbol{Y}) \neq 0$. Also, we use $P(X)$ to denote $P(X \mid \emptyset)$.

Definition. Let P be a probability distribution over U. A directed acyclic graph D (i.e., a directed graph with no directed cycles) is a *Bayesian network of P*, if D is constructed from P by the following steps. Assign a *construction order* $u_1, u_2, \ldots, u_n$ to the variables of U and designate a node u_i for each variable u_i.[3] For each variable u_i in U, identify a set $\pi(u_i) \subseteq \{u_1, \ldots, u_{i-1}\}$ such that

$$P(u_i \mid \pi(u_i)) = P(u_i \mid u_1, \ldots, u_{i-1}) \tag{4}$$

(for all values of $u_1, \ldots, u_i$). Assign a direct link from every node in $\pi(u_i)$ to node u_i. This network is *minimal* if for each $u_i \in U$, no proper subset of $\pi(u_i)$ satisfies Eq. (4).

For example, if $u_1, \ldots, u_5$ is a construction order, and if P satisfies, $P(u_3 \mid u_1) = P(u_3 \mid u_1, u_2)$, $P(u_4 \mid u_2, u_3) = P(u_4 \mid u_1, u_2, u_3)$, and $P(u_5 \mid u_4) = P(u_5 \mid u_1, u_2, u_3, u_4)$, then the network of Fig. 2 is a Bayesian network of P.

[3] We deliberately denote with u_i the node that corresponds to variable u_i. It will be immaterial or clear from the context whether we talk about a variable or its associated node.

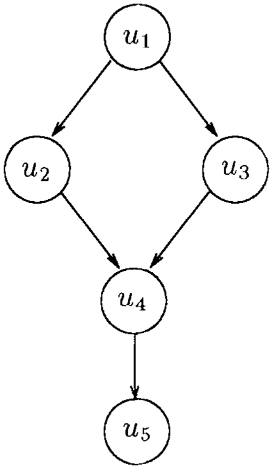

Fig. 2. An abstract example of a Bayesian network.

The number of parameters that a Bayesian network requires and the complexity of its topology depends on the construction order, which is not dictated by its definition. In practice, cause-and-effect and time-order relationships often suggest construction orders that yield simple networks.

Definition. Let P be a probability distribution over U. Let X, Y, and Z be three disjoint subsets of U, and $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{Z}$ be arbitrary respective values. Then $\mathbf{X}$ is *conditionally independent* of $\mathbf{Y}$ given $\mathbf{Z}$, denoted by $I(\mathbf{X},\mathbf{Y} \mid \mathbf{Z})$, if

$$P(\mathbf{X} \mid \mathbf{Z},\mathbf{Y}) = P(\mathbf{X} \mid \mathbf{Z}) \quad \text{or} \quad P(\mathbf{Z}) = 0.$$

$I(\mathbf{X},\mathbf{Y} \mid \mathbf{Z})$ is called an *independence assertion.*

Definition. A set X is *conditionally independent* of Y given Z, denoted by $I(X,Y \mid Z)$, if $I(\mathbf{X},\mathbf{Y} \mid \mathbf{Z})$ holds for every respective value of X, Y and Z. $I(X,Y \mid Z)$ is called a *symmetric* independence assertion.

Using the above notation, the Bayesian network depicted in Fig. 2 satisfies $I(u_3,u_2 \mid u_1)$, $I(u_4,u_1 \mid \{u_2,u_3\})$, and $I(u_5,\{u_1,u_2,u_3\} \mid u_4)$. (For simplicity, we use u_i to denote the singleton $\{u_i\}$.)

When $I(\mathbf{X},\mathbf{Y} \mid \mathbf{Z})$ holds for some but not all the values of the variables involved, then this independence assertion is called *asymmetric*. This term is adapted from the literature on decision analysis, where asymmetric independence corresponds to asymmetries in decision trees [19]. Asymmetric independence assertions are not represented in the topology of Bayesian networks, whereas the representations described in this paper do explicitly encode such assertions. As we will see, this difference makes our new representations better suited for the representation and solving of diagnostic problems.

2.3. Semantics of Bayesian networks

The criteria of d-separation, defined below, characterizes all independence assertions implied by the topology of a Bayesian network.

Definition. The *underlying graph* of a Bayesian network is an undirected graph obtained from the network by replacing every link with an undirected edge.

Definition. A *trail* in a Bayesian network is a sequence of links that form a cycle-free path in the underlying graph.

Definition (*Pearl* [32]). A node b is called a *head-to-head* node w.r.t. (with respect to) a trail t if there are two consecutive links $a \rightarrow b$ and $b \leftarrow c$ on t.

For example, $u_2 \rightarrow u_4 \leftarrow u_3$ is a trail in Fig. 2 and u_4 is a head-to-head node with respect to this trail.

Definition (*Pearl* [32]). A trail t is *active* w.r.t. a set of nodes Z if (1) every head-to-head node w.r.t. t either is in Z or has a descendant in Z, and (2) every other node along t is outside Z. Otherwise, the trail is said to be *blocked* (or *d-separated*) by Z.

In Fig. 2, for example, both trails between $\{u_2\}$ and $\{u_3\}$ are d-separated by $Z = \{u_1\}$; the trail $u_2 \leftarrow u_1 \rightarrow u_3$ is d-separated by Z because node u_1, which is not a head-to-head node w.r.t. this trail, is in Z whereas the trail $u_2 \rightarrow u_4 \leftarrow u_3$ is d-separated by Z, because node u_4 and its descendant u_5 are outside Z. The trail, $u_2 \rightarrow u_4 \leftarrow u_3$, however, is not d-separated by $Z' = \{u_1, u_5\}$, because u_5 is in Z'.

Theorem 1. *Let D be a Bayesian network of a probability distribution P over U and let X, Y, and Z be three disjoint subsets of U. Then:*

Soundness (Verma and Pearl [38]): *If all trails between a node in X and a node in Y are d-separated by Z, then X and Y are conditionally independent given Z in P.*

Completeness (Geiger and Pearl [11]): *The criterion above lists all independence assertions holding in P that can be identified from the topology of D.*

This theorem is extremely useful. It implies, for example, that each variable is independent of all its non-descendants, conditioned on its parents, because the parents of each node d-separates all trails between a node and its non-descendants.[4] It also implies that each variable is independent, given its parents, children, and its children's parents, of all other variables in the network [31]. Such independence assertions are the cornerstone of efficient computations. A generalization of this theorem is given in [12].

[4] This observation was first made by Howard and Matheson [19], and then proven by Olmsted [29].

2.4. *Bayesian networks and inference*

Several algorithms exist to compute posterior distributions given that the values of some variables are observed. These algorithms are collectively called *inference algorithms* and they all rely on independence relationships encoded in the network. For example, each of these inference algorithms can compute the posterior distribution for sore-throat diseases given that glands swollen and high fever are observed, based on the prior distribution, and the independence relationships encoded in Fig. 1.

Pearl [30] and Kim and Pearl [23] developed inference algorithms for networks in which every two nodes are connected with at most one trail. Pearl [31] extended the algorithm to general networks.

Another inference algorithm is that of Lauritzen and Spiegelhalter [26], which initially compiles a given network into a *clique tree.* Each node in a clique tree represents a cluster of variables that are collapsed into a single variable whose domain is the Cartesian product of its constituents. The algorithm minimizes as much as possible the size of the largest cluster. Computations of posterior probabilities are done using Kim and Pearl's algorithm on the clique tree. Improvements to this algorithm are described in [21,22].

Shachter [35] developed an inference algorithm based on two types of transformations: *node removal* and *arc reversal.* Node removal is the elimination of a node that has no descendants from the network. This operation corresponds to summing over its possible values. Arc reversal refers to the reversal of a particular arc after the addition of other arcs. The parameters in the transformed network are computed via a simple closed-form formula based on Bayes rule. Both transformations preserve the joint distribution over the remaining variables and can therefore be applied repeatedly for inference [29].

The time complexity of the above three algorithms is exponential, a fact that is not surprising since inference in Bayesian networks is NP-hard [4]. Nevertheless, these algorithms are efficient enough for many real-world applications [1,2,16].

3. Bayesian multinets

3.1. *Definition and representational advantages*

Although Bayesian networks considerably extend traditional approaches, they are still not expressive enough to encode every piece of information that may reduce computations. The following example demonstrates the inadequacy of Bayesian networks for representing asymmetric independence:

> A guard of a secured building expects three types of persons to approach the building's entrance: workers in the building, approved visitors, and spies. As a person approaches the building, the guard can note its gender and whether or not the person wears a badge. Spies are mostly men. Spies always wear badges in an attempt to fool the guard. Visitors don't wear badges because they don't have one. Female workers tend to wear badges more often than do male workers. The task of the guard is to identify the type of person approaching the building.

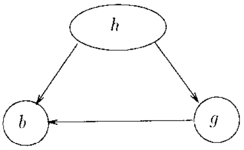

Fig. 3. A Bayesian network for the secured-building example.

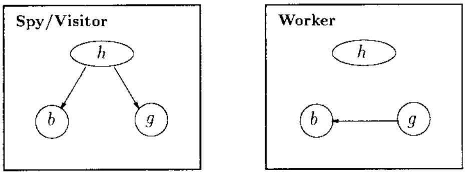

Fig. 4. A Bayesian multinet representation of the secured-building story.

A Bayesian network that represents this story is shown in Fig. 3. Variable h in the figure represents the correct identification. It has three values *worker*, *visitor*, and *spy*. Variables g and b are binary variables representing, respectively, the person's gender and whether or not the person wears a badge. The links from h to g and from h to b reflect the fact that both gender and badge worn are clues for correct identification. The link from g to b encodes the relationship between gender and badge worn.

Unfortunately, the topology of this network hides the fact that gender and badge worn are conditionally independent, given that the person is a spy or a visitor (this assertion holds because, independent of gender, spies always wear badges, and visitors never do). The link between g and b is drawn only because gender and badge worn are related variables when the person is a worker.

We can represent the independence assertions in this story more explicitly using the two Bayesian networks shown in Fig. 4. The first network represents the cases where the person approaching the entrance is either a spy or a visitor. In these two cases, badge worn depends only on the type of person approaching, and not on his or her gender. Consequently, nodes b and g are shown to be conditionally independent (node h blocks the trail between them). The links from h to b and from h to g in this network reflect the fact that badges and gender are relevant clues that help to discriminate spies from visitors. The second network represents the situation where the person is a worker, in which case gender and badge worn are related.

Fig. 4 is a better representation of our story than is Fig. 3, because it shows the dependence of badge worn on gender only in the context in which such a relationship exists—namely, for workers. Moreover, the former representation requires 11 probabilities whereas the representation of Fig. 4 requires only 9 probabilities. This gain, due to the explicit representation of asymmetric independence, can be substantially larger for real-world problems, because the number of probabilities needed can grow exponentially in the number of variables, whereas the overhead of representing multiple networks grows linearly in the number of variables.

We call the representation scheme of Fig. 4 a *Bayesian multinet*. In the remainder of this paper, we often refer to h as the *hypothesis variable*; and we refer to the values of h as *hypotheses*. Furthermore, the variable h will be the focus of construction for Bayesian multinets and for similarity networks, and thus sometimes we call h the distinguished variable. We refer to other variables in a given domain as non-distinguished variables.

Let A_i be a subset of the values of h and let the event $[\![A_i]\!]$ stand for "one of the hypotheses in A_i holds true".

Definition. Let $P(h, u_1, \ldots, u_n)$ be a probability distribution and $A_1, \ldots, A_k$ be non-empty mutually disjoint sets whose union is equal to the set of all values of h (i.e., a partition of the domain of h). A directed acyclic graph D_i is called a *comprehensive local network of P associated with A_i* if D_i is a Bayesian network of $P(h, u_1, \ldots, u_n \mid [\![A_i]\!])$. The set of k local networks is called a *Bayesian multinet* of P. When each A_i is a singleton, we say the Bayesian multinet is *hypothesis-specific*.

In the secured-building example of Fig. 4, $\{\{spy,\ visitor\},\ \{worker\}\}$ is a partition of the values of the hypothesis node h, one local network is a Bayesian network of $P(h, b, g \mid worker)$, and the other local network is a Bayesian network of $P(h, b, g \mid [\![spy, visitor]\!])$ where the conditioning event $[\![spy, visitor]\!]$ is a short-hand notation for saying that either $h = spy$ or $h = visitor$.

The fundamental concept associated with Bayesian multinets is that of *conditioning*; each local network represents a distinct situation where hypotheses are restricted to a specified subset. As a result of such conditioning, asymmetric independence assertions are encoded in the topology of the local networks. Consequently, savings in computations and memory requirements result. In our example, conditional independence between gender and badge worn is encoded as a result of conditioning on h.

Conditioning may also destroy independence relationships rather then create them [32]. Nonetheless, if the distinguished variable is a root node (i.e., a node with no incoming links), then, according to d-separation, conditioning on its values never decreases and often increases the number of independence relationships. We address situations where the hypothesis variable is not a root node and where more than one node represents hypotheses in Sections 3.3 and 5, respectively.

The relationship between gender and badge worn is an example of *hypothesis-specific* independence, wherein two variables are independent given some hypotheses ($\{spies,\ visitors\}$) but dependent given others (*workers*). In general, a hypothesis-specific independence assertion is represented in a Bayesian multinet whenever a link between two non-distinguished variables exists in some local networks but does not exist in other local networks.

The following variation of the secured-building example demonstrates an additional type of asymmetric independence that can be represented by Bayesian multinets.

> The guard of the secured building now expects *four* types of persons to approach the building's entrance: executives, regular workers, approved visitors, and spies. The guard can note gender, whether or not the person is wearing a badge, and whether or not the person arrives in a limousine (l). We assume that only executives arrive

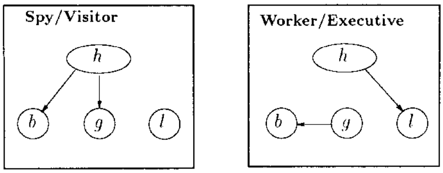

Fig. 5. A Bayesian multinet representation of the augmented secured-building story.

in limousines and that male and female executives wear badges just as do regular workers (to serve as role models).

This story is represented by the two local networks shown in Fig. 5. One network represents a situation where either a spy or a visitor approaches the building, and the other network represents a situation where either a worker or an executive approaches the building. The link from h to l in the latter network reflects the fact that arrival in a limousine is a relevant clue that helps to discriminate workers from executives. The absence of this link in the former network reflects the fact that arrival in a limousine does not help to discriminate spies from visitors.

The relationship between arrival in limousines and the hypothesis variable h is an example of *subset* independence, wherein a non-distinguished variable (l) is independent of h given h draws its values from a subset of hypotheses $\{$*spy*, *visitor*$\}$. In general, a subset independence assertion is represented in a Bayesian multinet whenever a link between the hypothesis node and a non-distinguished variable exists in some local networks but not in all. The relationship between the set of variables $\{$badge worn, gender$\}$ and h, when h is restricted to $\{$*worker*, *executive*$\}$ is another example of subset independence.[5]

The next theorem demonstrates that a Bayesian multinet always encodes the conditional distribution $P(u_1,\ldots,u_n \mid h)$. Its proof provides us with an inference algorithm.

Theorem 2. *Let M be a Bayesian multinet of $P(h,u_1,\ldots,u_n)$ based on a partition $A_1,\ldots,A_k$ of the values of h. Then, the distribution $P(u_1,\ldots,u_n \mid h)$ can be computed from the parameters encoded in M.*

Proof. The distribution $P(u_1,\ldots,u_n,h \mid [\![A_i]\!])$ is encoded in the local network associated with A_i because according to the definition of a local network,

$$P(u_1,\ldots,u_n,h \mid [\![A_i]\!]) = \prod_{v} P(v \mid \pi(v), [\![A_i]\!]) \qquad (5)$$

where v is either h or some u_i and $\pi(v)$ are v's parents.

Let $\boldsymbol{h}$ be an hypothesis in A_i. The distribution $P(u_1,\ldots,u_n \mid \boldsymbol{h}, [\![A_i]\!])$ is computed from $P(u_1,\ldots,u_n,h \mid [\![A_i]\!])$. Moreover, the distribution $P(u_1,\ldots,u_n \mid \boldsymbol{h}, [\![A_i]\!])$ is equal to $P(u_1,\ldots,u_n \mid \boldsymbol{h})$ because the assertion "h is assigned the value $\boldsymbol{h}$" logically implies

[5] Heckerman [14] coined the terms subset independence and hypothesis-specific independence. A hypothesis-specific Bayesian multinet is similar to hypothesis-specific similarity network defined in [14] except that it contains all the variables in U.

the assertion "h draws its value from A_i" whenever A_i includes $\boldsymbol{h}$. Thus $P(u_1, \ldots, u_n \mid h)$ can be computed from the parameters of M. □

The parameters encoded in a Bayesian multinet can be used to compute the relative posterior probability between every pair of hypotheses within each A_i. In order to compute the absolute value of the posterior probability of each hypothesis, however, one must have information about the prior distribution $P(h)$ in addition to the Bayesian multinet because $P(h)$ cannot be computed from the parameters encoded in the local networks.

3.2. *Bayesian multinets and inference*

The proof of Theorem 2 and the comment that follows suggest an inference algorithm for computing the posterior distribution of h from a Bayesian multinet of P and from the prior distribution of h. The inference algorithm uses a procedure called *INFER* which has two parameters, one specifying a query of the form "compute $P(X \mid \boldsymbol{Y})$" and the second is a Bayesian network where X and Y are sets of variables that appear in the network and $\boldsymbol{Y}$ is a value of Y. As we have discussed in Section 2.4, there are many ways to realize *INFER* and we do not need to specify *INFER*'s operational details in order to demonstrate how this procedure is extended to operate on Bayesian multinets. The inference algorithm is described below.

Algorithm (*Bayesian multinet inference*).

Input:

- A Bayesian multinet of $P(h, u_1, \ldots, u_n)$ based on a partition $A_1, \ldots, A_k$ of h's values. The local network associated with A_i is denoted by D_i.
- A priori probability distribution $P(h)$.
- Instances $\boldsymbol{u}'_1, \ldots, \boldsymbol{u}'_m$ for a set of variables $\{u'_1, \ldots, u'_m\} \subseteq \{u_1, \ldots, u_n\}$.

Output: The posterior probability distribution $P(h \mid \boldsymbol{u}'_1, \ldots, \boldsymbol{u}'_m)$.

```
1  For each partition element $A_j$
2    For each hypothesis $\boldsymbol{h}_i \in A_j$
3      $\alpha_{i,j} = INFER(\, P(\boldsymbol{u}'_1, \ldots, \boldsymbol{u}'_m \mid \boldsymbol{h}_i, [\![A_j]\!]),\ D_j)$
4  For each $\boldsymbol{h}_i$
5    Compute $P(\boldsymbol{h}_i \mid \boldsymbol{u}'_1, \ldots, \boldsymbol{u}'_m) = P(\boldsymbol{h}_i) \cdot \alpha_{i,j} / (\sum_i P(\boldsymbol{h}_i) \cdot \alpha_{i,j})$
```

Line 3 is the normal computation performed by an inference algorithm for Bayesian networks. Lines 4 and 5 encode Bayes rule together with the fact that the distribution $P(u_1, \ldots, u_n \mid \boldsymbol{h}_i, [\![A_j]\!])$ is equal to $P(u_1, \ldots, u_n \mid \boldsymbol{h}_i)$ which is computed on line 3 and assigned to $\alpha_{i,j}$. This equality follows from the fact that $\boldsymbol{h}_i$ implies $[\![A_j]\!]$ whenever $\boldsymbol{h}_i \in A_j$.

The advantage of computing $P(u_1, \ldots, u_n \mid \boldsymbol{h})$ via this algorithm versus using *INFER* on a Bayesian network of P arises from the fact that independence assertions are represented in some local networks, but not in the Bayesian network. For example, suppose the guard of our secured-building problem sees a person wearing a badge ($\boldsymbol{b}$) approach the building but does not notice the person's gender. Using the Bayesian network of

Fig. 3, *INFER* computes the posterior probability of each possible identification (*worker*, *visitor*, *spy*) as follows:

$$P(h \mid \boldsymbol{b}) = k \cdot P(h) \cdot \sum_{g} P(g \mid h) \cdot P(\boldsymbol{b} \mid g, h) \tag{6}$$

where k is the normalizing constant that makes $P(h \mid \boldsymbol{b})$ sum to unity. Since the Bayesian network representing this problem does not encode any statement of conditional independence the above computation is done by any realization of *INFER*.

Alternatively, our inference algorithm computes the posterior probability of each hypothesis more efficiently, using the Bayesian multinet of Fig. 4, as follows:

$$P(spy \mid g, \boldsymbol{b}) = k \cdot P(spy) \cdot P(\boldsymbol{b} \mid spy), \tag{7}$$

$$P(visitor \mid g, \boldsymbol{b}) = k \cdot P(visitor) \cdot P(\boldsymbol{b} \mid visitor), \tag{8}$$

$$P(worker \mid g, \boldsymbol{b}) = k \cdot P(worker) \cdot \sum_{g} P(g \mid worker) \cdot P(\boldsymbol{b} \mid g, worker). \tag{9}$$

Eqs. (7) and (8) take advantage of hypothesis-specific independence. In particular, the two equations incorporate the fact that g and b are conditionally independent given $h = spy$ and $h = visitor$, respectively. Thus, we do not have to sum over the variable gender as we do when using a Bayesian network (Eq. (6)). These savings are achieved by the inference algorithm for Bayesian multinets because the computations of line 3 are done on the local network that encodes this independence information. If we were to use the same inference algorithm used by line 3 on the Bayesian network of Fig. 3, where this independence assertion is not displayed, then the more costly computation done by Eq. (6) would have been performed.

3.3. Nonroot hypothesis variables

The multinet approach described thus far is especially beneficial when the hypothesis variable can be modeled as a root node because, then, no new links are ever introduced by conditioning on the different hypotheses. Nonetheless, there are situations where modeling the hypothesis node as a root node is awkward. For example, in the secured-building story, suppose there are two independent reports indicating possible spying—say, for military and economical reasons. Such a priori factors for correct identification are best modeled as parent nodes of h, called—say—*economics* and *military*. The resulting subnetwork among these variables is $economics \rightarrow h \leftarrow military$, which represents the reasonable assertion that *economics* and *military* are marginally independent.

When h assumes the value *spy*, however, an induced dependency is introduced between its parents *economics* and *military*; For example, a military explanation for a confirmed spy makes less likely an economical explanation, because the former explains the presence of the spy. Consequently, a link must be drawn between the *economics* and *military* nodes in the local network for spies versus visitors. This link would not appear in the full Bayesian network because *economics* and *military* can reasonably be assumed independent. They only become dependent when conditioning on $h = spy$. The

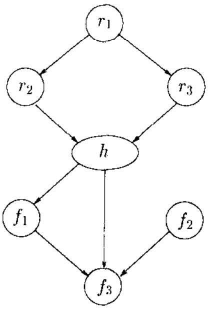

Fig. 6. A Bayesian network where all trails between a priori factors r_i and evidential clues f_i pass through h.

probability distributions associated with such induced links are difficult to assess (e.g., $P(economics \mid h, military)$. Thus, in this example, constructing a local network is harder than constructing the full network.

One approach to handle this problem is as follows. First, construct a Bayesian network that represents only a priori factors that influence the hypotheses, ignoring any evidential variables (such as gender, badge worn, and arrival in limousine). In our example, this network would be *economics* → h ← *military*. Then, use this network to revise the a priori probabilities of the different hypotheses. Finally, construct local networks ignoring a priori factors (as is done in Fig. 4) and use the resulting multinet with the revised priors of h to compute the posterior probability of h as determined by the evidential clues. This decomposition technique works best if a priori factors are independent of all evidential clues conditioned on the different hypotheses. That is, in situations that can be modeled with Bayesian networks of the form shown in Fig. 6, where all trails between a priori factors r_i and evidential clues f_i pass through h.

When a network of this form cannot serve as a justifiable model, another approach can be used instead. First, compose a Bayesian multinet ignoring a priori factors, construct a Bayesian network from the local networks by taking the union of all their links (e.g., the union of all links in Fig. 4 yields the Bayesian network of Fig. 1). Then, add a priori factors to the resulting network. This approach is described in [14]. The disadvantage of this method is that in the process of generating a Bayesian network from a multinet, one encodes asymmetric independence in the parameters rather than in the topology of the Bayesian network. Consequently, these asymmetric assertions cannot speed up the computations of known inference algorithms. Nevertheless, this approach is still the best alternative for decomposing the construction of large Bayesian networks having topologies more complex than that of Fig. 6.

4. Similarity networks

4.1. Definition and representational advantages

In Bayesian multinets, we required that every variable be included in each local network. This requirement stands in contrast to the observation that in many domains each measurement often helps to discriminate only a specialized class of hypotheses. Symptoms are often related to narrow classes of diseases, and systems' faults often isolate a specific class of potential malfunctions. Assessing the dependence between two variables under assumptions unrelated to their semantics can present an insurmountable burden on the model builder. This difficulty was realized during the construction of an expert system for surgical pathology diagnosis [14]. When the expert pathologist was asked by the model builder: Given a particular disease, does observing symptom x change your belief that you will observe symptom y? The pathologist would sometimes reply:

> I've never thought about these two symptoms at the same time before. Symptom x is relevant to only one set of diseases, while symptom y is only relevant to another set of diseases. These sets of diseases do not overlap, and I never confuse the first set of diseases with the second.

An erroneous solution to this difficulty is to include in each local network of a Bayesian multinet only those variables that help to discriminate among the hypotheses covered by that local network. In doing so, however, valuable information for correct identification may be lost.

For example in the secured-building problem gender (g) and badge worn (b) do not help to discriminate workers from executives. If these variables would not have been depicted in the local network for $\{$*worker,executive*$\}$ in the Bayesian multinet of Fig. 5 then this multinet would have failed to represent the genuine relationship between badge worn and gender.

As we will see, a correct solution to this difficulty is indeed to include in each local network only those variables that help to discriminate among the hypotheses covered by that local network, but also to construct additional local networks to compensate for lost information. The structure that results is called a *similarity network* [14]. For example, the secured-building problem can be represented by a similarity network shown in Fig. 7. Whereas the Bayesian multinet of Fig. 5 contains two local networks, the similarity network contains three local networks: one local network helps to discriminate spies from visitors, another local network helps to discriminate visitors from workers, and a third local network helps to discriminate workers from executives. In each local network, we include only those variables that help to discriminate among the hypotheses covered by that local network. For example, in Fig. 7, the dependence between badge worn and gender is not included in the local network for workers versus executives. This dependence, however, is included in the local networks for visitors versus workers, because badge worn helps to discriminate between these two hypotheses.

The main advantage of similarity networks, from the perspective of knowledge acquisition, is that a domain expert who provides the parameters of the network is not required to quantify the dependence between variables that are not related to the hypotheses under consideration. In order not to loose information needed for correct diagnosis we will see that the local networks must be based on a *connected cover* of hypotheses.

Definition. A *cover* of a set of hypotheses H is a collection $\{A_1, \ldots, A_k\}$ of nonempty subsets of H whose union is H. Each cover is a hypergraph, called a *similarity hyper-*

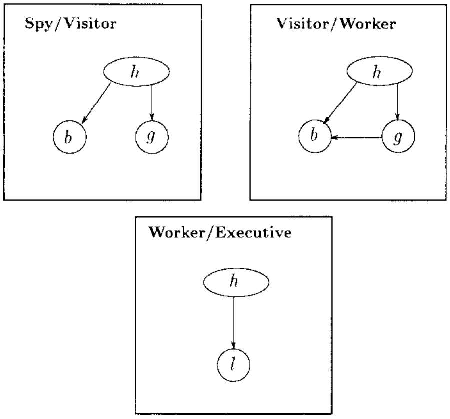

Fig. 7. A similarity network representation of the secured-building story.

graph, where the A_i are hyperedges and the hypotheses are nodes. A cover is *connected* if the similarity hypergraph is connected.

In Fig. 7, {*spy, visitor*}, {*visitor, worker*}, {*worker, executive*} is a cover of the hypotheses set. This cover is connected because it consists of the three links *spy–visitor–worker–executive* which form a connected hypergraph (as well as a connected graph). The set {*spy, visitor*}, {*worker, executive*} is also a cover but it is not connected. The set {*worker, executive, visitor*}, {*visitor, spy*} is an example of a connected cover that is a hypergraph but not a graph.

Definition. Let $P(h, u_1, \ldots, u_n)$ be a probability distribution and $A_1, \ldots, A_k$ be a connected cover of the values of h. A directed acyclic graph D_i is called a *local network of P associated with A_i* if D_i is a Bayesian network of $P(h, v_1, \ldots, v_m \mid [\![A_i]\!])$ where $\{v_1, \ldots, v_m\}$ is the set of all variables in $\{u_1, \ldots, u_n\}$ that "help to discriminate" the hypotheses in A_i. The set of k local networks is called a *similarity network* of P.

We define "help to discriminate" formally in the next section.

The definition of similarity networks does not specify how to select a connected cover of hypotheses. Although any selection of a connected cover yields a valid similarity network, some selections yield similarity networks that display more asymmetric independence assertions than do other selections. An analogous situation exists when constructing a Bayesian network where some construction orders yield Bayesian networks that display more symmetric independence assertions than do other Bayesian networks. The practical solution for constructing a Bayesian network is to choose a construction order according to cause-effect relationships.[6] This selection tends to maximize the information about symmetric independence encoded in the resulting network.

[6] Bayesian networks are often called *causal networks*.

The practical solution for constructing the similarity hypergraph is to choose a connected cover by grouping together hypotheses that are "similar" to each other by some accessible criteria (e.g., spies and visitors are outsiders). This choice tends to maximize the number of asymmetric independence assertions encoded in a similarity network. Hence the name for this representation.

Similarity networks have another important advantage not mentioned so far: This representation helps to prevent the model builder from omitting relevant information. For example, suppose workers and executives often arrive at work with a smile, whereas spies and visitors often arrive at work without a smile. This information is likely to be forgotten when constructing the local networks for spies versus visitors and for visitors versus executives because it does not help to discriminate between these pairs of hypotheses. When constructing the similarity network of Fig. 7, however, the builder of the network is likely to recall the information about smile because he must compose a local network for discriminating visitors from workers—a task for which this information is important. In general, whenever the cover is connected, every variable that is useful for discriminating some pair of hypotheses will appear in at least one local network. This property of similarity network was called *exhaustiveness* by Heckerman [14].

4.2. Relevance relations

The definition of similarity networks is not complete without attributing a precise meaning to the utterance "helps to discriminate" used in the definition of a local network. Below we give several possibilities.

Definition. Let $P(u_1, \ldots, u_n \mid e)$ be a probability distribution where e is a fixed event. Variables u_i and u_j are *unrelated given* e if u_i and u_j are disconnected in every minimal Bayesian network of $P(u_1, \ldots, u_n \mid e)$. Otherwise, u_i and u_j are *related given* e, denoted *related*$(u_i, u_j \mid e)$.

This definition states that two variables u_i an u_j are unrelated given e if there exists no trail connecting them, i.e., there exists no sequence of variables $u_i, \ldots, u_j$ such that every two consecutive variables in this sequence are connected with a link. The requirement that u_i and u_j be disconnected in *every* minimal network is not as strong as it may seem because if u_i and u_j are disconnected in one minimal Bayesian network of P then u_i and u_j are disconnected in every minimal Bayesian network of P [9]. Furthermore, one can phrase the definition of relatedness as follows.

Theorem 3 (Geiger and Heckerman [9]). *Let $P(u_1, \ldots, u_n \mid e)$ be a probability distribution over $U = \{u_1, \ldots, u_n\}$ and e be a fixed event. Then, u_i and u_j are unrelated given e iff there exist a partition U_1, U_2 of U such that $u_i \in U_1$, $u_j \in U_2$, and $P(U_1, U_2 \mid$* e) $= P(U_1 \mid e) P(U_2 \mid e)$.[7]

[7] In [9], $P(U \mid e)$ is replaced with $P(U)$. Since e is a fixed event, this shift of notation does not alter this theorem's proof.

An immediate consequence of this theorem is that the relation *related* is transitive, namely, for every three variables u_i, u_j and u_k,

$$(\mathit{related}(u_i, u_j \mid e) \text{ and } \mathit{related}(u_j, u_k \mid e)) \Rightarrow \mathit{related}(u_i, u_k \mid e). \tag{10}$$

The second definition is the more appealing one from a knowledge engineering point of view. It states that two variables u_i and u_j are unrelated if, in any context, knowing the value of one variable does not change the knowledge about the values of the other.

Definition. Let $P(u_1, \ldots, u_n \mid e)$ be a probability distribution where e is a fixed event. Variables u_i and u_j are *pairwise irrelevant* given e if

$$P(u_i \mid u_j, v_1 \in V_1, \ldots, v_m \in V_m, e) = P(u_i \mid v_1 \in V_1, \ldots, v_m \in V_m, e)$$

or

$$P(v_1 \in V_1, \ldots, v_m \in V_m \mid e) = 0$$

for every subset of values $V_1, \ldots, V_m$ for $v_1, \ldots, v_m$, respectively, where $\{v_1, \ldots, v_m\}$ is an arbitrary subset of $\{u_1, \ldots, u_n\} \setminus \{u_i, u_j\}$.

This definition states that u_i and u_j are pairwise irrelevant if they are independent given any possible context. That is, if u_i and u_j are independent given that the value of each v_i $(i = 1, \ldots, m)$ is taken from a subset V_i of v_i's domain and that u_i and u_j remain independent under each selection of variables $v_1, \ldots, v_m$ and for each restriction of their values.

When $u_1, \ldots, u_n$ are all binary variables, then the sets V_i are singletons, namely, a single specific assignment for v_i. In [9], we have shown that pairwise irrelevance and unrelatedness are equivalent when all variables are binary and when the distribution P is strictly positive. We conjecture that the equivalence between pairwise irrelevance and relatedness holds even when these restrictions are lifted.

In the remainder of this paper we use the concept of relatedness for defining which variables are excluded from a local network. We do so by distinguishing one variable as the hypothesis variable (call it h) and defining the event e to be $[\![A_i]\!]$, namely, a disjunction over a subset of the values of h. A variable x is then excluded from a local network for A_i iff x and h are unrelated given $[\![A_i]\!]$.

4.3. Inference using similarity networks

Similarity networks are designed for the task of diagnosis or discrimination. In particular, they are designed to compute the posterior probability of each possible hypothesis given a set of observations. In this section, we show that under reasonable assumptions, the computation of the posterior probability of each hypothesis can be done in each local network and then be combined coherently according to the axioms of probability theory. We analyze the complexity of our algorithm demonstrating its superiority over inference algorithms that operate on Bayesian networks.

We assume that any instantiation of the variables in a similarity network of P has a nonzero probability to occur. Such a probability distribution is said to be *strictly positive.*

This assumption is reasonable for some domains of medical diagnosis, where given an arbitrary collection of clinical findings, the existence of each disease retains a nonzero probability. Subject to this assumption, we develop an inference algorithm that operates directly on similarity networks. We will remove this assumption later at the cost of higher complexity.

The inference problem at hand can be stated as follows: Given a similarity network of $P(h, u_1, \ldots, u_n)$ that is based on a partition $\mathcal{A} = \{A_1, \ldots, A_k\}$ of the values of h, and given a set of assignments $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m$ for a set $v_1, \ldots, v_m$ of variables that is a subset of $\{u_1, \ldots, u_n\}$ compute $P(\boldsymbol{h}_j \mid \boldsymbol{v}_1, \ldots, \boldsymbol{v}_m)$—the posterior probability of $\boldsymbol{h}_j$—for every $\boldsymbol{h}_j$.

In order to compute the posterior probability of each $\boldsymbol{h}_j$ we use the procedure *INFER*. As in Section 3.2, this procedure has two parameters, one specifying a query of the form "compute $P(X \mid \boldsymbol{Y})$" and the second is a Bayesian network where X and Y are sets of variables that appear in the network and $\boldsymbol{Y}$ is a value of Y. As before we do not need to specify *INFER*'s operational details in order to demonstrate how this procedure is extended to operate on similarity networks. The new inference algorithm is described below.

First, for each $\boldsymbol{h}_i$ we identify a set of hypotheses $A_j \in \mathcal{A}$ to which $\boldsymbol{h}_i$ belongs and compute the posterior probability of hypothesis $\boldsymbol{h}_i$ under the additional assumption that one of the hypotheses in A_j holds true. In other words, we compute $P(\boldsymbol{h}_i \mid \boldsymbol{v}_1, \ldots, \boldsymbol{v}_m, [\![A_j]\!])$. Second, we compute the posterior probabilities $P(\boldsymbol{h}_j \mid \boldsymbol{v}_1, \ldots, \boldsymbol{v}_m)$ from the probabilities $P(\boldsymbol{h}_j \mid \boldsymbol{v}_1, \ldots, \boldsymbol{v}_m, [\![A_i]\!])$, by solving a set of linear equations:

$$P(\boldsymbol{h}_j \mid \boldsymbol{v}_1, \ldots, \boldsymbol{v}_m) = P(\boldsymbol{h}_j \mid \boldsymbol{v}_1, \ldots, \boldsymbol{v}_m, [\![A_i]\!]) \cdot \sum_{\boldsymbol{h}_j \in A_i} P(\boldsymbol{h}_j \mid \boldsymbol{v}_1, \ldots, \boldsymbol{v}_m)$$

that relate these quantities. We will see later that these equations have a unique solution.

It remains to show how to compute the query $P(\boldsymbol{h}_i \mid \boldsymbol{v}_1, \ldots, \boldsymbol{v}_m, [\![A_j]\!])$. It seems that one can merely call the procedure *INFER* to compute this query using the local network D_j which corresponds to A_j. The query $P(\boldsymbol{h}_i \mid \boldsymbol{v}_1, \ldots, \boldsymbol{v}_m, [\![A_j]\!])$, however, may include variables that do not appear in D_j in which case *INFER* is not applicable.

Fortunately, the following equality will be shown to hold:

$$P(\boldsymbol{h}_i \mid \boldsymbol{v}_1, \ldots, \boldsymbol{v}_l, [\![A_j]\!]) = P(\boldsymbol{h}_i \mid \boldsymbol{v}_1, \ldots, \boldsymbol{v}_m, [\![A_j]\!]) \tag{11}$$

where $v_1, \ldots, v_l$ are the variables in $\{v_1, \ldots, v_m\}$ that appear in D_j and $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_l$ are their values. Thus to compute $P(\boldsymbol{h}_i \mid \boldsymbol{v}_1, \ldots, \boldsymbol{v}_m, [\![A_j]\!])$ we use the procedure *INFER* to compute the query $P(\boldsymbol{h}_i \mid \boldsymbol{v}_1, \ldots, \boldsymbol{v}_l, [\![A_j]\!])$ using the network D_j. Eq. (11) tells us that the two computations yield identical answers.

Eq. (11) states that $v_{l+1}, \ldots, v_m$ are conditionally independent of $\boldsymbol{h}_j$ given every value of the variables $v_1, \ldots, v_l$ that appear in D_j where $v_{l+1}, \ldots, v_m$ are the variables in $\{v_1, \ldots, v_m\}$ that do not appear in D_j. If Eq. (11) does not hold, some of the variables in $\{v_{l+1}, \ldots, v_m\}$ would appear in the local network D_j, contrary to our assumption that D_j contains only $v_1, \ldots, v_l$.

This algorithm is summarized below.

Algorithm (*Inference in similarity networks*).

Input: A similarity network of $P(u_1, \ldots, u_n, h)$ based on a connected cover $A_1, \ldots, A_k$ of h's values.

Output: $P(h \mid v_1, \ldots, v_m)$ where $v_1, \ldots, v_m$ are values of variables $v_1, \ldots, v_m$ and $\{v_1, \ldots, v_m\}$ is a subset of $\{u_1, \ldots, u_n\}$.

Notation: D_j denotes the local network that corresponds to A_j and V_j are the variables that appear in D_j.

```
1  For each $A_j$
2    Let $\{v_1, \ldots, v_l\}$ be the variables in $V_j \cap \{v_1, \ldots, v_m\}$
3    For each $\boldsymbol{h}_i \in A_j$
4      $\alpha_{i,j} := INFER(P(\boldsymbol{h}_i \mid v_1, \ldots v_l, [\![A_j]\!]), D_j)$
5      If $\alpha_{i,j} = 0$, then Return "$P$ is not strictly positive"
6  Solve the following set of linear equations:
7    For all $i$ and $j$, $P(\boldsymbol{h}_i \mid v_1, \ldots, v_m) = \alpha_{i,j} \cdot \sum_{\boldsymbol{h}_i \in A_j} P(\boldsymbol{h}_i \mid v_1, \ldots, v_m)$
8    $\sum_i P(\boldsymbol{h}_i \mid v_1, \ldots, v_m) = 1$
9  Return $P(h \mid v_1, \ldots, v_m)$
```

We have argued already that the solution to the equations listed in lines 7 and 8 provide the desired posterior probability. It remains to show that there exists a unique solution. Let us examine a local network D_j that corresponds to A_j. Assume A_j consists of $\boldsymbol{h}_1, \ldots, \boldsymbol{h}_r$. Since $v_1, \ldots, v_m$ remain fixed throughout the computations we denote $P(\boldsymbol{h}_i \mid v_1, \ldots, v_m)$ by $Q(\boldsymbol{h}_i)$. Consider the following equations:

$$Q(\boldsymbol{h}_1) = \alpha_{1,j}\,[Q(\boldsymbol{h}_1) + Q(\boldsymbol{h}_2) + \cdots + Q(\boldsymbol{h}_r)]\,, \tag{12}$$

$$Q(\boldsymbol{h}_2) = \alpha_{2,j}\,[Q(\boldsymbol{h}_1) + Q(\boldsymbol{h}_2) + \cdots + Q(\boldsymbol{h}_r)]\,, \tag{13}$$

$$\vdots$$

$$Q(\boldsymbol{h}_r) = \alpha_{r,j}\,[Q(\boldsymbol{h}_1) + Q(\boldsymbol{h}_2) + \cdots + Q(\boldsymbol{h}_r)]\,. \tag{14}$$

These are the subset of the equations defined in line 7 which correspond to the local network D_j. By dividing every pair of consecutive equations, we obtain the following ratios:

$$\begin{aligned} &Q(\boldsymbol{h}_r) = \frac{\alpha_{r,j}}{\alpha_{r-1,j}} Q(\boldsymbol{h}_{r-1}), \qquad Q(\boldsymbol{h}_{r-1}) = \frac{\alpha_{r-1,j}}{\alpha_{r-2,j}} Q(\boldsymbol{h}_{r-2}), \\ &\ldots, Q(\boldsymbol{h}_2) = \frac{\alpha_{2,j}}{\alpha_{1,j}} Q(\boldsymbol{h}_1). \end{aligned} \tag{15}$$

Hence, the solution of these equations provides the ratios of the posterior probabilities between every pair of hypotheses in A_j. Since we repeat this process for every A_j and since the cover defined by $A_1, \ldots A_k$ is connected, the ratio of every pair of hypotheses is established. To obtain the absolute values of each $Q(\boldsymbol{h}_i)$, it remains to normalize their sum to one, using the equation on line 8 of the algorithm.

Consequently we have proven the following theorem.

Theorem 4. *Let $P(h, u_1, \ldots, u_n)$ be a strictly positive probability distribution and $\mathcal{A} = \{A_1, \ldots, A_k\}$ be a partition of the values of h. Let S be a similarity network based on $\mathcal{A}$. Let $v_1, \ldots, v_m$ be a subset of variables whose value is given. There exists a single solution for the set of equations defined by lines 7 and 8 of the above algorithm and this solution determines uniquely the conditional probability $P(h \mid v_1, \ldots, v_m)$.*

An important observation to make is that the equations on lines 7 and 8 are derived from a given probability distribution $P(h, u_1, \ldots, u_n)$. Consequently, although some equations might be redundant, these equations are always consistent. When the set of local networks is constructed from expert judgments, as done in practice, consistency is not guaranteed. Heckerman [14] describes an algorithm that helps a user to construct a consistent set of local networks by prompting to his attention all probabilities that have already been assigned previously in another local network and verifying with him that these probabilities are acceptable.

It remains to analyze the complexity of this inference algorithm. For simplicity, we assume that all variables are binary in which case the procedure *INFER* has a worst-case complexity of $O(2^n)$. In the worst case, the proposed inference algorithm may not perform more efficiently, because all n variables may appear in each local network. In practice, however, each local network contains a small percentage, say c, of the n variables because all other variables are irrelevant given the context of a specific local network.[8] If $O(n)$ local networks are given, the worst-case complexity of applying *INFER* to these local networks is $O(n \cdot 2^{cn})$, which is smaller than $O(2^n)$ obtained by applying *INFER* on a single Bayesian network generated from these local networks. The complexity of solving the equations on lines 7 and 8 is ignored because it is linear in n. Thus from a worst-case of 2^{100} calculations, for example, we reduce the number of calculations to $100 \cdot 2^{20}$.

4.4. Inferential and diagnostic completeness

An important property of Bayesian networks is that their parameters encode the entire joint distribution through the product rule (Eq. (3)). This property guarantees that any inference task can in principle be computed from the parameters encoded in a Bayesian network. Motivated by this observation we establish the following definition.

Definition. A similarity network S for $P(u_1, \ldots, u_n, h)$ is *inferentially complete* if the distribution $P(u_1, \ldots, u_n, h)$ can be recovered from the parameters of S.

Not all similarity networks are inferentially complete. For example, if $P(u_1, \ldots, u_n, h)$ factors into the product $P(u_1)P(u_2 \ldots, u_n, h)$ then the variable u_1 will not be included in any local network. Therefore, it will be impossible to recover $P(u_1)$ from the parameters encoded in the similarity networks of P. The information about $P(u_1)$ that is lost in the process of producing a similarity network of P, however, is never needed in order to compute the posterior probability of any hypothesis. Evidently, inferential

[8] An approximate number for c in the lymph-node pathology domain is 0.2.

completeness is too strong a requirement for the purpose of computing the posterior probability of each hypothesis.

Definition. A similarity network S for $P(h, u_1, \ldots, u_n)$ is *diagnostically complete* if the conditional distribution $P(h \mid v_1, \ldots, v_m)$ can be recovered from the parameters of S for every subset $\{v_1, \ldots, v_m\}$ of $\{u_1, \ldots, u_n\}$.

In the previous section, we showed that every similarity network of a strictly positive probability distribution P is diagnostically complete (Theorem 4). The inference algorithm we presented shows how to compute $P(h \mid v_1, \ldots, v_m)$ for every value of $v_1, \ldots, v_m$. If P is not strictly positive, then one can construct examples where the equations defined by lines 7 and 8 of our inference algorithm do not determine the probability $P(h \mid v_1, \ldots, v_m)$. Nevertheless, we will prove that, under minor restrictions, every similarity network is diagnostically complete.

Before proving diagnostic completeness we resort to an example where our inference algorithm fails, and examine how the posterior probability can be computed in an alternative way. This computation highlights the general approach. Suppose S is a similarity network for $P(h, y)$ where h has three values $\{\boldsymbol{h}_1, \boldsymbol{h}_2, \boldsymbol{h}_3\}$ having equal a priori probability and suppose that y has two values $+y$, $-y$. Also assume that S is based on the cover $\{\{\boldsymbol{h}_1, \boldsymbol{h}_2\}, \{\boldsymbol{h}_2, \boldsymbol{h}_3\}\}$ and that $P(+y \mid \boldsymbol{h}_2) = 0$.

When we apply our algorithm to compute $P(\boldsymbol{h}_i \mid +y)$, the algorithm generates three equations $P(\boldsymbol{h}_1 \mid +y, [\![\boldsymbol{h}_1, \boldsymbol{h}_2]\!]) = 1$, $P(\boldsymbol{h}_2 \mid +y, [\![\boldsymbol{h}_2, \boldsymbol{h}_3]\!]) = 0$, and $P(\boldsymbol{h}_3 \mid +y, [\![\boldsymbol{h}_2, \boldsymbol{h}_3]\!]) = 1$. From these three equations, we cannot compute the relative magnitude of the posterior probability of $\boldsymbol{h}_1$ versus $\boldsymbol{h}_3$. All three equations merely show that $P(\boldsymbol{h}_2 \mid +y)$ is zero.

Nonetheless, $P(\boldsymbol{h}_i \mid +y)$ can be computed from the parameters that quantify S. These parameters include the following: $P(\boldsymbol{h}_1 \mid \boldsymbol{h}_1 \vee \boldsymbol{h}_2)$, $P(\boldsymbol{h}_2 \mid \boldsymbol{h}_2 \vee \boldsymbol{h}_3)$, $P(\boldsymbol{h}_3 \mid \boldsymbol{h}_2 \vee \boldsymbol{h}_3)$, and $P(+y \mid \boldsymbol{h}_1, \boldsymbol{h}_1 \vee \boldsymbol{h}_2)$, $P(+y \mid \boldsymbol{h}_2, \boldsymbol{h}_1 \vee \boldsymbol{h}_2)$ and $P(+y \mid \boldsymbol{h}_3, \boldsymbol{h}_2 \vee \boldsymbol{h}_3)$. From the first three parameters, $P(\boldsymbol{h}_i)$, $i = 1, \ldots, 3$, can be recovered provided none of the prior probabilities is zero. The restriction that all prior probabilities are nonzero is quite reasonable. If the prior probability of some hypothesis were zero, there would be little reason to include that hypothesis in the model.

The other three parameters are equal to $P(+y \mid \boldsymbol{h}_1)$, $P(+y \mid \boldsymbol{h}_2)$, and $P(+y \mid \boldsymbol{h}_3)$, respectively, because $\boldsymbol{h}_i$ entails $\boldsymbol{h}_i \vee \boldsymbol{h}_j$. Thus $P(\boldsymbol{h}_i \mid +y)$ can be computed by Bayes rule:

$$P(\boldsymbol{h}_i \mid +y) = \frac{P(+y \mid \boldsymbol{h}_i) P(\boldsymbol{h}_i)}{\sum_{j=1}^{3} P(+y \mid \boldsymbol{h}_j) P(\boldsymbol{h}_j)}.$$

This example suggests a general methodology for computing the posterior probability of each hypothesis. The general method is based on the proof of the following two theorems.

Theorem 5 (restricted inferential completeness). *Let S be a similarity network of $P(h, u_1, \ldots, u_n)$ based on the connected cover $A_1, \ldots, A_k$ of the values of h. Let $\{v_1, \ldots, v_l\}$ be a subset of variables in $\{u_1, \ldots, u_n\}$ that satisfy related(v_i, h). Then,* *the distribution $P(h, v_1, \ldots, v_l)$ can be computed from the parameters encoded in S provided $P(\boldsymbol{h}_i) \neq 0$ for every value $\boldsymbol{h}_i$ of h.*

Proof. To show that the distribution $P(h, v_1, \ldots, v_l)$ can be computed from the parameters of S, we will show how to compute $P(h)$ and then we will show how to compute $P(v_1, \ldots, v_l \mid h)$. The product of these two probability distributions is equal to $P(h, v_1, \ldots, v_l)$.

For each hypothesis $\boldsymbol{h}_i$, let $\alpha_{i,j}$ equal $INFER(P(\boldsymbol{h}_i \mid [\![A_j]\!]), D_j)$, where A_j contains $\boldsymbol{h}_i$ and D_j is the local network corresponding to A_j. The prior probability of each $\boldsymbol{h}_i$ is computed by solving the following set of linear equations:

$$P(\boldsymbol{h}_i) = \alpha_{i,j} \cdot \sum_{\boldsymbol{h}_i \in A_j} P(\boldsymbol{h}_i), \quad \sum_{1}^{n} P(\boldsymbol{h}_i) = 1.$$

In the previous section, we solved these equations and showed that the solution (Eq. (15)) is unique provided $P(\boldsymbol{h}_i) \neq 0$ for all $\boldsymbol{h}_i$.

Due to the chaining rule, $P(v_1, \ldots, v_l \mid \boldsymbol{h}_i)$ can be factored as follows:

$$P(v_1, \ldots, v_l \mid \boldsymbol{h}_i) = P(v_1 \mid \boldsymbol{h}_i) \cdot P(v_2 \mid v_1, \boldsymbol{h}_i) \cdots P(v_l \mid v_1, \ldots, v_{l-1}, \boldsymbol{h}_i). \tag{16}$$

Thus, it suffices to show that for each variable v_j, $P(v_j \mid v_1, \ldots, v_{j-1}, \boldsymbol{h}_i)$ can be computed from the parameters encoded in S. Furthermore we can assume that the conditioning event is possible, namely, $P(v_1, \ldots, v_{l-1}, \boldsymbol{h}_i) > 0$, lest the entire product is zero and the equality holds.

Let D_i denote a local network in S, A_i be the hypotheses associated with D_i, and $\boldsymbol{h}_i$ be an hypothesis in A_i. Each variable v_j is depicted in some local network because it satisfies *related*(v_j, h). Let $A_i, A_{i+1}, \ldots, A_m$ be a path in the similarity hypergraph where A_m is the only hyperedge on this path associated with a local network that depicts v_j as a node. Such a path exists because the similarity hypergraph is connected and v_j is depicted in one of the local networks. If v_j is depicted in A_i (i.e., $m = i$) then $P(v_j \mid v_1, \ldots, v_{j-1}, \boldsymbol{h}_i)$ can be computed from the local network that corresponds to A_i.

Suppose that $m > i$. Let D_k be the local network associated with A_k for $k = i+1, \ldots, m$ and let $\boldsymbol{h}_{i+1}, \boldsymbol{h}_{i+2}, \ldots, \boldsymbol{h}_m$ be a sequence of hypotheses such that $\boldsymbol{h}_k \in A_{k-1} \cap A_k$. Due to the definition of a similarity network, since v_j is not depicted in D_k where $k < m$, v_j is unrelated to h given $[\![A_k]\!]$. Thus,

$$P(v_j \mid v_1, \ldots, v_{j-1}, \boldsymbol{h}_{k-1}, [\![A_k]\!]) = P(v_j \mid v_1, \ldots, v_{j-1}, \boldsymbol{h}_k, [\![A_k]\!]).$$

Since h implies $[\![A]\!]$ whenever $h \in A$ it follows that

$$P(v_j \mid v_1, \ldots, v_{j-1}, \boldsymbol{h}_{k-1}) = P(v_j \mid v_1, \ldots, v_{j-1}, \boldsymbol{h}_k).$$

This equation holds for every k between $i + 1$ and m, thus we obtain,

$$P(v_j \mid v_1, \ldots v_{j-1}, \boldsymbol{h}_i) = P(v_j \mid v_1, \ldots v_{j-1}, \boldsymbol{h}_m).$$

Furthermore,

$$P(v_j \mid v_1, \ldots v_{j-1}, \boldsymbol{h}_m) = P(v_j \mid v_1', \ldots v_l', \boldsymbol{h}_m) \tag{17}$$

where $v'_1, \ldots, v'_l$ are the subset of variables of $v_1, \ldots, v_{j-1}$ which are depicted in D_m.

Eq. (17) holds lest *related*$(v, v_j \mid [\![A_m]\!])$ would hold, where v is some variable in $\{v_1, \ldots, v_{j-1}\}$ not appearing in D_m. However, together with *related*$(v_j, h \mid [\![A_m]\!])$ which holds because v_j is depicted in D_m, these two assertions would imply by transitivity that *related*$(v, h \mid [\![A_m]\!])$ holds too, contradicting the fact that v is assumed not to be included in D_m. Finally,

$$P(v_j \mid v'_1, \ldots v'_l, \boldsymbol{h}_m) = P(v_j \mid v'_1, \ldots v'_l, \boldsymbol{h}_m, [\![A_m]\!]), \tag{18}$$

because $\boldsymbol{h}_m$ logically implies the disjunction over all hypotheses in A_m.

The latter probability can be computed using *INFER* on the local network D_m. Thus, due to the equalities above, $P(v_j \mid v_1, \ldots, v_{j-1}, \boldsymbol{h}_i)$ can be computed as needed. □

The above theorem shows that similarity networks are inferentially complete subject to the restriction that only features that help to discriminate between some hypotheses are included in the model and that all hypotheses which are included in the model have a probability greater than zero. Consequently, diagnostic completeness is guaranteed too.

Theorem 6 (diagnostic completeness). *Let S be a similarity network of $P(h, u_1, \ldots, u_n)$. Then the conditional distribution $P(h \mid v_1, \ldots, v_m)$ can be computed from the parameters of S for every subset $\{v_1, \ldots, v_m\}$ of $\{u_1, \ldots, u_n\}$ provided $P(\boldsymbol{h}_i) \neq 0$ for every value $\boldsymbol{h}_i$ of h.*

Proof. To compute $P(h \mid v_1, \ldots, v_m)$ observe that $P(h \mid v_1, \ldots, v_m) = P(h \mid v'_1, \ldots, v'_l)$ where $v'_1, \ldots, v'_l$ is the subset of variables in $v_1, \ldots, v_m$ that are related to h. Theorem 5 states that the joint distribution $P(h, v'_1, \ldots, v'_l)$ can be computed from the parameters of S. The conditional probability $P(h \mid v'_1, \ldots, v'_l)$ can be computed from this joint distribution. □

The above two theorems provide us with a naive computation of the posterior probability of each hypothesis. This computation does not take into account the fact that $P(h, v'_1, \ldots, v'_l)$ might be too large to be explicitly computed or stored as a table. Moreover, the computation suggested by these proofs ignores the crucial observation that, in practice, all local networks are often constructed according to a common order, say $h, v'_1, \ldots, v'_l$, which usually reflects cause-effect relations or time constrains.

When such a common order exists some computations become much easier. In particular, Eq. (18) can be further developed, that is,

$$P(v_j \mid v'_1, \ldots v'_l, \boldsymbol{h}_m, [\![A_m]\!]) = P(v_j \mid \pi_{D_m}(v_j), [\![A_m]\!]) \tag{19}$$

where $v'_1, \ldots, v'_l$ are the variables depicted in the local network D_m and $\pi_{D_m}(v_j)$ are the parents of v_j in D_m. Consequently, $P(v_j \mid v'_1, \ldots v'_l, \boldsymbol{h}_m, [\![A_m]\!])$ need not be computed using *INFER* as done in the proof. It is stored explicitly at node v_j in the local network D_m.

Eq. (19) defines a Bayesian network M_i of $P(v_1, \ldots, v_m \mid \boldsymbol{h}_i)$ because, for each v_j, π_{M_i} is set to be v_j parents' set in D_m excluding h, and the parameters associated with v_j in M_i are merely those associated with v_j in D_m. The collection of these local networks, one network for each hypothesis $\boldsymbol{h}_i$, forms an hypothesis-specific Bayesian network of $P(h, v_1, \ldots, v_m)$. We can now use our inference algorithm developed for Bayesian multinets to compute the posterior probability of each hypothesis.

The algorithm below summarizes this technique. Its first step uses arc-reversal transformations in order to reorient all local networks according to a common construction order. This step is given for the purpose of completeness, namely, to enable the algorithm to process similarity networks that are not constructed according to a common construction order. In practice, however, this step is usually not needed because similarity networks are constructed according to a common order of all relevant variables.

Algorithm (*Similarity network to Bayesian multinet conversion*).

Input: A similarity network S of $P(h, u_1, \ldots, u_n)$ based on a connected cover $A_1, \ldots, A_k$ of the values of h.

Output: A hypothesis-specific Bayesian multinet of $P(h, v_1, \ldots, v_l)$ where each v_i is depicted in some local network of S.

Notation:

- M_i is the comprehensive local network for hypothesis $\boldsymbol{h}_i$,
- D_i is the local network associated with A_i,
- $\pi_G(u)$ are the parents of u in a graph G,
- the probability associated with node u in M_i is $P_{M_i}(u \mid \pi_{M_i}(u), \boldsymbol{h}_i)$ and the probability associated with node u in D_m is $P_{D_m}(u \mid \pi_{D_m}(u), \boldsymbol{h}_m)$.

```
Reorient all local networks in $S$ according to a common construction order
For each $\boldsymbol{h}_i$ construct $M_i$ as follows
  For each $v_j$ taken in order $v_1, \ldots, v_l$
    Find a path $A_i, \ldots, A_m$ such that $\boldsymbol{h}_i \in A_i$ and $v_j$ is depicted only in $A_m$
    Set $\pi_{M_i}(v_j)$ to be $\pi_{D_m}(v_j) \setminus \{h\}$
    Set $P_{M_i}(v_j \mid \pi_{M_i}(v_j), \boldsymbol{h}_i)$ to be $P_{D_m}(v_j \mid \pi_{D_m}(v_j), \boldsymbol{h}_m)$
```

As an example, let us examine how the algorithm processes the similarity network S in Fig. 7. Because the node ordering h, g, b, l is common to all local networks of S, the algorithm performs no arc reversals. Suppose the algorithm first builds the comprehensive local network M_e for the hypothesis *executive*. Because l appears in the local network for $\{\textit{worker}, \textit{executive}\}$ with only h as a parent, the algorithm makes l a root note in M_e, and sets $P_{M_e}(l \mid \textit{executive})$ to be $P_{w \vee e}(l \mid \textit{executive})$, where $w \vee e$ denotes the local network for $\{\textit{worker}, \textit{executive}\}$. The local network for $\{\textit{visitor}, \textit{worker}\}$ is the closest neighbor to the local network for $\{\textit{worker}, \textit{executive}\}$ that depicts g and b. Because the only parent of g in the local network for $\{\textit{visitor}, \textit{worker}\}$ is h, the algorithm makes g a root node in M_e. Because g and h are the parents of b in the local network for $\{\textit{visitor}, \textit{worker}\}$, the algorithm makes g a parent of b in M_e. The algorithm sets $P_{M_e}(g \mid \textit{executive})$ to be $P_{v \vee w}(g \mid \textit{worker})$ and $P_{M_e}(b \mid g, \textit{executive})$ to be $P_{v \vee w}(b \mid g, \textit{worker})$, where $v \vee w$ denotes the local network for $\{\textit{visitor}, \textit{worker}\}$. The algorithm constructs the comprehensive local networks for *worker*, *visitor* and *spy* similarly.

4.5. Similarity networks in the real world

The practical use of similarity networks for constructing and reasoning with probabilistic models consists of a few straightforward steps. In particular, to construct a joint probability distribution for $h, u_1, \ldots, u_n$, a user (1) constructs a similarity hypergraph for the values of h, (2) constructs a local network for each hyperedge in the hypergraph, and (3) assesses each local network. The inference algorithm described in Sections 4.2–4.4 can then compute the most likely hypothesis. The quality of a similarity network is strongly influenced by the first step, wherein the user identifies sets of similar hypotheses.

In addition to the theoretical arguments in favor of the use of similarity networks, this representation has also proven itself in practice. As mentioned in the introduction, the similarity network representation has facilitated the construction of several real-world expert systems including Pathfinder—an expert system that assists pathologists with the diagnosis of lymph-node diseases [14, 16, 18]. Pathfinder reasons about over 60 diseases (25 benign diseases, 9 Hodgkin's lymphomas, 18 non-Hodgkin's lymphomas, and 10 metastatic diseases) and over 140 features of disease, including morphologic, clinical, laboratory, immunological, and molecular biological findings. A formal evaluation of Pathfinder has demonstrated that its diagnostic accuracy is at least as good as that of the expert consulted to build it [17].

5. Generalized similarity networks

In previous sections we assume all hypotheses are mutually exclusive and are, therefore, represented as values of a single hypothesis variable h. In this section, we outline a way to relax this assumption, introducing a representation that allows several variables to represent hypotheses.

Definition. Let P be a probability distribution over $\{h_1, \ldots, h_r, u_1 \ldots u_n\}$ where $H = \{h_1, \ldots, h_r\}$ is a set of distinguished variables each representing a set of hypotheses. Denote the Cartesian product of the sets of values of the distinguished variables by *domain*(H). Let $A_1, \ldots, A_k$ be a connected cover of *domain*(H). A directed acyclic graph D_i is called a *local network of P associated with A_i* if D_i is a Bayesian network of $P(h_1, \ldots, h_r, v_1, \ldots, v_m \mid [\![A_i]\!])$ where $\{v_1, \ldots, v_m\}$ is the set of all variables in $\{u_1, \ldots, u_n\}$ that "help to discriminate" the values of A_i. The set of k local networks is called a *generalized similarity network* of P.

The generalized similarity network of Fig. 8, for example, represents the following problem:

> A *pair* of people approach the secured building and the guard tries to classify them as they approach. Each approaching person is either a worker (w), a visitor (v), or a spy (s). Assume that only workers converse (c) and that workers often arrive with other workers (because they car-pool).

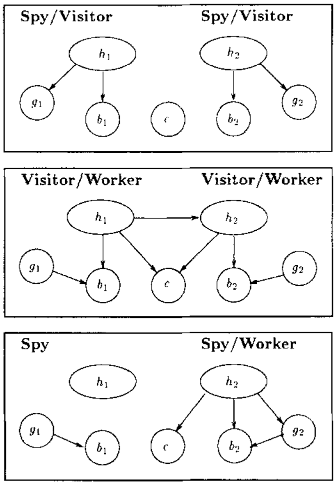

Fig. 8. A generalized similarity network with two hypothesis nodes.

Note, $H = \{h_1, h_2\}$ and *domain*(H) consists of nine elements (x, y) where both x and y are drawn from the set $\{worker, visitor, spy\}$ whose elements we denote by $\{w,\ v,\ s\}$. The connected cover of *domain*(H) corresponding to Fig. 8 is

$$\{(s,s),(v,s),(s,v),(v,v)\},$$
$$\{(v,v),(w,v),(v,w),(w,w)\},$$
$$\{(s,s),(s,w)\}.$$

The absence of a link between h_1 and h_2 in the top network encodes the fact that if the guard knew that one person is a spy or visitor, then this knowledge would not help him to decide whether the other person is a spy or a visitor. The existence of a link between h_1 and h_2 in the middle network encodes the fact that workers come in pairs more often than do visitors.

Node c should not have been included in the top local network, but it is drawn merely to highlight the independencies involving c. For the same reason, nodes g_1 and b_1 are drawn in the bottom local network. The remaining independence assertions encoded in Fig. 8 were described in previous sections or are obvious from the verbal description of the story.

The relationship between the hypothesis variables h_1 and h_2 in case of spies versus visitors is an example of *inter-hypothesis independence*, wherein two distinguished variables are independent given some hypotheses, but dependent given others. An inter-hypothesis independence assertion is represented in a generalized similarity network whenever a link between two distinguished variables exists in some local networks, but does not exist in other local networks. Such asymmetric independence assertions cannot be encoded in a non-generalized similarity network.

The results about similarity networks presented in the previous section can be extended in a straightforward manner to generalized similarity networks. An analogous definition for generalized Bayesian multinet is also self-evident.

6. Summary

In this article, we provide several enhancements to the similarity network representation originated by Heckerman [14].

First, we introduced the Bayesian multinet. We showed how the representation uses multiple Bayesian networks to encode asymmetric independence assertions, and how we can use these assertions to decrease storage requirements and increase the efficiency of inference. Next, we offered a definition of similarity networks which emphasizes the advantages of similarity networks compared to Bayesian multinets for knowledge acquisition. Then, we introduced an algorithm that converts a similarity network into a Bayesian multinet, thereby providing a general inference algorithm for similarity networks. In addition, we described generalized similarity networks which facilitate the representation of non-mutually exclusive hypotheses. We hope that this work will encourage a line of research that strives to devise additional graphical representation schemes of salient types of asymmetric independence that further simplify knowledge acquisition and inference.

Finally, we note that the computational advantages that result from sparse matrix manipulations, as suggested in [20], can also be combined with knowledge about asymmetric independence. In fact, we have argued that any inference algorithm for Bayesian networks can also be applied to a similarity network. One should emphasize, however, that these two sources of computational savings are disjoint since inference methods that rely on sparse matrices arise from the presence of zero probabilities whereas inference methods that rely on asymmetric independence constrains arise due to equalities between certain probabilities.

Acknowledgments

An earlier version of this article has been presented in the Seventh Conference on Uncertainty in Artificial Intelligence. Some of this work was done while the first author was at Northrop Research and Technology Center and the second author was at the University of California at Los Angeles. We thank Mark Peot for suggesting an early version of the inference algorithm for strictly positive distributions.

References

[1] S. Andreassen, M. Woldbye, B. Falck and S. Andersen, MUNIN: a causal probabilistic network for interpretation of electromyographic findings, in: *Proceedings IJCAI-87*, Milan (Morgan Kaufmann, San Mateo, CA, 1987) 366-372.

[2] J. Breese, E. Horvitz, M. Peot, R. Gay and G. Quentin, Automated decision-analytic diagnosis of thermal performance in gas turbines, in: *Proceedings International Gas Turbine and Aeroengine Congress and Exposition* (American Society of Mechanical Engineers, Cologne, 1992).

[3] W. Buntine, Theory refinement on Bayesian networks, in: *Proceedings Seventh Conference on Uncertainty in Artificial Intelligence*, Los Angeles, CA (Morgan Kaufmann, Los Altos, CA, 1991) 52-60.

[4] G.F. Cooper, Computational complexity of probabilistic inference using Bayesian belief networks, *Artif. Intell.* **42** (1990) 393–405.

[5] G.F. Cooper and E. Herskovits, A Bayesian method for constructing Bayesian belief networks from databases, in: *Proceedings Seventh Conference on Uncertainty in Artificial Intelligence*, Los Angeles, CA (Morgan Kaufmann, Los Altos, CA, 1991) 86–94.

[6] G.F. Cooper and E. Herskovits, A Bayesian method for the induction of probabilistic networks from data, *Mach. Learn.* **9** (1992) 309–347.

[7] R. Cox, Probability, frequency and reasonable expectation, *Am. J. Phys.* **14** (1946) 1–13.

[8] D. Geiger, Probabilistic networks, in: S. Shapiro, ed., *Encyclopedia of Artificial Intelligence* (Wiley, New York, 2nd ed., 1992) 1201–1209.

[9] D. Geiger and D. Heckerman, Separable and transitive graphoids, in: *Proceedings Sixth Conference on Uncertainty in Artificial Intelligence*, Boston, MA (Association for Uncertainty in Artificial Intelligence, Mountain View, CA, 1990) 538–545.

[10] D. Geiger and D. Heckerman, Learning Gaussian networks, in: *Proceedings Tenth Conference on Uncertainty in Artificial Intelligence*, Seattle, WA (Morgan Kaufmann, Los Altos, CA, 1994) 235–243.

[11] D. Geiger and J. Pearl, On the logic of causal models, in: *Proceedings Fourth Workshop on Uncertainty in Artificial Intelligence*, Minneapolis, MN (Association for Uncertainty in Artificial Intelligence, Mountain View, CA, 1988) 136–147; also in: R. Shachter, T. Levitt, L. Kanal and J. Lemmer, eds., *Uncertainty in Artificial Intelligence* **4** (North-Holland, New York, 1990) 3–14.

[12] D. Geiger, T. Verma and J. Pearl, Identifying independence in bayesian networks, *Networks* **20** (1990) 507–534.

[13] M. Goldbaum, B. Cote, A. Listhaus, D. Heckerman, E. Horvitz and J. Breese, IntellEye, an expert system for diagnostic ophthalmologic diseases from images of the ocular fundus, in: *Investigative Ophthalmology and Visual Science*, Supplement (1991).

[14] D. Heckerman, *Probabilistic Similarity Networks* (MIT Press, Cambridge, MA, 1991).

[15] D. Heckerman, D. Geiger and D. Chickering, Learning Bayesian networks, *Mach. Learn.* **20** (1995) 197–243.

[16] D. Heckerman, E. Horvitz and B. Nathwani, Toward normative expert systems, Part I: The Pathfinder project, *Methods Inf. Medicine* **31** (1992) 90–105.

[17] D. Heckerman and B. Nathwani, An evaluation of the diagnostic accuracy of Pathfinder, *Comput. Biomed. Res.* **25** (1992) 56–74.

[18] D. Heckerman and B. Nathwani, Toward normative expert systems, Part II: Probability-based representations for efficient knowledge acquisition and inference, *Methods Inf. Medicine* **31** (1992) 106–116; also in: J. van Bemmel and A. McCray, eds., *Yearbook of Medical Informatics* (International Medical Informatics Association, Rotterdam, 1993) 430–440.

[19] R. Howard and J. Matheson, Influence diagrams, in: R. Howard and J. Matheson, eds., *Readings on the Principles and Applications of Decision Analysis*, Vol. II (Strategic Decisions Group, Menlo Park, CA, 1981) 721–762.

[20] F. Jensen and S. Andersen, Approximations in Bayesian belief universes for knowledge based systems, in: *Proceedings Sixth Conference on Uncertainty in Artificial Intelligence*, Cambridge, MA (Morgan Kaufmann, Los Altos, CA, 1990) 162–169.

[21] F. Jensen, S. Lauritzen and K. Olesen, Bayesian updating in recursive graphical models by local computations, *Comput. Stat. Quart.* **4** (1990) 269–282.

[22] F. Jensen, K. Olesen and S. Andersen, An algebra of Bayesian belief universes for knowledge-based systems, *Networks* **20** (1990) 637–660.

[23] J. Kim and J. Pearl, A computational model for causal and diagnostic reasoning in inference engines, in: *Proceedings IJCAI-83*, Karlsruhe (1983) 190–193.

[24] H. Kyburg and H. Smokler, *Studies in Subjective Probability* (Wiley, New York, 1980).

[25] W. Lam and F. Bacchus, Using causal information and local measures to learn Bayesian networks, in: *Proceedings Ninth Conference on Uncertainty in Artificial Intelligence*, Washington, DC (Morgan Kaufmann, Los altos, CA, 1993) 243–250.

[26] S. Lauritzen and D. Spiegelhalter, Local computations with probabilities on graphical structures and their application to expert systems, *J. Roy. Stat. Soc. B* **50** (1988) 157–224.

[27] B. Nathwani, D. Heckerman, E. Horvitz and T. Lincoln, Integrated expert systems and videodisc in surgical pathology: an overview, *Human Path.* **21** (1990) 11–27.

[28] G. Nino-Murcia and M. Shwe, An expert system for diagnosis of sleep disorders, in: M. Chase, R. Lydic and C. O'Connor, eds., *Sleep Research* **20** (Brain Information Service, Los Angeles, CA, 1991) 433.

[29] S. Olmsted, On representing and solving decision problems, Ph.D. Thesis, Department of Engineering-Economic Systems, Stanford University, Stanford, CA (1983).

[30] J. Pearl, Reverend Bayes on inference engines: a distributed hierarchical approach, in: *Proceedings AAAI-82*, Pittsburgh, PA (AAAI Press, Menlo Park, CA, 1982) 133–136.

[31] J. Pearl, Fusion, propagation, and structuring in belief networks, *Artif. Intell.* **29** (1986) 241–288.

[32] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference* (Morgan Kaufmann, San Mateo, CA, 1988).

[33] J. Pearl and T. Verma, A theory of inferred causation, in: J. Allen, R. Fikes and E. Sandewall, eds., *Knowledge Representation and Reasoning: Proceedings Second International Conference* (Morgan Kaufmann, San Mateo, CA, 1991) 441–452.

[34] L. Savage, *The Foundations of Statistics* (Dover, New York, 1954).

[35] R. Shachter, Evaluating influence diagrams, *Oper. Res.* **34** (1986) 871–882.

[36] R. Shachter and C. Kenley, Gaussian influence diagrams, *Manage. Sci.* **35** (1989) 527–550.

[37] P. Spirtes, C. Glymour and R. Scheines, *Causation, Prediction, and Search* (Springer, New York, 1993).

[38] T. Verma and J. Pearl, Causal networks: semantics and expressiveness, in: *Proceedings Fourth Workshop on Uncertainty in Artificial Intelligence*, Minneapolis, MN (Association for Uncertainty in Artificial Intelligence, Mountain View, CA, 1988) 352–359; also in: R. Shachter, T. Levitt, L. Kanal and J. Lemmer, eds., *Uncertainty in Artificial Intelligence* 4 (North-Holland, New York, 1990) 69–76.