Distributed Revision of Composite Beliefs*

Judea Pearl
Cognitive Systems Laboratory, UCLA Computer Science Department, Los Angeles, CA 90024, U.S.A.

Recommended by Richard Duda

ABSTRACT

This paper extends the applications of belief network models to include the revision of belief "commitments," i.e., the categorical acceptance of a subset of hypotheses which, together, constitute the most satisfactory explanation of the evidence at hand. A coherent model of nonmonotonic reasoning is introduced, and distributed algorithms for belief revision are presented. We show that, in singly connected networks, the most satisfactory explanation can be found in linear time by a message-passing algorithm similar to the one used in belief updating. In multiply connected networks, the problem may be exponentially hard but, if the network is sparse, topological considerations can be used to render the interpretation task tractable. In general, finding the most probable combination of hypotheses is no more complex than computing the degree of belief for any individual hypothesis. Applications to circuit and medical diagnosis are illustrated.

1. Introduction

People's beliefs are normally cast in categorical terms, often involving not just one, but a composite set of propositions which, stated together, offer a satisfactory account of the observed data. For example, a physician might state, "This patient apparently suffers from two simultaneous disorders A and B which, due to condition C, caused the deterioration of organ D." Except for the hedging term "apparently," such a composite statement conveys a sense of unreserved commitment (of beliefs) to a set of four hypotheses. The individual components in the explanation above are meshed together by mutually enforced cause-effect relationships, forming a cohesive whole; the removal of any one component from the discourse would tarnish the completeness of the entire explanation.

Such a sense of cohesiveness normally suggests that a great amount of

*This work was supported in part by the National Science Foundation, Grant DCR 83-13875. An earlier version of this paper was presented at the 2nd AAAI Workshop on Uncertainty in Artificial Intelligence, Philadelphia, PA, August 1986.

refuting evidence would have to be gathered before the current interpretation would undergo a revision. Moreover, once a revision is activated, it will likely change the entire content of the interpretation, not merely its level of plausibility. Another characteristic of coherent explanations is that they do not assign degrees of certainty to any individual hypothesis in the argument; neither do they contain information about alternative, next-to-best combinations of hypotheses.

Even the certainty of the accepted composite explanation is only seldom consulted; most everyday activities are predicated upon beliefs which, despite being provisional do not seem to be muddled with varying shades of uncertainty. Consider, for example the sentence: "John decided to have a bowl of cereal but, finding the cupboard empty, he figured out that Mary must have finished it at breakfast." Routine actions such as reaching for the cupboard are normally performed without the slightest hesitation or reservation, thus reflecting adherence to firmly held beliefs (of finding cereal there). When new facts are observed, refuting current beliefs, a process of belief revision takes place; new beliefs replace old ones, also to be firmly held, until refuted.

These behavioral features are somewhat at variance with past work on belief network models of evidential reasoning [21]. Thus far, this work has focussed on the task of *belief updating*, i.e., assigning each hypothesis in a network a degree of belief, BEL($\cdot$), consistent with all observations. The function BEL changes smoothly and incrementally with each new item of evidence.

This paper extends the applications of Bayesian analysis and belief network models to include revision of belief *commitments*, i.e., the tentative categorical acceptance of a subset of hypotheses which, together, constitute the most satisfactory explanation of the evidence at hand. Using probabilistic terminology, that task amounts to finding the most probable instantiation of all hypothesis variables, given the observed data. The resulting output is an optimal list of jointly accepted propositions that may vary dynamically as more evidence obtains.

In principle, this optimization task seems intractable because enumerating and rating all possible instantiations is computationally prohibitive and, instead, many heuristic techniques have been developed in various fields of application. In pattern recognition the problem became known as the "multimembership problem" [2]; in medical diagnosis it is known as "multiple disorders" [1, 5, 24–26] and in circuit diagnosis as "multiple faults" [6, 27].

This paper departs from previous work by emphasizing a *distributed* computation approach to belief revision. The impact of each new piece of evidence is viewed as a perturbation that propagates through the network via local communication among neighboring concepts, with minimum external supervision. At equilibrium, each variable will be bound to a definite value which, together with all other value assignments, is the best interpretation of the evidence. The main reason for adopting this distributed message-passing paradigm is that it provides a natural mechanism for exploiting the independencies embodied in sparsely constrained systems and translating them, by subtask decomposition, into substantial reduction in complexity. Additionally, distributed propagation is inherently "transparent," namely, the intermediate steps, by virtue of their reflecting interactions only among semantically related variables, are conceptually meaningful. This facilitates the use of natural, object-oriented programming tools and helps establish confidence in the final result.

We show that, in singly connected networks, the most satisfactory explanation can be found in linear time by a message-passing algorithm similar to the one used in belief updating. In multiply connected networks, the problem may be exponentially hard but, if the network is sparse, topological considerations can be used to render the interpretation task tractable. In general, assembling the most believable combination of hypotheses is no more complex than computing the degree of belief for any individual hypothesis.

This paper comprises seven sections. Section 2 provides a brief summary of belief updating in Bayesian networks, as described in [21]. It defines the semantics of network representation, describes the tasks of belief updating and summarizes the propagation rules which lead to coherent updating in singly connected networks. Section 3 illustrates the propagation scheme using a simple example of circuit diagnosis, and compares belief updating with belief revision on the same example. Section 4 develops the propagation rules for belief revision in singly connected networks and compares them to those governing belief updating. Section 5 extends the propagation scheme to multiply connected networks using two methods, clustering and conditioning. Section 6 illustrates the method of conditioning on a simple medical diagnosis example, involving four diseases and four symptoms. Section 7 relates the revision process described in this paper to previous philosophical work on belief acceptance, discusses the adequacy of the "most probable" criterion, and touches on the issues of hysteresis and consistency in belief revision.

2. Review of Belief Updating in Bayesian Belief Networks

Bayesian belief networks [21] are directed acyclic graphs (DAGs) in which the nodes represent propositional variables, the arcs signify the existence of direct causal influences between the linked propositions, and the strengths of these influences are quantified by the conditional probabilities of each variable given the state of its parents. Thus, if the nodes in the graph represent the ordered variables $X_1, X_2, \ldots, X_n$, then each variable X_i draws arrows from a subset S_i of variables perceived to be "direct causes" of X_i, i.e., S_i is a set of X_i's predecessors satisfying $P(x_i|s_i) = P(x_i|x_1, x_2, \ldots, x_{i-1})$. A complete and consistent parametrization of the model can be obtained by specifying, for each X_i, an assessment of $P(x_i|s_i)$. The product of all these local assessments,

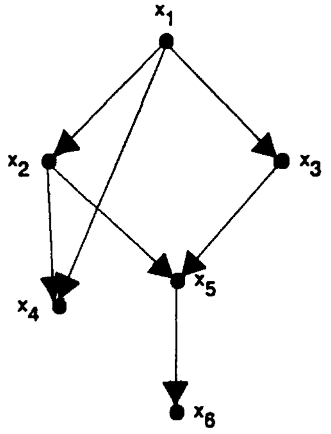

FIG. 1. A typical Bayesian network representing the distribution $P(x_6|x_5)P(x_5|x_2, x_3)$ $P(x_4|x_1, x_2)P(x_3|x_1)P(x_2|x_1)P(x_1)$.

$$P(x_1, x_2, \ldots, x_n) = \prod_i P(x_i|s_i)\,,$$

constitutes a joint probability model consistent with the assessed quantities. Thus, for example, the distribution corresponding to the network of Fig. 1 can be written by inspection:[1]

$$\begin{aligned} &P(x_1, \ldots, x_6) \\ &\quad = P(x_6|x_5)P(x_5|x_2, x_3)P(x_4|x_1, x_2)P(x_3|x_1)P(x_2|x_1)P(x_1)\,. \end{aligned}$$

A Bayesian network provides a clear graphical representation for the essential independence relationships embedded in the underlying causal model. These independencies can be detected by the following *DAG-separation* criterion: if all paths between X_i and X_j are "blocked" by a subset S of variables, then X_i is independent of X_j, given the values of the variables in S. A path is "blocked" by S if it contains a member of S between two diverging or two cascaded arrows or, alternatively, if it contains two arrows converging at node X_k, and neither X_k nor any of its descendants is in S. In particular, each variable X_i is independent of both its grandparents and its nondescendant siblings, given the values of the variables in its parent set S_i. In Fig. 1, for

[1] Probabilistic formulae of this kind are shorthand notation for the statement that for any instantiation $x_1, x_2, \ldots, x_n$ of the variables $X_1, X_2, \ldots, X_n$, the probability of the joint event $(X_1 = x_1)$ & $\cdots$ & $(X_n = x_n)$ is equal to the product of the probabilities of the corresponding conditional events $(X_1 = x_1)$, $(X_2 = x_2$ if $X_1 = x_1)$, $(X_3 = x_3$ if $X_2 = x_2$ & $X_1 = x_1)$, $\ldots$.

example, X_2 and X_3 are independent, given either $\{X_1\}$ or $\{X_1, X_4\}$, because the two paths between X_2 and X_3 are blocked by either one of these sets. However, X_2 and X_3 may not be independent given $\{X_1, X_6\}$ because X_6, as a descendant of X_5, "unblocks" the head-to-head connection at X_5, thus opening a pathway between X_2 and X_3.

Once a Bayesian network is constructed, it can be used as an interpretation engine, namely, newly arriving information will set up a parallel constraint-propagation process which ripples multidirectionally through the networks until, at equilibrium, every variable is assigned a measure of belief consistent with the axioms of probability calculus. Incoming information may be of two types: *specific evidence* and *virtual evidence*. Specific evidence corresponds to direct observations which validate, with certainty, the values of some variables already in the network. Virtual evidence corresponds to judgment based on undisclosed observations which affect the belief of some variables in the network. Such evidence is modeled by dummy nodes representing the undisclosed observations connected to the variables affected by the observations.

The objective of updating beliefs coherently by purely local computations can be fully realized if the network is singly connected, namely, if there is only one undirected path between any pair of nodes. These include causal trees, where each node has a single parent, as well as networks with multi-parent nodes, representing events with several causal factors. We shall first review the propagation scheme in singly connected networks and then discuss (in Section 5) how it can be extended to multiply connected networks.

Let variable names be denoted by capital letters, e.g. U, V, X, Y, Z and their associated values by lower case letters, e.g., u, v, x, y, z. All incoming information, both specific and virtual, will be denoted by $\boldsymbol{e}$ to connote *evidence* and will be represented by nodes whose values are held constant. For the sake of clarity, we will distinguish between the fixed conditional probabilities that label the links, e.g., $P(x|u, v)$, and the dynamic values of the updated node probabilities. The latter will be denoted by BEL(x), which reflects the overall belief accorded to the proposition $X = x$ by all evidence so far received. Thus,

$$\mathrm{BEL}(x) \overset{\Delta}{=} P(x|\boldsymbol{e}) , \tag{1}$$

where $\boldsymbol{e}$ is the value combination of all instantiated variables.

Consider a fragment of a singly connected Bayesian network, as depicted in Fig. 2. The link $U \rightarrow X$ partitions the graph into two: a *tail* subgraph, G^+_{UX}, and a *head* subgraph, G^-_{UX}, the complement of G^+_{UX}. Each of these two subgraphs may contain a set of evidence, which we shall call respectively $\boldsymbol{e}^+_{UX}$ and $\boldsymbol{e}^-_{UX}$. Likewise, the links $V \rightarrow X$, $X \rightarrow Y$ and $X \rightarrow Z$ respectively define the subgraphs G^+_{VX}, G^-_{XY}, and G^-_{XZ}, which may contain the respective evidence sets $\boldsymbol{e}^+_{VX}$, $\boldsymbol{e}^-_{XY}$ and $\boldsymbol{e}^-_{XZ}$.

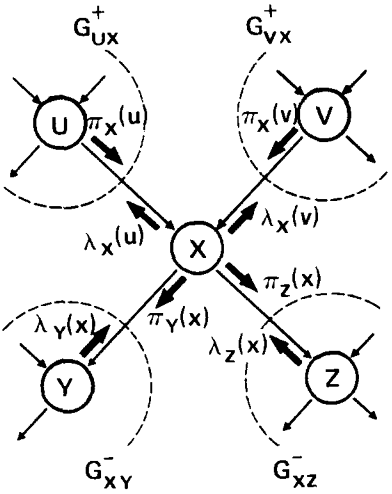

FIG. 2. Fragment of a singly connected network with multiple parents, illustrating graph partitioning and message parameters.

The belief distribution of each variable X in the network can be computed if three types of parameters are made available:

(1) the current strength of the *causal* support, π, contributed by each incoming link to X:

$$\pi_X(u) = P(u|e^+_{UX}) , \tag{2}$$

(2) the current strength of the *diagnostic* support, λ, contributed by each outgoing link from X:

$$\lambda_Y(x) = P(e^-_{XY}|x) , \tag{3}$$

(3) the fixed conditional-probability matrix, $P(x|u, v)$, which relates the variable X to its immediate parents.

Using these parameters, local belief updating can be accomplished by the following three steps, to be executed in any order:

Step 1: *Belief updating*. When node X is activated to update its parameters, it simultaneously inspects the $\pi_X(u)$ and $\pi_X(v)$ communicated by its parents and the messages $\lambda_Y(x)$, $\lambda_Z(x), \ldots$ communicated by each of its sons. Using this input, it then updates its belief measure as follows:

$$\mathrm{BEL}(x) = \alpha\lambda_Y(x)\lambda_Z(x) \sum_{u,v} P(x|u, v)\pi_X(u)\pi_X(v) , \tag{4}$$

where α is a normalizing constant, rendering

$$\sum_x \mathrm{BEL}(x) = 1 .$$

Step 2: *Updating* λ. Using the messages received, each node computes new λ messages to be sent to its parents. For example, the new message $\lambda_X(u)$ that X sends to its parent U is computed by:

$$\lambda_X(u) = \alpha \sum_v \left[\pi_X(v) \sum_x \lambda_Y(x) \lambda_Z(x) P(x|u, v) \right] . \tag{5}$$

Step 3: *Updating* π. Each node computes new π messages to be sent to its children. For example, the new $\pi_Y(x)$ message that X sends to its child Y is computed by:

$$\pi_Y(x) = \alpha \lambda_Z(x) \left[\sum_{u,v} P(x|u, v) \pi_X(u) \pi_X(v) \right] . \tag{6}$$

These three steps summarize the six steps described in [21] and can be executed in any order. (Step 1 can be skipped when BEL(x) is of no interest.) An alternative way of calculating BEL(x) would be to multiply the incoming and outgoing messages on some link from X to any of its children, e.g.,

$$\mathrm{BEL}(x) = \alpha \pi_Y(x) \lambda_Y(x) , \tag{7}$$

where $\pi_Y(x)$ is calculated via (6).

This concurrent message-passing process is both initiated and terminated at the peripheral nodes of the network, subject to the following boundary conditions:

(1) *An anticipatory node* represents an uninstantiated variable with no successors. For such a node, X, we set $\lambda_Y(x) = (1, 1, \ldots, 1)$.

(2) *An evidence node* represents a variable with instantiated value. If variable X assumes the value x', we introduce a dummy child Z with

$$\lambda_Z(x) = \begin{cases} 1, & \text{if } x = x' , \\ 0, & \text{otherwise} . \end{cases}$$

This implies that, if X has children, $Y_1, \ldots, Y_m$, each child should receive the same message $\pi_{Y_j}(x) = \lambda_Z(x)$ from X.

(3) A *root node* represents a variable with no parents. For each root variable X, we introduce a dummy parent U, permanently instantiated to $U = 1$, and set the conditional probability on the link $U \rightarrow X$ equal to the prior probability of X, i.e., $P(x|u) = P(x)$.

In [21], it is shown that, in singly connected networks, the semantics of the messages produced via (4)–(6) are preserved, namely,

$$\lambda_X(u) = P(e^-_{UX} | u)\ , \qquad \pi_Y(x) = P(x | e^+_{XY})\ , \tag{8}$$

and

$$\text{BEL}(x) = P(x | e)\ . \tag{9}$$

3. Illustrating the Propagation Scheme

The simple circuit of Fig. 3(a) will be used to illustrate the propagation pattern of the proposed scheme, the semantics of the messages involved, as well as the difference between belief updating and belief revision. The circuit consists of three AND gates in tandem. X_1, X_2 and X_3 are binary input variables $(X_i \in \{0, 1\})$, Y_3 is the circuit's output $(Y_3 = X_1 \wedge X_2 \wedge X_3)$ and Y_1, Y_2 are intermediate, unobserved variables $(Y_i = Y_{i-1} \wedge X_i)$. Under normal operation all inputs are ON and so is the output Y_3. A failure occurs when any of the inputs is OFF which would be reflected in $Y_3 = 0$ (the circuits are assumed to be operational). The problem is to infer which input is faulty given the simultaneous observations $\{Y_3 = 0, X_2 = 1\}$ and assuming that failures are independent events with prior probabilities

$$q_i = 1 - p_i = P(X_i = 0)\ , \quad i = 1, 2, 3\ . \tag{10}$$

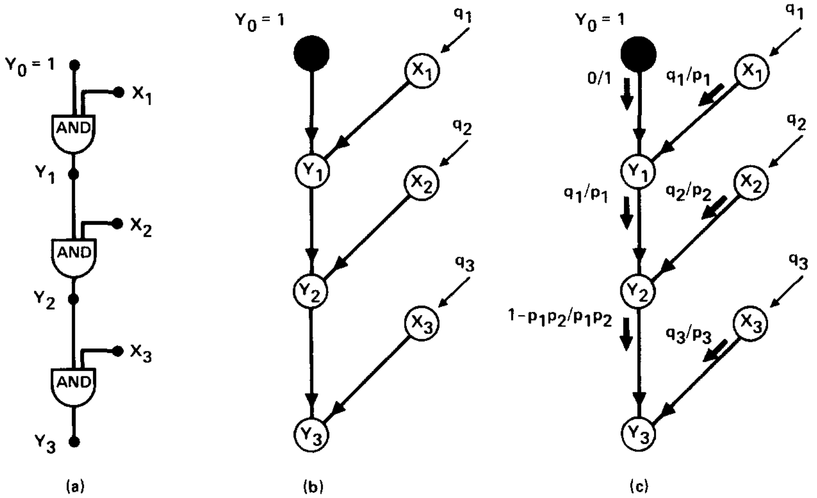

FIG. 3. (a) Logic circuit used to demonstrate the process of belief updating. (b) The Bayesian network corresponding to the circuit in (a). (c) Profile of π messages in the initial state of the network; the λ messages (not shown) are unit vectors.

The Bayesian network corresponding to this circuit diagram is shown in Fig. 3(b). Since the output of each component is functionally determined by the state of its two inputs, X_i and Y_{i-1} are identified as the parents of Y_i, $i = 1, 2, 3$, and the conditional probabilities which characterize these child-parents relationships are given by:

$$P(y_i | y_{i-1}, x_i) = \begin{cases} 1, & \text{if } y_i = y_{i-1} \wedge x_i , \\ 0, & \text{otherwise} . \end{cases} \tag{11}$$

3.1. Distributed belief updating

In the initial, quiescent state (Fig. 3(c)), all λ are unit vectors, $\lambda = (1, 1)$, since no variable has any observed descendant (see (3)) and, so, there exists no evidence favoring the state 1 over the state 0. The π messages on the links are computed from (6) and (10) and (11), and are given by:

$$\pi_{Y_i}(x_i) = \begin{cases} q_i , & x_i = 0 , \\ p_i , & x_i = 1 ; \end{cases} \tag{12}$$

$$\pi_{Y_2}(y_1) = \begin{cases} q_1 , & y_1 = 0 , \\ p_1 , & y_1 = 1 ; \end{cases} \tag{13}$$

$$\pi_{Y_3}(y_2) = \begin{cases} 1 - p_1 p_2 , & y_2 = 0 , \\ p_1 p_2 , & y_2 = 1 . \end{cases} \tag{14}$$

They simply describe the prior ON-OFF probabilities of the corresponding variables. The belief measures can be computed from these messages using (4) or (7) and, they, too, stand for the prior probabilities associated with the individual variables. For example,

$$\mathrm{BEL}(x_2) = \begin{cases} q_2 , & x_2 = 0 , \\ p_2 , & x_2 = 1 ; \end{cases} \tag{15}$$

$$\mathrm{BEL}(y_3) = \begin{cases} 1 - p_1 p_2 p_3 , & y_3 = 0 , \\ p_1 p_2 p_3 , & y_3 = 1 . \end{cases} \tag{16}$$

Now imagine that two observations are conducted simultaneously, giving $Y_3 = 0$, $X_2 = 1$. The first implies that at least one input is faulty while the second exonerates X_2, leaving either X_1 or X_3 (or both) as candidate culprits. The problem is small enough to permit an immediate global computation of all probabilities. For example, the probability that input X_1 is faulty is given by:

$$\begin{aligned} P(X_1 = 0 | e) &= \mathrm{BEL}(x_1 = 0) \\ &= \frac{q_1 p_3 + q_1 q_3}{q_1 q_3 + q_1 p_3 + p_1 q_3} = \frac{q_1}{1 - p_1 p_3} , \end{aligned} \tag{17}$$

while

$$P(X_3 = 0|e) = \mathrm{BEL}(x_3 = 0)$$

$$= \frac{q_1q_3 + p_1q_3}{q_1q_3 + q_1p_3 + p_1q_3} = \frac{q_3}{1 - p_1p_3} . \tag{18}$$

In a large network the problem may not be as easy and we shall next demonstrate how the results (17) and (18) can be obtained by distributed computation.

Figure 4 illustrates three successive stages of the propagation process triggered by the two observations, assuming that a processor is assigned to each variable and that each processor is activated if any change occurs in the messages incident on that processor. Each diagram displays the messages updated at the corresponding stage; the top-down arrows represent π messages and bottom-up arrows represent λ messages.

In Fig. 4(a) the instantiation of Y_3 and X_2 triggers the update of three messages: $\lambda_{Y_3}(x_3)$, $\lambda_{Y_3}(y_2)$ and $\pi_{Y_2}(x_2)$. Their magnitudes are computed locally from (5) and (6), using the π values in (12)–(14), giving:

$$\pi_{Y_2}(x_2) = (0, 1) , \tag{19}$$

$$\lambda_{Y_3}(x_3) = (1, 1 - p_1p_2) , \tag{20}$$

$$\lambda_{Y_3}(y_2) = (1, q_3) . \tag{21}$$

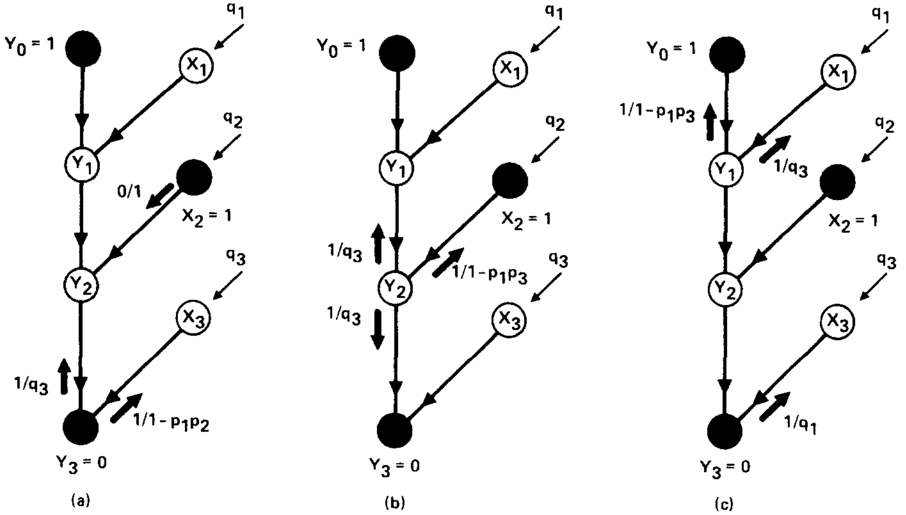

FIG. 4. Propagation of updated π and λ messages after observing $X_2 = 1$ and $Y_3 = 0$. (a) Messages generated immediately after the observations. (b) Messages generated by the activation of Y_2. (c) Messages generated by the activation of Y_1 and Y_3.

At the next phase of propagation (Fig. 4(b)), Y_2 is activated and generates three new messages: $\lambda_{Y_2}(y_1)$, $\lambda_{Y_2}(x_2)$ and $\pi_{Y_3}(y_2)$. The first incorporates the changes observed in both $\lambda_{Y_3}(y_2)$ and $\pi_{Y_2}(x_3)$ while the latter two reflect the recent changes in $\lambda_{Y_3}(y_2)$ and $\pi_{Y_2}(x_3)$, respectively. Their magnitudes are given by:

$$\lambda_{Y_2}(y_1) = (1, q_3)\,, \tag{22}$$

$$\lambda_{Y_2}(x_2) = (1, 1 - p_1p_3)\,, \tag{23}$$

$$\pi_{Y_3}(y_2) = (1 + q_3)^{-1}(1, q_3)\,, \tag{24}$$

The final phase of propagation is depicted in Fig. 4(c). Processors Y_1 and Y_3 are activated simultaneously and generate the messages $\lambda_{Y_1}(y_0)$, $\lambda_{Y_1}(x_1)$ and $\lambda_{Y_3}(x_3)$. The first is superfluous since Y_0 is "clamped" to 1. The latter two are computed via (5), giving:

$$\lambda_{Y_1}(x_1) = \lambda_{Y_2}(y_1) = (1, q_3)\,, \tag{25}$$

$$\lambda_{Y_3}(x_3) = (1, q_1)\,. \tag{26}$$

They reflect the relative probabilities of the total evidence observed, conditioned on the two possible values, 0 and 1, of the variables X_1 and X_3, respectively. For example, under the assumption $X_1 = 0$ the probability of the total evidence $e = \{X_2 = 1, Y_3 = 0\}$ is p_2; while under the assumption $X_1 = 1$ that probability becomes p_2q_3 (X_3 must be faulty to explain $Y_3 = 0$).

From these final values of the λ and π messages the belief distribution BEL can be computed for each variable in the system, (7). For example, for X_1 we obtain

$$\begin{aligned}\mathrm{BEL}(x) &= \alpha\pi_{Y_1}(x)\lambda_{Y_1}(x) = \alpha(q_1, p_1)(1, q_3) \\ &= \alpha(q_1, p_1q_3) = \left(\frac{q_1}{q_1 + p_1q_3}, \frac{p_1q_3}{q_1 + p_1q_3}\right),\end{aligned} \tag{27}$$

identically to (17).

3.2. Distributed belief revision

The aim of belief *revision* is not to associate a measure of belief with each individual proposition but, rather, to identify a composite set of propositions (one from each variable) which "best" explains the evidence at hand. In the example of Fig. 3(a), the aim is to find a consistent assignment of values to the set of uninstantiated variables, $\{X_1, X_3, Y_2\}$, which best explains the evidence

$e = \{Y_3 = 0, X_2 = 1\}$. Since Y_2 is functionally dependent on X_1 the space of consistent assignments is determined by the values assigned to X_1 and X_3 and, since $(X_1 = 1, X_2 = 1)$ is incompatible with $Y_3 = 0$, the choice is between three candidates:

$$I_1 = \{X_1 = 0, X_3 = 0\} ,$$
$$I_2 = \{X_1 = 0, X_3 = 1\} ,$$
$$I_3 = \{X_1 = 1, X_3 = 0\} .$$

We shall refer to such assignments, interchangeably, as *explanations*, *extensions* or *interpretations*.

Basic probabilistic considerations dictate

$$P(I_1|e) = \frac{q_1 q_3}{1 - p_1 p_2} ,$$
$$P(I_2|e) = \frac{q_1 p_3}{1 - p_1 p_2} ,$$
$$P(I_3|e) = \frac{p_1 q_3}{1 - p_1 p_2} , \tag{28}$$

where, assuming for simplicity

$$\tfrac{1}{2} > q_1 > q_2 > q_3 , \tag{29}$$

$I_2 = \{X_1 = 0, X_3 = 1\}$ is identified as the "best" explanation of the evidence e. However, this optimal assignment cannot be obtained by simply optimizing the beliefs of the individual variables. For example, taking $q_1 = 0.45$ and $q_3 = 0.4$ yields ((17) and (18)).

$$\mathrm{BEL}(x_1 = 0) = 0.672 > \mathrm{BEL}(x_1 = 1) = 0.328 ,$$
$$\mathrm{BEL}(x_3 = 0) = 0.597 > \mathrm{BEL}(x_3 = 1) = 0.403 .$$

Yet, choosing the most probable value of each variable separately yields the assignment $I_1 = \{X_1 = 0, X_3 = 0\}$ which is the *least* probable explanation, with $P(I_1|e) = 0.268$ compared with $P(I_2|e) = 0.403$ and $P(I_3|e) = 0.328$.

We shall now demonstrate how the optimal explanation can be assembled by a distributed message-passing scheme similar to that used in belief updating (Fig. 4). Clearly, the messages used in this scheme should carry a summarized description of the entire network, sufficient to guarantee that local choices of individual variables constitute a globally optimal explanation. In our example (Fig. 4(a)), the final messages incidenting on processor X_1 should locally determine the choice $X_1 = 0$ and, simultaneously, those incidenting on X_3 should dictate $X_3 = 1$.

To meet this goal we associate with each variable X a new function, BEL*(x), which, for each value x represents *the probability of the best interpretation of the proposition* $X = x$, i.e., the interpretation in which the values of all other variables were adjusted so as to attain their most probable combination. For example, in Fig. 3(b), the best interpretation of the proposition $Y_1 = 0$ is the assignment $\{X_1 = 0, X_2 = 1, X_3 = 1\}$, with probability $q_1 p_2 p_3$, while the proposition $Y_2 = 1$ is best interpreted by the no-fault condition $\{X_1 = 1, X_2 = 1, X_3 = 1\}$, with probability $p_1 p_2 p_3$. Thus, the BEL* function associated with Y_2 will be

$$\mathrm{BEL}^*(y_2) = \begin{cases} q_1 p_2 p_3 , & \text{if } y_2 = 0 , \\ p_1 p_2 p_3 , & \text{if } y_2 = 1 , \end{cases} \tag{30}$$

and, since $q_1 < p_1$ (see (29)), the local choice $y_2 = 1$ is guaranteed to be part of the globally optimal explanation.

The computation of BEL* can be accomplished by a local, message-passing scheme similar to that of belief updating. The propagation dynamics is identical to that depicted in Fig. 4, except that the information carried by the messages has different meaning and the computations paralleling those of (4)–(6) involve maximization rather than summation.

The ability to assemble a globally optimal solution by local computations rests on the many conditional-independence relations embodied in the system, as is reflected in the network topology (see [21, 22] for formal treatment of conditional independence and its graphical representations). These permit us to decompose the task of finding a best overall explanation into smaller subtasks of finding best explanations in subparts of the network, then combining them together. In Fig. 3(b), for example, finding the best explanation for $Y_2 = 0$ can be decomposed into two independent subtasks:

(1) Find a best subexplanation for $Y_2 = 0$ in the *tail* subgraph of the link $Y_2 \rightarrow Y_3$ (i.e., comprising $\{X_1, Y_1, X_2\}$).

(2) Find a best subexplanation for $Y_2 = 0$ in the *head* subgraph of the link $Y_2 \rightarrow Y_3$ (i.e., comprising $\{Y_3, X_3\}$).

The fact that these two subgraphs are joined only by the link $Y_2 \rightarrow Y_3$ guarantees that the overall best explanation (for $Y_2 = 0$) is composed precisely of the two subexplanations found in (1) and (2) above. Moreover, the degree of support that the overall best explanation extends to $Y_2 = 0$ can be computed (locally) from those extended by the two subexplanations. Thus, both BEL*$(Y_2 = 0)$ and BEL*$(Y_2 = 1)$ can be computed locally and the best value for Y_2 decided by choosing the one with the highest value of BEL* (in our case $Y_2 = 0$). The messages carrying these partial degrees of support will be denoted by π^* and λ^*, respectively, formally defined as

$$\lambda_Y^*(x) = \max_{w_{XY}^-} P(w_{XY}^- | x, e) , \tag{31}$$

$$\pi_Y^*(x) = \max_{w_{XY}^+} P(x, w_{XY}^+ | e) , \tag{32}$$

where w_{XY}^- and w_{XY}^+ stand, respectively, for any head extension and tail extension of $\{X = x, e\}$, relative to the link $X \rightarrow Y$ (see Fig. 2). For example, in Fig. 5(a), $e = \emptyset$ and the best tail extension of $Y_2 = 0$ is $w_{Y_2Y_3}^+ = \{X_1 = 0, Y_1 = 0, X_2 = 1\}$, with

$$\pi_{Y_3}^*(y_2 = 0) = P(Y_2 = 0, X_1 = 0, Y_1 = 0, X_2 = 1) = q_1 p_2 , \tag{33}$$

while its best head extension is $w_{Y_2Y_3}^- = \{Y_3 = 0, X_3 = 1\}$ with

$$\lambda_{Y_3}^*(y_2 = 0) = P(Y_3 = 0, X_3 = 1 | Y_2 = 0) = p_3 . \tag{34}$$

By similar considerations we obtain:

$$\pi_{Y_3}^*(y_2 = 1) = P(Y_2 = 1, X_1 = 1, Y_1 = 1, X_2 = 1) = p_1 p_2 , \tag{35}$$

$$\lambda_{Y_3}^*(y_2 = 1) = P(Y_3 = 1, X_3 = 1 | Y_2 = 1) = p_3 , \tag{36}$$

thus yielding the messages:

$$\pi_{Y_3}^*(y_2) = (q_1 p_2, p_1 p_2) , \tag{37}$$

$$\lambda_{Y_3}^*(y_2) = (p_3, p_3) . \tag{38}$$

In Section 4, we shall demonstrate that

(1) the π^* and λ^* messages defined above can be propagated by local computations, simply replacing the summations in (5) and (6) by maximizations (over the same set of variables), as in (53) and (56);

(2) the BEL* functions can be computed from the π^* and λ^* messages by simple products, e.g.,

$$\text{BEL}^*(x) = \alpha \pi_Y^*(x) \lambda_Y^*(x) , \tag{39}$$

or, alternatively, using a modification of (4) with maximization replacing the summation (see (52)).

The rest of this section provides a qualitative description of how the best explanations in the example of Fig. 3 are found by a message-passing process. Quantitative account will be postponed until the propagation rules are estab-

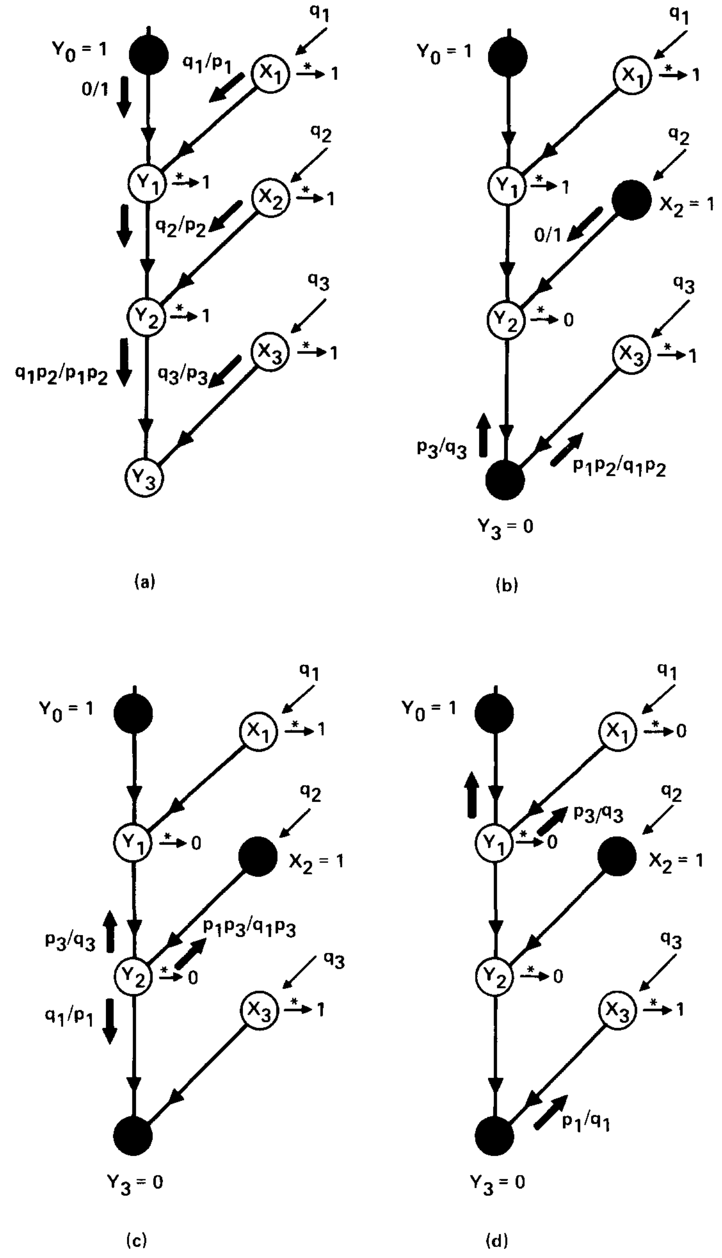

FIG. 5. π^* and λ^* message propagation under belief revision. The observation $\{Y_3 = 0, X_2 = 1\}$ causes a switch from the initial default explanation $\{X_1 = X_2 = X_3 = 1\}$ in (a) to a new stable (and maximally probable) explanation $\{X_1 = 0, X_2 = 1, X_3 = 1\}$ in (d). The intermediate states in (b) and (c) yield temporary belief commitments based on incomplete transient information.

lished in Section 4 (see equations (52), (54) and (56) or, more generally, (59)–(61)).

Initially, all λ^* are unit vectors and the π^* messages are given in Fig. 5(a). These are almost identical to the π messages of Fig. 3(c) except for $\pi^*_{Y_3}(y_2)$ (see (14), (33) and (35)). The difference stems from the fact that while $\pi_{Y_3}(y_2)$ represents the total probability of all tail extensions of $Y_2 = y_2$, $\pi^*_{Y_3}(y_2)$ represents the probability of only *one* such tail extension, namely, the most probable one. The implications $\overset{*}{\rightarrow} 1$ indicate the current commitments made on the basis of BEL* (see (39)) which, at this stage, represent the default state $\{X_1 = X_2 = X_3 = 1\}$. Note, however that the initial π^* values represent, not just the currently committed explanation, but a whole set of possible system behaviors, each being a best explanation for some possible future observation of the form $Y = y$ or $X = x$.

When nodes Y_3 and X_2 are instantiated (Fig. 5(b)) they set up new π^* and λ^* messages which, temporarily, yield suboptimal and inconsistent belief commitments, such as $\{X_1 = 1, Y_1 = 1, Y_2 = 0, X_3 = 1\}$ in Fig. 5(b) and $\{X_1 = 1, Y_1 = 0, Y_2 = 0, X_3 = 1\}$ in Fig. 5(c). Eventually, however, all messages are absorbed at the peripheral nodes and a new consistent explanation emerges, $\{X_1 = 0, Y_1 = 0, Y_2 = 0, X_3 = 1\}$, which is also globally optimal. In general, the propagation process can be activated concurrently, it subsides in time proportional to the network diameter and, at equilibrium, all belief commitments are optimal.

4. Belief Revision in Singly Connected Networks

Let $\boldsymbol{W}$ stand for the set of all variables considered, including those in $\boldsymbol{e}$. Any assignment of values to the variables in $\boldsymbol{W}$ consistent with $\boldsymbol{e}$ will be called an *extension*, *explanation* or *interpretation* of $\boldsymbol{e}$. Our problem is to find an extension $\boldsymbol{w}^*$ which maximizes the conditional probability $P(\boldsymbol{w}|\boldsymbol{e})$. In other words, $\boldsymbol{W} = \boldsymbol{w}^*$ is the *most probable explanation* (*MPE*) of the evidence at hand if

$$P(\boldsymbol{w}^*|\boldsymbol{e}) = \max_{\boldsymbol{w}} P(\boldsymbol{w}|\boldsymbol{e}) . \tag{40}$$

The task of finding $\boldsymbol{w}^*$ will be executed locally, by letting each variable X compute the function

$$\text{BEL}^*(x) = \max_{\boldsymbol{w}'_X} P(x, \boldsymbol{w}'_X|\boldsymbol{e}) , \tag{41}$$

where $\boldsymbol{W}'_X = \boldsymbol{W} - X$. Thus, BEL*$(x)$ stands for the probability of the most probable extension of $\boldsymbol{e}$ which is also consistent with the hypothetical assignment $X = x$. Unlike BEL(x) (see equation (1)), BEL*(x) need not sum to unity over x.

The propagation scheme presented below is based on the following philosophy: For every value x of a singleton variable X, there is a best extension of the complementary variables W'_X. Due to the many independence relationships embedded in the network, the problem of finding the best extension of $X = x$ can be decomposed into that of finding the best complementary extension to each of the neighboring variables, then using this information to choose the best value of X. This process of decomposition (resembling the principle of optimality in dynamic programming) can be applied recursively until, at the network's periphery, we meet evidence variables whose values are predetermined, and the process halts.

4.1. Deriving the propagation rules

We consider again the fragment of a singly connected network in Fig. 2 and denote by W^+_{XY} and W^-_{XY} the subset of variables contained in the respective subgraphs G^+_{XY} and G^-_{XY}. Removing any node X would partition the network into the subgraphs G^+_X and G^-_X containing two sets of variables, W^+_X and W^-_X, and (possibly) two sets of evidence, e^+_X and e^-_X, respectively.

Using this notation, we can write

$$P(w^*|e) = \max_{x, w^+_X, w^-_X} P(w^+_X, w^-_X, x | e^+_X, e^-_X) . \tag{42}$$

The conditional independence of W^+_X and W^-_X, given X, and the entailments $e^+_X \subseteq W^+_X$ and $e^-_X \subseteq W^-_X$ yield:

$$\begin{aligned} P(w^*|e) &= \max_{x, w^+_X, w^-_X} \frac{P(w^+_X, w^-_X, x)}{P(e^+_X, e^-_X)} \\ &= \alpha \max_{x, w^+_X, w^-_X} P(w^-_X | x) P(x | w^+_X) P(w^+_X) , \end{aligned} \tag{43}$$

where $\alpha = [P(e^+_X, e^-_X)]^{-1}$ is a constant, independent of the uninstantiated variables in W and would have no influence on the maximization in (43). From here on we will use the symbol α to represent any constant which need not be computed in practice, because it does not affect the choice of w^*.

Equation (43) can be rewritten as a maximum, over x, of two factors:

$$\begin{aligned} P(w^*|e) &= \alpha \max_x [\max_{w^-_X} P(w^-_X | x)] [\max_{w^+_X} P(x | w^+_X) P(w^+_X)] \\ &= \alpha \max_x \lambda^*(x) \pi^*(x) , \end{aligned} \tag{44}$$

where

$$\lambda^*(x) = \max_{w^-_X} P(w^-_X | x) , \tag{45}$$

$$\pi^*(x) = \max_{w_X^+} P(x, w_X^+) . \tag{46}$$

Thus, if, for each x, an oracle were to provide us the MPE values of the variables in W_X^-, together with the MPE values of the variables in W_X^+, we would be able to determine the best value of X by computing $\lambda^*(x)$ and $\pi^*(x)$ and, then, maximize their product, $\lambda^*(x)\pi^*(x)$.

We now express $\lambda^*(x)$ and $\pi^*(x)$ in such a way that they can be computed at node X from similar parameters available at X's neighbors. Writing

$$W_X^- = W_{XY}^- \cup W_{XZ}^- , \qquad W_X^+ = W_{UX}^+ \cup W_{VZ}^+ ,$$

$$W_{U'X}^+ = W_{UX}^+ - U , \qquad W_{V'X}^+ = W_{VX}^+ - U ,$$

we obtain

$$\lambda^*(x) = \max_{w_{XY}^-} P(w_{XY}^- | x) \max_{w_{XZ}^-} P(w_{XZ}^- | x) = \lambda_Y^*(x)\lambda_Z^*(x) \tag{47}$$

and

$$\pi^*(x) = \max_{u, v, w_U^+, w_V^+} [P(x|u, v)P(u, v, w_U^+, w_V^+)]$$

$$= \max_{u,v}[P(x|u, v) \max_{w_{U'X}^+} P(u, w_{U'X}^+) \max_{w_{V'X}^+} P(v, w_{V'X}^+)]$$

$$= \max_{u,v} P(x|u, v)\pi_X^*(u)\pi_X^*(v) , \tag{48}$$

where $\lambda_Y^*(x)$ (and, correspondingly, $\lambda_Z^*(x)$) can be regarded as a message that a child, Y, sends to its parent, X:

$$\lambda_Y^*(x) = \max_{w_{XY}^-} P(w_{XY}^- | x) . \tag{49}$$

Similarly,

$$\pi_X^*(u) = \max_{w_{U'X}^+} P(u, w_{U'X}^+) \tag{50}$$

can be regarded as a message that a parent U sends to its child X. Note the similarities between λ^* and π^* and λ and π in (2) and (3).

Clearly, if these λ^* and π^* messages are available to X, it can compute its best value x^* using (44)–(46). What we must show now is that, upon receiving these messages, it can send back to its neighbors the appropriate $\lambda_X^*(u)$, $\lambda_X^*(v)$, $\pi_Y^*(x)$ and $\pi_Z^*(x)$ messages, while preserving their probabilistic definitions according to (49) and (50).

Updating π^*

Rewriting (41) as

$$\mathrm{BEL}^*(x) = P(x, w_X^{+*}, w_X^{-*} | e) \tag{51}$$

and using (45)–(50), we have

$$\begin{aligned} \mathrm{BEL}^*(x) &= \alpha\lambda^*(x)\pi^*(x) \\ &= \alpha\lambda_Y^*(x)\lambda_Z^*(x) \max_{u,v} P(x|u, v)\pi_X^*(u)\pi_X^*(v) \,. \end{aligned} \tag{52}$$

Comparing this expression to the definition of $\pi_Y^*(x)$, we get

$$\begin{aligned} \pi_Y^*(x) &= \max_{w_{X'Y}^+} P(x, w_{X'Y}^+) = \max_{w_X^+, w_{XZ}^-} P(x, w_X^+, w_{XZ}^-) \\ &= \lambda_Z^*(x) \max_{u,v} P(x|u, v)\pi_X^*(u)\pi_X^*(v) \,. \end{aligned} \tag{53}$$

Alternatively, $\pi_Y^*(x)$ can be obtained from $\mathrm{BEL}^*(x)$ by setting $\lambda_Y^*(x) = 1$ for all x. Thus,

$$\pi_Y^*(x) = \alpha \, \mathrm{BEL}^*(x)|_{\lambda_Y^*(x)=1} = \alpha \, \frac{\mathrm{BEL}^*(x)}{\lambda_Y^*(x)} \,. \tag{54}$$

The division by $\lambda_Y^*(x)$ in (54) amounts to discounting the contribution of all variables in G_{XY}^-. Note that $\pi_Y^*(x)$, unlike $\pi_Y(x)$, need not sum to unity over x.

Updating λ^*

Starting with the definition

$$\lambda_X^*(u) = \max_{w_{UX}^-} P(w_{UX}^- | u) \tag{55}$$

we partition W_{UX}^- into its constituents

$$W_{UX}^- = X \cup W_{XY}^- \cup W_{XZ}^- \cup W_{V'X}^+ \cup V$$

and obtain

$$\begin{aligned} \lambda_X^*(u) &= \max_{x, w_{XY}^-, w_{XZ}^-, w_{V'X}^+, v} P(x, w_{XY}^-, w_{XZ}^-, v, w_{V'X}^+) \\ &= \max_{x, v, w_{XY}^-, w_{XZ}^+, w_{V'X}^+} P(w_{XY}^-, w_{XZ}^- | w_{V'X}^+, x, v, u) P(x, v, w_{V'X}^+ | u) \\ &= \max_{x,v} [\lambda_Y^*(x)\lambda_Z^*(x) P(x|u, v) \max_{w_{V'X}^+} P(v, w_{V'X}^+ | u)] \,. \end{aligned}$$

Finally, using the marginal independence of U and W^+_{VX}, we have

$$\lambda^*_X(u) = \max_{x,v}[\lambda^*_Y(x)\lambda^*_Z(x)P(x|u, v)\pi^*_X(v)] . \tag{56}$$

4.2. Summary of propagation rules

In general, if X has n parents, $U_1, U_2, \ldots, U_n$, and m children, $Y_1, Y_2, \ldots, Y_m$, node X receives the messages $\pi^*_X(u_i)$, $i = 1, \ldots, n$, from its parents and $\lambda^*_{Y_j}(x)$, $j = 1, \ldots, m$, from its children.

$\pi^*_X(u_i)$ stands for the probability of the most probable tail extension of the proposition $U_i = u_i$ relative the link $U_i \rightarrow X$. This subextension is sometimes called an "explanation," or a "causal argument."

$\lambda^*_{Y_j}(x)$ stands for the conditional probability of the most probable head extension of the proposition $X = x$ relative of the link $X \rightarrow Y_j$. This subextension is sometimes called a "prognosis" or a "forecast."

Using these $n + m$ messages together with the fixed probability $P(x|u_1, \ldots, u_n)$, X can identify its best value and further propagate these messages using the following three steps:

Step 1: *Updating* BEL*. When node X is activated to update its parameters, it simultaneously inspects the $\pi^*_X(u_i)$ and $\lambda^*_{Y_j}(x)$ messages communicated by each of its parents and children and forms the product

$$F(x, u_1, \ldots, u_n) = \prod_{j=1}^{m} \lambda^*_j(x)P(x|u_1, \ldots, u_n) \prod_{i=1}^{n} \pi^*_X(u_i) . \tag{57}$$

This F function enables X to compute its BEL*(x) function and, simultaneously, identify the best value x^* from the domain of X:

$$x^* = \max_x{}^{-1} \text{BEL}^*(x) , \tag{58}$$

where

$$\text{BEL}^*(x) = \alpha \max_{u_k:\, 1 \leq k \leq n} F(x, u_1, \ldots, u_n) \tag{59}$$

and α is a constant, independent of x, which need not be computed in practice.

Step 2: *Updating* λ^*. Using the F function computed in Step 1, node X computes the parent-bound messages by performing n vector maximizations, one for each parent:

$$\lambda^*_X(u_i) = \max_{x, u_k:\, k \neq i} [F(x, u_1, \ldots, u_n)/\pi^*_X(u_i)] , \quad i = 1, \ldots, n . \tag{60}$$

Step 3: *Updating* π^*. Using the BEL*(x) function computed in Step 1, node X computes the children-bound messages:

$$\pi^*_{Y_j}(x) = \alpha \frac{\text{BEL}^*(x)}{\lambda^*_j(x)} \tag{61}$$

and posts these on the links to $Y_1, \ldots, Y_m$.

These steps are identical to those governing belief updating, equations (4)–(6), with maximization replacing the summation. They can be viewed as tensor operations, using max for inner product, i.e., $\langle AB \rangle_{ik} = \max_j A_{ij}B_{jk}$ [4]; each outgoing message is computed by taking the max inner products of the tensor $P(x|u_1, \ldots, u_n)$ with all incoming messages posted on the other links.

The boundary conditions are identical to those of belief updating and are summarized below for completeness:

(1) *An anticipatory node* represents an uninstantiated variable with no successors. For such a node, X, we set $\lambda^*_{Y_j}(x) = (1, 1, \ldots, 1)$.

(2) *An evidence node* represents a variable with instantiated value. If variable X assumes the value x', we introduce a dummy child Z with

$$\lambda^*_Z(x) = \begin{cases} 1\,, & \text{if } x = x'\,, \\ 0\,, & \text{otherwise}\,. \end{cases}$$

This implies that, if X has children, $Y_1, \ldots, Y_m$, each child should receive the same message $\pi^*_{Y_j}(x) = \lambda^*_Z(x)$ from X.

(3) *A root node* represents a variable with no parents. For each root variable X, we introduce a dummy parent U, permanently instantiated to $U = 1$, and set the conditional probability on the link $U \rightarrow X$ equal to the prior probability of X, i.e., $P(x|u) = P(x)$.

These boundary conditions ensure that the messages defined in (49) and (50) retain their semantics on peripheral nodes.

4.3. Reaching equilibrium and assembling a composite solution

To prove that the propagation process terminates, consider a parallel and autonomous control scheme whereby each processor is activated whenever any of its incoming messages changes value. Note that, since the network is singly connected, every path must eventually end at either a root node having a single child or a leaf node having a single parent. Such single-port nodes act as absorption barriers; updating messages received through these ports get absorbed and do not cause subsequent updating of the outgoing messages. Thus, the effect of each new piece of evidence would subside in time proportional to the longest path in the network.

To prove that, at equilibrium, the selected x^* values do, indeed, represent the most likely interpretation of the evidence at hand, we can reason by induction on the depth of the underlying tree, taking an arbitrary node X as a root. The λ^* or π^* messages emanating from any leaf node of such a tree certainly comply with the definitions of (49) and (50). Assuming that the λ^* (or π^*) messages at any node of depth k of the tree comply with their intended definitions of equations (49) and (50), the derivation of (51)–(56) guarantees that they continue to comply at depth $k-1$, and so on. Finally, at the root node, $\alpha\mathrm{BEL}^*(x^*)$ actually coincides with $P(\boldsymbol{w}^*|\boldsymbol{e})$, as in (42), which means that $\mathrm{BEL}^*(x)$ computed from (59) must induce the same rating on x as does $\max_{\boldsymbol{w}'_X} P(x, \boldsymbol{w}'_X|\boldsymbol{e})$ (see (41)). This proves that each local choice of x^* is part of some optimal extension $\boldsymbol{w}^*$.

Had the choice of each x^* value been unique, this would also guarantee that the assembly of x^* values constitutes the (unique) most probable extension $\boldsymbol{w}^*$. However, when several assignments $X = x$ yield the same optimal $\mathrm{BEL}^*(x)$, a pointer system must be consulted to ensure that the ties are not broken arbitrarily but cohere with choices made at neighboring nodes.

For example, in the circuit of Fig. 3, had we assumed $q_1 = q_3 < \frac{1}{2}$, the optimal interpretations $\{X_1 = 1, X_3 = 0\}$ and $\{X_1 = 0, X_1 = 1\}$ would be equally meritorious, both yielding $\mathrm{BEL}^*(X_3 = 0) = \mathrm{BEL}^*(X_3 = 1) = q_3 p_1 = p_3 q_1$, as reflected in the $\pi^*(x_3)$ and $\lambda^*_{Y_3}(x_3)$ messages of Fig. 5(d). Breaking the tie arbitrarily might result in choosing a suboptimal extension $\{X_1 = 0, X_3 = 0\}$ or even an inconsistent one $\{X_1 = 1, X_3 = 1\}$. To enforce a selection of values within the *same* optimal extension, local pointers should be saved to mark the neighbor's values at which the maximization is achieved. (In singly connected networks the relevant neighborhood consists of parents, children and spouses, i.e., parents of common children [21].) For example, node Y_3 (Fig. 5(d)) should maintain a pointer from $Y_2 = 0$ to $X_3 = 1$ and another pointer from $Y_2 = 1$ to $X_2 = 0$, to indicate that these two value pairs are compatible members in the same optimal extension. These compatible combinations are found during the local maximization required for calculating $\lambda^*_{Y_3}(x_3)$, as in (56) or (61).

Having these pointers available at each node provides a simple mechanism for retrieving the overall optimal extension; we solve for x^* at some arbitrary node X and then recursively follow the pointers attached to x^*. Additionally, we can retrieve an optimal extension compatible with *any* instantiation (say second best) of some chosen variable X and, comparing the merits of several such extensions, the globally second-best explanation can be identified [12]. Another use of this mechanism is facilitating sensitivity analysis; to analyze the merit of testing an unknown variable, we can simply follow the links attached to each of its possible instantiations and examine its impact on other propositions in the system.

4.4. Comparison to belief updating

The propagation scheme described in this section bears many similarities to that used in belief updating (equations (4)–(6)). In both cases, coherent global equilibria are obtained by local computations in time proportional to the network's diameter. Additionally, the messages π^* and λ^* bear both formal and semantic similarities to their π and λ counterparts, and the local computations required for updating them involve, roughly, the same order of complexity.

It is instructive, however, to highlight the major differences in the two schemes. First, belief updating involves *summation*, whereas in belief revision, *maximization* is the dominant operation. Second, belief updating involves more absorption centers than belief revision. In the former, every anticipatory node acts as an absorption barrier in the sense that it does not permit the passage of messages between its parents. This is clearly shown in (5); substituting $\lambda_Y(x) = \lambda_Z(x) = 1$ yields $\lambda_X(u) = 1$, which means that evidence in favor of one parent (V) has no bearing on another parent (U) as long as their common child (X) receives no evidential support ($\lambda(x) = 1$). This matches our intuition about how frames should interact; data about one frame (e.g., seismic data indicating the occurrence of an earthquake) should not evoke a change of belief about another unrelated frame (say, the possibility of a burglary in my home) just because the two may give rise to a common consequence sometimes in the future (e.g., triggering the alarm system). This frame-to-frame isolation no longer holds for belief revision, as can be seen from (56). Setting $\lambda^*_Y(x) = \lambda^*_Z(x) = 1$ still renders $\lambda^*_X(u)$ sensitive to $\pi^*_X(v)$.

Such endless frame-to-frame propagation raises both psychological and computational issues. Psychologically, in an attempt to explain a given phenomenon, the mere mental act of imagining the likely consequences of the hypotheses at hand will activate other, remotely related, hypotheses just because the latter could also cause the imagined consequence. We simply *do not encounter* that mode of behavior in ordinary reasoning; in trying to explain the cause of a car accident, we do not interject the possibility of lung cancer just because the two (accidents and lung cancer) could lead to the same eventual consequence—death.

Computationally, it appears that, in large systems, the task of finding the most satisfactory explanation would require an excessive amount of computation; the propagation process would spread across loosely coupled frames until every variable in the system reexamines its selected value x^*.

These considerations, together with other epistemological issues (see Section 7), require that the set of variables, $\boldsymbol{w}$, over which P is maximized be *circumscribed* in advance to a privileged set called *explanation corpus*. In addition to the evidence $\boldsymbol{e}$, $\boldsymbol{W}$ should contain those variables only which both stand in clear causal relation to $\boldsymbol{e}$ (i.e., ancestors of $\boldsymbol{e}$) and have significant impact on pending decisions. For example, if Y_2 were the only observed variable in Fig. 3, then the explanation corpus would consist of $\boldsymbol{W} = \{X_1, X_2, Y_2\}$, excluding X_3 and Y_3. If, in addition, X_1 and X_2 were outputs of two complex digital circuits and our only concern were to find out whether any of these circuits should be replaced, then the ancestors of X_1 and X_2, too, should be excluded from $\boldsymbol{W}$. In other words, if there is no practical utility in finding which particular gate in those circuits is faulty, then it would be both wasteful and erroneous to enter these ancestors into $\boldsymbol{W}$ (see Section 7).

Circumscribing an explanation corpus partitions the variables in the system into two groups, $\boldsymbol{W}$ and its complement $\boldsymbol{W}'$. The computation of $\boldsymbol{w}^*$ now involves mixed operations; maximization over $\boldsymbol{W}$ and summation over $\boldsymbol{W}'$:

$$P(\boldsymbol{w}^*|\boldsymbol{e}) = \max_{\boldsymbol{w}} P(\boldsymbol{w}|\boldsymbol{e}) = \max_{\boldsymbol{w}} \sum_{\boldsymbol{w}'} P(\boldsymbol{w}|\boldsymbol{w}', \boldsymbol{e}) P(\boldsymbol{w}'|\boldsymbol{e}) .$$

The propagation rules, likewise, should be mixed; variables in $\boldsymbol{W}$ should follow the revision rules of (57)–(61), while those in $\boldsymbol{W}'$ the updating rules of (5)–(6). The interaction between the λ^* and π^* messages produced by the former and the λ and π messages produced by the latter should conform to their probabilistic semantics and will not be elaborated here.

5. Coping with Loops

Loops are undirected cycles in the underlying network, i.e., the Bayesian network without the arrows. When loops are present, the network is no longer singly connected, and local propagation schemes invariably run into trouble. The two major methods for handling loops while still retaining some of the flavor of local computation are: *clustering* and *conditioning* (also called *assumption-based* reasoning).

5.1. Clustering methods

Clustering involves forming compound variables in such a way that the topology of the resulting network is singly connected. For example, if in the network of Fig. 1 we define the compound variables,

$$Y_1 = \{X_1, X_2\} , \qquad Y_2 = \{X_2, X_3\} ,$$

the following tree ensues: $X_4 \leftarrow Y_1 \rightarrow Y_2 \rightarrow X_5 \rightarrow X_6$. In the network of Fig. 7 (Section 6) defining the variables $D_{234} = \{D_2, D_3, D_4\}$ and $M_{123} = \{M_1, M_2, M_3\}$, we obtain a singly connected network of the form:

$$D_1 \rightarrow M_{123} \leftarrow D_{234} \rightarrow M_4 .$$

A popular method of selecting clusters is to form *clique-trees* [3, 30, 32]. If the clusters are allowed to overlap each other until they cover all the links of the original network, then the interdependencies between any two clusters are mediated solely by the variables which they share. If we further insist that these clusters grow until their interdependencies form a tree structure (called a *join tree* [3]) then the tree-propagation scheme of Section 4 will be applicable. For example, in the network of Fig. 1, if we define $Z_1 = \{X_1, X_2, X_4\}$, $Z_2 = \{X_1, X_2, X_3\}$, $Z_3 = \{X_2, X_3, X_5\}$ and $Z_4 = \{X_5, X_6\}$, the dependencies among the Z variables will be described by the chain,

$$\begin{tikzcd} Z_1 \arrow[r, dash, "\{X_1,X_2\}"] & Z_2 \arrow[r, dash, "\{X_2,X_3\}"] & Z_3 \arrow[r, dash, "\{X_5\}"] & Z_4 , \end{tikzcd}$$

where the X symbols on the links identify the set of elementary variables common to any pair of adjacent Z clusters.

These clustered networks can be easily processed with the propagation techniques of Section 4, except that the multiplicity of each compound variable increases exponentially with the number of elementary variables it contains. Consequently, the size of either the link matrices or the messages transmitted may become prohibitively large.

An extreme case of clustering would be to represent all ancestors of the observed findings by *one* compound variable. For example, if X_6 and X_4 are the observed variables in the network of Fig. 1, we can define the compound variable $Z = \{X_1, X_2, X_3, X_5\}$ and obtain the tree $X_4 \leftarrow Z \rightarrow X_6$. Assigning a definite value to the compound variable Z would constitute an explanation for the findings observed. Indeed, this is the approach taken by Cooper [5] and Peng and Reggia [24]. To search for the best explanation through the vast domain of possible values associated with the explanation variable, admissible heuristic strategies had to be devised, similar to that of the A* algorithm [15]. Yet the complexity of these algorithms is still exponential [20] since they do not exploit the interdependencies among the variables in Z. Another disadvantage of this technique is the loss of conceptual flavor; the optimization procedure does not reflect familiar mental processes and, consequently, it is hard to construct meaningful arguments to defend the final conclusions.

5.2. The method of conditioning (reasoning by assumptions)

Conditioning is an attempt to both reduce complexity by exploiting the structural independencies embodied in the network and preserve, as much as possible, the conceptual nature of the interpretation process. This is accomplished by performing the major portion of the optimization using local computations *at the knowledge level itself*, i.e., using the links provided by the network as communication channels between simple, autonomous and semantically related processors.

The basic idea behind conditioning can be illustrated using Fig. 1. It is based on our ability to change the connectivity of a network and render it singly connected by instantiating a selected group of variables, called a cycle cutset. For example, instantiating node X_1 to some value would block all pathways through X_1 and would render the rest of the network singly connected, amiable to the propagation technique of Section 4. Thus, if we wish to find the most likely interpretation of some evidence $\boldsymbol{e}$, say $\boldsymbol{e} = \{X_6 = 1\}$, we first assume $X_1 = 0$ (as in Fig. 6(a)), propagate λ^* and π^* to find the best interpretation, I_0, under this assumption, repeat the propagation to find the best interpretation, I_1, under the assumption $X_1 = 1$ (as in Fig. 6(b)) and, finally, compare the two interpretations and choose the one with the highest probability. For example, if I_0 and I_1 are realized by the vectors

$$\begin{aligned} I_0 &= (X_1 = 0, x_2^0, x_3^0, x_4^0, x_5^0)\,, \\ I_1 &= (X_1 = 1, x_2^1, x_3^1, x_4^1, x_5^1)\,, \end{aligned} \tag{62}$$

then the best overall interpretation is determined by comparing the two products

$$\begin{aligned} P(I_0|\boldsymbol{e}) = {} & \alpha P(X_6 = 1|x_5^0)P(x_5^0|x_2^0, x_3^0)P(x_4^0|X_1 = 0, x_2^0) \\ & \cdot P(x_3^0|X_1 = 0)P(x_2^0|X_1 = 0)P(X_1 = 0)\,, \\ P(I_1|\boldsymbol{e}) = {} & \alpha P(X_6 = 1|x_5^1)P(x_5^1|x_2^1, x_3^1)P(x_4^1|X_1 = 1, x_2^1) \\ & \cdot P(x_3^1|X_1 = 1)P(x_2^1|X_1 = 1)P(X_1 = 1)\,, \end{aligned}$$

where $\alpha = [P(\boldsymbol{e})]^{-1}$ is a constant. Since, all the factors in the products above are available from the initial specification of the link's probabilities, the comparison can be conducted by simple computations.

Such globally supervised comparisons of products are the basic computational steps used in the diagnostic method of Peng and Reggia [24]. However, we use them to compare only two candidates from the space of 2^5 possible

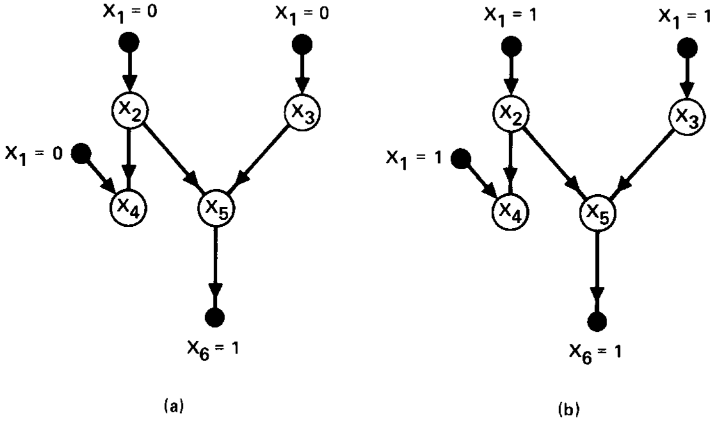

FIG. 6. Instantiating variable X_1 renders the network of Fig. 1 singly connected.

value combinations. Most of the interpretation work was conducted by local propagation, selecting the appropriate match for each of the two assumptions $X_1 = 0$ and $X_1 = 1$. Thus, we see that, even in multiply connected networks, local propagation provides computationally effective and conceptually meaningful method of trimming the space of interpretations down to a manageable size.

The effectiveness of conditioning depends heavily on the topological properties of the network. In general, a set of several nodes (a cycle cutset) must be instantiated before the network becomes singly connected. This means that 2^c candidate interpretations will be generated by local propagation, where c is the size of the cycle cutset chosen for conditioning. Since each propagation phase takes only time linear with the number of variables in the system (n), the overall complexity of the optimal interpretation problem is exponential with the size of the cycle cutset that we can identify. If the network is sparse, topological considerations can be used to find a small cycle cutset and render the interpretation task tractable. Although the problem of finding the minimal cycle cutset is NP hard, simple heuristics exist for finding close-to-minimal sets [18]. Identical complexity considerations apply to the task of belief updating [22], so finding the globally best explanation is no more complex than finding the degree of belief for any individual proposition.

A third method of sidestepping the loop problem is that of stochastic simulation [23]. It amounts to generating a random population of scenarios agreeing with the evidence, then selecting the most probable scenario from that population. This is accomplished distributedly by having each processor inspect the current state of its neighbors, compute the belief distribution of its host variable, then randomly select one value from the computed distribution. The most likely interpretation is then found by identifying either the global state which has been selected most frequently or the one possessing the highest probability (computed by taking the product of n conditional probabilities).

It is important to note that the difficulties associated with the presence of loops are not unique to probabilistic formulations but are inherent to any problem where globally defined solutions are produced by local computations, be it probabilistic, deterministic, logical, numerical or hybrids thereof. Identical computational issues arise in Dempster–Shafer's formalism [29], constraint-satisfaction problems [7], truth maintenance systems [9], diagnostic reasoning [6], database management [3], matrix inversion [31], distributed optimization [10] and logical deduction.

The importance of network representation, though, is that it uncovers the core of these difficulties, and provides a unifying abstraction that encourages the exchange of solution strategies across domains. The cycle-cutset conditioning method, for example, has been used successfully in nonprobabilistic circuit diagnostics [11] and for improving the efficiency of backtracking in constraint-satisfaction problems [8].

6. A Medical Diagnosis Example

6.1. The model

To illustrate the mechanics of the propagation scheme described in Section 4, let us consider the diagnosis network of Fig. 7 (after Peng and Reggia [24]), where the nodes at the top row, $\{D_1, D_2, D_3, D_4\}$, represent four hypothetical diseases and the nodes at the bottom row, $\{M_1, M_2, M_3, M_4\}$, four manifestations (or symptoms) of these diseases. The parameters c_{ij}, shown on the links of Fig. 7, represent the strength of causal connection between disease D_i and symptom M_j,

$$c_{ij} = P(M_j \text{ observed} \mid \text{only } D_j \text{ present}) . \tag{63}$$

All four diseases are assumed to be independent and their prior probabilities, $\pi_i = P(D_i = \text{TRUE})$, are shown in Fig. 7. When several diseases give rise to the same symptom, their combined effect is assumed to be of the "noisy OR-gate" type [21], i.e.,

(1) a symptom can be triggered only if at least one of its causes is present (mandatory causation),

(2) the mechanism capable of masking a symptom in the presence of one disease is assumed to be independent of that masking it in the presence of another (exception independence).

Given this causal model, we imagine a patient showing symptoms $\{M_1, M_3\}$ but *not* $\{M_2, M_4\}$. Our task is to find that disease *combination* which best explains the observed findings, namely, to find a TRUE–FALSE assignment to variables $\{D_1, D_2, D_3, D_4\}$ which constitutes the most probable extension of the evidence

$$e = \{M_1 = \text{TRUE}, M_2 = \text{FALSE}, M_3 = \text{TRUE}, M_4 = \text{FALSE}\} .$$

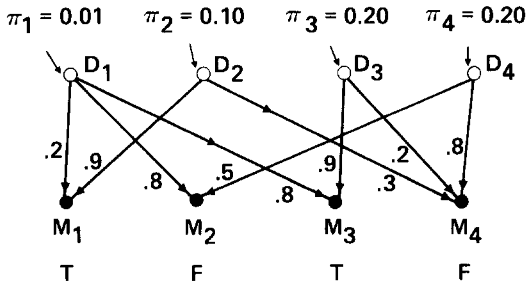

FIG. 7. Network representing causal relations between four diseases and four manifestations. The link parameters, c_{ij}, measure the strength of causal connection.

Let D_i and M_j denote the propositional variables associated with disease D_i and manifestation M_j, respectively; each may assume a TRUE or FALSE value. Additionally, for each propositional variable X, we let $+x$ and $\neg x$ denote the propositions $X = \text{TRUE}$ and $X = \text{FALSE}$, respectively. Thus, for example,

$$P(\neg m_j | + d_j) = P(M_j = \text{FALSE} | D_i = \text{TRUE})$$

would stand for the probability that a patient definitely having disease D_i will *not* develop symptom M_j.

Let X stand for some manifestation variable and $\{U_1, \ldots, U_n\}$ its parents set. The OR-gate interaction assumed above permits us to construct the combined parents-child relationship $P(x|u_1, \ldots, u_n)$ from the individual parent-child relations parametrized by c_{iX} (equation (63)). If I_T stands for the set of (indices of) parents with value TRUE,

$$I_\text{T} = \{i: U_i = \text{TRUE}\}\,, \tag{64}$$

then X is FALSE iff all its TRUE parents simultaneously fail to trigger the manifestation corresponding to X. Thus,

$$P(\neg x|u_1, \ldots, u_n) = \prod_{i \in I_\text{T}} q_{iX} \tag{65}$$

and

$$P(+x|u_1, \ldots, u_n) = 1 - \prod_{i \in I_\text{T}} q_{iX}\,, \tag{66}$$

where

$$q_{iX} = 1 - c_{iX}\,. \tag{67}$$

Substitution in (57), the function $F(x, u_1, \ldots, u_n)$ obtains the form:

$$F(+x, u_1, \ldots, u_n) = \left[1 - \prod_{i \in I_\text{T}} q_{iX}\right] \prod_{j=1}^{m} \lambda^*_{Y_j}(+x) \prod_{i=1}^{n} \pi^*_X(u_i)\,, \tag{68}$$

$$F(\neg x, u_1, \ldots, u_n) = \prod_{i \in I_\text{T}} q_{iX} \prod_{j=1}^{m} \lambda^*_{Y_j}(\neg x) \prod_{i=1}^{n} \pi^*_X(u_i)\,. \tag{69}$$

These product forms would permit the calculation of the π^* and λ^* messages according to (58)–(61). In particular, for every negatively instantiated symptom node X we have

$$\frac{\lambda^*_X(+u_i)}{\lambda^*_X(\neg u_i)} = q_{iX} \tag{70}$$

independently of $\pi^*_X(u_k)$, $k \neq i$. For every disease node X, setting $P(x|u_1, \ldots, u_n)$ to the prior probability $\pi(x)$, (53) yields

$$\pi^*_{Y_j}(x) = \pi(x) \prod_{k \neq j} \lambda^*_{Y_k}(x) . \tag{71}$$

6.2. Message propagation

For convenience, let us adopt the following notation:

$$\lambda_{ji} = \lambda^*_{M_j}(+d_i)/\lambda^*_{M_j}(\neg d_i) , \tag{72}$$

$$\pi_{ij} = \pi^*_{M_j}(+d_i)/\pi^*_{M_j}(\neg d_i) . \tag{73}$$

The network in Fig. 7 becomes singly connected upon instantiating D_1. We shall first instantiate D_1 to TRUE, find its best extension, then repeat the process under the assumption $D_1 =$ FALSE. Figure 8(a) shows the network's message-passing topology, together with the initial messages posted by the instantiated variables $\{+d_1, e\}$:

$$\lambda_{12} = \frac{(1 - q_{11}q_{12})}{(1 - q_{11})} = \frac{(1 - 0.8 \cdot 0.1)}{(1 - 0.8)} = 4.600 ,$$

$$\lambda_{33} = \frac{(1 - q_{13}q_{33})}{(1 - q_{13})} = \frac{(1 - 0.2 \cdot 0.1)}{(1 - 0.2)} = 1.225 ,$$

$$\lambda_{24} = q_{42} = 0.5 , \qquad \lambda_{43} = q_{34} = 0.8 ,$$

$$\lambda_{44} = q_{44} = 0.2 , \qquad \lambda_{42} = q_{24} = 0.7 .$$

The last four values are direct consequence of (70).

At the second phase, each D_i processor inspects the λ^* messages posted on its links and performs the operation specified in (71). This leads to the message distribution shown in Fig. 8(b), with:

$$\pi_{24} = \frac{\pi_2\lambda_{12}}{(1 - \pi_2)} = 0.510 , \qquad \pi_{21} = \frac{\pi_2\lambda_{42}}{(1 - \pi_2)} = 0.077 ,$$

$$\pi_{34} = \frac{\pi_3\lambda_{33}}{(1 - \pi_3)} = 0.305 , \qquad \pi_{33} = \frac{\pi_3\lambda_{43}}{(1 - \pi_3)} = 0.200 ,$$

$$\pi_{44} = \frac{\pi_4\lambda_{24}}{(1 - \pi_4)} = 0.125 , \qquad \pi_{42} = \frac{\pi_4\lambda_{44}}{(1 - \pi_4)} = 0.050 .$$

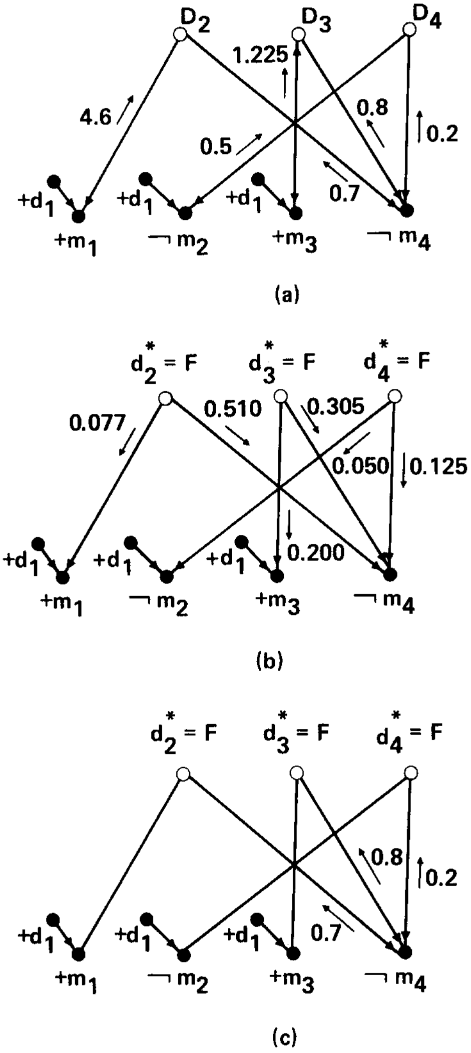

FIG. 8. (a) λ^* messages after instantiating D_1 and all four symptoms. (b) π^* messages after activating all D nodes. (c) λ^* messages after activating M_4; the best explanation is $d_2^* = d_3^* = d_4^* =$ FALSE.

The x^* value chosen by each of the D_i processors is FALSE (see (58)) because, for each $i = 2, 3, 4$, we have

$$\frac{\mathrm{BEL}^*(+d_i)}{\mathrm{BEL}^*(\neg d_i)} = \prod_{j=1}^{4} \lambda_{ji} \frac{\pi_i}{1-\pi_i} < \tfrac{1}{2} .$$

For example, processor D_2 receives: $\lambda_{12} = 4.6$, $\lambda_{42} = 0.7$; so,

$$\frac{\mathrm{BEL}^*(+d_2)}{\mathrm{BEL}^*(\neg d_2)} = \frac{\lambda_{12} \cdot \lambda_{42} \cdot \pi_2}{1 \cdot 1 \cdot (1 - \pi_2)} = \frac{4.6 \cdot 0.7 \cdot 0.1}{1 \cdot 1 \cdot 0.9} = 0.358 < \tfrac{1}{2} . \tag{74}$$

The messages π_{21}, π_{33} and π_{42} will eventually get absorbed at node D_1, while π_{24}, π_{34} and π_{44} are now posted on the ports entering node M_4. Again, since M_4 is instantiated to $\neg m_4$, the λ^* messages generated by M_4 on the next activation phase remain unchanged (Fig. 8(c)), and the process halts with the current $\boldsymbol{w}^*$ values: $D_2 = D_3 = D_4 = \mathrm{FALSE}$.

Let us now retract the assumption $D_1 = \mathrm{TRUE}$ and posit the converse: $D_1 = \mathrm{FALSE}$. This results in the messages $\pi_{11} = \pi_{12} = \pi_{13} = \infty$ being posted on all those links emanating from node D_1 which get translated to $\lambda_{12} = \infty$, $\lambda_{13} = \infty$ and $\lambda_{24} = q_{42} = 0.5$. This means that D_2 and D_3 will switch simultaneously and permanently to state TRUE while D_4, by virtue of

$$\frac{\mathrm{BEL}^*(+d_4)}{\mathrm{BEL}^*(\neg d_4)} = \frac{\lambda_{24} \cdot \lambda_{44} \cdot \pi_4}{(1 - \pi_4)} = \frac{0.50 \cdot 0.20 \cdot 0.20}{0.80} = 0.025 < \tfrac{1}{2} ,$$

tentatively remains at the state FALSE, as illustrated in Fig. 9(a).

During the next activation phase (Fig. 9(b)), D_2 and D_3 post the messages $\pi_{24} = \pi_{34} = \infty$, which M_4 inspects for possible updating of λ_{44}. However, these new messages will not cause any change in λ_{44} because, according to (60) and (65), the ratio λ_{44} remains

$$\lambda_{44} = \frac{P(\neg m_4 | + d_4, d_2, d_3)}{P(\neg m_4 | \neg d_4, d_2, d_3)} = q_{44} ,$$

independently of π_{24} and π_{34} .

Thus, under the current premise $\neg d_1$, the best interpretation of the observed symptoms is $\{+d_2, +d_3, \neg d_4\}$, which is to be expected, in view of the network topology.

6.3. Choosing the best interpretation

We see that the assumption $+d_1$ yields the interpretation $\{\neg d_2, \neg d_3, \neg d_4\}$, while $\neg d_1$ yields $\{+d_2, +d_3, \neg d_4\}$. The question now is to decide which of the two interpretations is more plausible or, in other words, which has the highest posterior probability given the evidence $\boldsymbol{e} = \{+m_1, \neg m_2, +m_3, \neg m_4\}$ at hand. A direct way to decide between the two candidates is to calculate the two posterior probabilities, $P(I^+ | \boldsymbol{e})$ and $P(I^- | \boldsymbol{e})$, where

$$I^+ = \{+d_1, \neg d_2, \neg d_3, \neg d_4\} \quad \text{and} \quad I^- = \{\neg d_1, +d_2, +d_3, \neg d_4\} .$$

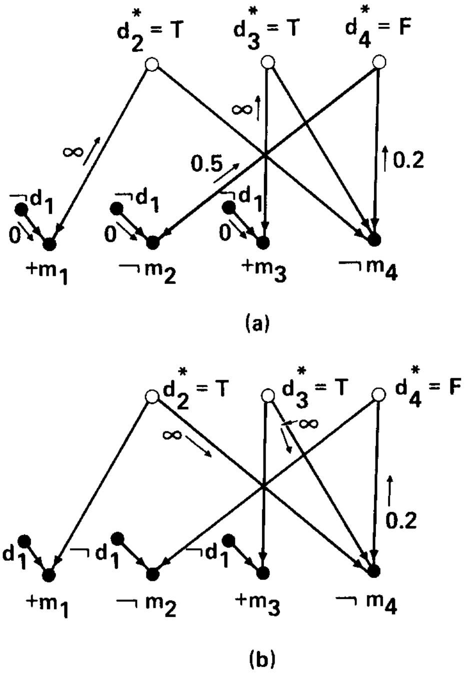

FIG. 9. (a) Message profile after instantiating D_1 to FALSE; the best explanation switches to $\{d_2^* = d_3^* = \text{TRUE},\ d_4^* = \text{FALSE}\}$. (b) Message profile after activating M_4; the best explanation remains $\{d_2^* = d_3^* = \text{TRUE},\ d_4^* = \text{FALSE}\}$.

These calculations are quite simple, because instantiating the D variables *separates* the M variables from each other, so that the posterior probabilities involve only products of $P(m_j \mid \text{parents of } M_j)$ over the individual symptoms and a product of the prior probabilities over the individual diseases. For example,

$$
\begin{aligned}
P(I^+ \mid e) &= \alpha P(I^+) P(e \mid I^+) \\
&= \alpha \pi_1 (1-\pi_2)(1-\pi_3)(1-\pi_4)(1-q_{11}) q_{12} (1-q_{13}) \\
&= \alpha \cdot 0.01 \cdot 0.90 \cdot 0.80 \cdot 0.80 \cdot 0.20 \cdot 0.90 \cdot 0.80 \\
&= \alpha \cdot 8.2944 \cdot 10^{-4}\,, \\
P(I^- \mid e) &= \alpha P(I^-) P(e \mid I^-) \\
&= \alpha (1-\pi_1) \pi_2 \pi_3 (1-\pi_4)(1-q_{21})(1-q_{33}) q_{24} q_{34} \\
&= \alpha \cdot 0.99 \cdot 0.10 \cdot 0.20 \cdot 0.80 \cdot 0.90 \cdot 0.90 \cdot 0.70 \cdot 0.80 \\
&= \alpha \cdot 7.18 \cdot 10^{-3}\,.
\end{aligned}
$$

Since $\alpha = [P(\boldsymbol{e})]^{-1}$ is a constant, we conclude that I^- is the most plausible interpretation of the evidence $\boldsymbol{e}$.

6.4. Generating explanations

The propagation pattern of the λ^* and π^* messages can also be instrumental in mechanically generating verbal explanations. When belief in a certain proposition is supported (or undermined) from several directions, the π^* and λ^* messages can be consulted to determine the factors most influential in the current selection of x^*. Tracing the most influential π^* and λ^* messages back to the generating evidence would yield a skeleton subgraph from which verbal explanation can be structured. For example, the messages loading the graphs of Figs. 9(a) and (b) should be summarized by:

> Since we have ruled out disease D_1, the only possible explanation for observing symptoms M_1 and M_3 is that the patient suffers, simultaneously, from D_2 and D_3. The fact that M_2 and M_4 both came out negative indicates that disease D_4 is absent. Moreover, even if M_4 were positive, it would be completely explained away by D_2 and D_3.

The last sentence is a result of running a hypothetical positive instantiation of M_4 and realizing that, due to the strong (∞) π^* messages from D_2 and D_3, M_4 cannot deliver a λ_{42} high enough to switch D_4 to TRUE.

Conflicting evidence is identified by the presence of strongly supportive and strongly opposing messages, simultaneously impinging on the same proposition. For example, the proposition $D_2 = \text{TRUE}$ in Fig. 8(a) receives a strong support from $\lambda^*_{12} = 4.6$ and a strong denial from $\pi_2 = 0.2$. The two balance each other out and yield $\text{BEL}^*(+d_2)$ very close to $\text{BEL}^*(\neg d_2)$ (see (74)). The following explanation would then be appropriate:

> Although symptom M_1 strongly suggests D_2, it is partly explained by D_1 (which we assumed TRUE) and, in view of the rarity of D_2, this patient probably does not suffer from D_2.

6.5. Reversibility versus perseverance

It is interesting to note that there is a definite threshold value for π_1, $\pi_1 = 0.0804$, at which the two interpretations, I^+ and I^-, are equiprobable. That means that, as evidence in favor of $+d_1$ accumulates and π_1 increases beyond the value 0.0804, the system will switch abruptly from interpretation I^- to interpretation I^+. This abrupt "change of view" is a collective phenomenon, characteristic of massively parallel systems, and is reminiscent of the way people's beliefs undergo complete reversal in response to a minor clue. Note, though, that the transition is *reversible*, i.e., as π_1 decreases, the system will switch back to the I^- interpretation at exactly the same threshold value,

$\pi_1 = 0.0804$. No hysteresis occurs because, although the computations are done locally, w^* is globally optimal and is, therefore, a unique function of all systems' parameters.

This reversibility differs from human behavior in that, once we commit our belief to a particular interpretation, it often takes more convincing evidence to make us change our mind than the evidence which got us there in the first place. Simply discrediting a piece of evidence would not, in itself, make us abandon the beliefs which that evidence induced [13, 28]. The phenomenon is very pronounced in perceptual tasks; once we adopt one view of Necker's cube or an Escher sketch, it takes a real effort to break ourselves loose and adopt alternative interpretations. Irreversibility (or hysteresis) of that kind is characteristic of systems with local feedback, similar to the one responsible for magnetic hysteresis in metals. If the magnetic spin of one atom heads north, it sets up a magnetic field which encourages its neighbors to follow suit; when the neighbors' spins eventually turn north, they generate a magnetic field which further "locks" the original atom in its north-pointing orientation.

The hysteresis characteristic of human belief revision may have several sources. One possibility is that local feedback loops are triggered between evoked neighboring concepts; e.g., if I suspect fire, I expect smoke, and that very expectation of smoke reinforces my suspicion of fire—as if I actually saw smoke. This is a rather unlikely possibility because it would mean that even in simple cases (e.g., the fire and smoke example), people are likely to confuse internal thinking with genuine evidence. A more reasonable explanation is that, by and large, the message-passing process used is feedback-free and resembles that of Section 4, where the π^* and λ^* on the same link are orthogonal to each other. However, in complex situations, where loops are rampant, people simply cannot afford the overhead computations required by conditioning or clustering. As an approximation, then, they delegate the optimization task to local processes and continue to pass messages as if the belief network were singly connected. The resultant interpretation, under these conditions, is locally, not globally, optimal, and this yields irreversible belief revision.

Another source of belief perseverance may lie in the difficulty of keeping track of all justifications of ones beliefs and tracing them back to all evidence, past and present, upon which beliefs are founded [13]. For computational reasons people simply forget the evidence and retain its conclusion. More formally, propositional networks such as those treated in this paper, are not maintained as stable mental constructs but, rather, are created and destroyed dynamically, to meet pragmatic needs. For example, connections may be formed for the immediate purpose of explaining some strange piece of evidence or for supporting a hypothesis of high immediate importance. Once the evidence imparts its impact onto other propositions, we tend to break the mediating connection, forget the evidence itself and retain only the conclusion. When that evidence is later discredited, the connection to the induced conclusions is no longer in vivid memory while the discrediting information, in itself, may not be perceived to be of sufficient pragmatic importance for reestablishing old connections.

7. Discussion

7.1. Accepting versus assessing beliefs

The method described in this paper constitutes a bridge between probabilistic reasoning and nonmonotonic logics. Like the latter, the method provides systematic rules that lead from a set of factual sentences (the evidence) to a set of conclusion sentences (the accepted beliefs) in a way that need not be truth-preserving. For example, in Fig. 3, we start with the sentence "all inputs are ON," we obtain the sentence "$Y_3 = \text{OFF}$, $X_2 = \text{ON}$," and we output the conclusion "$X_1 = \text{OFF}$, $X_3 = \text{ON}$." True, the medium through which these inferences are made is probabilistic (e.g., the assumption $q_1 > q_3$ was critical for the conclusion) but the input-output pairs are categorical. Seeking the most probable extension parallels the default-logic aim of minimizing abnormal assumptions and this paper shows how and when the minimization can be realized by local computations.

Unlike the scheme expounded in this paper, the dominant "logicist" paradigm has been the formalization of belief revision as direct logical relationships between evidence and conclusions, unmediated by numerical measures of belief. The entire notion of "degree of belief" plays only a minor role in these endeavors and likelihood judgments are often regarded as secondary byproducts of symbolic manipulations on categorical knowledge bases.

Many philosophers of science, especially those studying inductive logic, have taken a different tack. They hold that the bulk of human knowledge is probabilistic in nature, i.e., there is a set Q of confirmation functions which measure the degree to which statements are confirmed by the evidence at hand. Some statements, however, enjoy a special status—they are accepted as true in almost every respect, except for the fact that they can have this status revoked at a later time, perhaps in light of new evidence. This corpus of statements, K, is called "accepted beliefs" and the essential test testifying its formation is that an agent accepting K would behave as though all statements in K were "practically certain," for example, behavior predicated on any accepted statement will not be different if more evidence were to support that statement [19].

In this view, accepted beliefs can be regarded as local and temporary crystallization in a continuous fluid of partial beliefs. Belief revision is viewed as the rules that govern the dynamics of this crystallization process, namely, the condition under which a given statement would be promoted to the privileged membership in the "acceptance" corpus.

Philosophers disagree over what constitutes a "good," rational rule of acceptance. At first glance it appears that knowing the confirmation functions in Q would, in itself, be sufficient for defining acceptance rules. This, however, turns out not to be the case. The obvious rule of acceptance is the high-probability or "thresholding" rule: Accept a statement h iff $P(h|e) > 1 - \varepsilon$, for some small ε. The problem is that, for any nonzero ε, this rule leads to knowledge bases that are grossly inconsistent. This came to be known as the "lottery paradox" [16]: A large number of people buy tickets to a lottery having a single winner. The probability that the ith person wins the lottery is clearly very small and, by thresholding, should lead us to accept the statement "person i is not the winner," for every i. Yet, this collection of statements is inconsistent with the given fact that one person is definitely destined to be the winner. Many other acceptance rules have been proposed but none seem to satisfy both our criterion that behavior should remain invariant to evidence confirming an accepted statement and our desire that the acceptance corpus, to some extent, be deductively closed and consistent.

Levi [17] argues that rules of acceptance cannot be formulated on the basis of confirmation functions alone, but must take into account pragmatic considerations as well, namely, what is going to be done with the statement once it is accepted. An extreme example for the importance of pragmatics can be found in betting behavior. No matter how sure a person is in the truth of a (factual) statement, if sufficient heavy penalties are imposed on wrong decisions, that person is bound to show hesitation acting in accordance with his/her beliefs.

Harsanyi [14] and Loui [19] include computational considerations within the pragmatics of belief acceptance. The crystallization of partial beliefs into crisp corpus of logical statements has computational advantages which overshadow the incurred loss of details. An obvious advantage is the economy gained in both storage and communication. A more subtle advantage is the utilization of beliefs in inferential schemata. In many reasoning tasks, use is made of prestored schemata that turned successful in the past. These schemata need to be matched to distinct classes of situations, e.g., the antecedent part of any decision rule identifies the situations to which the action part is applicable. Commitment to a categorical set of beliefs facilitates an efficient symbolic encoding of the classes of situations to which the inferential schemata are applicable.

Pragmatic considerations of this sort help explain the vast disparity between AI and the management sciences, in their treatment of uncertainty. The reason that the management sciences have embraced probabilistic approaches and have emphasized *measures* of beliefs and uncertainty is that the domain of management decisions involves a wide spectrum of critical payoffs and penalties. Thus, even very unlikely events cannot be ignored off-hand but must be brought into comparison against the more likely (and moderately paidoff) events.

In AI applications, on the other hand, the variability of the payoffs is often rather narrow and the number of decisions enormous, so, even just *likely* events can be treated as a sure thing. For example, John's walk towards the cupboard reaching for the box of cereal is an action involving no high risk; the cost of failure is a meager exertion of a few extra steps. Had the stakes been higher, John could have embarked on lengthier deliberation prior to taking the action. For example, recalling Mary's breakfast he could have possibly assessed the chances that the cereal is finished. However, given the noncritical nature of the actions involved, there is no pressing need for such deliberation, John can safely proceed toward the cupboard without considering *all* relevant evidence, i.e., refraining from propagating the evidence to the entire belief network. When contradictory evidence arrives, some beliefs switch abruptly from "almost surely true" to "almost surely false," apparently skipping the phase of numerical evaluation. In summary, common everyday activities are characterized by firmly held beliefs because violated expectations involve relatively mild risks, there is a definite computational advantage for accepting these expectations as firm beliefs and there is no practical danger of even letting some of these beliefs turn inconsistent.

7.2. Is a most probable explanation adequate?

The most probable explanation (MPE) criterion used in this paper reflects the following acceptance rule:

> A statement h is accepted iff h is entailed by I^*, where I^* is a conjunction of primitive sentences forming the most probable explanation of the available evidence.

An equivalent acceptance rule is:

> Out of all world models consistent with the evidence, choose the one with the highest overall probability.

In the case of Kyburg's lottery, for example, the set of consistent world models consists of all choices of a single winner from the population of ticket holders. If the lottery is absolutely fair, all models are equally likely and no acceptance can be invoked. However, assuming that one person is known to have a higher chance of winning than the rest (say by virtue of possessing a larger number of tickets), we fully commit our entire belief to the world model in which that person is the one and only winner.

Like every acceptance rule based solely on confirmation, this, too, has its drawbacks. For example, if I were asked to bet $1,000 on a would-be winner, I would resist endorsing even the one holding dozens of tickets. Indeed, if payoff information is available, the MPE criterion loses its viability. It should give way, then, to maximum-utility or minimum-risk type of alternatives. However, when the payoffs are either unknown or insignificant, the MPE criterion offers a reasonable compromise. It is not uncommon for people to adopt the following hypotheses based on rather tenuous evidence:

> *h*: I am going to be the winner, because I feel I have a slightly better chance than anybody else.

or,

> *h*: Did you say my uncle bought a ticket too, the bastard? I bet he is going to be the winner, he's been just damned lucky all his life.

Similarly, it is not uncommon for people to switch abruptly from one interpretation of Necker's cube to the opposite, as a result of a slight change in visual clues.

There are two situations where the MPE criterion is justifiable even on pragmatic grounds:

(1) If the difference between the best and the second-best explanation is appreciable.

(2) When one is forced to choose a definite, terminal action and the risks associated with wrong choices are all equal.

For example, in answering multiple-choice examinations, the student's best strategy is to select the answer most likely to be correct. Similarly, technicians engaged in troubleshooting electronic circuits would do well if, at every phase of the diagnosis, they replace or test the unit most likely to account for the faults observed. The truth maintenance strategy of resolving contradictions by retracting the minimum number of assumptions is also a variant of the MPE criterion, where all assumptions are assumed to be equally probable.

When it comes to scientific or causal explanations, the MPE criterion carries a special weight of yielding a *neutral* explanation, i.e. unbiased by pragmatic considerations such as payoffs and risks. Still, even in purely scientific settings, an explanation is always translated into some action and actions always lead to consequences and payoffs. It is desirable, therefore, that the most likely explanation not be issued in isolation, but be accompanied with additional information such as: the absolute probability of the best explanation, the probability of the second-best explanations and how likely are these measures to change in the face of new, pending tests. Having an efficient way of propagating beliefs in causal networks is a necessary and, often, sufficient step toward computing these auxiliary measures, [12].

7.3. Circumscribing explanations

So far we have assumed that every consistent instantiation of the variables in the system constitutes an *explanation* of the evidence and, consequently, the optimization was conducted over all nonevidence variables in the network. Unfortunately, this leads to both computational and conceptual difficulties.

Computationally, this means that one needs to propagate the impact of every piece of evidence to the entire network and, since unconfirmed consequences no longer serve as absorption centers (as they do in belief updating, see Section

4), we are running the risk of spending valuable resources in totally irrelevant sections of one's knowledge base. Conceptually, this unchecked frame-to-frame propagation might lead to paradoxical results.

Suppose I am very concerned about having a certain fatal disease and, by a strike of good luck, the results of a medical test reveal that there is an 80% chance that I am totally healthy. According to the MPE acceptance rule, I should commit all my belief to a world model in which I am healthy. So, I start imagining all kinds of possible scenarios associated with my newly adopted belief. For example, I imagine 10 mutually exclusive scenarios, $S_1, S_2, \ldots, S_{10}$ (e.g. trip to Bahama, trip to Afghanistan, . . .), each having a probability 0.1 of getting realized. Now I repeat the MPE exercise, but this time on a larger world scale, consisting of the earlier facts about the disease, plus the newly imagined scenarios. Lo and behold, any world model in which I am healthy now receives only an 8% chance of getting realized so, the most probable "explanation" of the evidence is that I *do* suffer from that horrible disease, and all for being a hasty daydreamer!

In everyday discourse we would exclude such scenarios from being part of the explanation because they do not stand in causal relation to observed evidence, i.e., the symptoms or the test outcome. This is indeed a form of circumscription which is easily implementable in belief networks by insisting that an "explanation" of evidence e should consist only of ancestors of e, all other variables excluded. Thus, the maximization exercise should range only over the set of variables that are ancestors of some observed facts.

This circumscription, however, does not go far enough as it does not insist that the relation between e and its causal ancestors pass some test of strength or rationality. The spectrum of everyday observations is so rich that with a little stretch of the imagination one can always proclaim any proposition h to be supported by some observation e', however feeble the support. Thus, at least in theory, every proposition would qualify for admittance into the explanation corpus, and we are back where we started.

The solution to this dilemma relies, again, on the notion of payoffs. In every practical situation, when we seek an explanation for some experience e, we have a fairly good idea which collections of hypotheses should be included in that explanation, namely, what set of variables should be subjected to optimization. This is determined by pragmatic considerations: we know which set of hypotheses would influence those consequences that stand at the center of our concern. For example, partitioning the hypothesis "I am healthy" into ten specific scenarios would be a rational thing to do if I were pressed by my travel agent to make payment on a plane ticket and must decide on a specific destination. It would not be rational if my main concern were knowing whether I am sick or not.

In order to apply the MPE scheme described in this paper, one must assume that the network contains only "interesting" partitions of world models;

"interesting" being relevant to a set of consequences at the center of one's practical concerns. For example, if disease H_1 has ten (equally likely and mutually exclusive) prognoses while a competing hypothesis H_2 has only one clear prognosis, should we partition H_1 prior to maximization, or should we maximize over the space H_1 versus H_2, then, if H_1 turns out to be accepted, maximize again over its components? That depends on the circumstances: if the ten prognoses call for only minor variations in treatment we should do the latter, else, if each calls for drastically different action and every error can be devastating, we should do the former. (Harsanyi [14] calls it "the principle of small disparate risks.")

In conclusion, one can justify the MPE acceptance rule only relative to small worlds, precompiled by a fairly astute decision maker. Such compilation involves preanalytical judgments to decide the tradeoff between the gain in simplicity and loss of utility associated with carving and delimiting these worlds.

8. Conclusions

This paper develops a distributed scheme for finding the most probable composite explanation of a body of evidence.

We show that, in singly connected networks, globally optimal explanations can be configured by local and autonomous message-passing processes, similar to those used in belief updating; conceptually related propositions communicate with each other via a simple protocol, and the process converges to the correct solution in time proportional to the network diameter. In multiply connected networks, the propagation method must be assisted either by clustering (i.e., locally supervised groups of variables) or conditioning (i.e. reasoning by assumptions); each exploiting different aspects of the network topology.

The implications of these results are several. First, from a psycho-philosophical viewpoint, they provide a clear demonstration of how cognitive constructs exhibiting global coherence can be assembled by local, neuron-like processors without external supervision. Second, along a more practical dimension, the message-passing method developed offers substantial reduction in complexity compared with previous optimization techniques, achieving linear complexity in singly connected networks and $\exp(c)$ in general networks, where c is the size of the cycle cutset. This is accomplished by subtask decomposition, supervised solely by the network topology. Third, the paper establishes a clear paradigmatic link between probabilistic and logical formalisms of nonmonotonic reasoning. It demonstrates how numerical probabilities can be used as a concealed inferencing fuel for performing coherent transformations between evidence sentences and conclusion sentences. It also identifies the kind of structures where such transformations can be executed by autonomous production rules and those that invite problems of intractability and/or instability, unless treated with care.

ACKNOWLEDGMENT

I thank Hector Geffner for a thorough review of the manuscript, for discovering a flaw in an earlier "proof" of optimality and for adopting and expanding the proposed scheme to perform circuit diagnosis [12]. A large portion of Section 7 is based on [19] and has crystallized to its current form after many discussions with Ron Loui. Glen Shafer and Richard Duda have offered many suggestions that helped improve the manuscript.

REFERENCES

1. Ben Bassat, M., Carlson, R.W., Puri, V.K., Lipnick, E., Portigal, L.D. and Weil, M.H., Pattern-based interactive diagnosis of multiple disorders: The MEDAS system, *IEEE Trans. Pattern Anal. Mach. Intell.* **2** (1980) 148–160.
2. Ben Bassat, M., Multimembership and multiperspective classification: Introduction, applications and a Bayesian model, *IEEE Trans. Syst. Man Cybern.* **10** (1980) 331–336.
3. Beeri, C, Fagin, R., Maier D. and Yannakakis, M., On the desirability of acyclic database schemes, *J. ACM* **30** (1983) 479–513.
4. Booker, L., Personal communication, 1987.
5. Cooper, G.F., NESTOR: A computer-based medical diagnostic aid that integrates causal and probabilistic knowledge, Ph.D. Dissertation, Department of Computer Science, Stanford University, Stanford, CA, 1984.
6. de Kleer, J. and Williams, B.C., Reasoning about multiple faults, in: *Proceedings AAAI-86*, Philadelphia, PA (1986) 132–139.
7. Dechter, R. and Pearl, J., The anatomy of easy problems: A constraint-satisfaction formulation, in: *Proceedings IJCAI-85*, Los Angeles, CA (1985) 1066–1072.
8. Dechter, R. and Pearl, J., The cycle-cutset method for improving search performance in AI applications, in: *Proceedings Third IEEE Conference on AI Applications*, Orlando, FL (1987) 224–230.
9. Doyle, J., A truth maintenance system, *Artificial Intelligence* **12**(3) (1979) 231–272.
10. Gafni, E. and Barbosa, V.C., Optimal snapshots and the maximum flow in precedence graphs, *Proceedings 24th Allerton Conference*, 1986.
11. Geffner, H. and Pearl, J., An improved constraint propagation algorithm for diagnosis, Tech. Rept. R-73. UCLA, CS Department, Cognitive System Laboratory, Los Angeles, CA, 1986; also in: *Proceedings IJCAI-87*, Milan, Italy, 1987.
12. Geffner, H. and Pearl, J., A distributed approach to diagnosis, Tech. Rept. R-66, UCLA, CS Department, Cognitive Systems Laboratory, Los Angeles, CA, short version in: *Proceedings Third IEEE Conference on AI Applications*, Orlando, FL (1987) 156–162.
13. Harman, G., *Change in View* (MIT Press, Cambridge, MA, 1986).
14. Harsanyi, J.C., Acceptance of empirical statements: A Bayesian theory without cognitive utilities, *Theor. Decis.* **18** (1985) 1–30.
15. Hart, P.E., Nilsson, J.J. and Raphael, B., A formal basis for the heuristic determination of minimum cost paths, *IEEE Trans. Syst. Sci. Cybern.* **4** (1968) 100–107.
16. Kyburg, H.E., Jr., *Probability and the Logic of Rational Belief* (Weslyan University Press, Middletown, CT, 1961).
17. Levi, I., *The Enterprise of Knowledge* (MIT Press, Cambridge, MA, 1980).

18. Levy, H. and Low, D.W., A new algorithm for finding small cycle cutsets, Rept. G 320-2721, IBM Los Angeles Scientific Center, Los Angeles, CA, 1983.

19. Loui, R., Real rules of inference: Acceptance and non-monotonicity in AI, TR 191, Department of Computer Science, University of Rochester, Rochester, NY, 1986.

20. Pearl, J., *Heuristics: Intelligent Search Strategies for Computer Problem Solving* (Addison-Wesley, Reading, MA, 1984).

21. Pearl, J., Fusion, propagation and structuring in belief networks, *Artificial Intelligence* **29** (1986) 241–288.

22. Pearl, J., A constraint-propagation approach to probabilistic reasoning, in: L.N. Kanal and J. Lemmer (Eds.), *Uncertainty in Artificial Intelligence* (North-Holland, Amsterdam, 1986) 357–370.

23. Pearl, J., Evidential reasoning using stochastic simulation of causal models, *Artificial Intelligence* **32** (1987) 245–257.

24. Peng, Y. and Reggia, J., Plausibility of diagnostic hypotheses, in: *Proceedings AAAI-86*, Philadelphia, PA (1986) 140–145.

25. Pople, H., Heuristic methods for imposing structures on ill-structured problems, in P. Solovits (Ed.), *AI in Medicine* (Westview Press, Boulder, CO, 1982).

26. Reggia, J.A., Nau, D.S. and Wang, Y., Diagnostic expert systems based on a set-covering model, *Int. J. Man-Mach. Stud.* **19** (1983) 437–460.

27. Reiter, R., A theory of diagnosis from first principles, *Artificial Intelligence* **32** (1987) 57–95.

28. Ross, L. and Anderson, C.A., Shortcomings in the attribution process: On the origins and maintenance of erroneous social assessments, in: D. Kahneman, P. Slovic, A. Tversky (Eds.) *Judgement under Certainty: Heuristics and Biases* (Cambridge University Press Cambridge, 1982) 129–152; 331–336.

29. Shenoy, P. and Shafer, G., Propagating belief functions with local computations, *IEEE Expert* **1** (3) (1986) 43–52.

30. Spiegelhalter, D.J., Probabilistic reasoning in predictive expert systems, in: L.N. Kanal and J. Lemmer (Eds.), *Uncertainty in Artificial Intelligence* (North-Holland, Amsterdam, 1986) 47–68.

31. Tarjan, R.E., Graph theory and Gaussian elimination, in: J.R. Bunch and D.J. Rose (Eds.), *Sparse Matrix Computations* (Academic Press, New York, 1976) 3–22.

32. Tarjan, R.E. and Yannakakis, M., Simple linear-time algorithms to test chordality of graphs, test acyclicity of hypergraphs, and selectively reduce acyclic hypergraphs, *SIAM J. Comput.* **13** (1984) 566–579.

Received July 1986; revised version received February 1987