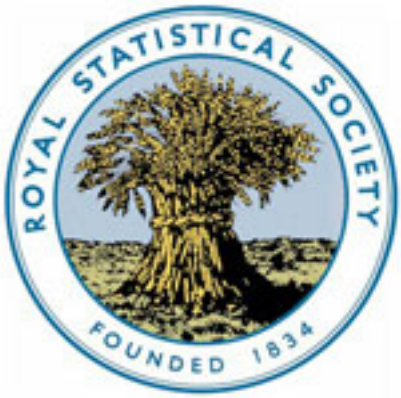


J. R. Statist. Soc. B (1988)
50, *No. 2, pp.* 157–224

Local Computations with Probabilities on Graphical Structures and their Application to Expert Systems

By S. L. LAURITZEN and D. J. SPIEGELHALTER†

Aalborg University, Denmark *MRC Biostatistics Unit, Cambridge, UK*

[*Read before the* Royal Statistical Society *at a meeting organised by the* Research Section *on Wednesday, January 13th, 1988,* Professor A. P. Dawid *in the Chair*]

SUMMARY

A causal network is used in a number of areas as a depiction of patterns of 'influence' among sets of variables. In expert systems it is common to perform 'inference' by means of local computations on such large but sparse networks. In general, non-probabilistic methods are used to handle uncertainty when propagating the effects of evidence, and it has appeared that exact probabilistic methods are not computationally feasible. Motivated by an application in electromyography, we counter this claim by exploiting a range of local representations for the joint probability distribution, combined with topological changes to the original network termed 'marrying' and 'filling-in'. The resulting structure allows efficient algorithms for transfer between representations, providing rapid absorption and propagation of evidence. The scheme is first illustrated on a small, fictitious but challenging example, and the underlying theory and computational aspects are then discussed.

Keywords: ARTIFICIAL INTELLIGENCE; BAYESIAN METHODS; CAUSAL MARKOV RANDOM FIELD; DECOMPOSABLE GRAPHS; EXPERT SYSTEMS; LOCAL POTENTIALS; MARKOV RANDOM FIELD; MAXIMUM CARDINALITY SEARCH; PROBABILISTIC REASONING; TRIANGULATED GRAPHS

1. INTRODUCTION

The label 'expert system' is, broadly speaking, given to a computer program intended to make reasoned judgements or give assistance in a complex area in which human skills are fallible or scarce: this paper, in common with many others, will use medicine as a domain of application. A general characteristic of such systems is the explicit separation of the 'knowledge-base', summarising assumptions about the domain in a flexible and modular format, from the data available on a new patient, with an 'inference-engine' acting as a control mechanism applying the knowledge to the new case. A knowledge-base will contain a structure of propositions, or facts, related in one of a number of forms, such as rules, frames or networks. Each structure hinges on local relationships between entities, enabling graphical displays of the knowledge base to be used in communications with the user. For a general review, see Barr and Feigenbaum (1981).

In many domains, such as computerising a consistent set of legislation, the analysis of a new case may proceed in entirely logical steps. However, areas such as medicine are characterised by both the knowledge and the data on a new patient being incomplete, the relationships being inexact, and terms being imprecisely defined. This paper can be seen as an attempt to exploit maximally the local nature of the expressed relationships in order to handle this pervading uncertainty in a coherent probabilistic manner.

† *Address for correspondence*: MRC Biostatistics Unit, 5 Shaftesbury Road, Cambridge CB2 2BW, UK.

 0035–9246/88/50157 $2.00

We shall not attempt to review the long, and sometimes acrimonious, debate as to whether probability theory is an appropriate tool in this context; see, for example, Szolovits and Pauker (1978), Spiegelhalter and Knill-Jones (1984), Cheeseman (1985), Spiegelhalter (1986a), Henrion (1987) and, in particular, the articles in Kanal and Lemmer (1986) for a wide range of diverging opinions. Very briefly, there are four main schools of thought. The 'logical' model follows the tenets of artificial intelligence most closely in using only symbolic reasoning and avoiding numerical assessments (Cohen, 1985; Fox, 1986). The 'linguistic' model uses fuzzy reasoning to quantify the extent to which the imprecise statements used in common language match formally defined propositions (Zadeh, 1983, 1986). A 'legal' model uses Shafer–Dempster belief functions to construct arguments for interval-valued beliefs based on evidence whose reliability is given a numerical assessment (Shafer, 1976, 1987).

Finally the 'statistical/engineering' model adheres to the probability calculus, justified both from a theoretical perspective (Lindley, 1982, 1987) and from the pragmatic claim that it alone provides flexible and operational means of assessment, criticism and learning (Cheeseman, 1985; Spiegelhalter, 1987). Pearl (1986a) also argues for probabilistic structuring in expert systems as providing a good model for human understanding and memory.

However, in most implementations of expert systems that acknowledge uncertainty, none of these four paradigms is rigorously adopted. Instead, the somewhat informal numerical schemes used in early, now classical applications—such as 'certainty factors' in MYCIN (Buchanan and Shortliffe, 1984), quasi-probabilistic calculus in PROSPECTOR (Duda *et al.*, 1977), causal weights in CASNET (Weiss *et al.*, 1978) and frequency weights and evoking weights in INTERNIST (Miller *et al.*, 1982)—have been built into 'shells' for general expert system development (and hence widely adopted. Hajek (1985) has provided a theory for reasonable properties for 'degrees of certainty' attached to rules of the type 'If a is true, then b is true (with certainty x)'; these concern *propagation* (with a further rule 'If b then c (with certainty y)', how should knowing a to be true affect our belief in c?) and *combination* (with another rule 'If d then b (with certainty z)'). Probabilistic interpretations have been given to such degrees of certainty (Heckerman, 1986; Hajek, 1985) but there appears to be considerable under-specification (how can one propagate without knowing the effect of b being false?; what is the relationship between a and d?) and the view that ill-defined numerical measures could be described as 'currently fashionable *ad-hoc* quantitative mumbo-jumbo' (Smith, 1984) may be held by many.

In this paper we shall not attempt a comparison of alternative approaches for handling uncertainty, but will explore how far exact probabilistic manipulations can deal with the important challenges raised by the objectives of expert systems.

Much of our interest in this area has been motivated by the problems posed by the MUNIN system for diagnosis and test planning in electromyography (EMG) (Section 2) which, in common with many other medical expert systems, uses a 'causal network' representation to express clinical knowledge in a graphical form from which qualitative dependency relationships may be read. MUNIN stands out, however, in attempting to use strict probabilistic reasoning, and in Section 3 we list a number of issues raised in this and similar applications. These are related to problems in probability manipulations on complex graphical structures such as those used in genetics, and we briefly summarise our own and previous approaches.

For illustrative purposes, a simple fictitious example is introduced in Section 4,

and the steps of our argument are traced in a non-formal manner in Section 5. It is intended that a grasp of our methods can be obtained from Sections 1 to 5 alone, while Sections 6 to 9 provide the formal graph-theoretic and probabilistic underpinning of the methods. In Section 10, we briefly discuss aspects of implementation of the methods including some considerations of the computational complexity, and emphasise the suitability of the methods for programming in an object-oriented environment. Perspectives for future development are given in Section 11.

Our emphasis throughout is on efficient computational schemes for exact probabilistic manipulations in large complex networks defined by local relationships, and some limitations of this paper need to be acknowledged. Firstly, we make no claim that all concerns with 'inference' in expert systems and artificial intelligence can be dealt with in the structure we describe. Much work in artificial intelligence is concerned with maximally exploiting logical 'knowledge' using symbolic reasoning techniques, and explicitly handling, for example, temporal, spatial, structural and taxonomic relationships with respective rules for inference. It is an open question to what extent such qualitative structure could be handled in a probabilistic framework with good explanation facilities.

Secondly, our methods only cover one aspect of the standard statistical process of iterative model development and criticism. That is, we assume a fixed model is currently being entertained and the numerical assessments are precisely specified, and our objective is to draw conclusions valid within this current structure, without any claim that the model is 'true'. Some references concerning initial structuring are given in the next section, and in Section 11 we outline our current views on the vital issue of criticism and refinement of both structure and quantitative assessment as data accumulate. These problems are bound to be the subject of considerable future research.

Our interpretation of probabilities is that of a subjectivist Bayesian (Savage, 1972; de Finetti, 1974; Lindley, 1982). This seems a convenient and appropriate view in an area concerned with the rational structuring and manipulation of opinion, and the subjectivist objectives of a coherent system of probabilities representing belief in verifiable propositions, successively updated on the basis of available evidence, appears to fit remarkably the objectives of expert systems research. However, many of the techniques presented here are appropriate in disciplines where graphical structures are used and a frequentist interpretation is more appropriate, such as in complex pedigree analysis in genetics.

2. MUNIN—AN EXAMPLE OF A CAUSAL NETWORK

A recent leading article in the *New England Journal of Medicine* (Schwartz *et al.*, 1987) describes the growing use of causal networks to encapsulate medical knowledge, where the nodes of the network represent clinical entities that may take on one of a number of values, and a directed link between two nodes represents a causal relationship. Initial structuring of such networks has been discussed by Pople (1982), Kuipers and Kassirer (1983) and Patil *et al.* (1983) for example.

EMG concerns the diagnosis of neurological disease through analysis of bioelectrical signals from muscle and nerve tissue. Andersen *et al.* (1986) argue that much of the knowledge necessary to carry out an EMG examination can be structured as a causal network, and as part of an ESPRIT project a prototype expert system MUNIN

(MUscle and Nerve Inference Network) has been developed at Nordjysk Udviklingscenter in Aalborg, Denmark. Fig. 1 shows a 25-node representation of a single muscle as it appears to the user of the system (apart from the arrows on the links). The left-hand node is the true disease, assumed to take on one of the 11 states shown. The disease may lead to distinct intermediate pathophysiological disorders, which in turn may influence 15 actual EMG investigations. The values each investigation or clinical state may take on are shown in the box for each node, and the pre-measurement beliefs represented as a histogram. Two continuous measurements are included, and many of the nodes may take on the value 'normal', while 'other' allows for patients not falling within the spectrum for which the system is currently designed (see Andreassen *et al.* (1987)). As the findings are obtained, the histograms change dynamically throughout the network, providing attractive display of the revised beliefs. We emphasise that the final system will comprise many such muscle and nerve networks.

From a probabilistic perspective, the basis for the representation of beliefs within MUNIN is the identification of the causal network (Fig. 1) with the construction of a joint probability distribution over all 25 variables (nodes) from local elements. Specifically, we term a node a 'parent' if it has a directed link away from it, and a 'child' if it has a link coming into it. The construction of MUNIN required the specification, for each node v, of the probability of each of its states occurring given all combinations of values its parents could take on. For the 'DISEASE' node, with no parents, this is simply a prior distribution on the 11 states shown in Fig. 1, but for the 'FORCE' node, say, this table comprises $6 \times 5 \times 9 = 270$ values (see Andreassen *et al.* (1987) for details of the assessment procedure). Having specified this conditional probability table for each node, we then make the crucial assumption (Kiiveri *et al.*, 1984) that our joint probability for a particular set of 25 states equals the product over the entries in the 25 conditional probability tables that feature the appropriate states.

From a qualitative point of view, this represents our judgement that if we know the values of parents of a node v whose value is currently unknown, then no other knowledge (except concerning descendants of v) will influence our opinion concerning the true value of v (Kiiveri *et al.*, 1984); this reference, as well as Pearl (1986a, b), provides discussion of additional independence relations that can be read off a causal graph.

The use of a causal network as our starting point may initially appear unduly restrictive and the appropriate means of dealing with feedback mechanisms needs to be explored. However, we should emphasise that 'causality' has a broad interpretation as any natural ordering in which knowledge of a parent influences opinion concerning a child—this influence could be logical, physical, temporal or simply conceptual in that it may be most appropriate to think of the probability of children given parents. Similar diagrams have occurred in 'path analysis' (Wright, 1921, 1934), causal econometric and social models (Wold, 1954; Blalock, 1971; Jöreskog, 1973), as 'influence diagrams' in decision analysis (Shachter, 1986, 1987; Smith, 1987), as 'recursive models' in contingency table analysis (Wermuth and Lauritzen, 1983) as 'Bayes networks' in artificial intelligence (Pearl, 1986a), and as pedigrees in genetics. Among medical applications, such structures appear particularly appropriate in causality assessment in adverse drug reactions (Spiegelhalter, 1986c) and in models underlying cancer screening using epidemiological data and past screening history.

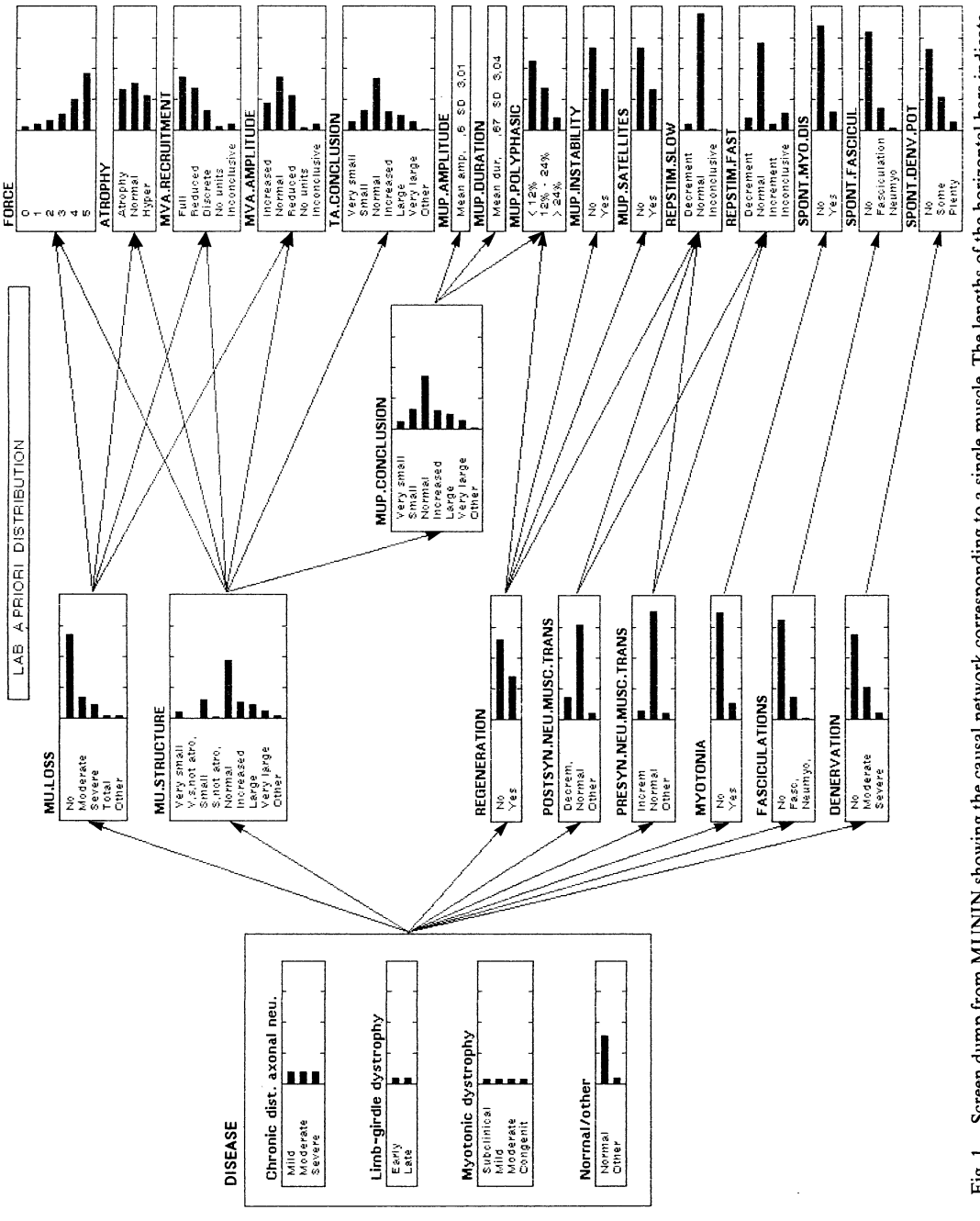


Fig. 1. Screen dump from MUNIN showing the causal network corresponding to a single muscle. The lengths of the horizontal bars indicate the prior probabilities of the possible states of each node. The nodes on the far right correspond to EMG findings, those in the middle layer to pathophysiological disorders. The disease node is to the far left. Direction of causality is from left to right in the figure.

3. ISSUES IN THE ANALYSIS OF NETWORKS

We have assumed that our knowledge is expressed as a causal network with appropriate conditional probability tables attached to each node. We now consider six reasonable requirements of a system such as MUNIN for which efficient solutions will be provided. It should be kept in mind that while some of the necessary probability manipulations may seem trivial on a small scale, efficient algorithms are required to handle envisaged systems of hundreds or thousands of nodes.

1. *Initialisation*: from the conditional probability tables we need to be able to generate an internal representation of our beliefs from which the marginal distributions on individual nodes may be easily obtained and displayed (Fig. 1), and which allow the operations listed below to take place. There should be no problem with 'logical' links, in which a probability is either 0 or 1 conditional on the values of the parents.

2. *Absorption of evidence*: in common with other clinical situations, a number of EMG findings arrive simultaneously. For reasons of efficiency, we wish to be able to absorb such evidence in a simple fashion, so that its effect may later be propagated through the graph in combination, rather than in single items. The use of strict probabilistic methods automatically ensures that the effect of multiple pieces of evidence does not depend on their order of arrival.

3. *Global propagation*: when required, we need to be able to propagate the effects of received evidence through the network, coherently revising our beliefs in the nodes that are still not established. We need to be able to cope with multiple parents, and 'loops' due, for example, to two alternative explanations for a disease causing a finding. Propagation in MUNIN, for example, needs to go 'up' the network for diagnostic purposes, (i.e. against the direction of causality) then 'down' again to revise beliefs in tests that are not yet done.

4. *Hypothesising and propagating single items of evidence*: we need to be able to condition on a node taking on a particular value, and rapidly observe its effect throughout the network.

5. *Planning*: usually some nodes are of particular interest, say the true disease, and we need to be able efficiently to assess the informational value in eliciting the response to nodes corresponding to potentially obtainable data.

6. *Influential findings*: after data are in, we need to be able to 'retract' their effect in order to identify particularly influential items. This is related to non-monotonic reasoning (Doyle, 1979).

The issues 1 to 6 are thus concerned with fast calculation of conditional and marginal probabilities on variables denoted by nodes on a graph. The basis for our approach is to exploit a number of different representations of a joint probability distribution, each representation being *local* in the sense defined by the topology of the graph. Fast algorithms for transferring between the representations allow efficient means of calculating conditional and marginal distributions for specified parts of large, sparse networks.

Others have made use of local computations for probability manipulations on directed graphs. Kelly and Barclay (1973) and Pearl (1982) consider 'trees' in which each node has at most one parent and there are no 'loops' (i.e. removal of any edge disconnects the graph). Kim and Pearl (1983) and Pearl (1986a) describe an elegant scheme for 'generalised Chow trees' in which multiple parents are permitted, and revision of beliefs is achieved by keeping track of 'messages' (essentially Bayes factors) being passed both up and down edges; Jensen *et al.* (1987) describe an implementation of this method in MUNIN, which requires a combination of nodes in order to remove loops. When loops are unavoidable, Pearl (1986a) recommends conditioning on each value combination of nodes that will decompose the graph into trees and then averaging over the results obtained from each of the propagations, or restructuring into a tree by introducing intermediate states that 'explain' dependencies. These concepts all have parallels in genetics, and Pearl's first suggestion is generally adopted to cope with intermarriage.

Shachter (1986, 1987) allows logical links and loops, and provides an algorithm for 'partial propagation' from a specified set of observed nodes to a set of nodes of interest. The topology of the original directed graph is adapted in a somewhat complex sequence of node removals and arc reversals and additions, which need to be repeated for each subsequent propagation.

Spiegelhalter (1986b, 1987) introduced the notion of an initial change to an 'undirected' graphical representation by adding additional 'dummy' edges for the internal machine representation allowing the joint distribution to be expressed in terms of marginals on small sets of nodes, although we shall see that his algorithm may add unnecessary edges. Goldman and Rivest (1986) suggest a similar change in the topology of the graph, but neither proposal deals with logical links. We note that a similar 'local' approach has been taken by Shafer *et al.* (1986) and Kong (1986) in handling belief functions on trees and general networks, respectively. Kong (1986) recognises the computational importance of having a triangulated graph, and presents a scheme for conditioning and marginalising using Dempster's rule of combination. See also Dempster and Kong (1986).

4. SIMPLE EXAMPLE

It is not feasible to step through our procedure with as complex an example as MUNIN so, for illustrative purposes, we shall consider a small piece of fictitious qualitative medical 'knowledge':

> Shortness-of-breath (dyspnoea) may be due to tuberculosis, lung cancer or bronchitis, or none of them, or more than one of them. A recent visit to Asia increases the chances of tuberculosis, while smoking is known to be a risk factor for both lung cancer and bronchitis. The results of a single chest X-ray do not discriminate between lung cancer and tuberculosis, as neither does the presence or absence of dyspnoea.

This needs to be applied to the following hypothetical situation. A patient presents at a chest clinic with dyspnoea, and has recently visited Asia. Smoking history and chest X-ray are not yet available. The doctor would like to know the chance that

each of the diseases is present, and if tuberculosis were ruled out by another test, how would that change the belief in lung cancer? Also, would knowing smoking history or getting an X-ray contribute most information about cancer, given that smoking may 'explain away' the dyspnoea since bronchitis is considered a possibility? Finally, when all information is in, can we identify which was the most influential in forming our judgement?

We acknowledge the stylised nature of this example, but it is contrived to illustrate as many issues as possible in a small problem. In particular, the event 'the patient presented to the clinic' is implicitly conditioned upon, and perhaps should be explicitly represented.

The structure of our knowledge-base is represented by the directed graph in Fig. 2. Note that the graph contains logical links: in this case to 'either τ or λ'. This is an economical way, in terms of storage space, of expressing our judgement that ξ and δ do not discriminate between τ and λ.

Our quantitative knowledge comprises numerical assessments of the probability of each state conditional on all possible parent states. Assessments are given in Table 1, representing a fictitious population coming to a chest clinic. Our notation uses a to indicate a positive response on the node α 'visit to Asia?', $\bar{a}$ to indicate a negative response, and $p(a)$ to stand for $\Pr(\alpha = a)$. Similarly, t stands for the presence of 'tuberculosis'; s, 'smoker'; l, 'lung cancer'; b, 'bronchitis'; e, 'lung cancer or bronchitis'; x, 'positive X-ray'; and d, 'dyspnoea'. The probability tables for negative responses may be derived from Table 1; for example, $p(\bar{d} \mid e, \bar{b}) = 0.30$. Fourteen probabilities require assessment, showing 'visit to Asia' as a fivefold risk factor for tuberculosis, 'smoking' as tenfold risk for lung cancer, and twofold for bronchitis, and the X-ray having 98% sensitivity and 95% specificity. (We again acknowledge that more realistic assessments may be made without first conditioning on arrival at the clinic.)

The graph expresses our fundamental assumption that the joint distribution

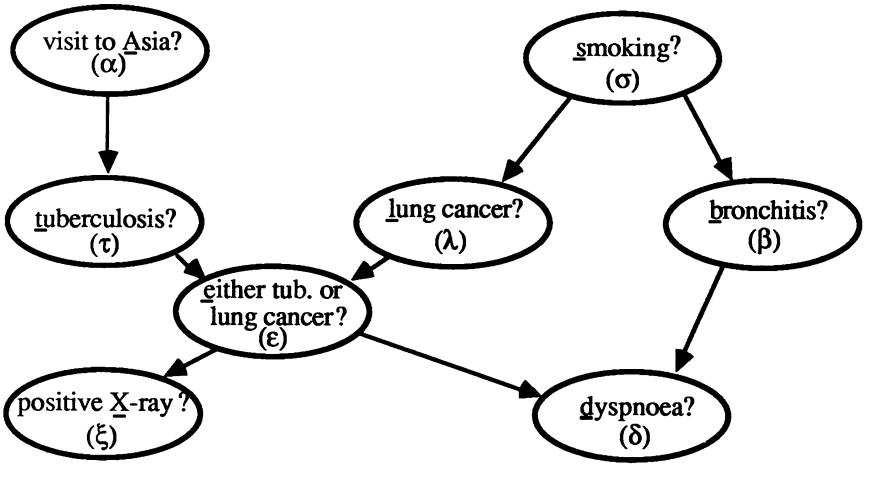


Fig. 2. Causal network in our fictitious example with short node names (greek letters) to be used in the text. Each node has two possible states representing responses 'yes' and 'no'. Direction of causality is from top to bottom.

TABLE 1
Fictitious conditional probability tables for our example

α:	$p(a)$	$=.01$	ε:	$p(e \mid l, t) = 1$
				$p(e \mid l, \bar{t}) = 1$
τ:	$p(t \mid a)$	$=.05$		$p(e \mid \bar{l}, t) = 1$
	$p(t \mid \bar{a})$	$=.01$		$p(e \mid \bar{l}, \bar{t}) = 0$
σ:	$p(s)$	$=.50$	ξ:	$p(x \mid e) = .98$
				$p(x \mid \bar{e}) = .05$
λ:	$p(l \mid s)$	$=.10$		
	$p(l \mid \bar{s})$	$=.01$	δ:	$p(d \mid e, b) = .90$
				$p(d \mid e, \bar{b}) = .70$
β:	$p(b \mid s)$	$=.60$		$p(d \mid \bar{e}, b) = .80$
	$p(b \mid \bar{s})$	$=.30$		$p(d \mid \bar{e}, \bar{b}) = .10$

$p(\alpha, \tau, \xi, \varepsilon, \delta, \lambda, \beta, \sigma)$ can be expressed as the product

$$p(\alpha)p(\tau \mid \alpha)p(\xi \mid \varepsilon)p(\varepsilon \mid \tau, \lambda)p(\delta \mid \varepsilon, \beta)p(\lambda \mid \sigma)p(\beta \mid \sigma)p(\sigma) \tag{4.1}$$

Simplistically, if we wanted to calculate, say, the revised probability $p(x \mid a, d)$ of a positive X-ray for our patient with dyspnoea who has visited Asia, we would require $p(x, a, d)/p(a, d)$. A brute-force evaluation of these probabilities will involve first performing the actual multiplications in (4.1) to obtain $2^8 = 256$ joint probabilities and then performing the relevant summations. Such a procedure is clearly inefficient and becomes prohibitive for large networks. A more efficient summation to obtain $p(x, a, d)$ may be performed, for example, as

$$p(a) \sum_{\tau} p(\tau \mid a)\left[\sum_{\varepsilon} p(x \mid \varepsilon)\left[\sum_{\lambda} p(\varepsilon \mid \tau, \lambda)\left[\sum_{\beta} p(d \mid \varepsilon, \beta)\left[\sum_{\sigma} p(\lambda \mid \sigma)p(\beta \mid \sigma)p(\sigma)\right]\right]\right]\right] \tag{4.2}$$

where the terms in parentheses are calculated one at a time from the 'inside'. Our approach may be viewed as exploiting an adapted topology of the graph in order to perform such efficient calculations systematically; see Cheeseman (1983) for a related argument.

Another important point is that—as has been recognised for many years in Bayesian statistics—full calculation of conditional distributions in a multiparameter context is often either unnecessary or can be left until the end of a sequence of uses of Bayes theorem: essentially, one can work with proportionality and not calculate the normalising constant unless strictly necessary (DeGroot, 1970). In our context, when states of particular nodes are revealed to us, we just have to insert the observed states into (4.1) and the expression will be correct, up to a normalising factor, for the conditional probability distribution of the states at the remaining nodes given those observed. The interpretation of each multiplicative term as the current conditional probability is, however, no longer valid. A formal way of expressing this is to think of (4.1) as a less structured expression, such as

$$\psi(\alpha)\psi(\tau, \alpha)\psi(\xi, \varepsilon)\psi(\varepsilon, \tau, \lambda)\psi(\delta, \varepsilon, \beta)\psi(\lambda, \sigma)\psi(\beta, \sigma)\psi(\sigma) \tag{4.3}$$

where the ψs are, initially, the conditional probabilities as in (4.1); for example, $\psi(\alpha) = p(\alpha)$, $\psi(t, \alpha) = p(\tau \mid \alpha)$, and so on.

To keep track of groups of variables entering into ψ functions, it becomes convenient to consider an *undirected* graph—that formed by providing a link between unjoined

parents of a common child and forgetting the original directions. We call this the *moral graph* (formed by 'marrying' parents), and this is shown in Fig. 3.

Specifically, (4.3) only involves functions on sets of nodes which are *complete* (all nodes joined) subgraphs of the moral graph. We shall call the ψs *evidence potentials.* Such a transfer from a directed to an undirected representation, exploiting proportionality, formed the basis of the program GENEX for obtaining probability formulae in pedigree analysis (Hilden, 1970, 1982).

The evidence potentials are not uniquely determined and may not, in general, be straightforward to interpret other than through the joint probability being proportional to expression (4.3), nor to enable simple node marginals to be obtained. In addition, when summing out σ in (4.2) an intermediate function on (λ, β) is formed involving pairs of nodes that are not joined in the graph, and hence would break our

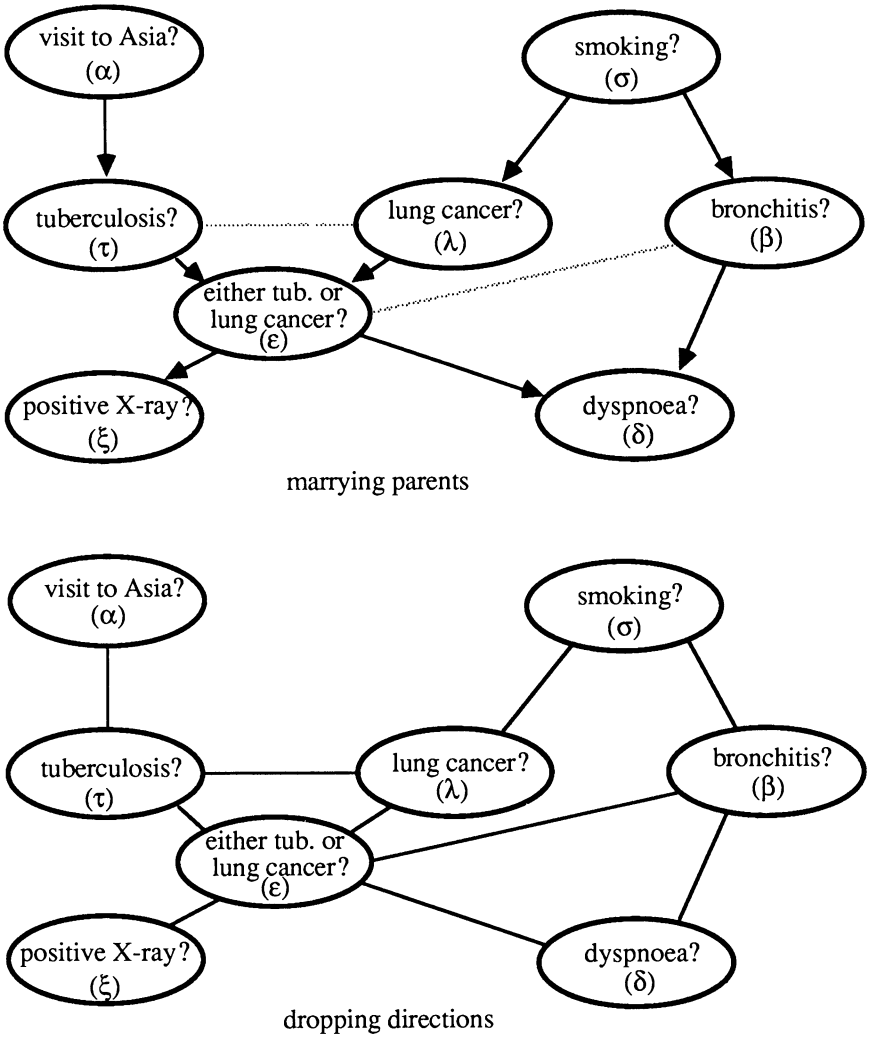


Fig. 3. Forming the moral graph: connections are introduced between nodes with common children and directions are then ignored.

rules for representation. (β could have been summed out first, but then a function on (σ, ε) would have been created.) Fortunately, there is a simple condition on the undirected graph that ensures that we will be able to perform all such summations, after choosing a suitable ordering, without creating new functions that depend on groups that may not be joined in the graph. This condition is that the undirected graph is *triangulated*, in that there are no cycles of length 4 or more without a *chord* or 'short-cut' (see Section 6 for precise graph-theoretic concepts).

If the moral graph is not already triangulated, it will need to be 'filled in' to make it so. In our example, we need to short-cut the cycle (σ, λ, ε, β). Two small fill-ins are possible. One can either add an edge between β and λ or add an edge between σ and ε. We add one between β and λ to obtain Fig. 4.

In larger networks, automatic algorithms for filling-in may be appropriate. These are discussed in Section 6, in which a technique for ordering the nodes—maximum cardinality search—is introduced. Note that the method of filling-in described by Spiegelhalter (1987)—by not first forming the moral graph—is not as efficient as the method described in the present paper. In our example, the method would have led to adding an unnecessary edge between α and λ, or σ and τ. Some savings may be possible by doing the graph manipulations dynamically at run-time, in that, for example, if evidence comes in at λ, say, a fill-in is unnecessary. But we imagine that there is more to be gained by optimising this step at an early stage (see Section 6), such that the fill-in is done at the construction of the expert system.

Of the many possible potential representations of the joint distribution, it will be most convenient to adopt one whose ψ functions are defined on the *cliques* (i.e. maximal complete subsets) of the filled-in graph, and hence whose domains will cope with all possible future computations. Such a representation is given by

$$p \propto \psi(\alpha, \tau)\psi(\tau, \lambda, \varepsilon)\psi(\lambda, \varepsilon, \beta)\psi(\lambda, \beta, \sigma)\psi(\varepsilon, \beta, \delta)\psi(\varepsilon, \xi) \tag{4.4}$$

The potential functions ψ are obtained by matching relevant expressions in (4.1): $\psi(\alpha, \tau) = p(\alpha)p(\tau \mid \alpha)$, $\psi(\tau, \lambda, \varepsilon) = p(\varepsilon \mid \tau, \lambda)$, $\psi(\lambda, \beta, \sigma) = p(\lambda \mid \sigma)\ p(\beta \mid \sigma)\ p(\sigma)$, $\psi(\varepsilon, \beta, \delta) = p(\delta \mid \varepsilon, \beta)$, $\psi(\varepsilon, \xi) = p(\xi \mid \varepsilon)$. $\psi(\lambda, \varepsilon, \beta)$ is not yet defined and may be assumed to take on any

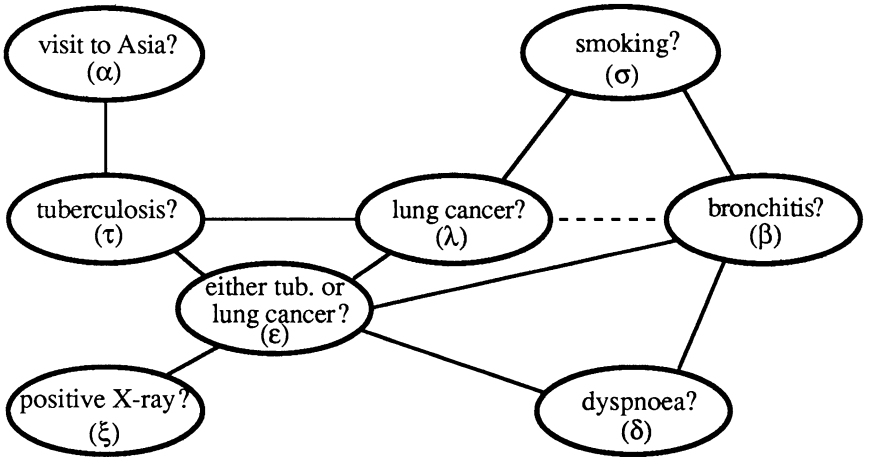


Fig. 4. The full moral graph, obtained from the original causal network by marrying parents, dropping directions and adding a connection between nodes representing lung cancer and bronchitis to make the final graph triangulated.

non-zero constant value, say 1. The other potentials may be calculated from Table 1 and are displayed in Table 3, third column.

The important bonus of a distribution expressed as functions on cliques of a triangulated graph is that it is also possible to express the joint distribution as a simple function of the individual marginal distributions on these cliques. In fact, as we shall see in Section 7.4, the joint probability may be written

$$\frac{p(\alpha, \tau)p(\tau, \lambda, \varepsilon)p(\lambda, \varepsilon, \beta)p(\lambda, \beta, \sigma)p(\varepsilon, \beta, \delta)p(\varepsilon, \xi)}{p(\tau)p(\lambda, \varepsilon)p(\lambda, \beta)p(\varepsilon, \beta)p(\varepsilon)}. \tag{4.5}$$

Storage of clique marginals makes retrieval of probabilities of single nodes trivial, after each batch of evidence has been processed. The expressions (4.1), (4.4) and (4.5) correspond to different *local representations* (Section 7) of the joint distribution. The procedures for moving between these representations are illustrated in the next section.

5. WORKING THROUGH THE EXAMPLE

We now consider the six issues outlined in Section 3 with respect to our simple example.

5.1. *Initialisation*

In order to obtain a representation as clique marginals, we shall see an additional intermediate step is necessary, in which the potentials now take the special form of conditional probability tables on a particular chain of sets of nodes. This *set chain* exploits an attractive characterisation of triangulated graphs that underlies our algorithm.

We begin by labelling the nodes using maximum cardinality search (see Section 6), starting with an arbitrary member, say α, given number 1. This ordering is shown in Fig. 5.

If we rank according to the highest labelled node in each clique, we obtain the ordering of the cliques shown in Table 2, which we say form a set chain. The algorithm

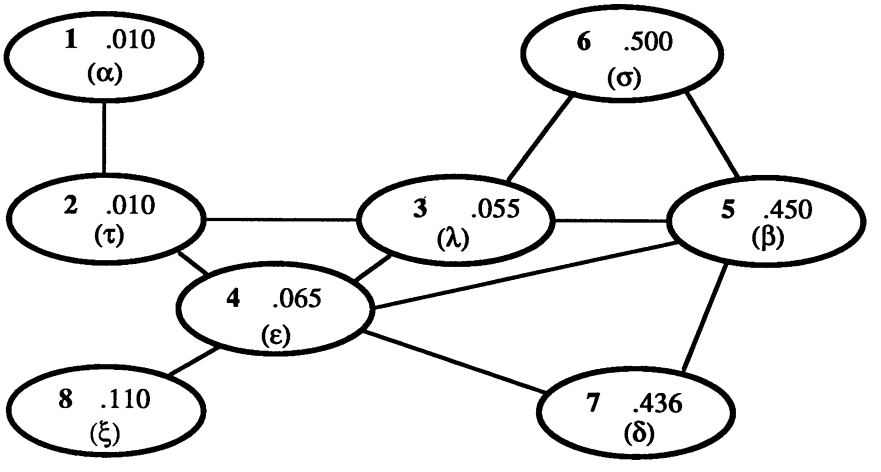


Fig. 5. Full moral graph for our fictitious example, showing a possible initial ordering of nodes and calculated marginals on each node. The full clique marginals are shown in Table 3.

TABLE 2
Initial set chain from full moral graph, Fig. 5, showing separators as nodes intersecting cliques of lower order

i	*Cliques* C_i	*Residuals* R_i	*Separators* S_i	*Possible parent cliques*
1	α, τ	α, τ	$\varnothing$	
2	$\tau, \lambda, \varepsilon$	λ, ε	τ	1
3	$\lambda, \varepsilon, \beta$	β	λ, ε	2
4	λ, β, σ	σ	λ, β	3
5	$\varepsilon, \beta, \delta$	δ	ε, β	3
6	ε, ξ	ξ	ε	2, 3, 5

for generating the set chain ensures the cliques have the *running intersection property*, in that the nodes of a clique (C_i) also contained in previous cliques ($C_1, \ldots, C_{i-1}$) are all members of *one* previous clique, known as a parent clique. This is demonstrated in Table 2, where—for example—$C_4 \cap (C_1 \cup C_2 \cup C_3) = \{\lambda, \beta\}$ is contained in C_3. We denote these separating nodes as $S_i = C_i \cap (C_1 \cup \ldots \cup C_{i-1})$, and the residual $C_i \backslash S_i$ as R_i.

The running intersection property ensures that our joint probability can also be expressed as

$$p(\alpha, \tau)p(\lambda, \varepsilon \mid \tau)p(\beta \mid \lambda, \varepsilon)p(\sigma \mid \lambda, \beta)p(\delta \mid \varepsilon, \beta)p(\xi \mid \varepsilon), \tag{5.1}$$

as can be seen directly from (4.5). In (5.1) p is simply a product of functions on cliques and hence is yet another potential representation, but one from which it will be easier to obtain clique marginals (see below). The problem lies in getting from (4.4) to (5.1).

Fortunately the running intersection property allows a straightforward algorithm to obtain (5.1) term by term. As we shall show in Section 9.2,

$$p(\xi \mid \varepsilon) = p(R_6 \mid S_6) = \psi(\xi, \varepsilon) \Big/ \sum_{\xi} \psi(\xi, \varepsilon)$$

and hence the last term in (5.1) is obtained directly from the potentials on C_6. Furthermore, the potential representation on the nodes except ξ is just as before except $\psi(C_5) = \psi(\varepsilon, \beta, \delta)$ is changed to $\psi(\varepsilon, \beta, \delta)\, \Sigma_\xi \psi(\xi, \varepsilon)$. In this particular instance $\Sigma_\xi \psi(\xi, \varepsilon) = 1$ but in general the procedure when i cliques remain is to transform $\psi(C_i)$ to

$$p(R_i \mid S_i) = \psi(C_i) \Big/ \sum_{R_i} \psi(C_i),$$

then multiply the potentials for a parent clique of C_i by $\Sigma_{R_i} \psi(C_i)$. The results can be traced in Table 3, fourth column. Note that we have displayed these numbers with more decimals than on the figures, to allow them to be used for calculations.

Having obtained our representation (5.1), it is possible to derive the clique marginals. From $p(C_1) = p(\alpha, \tau)$ we obtain $p(S_2) = p(\tau)$ by marginalisation. Then multiplication gives $p(C_2) = p(R_2 \mid S_2)p(S_2) = p(\lambda, \varepsilon \mid \tau)p(\tau)$. Chaining back through the graph gives the clique marginals displayed in Table 3, fifth column. The running intersection property ensures that at each step the required separator marginal can be obtained from a previously calculated clique marginal. The marginals on individual nodes are shown in Fig. 5.

To summarise, we have done two passes through the graph; the first to change our original potential representation derived from assessed conditional probability tables to one expressed as new conditional probability tables on a particular set chain, and the reverse pass to calculate marginals on the cliques. We shall find that these basic operations will be used in the other problems of interest.

5.2. *Absorption of Evidence*

We now suppose we elicit the responses a and d, a recent visitor to Asia with dyspnoea. If we wished to propagate their effects individually, a simple procedure using the clique marginals could be used, which will be illustrated in Section 5.4. However, it will be computationally more economical to 'save up' this information before propagating, and a potential representation is appropriate.

Thus, the required conditional distribution is

$$p(\tau, \lambda, \varepsilon, \beta, \sigma, \xi \mid a, d) \propto p(a, \tau, \lambda, \varepsilon, \beta, \sigma, d, \xi)$$
$$\propto \psi(a, \tau)\psi(\tau, \lambda, \varepsilon)\psi(\lambda, \varepsilon, \beta)\psi(\lambda, \beta, \sigma)\psi(\varepsilon, \beta, d)\psi(\varepsilon, \xi), \tag{5.2}$$

by (4.4). Removing the observed nodes from the graph gives Fig. 6, and a potential representation on the cliques of this graph

$$\psi^*(\tau, \lambda, \varepsilon)\psi^*(\lambda, \varepsilon, \beta)\psi^*(\lambda, \beta, \sigma)\psi^*(\varepsilon, \xi) \tag{5.3}$$

is obtained by matching terms with (5.2):

$$\psi^*(\tau, \lambda, \varepsilon) = \psi(\tau, \lambda, \varepsilon)\psi(a, \tau), \qquad \psi^*(\lambda, \varepsilon, \beta) = \psi(\lambda, \varepsilon, \beta)\psi(\varepsilon, \beta, d),$$
$$\psi^*(\lambda, \beta, \sigma) = \psi(\lambda, \beta, \sigma), \qquad \psi^*(\varepsilon, \xi) = \psi(\varepsilon, \xi).$$

Essentially we absorb evidence by conditioning our potential representation, which involves projecting the potentials involving each observed term onto either a new, reduced clique, or onto a neighbour if a clique may be removed. Starting with the

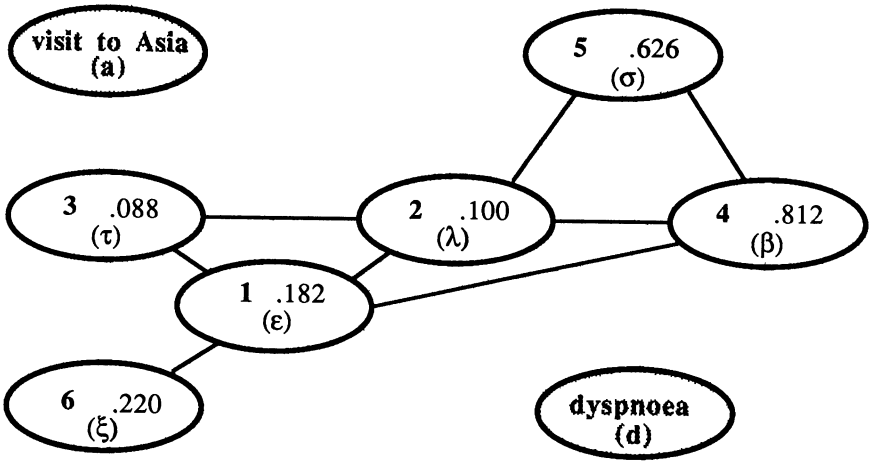


Fig. 6. Graph remaining after evidence about a visit to Asia and dyspnoea being present has been absorbed and propagated. The observed nodes are shaded and a possible new ordering of the remaining nodes is shown.

set chain representation of Table 3 (fourth column), the conditional potentials (5.3) are shown in Table 3, sixth column. For example $\psi^*(t, l, e) = \psi(t, l, e)\psi(a, t) = 0.055 \times 0.0005 = 0.000028$.

5.3. *Global Propagation*

We now have a local potential representation of our beliefs on the reduced graph shown in Fig. 6, and we can obtain the marginal distributions on the cliques in

TABLE 3

Stages in calculating clique marginals on full moral graph from original conditional probability tables, and potentials after absorbing evidence a, d

Original clique order	*Configuration*	*Potentials from conditional probability tables*	*Potentials from set chain*	*Clique marginals*	*Potentials after absorbing a, d*
			$\{p(\alpha, \tau)\}$		
$C_1 = \{\alpha, \tau\}$	a, t	.0005	.0005	.0005	
	$a, \bar{t}$	.0095	.0095	.0095	
	$\bar{a}, t$	.0099	.0099	.0099	
	$\bar{a}, \bar{t},$	.9801	.9801	.9801	
			$\{p(\lambda, \varepsilon \mid \tau)\}$		
$C_2 = \{\tau, \lambda, \varepsilon\}$	t, l, e	1	.0550	.00057	.000028
	$t, l, \bar{e}$	0	.0	.0	.0
	$t, \bar{l}, e$	1	.9450	.00983	.000473
	$t, \bar{l}, \bar{e}$	0	.0	.0	.0
	$\bar{t}, l, e$	1	.0550	.05443	.000523
	$\bar{t}, l, \bar{e}$	0	.0	.0	.0
	$\bar{t}, \bar{l}, e$	0	.0	.0	.0
	$\bar{t}, \bar{l}, \bar{e}$	1	.9450	.9352	.008978
			$\{p(\beta \mid \lambda, \varepsilon)\}$		
$C_3 = \{\lambda, \varepsilon, \beta\}$	l, e, b	1	.5727	.03150	.5154
	$l, e, \bar{b}$	1	.4273	.02350	.2991
	$l, \bar{e}, b$	1	.5727	.0	.4582
	$l, \bar{e}, \bar{b}$	1	.4273	.0	.0427
	$\bar{l}, e, b$	1	.4429	.00435	.3986
	$\bar{l}, e, \bar{b}$	1	.5571	.00548	.3899
	$\bar{l}, \bar{e}, b$	1	.4429	.4142	.3543
	$\bar{l}, \bar{e}, \bar{b}$	1	.5571	.5210	.0557
			$\{p(\sigma \mid \lambda, \beta)\}$		
$C_4 = \{\lambda, \beta, \sigma\}$	l, b, s	.0300	.9524	.0300	.9524
	$l, b, \bar{s}$	.0015	.0476	.0015	.0476
	$l, \bar{b}, s$	.0200	.8511	.0200	.8511
	$l, \bar{b}, \bar{s}$	.0035	.1489	.0035	.1489
	$\bar{l}, b, s$	.2700	.6452	.2700	.6452
	$\bar{l}, b, \bar{s}$	.1485	.3548	.1486	.3548
	$\bar{l}, \bar{b}, s$	.1800	.3419	.1800	.3419
	$\bar{l}, \bar{b}, \bar{s}$	.3465	.6581	.3464	.6581
			$\{p(\delta \mid \varepsilon, \beta)\}$		
$C_5 = \{\varepsilon, \beta, \delta\}$	e, b, d	.9	.9	.03227	
	$e, b, \bar{d}$	.1	.1	.00359	
	$e, \bar{b}, d$	.7	.7	.02029	
	$e, \bar{b}, \bar{d}$	.3	.3	.00869	
	$\bar{e}, b, d$	.8	.8	.3314	
	$\bar{e}, b, \bar{d}$	.2	.2	.08284	
	$\bar{e}, \bar{b}, d$	.1	.1	.05210	
	$\bar{e}, \bar{b}, \bar{d}$	.9	.9	.4689	
			$\{p(\xi \mid \varepsilon)\}$		
$C_6 = \{\varepsilon, \xi\}$	e, x	.98	.98	.06354	.98
	$e, \bar{x}$	.02	.02	.00130	.02
	$\bar{e}, x$	.05	.05	.04676	.05
	$\bar{e}, \bar{x}$	.95	.95	.8884	.95

precisely the same manner as initialising a full moral graph as described in Section 5.1. Maximum cardinality search, starting at any node, say ε, provides a node ordering shown in Fig. 6 and subsequent ordering of the cliques $C_1 = \{\tau, \lambda, \varepsilon\}$, $C_2 = \{\lambda, \varepsilon, \beta\}$, $C_3 = \{\lambda, \beta, \sigma\}$, $C_4 = \{\varepsilon, \xi\}$.

The first two conditional distributions in the set chain

$$p(R_4 \mid S_4) = p(\xi \mid \varepsilon), \quad p(R_3 \mid S_3) = p(\sigma \mid \lambda, \beta)$$

are obtained unchanged from the initialisation procedure, but in calculating $p(R_2 \mid S_2) = p(\beta \mid \lambda, \varepsilon)$ the revised potential function on $\{\lambda, \varepsilon, \beta\}$ has to be taken into account.

Continuing backwards we eventually obtain the marginal on $S_1 = \{\tau, \lambda, \varepsilon\}$. This then allows us to go forward through the graph to obtain marginals on all cliques and hence on the nodes as shown in Fig. 6.

We note that the symptomatic and 'epidemiological' evidence has raised the probabilities of all diseases, with a ninefold increase for tuberculosis. The dyspnoea has given us a higher expectation that the patient is a smoker, while the chance of a positive chest X-ray has doubled.

5.4. *Hypothesising and Propagating Single Items of Evidence*

An attractively simple propagation procedure is available if we wish to condition on either single nodes or nodes contained in a single clique. As an example, suppose we take the reduced graph (Fig. 6), and assume that the doctor wished to hypothesise that tuberculosis was ruled out by some external test. We use this as an opportunity to show direct global propagation staying within the marginal representation. In many applications this may be the only procedure required once initialisation has taken place.

Fig. 7 shows an ordering of the nodes which has been obtained by maximum cardinality search, and the consequent ordering of the cliques.

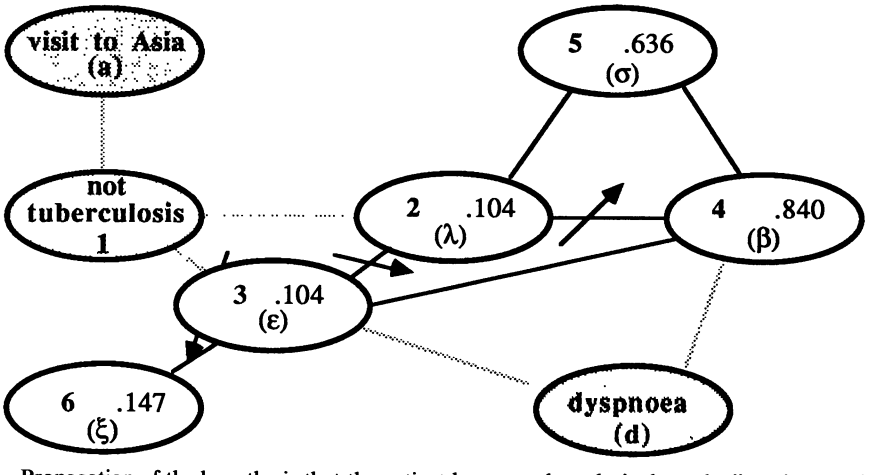


Fig. 7. Propagation of the hypothesis that the patient has not tuberculosis through clique intersections. Observed nodes are shaded darkly and the node hypothesised shaded lighter. A possible new ordering of the nodes involved is shown. The arrows indicate the flow of hypothesised evidence.

Following a similar argument to that leading to (4.5) the joint distribution $\tilde{p}$, having observed a and d, may be written in the form

$$\frac{\tilde{p}(\tau, \lambda, \varepsilon)\tilde{p}(\lambda, \varepsilon, \beta)\tilde{p}(\lambda, \beta, \sigma)\tilde{p}(\varepsilon, \xi)}{\tilde{p}(\lambda, \varepsilon)\tilde{p}(\lambda, \beta)\tilde{p}(\varepsilon)}.$$

After hypothesising $\bar{t}$, our revised distribution, denoted p^*, is obtained by dividing $\tilde{p}$ by $\tilde{p}(\bar{t})$ to give

$$\frac{\tilde{p}(\lambda, \varepsilon \mid \bar{t})\tilde{p}(\lambda, \varepsilon, \beta)\tilde{p}(\lambda, \beta, \sigma)\tilde{p}(\varepsilon, \xi)}{\tilde{p}(\lambda, \varepsilon)\tilde{p}(\lambda, \beta)\tilde{p}(\varepsilon)} \tag{5.4}$$

and may itself be represented in terms of revised marginals on the cliques $\{\lambda, \varepsilon, \beta\}$, $\{\lambda, \beta, \sigma\}$ and $\{\varepsilon, \xi\}$.

We may first find the new marginal distribution $p^*(\lambda, \varepsilon) = \tilde{p}(\lambda, \varepsilon \mid \bar{t})$ directly from the stored marginal $\tilde{p}(\tau, \lambda, \varepsilon)$. Then the new marginal on the first clique $\{\lambda, \varepsilon, \beta\}$ can be found by summing ξ and σ out of (5.4) and may be written

$$p^*(\lambda, \varepsilon, \beta) = \tilde{p}(\lambda, \varepsilon, \beta)p^*(\lambda, \varepsilon)/\tilde{p}(\lambda, \varepsilon).$$

From this $p^*(\lambda, \beta)$ may be found, and we can continue to the next clique $\{\lambda, \beta, \sigma\}$ by

$$p^*(\lambda, \beta, \sigma) = \tilde{p}(\lambda, \beta, \sigma)p^*(\lambda, \beta)/\tilde{p}(\lambda, \beta)$$

and so on.

This is an illustration of the updating scheme described in Spiegelhalter (1987) and discussed in Section 8.2. Each parent clique in the set chain passes evidence to its children through multiplying each term in the marginal distribution of the child by the ratio of new to old probability on the appropriate intersection term. The calculations are shown in detail in Table 4, where by 'update ratio' we mean the ratio $p^*(\lambda, \varepsilon)/\tilde{p}(\lambda, \varepsilon)$ propagated through the clique intersection. Lemmer (1983) describes a similar propagation rule derived from maximum entropy considerations.

We emphasise the trivial nature of the calculations, the ability to deal with logical zeros, and how the 'update ratio' seems to have the possibility to form the basis for

TABLE 4

Evidence propagation from clique $\{\tau, \lambda, \varepsilon\}$ *to* $\{\lambda, \varepsilon, \beta\}$ *after hypothesising* $\tau = \bar{t}$

Value combination	*Initial belief*	*'Update ratio'*	*Revised belief* p^*
$l\ e\ b$	.0631	1.0401	.0656
$l\ e\ \bar{b}$	.0366	1.0401	.0381
$l\ \bar{e}\ b$	0	0	0
$l\ \bar{e}\ \bar{b}$	0	0	0
$\bar{l}\ e\ b$	.0418	0	0
$\bar{l}\ e\ \bar{b}$	.0410	0	0
$\bar{l}\ \bar{e}\ b$	.7064	1.0964	.7745
$\bar{l}\ \bar{e}\ \bar{b}$	.1111	1.0964	.1218
	1.0000		1.0000

'explanation' of the flow of evidence through the network. The eventual node marginals are shown in Fig. 7, where it is clear that ruling tuberculosis out has minimal effect.

5.5. *Planning*

When a particular node is of interest, but currently unobservable, it is valuable to know which questions will most provide relevant information. Simulated global propagation of the effect of asking each possible question in turn is clearly inefficient, but if we use a symmetric measure of mutual information between two currently unobserved nodes, then we need only simulate *away* from the node of interest. A suitable measure between two nodes u, v is

$$M(u, v) = -E_{u,v} \log\left\{\frac{p(u)p(v)}{p(u, v)}\right\}$$

the 'mutual information' which can be interpreted as the Kullback–Leibler distance between the true joint distribution of u, v and the distribution were independence to be assumed. This was suggested as a planning device by Jensen *et al.* (1987).

Suppose the node of interest is lung cancer, λ. Then the mutual information from another node u may be written as

$$M(u, \lambda) = \sum_u \sum_\lambda \log\left\{\frac{p(u \mid \lambda)}{p(u)}\right\} p(u \mid \lambda)p(\lambda)$$

the expected 'change' in belief in u were λ to be elicited. We thus need to propagate the effect of l and $\bar{l}$ using the marginal propagation method described in Section 5.4, find revised beliefs $p(u \mid \bar{l})$ and $p(u \mid l)$, and hence calculate the mutual information measure for each node u. We obtain $M(\sigma, \lambda) = 0.02$ and $M(\xi, \lambda) = 0.16$, showing the limited value of smoking history with respect to lung cancer, since bronchitis is an alternative explanation for the symptoms and smoking is also a risk factor for bronchitis.

5.6. *Influential Findings*

Suppose that further investigations reveal that the patient has not smoked and has a negative X-ray. The joint conditional distribution $p(\tau, \lambda, \varepsilon, \beta \mid a, d, \bar{x}, \bar{s})$ is then obtained by projecting the potentials ψ^* from (5.3) onto the two remaining cliques to give a potential representation

$$\tilde{\psi}(\tau, \lambda, \varepsilon)\tilde{\psi}(\lambda, \varepsilon, \beta),$$

where

$$\tilde{\psi}(\tau, \lambda, \varepsilon) = \psi^*(\tau, \lambda, \varepsilon)\psi^*(\varepsilon, \bar{x}) \text{ and } \tilde{\psi}(\lambda, \varepsilon, \beta) = \psi^*(\lambda, \varepsilon, \beta)\psi^*(\lambda, \beta, \bar{s}).$$

Alternatively, but less efficiently, these could have been obtained directly from (4.4). The techniques of Section 5.3 then provide clique marginals and hence node marginals shown in Fig. 8. We see that the circumstantial evidence for tuberculosis has been overwhelmed by the negative X-ray, leaving bronchitis as the only serious contender.

For explanatory purposes, it may be useful to indicate how influential items of evidence are in changing our opinion concerning, say, bronchitis. One approach is to remove items one at a time from the list of findings and measure 'change' in terms of

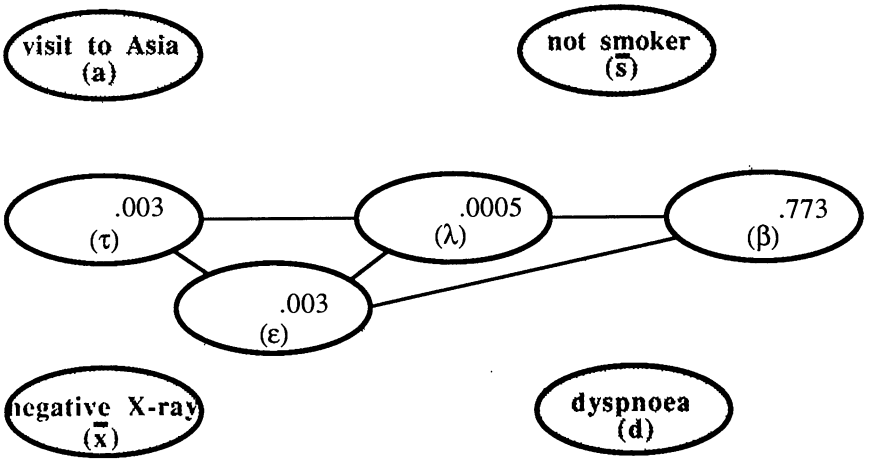


Fig. 8. Status of the system after further establishing smoking history and negative chest X-ray. Information about shaded nodes have been revealed to the clinic.

the Kullback–Leibler distance used above. For example, we may define the 'influence' of the negative smoking history $\bar{s}$ on our belief in β as

$$I(\beta : \bar{s} \mid a, d, \bar{x}) = -\sum_{\beta} \log\left\{\frac{p(\beta \mid a, d, \bar{x})}{p(\beta \mid a, d, \bar{x}, \bar{s})}\right\} p(\beta \mid a, d, \bar{x}, \bar{s}).$$

Quick computation of the necessary 'de-conditioned' distribution is possible through changing the potential representation for $p(\tau, \lambda, \varepsilon, \beta \mid a, d, \bar{x}, \bar{s})$ to one for $p(\tau, \lambda, \varepsilon, \beta \mid a, d, \bar{x})$. Specifically, we have

$$\begin{aligned} p(\tau, \lambda, \varepsilon, \beta \mid a, d, \bar{x}) &\propto \sum_{\sigma} p(a, \tau, \lambda, \varepsilon, \beta, \sigma, d, \bar{x}) \\ &\propto \sum_{\sigma} \psi(a, \tau)\psi(\tau, \lambda, \varepsilon)\psi(\lambda, \varepsilon, \beta)\psi(\lambda, \beta, \sigma)\psi(\varepsilon, \beta, d)\psi(\varepsilon, \bar{x}) \\ &\propto \left\{\frac{\psi(\lambda, \beta, s) + \psi(\lambda, \beta, \bar{s})}{\psi(\lambda, \beta, \bar{s})}\right\}\psi^*(\tau, \lambda, \varepsilon)\psi^*(\lambda, \varepsilon, \beta). \end{aligned}$$

(Note that the term in curly brackets is the reciprocal of $p(\bar{s} \mid \lambda, \beta)$, see Section 11.3.) Hence we have the same potential representation as when all evidence is taken into account except the neighbouring clique to σ, $\{\lambda, \varepsilon, \beta\}$, has its potentials multiplied by $1 + \psi(\lambda, \beta, s)/\psi(\lambda, \beta, \bar{s})$. Going through the process of finding clique marginals reveals a revised probability $p(b \mid a, d, \bar{x}) = 0.863$ and hence an influence of not smoking of 0.03. Repeating this operation for each item reveals influences of the visit to Asia 0.00, dyspnoea 0.48, and negative X-ray 0.01. Thus, not surprisingly, the presence of dyspnoea has the greatest impact on our belief in bronchitis, allowing for other evidence available.

6. GRAPHS AND NOTATION

The problems of interest are expressed using networks and our methods are based on graph-theoretic algorithms. We shall refer to the literature for a more precise

description of the concepts involved, but some terminology is necessary here. Good standard references are the books by Berge (1973) and Golumbic (1980).

A *graph* consists of a finite set V of *nodes* and a set of *edges* between pairs of these. The edges can either be present or absent and *directed* or *undirected*. In the present paper, no graphs will have both types of edges. The notation $v \to w$, $v \not\to w$, $v \sim w$, $v \not\sim w$ should be immediately transparent. A graph is represented with nodes as circles and edges either arrows or lines as in Fig. 7. If $v \to w$, w is a *child* of v, v is a *parent* of w and Π_w denotes the set of such parents. If $v \sim w$, we say that v and w are *neighbours* and $\mathrm{bd}(v)$ is the boundary of v, $\mathrm{bd}(v) = \{w \mid w \sim v\}$. The notions are illustrated in Fig. 9. A subset $C \subseteq V$ is *complete* if there are edges between all nodes in C. A subset which is maximal with this property is called a *clique*.

A *path* of length k from v to w is a sequence of nodes $(v_0, v_1, \ldots, v_k)$ with $v_0 = v$, $v_k = w$ such that either $v_{i-1} \to v_i$ or $v_{i-1} \sim v_i$. A *cycle* is a path $(v_0, \ldots, v_k)$ with $v_0 = v_k$. An *acyclic directed graph* is the formal equivalent of what we have earlier termed a causal network. For each acyclic directed graph, we define the corresponding *moral graph* as the one obtained by 'marrying' parents, i.e. by adding edges $u \sim w$ between all pairs $u, w \in \Pi_v$ of parents where neither $u \to w$ nor $w \to u$ and then changing all directed edges to undirected. This moral graph is thus an undirected graph.

The class of *triangulated graphs* plays a central role in our methods. These are undirected graphs, characterised by the property that all cycles $(v = v_0, v_1, \ldots, v_k = v)$ of length $k \geqslant 4$ possess a *chord*, i.e. $v_i \sim v_j$ for some i, j with $j \neq i \pm 1 \bmod (k+1)$. Triangulated graphs are described by Berge (1973) and Golumbic (1980), but have also been discussed under other names as, for example, *rigid circuit* (Dirac, 1961), *chordal* (Gavril, 1972) and *decomposable* (Lauritzen *et al.*, 1984). The moral graph corresponding to a 'generalised Chow-tree' is always triangulated, see Goldman and Rivest (1986).

A numbering of nodes V in an undirected graph is called *perfect* if for all i

$$A_i = \mathrm{bd}(i) \cap \{1, \ldots, i-1\} \text{ is complete,} \tag{6.1}$$

where $V = \{1, \ldots, k\}$ is the numbered node set. *A graph is triangulated if and only if it admits a perfect numbering* (see references above for proofs). If a numbering is not perfect, it can be made so by 'filling-in', i.e. by adding edges between nodes in the sets A_i above such as to satisfy (6.1). A fast method for computing this fill-in is

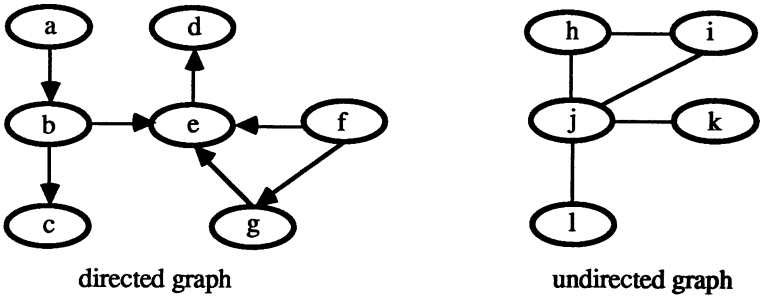


Fig. 9. Illustration of graph-theoretic concepts: we have, for example, $e \to d$, $c \not\to e$, $j \sim i$, $k \not\sim i$. The set of parents of e is $\{b, f, g\}$, b has children $\{e, c\}$, the boundary $\mathrm{bd}(i)$ of i is $\{h, j\}$. The undirected graph has cliques $\{i, j, h\}$, $\{j, k\}$, $\{j, l\}$.

described by Tarjan and Yannakakis (1984). A fill-in is said to be *minimal* if no renumbering of the nodes gives a fill-in which is a subset of the given fill-in, whereas a fill-in is *minimum* if it contains a minimum number of edges.

A fast algorithm for testing the triangulatedness of an undirected graph has been developed by Tarjan and Yannakakis (1984). Called *maximum cardinality search*, it runs in $O(n+e)$ time, where n is the number of nodes and e the number of edges. The algorithm is as follows.

> Give number 1 to an arbitrary node. Number the nodes consecutively, choosing as the next to number a node with a maximum number of previously numbered neighbours. Break ties arbitrarily.

If the graph is triangulated, the ordering so obtained will be perfect. The test for triangulatedness is therefore completed by computing the fill-in and checking whether or not this is empty. Fig. 10 shows the numbering of the nodes of the MUNIN graph obtained by maximum cardinality search.

The size of the state spaces at the various nodes in this example varies from 2 to 11. It is apparent from this application that many of the procedures described can be improved by taking into account special features of the particular system. One could for example have filled in an edge between node 1 and 16 instead of one between 7 and 10. But since the number of states at nodes 1, 7, 10, 16 were 11, 2, 9, 7, the maximal clique state size in the two cases will be 198 in the fill-in used but 693 in the other. Since maximal clique state size is vital for the computing time (see Section 10), this saving is substantial.

It might happen that the filled-in graph has such large cliques that our methods are not feasible; for example, lattice structures in types of image processing. However, Pearl (1986a) argues that sparse, irregular, causal networks are generally appropriate, and this has been the case in the MUNIN application.

The fill-in corresponding to maximum cardinality search will in general have no optimality properties. *Lexicographic search* as described by Rose *et al.* (1976) gives a minimal fill-in but runs in $O(ne)$ time. The problem of computing a minimum fill-in is NP complete, as shown by Yannakakis (1981). See, for example, Golumbic (1980) for the notion of NP completeness.

A *hypergraph* is a set V together with a set Γ of subsets of V. A hypergraph is *acyclic* if no elements in Γ are subsets of other elements, and if the elements of Γ can be ordered $(C_1, \ldots, C_p)$ to have the *running intersection property*:

$$\forall j \geqslant 2,\ \exists i < j : C_i \supseteq C_j \cap (C_1 \cup \ldots \cup C_{j-1}),$$

cf. Beeri *et al.* (1981, 1983), Lauritzen *et al.* (1984), Tarjan and Yannakakis (1984).

It is convenient to introduce the sets for $j = 1, \ldots, p$:

$$S_j = C_j \cap (C_1 \cup \ldots \cup C_{j-1}),\ R_j = C_j \backslash S_j$$

where $S_1 = \varnothing$, $R_1 = C_1$. The sets S_j *separate* the *residual* R_j from $(C_1 \cup \ldots \cup C_{j-1}) \backslash S_j$. Any clique C_i containing S_j with $i < j$ shall be called a possible *parent* clique of C_j.

A hypergraph Γ is acyclic if and only if it can be considered to be the set of cliques of a triangulated graph (see the above references). If a hypergraph is acyclic, and the nodes of the corresponding triangulated graph are numbered by either maximum

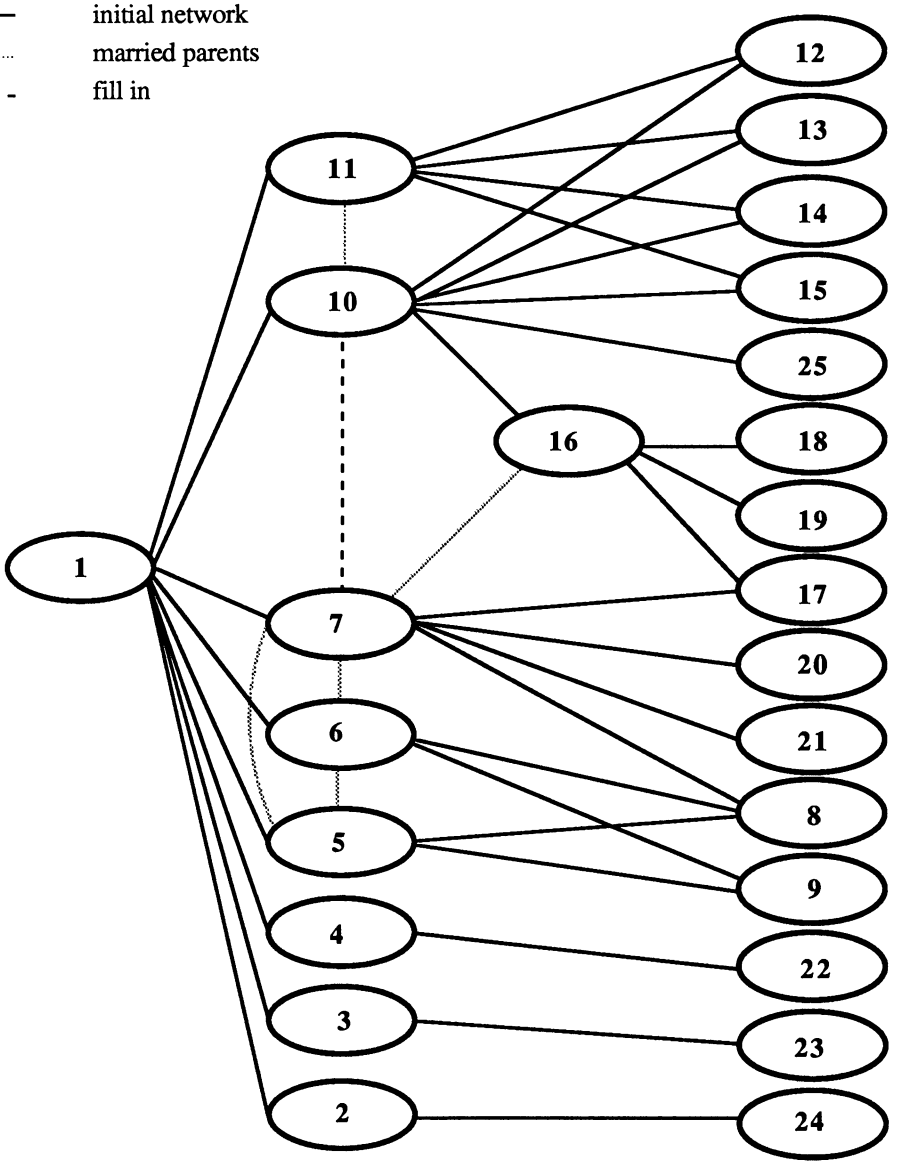


Fig. 10. The MUNIN network from Fig. 1 with nodes ordered by maximum cardinality search, performed after parents have been married, directions have been dropped and the graph has been filled in to make it triangulated.

cardinality search or by lexicographic search, the cliques can then be ordered according to the maximal note in each. *The ordering so obtained will have the running intersection property*, see Tarjan and Yannakakis (1984), Leimer (1985, 1988).

Having covered the basic graph-theoretic notions, we now need some probability notation. With each node $v \in V$ a *state space* Ω_v (finite) is associated. The total set of *configurations* is the set

$$\Omega = \underset{v \in V}{\times} \Omega_v.$$

Typical elements of Ω_v are denoted x_v and elements of Ω are $(x_v,\ v \in V)$. We are sometimes interested in *partial configurations*

$$x_A = (x_v,\ v \in A) \in \Omega_A = \underset{v \in A}{\times} \Omega_v.$$

We shall repeatedly use the short notation $p(A)$ for the probability

$$p(A) = \Pr\{X_v = x_v,\ v \in A\}$$

where X_v are random variables taking values in the spaces Ω_v. Similarly we write $p(A \mid B)$ for

$$p(A \mid B) = \Pr\{X_v = x_v,\ v \in A \mid X_w = x_w,\ w \in B\}$$

and, for example, $\Sigma_A p(A, B)$ for

$$\sum_{x_A \in \Omega_A} \Pr(X_A = x_A,\ X_B = x_B).$$

7. LOCAL REPRESENTATIONS

In the present section we shall discuss four related but different local representations of a joint probability distribution. Every representation consists of a system of functions, which each depend only on a limited number of variables, together with a rule for combining these to form the joint probability. It is important that, in general, this joint probability does not have to be computed. It is a mathematical object and has the role of ensuring consistency of manipulations. Each of the four representations has its advantages depending on the issue to be considered, cf. the discussion in Sections 4 and 5.

7.1. *Conditional Probability Tables*

A system of *conditional probability tables* corresponding to a causal network consists of a list of *tables* k_v, for all nodes v. The tables express, for each combination of states at the parent nodes Π_v, the conditional probability of the given states at node v, and must thus satisfy

$$k_v(v \mid \Pi_v) \geqslant 0, \quad \sum_v k_v(v \mid \Pi_v) = 1.$$

Since we have assumed the corresponding directed graph to be *acyclic* (that is, the network contains no feedback), an easy induction argument shows that the expression

$$p(V) = \prod_{v \in V} k_v(v \mid \Pi_v) \tag{7.1}$$

defines a probability on the set of joint configurations and, in fact,

$$p(v \mid \Pi_v) = k_v(v \mid \Pi_v).$$

If $k_v > 0$, (7.1) defines a *causal Markov random field* as discussed by Kiiveri *et al.* (1984). We have chosen to allow $k_v = 0$ to be able to deal with logical links, and we will refer to (7.1) as the *causal Markov property*.

Following Kiiveri *et al.* (1984), we define A_v as the set of nodes 'anterior' to v in the sense that $v \notin A_v$ and if $w \in A_v$ there is no path from v to w. Clearly Π_v is a subset of A_v. The causal Markov property (7.1) then implies $p(v \mid A_v) = p(v \mid \Pi_v)$, i.e. that the variable v is conditionally independent of the anterior variables given the parent variables.

7.2. *Evidence Potentials*

A *potential representation* consists of a *list* Δ of subsets A of nodes and corresponding *evidence potentials* ψ_A, $A \in \Delta$, each being real-valued functions depending only on the states of variables in A. The potentials are assumed to be non-negative.

The equation

$$p(V) = Z^{-1} \prod_{A \in \Delta} \psi_A, \tag{7.2}$$

where

$$Z = \sum_V \prod_{A \in \Delta} \psi_A$$

is a normalisation constant, now defines a joint probability if $Z \neq 0$. If $Z = 0$, the assumptions made are mutually incompatible. Apart from the non-negativity and the condition $Z \neq 0$, the functions ψ_A and the list Δ can be specified freely. In general different potential representations can give the same probability function p but, as we shall see, this can be used to our advantage. Note that a transition kernel representation automatically gives a potential representation if we let

$$\psi_A = k_v(v \mid \Pi_v) \text{ for } A = \{v\} \cup \Pi_v,\ v \in V,$$

as in (4.3). In this case, we have $Z = 1$, as follows directly from (7.1).

If an undirected graph is given such that all subsets A in the list Δ are *complete* subsets, we say that (Δ, ψ) is a *nearest neighbour potential.* If p has a representation of the form (7.2) with nearest neighbour potentials, we say that p is *Markov* with respect to the given undirected graph. Since the sets $\{v\} \cup \Pi_v$ are complete in the moral graph, it follows that if *p is causal Markov, it is Markov on the corresponding moral graph.*

Note that our notion of a Markov probability is slightly non-standard. If $\psi_A > 0$, the 'usual' potentials will be logarithms of ψ and the representation (7.2) would mean that p is nearest neighbour Gibbs. It is then a theorem that this is equivalent to a Markov property, expressed in terms of conditional independence; see, for example, Speed (1979). In the case where zero probabilities are allowed, things are slightly more complicated, see, for example, Moussouris (1974) or Averintsev (1975). But, to be able to cope with logical links, we shall allow zero potentials and refer to the factorisation (7.2) as the Markov property on an undirected graph, without discussing conditional independence directly.

The advantages of potential representations are partly that they can be freely specified, and partly that they behave extremely well under conditioning, which in this context means absorption of incoming evidence. This way of representing a probability is related to that used in log-linear models, see Darroch *et al.* (1980), and has, in the context of expert systems, been exploited by Cheeseman (1983); see also Pearl (1986b).

7.3. *Set Chains*

A set chain representation is based on an ordered chain of sets $(C_1, \ldots, C_p)$ with union V, having the running intersection property (see Section 6). Together with such a list of sets, we need *kernels* $g(R_i \mid S_i)$ with

$$g(R_i \mid S_i) \geq 0, \quad \sum_{R_i} g(R_i \mid S_i) = 1$$

and we can again define a joint probability

$$p(V) = \prod_{i=1}^{p} g(R_i \mid S_i) \tag{7.3}$$

where $g(R_i \mid S_i)$ now become conditional probabilities satisfying

$$p(R_i \mid S_i) = g(R_i \mid S_i).$$

In fact, if $p(S_i^*) = 0$ for some particular configuration S_i^* states at nodes S_i, the corresponding kernel $g(R_i \mid S_i^*)$ does not have to satisfy the constraints above and can take on any value, $p(V)$ being zero for any configuration containing S_i^*.

Set chain representations are similar to the conditional probability tables, except that they involve sets of nodes rather than single nodes. They were more or less implicit in the work of Goodman (1970), see Darroch *et al.* (1980). In the present paper, the set chains will play a role only as an intermediate representation used in the calculations.

7.4. *Clique Marginals*

A clique marginal representation consists of a list Γ of subsets C of nodes such that

$$V = \bigcup_{C \in \Gamma} C.$$

We shall assume that the list Γ is such that the hypergraph (V, Γ) is *acyclic*, implying that Γ can be considered to be the set of cliques of an undirected, triangulated graph (see Section 6). Together with the list Γ we need *marginal probabilities* μ_C satisfying

$$\mu_C \geq 0, \quad \sum \mu_C = 1$$

but also *the consistency condition* that if $C \cap D \neq \varnothing$ then

$$\mu_C(C \cap D) = \sum_{C \setminus D} \mu_C = \sum_{D \setminus C} \mu_D = \mu_D(C \cap D).$$

The problem of the existence of a joint probability with given marginals to a system of sets was considered by Vorob'ev (1962, 1963) and Kellerer (1964a, b), who showed that the answer is positive when (V, Γ) forms an acyclic hypergraph. If one further assumes the joint probability to be Markov or to have maximal entropy, the answer is unique and is given as follows.

Order the sets C to form a chain $(C_1, \ldots, C_p)$ with the running intersection property—this can be done because the hypergraph is acyclic. Then define

$$g(R_i \mid S_i) = \mu_{C_i}(R_i \cup S_i)/\mu_{C_j}(S_i).$$

The clique C_j is any parent clique of C_i, i.e. $S_i \subseteq C_j$, and 0/0 is defined to be zero.

The joint probability p can now be defined by (7.3) and thus satisfies

$$p(V) = \prod_{i=1}^{p} \frac{p(C_i)}{p(S_i)} \tag{7.4}$$

as well as

$$p(C_i) = \mu_{C_i}.$$

In the present context this has been utilised by Goldman and Rivest (1986) and Jirousek and Perez (1986), where the set marginals primarily play the role of being constraints used to initialise the system. Malvestuto (1987) recommends approximating high dimensional databases in terms of clique marginals to allow efficient storage and retrieval. Alternatively, Geman (1985) suggests that a list of (not necessarily consistent) conditional and unconditional probability assessments are obtained, and the 'best' potential representation, in the sense of maximising entropy, is then handled using simulated annealing techniques (Geman and Geman, 1984). In our paper, the clique marginals serve a slightly different purpose, mainly that of being very convenient when current beliefs about states at a given node have to be calculated. These can be obtained easily by further marginalisation of μ_C for a set C containing the variable v.

To summarise this section, we have described four different ways of specifying a joint probability, represented by the expressions (7.1) to (7.4). We now consider the conditioning and marginalisation procedures in the two primary representations: potentials and marginals.

8. BASIC OPERATIONS

Fundamental to our methods are recursive procedures for updating the potential and marginal representations given by (7.2) and (7.4), where 'updating' covers both conditioning on a particular combination of states x_E^* at a set of nodes E, and finding a representation for the marginal probability of configurations at a set of nodes $D \subseteq V$. For convenience, we will always assume $V = D \cup E$.

8.1. *Conditioning in Potentials*

This updating is extremely easy as was demonstrated in Section 5.2. It is based on the fact that

$$p(D \mid E^*) \propto p(D, E^*)$$

i.e. that the conditional probability of the configurations at nodes D, given particular states at nodes E, is proportional to the joint probability with the given value inserted at nodes E. Since we do not bother with computing the normalising constant, the updating is trivial. The new representation (Δ^*, ψ^*) concerns only nodes in $V^* = V \backslash E$ and is obtained as follows. Go through the list Δ, and on finding a set A having $E \cap A \neq \varnothing$, replace A with $A \backslash E$ (if this is not already present), and let

$$\psi^*_{A \backslash E}(.) = \psi_{A \backslash E}(.) \times \psi_A(., x^*_{E \cap A}), \tag{8.1}$$

and leave everything else unchanged. (Here and in the following we have adopted the convention that if there is at any time no potential at a set B, i.e. $B \notin \Delta$, ψ_B is interpreted as being identically equivalent to 1). It follows from the definition that this will be a potential representation of the conditional probability. Note that as a

by-product we have proved that *if p is Markov on a graph, the conditional distribution is Markov on the sub-graph induced by* $V^* = V \backslash E$.

8.2. *Conditioning in Marginals*

This is more complicated and no direct updating method which is computationally feasible seems to exist without restrictions on the set E. Our method here was described by Spiegelhalter (1987) and works if

$$E \subseteq C \text{ for some } C \in \Gamma. \tag{8.2}$$

The procedure was demonstrated in Section 5.4. The first step is to find an ordering of the elements of Γ, say $(C_1, \ldots, C_p)$, such that $E \subseteq C_1$ and the sets have the running intersection property. This is done by maximum cardinality search, as described in Section 6.

The marginal representation for the revised probability p^* can then be computed recursively by letting $\tilde{C}_1 = C_1 \backslash E$ and

$$p^*(\tilde{C}_1) = p(\tilde{C}_1 \mid E^*) = p(\tilde{C}_1, E^*)/p(E^*)$$

and continuing for $k = 2, \ldots, p$:

$$p^*(C_k) = p(C_k)p^*(S_k)/p(S_k) \tag{8.3}$$

again with the convention $0/0 = 0$. Note that because of the running intersection property, $p^*(S_k)$ can be obtained from previously calculated p^*s.

The following shows this *is* a marginal representation for the conditional distribution:

$$\begin{aligned} p^*(\tilde{C}_1) \prod_{i=2}^{p} \frac{p^*(C_i)}{p^*(S_i)} &= p^*(\tilde{C}_1) \prod_{i=2}^{p} \frac{p(C_i)p^*(S_i)}{p^*(S_i)p(S_i)} \\ &= \frac{p^*(\tilde{C}_1, E^*)}{p(E^*)} \prod_{i=2}^{p} \frac{p(C_i)}{p(S_i)} \\ &= p(D \mid E^*). \end{aligned}$$

Here we have used that if $p(S_i) = 0$, then $p^*(S_i) = p(C_i) = p^*(C_i) = 0$, as well as the fact that $p(D \mid E^*) = 0$ if $p^*(S_i) = 0$ for some $i = 2, \ldots, p$. If $\tilde{C}_1 = S_1$, it can be dropped from the new list Γ^* (as is the case in our example, see Fig. 7).

8.3. *Marginalising in Potentials*

From a potential representation (Δ, ψ) we now consider the problem of finding a potential representation $(\bar{\Delta}, \bar{\psi})$ of the marginal distribution on the set of nodes D, i.e. marginalising over E. Letting $\Delta_1 = \{A \mid A \cap E = \varnothing\}$ and $\Delta_2 = \Delta \backslash \Delta_1$, we see from the calculation

$$\begin{aligned} p(D) &= \sum_{E} p(D, E) \\ &= \sum_{E} Z^{-1} \prod_{A \in \Delta} \psi_A \\ &= Z^{-1} \prod_{A \in \Delta_1} \psi_A \sum_{E} \prod_{A \in \Delta_2} \psi_A \end{aligned} \tag{8.4}$$

that the new potential representation can be obtained by letting

$$B = \bigcup_{A \in \Delta_2} (A \backslash E),$$

adding B to the list Δ (if it is not already there) to obtain $\bar{\Delta}$ and calculate the function

$$\phi_B(x_B) = \sum_E \prod_{A \in \Delta_2} \psi_A(x_{A \backslash E}, x_E), \tag{8.5}$$

whereafter we let $\bar{\psi}_B = \psi_B \phi_B$ leaving all other potentials unchanged. This recursion is related to that of Ekholm (1985).

Note that the normalising constant Z is unchanged in this operation and that in the special case where $E = V$, $D = B = \varnothing$, we obtain from (8.5) that $\varnothing_B = Z$.

8.4. *Marginalising in Marginals*

We consider a marginal representation (Γ, μ_C), where Γ is the set of cliques of a (triangulated) graph and we wish to find a similar representation for the marginal distribution at a set of nodes D, marginalising over E.

There is no easy solution in full generality but, as the section heading suggests, in important special cases the problem is trivial. Suppose for example that D is contained in some clique $C \in \Gamma$. Then, obviously

$$p(D) = \sum_{C \backslash D} p(C) = \sum_{C \backslash D} \mu_C(C). \tag{8.6}$$

9. TRANSFER BETWEEN REPRESENTATIONS

As we have seen in Section 8, each representation has its advantages depending on the issue to be considered, and it is therefore desirable to have fast algorithms to move freely between them. In the present section, we shall develop such algorithms corresponding to Fig. 11.

Fig. 11 shows there are algorithms to go from anywhere to anywhere. This means that in principle one can initialise the system by potentials or by marginals, corresponding respectively to the approaches of Cheeseman (1983) and Goldman and Rivest (1986).

9.1. *From Conditional Probability Tables to Potentials*

This has already been described in Sections 4 and 7.2, in that

$$\psi_{\{v\} \cup \Pi_v} = k_v(v \mid \Pi_v), \quad v \in V$$

is a potential representation for p.

9.2. *From Potentials to Set Chain*

We proceed as follows, as exemplified in Section 5.1. First we number the nodes by maximum cardinality search. This in turn orders the cliques of the full moral graph with the running intersection property.

We then use the procedure for marginalising potentials given in Section 8.3 with $E = R_p = C_p \backslash S_p$. This gives us a potential representation of the marginal distribution

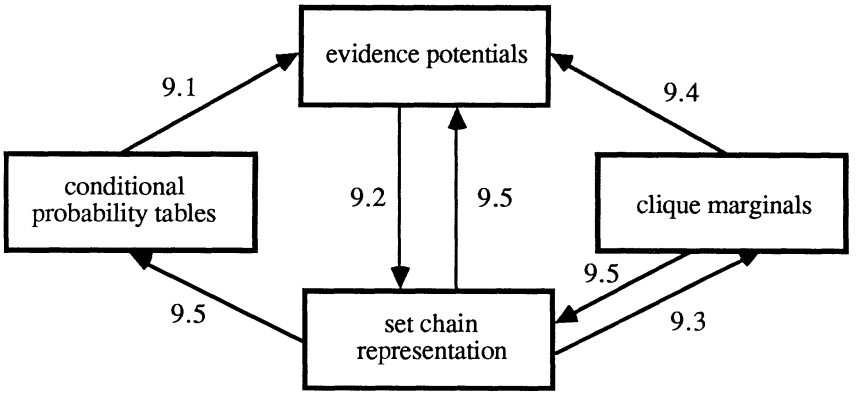


Fig. 11. Transfer between the four representations. The numbers refer to the subsections in which the corresponding algorithm is described.

of the nodes $C_1, \ldots, C_{p-1}$. From (8.4) and (8.5) we obtain that

$$\begin{aligned} p(R_p \mid S_p) &= p(R_p \mid C_1 \ldots C_{p-1}) \\ &= p(V)/p(C_1 \ldots C_{p-1}) \\ &= Z^{-1} \prod_{A \in \Delta} \psi_A \Big/ \left(Z^{-1} \prod_{A \in \Delta_1} \psi_A \sum_E \prod_{A \in \Delta_2} \psi_A \right) \\ &= \prod_{A \in \Delta_2} \psi_A / \phi_B \end{aligned} \tag{9.1}$$

where Δ_2 and ϕ_B are defined in (8.4) and (8.5). Successively repeating this step until $p = 1$ gives the set chain probabilities

$$p(R_i \mid S_i),\ i = 1, \ldots, p.$$

Remembering that $S_1 = \varnothing$ gives that the last calculation of $\phi_B = \phi_\varnothing$ will yield

$$\phi_\varnothing = Z, \tag{9.2}$$

the normalisation constant (see Section 8.3). This is well worth remembering for other purposes.

9.3. *From Set Chain to Marginals*

Here we proceed recursively in the opposite direction by letting

$$p(C_i) = p(R_i \mid S_i) p(S_i) \qquad i = 1, \ldots, p \tag{9.3}$$

and noting that $p(S_i)$ can always be computed from a previously obtained marginal because of the running intersection property. See Section 5.1 for an illustration.

9.4. *From Marginals to Potentials*

It follows directly from (7.4) that a potential representation can be obtained from the marginals by letting

$$\Delta = (C_1, \ldots, C_p, S_1, \ldots, S_m)$$

where S_i, $i = 1, \ldots, m$ are the separators, listed without repetitions. Further, let

$$\psi_C = p(C),\ \psi_S = p(S)^{-\nu(S)}$$

where $\nu(S)$ is the multiplicity of S in the sequence of separators, and zero raised to a negative power is defined to be zero. This representation has $Z = 1$. For the combinatorial properties of the index $\nu(S)$ see Darroch *et al.* (1980) as well as Lauritzen (1984).

9.5. *Other Transfers*

The last transfers (from marginals to set chain, set chain to conditional probability tables, and set chain to potentials) are less interesting for purposes related to the issues in the present paper, and we shall therefore only briefly mention that ordering the sets in Γ with running intersection property and letting

$$p(R_i \mid S_i) = p(C_i)/p(S_i),\ i = 1, \ldots, p$$

gets from marginals to set chain. Ordering the vertices within each set R_i gives a system of conditional probability tables for $v \in R_i$

$$\begin{aligned} k_v(v \mid \Pi_v) &= p(v \mid \Pi_v) \\ &= p(\{v\} \cup \Pi_v \backslash S_i \mid S_i)/p(\Pi_v \backslash S_i \mid S_i) \end{aligned}$$

where $\Pi_v = S_i \cup \{v_{1i}, \ldots, v_{si}\}$ and R_i is ordered as

$$R_i = (v_{1i}, \ldots, v_{si}, v, v_{s+1,i}, \ldots, v_{ti})$$

From this we get the result also shown by Wermuth and Lauritzen (1983) and Kiiveri *et al.* (1984) that *any Markov random field on a triangulated graph is causal Markov on a suitable directed version of the graph.*

A potential representation may be directly obtained from a set chain, without computation, by letting $\Delta = \{C_1, \ldots, C_p\}$, and $\psi_{C_i} = p(R_i \mid S_i)$, cf. also Section 5.1.

10. COMPUTATIONAL ASPECTS

We would claim that our methods can be implemented in a computationally feasible manner in some real-life expert systems, and experience with MUNIN supports this. To give further support, we shall briefly discuss the computational complexity of the methods by giving upper bounds for the number of elementary arithmetic operations needed to perform the various tasks. Recall that the graph-theoretic manipulations could be done in $O(n + e)$ time, where n is the number of nodes and e the number of edges.

We introduce the quantities:

$g = |\Gamma| =$ the number of cliques

$\gamma = \sup_{C \in \Gamma} |C| =$ the maximum number of nodes in a clique

$K = \sum_{C \in \Gamma} |\Omega_C| =$ the total state space

$\Theta = \sup_{C \in \Gamma} |\Omega_C| =$ the largest state space of a clique

and mention that in the MUNIN example (Figs 1 and 9), we have

$$g = 25,\ \gamma = 4,\ K = 2233,\ \Theta = 495.$$

We shall first consider the process of *initialisation*. A potential representation is assumed given, based for example on the conditional probability tables, and the procedures described in Sections 9.2 and 9.3 shall be applied to obtain first a set chain representation and then the clique marginals. Consider first formula (9.1) where the last expression is to be used for calculations, and assume we are about to find $p(R \mid S)$ for a particular clique $C = R \cup S$.

Because of the running intersection property, all elements of $\Delta_2 = \{A \mid A \cap R \neq \varnothing\}$ must be subsets of C and $B = S$. Thus, if we first calculate the products, this will require no more multiplications to be executed than

$$|\Omega_C|(|\Delta_2| - 1) \leqslant |\Omega_C|(2^{|C|} - 2).$$

The quantity ϕ_B can now be obtained from these by adding over $x_E = x_R$. This demands at most

$$|\Omega_B| \times |\Omega_R| = |\Omega_C|$$

additions. Dividing by ϕ_B everywhere uses an additional $|\Omega_C|$ operations. Adding up these terms and recalling that this has to be done once for each clique, we obtain that the whole process can be done with less than

$$\sum_{C \in \Gamma} 2^{|C|}|\Omega_C| \leqslant 2^{\gamma} K$$

elementary arithmetic operations. Here 2^{γ} is considerably over-estimated if the cliques are not almost equal in size.

To obtain the clique marginals, we first compute the separator marginal $p(S)$ from the previously calculated clique marginal $p(D)$ of a possible parent clique D. This demands that for each x_S we must add over $x_{D \setminus S}$, i.e. the number of additions is

$$|\Omega_S| \times |\Omega_{D \setminus S}| = |\Omega_D|.$$

Then the marginal of the clique C is calculated by (9.3) requiring $|\Omega_D|$ operations. A worst case is when $|\Omega_D|$ is equal to Θ for every such D. Then the number of operations needed is no larger than

$$\sum_{C \in \Gamma} (|\Omega_C| + \Theta) = K + g\Theta.$$

Thus the total number of operations needed for the initialisation cannot exceed

$$(2^{\gamma} + 1)K + g\Theta.$$

Here the term $g\Theta$ will be a pessimistic bound in most cases. But the above expression nevertheless indicates that it is worthwhile spending some effort on the initial graph manipulations, keeping K and Θ low, since these are critical for computational feasibility.

It also seems convenient to substitute the original potential representation with that corresponding to the set chain (cf. Sections 5.1 and Section 9.5), i.e.

$$\Delta = \Gamma,\ \phi_C = p(R \mid S) \text{ for } C \in \Gamma.$$

This reduces the number of multiplications to be performed in subsequent applications of the system. We assume in the following that this has been done.

Absorption of evidence takes place via formula (8.1). The worst case appears when E consists of a single node with two states and this node is a member of *all* cliques. Using the particular representation above, (8.1) takes less than $\frac{1}{2}K$ operations.

The *global propagation* is no worse than the initialisation. The *hypothesising* on a set of nodes E that a member of a clique C is done via (8.3) involving calculating $p^*(S)$ and $p(S)$ as well as a multiplication and a division for each element x_C. The total number of operations thus stays less than $\Sigma_{C\in\Gamma}(|\Omega_C| + \Theta + 2|\Omega_C|) = 3K + g\Theta$.

We shall abstain from further calculations of such quantities, and just mention that K and Θ persistently turn up as the critical entities.

It seems that further time can be saved by using an object-oriented programming approach, where objects are cliques, structured in a 'computational tree', generated from an ordering with the running intersection property. Our operations then take the form of cliques sending messages to neighbours in the tree. This avoids, for example, repeated calculation of the same marginals. Andersen *et al.* (1987) give a preliminary account of this approach.

11. EXTENSIONS

In the preceding sections, we have shown how to tackle some basic and important issues in handling causal probabilistic networks. We wish in the present section to indicate possibilities both for computational savings and for solving other problems than those discussed.

11.1. *Collapsibility*

There are situations in which one can calculate marginals fast. The key notion here is *collapsibility* as introduced by Lauritzen (1982) and investigated in detail by Asmussen and Edwards (1983).

Suppose we want to find a clique marginal representation of $p(D)$ for a subset of nodes D. Let $E_1, \ldots, E_k$ be the *connected components* of the subgraph induced by $E = V\backslash D$, i.e. equivalence classes corresponding to the relation 'v and w are equivalent if and only if there is a path from v to w not intersecting nodes in D'. Suppose that each of the sets E_k has a complete boundary in the original graph. Then, p is *collapsible* onto D — meaning that the system $(\bar{\Gamma}, \bar{\mu}_C)$, with

$$\bar{\Gamma} = \{\text{cliques of the subgraph formed by the nodes of } D \text{ with corresponding edges}\}$$
$$\text{and } \bar{\mu}_C = p(C)$$

is a marginal representation of $p(D)$.

Special cases of this collapsibility occur if D is contained in one clique but also if E is a member of one clique only.

11.2. *Calculating a Specific Probability*

Imagine that we have both a potential representation and a marginal representation of a probability p, and assume that we know the normalising constant Z, for example because we have just calculated the marginals by using the procedures described in Section 9.2.

We now want to calculate the probability of a particular configuration at nodes D, say, x_D^*. If D is contained in a clique, we can obtain it directly from the marginal

representation. If this is not the case we should use *restricted* maximum cardinality search as defined in Tarjan and Yannakakis (1984). This is done by first numbering the nodes in D and then proceeding by numbering nodes with the maximum number of numbered neighbours. A corresponding fill-in is then computed. Suppose that this is small, in the sense that only edges between nodes in D have to be filled in. This corresponds to the graph, with edges between D added, being triangulated.

We first look at the case where D is very large compared to the network. In this case, very few marginalisations have to be made and they can be performed on the potentials, bearing in mind that only values corresponding to the state x_D^* have to be calculated. Having quickly obtained a marginalised potential representation (Δ^*, ψ^*) on subsets of D, the desired probability can be calculated directly by

$$p(D^*) = Z^{-1} \prod_{A \in \Delta^*} \psi_A^*$$

since we know Z, and this will be of complexity proportional to the length of the list Δ^*.

If D is relatively small, it might be advantageous to proceed otherwise. Let the nodes be numbered as

$$(1, \ldots, d, d+1, \ldots, d+e, d+e+1, \ldots, d+e+f)$$

where $D = \{1, \ldots, d\}$, $(d+e+1, \ldots, d+e+f)$ are nodes not giving rise to fill-ins; i.e. each of these is a member of exactly one clique when nodes with higher numbers are deleted; and we assume that f is large. Then, as in Section 11.1, the system of marginals is collapsible onto $(1, \ldots, d, d+1, \ldots, d+e)$ and we can quickly compute a marginal representation for the marginal distribution. Using the transfer procedure described in Section 9.4, we obtain a potential representation for the same probability with $Z = 1$ and we can then proceed as described above.

11.3. *Withdrawing the Effect of an Item of Evidence*

Suppose we have observed the nodes in E to take on values x_E^*, and have used this to obtain a potential representation (Δ^*, ψ^*) for $p(D \mid E^*)$ where $D = V \backslash E$. For some reason (perhaps to identify influential observations as in Section 5.6), we want to retract the information that a particular node $e \in E$ is in the state observed, and see how this modifies the probability distribution on the remaining nodes. Introduce the notation

$$F = E \backslash \{e\}, \ \Delta_0 = \{A \in \Delta \mid e \in A\},$$

$$D_0 = \bigcup_{A \in \Delta_0} (A \cap D), \ F_0 = \bigcup_{A \in \Delta_0} (A \cap F),$$

where (Δ, ψ) is the original potential representation. From the factorisation in (7.2) (conditional independence), we obtain that

$$p(D \mid E) \propto p(D, E) = p(D, F) p(e \mid D_0, F_0).$$

If this is different from zero, we obtain from the fact that $p(D \mid F) \propto p(D, F)$ that

$$p(D \mid F) \propto p(D \mid E) / p(e \mid D_0, F_0).$$

But the denominator can be obtained from the original potential representation as

$$p(e \mid D_0, F_0) = \prod_{A \in \Delta_0} \psi_A \Big/ \sum_e \prod_{A \in \Delta_0} \psi_A.$$

Thus a potential representation $(\tilde{\Delta}, \tilde{\psi})$ for $p(D \mid F^*)$ is obtained by letting

$$\tilde{\psi}_{D_0} = \psi^*_{D_0} / p(e^* \mid D_0, F^*_0)$$

and leaving everything else unchanged. The other representations can now be obtained as usual.

11.4. *Some Future Perspectives*

Although we believe we have tackled some basic problems in analysis of probabilistic causal networks, there are still a number of interesting challenges that we shall only touch upon in a tentative fashion.

There may be a demand for the incorporation of *imprecision* of probabilities, propagated to form error bounds on probability outputs, whether the numerical assessments come from expert opinion or from data, or a mixture of both. Inseparable from this is a requirement for automatic *updating* of probabilities as data arrive. An ideal solution would be to consider the probabilities as parameters (Fisher, 1957; Good, 1965; Fung and Chong, 1986) with a prior form which reflected the graphical structure. Standard Bayesian methods would then adjust second-order beliefs as data accumulated. Spiegelhalter (1986b) has suggested introducing unobservable 'probability' nodes as a parent of all nodes in each clique of the graph, generating a discrete marginal distribution on the clique. The usual propagation scheme means that the belief in the probability node can be updated, and retained for the next case. Methods based on a full conjugate prior would, however, be preferable. *Uncertain observations* may be handled by dynamically adding additional nodes (Spiegelhalter, 1986b).

Explanation is a vital objective in expert systems, and it may be thought at first that our approach is somewhat complex in comparison with less formal procedures. However, explanation can take place at many levels. At the most superficial, only the directed graph would be available to the user, with graphical display of probabilities as described in Andreassen *et al.* (1987). Flow of evidence might then be displayed pictorially making use of the update ratios of Table 4. Identification of influential items of evidence is possible as indicated in Section 5.6, and automatic generation of explanatory text could be based on such results (Pearl, 1986a). Only the most determined user need penetrate to the internal manipulations and re-representations being carried out.

A *meta-level* of control may be required to take advantage of specific aspects of the application. Global propagation may be unnecessary, or the flow of evidence through the graph can be pre-structured through the computational tree mentioned in Section 10. It is possible that with the formal underpinning of coherent probabilistic reasoning, heuristic tools may well be appropriate for finding approximations and short-cuts.

Parallel with the problem of initial structuring comes the need for qualitative monitoring and criticism. Additional outputs of a probabilistic system may indicate the area in which attention is required, since a great advantage of probabilistic output is that the system always has a predictive probability for the response to the question it is asking. Hence poorly calibrated prediction may be monitored (Dawid, 1985), say, using proper scoring rules (Winkler, 1969), and consistent 'surprise' for a particular case would indicate an 'outlier', while over a number of cases, build-up of 'surprise' in a particular part of the system may indicate either faults in the numerical assessments or the structuring. The response to either would not be completely automatic, requiring

either *tuning* of the probabilities (non-Bayesian changes) or *learning* about the qualitative structure. Adaptive monitoring and improvement of graphical models is clearly an area in need of study, and may have elements in common with current research into connectionist models for learning (Rumelhart *et al.*, 1986).

Finally, many expert systems incorporate a degree of *taxonomic* structure, with evidence coming in at different levels. Pearl (1986c) has shown how this can be handled probabilistically, and our allowance of logical links appears to circumvent theoretical problems.

On the more technical side, it seems of interest to extend the methods to networks with nodes representing *continuous measurements* and/or with *both directed and symmetric* influences. The results of Lauritzen and Wermuth (1984, 1987) should be applicable. Furthermore it seems that the ideas in Andersen *et al.* (1987), can be used to simplify and extend the theoretical as well as the practical development of the methods.

ACKNOWLEDGEMENTS

We are indebted to members of the MUNIN group, especially S. K. Andersen and F. V. Jensen, for stimulating discussions and permission to discuss their implementation. Jørgen Hilden provided many valuable comments on an earlier version, and the referees made many useful suggestions. We are grateful to F. P. Kelly for reading and commenting on this version and especially to Iris Castleton and Margaret Cowling for preparation of the manuscript.

REFERENCES

Andersen, S. K., Andreassen, S. and Woldbye, M. (1986) Knowledge representation for diagnosis and test planning in the domain of electromyography. In *Proc. 7th European Conference on Artificial Intelligence, Brighton,* pp. 357–368.

Andersen, S. K., Jensen, F. V. and Olesen, K. G. (1987) The HUGIN core—preliminary considerations on systems for fast manipulations of probabilities. In *Proc. Workshop on Inductive Reasoning: Managing Empirical Information in AI-Systems, Risø.*

Andreassen, S., Woldbye, M., Falck, B. and Andersen, S. K. (1987) MUNIN – a causal probabilistic network for interpretation of electromyographic findings. In *Proc. 10th International Joint Conference on Artificial Intelligence, Milan,* pp. 366–372. Kaufmann.

Asmussen, S. and Edwards, D. (1983) Collapsibility and response variables in contingency tables. *Biometrika,* **70,** 567–578.

Averintsev, M. V. (1975) Gibbs description of random fields whose conditional probabilities may vanish. *Prob. Per. Inform.,* **11,** 86–96.

Barr, A. and Feigenbaum, E. A. (1981) *Handbook of Artificial Intelligence,* vols 1 and 2. Los Altos: Kaufmann.

Beeri, C., Fagin, R., Maier, D., Mendelzon, A., Ullman, J. and Yannakakis, M. (1981) Properties of acyclic database schemes. In *Proc. 13th Annual ACM Symposium on the Theory of Computing, Milwaukee.* New York: Association of Computing Machines.

Beeri, C., Fagin, R., Majer, D. and Yannakakis, M. (1983) On the desirability of acyclic database schemes. *J. Ass. Comput. Mach.,* **30,** 479–513.

Berge, C. (1973) *Graphs and Hypergraphs* Transl. from French by E. Minieka. Amsterdam: North-Holland.

Blalock, H. M. (1971) *Causal Models in the Social Sciences.* London: Macmillan.

Buchanan, B. G. and Shortliffe, E. H. (1984) *Rule-based Expert Systems: the MYCIN Experiment of the Stanford Heuristic Programming Project.* Reading: Addison-Wesley.

Cheeseman, P. (1983) A method of computing generalised Bayesian probability values for expert systems. In *Proc. 8th International Joint Conference on Artificial Intelligence, Karlsruhe,* pp. 198–202.

———(1985) In defense of probability. In *Proc. 9th International Joint Conference on Artificial Intelligence, Los Angeles,* pp. 1002–1009.

Cohen, P. R. (1985) *Heuristic Reasoning about Uncertainty: an Artificial Intelligence Approach.* Boston: Pitman.

Darroch, J. N., Lauritzen, S. L. and Speed, T. P. (1980) Markov fields and log-linear models for contingency tables. *Ann. Statist.,* **8,** 522–539.

Dawid, A. P. (1985) Calibration-based empirical probability. *Ann. Statist.*, **13**, 1251–1285.
DeGroot, M. H. (1970) *Optimal Statistical Decisions.* New York: McGraw-Hill.
Dempster, A. P. and Kong, A. (1986) Uncertain evidence and articial analysis. *Research Report S-108.* Department of Statistics, Harvard University.
Dirac, G. A. (1961) On rigid circuit graphs. *Abh. Math. Sem. Univ. Hamburg*, **25**, 71–76.
Doyle, J. (1979) A truth maintenance system. *Artific. Intell.*, **12**, 231–272.
Duda, R. O., Hart, P. E. and Nilsson, N. J. (1976) Subjective Bayesian methods for rule-based inference systems. *Proc. AFIPS Natl Comput. Conf.*, **47**, 1075–1082.
Ekholm, A. (1985) A recursion formula for the log-linear parameters of a collapsed contingency table. *Research Report 53.* Department of Statistics, University of Helsinki.
de Finetti, B. (1974) *Theory of Probability*, vols 1 and 2. New York: Wiley.
Fisher, R. A. (1957) The underworld of probability. *Sankhya*, **18**, 201–210.
Fox, J. (1986) Three arguments for extending the framework of probability. In *Uncertainty in Artificial Intelligence* (eds L. N. Kanal and J. Lemmer), pp. 447–458. Amsterdam: North-Holland.
Fung, R. M. and Chong, C. Y. (1986) Metaprobability and Dempster-Shafer in evidential reasoning. In *Uncertainty in Artificial Intelligence* (eds L. N. Kanal and J. Lemmer), pp. 295–302. Amsterdam: North-Holland.
Gavril, T. (1972) Algorithms for minimum coloring, maximum clique, minimum coloring by cliques and maximum independent set of a chordal graph. *SIAM J. Comput.*, **1**, 180–187.
Geman, S. (1985) Stochastic relaxation methods for image restoration and expert systems. In *Automated Image Analysis: Theory and Experiments* (eds D. B. Cooper, R. L. Launer and D. E. McClure). New York: Academic Press.
Geman, S. and Geman, D. (1984) Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Machine Intell.*, **6**, 721–741.
Goldman, S. A. and Rivest, R. L. (1986) Making maximum entropy computations easier by adding extra constraints. In *Proc. 6th Annual Workshop on Maximum Entropy and Bayesian Methods in Applied Statistics.*
Golumbic, M. C. (1980) *Algorithmic Graph Theory and Perfect Graphs.* London: Academic Press.
Good, I. J. (1965) *The Estimation of Probabilities.* Cambridge: Massachusetts Institute of Technology Press.
Goodman, L. A. (1970) The multivariate analysis of qualitative data: interaction among multiple classifications. *J. Amer. Statist. Ass.*, **65**, 226–256.
Hajek, P. (1985) Combining functions for certainty degrees in consulting systems. *Int. J. Man-Machine Studies*, **22**, 59–65.
Heckerman, D. (1986) Probabilistic interpretations for MYCIN's certainty factors. In *Uncertainty in Artificial Intelligence* (eds. L. N. Kanal and J. Lemmer), pp. 167–196. Amsterdam: North-Holland.
Henrion, M. (1987) Uncertainty in artificial intelligence: is probability epistemologically and heuristically adequate? In *Expert Systems and Expert Judgement* (ed. J. Mumpower). New York: Springer.
Hilden, J. (1970) GENEX—an algebraic approach to pedigree probability calculus. *Clinical Genetics*, **1**, 319–348.
———(1982) Computerized derivations of Mendelian probability formulae: the GENEX processor. In *Nordic Symposium in Applied Statistics and Data Processing* (eds A. Höskuldsson *et al.*), pp. 395–410. Lyngby: NEUCC – Technical University of Denmark.
Jensen, F. V., Andersen, S. K., Kjaerulff, U. and Andreassen, S. (1987) MUNIN: on the case for probabilities in medical expert systems—a practical exercise. In *Proc. 1st Conference of European Society for Artificial Intelligence in Medicine* (eds. J. Fox, M. Fieschi and R. Engelbrecht), pp. 149–160. Heidelberg: Springer.
Jirousek, R. and Perez, A. (1986) A partial solution of the marginal problem. In *Trans. 10th Prague Conference on Information Theory.*
Jöreskog, K. G. (1973) Analysis of covariance structures. In *Proc. 3rd Symposium on Multivariate Analysis* (ed. Krishnaiah), pp. 263–283. New York: Academic Press.
Kanal, L. N. and Lemmer, J. (eds) (1986) *Uncertainty in Artificial Intelligence.* Amsterdam: North-Holland.
Kellerer, H. G. (1964a) Masstheoretische Marginalprobleme. *Math. Ann.*, **153**, 168–198.
———(1964b) Verteilungsfunktionen mit gegebenen Marginalverteilungen. *Z. Wahrsch. Geb.*, **3**, 247–270. *verw. Geb.*, **3**, 247–270.
Kelly, C. W. and Barclay, S. (1973). A general Bayesian model for hierarchical inference. *Organizational Behaviour and Human Performance*, **10**, 388–403.
Kiiveri, H., Speed, T. P. and Carlin, J. B. (1984) Recursive causal models. *J. Aust. Math. Soc.*, **36**, 30–52.
Kim, J. H. and Pearl, J. (1983) A computational model for causal and diagnostic reasoning in inference systems. In *Proc. 8th International Joint Conference on Artificial Intelligence, Karlsruhe*, pp. 190–193.
Kong, A. (1986) *Multivariate Belief Functions and Graphical Models.* PhD Thesis, Department of Statistics, Harvard University.
Kuipers, B. and Kassirer, J. P. (1983) How to discover a knowledge representation for causal reasoning by studying an expert physician. In *Proc. 8th International Joint Conference on Artificial Intelligence, Karlsruhe*, pp. 49–56.
Lauritzen, S. L. (1982) *Lectures on Contingency Tables*, 2nd edn. University of Aalborg Press.
Lauritzen, S. L., Speed, T. P. and Vijayan, K. (1984) Decomposable graphs and hypergraphs. *J. Aust. Math. Soc.* **A, 36**, 12–29.

Lauritzen, S. L. and Wermuth, N. (1984) Mixed interaction models. *Research Report R-84-8.* Institute of Electronic Systems, Aalborg University.

———(1987) Graphical models for associations between variables, some of which are qualitative and some quantitative. *Research Report R-87-10.* Department of Mathematics and Computer Science, Aalborg University.

Leimer, H.-G. (1985) Strongly decomposable graphs and hypergraphs. Thesis. *Ber. Stochast. Verw. Geb.* 85-1 University of Mainz.

———(1988) Optimal decomposition by complete separators. *Discrete Mathematics* (to appear).

Lemmer, J. (1983) Generalized Bayesian updating of incompletely specified distributions. *Large Scale Systems,* **5**, 51–68.

Lindley, D. V. (1982) Scoring rules and the inevitability of probability. *International Statistical Review,* **50**, 1–26.

———(1987) The probability approach to the treatment of uncertainty in artificial intelligence and expert systems. *Statistical Science,* **3**, 17–24.

Malvestuto, F. M. (1987) Answering queries in categorical data bases. *Proc. 6th ACM Symposium on Principles of Database Systems, San Diego.*

Miller, R. A., Pople, H. E. Jr and Myers, J. D. (1982) INTERNIST-1, an experimental computer-based diagnostic consultant for general internal medicine. *New England Journal of Medicine,* **307**, 468–476.

Moussouris, J. (1974) Gibbs and Markov random systems with constraints. *J. Statistical Physics,* **10**, 11–33.

Patil, R. S., Szolovits, P. and Schwartz, W. B. (1982) Modelling knowledge of the patient in acid-base and electrolyte disorders. In *Artificial Intelligence in Medicine* (ed. P. Szolovits), pp. 191–226. Colorado: Westview.

Pearl, J. (1982) Reverend Bayes on inference engines: a distributed hierarchical approach. *Proc. AAAI National Conference on AI, Pittsburgh,* pp. 133–136.

———(1986a) Fusion, propagation and structuring in belief networks. *Artificial Intelligence,* **29**, 241–288.

———(1986b) Bayes and Markov networks: a comparison of two graphical representations of probabilistic knowledge. *Technical Report R-46.* Cognitive Systems Laboratory, UCLA.

———(1986c) On evidential reasoning in a hierarchy of hypotheses. *Artificial Intelligence,* **28**, 9–15.

Pople, H. E. (1982) Heuristic methods for imposing structure on ill structured problems: the structuring of medical diagnosis. In *Artificial Intelligence in Medicine* (ed. P. Szolovits), pp. 119–185. Colorado: Westview.

Rose, D. J., Tarjan, R. E. and Lueker, G. S. (1976) Algorithmic aspects of vertex elimination on graphs. *SIAM J. Comput.,* **5**, 266–283.

Rumelhart, D. E., Hinton, G. E. and Williams, R. J. (1986) Learning representations by back-propagating errors. *Nature,* **323**, 533–536.

Savage, L. J. (1972) *The Foundations of Statistics,* 2nd edn. New York: Dover Publications.

Schwartz, W. B., Patil, R. S. and Szolovits, P. (1987) Artificial intelligence in medicine: where do we stand? *New England Journal of Medicine,* **316**, 685–688.

Shachter, R. D. (1986) Intelligent probabilistic inference. In *Uncertainty in Artificial Intelligence* (eds L. N. Kanal and J. Lemmer), pp. 371–382. Amsterdam: North-Holland.

———(1987) Probabilistic inference and influence diagrams. *Operations Research* (to appear).

Shafer, G. (1976) *A Mathematical Theory of Evidence.* Princeton: Princeton University Press.

———(1987) Probability judgement in articial intelligence and expert systems. *Statistical Science,* **3**, 3–16.

Shafer, G., Shenoy, P. P. and Mellouli, K. (1986) Propagating belief functions in qualitative Markov trees. *Int. J. Approx. Reasng,* **1**, 349–400.

Smith, A. F. M. (1984) Discussion on Statistical and knowledge-based approaches to clinical decision-support systems, with an application in gastroenterology. *J. R. Statist. Soc.* A, **147**, 61.

Smith, J. Q. (1987) Influence diagrams for statistical modelling. *Research Report 117.* Department of Statistics, University of Warwick.

Speed, T. P. (1979) A note on nearest-neighbour Gibbs and Markov probabilities. *Sankhya* A, **41**, 184–197.

Spiegelhalter, D. J. (1986a) A statistical view of uncertainty in expert systems. In *Artificial Intelligence and Statistics* (ed. W. Gale), pp. 17–56. Reading: Addison-Wesley.

———(1986b) Probabilistic reasoning in predictive expert systems. In *Uncertainty in Artificial Intelligence* (eds L. N. Kanal and J. Lemmer), pp. 47–68. Amsterdam: North-Holland.

———(1986c) Computers, expert systems and adverse drug reactions: can causality assessment be automated? *Drug Information Journal,* **20**, 543–550.

———(1987) Coherent evidence propagation in expert systems. *Statistician,* **36**, 201–210.

Spiegelhalter, D. J. and Knill-Jones, R. P. (1984) Statistical and knowledge-based approaches to clinical decision-support systems, with an application in gastroenterology *J. R. Statist. Soc.* A, **147**, 35–77.

Szolovits, P. and Pauker, S. G. (1978) Categorical and probabilistic reasoning in medical diagnosis. *Artificial Intelligence,* **11**, 115–144.

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

Vorob'ev, N. N. (1962) Consistent families of measures and their extensions. *Theory of Probability and Applications,* **7**, 147–163.

———(1963) Markov measures and Markov extensions. *Theory of Probability and Applications,* **8**, 420–429.

Weiss, S. M., Kulikowski, C. A., Amarel, S. and Safir, A. (1978). A model-based method for computer-aided decision-making. *Artificial Intelligence,* **11**, 145–172.

Wermuth, N. and Lauritzen, S. L. (1983) Graphical and recursive models for contingency tables. *Biometrika*, **70**, 527–552.
Winkler, R. L. (1969) Scoring rules and evaluation of probability assessors. *J. Amer. Statist. Ass.*, **64**, 1073–1078.
Wold, H. D. A. (1954) Causality and econometrics. *Econometrica*, **28**, 443–463.
Wright, S. (1921) Correlation and causation. *J. Agric. Res.*, **20**, 557–585.
———(1934) The method of path coefficients. *Ann. Math. Statist.*, **5**, 161–215.
Yannakakis, M. (1981) Computing the minimum fill-in is NP-complete. *SIAM J. Algebraic Discrete Methods*, **2**, 77–79.
Zadeh, L. A. (1983) The role of fuzzy logic in the management of uncertainty in expert systems. *Fuzzy Sets and Systems*, **11**, 199–228.
———(1986) Is probability theory sufficient for dealing with uncertainty in AI: a negative view. In *Uncertainty in Artificial Intelligence* (eds L. N. Kanal and J. Lemmer), pp. 103–116. Amsterdam: North-Holland.

DISCUSSION OF THE PAPER BY LAURITZEN AND SPIEGELHALTER

Dr F. P. Kelly (University of Cambridge): This evening's paper is an important step towards the extension of the powerful knowledge representation techniques of expert systems to handle uncertainty, and the authors should be particularly congratulated on their positive and constructive approach. The arguments for and against the probability calculus paradigm have been well rehearsed: I share the authors' belief that the best argument in favour of probability will be a demonstration that it can be made to work.

The knowledge-base of an expert system is usually expressed in terms of a collection of if–then statements, and this makes natural the representation of probabilistic knowledge by a causal network, together with a system of conditional probability tables. However, the structure at the heart of the computational techniques described is a Markov field on a triangulated graph, with the causal network appearing as just a favoured method of specifying the field. In between these two structures is the minimal Markov field, defined on the minimal graph with respect to which the joint probability distribution is a Markov field. Uniqueness of this minimal graph follows from the factorisation theorem contained in the early work of Brook (1964) on the relationship between conditional and joint probability specifications. For the example of Fig. 2 and Table 1 the minimal graph is the moral graph of Fig. 3. The minimal graph may well be smaller than the moral graph: for example in Table 1 if $p(\delta \mid \varepsilon, \beta)$ admits a factorisation

$$p(\delta \mid \varepsilon, \beta) \propto \psi(\delta, \varepsilon)\psi(\delta, \beta) \quad \text{over } \delta$$

then the minimal graph would not have a link between nodes ε and β. It is interesting here to contrast the instincts of the system modeller and the statistician: the former may prefer a general conditional probability, while the latter may lean towards more parsimonious representations and may be prepared to accept a factorisation unless there is prior evidence for something more complex. Factorisation will lead to fewer edges in the minimal graph and may ease the computational burden.

The authors make clear that for their computational procedures to be feasible the cliques of the triangulated graph should not be too large. The MUNIN example suggests that this may often be the case, but what if it is not? Pearl, an important writer in this area, has a number of suggestions. One of his suggestions (Pearl, 1986a) is that we condition on a separating subset of nodes sufficient to break long cycles; this will allow the authors' procedures to be applied for each conditioning instance, and will be attractive provided that the state space for the subset of separating nodes is not too big. There are various ways in which the approach can be extended, but it remains the case that any general exact analysis will need to deal with probability distributions over large state spaces. This raises the question of whether there are approaches other than an exact analysis. We mention two.

We could, as proposed by Pearl (1987a), simulate a stochastic process which has the minimal Markov field as its stationary distribution. Sampling from the process will then give noisy estimates of the quantities of interest. It is easy to construct an appropriate stochastic process as a locally interacting particle system on the minimal graph, and the method is ideal for a parallel processing implementation. The method deals readily with the absorption of evidence: to condition on the values taken at a set of nodes just hold the nodes fixed at the given values throughout a simulation.

In statistical physics and communication network modelling Markov fields on multiply connected graphs are common, and in both these areas approximations have often been very successful. In the present context an obvious approximation technique would be to iterate around long cycles. There is concern that such iterations may not converge or may lead to instabilities. Experience in other areas suggests that it is quite easy to construct rapidly convergent iterative schemes. Multiple solutions can

occur; however, they are generally not artefacts of the approximation, but indicate potential instabilities in the Markov field. Such instabilities are unlikely in the present context, where they would correspond to a small amount of evidence causing a wholesale revision of beliefs.

While I have been speaking there has not been a wholesale revision of my own beliefs. I still consider the paper to be an important and positive step towards the implementation of expert systems that can genuinely deal with probabilistic information, and it gives me great pleasure to propose the vote of thanks to Professor Lauritzen and Dr Spiegelhalter.

Professor D. J. Hand (The Open University, Milton Keynes): My comments will be restricted to the expert systems and artificial intelligence side of the paper.

First, there is a slowly growing realisation that expert systems, as conventionally described, have a far more limited applicability than the proponents would like us to believe.

At least two reasons have been proposed for this. White (1987) suggests that success is limited to linguistically constrained, relatively mechanistic, and logical domains, e.g. areas describing systems built by humans, such as fault diagnosis or robotics. He cites the Spang Robinson (1986) survey of expert systems products in the USA as showing that two-thirds of applications work is in such areas. The point is that the conventional expert systems approach may not be or at least has not yet been shown to be well matched to more complex, or less well-defined, problems such as medical diagnosis.

The second suggested reason is described by Coombs and Alty (1984) (see also Hand (1987)). They suggest that there is a mismatch between the function that human experts carry out and the role that current expert systems are being built to play. In particular, while expert systems are typically intended to produce solutions to complex but essentially well-defined problems, human experts are more often called on to provide advice and conceptual guidance, such things as showing context, identifying important topics and indicating what would happen under different conditions.

One of the exciting things about the present paper is that the methods that the authors describe may well provide the basic substratum representation for a system which can act in this more flexible manner.

The authors only briefly refer to the actual construction of the network, and yet surely in building an expert system this is the most important part. Certainly rule elicitation is proving to be one of the most challenging aspects of current expert systems research, with much effort going into automatic rule induction methods. No really satisfactory solution has yet been found.

The situation is aggravated by the fact that real world problems change over time and that the much advertised extensibility of conventional expert systems is more fragile than it seems. White (1987), for example, cites the story of what is one of the most successful expert systems in the world, the R1 system for configuring DEC computer systems: 'Originally only 450 OPS rules, it grew to an effective program of about 5000, configuring DEC/VAX computer installations profitably! It subsequently expanded to about 7000 rules whereupon it became unmanageable, unmaintainable, and unmodifiable'.

Have the authors any ideas on how such difficulties can be avoided?

The authors stress the relevance of the large sparse network representation. However, I am not convinced that it is such a ubiquitously good representation. Although it may indeed be the case that, as Pearl (1986a) argues, humans implicitly use sparse networks, it is not clear to me that this is necessarily the best way to represent the structure of problems. In particular, many problems better match a broad shallow network with many parents for each child and with a short maximum path length. If diagnosis is the aim, such representations fall more immediately into the simple pattern matching category of problems. The issue is one of whether we are trying to solve problems or trying to emulate the way that humans solve problems.

A second point is that, although the large sparse network approach is a very reasonable way of retaining great flexibility in the form of the model, I am anxious that the flexibility may be too great and that there may be an overfitting problem. Similar issues have cropped up elsewhere in rule-based expert systems and in other areas of statistical classification such as nonparametric methods (e.g. Titterington *et al.* (1981)). I look forward with great interest to some comparisons of the authors' system with more conventional statistical diagnostic approaches.

Finally, in a broader context, many of the objections to artificial intelligence involve rather philosophical issues, such as whether a machine can be said to 'know' or to 'understand', etc. However, there is one genuine technical problem which is a real stumbling-block. This, termed the 'frame problem', is the question of how to separate the relevant from the irrelevant in 'intelligent' programs. Thus, when I pick up my brief-case from my desk I do not want the program to examine every object in the universe to see how each is affected. Yet it must know that the papers within the brief-case are also picked up,

while the desk remains where it is, as does the chair on the other side of the room, etc. The authors' approach to maintaining consistency of a large collection of entities is intimately connected with this. I would be very interested in the authors' comments.

The techniques that the authors have developed are elegant and convincing. It is clear that they will have a big impact on the expert systems community and in doing so they will serve to strengthen and promote the discipline of statistics. The authors thus deserve hearty congratulations for producing this work and I can sincerely say that it gives me great pleasure to second the vote of thanks.

The vote of thanks was passed by acclamation.

Mr A. R. Thatcher (New Malden): The success of the authors' method depends on three assumptions.

(a) A causal network exists and is known. This also assumes that at least some of the causes or diseases or symptoms are independent.
(b) There are experts who can supply all the required probabilities.
(c) These probabilities are all conditioned on the event that the patient was presented at the clinic.

These assumptions are all satisfied in MUNIN, but there may be other applications where they are not and it is worth spending a moment to consider whether anything can then be done to try to avoid them.

If it is not easy to find experts who can provide all the probabilities reliably, an alternative is to infer the probabilities from actual data, if we have sufficient. If the records of the previous patients at the clinic are available for this, these will have the advantage that they automatically satisfy assumption (c). Also, they will not involve any assumptions about independence as in (a).

There is no need to estimate every probability in the network. All that we require, for a particular new patient, are the probabilities which answer the particular questions we wish to ask about him and about the possible choices for the next step. If we work through the calculations completely from first principles, applying Bayes' theorem without assuming independence and taking joint probability distributions directly from the data, the result which finally emerges is very simple indeed (Thatcher, 1988). As an example, suppose that our new patient is the one in Fig. 8 of the paper and that our first question is about the probability that he has tuberculosis. All we have to do is to sort through the records to find the past patients who were like him, i.e. those who had visited Asia, had a negative X-ray, etc. The proportion of these matching past patients who had tuberculosis now gives an immediate estimate of the probability that our new patient will have tuberculosis.

The confidence limits for this proportion can be calculated by the standard method and give an immediate measure of the precision of the estimate. As regards complexity, the number of computer operations per question per new patient is of the order of NAP where N is the number of nodes, A is the average number of states per node and P is the number of past patients.

Dr J. Q. Smith (University of Warwick): By using 'evidence potentials' the authors have successfully combined Markov field theory with *ad hoc* but practically more useful graphical methods for manipulating probabilities, providing a theoretically sound and usable methodology for probabilistic expert systems. Potentials are attractive because they both enable the manipulation of conditional independence (CI) and also provide a structure for the calculation of useful probabilities. Unlike alternative direct methods of CI manipulation (Pearl, 1986b; Smith, 1987a, b), however, they are not able to cope with, for example, mixed distributed nodes.

Causal networks are useful in at least three areas.

(a) Although the theory of probability elicitation is well developed the *elicitation of model structure* has received little attention. In this paper it was shown how a set of modelling statements could be transformed into a causal network which can be manipulated so that the consequences of implicit CI statements can be presented to a client to enable him to modify his model (Smith, 1987a). Only after an elicitation of the structure of a problem should the elicitation of probabilities take place. Incidentally graphical methods etc. are now theoretically developed (Pearl, 1986b; Smith, 1987a, b, 1988).
(b) Casual networks can be used to help to *prove* theorems about complex decision problems—e.g. see Smith (1988).
(c) Networks help to direct the efficient storage and manipulation of probabilities as discussed here.

Although causal networks are useful they are not universally so. In practice useful models sometimes only have sparse networks *conditional on* some other variable(s).

Example. Identically distributed random variables $X_1, \ldots, X_n$ conditional on $X_0 = nk + r$, $1 \leqslant k \leqslant n$, $1 \leqslant r \leqslant n$, have the property $X_k = X_r$ and $\perp\!\!\!\perp X_i$, $i = \{1, \ldots, n\} \backslash r$.

This idealised distribution is inefficiently stored on a complete network. There is a need for the development of more general structures combining causal networks with probability trees. Can computational efficiency determine when to use such a hybrid system?

Finally the introduction of extra variables (like ε) is very useful, indeed central to Bayesian modelling (Smith, 1987b). Do the authors have a computational algorithm for determining when functional nodes should be introduced?

Dr Frank Critchley (University of Warwick): The field of expert systems is at the beginning of the long road towards maturity. The spirit of my remarks is therefore twofold: first, to comment on some possibly fruitful directions in which to make this journey (cf. Section 11.4) and, secondly, to sound some cautionary notes about potential dangers.

Let N denote the total number of mathematically independent quantities needed to specify a probability distribution on all possible combinations of node states in a causal network. Even for the modest (one muscle, 25 nodes) MUNIN network, I made N to be 1161. For the kinds of network ultimately envisaged, values of N of the order of 10^5 or 10^6 would be commonplace, i.e. without *further* structure, probability models for causal networks are not parsimonious, and so the prior distribution does matter. It would take an enormous amount of data to swamp it. Equally, an empirically based initial distribution would need an enormous amount of pre-existing data for estimability, let alone accuracy. Moreover there is the danger that *any* (however unreliable) input produces apparently precise output. What naive user is going to argue when 0.78653 flashes up?

I was pleased to note that, perhaps under the influence of a recent Society discussion paper, the authors place a great weight on morals. I hope that they will bear with me, then, if I draw a few morals from the above perceived dangers. We should

(a) ensure that the prior/initial distribution is specified as accurately as possible ... and, even then, ...
(b) express our uncertainty about these probabilities. It is simply unrealistically optimistic to assume precise knowledge about 10^6 quantities. The scale of this uncertainty, in practical problems trivial by comparison with those in the paper, can be enormous: see, for example, Critchley and Ford (1985) and Critchley *et al.* (1988).

Both (a) and (b) will be aided if we

(c) use intelligent parsimonious submodels, e.g. of particular conditional distributions, based on ...
(d) ... rich families of discrete distributions for probabilities: helpful progress here is reviewed in Aitchison (1986).

Nevertheless,

(e) we will still need *criticism*, for our own good. There is no virtue in deceiving ourselves! Because of its discrete nature, studying the influence of the network structure itself is a challenging new research problem within this domain.

Finally,

(f) *beware computer power*! ... in that, as this power increases, so does the scope for the dangerous shift in emphasis from the question 'is the model sensible?' to 'will it compute?' The danger in making this shift unthinkingly is to mistake a necessary condition for a sufficient one and wrongly to identify the complexity of a model with its worth.

In sum, I found the paper a theoretically and computationally impressive contribution to a subject that is still in its infancy. May it prosper on its journey towards maturity!

Professor A. F. M. Smith (University of Nottingham): From a Bayesian perspective, *planning* (or design) is a preposterior activity—what questions to ask and in what order?—and assessing *influence* is a posterior activity—which answers had the most (or least) effect on beliefs, given the other data? Both are of great importance, but what they have in common, in addition to this quasi-dual conceptual relationship, is that, in the context of the authors' models, they pose extremely challenging computational problems. These are not fully recognised or discussed in Section 10. Do the authors see here a potential use for novel simulated annealing strategies?

Although the authors make clear that it is not the central issue of this paper, the problem of *model comparison* within the structures envisaged here is going to be of great importance. In the context of more familiar models (e.g. regression) we typically have a clear understanding of the (e.g. linear space) mathematical operations that are implicit in simplifying or elaborating structure. In the present context, how is the interplay of graph representation and propagation algorithm affected by structural perturbations corresponding to common strategies for model simplification, e.g. replacing a set of nodes 'test $k = t$?' by '$\cup_k$\{test $k = t$?\}?'

Section 11.4 acknowledges the need for eventual modelling of *probabilities as parameters* and mentions the desirability of full conjugate analysis. Could the authors comment on the nature and role of exponential family types of model in the causal network framework? More generally, complex multiparameter structures often suggest the use of some form of hierarchical prior specification. However, the dependencies induced by collapsing stages in such a hierarchy would seem potentially to destroy the local conditional independence that is so fundamental to the authors' approach. Some further comments on probabilities as parameters would be helpful.

As with image processing, the area of expert systems has often been in danger of slipping away from the purview of the statistical community. In the image processing context, use of the Bayesian formalism, combined with creative modelling and algorithm development exploiting local dependency structures, has in recent years provided a clear demonstration of the relevance and power of the mathematical statistical perspective and has been a major influence on developments in that field. I believe that this fundamentally important paper will have a seminal influence on future developments in handling uncertainty in expert systems.

Dr Jørgen Hilden (Panum Institute, Copenhagen): The paper proves that outcome spaces which are structurally rich are valuable objects of scientific efforts. Probabilists have been much too fascinated by esoteric, almost unstructured mathematical spaces.

We also need *action nodes* with state spaces like {drug A, drug B, no drug}. Suitable dummy probabilities can be assigned, it being understood that only action-conditional results make sense. Computing the *expectation* $E\{f(X_v, v \in V)\}$, where $f(\cdot)$ may be survival time or perhaps *utility*, is easy if $f(\cdot)$ is a sum of a few terms, each depending on a single clique. However, $f(\cdot)$ can also be stored as a potential, and no new subroutines are then needed. For comparing expected utilities we do not even need Z.

As to the *predictive power* of a diagnostic test (Section 5.5), expected increments in logarithmic score (Kullback–Leibler entropy differences) are inappropriate for practical decision making, as opposed to inference (Glasziou and Hilden, 1988). They effectively assume that absolute diagnostic certainty is infinitely useful, which is false because medical utilities are always bounded. If, for instance, the probability of lung cancer is already high, we operate, and nothing is gained by refining the preoperative probability assessment slightly. For lack of a true utility function I suggest using the *quadratic scoring rule* (Hilden *et al.*, 1978). Its expected increment is

$$M_{\text{quadratic}}(u, \lambda) = \sum_u p(u) \sum_\lambda [p(\lambda \mid u) - p(\lambda)]^2,$$

which is preferable, also computationally, to its logarithmic counterpart.

The *update ratios* $UR_k(C_k) = p^*(S_k)/p(S_k)$, Section 8.3, are *p-weighted averages* of those of the parent clique (C_i, say):

$$UR_k(C_k) = E_p\{UR_i(C_i) \mid S_k\}, \text{ where } S_k = C_k \cap C_i.$$

Thus the effect, however measured, of observation E^* is necessarily attenuated as we move towards the branches of the clique tree. What is the worst case ill effect of breaking off the propagation when $|UR - 1| < \varepsilon$ (a chosen threshold)?

Program sentinels should be posted to look out not only for

(a) *incompatible data* ($Z = 0$), but also for
(b) *eliminable nodes* ($X_v = x_v^*$ with probability 1), and for
(c) other *accidental factorisations* ('breakable cliques').

Accidental factorisations may arise when 'conditional conditional' independencies exist and the outermost condition attains probability 1. I would welcome a formal apparatus for this. For uniform handling, the programmer will no doubt retain eliminated nodes as disconnected subgraphs. In Section

8.1 (equation (8.1)) and Section 8.3 (equation (8.3)) he may store

$$\psi_E^* = p^*(E) = I(E = E^*).$$

Dr Wilfrid S. Kendall (University of Strathclyde): Cheap but powerful personal computers, supporting computer algebra packages such as REDUCE and muMath, open up exciting possibilities for probabilists and statisticians. It would be interesting to apply computer algebra to the topic of this paper. Crucial probabilities and conditional probabilities would be left as undetermined parameters and carried through computations directed by the graphs and graphical constructions described in the paper. As always in computer algebra, it would be necessary to evade problems of overflow. Unnecessary computer algebra overflow corresponds loosely to prohibitively long numerical calculations of the kind mentioned after equation (4.1). The techniques described here should help considerably, when coupled to judicious inhibition of algebraic expansion algorithms. Nevertheless we would expect that the symbolic approach would be more limited than the numerical approach.

The symbolic approach offers practical rewards such as immediate and direct calculation of influence by means of symbolic differentiation. In the longer term we would envisage a theory of symbolic probability dealing with the interplay of abstract probability, graph theory and symbolic computation. On a mundane note there would be considerable pedagogic advantages in illustrating the fundamentals of probability theory by interacting with symbolic implementations of substantial expert systems. Such examples would counterbalance the trivialising influence of the ubiquitous coin, die and card!

Inspired by the paper, I have made a faltering beginning using the excellent muMath package. Combined with a cheap IBM compatible computer, this package yields a remarkably flexible small computer algebra system. Unfortunately my results are still far too incomplete to be discussed here.

On a different note I wonder whether it is practical to investigate methods for detecting possibilities for replacing a given expert system by another using fewer nodes and edges. Discriminant analysis and logistic regression offer possibilities, but lead away from the sparse graphical structure.

Finally here is a somewhat unconventional way to display imprecision of probabilities. Consider transformations of the probabilities (such as log-odds) as varying randomly with time according to an Ornstein–Uhlenbeck diffusion. Set the long-term mean at the original probability specification and set the diffusion parameters to reflect imprecision. A specified set of results can then be displayed in animation as fluctuating randomly about levels calculated on the basis of the original specifications. Fluctuation statistics could be derived using the Ito calculus (perhaps using the symbolic Ito calculus described in Kendall (1988)).

Mr K. G. Olesen and Dr S. K. Andersen (Aalborg University): The authors have made probabilistic propagation of evidence in causal networks with loops computationally feasible. We acknowledge the work as an essential prerequisite for the HUGIN core—a general tool for handling uncertainty in probabilistic networks.

We should like to address some possible practical simplifications in the scheme of evidence propagation.

In the 'simple example' outlined in Section 5 a maximum cardinality enumeration of vertices of the graph is made several times: first, to perform the initialisation and secondly when propagating evidence. In this way the linear running intersection property of the cliques in the triangulated graph is achieved. As an alternative, the maximum cardinality ordering performed at the initialisation can be used to establish a junction tree as described in the comment of Dr Jensen. Using this approach we obtain a static structure which yields an ordering that is usable whenever evidence becomes available.

The initial effort of creating a junction tree gives a structure of cliques tied together by separation sets in which all run time operations such as absorption and propagation are performed. The underlying qualitative structure remains the same independently of the representation used.

As an example the global propagation of multiple evidence (Section 5.3) can be performed in the static clique structure. All evidences are propagated to some clique where they are joined (Fig. 12). From this clique, now acting as a root in the junction tree, evidence is propagated to the whole tree, the basic operation being similar to the propagation of simple items of evidence (Section 5.4).

The crucial point is when evidence from different branches is merged. Here the order in which the different steps of the calculations are performed is essential.

In the 'simple example' outlined in Section 5 evidence arrives at a and d. This evidence is joined in one clique, say $(\tau, \lambda, \varepsilon)$, which now acts as a root and from here the whole system is calibrated. It is not necessary to propagate to the 'leaf' cliques which contain α and δ, though.

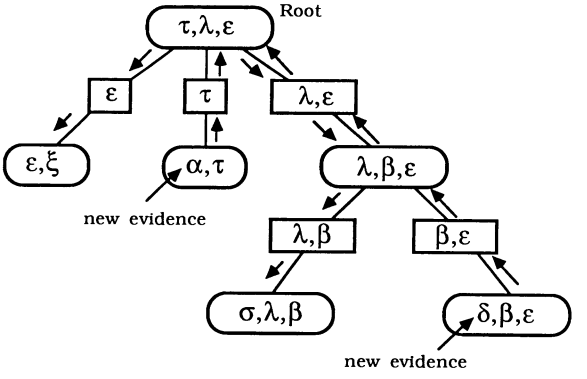


Fig. 12. Global propagation in a junction tree

In principle this way of updating works in the same way as described in the paper, but it reduces the computation involved and makes new maximum cardinality enumeration superfluous.

These methods can also be used in the initial transformation, i.e. the movement from a causal network described by nodes and their conditional probability tables to a junction tree consisting of cliques and their corresponding set marginals.

When the qualitative structure of the junction tree is settled a table is allocated to each clique and each separation set. These tables could hold either an evidence potential, a set marginal or a set chain representation. Initially the tables associated with the cliques hold an evidence potential composed of the conditional probabilities.

With the global propagation scheme on this set of tables and starting the propagation from all leaves, the joining of evidence in the root will leave the tables holding a set chain representation. The following calibration will then leave the tables holding a consistent set marginal representation and a run time environment is created.

Dr A. Gammerman (Heriot-Watt University): The paper is important not only because it gives us a neat and accurate description of a causal probabilistic reasoning method but also because it introduces us to some new ideas in the field of intelligent knowledge-based systems.

My contribution to the discussion will be limited to a comment on the development of a computational model of a causal network.

A working computational model of the causal probabilistic reasoning method suggested by Spiegelhalter (1986, 1987) has been designed and implemented at the Computer Science Department of Heriot-Watt University (Gammerman and Crabbe, 1987).

The model, a causal probabilistic reasoning system, at present consists of two main procedures. The first is a knowledge elicitation procedure through which an expert or a user may supply information about some area of knowledge in the form of a causal graph and associated conditional probabilities and store the information in a file.

The second part of the model is an evidence propagation procedure which deals with an individual case relating to an area of knowledge dealt with by the first procedure.

All programs in both procedures were written in the language C using UNIX operating environment on a VAX 11/750 computer.

The model is being tested and the question of how it may be used in forensic science is currently the subject of research.

The following contributions were received in writing, after the meeting.

Dr C. G. G. Aitken (University of Edinburgh): In the paper the authors have discussed a method for successively updating on the basis of available evidence a coherent system of probabilities representing belief in verifiable propositions. Another area, apart from the medical area, in which we are interested in updating beliefs in propositions is that of the interpretation of evidence in forensic science. The following artificial example illustrates the point.

Two people, A and B, are under suspicion of murdering a third person, V. Fibres from a jacket, similar to one found in the possession of A, are found at the scene, S, of the crime. Glass fragments of a refractive index similar to that of a window broken at S are found in the soles of pairs of shoes belonging to A and B. There is eye-witness evidence that shortly before the commission of the crime both A and B have had violent disagreements with V. Hearsay evidence is produced that B drives A's car frequently. This introduces the possibility of secondary transfer of fibres—a matter of current concern to forensic scientists—from A's jacket to clothes of B and from those clothes of B to S. Suppose, also, that A is a glazier. This affects the weight that may be attached to the glass evidence in his case since the glass in the soles of his shoes may have reached there in the natural course of his work. There is also the possibility that A will leave glass in his car which B may pick up while driving it and then deposit at S.

It is not difficult to see how these statements may be represented graphically and the techniques of today's paper applied. However, there are many problems of interpretation. Conditional probabilities need to be obtained and many of these will be subjective. The reliability of eye-witness and hearsay evidence needs to be considered. However, there are also exciting possibilities and extensions are not difficult to imagine.

Work is currently being done in collaboration with, and using causal computational models developed by, Dr A. J. Gammerman of Heriot-Watt University to investigate the performance of these techniques in the assessment of evidence in forensic science. We would be interested to know whether the authors have had any experience of such an application. The area seems most appropriate for ideas on imprecision and updating as outlined in Section 11.4.

Dr Jens Damgaard Andersen (University of Copenhagen): I recognise the importance of establishing a solid probabilistic foundation for expert systems and welcome the thorough analysis and the innovative approach taken in the paper.

The expert knowledge is stated as conditional probability tables (see Table 1) which implicitly define the structure of the casual network. The node marginals, which are really of interest to assess the probabilities of hypotheses, can be calculated directly in one pass without the need to go through the two passes (forwards and backwards) described in Section 5.1.

The binary relation $v \rightarrow w$ (w is a child of v) is antisymmetric and transitive. If we include a reflexivity property the set $(V, \rightarrow)$ is a partially ordered set (poset). Thus Fig. 2 may be redrawn as the Hasse diagram Fig. 13.

The unconditionally specified states correspond to the minimal elements of the poset (α and σ in Table 1). Given the structure of the poset it is possible to implement calculation of node marginals in one pass based directly on the table representations after the nodes have been sorted in antichains. Sorting in disjoint antichains requires fewer than mnp comparisons between node numbers, where m is the number of nodes, n the number of disjoint antichains and p the number of elements in the largest

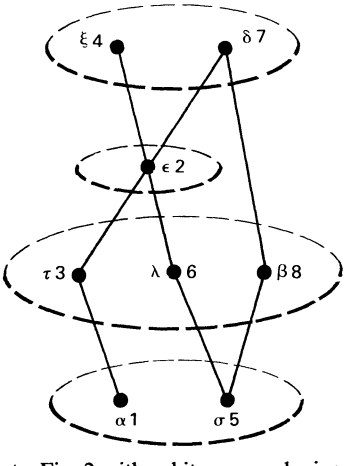


Fig. 13. Hasse diagram corresponding to Fig. 2 with arbitrary numbering of nodes and a partitioning into four disjoint antichains $\{\alpha, \sigma\}$, $\{\tau, \lambda, \beta\}$, $\{\varepsilon\}$, $\{\xi, \delta\}$

antichain. The poset $(V, \rightarrow)$ can be partitioned into n disjoint antichains if the length of the longest chain in V is n (the dual of Dillworth's theorem, see Liu (1987)). It is easy to implement a program for sorting in antichains. After the sorting procedure node marginals can be calculated directly from the covered node marginals and conditional probabilities for the states of the covering node given the states of the covered nodes.

Professor R. E. Barlow (University of California at Berkeley): Lauritzen and Spiegelhalter present a *general* method for computing conditional probabilities based on probabilistic influence diagrams. Their influence diagrams are probabilistic because they do not consider decision nodes.

Fault trees (or, more precisely, logic networks, since they are not trees in the graph theory sense) used in engineering reliability analysis are special cases of probabilistic influence diagrams. In a fault tree all probability nodes are root nodes and binary. All other nodes are deterministic logic nodes. It was pointed out by Rosenthal (1975) that for general fault trees the computation time for computing the probability of the top event will be exponential in the size of the fault tree. Even determining whether the fault tree top event can occur is non-polynomial (NP) difficult. In a fault tree a cut set is a minimal set of events which will cause the occurrence of the top event. Even finding the size of the smallest cut set is NP difficult. The point is that there can be no *general* efficient algorithm for even many easy sounding problems.

In discussing a general method, most authors including Lauritzen and Spiegelhalter consider some method for first finding modules (called cliques by the authors). This was also considered by Chatterjee (1975) for fault trees. However, this does not alter the basic fact that the problem is NP hard. Another general method is that of pivoting. This method has been used recently by Wood and McCullers (1988) who have also developed a computer program for fault trees.

The fact that no general algorithm can be devised which is efficient does not mean that efficient algorithms should not be sought for special structures. This has been the computational approach taken with respect to connectivity in network reliability problems and it has been quite successful. Therefore, I suggest that this is the approach that should be taken with respect to probabilistic influence diagrams, i.e. we should seek and define classes of influence diagrams of special structure for which efficient algorithms can be devised. Certainly such special structures will be at least in part defined by the graph structure. In a sense that is what the authors are doing.

Professor C. Berzuini and Professor M. Stefanelli (University of Pavia): Probabilistic coherence is of primary importance in the therapeutic decision field, but is it always important in diagnostic applications? Many recent efforts in the field of medical diagnostic expert systems have concentrated on modelling reasoning strategies in terms of abstract cognitive tasks and represent them in the system separately from medical facts and relations, so that the expert system could reflect the cognitive nature of the problem. If the diagnostic problem is very complex, exclusively worrying about probabilistic coherence may lead to an underestimation of the importance of the cognitive tasks, and this may result in a 'shallow' expert system. Typically, diagnostic reasoning proceeds by 'abduction–deduction–induction' cycles. Abduction takes initial observed manifestations as input and sets an initial belief scenario by outputting a list of admitted hypotheses; subsequently, conditionally on admitted hypotheses, 'deduction' produces a list of expected manifestations, indicating relevant additional tests to be performed on the patient. Iteration of such a cycle ought, on suitable monotonicity assumptions, to converge to the 'best' diagnostic explanation.

In the light of the global architecture of diagnostic reasoning, do we always require input–output from individual abduction–deduction steps to be of probabilistic type, or is it sometimes computationally and/or psychologically advantageous to use standard logical rules providing a categorical output?

The solution to this question perhaps relies on the notion of utility: in an individual abduction or deduction step there may be little utility in embarking in coherent probabilistic assessment if conclusions are only temporary. Since diagnosis is not deliberated in a single abduction step, we can pragmatically accept that at each abduction step all hypotheses surviving exclusion criteria are equally true, because they can have this status revoked at a later step, in the light of new evidence on the patient, perhaps containing pathognomonic signs. At each abduction step probabilities are temporarily 'crystallised' to 0, 1 values. The main effort is not in mediating the relationship between evidence and hypotheses by numerical probabilities, but in reproducing the global abduction–deduction architecture of reasoning followed by diagnosticians.

Perhaps an outstanding improvement would follow from combining the logistic and the probabilistic

paradigms. For example, if the behaviour of the system in a certain subproblem is unsatisfactory, we could translate the relevant portion of knowledge-base into a 'moral' network representation and, provided that relevant data are available, use the network representation to optimise the extraction of probabilistic parameters from the data and/or test hypotheses about the qualitative structure of the system.

We agree with the authors that the semantics of the network would be significantly enriched by symmetric links. For example, concepts of mutual exclusion or complementarity between diagnoses imply symmetric relationships, which are useful in reducing the space of probabilistic parameters since they imply zero conditional probabilities for certain disease combinations.

Dr Peter Cheeseman (RIACS, Moffett Field): I wish to call attention to the question: 'where are all the numbers coming from?'. The obvious answer is: 'from the expert', and the authors refer to Andreassen *et al.* (1987) for details of the assessment procedure. This raises the question: 'where did the expert get the numbers?' Again, the obvious answer is: 'from experience'. However, this answer raises a more fundamental question: 'does the expert have sufficient experience to justify all those numbers, or is he making them up?'.

For a rough quantitative grasp of this problem, consider the fragment of MUNIN shown in Fig. 1. As the authors show, the 'FORCE' node alone requires 270 values to be specified. Not only is this a lot of information to ask of the expert, it is not clear that the expert has this information to give. From my research on automatic induction of expert systems from data it is clear that there must be at least an order of magnitude more data (cases) than the number of parameters to be assessed—otherwise there are not sufficient data to distinguish real effects from noise. In this example, this implies about 3000 cases for the FORCE node alone. In some domains the ratio required is much higher. This is just a restatement of the well-known result in pattern recognition that a given amount of data will only justify induction of a limited number of parameters.

These rough estimates are assuming that the induction is done optimally (e.g. a Bayesian estimate with reasonable priors), but there is considerable psychological evidence that people (including experts) are not very good at estimating values. This poor estimation is especially clear when the signal is mixed with much irrelevant information, stretched over a long time period and in a numerical form, as in this case. Typically what happens is that the expert is unduly influenced by recent cases and ignores all but a handful of the strongest effects.

Since I do not know the experience of the expert used, or the strength of the effects captured in the causal tree, I can only speculate whether the expert went beyond his experience; however, the numbers suggest that he did. The only escape from this conclusion is the following.

(a) The expert used much more than his own knowledge: this is possible because text-books, papers, journals, teachers, collegues etc. potentially represent a much larger pool of knowledge than the expert's own experience. If this is a new domain, this loophole does not apply.
(b) Prior knowledge: in addition to observation, the expert knows a great deal about anatomy, physiology, pathology etc. This prior knowledge can contribute to the assessment of numbers beyond direct observation.
(c) It is not important anyway: the only point of a system like MUNIN is to aid clinical decision making. If the decision is not very sensitive to the numbers, then accurate assessment of the numbers is unnecessary—it depends on the problem.

Finally, I would plead that researchers drop the term 'causal' from graphical representations unless they mean it. The authors were careful to point out that they give the term a broad interpretation that includes 'logical, physical, temporal' etc., but in the rest of the paper the directed arcs are interpreted as conditional probability statements only. The usual meaning of the term causal can be given a directed graph representation that may or may not correspond to a given set of conditional probabilities. It only confuses the issue to use the same representation to express conditional probability information and causal influence.

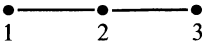
Gregory F. Cooper (Stanford University): The authors' primary goal is the development of an efficient algorithm for probabilistic inference using causal networks. Thus, I shall direct my remarks to computational efficiency issues. Without loss of generality, I use the term *probabilistic inference* to mean the global propagation of a single piece of evidence over a set of binary variables. For a causal network of size m (equal to the sum of the sizes of n nodes, d directed links and p prior and conditional probabilities),

the computational time complexity of probabilistic inference using the authors' algorithm has an upper bound of $3K + g\Theta$ elementary arithmetic operations, where K is the total size of the state space, g is the number of cliques and Θ is the size of the largest state space of a clique. In the worst case, when a clique has a state space of size 2^n the time complexity of probabilistic inference using their algorithm is also an exponential function of n. Consider an extreme case in which a causal network has a link between every node in the network. This network has a size m that is an exponential function of n. In this case, the authors' algorithm will produce only one clique of size 2^n, and the algorithm will have a time complexity of probabilistic inference that is an exponential function of n. This example only demonstrates that large causal networks require large amounts of computation. Of more interest is whether there are relatively small causal networks that require large amounts of computation. More specifically are there causal networks of size m for which the time complexity of probabilistic inference is an exponential function of m? This might be the case when the authors apply their algorithm to causal networks that have a lattice structure, as they suggest.

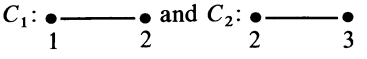
Cooper (1987) shows that probabilistic inference using causal networks is NP hard. This proof strongly suggests that, for *all* probabilistic inference algorithms, there are causal networks of size m for which the time complexity of probabilistic inference is an exponential function of m. The proof rests on showing that probabilistic inference using causal networks is at least as hard to compute as NP-complete problems, which have all eluded efficient algorithmic solutions. Knowing that a problem is NP hard is important because it suggests that any attempt at an exact, efficient solution is unlikely. Thus, expending a great deal of effort to develop such an algorithm should be given low priority.

The NP-hard proof for causal networks is a theoretical result that makes a statement about the infeasibility of efficient probabilistic inference over all possible causal networks. It does not address the extent to which an algorithm can perform probabilistic inference in a feasible amount of time on causal networks of *real* applications. The authors suggest that their approach might be practical for many applications because causal networks are often sparse and irregular; MUNIN is given as an example. However, MUNIN captures only a small portion of medical knowledge. Currently available probabilistic inference algorithms may not be sufficiently efficient for causal networks that represent a large amount of medical knowledge. Therefore, it will be important to characterise the types of causal networks encountered in large, complex real domains. If no current algorithm is adequately efficient when applied to large real networks, then perhaps some combination of current algorithms will be adequate; it also may be necessary to develop new exact or approximate algorithms. There is hope that we can eventually design a set of pragmatic algorithms for performing probabilistic inference on most real causal networks of interest.

A. P. Dawid (University College London): I have been collaborating with the authors on the problem of Bayesian learning about unknown probabilities in the kind of structure which they have described, and we have obtained some elegant results on the characterisation of 'conjugate' prior distributions. As a very simple illustration, consider three binary variables (X_1, X_2, X_3), with an unknown joint probability structure constrained to have X_1 independent of X_3, conditional on X_2. The (undirected) graphical representation is thus



The cliques of this graph are



and consequently the full structure is determined by the clique marginal probabilities $\theta_{ij} = \Pr(X_1 = i, X_2 = j)$ and $\phi_{jk} = \Pr(X_2 = j, X_3 = k)$ $(i, j, k = 0, 1)$, which are subject to the obvious consistency constraint $\Sigma_i \theta_{ij} = \Sigma_k \phi_{jk} = \psi_j$ say $(j = 0, 1)$. We can thus describe a prior distribution for the unknown overall probability structure in terms of a joint distribution for $(\boldsymbol{\theta}, \boldsymbol{\phi})$ over this constrained parameter space. For any specification of marginal prior distributions Π_1 for $\boldsymbol{\theta}$ and Π_2 for $\boldsymbol{\phi}$, satisfying the necessary compatibility condition that both induce the same distribution for $\boldsymbol{\psi}$, there will be a unique joint distribution for $(\boldsymbol{\theta}, \boldsymbol{\phi})$, over the constrained space, such that $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ have respective distributions Π_1 and Π_2 and, moreover, $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ are conditionally independent given $\boldsymbol{\psi}$. This conditional independence property will be preserved in the posterior distribution given a random sample of

observations on (X_1, X_2, X_3). If Π_1 and Π_2 are, furthermore, Dirichlet distributions, then the posterior is obtained by simply updating Π_1 (Π_2) using only the data on (X_1, X_2) $((X_2, X_3))$.

All these results generalise to an unknown Markov probability structure over a given triangulated undirected graph. There is a unique prior distribution having any assigned set of (compatible) distributions for the various clique marginals, and satisfying the additional requirement that, whenever $\{A, B\}$ is a *decomposition* of the full graph G, the marginal probability structures over the subgraphs A and B are conditionally independent, given that over their intersection C. (Here $\{A, B\}$ is a decomposition of G means that, as subsets, $A \cup B = G$, while $C = A \cap B$ is a *complete* subgraph of G which separates A from B in G: Lauritzen and Wermuth (1984).) Again, this conditional independence property is preserved under sampling, and, when the clique marginals have Dirichlet priors, each of these can be updated using only the relevant marginal data table. Such a prior distribution can be interpreted as an 'equivalent prior sample', to be combined additively with the data to yield an equivalent posterior sample, and, for this purpose, it is only necessary to store and update the relevant marginal tables for cliques.

Arthur P. Dempster and Russell G. Almond (Harvard University): Lauritzen and Spiegelhalter have presented an admirable exposition of a striking and important new branch of statistical technology to which they have made basic contributions. We have been independently developing a theory of belief functions on networks that closely parallels, and generalises, the Bayesian theory of Lauritzen and Spiegelhalter (Kong, 1986a, b; Dempster and Kong, 1986). Sometimes a more general mathematical viewpoint leads to simpler ways to understand and express a theory. We believe this to be so in this case.

For us, the basic representation is the natural belief function generalisation of *evidence potentials* (Section 7.2). A belief function is assigned to the states of each $A \in \Delta$, and these are combined by the product intersection rule which essentially multiplies probabilities, thus embodying independence or conditional independence assumptions, and intersects subsets, thus embodying logical conjunctions. When all component belief functions are simple distributions we recover Bayesian theory, but a belief function may also be a purely logical relation (including an observation), or a more general uncertainty representation as described in Shafer (1976). The rules for local computation by propagation and fusion described in Dempster and Kong (1986) are expressible in a simple and unified way via the language of belief functions.

A concept that we regard as central is left implicit by Lauritzen and Spiegelhalter, namely the idea of a *tree of cliques*. In the example in Section 4, after marrying and filling in, the tree of cliques is as shown in Fig. 14.

Edges indicate common nodes of the original network, and common nodes are always joined, at least indirectly via a sequence of cliques containing the node. The tree of cliques captures the fundamental Markov field structure of the model.

Directed graph representations of the original network are not especially fundamental to us, since they represent more-or-less arbitrary 'set chain' representations for constructing 'potentials'. However, a different type of arrow on the edges of the tree of cliques can be helpful for explaining the basic propagation and fusion algorithms that permit passage from potentials to marginals on the tree nodes (as are typically the basic outputs required in practice). Specifically, one can propagate from 'potentials

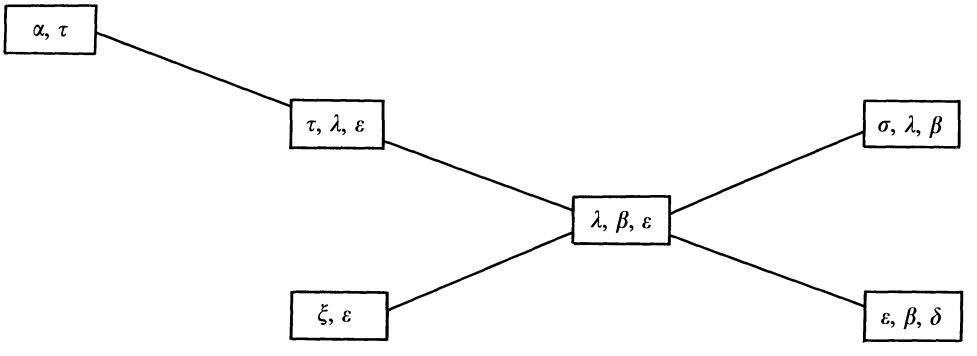


Fig. 14. Tree of cliques for dyspnoea diagnosis

to set chains' (Section 9.2) to flow through the tree in one direction ending at a preselected node, then from 'set chains to marginals' (Section 9.3) reversing all arrows to compute all margins simultaneously. Space limitations preclude numerical illustrations with flexible belief function potentials which are capable of representing a wide range of sensible uncertain knowledge.

Dr Didier Dubois and Dr Henri Prade (Université Paul Sabatier): Since symptoms and diseases are not exchangeable, links in causal networks attached to diagnosis problems are directed, whereas logic formulae are not. The meaning of so-called 'rules' of the form 'if A then B' is ambiguous, and part of the controversy between probabilists and logicians in artificial intelligence is related to this. To a logician, 'if A then B' means that 'either not A is true or B is true' and is expressed in a non-directed disjunctive form ($\neg A \vee B$). Interpreting 'if A then B' as 'A causes B' may lead to consider the conditional probability $P(B \mid A)$ (rather than $P(\neg A \vee B)$) as given by the expert, and to assume that the network obtained is acyclic. However, the statement 'if A then B' sometimes does not refer to causality, and $P(B \mid A)$ only reflects the amount of As that are Bs, as in 'most students are young'. In this case, the expert may wish to express knowledge about both $P(B \mid A)$ *and* $P(A \mid B)$ (adding that 'about 20% of young people are students') without being inconsistent. This type of situation introduces cycles in directed graphical representations, and the Bayesian methodology and the techniques described in the paper can no longer be applied. Especially if the *a priori* probability $P(A)$ is to be known, there is no longer any degree of freedom in the system since $P(B \mid A)$, $P(A \mid B)$ and $P(A)$ determine $P(B)$. A similar problem would occur if $P(A \mid B)$, $P(B \mid C)$ and $P(C \mid A)$ are the available data (without any causal interpretation). How would a joint probability distribution $P(A, B, C)$ be defined on such a basis? Only the causal interpretation of the network built from the expert knowledge can forbid this situation as being self-contradictory.

Only approaches based on upper and lower probabilities (Quinlan's (1983) INFERNO or Paass (1986)) can deal with cycles in knowledge networks, because these approaches assume that the degrees of probability provided by the experts define constraints on an unknown probability distribution, while Bayesian techniques always assume that the available knowledge is sufficient to define a joint probability distribution *uniquely*, as do Lauritzen and Spiegelhalter. To comply with this, additional assumptions such as conditional independence are extensively used. The justification, proposed by Bayesians, is strange. It seems as if they interpret it on a causal network. For instance knowing that C causes A ($P(A \mid C)$) and B ($P(B \mid C)$) leads us to consider $P(A \cap B \mid C) = P(A \mid C)\, P(B \mid C)$ as being natural, and that a joint distribution $P(A \mid C)\, P(B \mid C)\, P(C)$ follows. However, strictly $P(A \cap B \mid C)$ can be any number between $\max(0, P(A \mid C) + P(B \mid C) - 1)$ and $\min(P(A \mid C), P(B \mid C))$ and it is incautious to assume conditional independence because the expert has not yet said anything about the links between A and B. The authors consider that underspecification is a common flaw in current artificial intelligence techniques for uncertainty handling. In contrast we believe that a Bayesian representation is often implicitly overspecified. The requirement for a unique joint probability distribution leads to replacing missing information by strong *default* assumptions which ensure this unicity. Hence while it is true that Bayesian techniques are well fitted to absorb new evidence pertaining to the situation under study, they do not comply in a consistent way with the arrival of new links in the network, except by modifying the joint probability distribution in a non-monotonic way. In contrast, techniques based on upper and lower probabilities accommodate new knowledge by monotonically shrinking the probability intervals of derived conclusions. Bayesian techniques bear some similarities to default logics (e.g. Reiter (1980)) which are non-monotonic: *a priori* probabilities play the role of default values, and conditional independence assumptions stand as default rules. While upper and lower probability techniques provide safe responses that may be too imprecise to be useful, Bayesian techniques always provide *default* conclusions because adding new links to a causal network can completely change these conclusions.

Stephen E. Fienberg and Michael M. Meyer (Carnegie Mellon University): The paper combines and extends material in two different domains in which we have interest: the use of graphical representations for discrete probability structures and the application of (subjective) probability to expert system problems.

In the original work on graphical representations of Darroch *et al.* (1980), there were parallel structures for the class of 'graphical' log-linear models and the Markov random fields defined by graphs. In this parallel structure, the class of decomposable log-linear models (first described by Bishop (1971) and by Goodman (1970), and later extended by Haberman (1974)) played a very special role, and the graphical representation gave new insights into a variety of contingency table problems. In the present paper, these parallels lie primarily in the background, except for some indirect references in Sections 7–9.

Indeed, the principal role of the graphical structure is to develop computationally feasible approaches to expert systems problems. We believe that the ideas developed in this paper might have some important uses if they were translated back into the notation and language of log-linear models and contingency tables. For example, many asymptotic calculations involving Taylor series expansions for contingency table problems can be implemented only for decomposable log-linear moels (e.g. Lee (1977) and Bedrick (1983)). Perhaps ideas from this paper might be of use in deriving bounds for the related problems involving non-decomposable models. If the parallel is pursued in the other direction, then the structure of decomposable and graphical log-linear models may suggest a new approach to the expert system problem as well.

In a very different class of categorical data problems, Holland and Leinhardt (1981) developed log-linear representations for networks or directed graphs. We hope that the authors will comment on possible linkages of the ideas in the present paper with those of Holland and Leinhardt. Moreover, Fienberg *et al.* (1985) offer extensions of the log-linear representations to multivariate directed graphs. Have the authors explored analogous extensions?

I. J. Good (Virginia Polytechnic Institute and State University): Good (1984), among other things, drew attention to

(a) the idea of ignoring high order 'unexpectated' interactions between weights of evidence, as defined by Good (1960), and more needs to be done: this is a topic to which I returned in Good (1986),
(b) the weight of evidence provided by uncertain evidence (Good, 1981) and
(c) the probabilistic interpretation of a doctor's five-star representation of 'degrees of certainty'.

Causal networks can be used, as in the paper by Lauritzen and Spiegelhalter, for the estimation of the probabilities of disease states, or for more general hypotheses, and also for trying to assign a quantitative meaning to the degree to which one event tends to cause another. In the former activity it is sometimes convenient to consider hypotheses in pairs, as in a differential diagnosis, and to think in terms of the weight of evidence in favour of a hypothesis. But for the latter activity (exemplified by 'path analysis') the explanation that I support (e.g. Good (1961, 1988)) for the tendency of a disease F to cause a symptom E is the weight of evidence against F if E does not occur, given the state of the world just before F occurs.

During the last ten years or so there has been a revival of interest in the use of 'artificial neural networks' in research on artificial intelligence, the new name for this approach being 'connectionism'. It may be that real neural networks, even infra-human ones, embody unconscious iterative calculations both when we recognise an object and when we meditate on a problem, and when a non-mathematical doctor does a diagnosis. Some of my writing on such topics are more speculative than others, see my publication numbers 169, 183, 185, 243, 368, 397, 521, 525, 592, 615, 666, 753, 777, 938, 1212 and 1235. The meanings of these numbers may be found in Good (1983).

Dr Tomáš Havránek (Czechoslovak Academy of Science): Some questions arise if we leave the authors' probabilistic framework for some reason. I cannot agree with the view that probabilistic structuring provides a good model for human understanding and memory and hence is appropriate in expert system construction. The main support for this approach comes from statistics by considering cases in which we are able to base our knowledge on some statistical observations in addition to a fragment of a structural knowledge of an expert, as in the MUNIN example.

Human experts are rarely able to express their knowledge (uncertainty) in a numerical form respecting the nature of real numbers of probabilities fully. Relying on numbers can lead to great misunderstanding in constructing expert systems even if there is clearly a great temptation to use a numerical representation for its computational advantages. In this connection there is the algebraic theory of uncertainty. Hájek and Valdés (1987) apply finitely generated ordered Abelian groups as closely as possible to human expert understanding of uncertainties. This approach respects fully the fact that experts and users of a consultation system use only a small finite number of values for expressing the uncertainty. In this framework, for example, questions of stability of decisions of an expert system under changes of the designated values can be investigated. Such an approach restricts the expressive power of a knowledge representation, but as can be seen from the paper in the probabilistic approach the expressive power has to be restricted as well—from the computability point of view. The question is whether this restriction (triangulated graphs etc.) respects the nature of knowledge expressed by human experts. The answer is

yes for reliably established structural knowledge. We can consider incorporating logical links into probabilistic knowledge questionable, particularly in some 'second-order' unreliable cases.

Even if we only consider statistically generated knowledge, perhaps with some *a priori* structural expert knowledge, then some open questions remain: first, a non-Bayesian method using confidence region estimates of probabilities should be more realistic in many cases and secondly there is a question of establishing structural models (dependency graphs) from statistical data; data can support a great number of concurrent structural models (Edwards and Havránek, 1985, 1987) and some further research in model choice is indispensable.

In general, linking expert knowledge with knowledge obtained from data can be very dangerous.

Max Henrion (Carnegie Mellon University): The development of coherent and tractable inference methods for large causal or belief networks is an important goal. Its achievement is crucial for building diagnostic expert systems that are both practical and trustworthy. Lauritzen and Spiegelhalter present an elegant formulation and ingenious techniques, which constitute a major step towards this goal. The question is how much further do we have to go?

The authors point out that the computational complexity of their algorithm is exponential in γ, the size of the largest clique in the network. How large this is will depend on the degree of connectedness of the graph, and they cite Pearl (1986a), who argues that nets are often sparse. They illustrate this with MUNIN, whose largest clique has only four nodes. It can achieve this by treating the possible diseases as a single node with 11 mutually exclusive states. This may be entirely appropriate for MUNIN's domain, but in other domains more than one disease may be present simultaneously. This means that each disease must be treated as a separate node, and considerably increases the complexity of the net. For example, the medical expert system QMR, a descendant of INTERNIST-1 (Miller *et al.*, 1982), has almost 600 diseases and 4000 manifestations (symptoms, findings etc.). Some manifestations have as many as 150 possible causes. To 'moralise' the causal graph in such a case would require marriages among all the 150 parent nodes (ignoring the issue of the morality of group marriage on this scale!). Together these would form a single clique with their common child, producing a factor of 2^{151} and a computational impasse.

Such difficulties should not be too surprising: Cooper (1987) has shown that, in the general case, exact inference in a belief net is NP hard. This suggests that we will have to resort to approximate methods if our goal is to develop coherent methods for expert systems of the scale of QMR. While I believe that this paper will stand as an important landmark along the way, we still have some distance to go.

Dr F. V. Jensen (Judex Datasystemer, Aalborg): I see two major achievements in this work: initialisation and global propagation of evidence. I shall only comment on global propagation. The points made are reported in Jensen (1988).

Let U be the universe, and let G be a covering of U consisting of pairwise incomparable sets, each having a probability table. G is *consistent* if for any c, $d \in G$, the marginals for $c \cap d$ in the two tables coincide. As stated by Lauritzen and Spiegelhalter, it holds that if G is a decomposable hypergraph then there exists a probability function for U with the tables from G as marginals. This function is unique if maximal entropy is required. Now, they use that G is decomposable if and only if G can be ordered with a running intersection property (RIP). However, more flexible and efficient methods can be achieved using a tree ordering of G rather than the linear RIP.

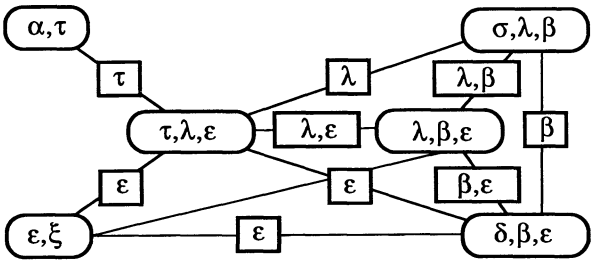


Fig. 15. Junction graph for the example in the report

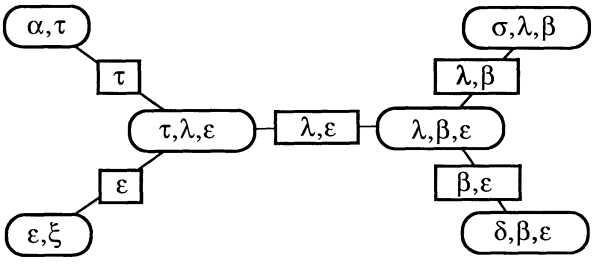


Fig. 16. Junction tree for the example

Definition. The *junction graph*, $J(G)$ for G has the elements from G as vertices. For each pair $c, d \in G$ with $c \cap d \neq \varnothing$ there is an edge with weight $c \cap d$ (Fig. 15).

A junction tree, T, for G is a spanning tree for $J(G)$ such that for any $c, d \in G$, all vertices on the path in T between c and d contain $c \cap d$ (Fig. 16).

Theorem. G is decomposable if and only if G has a junction tree.

Using junction trees, the existence of a probability function for U is fairly easy to prove. Furthermore, global propagation of evidence is straightforward: any vertex in the junction tree can be used as a root for a propagation. A change in the table is propagated by simply successively calibrating the neighbours in the tree to the recently calibrated vertices.

Since any maximal spanning tree is a junction tree (if they exist), effective algorithms for constructing junction trees exist, such that they can be established at run time if needed.

All methods based on junction trees are *local*: communication with neighbours in the junction tree only is sufficient. Therefore an object-orientated style of implementation is supported.

Finally I emphasise that, although junction trees might be a conceptual simplification and might form a basis for more flexible and efficient methods, the bulk of the theoretical work is contained in this paper.

Augustine Kong (University of Chicago): My comments will focus on some technical issues concerning marginalisation of potentials and fill-in algorithms.

In Section 8.3, Lauritzen and Spiegelhalter consider how to find a potential representation $(\bar{\Delta}, \bar{\Psi})$ of the marginal of the nodes D given the potential representation (Δ, Ψ) of the joint distribution over the nodes $D \cup E$ of a network. The suggested approach is natural if $|E| = 1$. However, when $|E| > 1$, it would often be important to proceed recursively through subsets of E. For example, suppose $E = \{v_1, v_2\}$. A recursive procedure could first marginalise over v_1, using the algorithm described in the paper, to derive (Δ', Ψ'), and then apply the same algorithm to marginalise over v_2 and obtain (Δ'', Ψ''). Compared with the one-step procedure, the recursive procedure in general leads to potentials that involve smaller sets of nodes. For example, if $D = \{2, 3, 4\}$, $E = \{1, 5\}$ and $\Delta = \{\{1, 2\}, \{2, 3\}, \{3, 4\}, \{4, 5\}\}$, then using the one-step procedure we obtain $\bar{\Delta} = \{\{2, 4\}, \{2, 3\}, \{3, 4\}\}$. Using the recursive approach, $\Delta' = \{\{2\}, \{2, 3\}, \{3, 4\}, \{4, 5\}\}$ and $\Delta'' = \{\{2\}, \{2, 3\}, \{3, 4\}, \{4\}\}$. The representation (Δ'', Ψ'') is more informative since Δ'' indicates that 2 and 4 are conditionally independent given 3. This recursive procedure is proposed in Kong (1986) as a fast algorithm for computing any joint marginal belief function that may be of interest.

In Section 6, the authors define minimum fill-in and note that computing a minimum fill-in is NP complete. This may give the impression that a minimum fill-in is preferred, and the only reason that it is not used is that it is expensive to compute. By contrast, Kong (1986) proposes searching for a fill-in that minimises the maximal clique state size. This proposal is supported by the comment of Lauritzen and Spiegelhalter that the maximal clique state size is vital to computational cost. Unfortunately, even in the simple case where the state sizes of individual nodes are all the same, Arnborg *et al.* (1987) show that finding such a fill-in is also NP complete. Hence, for applications, we do need fast fill-in algorithms such as the maximal cardinality search and the lexicographic search suggested by the authors, or a one-step look-ahead algorithm suggested in Kong (1986). We plan to report elsewhere on empirical tests of fast fill-in algorithms on different graphs with evaluations based on computation time and the maximal clique state size produced.

Professor K. V. Mardia (University of Leeds): I believe that the discussion in Sections 4 and 5 could be a simple result of obtaining the minimum Markov random field (MRF) through Brook's expansion. Have the authors tried this approach? In the case of a slightly more difficult situation, the approach works well (Mardia, 1988). For Section 11.4, we could use a prior for probabilities themselves using a Gaussian MRF either on a hyperplane or of the appropriate log-ratios (Kent and Mardia, 1988). These priors have some nice properties.

In the paper, it would have been nice to have seen how Geman's (1985) approach works for the given example. Also, a few more examples would have been helpful.

Dr Mary McLeish (University of Guelph): The paper sheds further light on the difficult and controversial problem of how to model uncertainty in expert systems. The authors provide a computationally feasible solution to a probabilistic approach which could also be applied to domains beyond the medical applications.

The computational complexity issue has benefited by the existence of an O(nodes + edges) solution to the required graph problem. The other quantities which enter into the time factor are γ, the maximum number of nodes in a clique, K, the total state space, θ, the largest state of a clique and g, the number of cliques. In the sample problems given here, these quantities are all fairly small and thus an expression like $(2^{\gamma} + 1)K + g\theta$ is not very large. At Guelph, we are working on computerised diagnostic aids for veterinary medicine involving discrete variables with usually five or six outcomes and sometimes as many as 50 (especially when continuous variables have been discretised). One project involves the diagnosis of liver disease in small animals and 100 test results are relevant to the diagnostic process. The maximum clique size could easily become large and the base of the exponential term would be greater than two. Thus, for certain applications, the exponential term in the complexity expressions could become significant.

A comment in the paper suggests how the use of object-oriented programming can reduce the time complexity. Another approach would be to investigate different machine architectures. New machines (SEQUENT, HYPERCUBE, etc.) offer parallelism with powerful central processor units (CPUs) at each node. The connection machine provides a massively parallel environment with minimal node capacity. There has already been considerable work on developing parallel graph algorithms (e.g. Quinn and Deo (1984)) and there is recent work related to medical diagnosis by Pearl (1987b). A recent implementation on a GAPP (parallel systolic array processor) assigns each rule in an inference net to a processor in the array, which performs a 'fuzzy' inference operation (Eshera and Lewis, 1987). Depending on the type of parallel machine, each node or clique of the graph in a probabilistic inference net could be assigned to a separate processor to do local computations. In larger applications, where the diagnosis of many related problems is being undertaken, each disease type could be handled by individual CPUs. Neural net architectures would also be worth investigating for Bayesian networks.

Judea Pearl (University of California, Los Angeles): The paper brings together two themes: the role of structural models in statistical analysis and the role of probabilistic analysis in expert reasoning. These carry a doubly important message to on-going work in artificial intelligence (AI).

Ever since McCarthy and Hayes (1969) proclaimed probabilities to be 'epistemologically inadequate', AI researchers have shunned probabilities. A few researchers have been trying to convince AI researchers that abandoning probability theory altogether might be premature (Spiegelhalter, 1986; Pearl, 1988). We have tried to communicate the understanding that 'Probability is not really about numbers; it is about the structure of reasoning' (a quote from G. Shafer). For example, the statement $P(B \mid A) = p$ is not so concerned with the precise magnitude of p as it is with specifying the permissible ways in which the conditioning context A can be transformed if p is to remain unchanged, and that the information needed for such transformations can be represented by graphs of no lesser stature than those networks of pointers and indices that decorate 'symbolic' programs in AI.

These arguments have remained tarnished by the realisation that the statistics community itself does not practice what we claim is the essence of statistical inference. For example, the statistical literature still treats graphical models as a mnemonic curiosity. This is attested by the observation that most text-books and journals in probability and statistics have hardly any diagrams in them or only depict shapes of density functions, not relationships between random entities.

This paper demonstrates to the AI community that local representations of probabilistic models are capable of meeting the computational demands of expert systems technology. Secondly, it establishes graphical models on sound theoretical foundations and, finally, it shows that such models are indispensable for inference tasks involving many variables.

The message to statistics is of no less value than that given to AI. Statisticians should rejoice that both directed and undirected graphs conform to the axiomatic structure that governs conditional independence (Dawid, 1979; Pearl and Verma, 1987), which renders graphical representation a useful tool in statistical analysis; graphs permit the statistician to verify swiftly whether one dependency follows from others, and they serve as communication to pass structural information economically and naturally.

I have two questions. The first relates to presentation. I have found it useful to explain the technique in the following manner:

(a) triangulate the graph;
(b) identify the maximal cliques of the triangulated graph;
(c) organise these cliques in a tree structure (i.e. a join-tree) and direct its links along the order used in the triangulation phase;
(d) treat each clique as a single compound variable and identify the conditional probabilities of the inter-clique links in terms of the conditional probabilities of the original network;
(e) update probabilities using the familiar method of propagation in causal trees (e.g. Pearl (1986a, 1988))

Is there a flaw in this or is there any advantage in keeping the join-tree undirected? One advantage in the directed tree approach is that the causal relationships used in the construction of the original network are kept explicit.

My second question raises the option of viewing the technique in the wider context of clustering methods. For example, consider Fig. 17(a): it can be organised into six overlapping clusters by the clique-decomposition method (Fig. 17(b)), but it can also be organised into a chain of five non-overlapping clusters (Fig. 17(c)). Moreover, the non-overlapping structure has the slight advantage that the largest cluster contains only two variables. Is it worth exploring the space of all singly connected clusterings, or can the authors' method now provide a systematic way of identifying the most useful structures in that vast space?

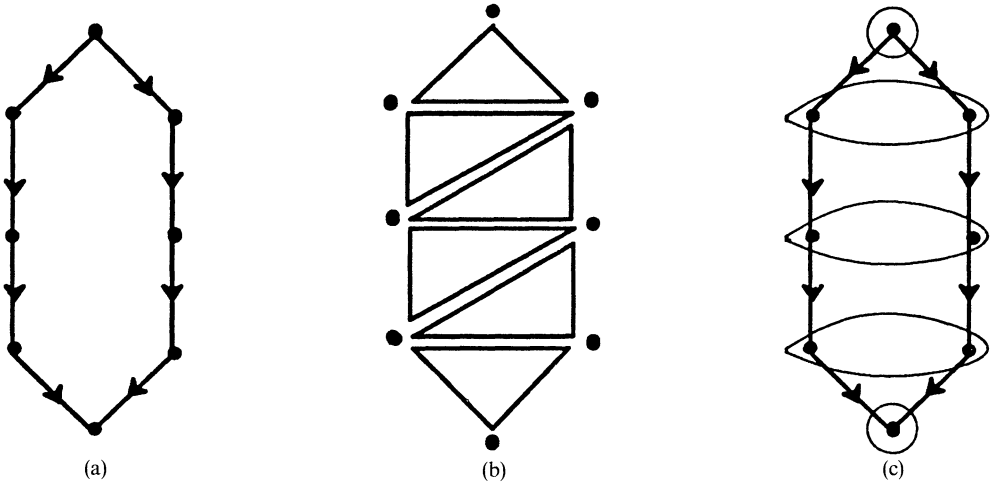


Fig. 17

Dr Lawrence D. Phillips (London School of Economics and Political Science): Any scheme that will simplify the potentially unmanageable computations required by a large inference structure is to be welcomed, and the contribution by Lauritzen and Spiegelhalter is particularly notable for its comprehensiveness and potential usefulness. A complementary approach has been taken in the pioneering work of psychologists Schum and Martin (1981) who have produced a 14-category taxonomy of possible patterns of data which may or may not interact in an inferential argument. Discussing the implications of this work, von Winterfeldt and Edwards (1986) speculate that

> '... these 14 categories of evidence structures can be thought of as the building blocks of a very general theory of evidence. In other words, we speculate that any inferential structure, no matter how complex, can be decomposed into various combinations of these 14 elements. Moreover, since these

> 14 patterns include all kinds of dependencies among items of evidence, each block should be conditionally independent of every other block. In that sense (if our speculation is correct), Schum and Martin have accomplished the extraordinary feat of providing a comprehensive taxonomy of inferential structures applicable to all possible instances of human (or machine) inference.'

(In his more recent work, Schum (1988) has reduced the taxonomy to 12 categories.) It appears that the computational schemes of Lauritzen and Speigelhalter are applicable to all these 12 evidence structures since each can be represented as a causal network. It is not clear to me whether the schemes can handle the subtle dependency issues raised by Schum and Martin (1981, 1982). If they can it would suggest that the next major development in causal networks might come about by marrying efficient and flexible computational schemes with generic structures so that combinations of these structures could usefully represent very complex inference structures in expert systems.

Dr B. T. Porteous (University of Durham): The idea of expressing a causal probability distribution as an equivalent distribution defined on a decomposable graph and then exploiting the special properties of decomposable graphs is, in my view, extremely ingenious.

Although the authors state that, for the purposes of their paper, they are concerned only with a fixed model, which is completely specified, this model would seem to be of crucial importance in the subsequent probability manipulations, and several questions seem worthy of future research. For example, what are the implications of misspecifying the initial causal network and how much effect do misspecifications of the conditional probability tables have on absorption and propagation?

Although the authors state that they interpret probabilities as subjective Bayesians, I feel that the frequentist approach, other than for initialisation perhaps, has many attractive features. Assuming that one is willing to work within the class of sparse decomposable graphs, model estimation, verification and updating or learning could all, in principle, be achieved locally using low dimensional margins of the data.

Finally, I would like to make some remarks on collapsibility. Collapsibility for the covariance selection models of Dempster (1972) has been studied in Porteous (1985). This and related studies, Lauritzen (1982), Asmussen and Edwards (1983), are useful not only for understanding the conditional independence structure of marginal models, but for constructing models capable of handling data consisting of both response and explanatory random variables.

Dr Ian Pratt (University of Manchester): Lauritzen and Spiegelhalter's method for performing probabilistic inference using causal networks assumes that effects of a single common cause in the network are probabilistically independent given that cause. I wish here to question the extent to which this assumption can be expected to hold.

The events and states referred to in the authors' examples admit of continuous variation in *degree*, *size*, *intensity* etc. Thus, headaches can occur with greater or less severity, smoking can be more or less heavy and prolonged, and growths can be larger or smaller. This variability generates problems for the assumption that joint effects will be independent given their causes.

Consider Fig. 2 of the paper. This tells us that smoking can cause both lung cancer and bronchitis. The question before us is: if we already know that the patient smokes (exact extent unspecified), does the *additional* information that he has lung cancer affect the likelihood that he has bronchitis? If so, we do not have the conditional independence claimed by the authors.

Let it be given that the patient smokes. This means that he has smoked a certain number of cigarettes for a certain time; but a wide variation in the rate and history of the patient's smoking is still possible. The crucial observation is that, if he has lung cancer, it is more likely that his smoking will be towards the heavy end of the spectrum. But if his smoking is heavy, it is more likely that he will have bronchitis than it would be if his smoking were light. Therefore, the observation of lung cancer increases the probability of bronchitis given only that the patient smokes (extent unspecified). If so, lung cancer and bronchitis are not independent given only that the patient smokes.

More generally, the problem is that most medical conditions are more-or-less affairs. You can smoke more or less heavily, have more or less exposure to radiation, be incubating more or fewer dread bacteria D, etc. Many of the propositions in the examples given in the paper concern conditions and circumstances which exhibit just such variability—variability which affects the probability with which various effects are brought about. This being so, the independence assumptions made by the authors will seldom hold in their chosen domain.

Dr A. L. Rector (University of Manchester): The models described require very large numbers of conditional probabilities. It is therefore important to determine their sensitivity to variations in the conditional probabilities and their behaviour under plausible approximations. Not all the conditional probabilities will be equally reliable or sensitive to local conditions. In general, probabilities of symptoms given diseases—those probabilities following the direction of the original directed graph—are reasonably well defined. However, probabilities of the symptoms given the absence of a disease depend heavily on the population from which the sample is drawn. For example, hospital clinic populations are often heavily preselected so that the probability of symptoms given the absence of disease is much lower than in the general population. Similarly, the probabilities of co-occurrence of common causes required for the 'moral graph' are likely to vary between populations. Some of the variation in the probability of co-occurrence is simply the effect of scaling of the overall probabilities—e.g. older people tend to have more diseases than do young people and it may be possible to compensate for these factors.

These models are interesting candidates for parallel implementation, since the propagation is essentially local. The discussion of 'object-orientated' programming in the text really applies to any parallel or pseudoparallel technique. The 'tuning' techniques discussed by Spiegelhalter (1986b) have obvious analogies with parallel distributed processing methods.

There are many cases where the causal graph is not naturally acyclic and some feedback is difficult to avoid. Whether a parallel processing approach might provide sufficient power to allow iterative solutions in these cases is a question worth pursuing.

The assumption of the causal Markov property corresponds to the closed world assumption in logic programming. The effect described, whereby evidence for one of two competing causes for a single manifestation decreases the probability of the other, is closely related to 'negation as failure' in logic programming, whereby any statement which cannot be proven is taken as false. It should therefore be emphasised that these techniques require a closed problem space. Many problems for medical expert systems are less well defined than those in the MUNIN system. Extension to systems in which possible causes for symptoms are added incrementally would be an important development.

Dr Ross D. Shachter (Stanford University): The algorithms in the paper can be viewed as a generalisation of algorithms for analysing singly connected directed graphs (Kim and Pearl, 1983; Pearl, 1986a). If we construct a 'supergraph' in which each clique is a node, then the running intersection property (enforced through triangulation) guarantees that we can build a singly connected directed graph to represent the dependence relationships among the cliques. When revising our distribution for a clique, we can use this supergraph to propagate the changes efficiently throughout the model while maintaining the original supergraph topology. This interpretation of the algorithm accounts for many of its desirable properties.

As an alternative approach, it *is* possible to perform all the required operations (with the possible exception of 'retraction' of evidence) within the directed graph representation (Howard and Matheson, 1981; Olmsted, 1983; Shachter; 1986a, b). Using influence diagram reductions, for example, global propagation can be performed through a combination of arc reversals and instantiation. In the process, the topology of the graph would be revised. This has the advantage of staying within the natural representation for the model builder, especially when the model is constructed and analysed dynamically. However, when the model can be precompiled we would expect the method presented in this paper to be more efficient, since it exploits the single connectedness of the supergraph.

The implicit reliance on causality in the network is unnecessary and potentially misleading. It is true that an expert's subjective model can incorporate causality and that this can be an invaluable paradigm in the elicitation of that expert's model. However, the construction of the original directed network merely expresses beliefs about conditional probability distributions and conditional independence; the directed arcs are not necessarily causal (Howard and Matheson, 1981; Shachter and Heckerman, 1987).

The use of an entropy measure such as Kullback–Leibler distance might be a useful criterion in some planning problems, but it does not indicate 'which questions will most provide relevant information' as the authors state. Such a measure might select a test which distinguishes between disorders for which the treatment decision would be the same. In general, the value of sample information, based on expected utility, is a more appropriate criterion for comparison. It also allows consideration of the differences in cost, pain and risk that might be associated with different test procedures.

The approach developed in this research will be useful in building expert systems under uncertainty, in concert with other methods for analysing and manipulating graphical probabilistic models. It seems most appropriate to domains (such as MUNIN) for which a detailed model can be constructed in advance.

In general, a combination of approaches is needed to provide a user with timely understandable support when a decision must be made.

Glenn Shafer and Prakash Shenoy (University of Kansas): The ideas on which Lauritzen and Spiegelhalter are working are of great importance not only for expert systems but also for our general understanding of probability.

A central lesson of the paper is that we can exploit for combining evidence the same conditional independence structure that we use in thinking about causation. This lesson is widely relevant. It applies not only to situations where sensible full probability distributions are available, but also to situations where the data are more fragmentary and we must make do with partial or qualitative judgements.

The conditional independence structure is more important than the probability numbers or the rules of calculation. We can replace the numbers by a verbal scale (likely, very likely, etc.), and we can replace Bayes's theorem by other rules (as we do in the theory of belief functions), but we cannot dispense with a structure that tells us what evidence bears on what propositions.

It is disturbing to note that a student must study probability for a long time before he or she will see this importance of structure. Markov trees and fields are topics for an advanced course in probability, and path analysis will usually appear only in a fourth or fifth course in statistical inference. Lauritzen and Spiegelhalter's paper suggests that we must change this if we want, as statisticians, to retain intellectual leadership in the area of probability. We must emphasise the importance of structure in our elementary teaching. One way of doing this is to recognise the need for structure to justify the rule of conditioning (Shafer, 1985).

Would the authors clarify further the relation between their work and the work of Pearl (1986a)? It is obvious from Fig. 2 that we can successively calculate the clique marginals for the cliques $\{\alpha, \tau\}$, $\{\lambda, \beta, \sigma\}$, $\{\varepsilon, \tau, \lambda\}$, $\{\varepsilon, \beta, \lambda\}$, $\{\delta, \varepsilon, \beta\}$ and $\{\varepsilon, \xi\}$ in one just pass through the graph. This natural ordering of the cliques can also be used for propagating new evidence efficiently, as Pearl has shown. The authors show us a more flexible approach, one which can use any triangulation and any ordering with the running intersection property. But what is gained by this flexibility?

Professor Philippe Smets (Université Libre de Bruxelles): Although the authors did not intend to compare the pros and cons of the various models they propose to handle uncertainty in a coherent manner, it is questionable whether their proposed Bayesian model is as meaningful as they claim. Once applicable, probabilities are perfect, but are they really applicable for medical diagnosis? For instance, can one defend the p value in $P(\text{MU.loss} = \text{other} \mid \text{disease} = \text{other}) = p$?

To simplify, suppose $B \equiv$ bronchitis and $D \equiv$ dyspnoea. I know what $P(D \mid B)$ is but what about $P(D \mid \bar{B})$? The difference is due to the 'well-defined' nature of the set B and the 'ill-defined' nature of the set $\bar{B}$. The causal relation between diseases and symptoms is usually clear and its translation into a conditional probability can be defended, but what about $P(D \mid \bar{B})$? Let the set of diseases be defined by the family of mutually exclusive and exhaustive diseases $B_1 B_2 \ldots B_n$ ($B = B_1$). For each B_i, I can define $P(D \mid B_i)$ but to define $P(D \mid \bar{B}_i)$ I need the *a priori* distribution of the B_is. If I have it, as in a well-defined population, then $P(D \mid \bar{B})$ is well defined. But medical diagnosis is not performed in a well-defined context.

In reality, the population is poorly defined. $P(D \mid B_i)$ can be used as essentially it does not depend on time and space, but the prior repartition $P(B_i)$ is not so constant. It varies in time and space. $\bar{B}$ is an every-varying hotchpotch of people. It is in practice unknown; therefore $P(D \mid \bar{B})$ is also unknown.

If $P(D \mid \bar{B})$ cannot be estimated the algorithm breaks down. We could accept the unavailability of $P(D \mid \bar{B})$, compute the limits between which $P(D \mid \bar{B})$ must be ... and derive upper and lower probabilities, or we could use the transferable belief model, which allocates parts of a total unitary belief to some subsets (as with probabilities) without requiring additivity (Smets, 1988). The part of belief assigned to a subset quantifies the amount of belief that supports that subset and does not support any strict subset due to lack of relevant information. If new relevant pieces of evidence become available, that amount of belief could be allocated to some subsets. This model implies implicitly the use of belief functions and Dempster's rules of conditioning and combination. This transferable belief model looks similar to Dempster and Shafer's model (Shafer, 1976) except that it does not require any underlying concept of probability and Dempster's rules are not arbitrary. In that model, $\text{bel}(D \mid B)$ is defined as in the Bayesian model but $\text{bel}(D \mid \bar{B})$ can be described by a vacuous belief function that describes adequately a state of total ignorance (which can hardly be done within the Bayesian framework). The use of the transferable belief model requires fewer conditional probabilities than the Bayesian model and the computation will be even easier than that needed within the Bayesian framework.

Dr Alun Thomas (University of Bath): As a mathematical geneticist I feel deeply concerned that the authors have failed to mention the method of 'joint-peeling' developed by Cannings *et al.* (1976, 1978) for the exact calculation of probability functions on arbitrarily complex pedigrees. This method has been used routinely for many years for efficient calculation of likelihoods and probabilities on large highly looped graphs, and user friendly computer programs for its application exist. The authors' statement that intermarriage is dealt with by decomposing the graph into trees, making repeated calculations and averaging the results is quite wrong.

There are many ideas presented in this paper which are already very familiar to anyone with experience of peeling. The efficient method of summation given in equation (4.2) is the very essence of peeling. The ψ functions introduced in equation (4.3) are the R functions of Cannings *et al.* The conditional probability tables of Section 7.1 are transmission probabilities, which in pedigree analysis would usually express Mendel's laws of inheritance. There are also clear links between the set chain with the 'running intersection property' and the cut set sequence which defines the order of calculation when peeling.

The notion of triangulating a graph is new to me, however, and while it is not necessary for calculating probabilities it may have a contribution to make in terms of efficiency, particularly in the case where repeated calculations are required in the face of new data.

The peeling method has been used exclusively in pedigree analysis, so it is extremely interesting that the authors have displayed its potential for applications in expert systems, but due recognition should be given to those who originally developed it.

Dr D. L. Tritchler (Ontario Cancer Institute, Toronto): The relationship of probability to frequency data seems a great advantage when such data are available. The 'Feigenbaum bottle-neck' or knowledge engineering problem is often mentioned as an obstacle to the construction of expert systems. Eliciting expert knowledge by dialogue is laborious, and research in computer induction of expert knowledge is aimed at automating this process (Michie, 1985). The statistical work cited on graphical models for contingency tables bears directly on estimating the networks treated in this paper. The triangulated model suggested by the authors is especially attractive since its parameters can be estimated without iteration and could be merged with a database which provided constantly updated probabilities. Some of the work on computer induction aims to formalise what a 'meaningful' relationship is (Michalski and Stepp, 1983) and might suggest useful ideas for evaluating and selecting statistical models.

I commend the authors' emphasis on explanation and feel that it is critical to prove the explanatory power of their approach in applications. For the example described by Table 4, the fact that propagating $\tau = \bar{t}$ from clique $\{\tau, \lambda, \varepsilon\}$ to $\{\lambda, \varepsilon, \beta\}$ implies that ε and λ are logically equivalent is clearly expressed by the update ratios of zero, but not by the ratios 1.0401 and 1.0964, suggesting that we can define principles for interpreting the update ratio. The flow of evidence between and through cliques seems more complex to me than the flow between nodes in a tree structure; perhaps explanation subtrees could be derived as a simplified representation for communicating explanations. The mutual information and the influence seem promising for providing higher levels of explanation.

The observation that the cliques can be programmed as communicating objects is important, because it means that the model lends itself to parallel computing, which promises to be widely used for artificial intelligence applications in the future.

Professor Nanny Wermuth (University of Mainz): The authors do not only achieve their main goal, to show that 'exact probabilistic methods are computationally feasible' to perform inference in expert systems, but, more importantly, their lucid presentation of concepts will bridge gaps between different fields of specialisation, between graph theory, probability theory, statistics and expert systems. Thereby they open the road for discussion, criticism and contributions to the proposed methods in a much broader community of scientists.

The importance of decomposable (or multiplicative) structures for multivariate statistical models had been recognised before connections to graph theoretic concepts were known (Goodman, 1970; Haberman, 1974; Andersen, 1974; Sundberg, 1975; Wermuth, 1976a). It is fascinating to see that decomposable structures have now found another, different application. It is a pity though that the late G. A. Dirac, who seems to have been the first to have studied decomposability of graphs, cannot witness the applications of his 'rigid circuits' as they appear now in expert systems, in multivariate statistical analyses and in efficient retrieval of information from databanks.

Two formulations in the paper are potentially misleading. There is no advantage in attaching the term causal to the described networks if—as is stated in Section 2—'causality' is to mean nothing but

a directed relation. Conversely, it would have been helpful to attach a qualifier like 'revised' each time the names 'marginal distribution' or 'marginals on all cliques' or on nodes are used in Sections 5.2–5.4. This would have alerted the reader that at this stage the authors are no longer concerned with marginal distributions implied by the joint distribution, as in Section 5.1, but with revised marginals that result from a conditional distribution of a subset of nodes given that responses to the remaining nodes are known.

The usefulness of the proposed methods will depend on how well the effects of changes in specifications will be understood and integrated into analyses. This comprises not only effects of different assessments of probabilities or of directions of influence and of added edges, but also of nodes not yet included in the graph. Problems analogous to effects of omitted nodes in a graph have been discussed by social scientists as 'moderating effects' of variables (Zedeck, 1971). A general statistical treatment of such effects is still awaited and will be important for expert systems, as well.

J. Whittaker (University of Lancaster): I should like to make one or two points more explicit from the perspective of log-linear modelling.

The initial specification of probabilities as conditional probabilities on the causal diagram is only one way to initialise the system. In the paper, it becomes clear that the joint distribution specified in terms of conditional independences and evidence potentials is more fundamental and these could be initialised directly. This is natural for the log-linear modeller because (apart from taking logarithms) the evidence potentials are just the interaction terms in the standard log-linear expansion.

That decomposable log-linear models have the runnning intersection property was used by Haberman (1974) to define decomposability, though he did not give an algorithm to generate the cliques in the correct order. That class is of great importance because it is the one for which maximum likelihood estimators have direct estimates (and similarly perform efficient probability calculations).

I am interested in the author's application of the Kullback–Leibler information divergence. In log-linear modelling, generalised log-likelihood ratio tests for testing certain interactions are zero, have the same mathematical structure as this divergence, and a more systematic treatment of the issues of planning and influence may be possible through evaluating the additive elements of the entropy. Such an approach for log-linear modelling has been briefly suggested by Whittaker (1984a, b).

The **authors** replied later, in writing, as follows.

We are extremely gratified both by the vigour of the argument concerning our paper and the positive nature with which it has been expressed.

We note that, despite our efforts, the reference list was clearly deficient, particularly with regard to the genetics literature. The fact that only around half of the contributors would label themselves as statisticians is indicative of the interdisciplinary nature of this area, and the similarity of many of the points raised by discussants bears witness to the concurrent international research effort. It is to be hoped that one side-effect of such a pooling of comments will be to help to prevent too much parallel processing of ideas.

We shall first cover the issues most commonly raised and then make necessarily brief replies to questions asked by particular discussants. The first general point concerns terminology.

What do we mean by 'causal'?

Our use of 'causal' has been criticised (Cheeseman, Shachter and Wermuth), and we agree that it will generally be interpreted too literally. Thus our first revision of opinion is to begin to use the term 'influence diagram' in place of 'causal network', in the hope that this will clarify our interpretation of the directed links. Since study of influence diagrams is not restricted to probabilistic interpretations (J. Q. Smith), we also obtain a stronger connection with the work of Dempster and Almond, Kong, and Shafer and Shenoy, who use a different calculus on the same qualitative structure.

Are the conditional independence properties expressed in our influence diagrams appropriate?

Drs Dubois and Prade consider that our representation may be overspecified through making independence assumptions by default and consider it strange that a point probability may well change after an adjustment to the structure. Dr Pratt correctly points out that conditional independence is lost if categories of a parent variable, say corresponding to the extent of a phenomenon, are combined, while Dr Rector points out that we assume exhaustive categorisation, and our use of 'not disease' and MUNIN's 'other disease' may lead, at a minimum, to problems in probability specification (see also Professor Smets's comments).

We can only re-emphasise our comments in Section 1; our approach is *model based*, in that we make no claim that either the qualitative structure or the quantitative probabilities are in any sense 'true'. The predictions made by the system, however, are logical consequences of our explicit assumptions, and hence those assumptions may be called into question by inadequacies in performance. This may lead to finer categorisation of variables, additional or fewer links, or revised probabilities, and afterwards the predictions will inevitably change. If $\bar{B}$ is a 'hotchpotch' (Smets) *and it matters*, then it can be refined either in construction or in response to errors. In a sense, Dubois and Prade are correct in viewing our predictions as defaults which hold unless something questions our assumptions, but these assumptions are not 'read off' the graph—they are the basis for constructing the graph and should be explicitly justified.

Where do the numbers come from, and how can we cope with their imprecision?

Drs Critchley, Cheeseman and Rector point out the large number of quantities that will need to be specified in a realistic system, and that available data cannot be expected to dominate the initial prior specification. As Critchley emphasises, this makes it all the more important not only to make explicit the imprecision in the assessments but also to allow them to improve as data accumulate. Fortunately, the hyper-Markov distribution described by Professor Dawid provides a conjugate prior on decomposable models that promises a firm theoretical foundation for future developments in this area, and our current thinking is to set up a particular type of hyper-Markov distribution as follows.

Above the *qualitative* 'core' of the influence diagram, which represents the independence assumptions, we place a graph termed the 'experience', which contains the *quantitative* specification. At its simplest, this extension consists of an additional parent node for each variable v in the core, which specifies belief about $p(v \mid \Pi_v)$, which is considered a random quantity. At initialisation, the expectation of the current belief concerning $p(v \mid \Pi_v)$ is dropped down into the core, and the operations described in our paper take place. When data on the patient are exhausted, the beliefs in $p(v \mid \Pi_v)$ are updated using standard Bayesian revision and retained for the next case. If full data are available, for example in a training set, or only particular configurations of missing data are allowed, then initial Dirichlet priors on the tables provide an exact, fully conjugate analysis in which all quantitative knowledge may be expressed in terms of counts on clique marginals (Porteous and Tritchler), and the operations are a simple extension of those in our paper.

The crucial point is that the strict probabilistic approach enables each of the conditional probabilities to be modelled with standard parametric techniques and allows us to use the full body of statistical methods. In particular, introducing quantitative nodes and using models such as described by Lauritzen and Wermuth (1987) make techniques of linear and logit regression available.

Several problems remain to be investigated in detail. Firstly, Professor A. F. M. Smith points out that, unless we assume initial marginal independence between the $p(v \mid \Pi_v)$s, the structure of the extended influence diagram may become extremely complex. Secondly, for many patterns of missing data, our continuing assumption of marginal independence becomes unfounded and needs to be monitored. Thirdly, Rector and Smets' observation that particular quantities may be rather tenuous and very sensitive to context is very important, and emphasises the need for precise specification of sources of training data. However, we would strongly deny Professor Smets' assertion that probabilities may be 'completely unknown' and believe that the dangers posed by data from an inappropriate context may be reduced by careful monitoring of any systematic deviation of updated probabilities from prior specifications, which could lead to questioning of the structure, the prior specification or the source of data.

In reply to Dr Cheeseman, the probabilities in MUNIN were largely based on physiological and anatomical knowledge.

How should we initialise, criticise and elaborate qualitative structure?

This is probably the most difficult issue and the one least amenable to technical solution. Both A. F. M. Smith and J. Q. Smith query when functional nodes should be introduced, while Professor Hand is concerned with both initial structuring and whether extensions will be smoothly accommodated. Both Professor Hand and Dr Kendall recognise the need for checking whether a simpler structure might suffice; Dr Porteous asks how much effect 'wrong' structure has, and Critchley, Havránek, A. F. M. Smith, Tritchler and Wermuth all emphasise formal mechanisms for model criticism and selection.

Our views on these crucial tasks are decidedly ill structured, and we consider this a vital area for future research. Data-based choice among graphical models has been studied by, for example, Wermuth (1976b) and Edwards and Havránek (1985, 1987) while the existence of proper priors supports the use of Bayes factors to compare models of differing dimensionality (Spiegelhalter and Smith, 1982).

Dr Cheeseman has also been investigating the use of stochastic complexity techniques (Cheeseman, 1984; Rissanen, 1987). However, we do not believe that structural initialisation or change should ever be based solely on statistical criteria, and hence cannot be fully automated. It is certainly advisable to have an independent store of full patient data and not to rely on just counting events on the cliques of an early representation.

For what kind of influence diagram is our approach appropriate?

Professor Hand and Professor Henrion point to applications in which a child may have a large number of parents and hence cliques become unmanageable, while Professors Barlow, Cooper and Shachter all identify the inefficiency of our technique in some diagrams, such as when two parents of a common child have a distant common ancestor. We can only agree with their universal opinion that it is fruitless to search for a single computational technique for handling all probabilistic influence diagrams. Other loop breaking techniques such as sketched by Pearl, Kelly and Cooper might be more appropriate for parts of a structure, and we hope that future years will generate sufficient experience to provide some reliable heuristics for matching techniques, both exact and approximate, to structure. We are grateful to Professor Barlow for providing a link to the fault diagram literature, and since our approach can happily handle logical links it will be interesting to make comparisons. Professor Henrion's nodes with 150 parents need careful consideration and presumably arise from trying to convert a network of *propositions* into a network of *variables*. Can the variables really be marginally independent, or could they be combined into a single node with a large number of states? Our heuristic is that, if it is unreasonable to think of the necessary quantities, then the structure is probably inappropriate.

How does our approach relate to trees of cliques?

Professors Pearl, Shachter, and Shafer and Shenoy all query whether we can consider our approach as just Kim and Pearl's (1983) algorithm on a tree of cliques. While this could be considered the case for propagation of single items of evidence (Section 8.2), the absorption then global propagation of multiple items requires a double pass with a varying root node and hence an undirected tree structure appears more appropriate. Drs Olesen and Andersen and Jensen briefly describe the process of absorbing and distributing evidence in a junction tree of cliques, and it is illuminating to find Professors Dempster and Almond putting forward an apparently identical scheme using the language of belief functions. We agree that the junction/join-tree formulation is more flexible than a linear structure with the running intersection property and see local propagation schemes in junction trees as the basic tool for efficient implementation. It has been particularly instructive for us to see the work of Olesen, Andersen and Jensen, which has clarified and simplified our original suggestions and fed back to provide new insight.

However, there are several reasons why it is advantageous to keep both the influence diagram and the clique trees. Firstly, the clique tree (i.e. the corresponding marginal representation) contains considerable redundancy. When introducing imprecision (see earlier), it is more natural and more efficient to do so on the original 'non-redundant' influence diagram, and this also holds when modifying and criticising the model. In addition, observations, or the lack of these, may occasionally break loops in the influence diagram and fill-ins can be avoided; in Fig. 8 the fill-in between λ and β is really unnecessary, because the loop is broken by revealing smoking history. It is also not completely true, such as hinted by Dempster and Almond, that the tree captures the Markov structure, since there are conditional independencies that cannot be read off the clique tree.

How can we express irrelevance?

Professor Hand raises the question about 'relevance', which leads to another reason for retaining explicit memory of the original directed graph. Suppose nodes E are observed and we want to give statements about nodes in F (and no other nodes): one situation in which this naturally occurs is when that data structure is 'nested' rather than 'crossed'; for example, in cervical cancer screening the part of the diagram relating to a smear result is irrelevant if no smear was taken. Then it is favourable first to reduce the diagram to involve only nodes in the smallest 'initial segment' containing E and F, where an initial segment is a subset A of V such that if $v \in A$ then all v's ancestors are in A. Then *after* this reduction of the diagram (which potentially could be dramatic) the triangulation and fill-in could be made and the clique tree constructed. In this clique tree we then only have to make a *partial propagation*, essentially moving evidence through the tree from cliques intersecting E to cliques intersecting F. The observation of Hilden that the update ratio is a martingale can also be used to stop propagation when information dies out.

Is our scheme restricted to a probabilistic interpretation?

Professors Dempster and Almond, Havránek, Kong, Shafer and Shenoy and Smets all consider our use of probabilities as a restrictive expression of uncertainty. Professors Pearl and Shafer and Shenoy have expressed very clearly that the influence diagram is the crucial reasoning tool and, although we have a strong bias towards probability, it is perfectly feasible to manipulate other measures of uncertainty using the tools described in our paper.

How is our approach related to connectionist models and parallel processing?

Drs McCleish, Rector and Tritchler all suggest implementation in parallel processing environments, and the computational structure of Olesen, Andersen and Jensen lends itself to local message passing between autonomous processors. As Dr Kelly points out, parallel processing would also be suitable for stochastic simulation approximations. Connectionist or neural network models (Good, McCleish, Rector) would generally be concerned with training a graph from scratch using data alone and no prior interpretation to internal nodes of the system. We might hypothesise that using expertise to initiate a structure, and then using parameter updating and model elaboration as described earlier, might be a more efficient means of developing a good multilevel classifier from limited training data. This remains to be investigated, but it is indisputable that the whole area of parallel distributed processing should be examined by statisticians, since graphical models form the natural link between radically different views on knowledge representation and learning.

We now consider points raised by individual discussants in order, omitting those that have already been covered under the general topics.

Dr Kelly questions whether the moral graph necessarily represents the minimal Markov field. Typically, this will be so, although it is not strictly necessary. Suppose, as Kelly hypothesises, that $p(\delta \mid \varepsilon, \beta) = \phi(\delta, \varepsilon)\, \psi(\delta, \beta)$ and $p(\varepsilon, \beta) = p(\varepsilon)\, p(\beta)$ since they are unjoined parents. Then we have $\varepsilon \perp\!\!\!\perp \beta \mid \delta$ as well as $\varepsilon \perp\!\!\!\perp \beta$, which together imply that either $\varepsilon \perp\!\!\!\perp (\beta, \delta)$ or $\beta \perp\!\!\!\perp (\varepsilon, \delta)$, which means a link could be removed from the original influence diagram. Thus, if the initial diagram is 'minimal', in the sense that no link may be removed without violating equation (7.1), then the moral graph will in most cases correspond to the minimal Markov field. In the area of approximations, Professors Henrion and Pearl are exploring stochastic simulation methods and it is clearly important to compare performance and computational efficiency, particularly in those diagrams with structures that will cause problems to exact methods.

Professor Hand raises many important issues. We are encouraged by his suggestion that an influence diagram representation may be able to deal with ill-structured problems, although it may need to be embedded in a structure similar to that proposed by Professors Berzuini and Steffanelli; experience will tell. As the diagram grows, we feel that 'local' elaborations in structure can be smoothly accommodated, but we are unsure of robustness to sudden inclusion of long-range links forming dramatic short-cuts. Hand, and other discussants, raises the issue of statistical parsimony against modelling the perceived world. We feel there should be a compromise; several widely different representations can often give roughly the same predictive ability, and it is then reasonable to choose that which corresponds most closely to human perception, to allow explanation and ease of expert input in the light of failures of performance, but we are aware of the dangers of unwarranted complexity.

Mr Thatcher's suggestion of eliminating structure altogether has been implemented in large medical databases such as ARAMIS (Fries, 1976). But, although computational complexity is low, data requirements are immense for even moderate dimensional problems. We feel that parsimonious modelling is essential, but the idea of only computing probabilities relevant to the particular case could be very important in implementing strategies for 'relevance' in our model-based approach.

We agree with Dr J. Q. Smith that the structuring and the properties of influence diagrams are important areas of study in their own right. The issue of 'conditional conditional independence' is also raised by Dr Hilden, and again reflects the need for an efficient mechanism for dynamic diagram alteration. We are unsure of the appropriate circumstances to introduce logical nodes.

We disagree with Professor A. F. M. Smith that 'planning' and 'influence', at least as described in our paper, will present challenging computational problems. Our planning suggestion merely involves a global propagation assuming each possible realisation of the node under scrutiny, and then possibly a use of Bayes theorem on each potentially observable node (see our reply to Dr Hilden). 'Influence' requires a propagation for each node to be removed, but, since only nodes on a certain path are relevant, computational short-cuts should be possible. Model comparison and probabilities as parameters have been previously covered, although we note that relevant references for exponential families on

influence diagrams include Wermuth and Lauritzen (1983, 1987), Lauritzen and Wermuth (1984, 1987) and Barndorff-Nielsen and Blæsild (1988).

Dr Hilden makes several important points. Introduction of decision nodes will further relate our work to classical influence diagrams, and we agree that a logarithmic scoring rule may be a misleading 'pseudo-utility' (see also Professor Shachter's comments). The quadratic scoring rule is better, and we were wrong to think it substantially more complex to calculate. However, as Dr Whittaker points out, monitoring influence through Kullback–Leibler distances should have a relationship to classic likelihood ratio statistics for model choice. Once again the issue of dynamic restructuring comes up, and we feel sure that Dr Hilden will be as full of novel ideas and critical suggestions as he has always been in the past.

Dr Kendall's suggestion of computer algebra relates to Hilden's (1970, 1982) work in manipulation of algebraic expressions in complex pedigrees, and we agree with the educational value of these manipulations. The suggestion of the Ito calculus is certainly novel.

The collaboration between Drs Gammerman and Aitken on applications in forensic science presents the methodology with an extremely challenging but important area of problems and will no doubt give rise to unforeseen difficulties. However, the papers by Schum and Martin (1981, 1982) referred to by Dr Phillips strongly suggests that legal reasoning *is* a feasible application.

The use of partially ordered sets by Dr J. D. Andersen is certainly an extremely elegant means of rapidly obtaining node marginals from the original probability tables, but we require more than just the node marginals for evidence propagation, and the technique does not appear to be repeatable once evidence has come in, since our beliefs are no longer causal Markov on the original graph.

Professors Berzuini and Steffanelli suggest a composite logical–probabilistic structure in which the influence diagram model may be embedded. Some kind of synthesis must be inevitable, and the use of logic in the abduction–hypothesis formation step, and probabilistic reasoning in the deduction–induction phases, echoes Szolovits and Pauker's (1978) view that 'categorical proposes, probability disposes'. We could perhaps view the logical 'abduction' step as simply the means of dynamically focusing on 'relevant' parts of the graph, which has repeatedly been stressed as a crucial task. The important question is whether the 'relevance' manipulations can be carried out locally within the influence diagram–junction tree representation, or whether a higher level of control is necessary. We hope that experiments on Berzuini and Steffanelli's excellent ANAEMIA project will provide insight into this.

We have already discussed Drs Dubois and Prade's comments on our independence assumptions, and we acknowledge their careful study of alternative representations of uncertainty. We definitely see a link $A \rightarrow B$ as meaning that it makes sense for an expert to think of $p(B \mid A)$. We do have a problem if an expert wishes to provide cyclical assessments, but we feel that these can be dealt with by thinking of the expert as directly providing a joint distribution on the clique A, B, C. If his probability assessments are not coherent, then this should be pointed out and they should be revised—an expert's assessments are not God given and revealing his incoherence should improve understanding, and may lead to revisions in structure. As other discussants have pointed out, our conditional probability tables are not the only way of initialising the system, although we hope it would usually be adequate as we would usually see, as J. Q. Smith recommends, the qualitative influence diagram being elicited before the quantitative 'experience' is assessed.

Professors Fienberg and Meyer suggest translation back into log-linear modelling, and we would also claim these techniques may be useful in developing efficient programs for the analysis of large contingency tables. Such work is in progress, see Badsberg (1986). Indeed, we hope that the future will see a breakdown in the distinction between 'statistics in artificial intelligence' and 'artificial intelligence in statistics', since a common set of techniques will exist for representing and manipulating complex multidimensional structures, combining the power of artificial intelligence representation with established statistical modelling. Although we do not see any direct linkage between our work and the sociological use of random networks to describe relationships between specific individuals, the possibility of using such models as prior distributions on the structure of the network deserves to be investigated; see also Frank and Strauss (1986).

We are grateful to Professor Good for his references and feel that his work on hierarchical models in contingency tables may yet have an important role in this area.

Professor Havránek is right to emphasise the elegant algebraic work on general measures of uncertainty, and we reiterate that most of our proposals do not necessarily require a probabilistic interpretation for uncertainty. His work with Edwards will be vital in criticism and elaboration of graphical structure, but while we agree that linking expert knowledge with knowledge obtained from data *can* be very dangerous, we also feel that developing appropriate methodology to do so is one of

the most vital challenges for future statistical/artificial intelligence research. *Not* combining knowledge from both sources would be even more dangerous.

We take note of Professor Kong's valuable point that potentials should be marginalised in connected subsets. If the junction tree representation of Jensen is used then this will occur automatically, again providing insight into the general theory. We agree that maximal clique state size is the crucial element in complexity and look forward to his recommendation for suitable algorithms.

We see our method as an alternative to, rather than a result of, Brook's expansion mentioned by Professor Mardia. We are sorry that we did not have room for more examples.

Professor Pearl's contribution is particularly welcome as he has consistently revealed astonishing insight into this subject area. We agree with both him and Hilden that *structure* between variables has received too little attention in statistics, although there is a non-negligible body of statistical literature on graphical models (see the references in the paper). His consideration of cut sets relates to the algorithms in pedigree analysis discussed in the following.

Professor Shachter makes several important points, most of which have been covered previously. We should note that direct manipulation of the original influence diagram representation can be attractive in prototyping, and within 30 minutes of receiving Shachter's DAVID program we were analysing our dyspnoea example. But in larger examples the revisions in topology can be rather bewildering and inefficient for repeated belief revisions.

We are apologetic for not referring to Schum's important work in the area of evidential reasoning and are grateful to Dr Phillips for the quote describing the taxonomy of generic structures. It certainly appears to us that all can be accommodated within an influence diagram framework, and we intend to explore examples in the hope that our graphical and computation scheme may illuminate legal issues. In particular such examples would seem to be good tests of any proposed automatic explanation facilities.

We apologise to Dr Thomas and others in mathematical genetics for our inadequate background research, and it is unfortunate that we did not acknowledge the important work in this area; see Thompson (1986) for a simple introduction which features expressions very similar to our equation (4.2). Complex pedigrees of the kind discussed by Cannings *et al.* (1978) could be handled using our technique, although the graphs would be somewhat more complex, having a node for each individual's observed phenotype, each of which has a single parent (in our sense) node standing for the unobserved genotype; these genotypes form the pedigree with transition probabilities governed by Mendelian laws of inheritance, while the genotype–phenotype transition probabilities are the 'penetration functions'. The objective may be to marginalise over the unobserved genotypes to obtain an overall likelihood for what has been observed, or to calculate conditional probabilities on sets of genotypes, or, in the case of genetic counselling, future phenotypes. The procedures of Cannings *et al.* (1978) for unlooped pedigrees are extremely similar to Pearl's (1986) procedure for trees, while their suggestions for loops involve a succession of cut sets as suggested in Pearl's contribution.

The R functions of Cannings *et al.* (1978) are the ϕ functions we calculate on the clique intersections when absorbing evidence onto a potential representation with $Z = 1$. In particular, we note that the overall probability of the pedigree trivially falls out of our calculations: if we have obtained a potential representation with $Z = 1$, the probability of any particular configuration of states x_E^* at nodes E can be calculated as the normalisation constant Z^*, see equation (9.2), when conditioning on x_E^* being observed and processing from potentials to set chain. This is because

$$p(V \setminus E \mid E^*) = p(V, E^*)/p(E^*)$$

and from Section 8.1

$$p(V \setminus E \mid E^*) = p(V, E^*)/Z^*$$

whereby

$$p(E^*) = Z^*.$$

This gives an alternative to the methods described in Section 11.2 and it would be of interest to see how our techniques compare in computational efficiency with those established in pedigree analysis.

We hypothesise that our approach may have considerable advantages over peeling methods, in that for similar computational cost we obtain conditional probabilities for all currently unknown genotypes and phenotypes. Unknown penetrance probabilities could also be modelled as additional nodes, and Kong's contribution should be relevant to automatic search procedures for cut set sequences (Thomas,

1986) and vice versa. In return, the statistical techniques developed for criticising and revising the structure of pedigrees may well be adaptable to our general formulation.

Dr Tritchler raises the important but difficult area of explanation. If we adopt the junction tree computation scheme then we do have a tree structure and flow of evidence could be graphically displayed on clique intersections of the full moral graph. If our structures can be broken down into the generic components of Schum, then perhaps condensed verbal explanation facilities may be more easily combined. In addition, when allowing imprecision in probabilities, it would be important to retain a description of the *source* of the experience to help to justify the conclusions. This remains a fascinating area to explore.

We are grateful to Professor Wermuth for pointing out the relation to the notion of moderating effects and see this as yet another benefit of the cross-fertilisation between disciplines obtained by relating the expert system problems to statistical modelling, as also pointed out by Whittaker, and Fienberg and Meyer. A slight amendment to the history of decomposable graphs: they were certainly studied by Wagner (1937) although Dirac (1961) seems to give the first comprehensive treatment.

In conclusion, the comments of the discussants show clearly that the issues that we have tackled only form a minor part of those that need to be eventually faced. We hope that the material collected here in the paper as well as in the discussion will stimulate statisticians to join in research efforts.

REFERENCES IN THE DISCUSSION

Aitchison, J. (1986) *The Statistical Analysis of Compositional Data.* London: Chapman and Hall.

Andersen, A. H. (1974) Multidimensional contingency tables. *Scand. J. Statist.*, **1**, 115–127.

Arnborg, S., Corneil, D. G. and Proskurowski, A. (1987) Complexity of finding embeddings in a k-tree. *SIAM J. Alg. Disc. Meth.*, **8**, 277–284.

Asmussen, S. and Edwards, D. (1983) Collapsibility and response variables in contingency tables. *Biometrika*, **70**, 567–578.

Badsberg, J. H. (1986) Kontingenstabeller (in Danish). *Thesis.* Aalborg University.

Barndorff-Nielsen, O. E. and Blæsild, P. (1988) Combination of reproductive models. *Ann. Statist.*, **16**, 323–341.

Bedrick, E. J. (1983) Adjusted chi-squared tests for cross-classified tables of survey data. *Biometrika*, **70**, 591–595.

Bishop, Y. M. M. (1971) Effects of collapsing multidimensional contingency tables. *Biometrics*, **27**, 119–128.

Brook, D. (1964) On the distinction between the conditional probability and the joint probability approaches in the specification of nearest-neighbour systems. *Biometrika*, **51**, 481–483.

Cannings, C., Thompson, E. A. and Skolnick, M. H. (1976) Recursive derivation of likelihoods on pedigrees of arbitrary complexity. *Adv. Appl. Probabil.*, **8**, 622–625.

———(1978) Probability functions on complex pedigrees. *Adv. Appl. Probabil.*, **10**, 26–61.

Chatterjee, P. (1975) Modularization of fault trees: a method to reduce the cost of analysis. *Reliability and Fault Tree Analysis* (eds Barlow, Fussell and Singpurwalla). Philadelphia: SIAM.

Cheeseman, P. (1984) Learning of expert systems from data. *Proc. IEEE Workshop Principles of Knowledge-based Systems, Denver*, pp. 115–122.

Coombs, M. and Alty, J. (1984) Expert systems: an alternative paradigm. In *Developments in Expert Systems* (ed. M. Coombs), pp. 135–157. London: Academic Press.

Cooper, G. F. (1987) Probabilistic inference using belief networks is NP-hard. *Research Report KSL-87-27.* Medical Computer Science Group, Stanford University.

Critchley, F. and Ford, I. (1985) Interval estimation in discrimination: the multivariate normal equal covariance case. *Biometrika*, **72**, 109–116.

Critchley, F., Ford, I. and Rijal, O. (1988) Interval estimation based on the profile likelihood: strong Lagrangian theory, with applications to discrimination. *Biometrika*, **75**, in the press.

Darroch, J. N., Lauritzen, S. L. and Speed, T. P. (1980) Markov fields and log-linear models for contingency tables. *Ann. Statist.*, **8**, 522–539.

Dawid, A. P. (1979) Conditional independence in statistical theory. *J. R. Statist. Soc.* B, **41**, 1–31.

Dempster, A. P. (1972) Covariance selection. *Biometrics*, **28**, 157–175.

Dempster, A. P. and Kong A. (1986) Uncertain evidence and artificial analysis. *Research Report S108.* Department of Statistics, Harvard University.

Edwards, D. and Havránek, T. (1985). A fast procedure for model search in multidimensional contingency tables. *Biometrika*, **72**, 339–351.

———(1987). A fast model selection procedure for large families of models *J. Amer. Statist. Ass.*, **82**, 205–211.

Eshera and Lewis (1987). *SPIE*, **786**.

Fienberg, S. E., Meyer, M. M. and Wasserman, S. (1985) Statistical analysis of multiple sociometric relations. *J. Amer. Statist. Ass.*, **80**, 51–67.

Frank, O. and Strauss, D. (1986) Markov graphs. *J. Amer. Statist. Ass.*, **81**, 832–842.

Fries J. F. (1976) A data bank for the clinician? *New Engl. J. Med.*, **294**, 1400–1402.

Gammerman, A. J. and Crabbe, W. J. (1987) Computational models of probabilistic reasoning in expert systems: a causal probabilistic reasoning system. *Technical Report 87/16.* Computer Science Department, Heriot-Watt University, Edinburgh.

Glasziou, P. and Hilden, J. (1988) Test selection measures. To be published.
Good, I. J. (1960) Effective sampling rates for signal detection: or can the Gaussian model be salvaged? *Inform. Control*, **3**, 116–140.
———(1961) A causal calculus. *Brit. J. Philos. Sci.*, **11**, 305–318; **12**, 43–51.
———(1981) The weight of evidence provided by uncertain testimony or from an uncertain event. *J. Statist. Comput. Simuln*, **13**, 56–60.
———(1983) *Good Thinking: the Foundations of Probability and its Applications*, pp. 253–263. Minneapolis: University of Minnesota Press.
———(1984) Weights of evidence in medical diagnosis. *J. Statist. Comput. Simuln*, **19**, 171–173.
———(1986) The whole truth. *Inst. Math. Statist. Bull.*, **15**, 366–373.
———(1988) Causal tendency: a review. In *Causation, Change and Credence* (eds W. Harper and B. Skyrms). Dordrecht: Reidel.
Goodman, L. A. (1970) The multivariate analysis of qualitative data. *J. Amer. Statist. Ass.*, **65**, 226–256.
Haberman, S. J. (1974) *The Analysis of Frequency Data*. Chicago: University of Chicago Press.
Hájek, P. and Valdés, J. J. (1987). Algebraic foundations of uncertainty processing in rule-based expert systems. *Research Report 28/1987*. Mathematical Institute, Czechoslovak Academy of Science, Prague.
Hand, D. J. (1987) A statistical knowledge enhancement system. *J. R. Statist. Soc.* A, **150**, 334–345.
Hilden, J. (1970) GENEX—an algebraic approach to pedigree probability calculus. *Clin. Genet.*, **1**, 319–348.
———(1982) Computerized derivations of Mendelian probability formulae: the GENEX processor. *Proc. Nordic Symp. Applied Statistics and Data Processing* (eds A. Höskuldson *et al.*), pp. 395–410. NEUCC–Technical University of Denmark.
Hilden, J., Habbema, J. D. F. and Bjerregaard, B. (1978) The measurement of performance in probabilistic diagnosis: III, methods based on continuous functions of the diagnostic probabilities. *Meth. Inform. Med.*, **17**, 238–246; **20**, 97–100.
Holland, P. W. and Leinhardt, S. (1981) An exponential family of probability distributions for directed graphs. *J. Amer. Statist. Ass.*, **76**, 33–50.
Howard, R. A. and Matheson, J. E. (1984) Influence diagrams. In *The Principles and Applications of Decision Analysis* (eds R. A. Howard and J. E. Matheson), vol. II. Menlo Park: Strategic Decisions Group.
Jensen, F. V. (1988) Junction trees—a new characterization of decomposable hypergraphs. *JUDEX Research Report*, to be published.
Kendall, W. S. (1988) Symbolic computation and the diffusion of shapes of triads. *Adv. Appl. Probabil.*, Dec., to be published.
Kent, J. T. and Mardia, K. V. (1988) Spatial classification using fuzzy membership models. *IEEE Trans. Pattern Anal. Mach. Intell.*, to be published.
Kim, J. H. and Pearl, J. (1983) A computational model for causal and diagnostic reasoning in inference systems. *Proc. 8th Int. Joint Conf. Artificial Intelligence, Karlsruhe*, pp. 190–193.
Kong, A. (1986a) Multivariate belief functions and graphical models. *PhD Thesis*. Department of Statistics, Harvard University.
———(1986b) Construction of a tree of cliques from a triangulated graph. *Technical Report S-118*. Department of Statistics, Harvard University.
Lauritzen, S. L. (1982) *Lectures on Contingency Tables*, 2nd edn. University of Aalborg Press.
Lauritzen, S. L. and Wermuth, N. (1984) Mixed interaction models. *Research Report R-84-8*. Institute of Electronic Systems, Aalborg University.
———(1987) Graphical models for association between variables, some of which are qualitative and some quantitative. *Research Report R-87-10*. Institute of Electronic Systems, Aalborg University.
Lee, S. K. (1977) On the asymptotic variances of u-terms in loglinear models of multidimensional contingency tables. *J. Amer. Statist. Ass.*, **72**, 412–419.
Liu, C. L. (1987) *Elements of Discrete Mathematics*, 2nd edn, p. 120. New York: McGraw-Hill.
Mardia, K. V. (1988) Multi-dimensional multivariate Gaussian Markov random fields. *J. Mult. Anal.*, to be published.
McCarthy, J. and Hayes, P. (1969) Some philosophical problems from the standpoint of artificial intelligence. In *Machine Intelligence 4* (eds B. Meltzer and D. Michie), pp. 463–502. Edinburgh: Edinburgh University Press.
Michalski, R. S. and Stepp, R. E. (1983) Automated construction of classifications: conceptual clustering versus numerical taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.*, **5**, 396–410.
Michie, D. (1985) Current developments in artificial intelligence and expert systems. *Zygon*, **20**, 375–389.
Miller, R. A., Pople, H. E., Jr, and Myers, J. D. (1982) INTERNIST-1, an experimental computer-based diagnostic consultant for general internal medicine. *New Engl. J. Med.*, **307**, 468–476.
Olmsted, S. M. (1983) On representing and solving decision problems. *PhD Thesis*. Stanford University.
Paass, G. (1986) Consistent evaluation of uncertain reasoning systems. *Proc. 6th Int. Workshop Expert Systems and their Applications, Avignon, April*, pp. 73–94.
Pearl, J. (1986a) Fusion, propagation and structuring in belief networks. *Artific. Intell.*, **29**, 241–288.
———(1986b) Bayes and Markov networks: a comparison of two graphical representations of probabilistic knowledge. *Technical Report R-46*. Cognitive Systems Laboratory, UCLA.
———(1987a) Evidential reasoning using stochastic simulation of causal models. *Artific. Intell.*, **32**, 245–257.
———(1987b) Distributed revision of composite beliefs. *Artific. Intell. J.*

———(1988) *Networks of Belief: Probabilistic Reasoning in Intelligent Systems.* Morgan and Kaufmann. To be published.
Pearl, J. and Verma, T. (1987) The logic of representing dependencies by directed graphs. *Proc. AAAI Conf., Seattle, July*, pp. 374–379.
Porteous, B. T. (1985) Properties of log linear and covariance selection models. *PhD Thesis.* Cambridge University.
Quinlan, J. R. (1983) INFERNO: a cautious approach to uncertain inference. *Comput. J.*, **26**, 255–269.
Quinn and Deo (1984). *Comput Surv.*, **16**.
Reiter, R. (1980) A logic for default reasoning. *Artific. Intell.*, **13**, 81–132
Rissanen, J. (1987) Stochastic complexity. *J. R. Statist. Soc.* B, **49**, 223–239.
Rosenthal, A. (1975) A computer scientist looks at reliability computations. In *Reliability and Fault Tree Analysis* (eds Barlow, Fussell and Singpurwalla), pp. 133–152. Philadelphia: SIAM.
Schum, D. (1988) *Evidence and Inference for the Intelligence Analyst*, to be published.
Schum, D. and Martin, A. (1981) Assessing the probative value of evidence in various inference structures. *Technical Report 81-02.* Department of Psychology, Rice University, Houston.
———(1982) Formal and empirical research on cascaded inference in jurisprudence. *Law Soc. Rev.*, **17**, 105–157.
Shachter, R. D. (1986a) Evaluating influence diagrams. *Oper. Res.*, **34**, 871–882.
———(1986b) Probabilistic inference and influence diagrams. *Oper. Res.*, to be published.
Shachter, R. D. and Heckerman, D. E. (1987) Thinking backwards for knowledge acquisition. *AI Mag.*, **8**, 55–61.
Shafer, G. (1976) *A Mathematical Theory of Evidence.* Princeton: Princeton University Press.
———(1985) Conditional probability. *Int. Statist. Rev.*, **53**, 261–277.
Smets, P. (1988) Belief functions. In *Non Standard Logics for Automated Reasoning* (eds P. Smets, A. Mamdani, D. Dubois and H. Prade), pp. 253–286. London: Academic Press.
Smith, J. Q. (1987a) Influence diagrams for Bayesian decision analysis. *Research Report 100.* Department of Statistics, Warwick University.
———(1987b) *Decision Analysis: A Bayesian Approach.* London: Chapman and Hall.
———(1988) Models, optional decisions and influence diagrams. *Proc. 3rd Conf. Bayesian Statistics, Valencia*, to be published.
Spang Robinson (1986) *Spang Robinson Report*, vol. 2, No. 10. Palo Alto: Spang Robinson.
Spiegelhalter, D. J. (1986) Probabilistic reasoning in predictive expert systems. In *Uncertainty in Artificial Intelligence* (eds L. N. Kanal and J. Lemmer), pp. 47–68. Amsterdam: North-Holland.
———(1987) Coherent evidence propagation in expert systems. *Statistician*, **36**, 201–210.
Spiegelhalter, D. J. and Smith, A. F. M. (1982) Bayes factors for linear and log-linear models with vague prior information. *J. R. Statist. Soc.* B, **44**, 377–387.
Sundberg, R. (1975) Some results about decomposable (or Markov-type) models for multidimensional contingency tables: distribution of marginals and partitioning of tests. *Scand. J. Statist.*, **2**, 71–79.
Szolovits, P. and Pauker, S. G. (1978) Categorical and probabilistic reasoning in medical diagnosis. *Artific. Intell.*, **11**, 115–144.
Thatcher, A. R. (1988) Computer models of probabilistic reasoning: Bayes' Theorem without assuming independence. *Technical Report 88/1.* Computer Science Department, Heriot-Watt University, Edinburgh.
Thomas, A. (1986) Optimal computation of probability functions for pedigree analysis. *IMA J. Math. Appl. Med. Biol.*, **3**, 167–178.
Thompson, E. A. (1986) Genetic epidemiology: a review of the statistical basis. *Statist. Med.*, **5**, 291–302.
Titterington, D. M., Murray, G. D., Murray, L. S., Spiegelhalter, D. J., Skene, A. M., Habbema, J. D. and Gelpke, G. J. (1981) Comparison of discrimination techniques applied to a complex data set of head injured patients. *J. R. Statist. Soc.* A, **144**, 145–175.
Wagner, K. (1937) Über eine Eigenschaft der ebenen Komplexe. *Math. Ann.*, 114.
Wermuth, N. (1976a) Analogies between multiplicative models in contingency tables and covariance selection. *Biometrics*, **32**, 95–108.
———(1976b) Model search among multiplicative models. *Biometrics*, **32**, 153–163.
Wermuth, N. and Lauritzen, S. L. (1983) Graphical and recursive models for contingency tables. *Biometrika*, **70**, 537–552.
———(1987) Conditional independence graphs, graphical chain models, and data. *Research Report 87-2.* University of Mainz.
White, I. (1987) W(h)ither expert systems? *Newsl. Brit. Comput. Soc.*, **17**, 40–44.
Whittaker, J. C. (1984a) Model interpretation from the additive elements of the likelihood function. *Appl. Statist.*, **33**, 52–64.
———(1984b) Fitting all decomposable and graphical models to multiway contingency tables. In *Compstat 1984* (eds T. Havránek *et al.*), pp. 401–406. Vienna: Physia.
von Winterfeldt, D. and Edwards, W. (1986) *Decision Analysis and Behavioral Research*, p. 178. Cambridge: Cambridge University Press.
Wood, K. and McCullers, W. T. (1988) Fault tree analysis by direct pivotal decomposition. *Proc. 3rd SIAM Conf. Discrete Mathematicics*, to be published.
Zedeck, S. (1971) Problems with the use of "moderator" variables. *Psychol. Bull.*, **76**, 295–310.