F1.2 Inference Methods

C Borgelt [1], *J Gebhardt* [2], *and R Kruse* [1]

[1]Otto-von-Guericke-University of Magdeburg, Germany
[2]University of Braunschweig, Germany

Abstract

This section investigates graphical modeling as a powerful framework for drawing inferences under imprecision and uncertainty. We survey the semantical background and relevant properties of relational, probabilistic, and possibilistic networks and consider evidence propagation in such networks as well as methods for learning them from data. Whereas the probabilistic Bayesian networks and Markov networks are well-known for a couple of years, we focus on possibilistic networks as a promising approach to the efficient treatment of information-compressed uncertain *and* imprecise knowledge.

Keywords

possibility theory, knowledge representation, graphical models, relational networks, probabilistic networks, possibilistic networks, evidence propagation, learning from data

F1.2.1 Introduction

The essential feature of inference is that a certain type of knowledge — knowledge about truth, probability, (degree of) possibility etc. — is transferred from given propositions, events, states etc. to other propositions, events, states etc. For example, in a (deductive) logical argument the knowledge about the truth of the premises is transferred to the conclusion; in probabilistic inference the knowledge about the probability of an event enters the calculation of the probability of other, connected events and is thus transferred to these.

For the transfer carried out in an inference three things are necessary: knowledge to start from (e.g. the knowledge that a given proposition is true), knowledge that provides a path for the transfer (e.g. an implication, of which the given proposition is the antecedent), and a mechanism to follow the path (e.g. the rule of *modus ponens* to establish the truth of the consequent of the implication). Only if all three are given and fit together, the inference can be carried out. Of course, the transfer need not always be direct. In logic, for example, arguments can be chained by using the conclusion of one as the premise for another, and several such steps may be necessary to arrive at the desired conclusion.

From this description we can understand what the main problems of modeling inference are. They consist in finding the paths along which knowledge can be transferred and providing the proper mechanisms for following them. (In contrast to this, the knowledge to start from is usually available right away, e.g. from observations.) Indeed, it is well-known from applications of classical logic, that automatic theorem provers spend most of their time searching for a path from the given facts to the desired conclusion.

But we do not discuss (fuzzy) logical inference in this section — this is the topic of chapter **??** —, but deal with a framework for knowledge representation and inference under uncertainty that is known as *graphical modeling* (Whittaker 1990). It owes its name to the fact that in it the paths along which knowledge can be transferred are represented by a (hyper)graph. We study the basic

ideas of this framework in section F1.2.3. Since it significantly simplifies multivariate data analysis, applications of graphical modeling can be found in all areas in which dependent observations have to be analyzed, for example, in regression analysis, spatial analysis, and expert systems.

As a consequence of its origin in multivariate statistics, the most advanced numerical approaches to such a structured handling of uncertain information have been developed in the field of probabilistic graphical models (Cheeseman and Oldford 1994). For example, *Bayesian networks* (Pearl 1988) are established as a powerful tool for reasoning in knowledge based systems. They provide a well-founded framework in the presence of *uncertain*, but *precise* data. Probabilistic graphical models are discussed in section F1.2.5.

But graphical modeling is not restricted to probability theory, but can be used with other ways of representing uncertain or imprecise information as well. In section F1.2.4 we inspect *relational graphical models* (Kruse and Schwecke 1990, Dechter and Pearl 1992, Kruse *et al* 1994), which are much simpler than probabilistic graphical models and closely related to the theory of relational databases (Maier 1983, Ullman 1988). In contrast to probabilistic graphical models they can deal with *imprecise*, i.e. multi-valued, but *certain* information.

Such models are important, since the explicit treatment of imprecise information is more and more claimed to be necessary for industrial practice. But, on the other hand, the treatment of uncertainty cannot be dispensed with completely. Therefore it is reasonable to investigate graphical models related to alternative uncertainty calculi, the more so, as extending purely probabilistic approaches to the treatment of imprecise information usually renders the corresponding inference mechanisms computationally intractable. Under the aspect of efficiency, such uncertainty calculi should provide a justifiable form of *information compression* and *simplification* in order to support reasoning under uncertainty *and* imprecision without essentially affecting the expressive power and the correctness of decision making procedures.

Possibility theory (Dubois and Prade 1988) is a good choice for systems that have to carry out *approximate* instead of exact reasoning and that are characterized by a low sensitivity w.r.t. slight changes of information. From this we may conjecture that possibility theory will grow up to play the same role in the field of uncertain reasoning in knowledge-based systems as nowadays *fuzzy control* plays as a tool for (information-compressed) interpolation between points in vague environments in the field of control engineering (Kruse *et al* 1994). After a brief discussion of the underlying interpretation of possibility distributions, which also clarifies our understanding of the terms *imprecision* and *uncertainty*, we study *possibilistic graphical models* in section F1.2.6.

Graphical models of whatever kind are powerful tools for reasoning and easy to handle — if they are available. But, unfortunately, constructing them by hand can be tedious and time-consuming. To increase the usefulness of the graphical model approach, it is therefore desirable to look for automatic induction methods for such models. Since there are promising results for all of the discussed graphical model types, we mention some of them in the corresponding sections.

F1.2.2 Notation and presuppositions

Notation. Let $V = \{A^{(1)}, \ldots, A^{(m)}\}$ be a finite set of attributes (or variables) $A^{(k)}$, which are used to describe the section of the world under consideration. We assume the domains $\mathrm{dom}(A^{(k)}) = \{a_1^{(k)}, \ldots, a_{n_k}^{(k)}\}$ of these attributes to be finite sets of categorical values. (I.e. in our discussion we confine to the important case of *discrete graphical models*.) With these presuppositions the space in which all inferences take place is the joint domain $\Omega = \mathrm{dom}(A^{(1)}) \times \cdots \times \mathrm{dom}(A^{(m)})$, which is sometimes called the *universe of discourse*. Each possible state of the world is described by a tuple (or vector) $\omega = (a_{i_1}^{(1)}, \ldots, a_{i_m}^{(m)})$, that contains the values which the attributes in V assume for this state. For simplicity (and because we cannot distinguish between states for which the describing tuples are identical) we identify each $\omega \in \Omega$ with a possible state of the world.

Several times we need to refer to subspaces of Ω and projections of tuples ω to these subspaces. A subspace $\Omega_W \subseteq \Omega$ is the joint domain of a subset $W \subseteq V$ of attributes, i.e. $\Omega_W = \times_{A \in W} \mathrm{dom}(A)$. A projection of a tuple $\omega \in \Omega$ to this subspace is a tuple $\mathrm{proj}_W^V(\omega) = \omega_W \in \Omega_W$, which contains only the values of the attributes contained in W.

Presuppositions. For a situation in which we are about to draw inferences, we assume that the considered section of the world is in a specific state, whose description $\omega_0 \in \Omega$ we do not know or do

not know completely. The inferences to be drawn aim at identifying this state, i.e. at determining the values in ω_0.

To be able to carry out such inferences, we assume *generic knowledge* about dependencies between the values of different attributes to be available, i.e. knowledge that provides paths for the inferences. This knowledge may have been obtained from experts, textbooks, databases etc. and is represented as a distribution $\mathcal{D}$ on Ω. This distribution assigns to each tuple $\omega \in \Omega$ a value d_ω, which expresses the probability or (degree of) possibility of the combination of values present in ω. Depending on the values d_ω can have and the interpretation of these values we distinguish between relations, probability distributions and possibility distributions. Details are given in the corresponding sections below.

In addition to generic knowledge we need knowledge to start the inferences from — also called *evidential knowledge* —, which consists in restrictions on the possible values of some of the attributes. This knowledge could be obtained e.g. from observations made about the current state ω_0. From the evidential knowledge about the values of some attributes we infer, using the generic knowledge, restrictions about the values of other attributes, thus narrowing the set of states that have to be considered possible or likely for ω_0.

It is obvious that storing the generic knowledge directly, i.e. the distribution $\mathcal{D}$, would make reasoning very simple, since then we only have to select all $\omega \in \Omega$ compatible with the given evidential knowledge and to combine the corresponding values d_ω appropriately. But, unfortunately, if there are more than only very few attributes, the number of values d_ω to be stored in this case would exceed any reasonable limit. Hence other ways of representing the generic knowledge have to be found. One of them is graphical modeling, which we discuss in the next section.

F1.2.3 Graphical modeling

As already indicated in the introduction, in graphical modeling a directed or undirected (hyper)graph is used to represent the generic knowledge about the domain in which the inferences take place. Each vertex corresponds to an attribute, each edge to a dependence between attributes. The edges are also the paths along which knowledge about the values of one attribute can be transferred to other attributes. This is understandable, since no information can be transferred from an attribute to another, which is independent of the first.

But even if attributes are dependent, they are sometimes unconnected in a graphical model. The idea underlying this is that an inference, as already mentioned above, need not be direct. If the dependence between two attributes is captured completely by the consecutive dependences on a path connecting the two attributes via other attributes, a direct connection is not necessary. All inferences from one of the attributes to the other can then be carried out along this path.

Conditional independence. Such situations can be characterized by the notion of *conditional independence* (Dawid 1979, Pearl 1988). If two attributes get independent, if certain other attributes are fixed, their dependence is not genuine, but only mediated through other attributes. Therefore they need not be connected directly in the graph. Thus the topology of the graph is used to represent an independence model, i.e. a set of conditional independence statements, of the domain under consideration (Pearl 1988, Spirtes *et al* 1993).

Of course, not just any notion of conditional independence will do, since, as stated above, the aim is to replace an inference along a direct connection between attributes by an indirect inference. In order to allow such a replacement, the used notion of conditional independence has to satisfy certain axioms, which are known as the *semi-graphoid axioms* (Dawid 1979, Spohn 1980, Pearl and Paz 1987, Smith 1989). If we denote the independence of a set of attributes X from a set of attributes Y given a set of attributes Z as $X \perp\!\!\!\perp Y \mid Z$, they can be written as

symmetry: $(X \perp\!\!\!\perp Y \mid Z) \Longrightarrow (Y \perp\!\!\!\perp X \mid Z)$

decomposition: $(W \cup X \perp\!\!\!\perp Y \mid Z) \Longrightarrow (W \perp\!\!\!\perp Y \mid Z) \wedge (X \perp\!\!\!\perp Y \mid Z)$

weak union: $(W \cup X \perp\!\!\!\perp Y \mid Z) \Longrightarrow (X \perp\!\!\!\perp Y \mid Z \cup W)$

contraction: $(W \perp\!\!\!\perp Y \mid Z) \wedge (X \perp\!\!\!\perp Y \mid Z \cup W) \Longrightarrow (W \cup X \perp\!\!\!\perp Y \mid Z)$

The *symmetry* axiom states that in any state of knowledge Z, if Y tells us nothing new about X, then X tells us nothing new about Y. The *decomposition* axiom asserts that if two combined items of information are judged irrelevant to X, then each separate item is irrelevant as well. The *weak*

union axiom states that learning irrelevant information W cannot help the irrelevant information Y become relevant to X. The *contraction* axiom states that if we judge X irrelevant to Y after learning some irrelevant information W, then X must have been irrelevant before we learned W. Together the weak union and contraction properties mean that irrelevant information should not alter the relevance of other propositions in the system; what was relevant remains relevant, and what was irrelevant remains irrelevant (Pearl 1988). It is plausible that any reasonable notion of conditional independence should satisfy these axioms.

Independence graphs. Given an appropriate notion of conditional independence, which must be chosen specifically for the uncertainty calculus under consideration, an *independence graph* can be defined. In such a graph the *conditional independence* of two sets of attributes given a third is expressed by *separation* of the corresponding node sets by the nodes corresponding to the conditioning attributes.

What is to be understood by 'separation' depends on whether the graph is directed or undirected. If it is undirected, separation is defined as follows: If X, Y, and Z are three disjoint subsets of nodes in an undirected graph (UG), then Z separates X from Y, iff after removing the nodes in Z and their associated edges from the graph there is no path, i.e. no sequence of consecutive edges, from a node in X to a node in Y. Or, in other words, Z separates X from Y, iff all paths from a node in X to a node in Y contain a node in Z.

For directed graphs, which have to be acyclic, the so-called *d-separation criterion* is used (Pearl 1988, Verma and Pearl 1990): If X, Y, and Z are three disjoint subsets of nodes in a directed acyclic graph (DAG), then Z is said to *d-separate* X from Y, iff there is no path, i.e. no sequence of consecutive edges (of any directionality), from a node in X to a node in Y along which the following two conditions hold:

1. every node with converging edges either is in Z or has a descendant in Z,
2. every other node is not in Z.

With the described notions of separation, we can define the so-called *Markov properties* of graphs (Whittaker 1990):

pairwise: Attributes, whose nodes are non-adjacent in the graph, are independent conditional on all remaining attributes.

local: Conditional only on the attributes corresponding to the adjacent nodes, an attribute is independent of all remaining attributes (see also Frydenberg 1990).

global: Any two subsets of attributes, whose corresponding node sets are separated by a third node set, are independent conditionally only on the attributes corresponding to the nodes in the third set (see also Lauritzen *et al* 1990).

Note that the local Markov property is contained in the global, and the pairwise Markov property in the local.

Since the pairwise Markov property refers to the independence of only two attributes, it would be most natural (at least for undirected graphs) to use it to define an independence graph: If two attributes are dependent given all other attributes, there is an edge between their corresponding nodes, otherwise there is no edge (Whittaker 1990). But, unfortunately, the three types of Markov properties are not equivalent in general, and it is obvious that we need the *global* Markov property for inferences from multiple observations. However, the above definition can be used, if — in addition to the semi-graphoid axioms — the following axiom holds:

intersection: $(W \perp\!\!\!\perp Y \mid Z \cup X) \wedge (X \perp\!\!\!\perp Y \mid Z \cup W) \Longrightarrow (W \cup X \perp\!\!\!\perp Y \mid Z)$

The semi-graphoid axioms together with this one are called the *graphoid axioms*. If they hold for a given notion of conditional independence, an independence graph can be defined via the pairwise Markov condition, since the intersection axiom allows us to infer the global Markov property from the pairwise. If the intersection axiom does not hold, the global Markov property has to be used to define an independence graph.

It is obvious that an independence graph for a given domain is easy to find. For example, the complete undirected graph, i.e. the graph in which every node is connected directly to every other, always is an independence graph. But using a complete graph would not reduce the amount of data to be stored (see below). Therefore, in graphical modeling, we have to add the condition that the independence graph has to be *sparse* or even *minimal*, i.e. should contain as few edges as possible.

It should be noted that directed acyclic graphs and undirected graphs represent conditional independence relations in fundamentally different ways. In particular, there are undirected graphs that represent a conditional independence that cannot be represented by a single directed acyclic graph, and vice versa.

The quantitative part of a graphical model. The independence graph is also called the *qualitative* part of a graphical model, since it specifies which attributes are dependent and which are independent, but not the details of the dependences. How the latter information, which is called the *quantitative* part of a graphical model, is described, depends again on the type of the graph. In a directed acyclic graph, it is represented as a set of conditional distributions: one for each attribute conditional on all of its parents in the graph. If an attribute does not have any parents, its associated distribution simplifies to an unconditional distribution.

For an undirected graph, the quantitative part is represented as a set of marginal distributions: one for each maximal clique of the independence graph, where a maximal clique is a fully connected subgraph that is not contained in another fully connected subgraph. Because of this representation an undirected *hyper*graph is often used instead of a normal undirected graph. The nodes of each maximal clique of the normal graph are then connected by one *hyper*edge in the hypergraph. Unfortunately, this approach suffers from the fact that the resulting hypergraph can have cycles. This causes problems, because during an inference process the same information can travel along more than one path and thus may be used several times to update the knowledge about an attribute. If the inference mechanism is not idempotent, i.e. if a second incorporation of already included information changes the result, this can invalidate the conclusions drawn.

In order to avoid these problems, the discussion is usually restricted to *triangulated* undirected graphs, i.e. to graphs in which each cycle of length four or larger contains a *chord*, where a chord is an edge between two non-consecutive nodes in the cycle. It can be shown that the maximal clique hypergraph of a triangulated undirected graph is always a *hypertree*, i.e. does not contain any cycles. In addition, this type of graphs is important, because it can be shown that a triangulated undirected graph is isomorphic to a directed acyclic graph. Thus, with the restriction to triangulated graphs, the difference between directed and undirected graphs in their capability to represent conditional independences is removed.

It is worth noting that especially the representation using undirected graphs suggests to view graphical modeling as a decomposition method: The (global) distribution $\mathcal{D}$ is decomposed into a set of (local) distributions $\{\mathcal{D}_{X_1}, \ldots, \mathcal{D}_{X_n}\}$ on subspaces, which are the cross-products of the domains of the attributes in a maximal clique. Because of this decomposition, global reasoning, i.e. drawing inferences using $\mathcal{D}$, can be replaced by local reasoning, which involves only the distributions $\mathcal{D}_{X_k}$.

Reasoning in graphical models. The reasoning process, which we describe here exemplary for an undirected graph, basically is this: Information obtained e.g. by observations about the values of an attribute is extended to the distributions on all hyperedges containing the attribute and then projected to the intersections of these hyperedges with other hyperedges. From there it is extended and projected again etc. until the information is distributed to all attributes.

A general local propagation algorithm for hypertrees has been developed for so-called *valuation-based systems (VBS)* (Shafer and Shenoy 1988, Shenoy and Shafer 1990). The axiomatic framework of a VBS (Shenoy 1989, Shenoy 1992a) can represent various uncertainty calculi such as probability theory, Dempster-Shafer theory, and possibility theory. Conditional independence in VBSs has been defined in (Shenoy 1991) and shown to satisfy the graphoid axioms in (Shenoy 1992b). The general algorithm, which we cannot describe in detail here, has been implemented in the PULCINELLA system (Saffiotti and Umkehrer 1991).

From this general discussion it follows that the main tasks of the following sections are to state the possible values d_ω of the distribution to represent by a graphical model and to explain their interpretation, to define the notion of a conditional distribution and using it the notion of conditional independence, and finally to give an outline of the mechanisms with which inferences can be drawn.

Learning graphical models from data. In addition, we make some remarks concerning methods for the automatic induction of graphical models from data. Concerning such methods, problems arise from the fact that various kinds of prior information can be available, expert knowledge as

well as a database of sample cases, both of which should be considered in a unified framework. However, we restrict ourselves to a purely data-oriented approach, i.e. we assume only a database of observations to be given.

Since constructive methods are rarely available, data oriented learning methods nearly always consist of two parts: a search method and an evaluation measure. The evaluation measure estimates the quality of a given (hyper)graph and the search method determines which (hyper)graphs are inspected. Often the search is guided by the value of the evaluation measure, since it is usually the goal to maximize (or to minimize) its value. Commonly used search methods include optimum weight spanning tree construction (Chow and Liu 1968, Dechter 1990) (for undirected graphs) and greedy parent selection (Cooper and Herskovits 1992) (for directed graphs). Evaluation measures depend on the underlying uncertainty calculus and are considered in the corresponding sections.

F1.2.4 Relational graphical models

In relational graphical models or *relational networks* the distribution $\mathcal{D}$, which specifies the generic knowledge about the domain under consideration, is the indicator function R of a relation. This function maps elements of Ω to the set $\{0,1\}$, i.e. $R : \Omega \to \{0,1\}$. We have $R(\omega) = 1$, if the combination of attribute values in ω is considered to be possible (i.e. if ω is contained in the relation), and $R(\omega) = 0$ otherwise.

A *conditional relation* is also represented by an indicator function. For two disjoint subsets X and Y of attributes the conditional relation over X given Y can be defined as follows:

$$R(\omega_X \mid \omega_Y) = \begin{cases} 1, & \text{if } \exists \omega \in \Omega : \text{proj}_X^V(\omega) = \omega_X \wedge \text{proj}_Y^V(\omega) = \omega_Y \wedge R(\omega) = 1, \\ 0, & \text{otherwise.} \end{cases}$$

With this definition established, it is easy to define *relational conditional independence*: Let X, Y, and Z be disjoint subsets of V. Then X is called *conditionally independent* of Y given Z w.r.t. R, abbreviated $X \perp\!\!\!\perp_R Y \mid Z$, iff

$$\forall \omega \in \Omega : \quad R(\omega_{X \cup Y} \mid \omega_Z) = \min\{R(\omega_X \mid \omega_Z), R(\omega_Y \mid \omega_Z)\} \tag{F1.2.1}$$

Since $R(\cdot \mid \cdot) \in \{0,1\}$, it is also possible to use the product instead of the minimum.

Reasoning in relational networks. We restrict our outline of the inference mechanism in relational networks to undirected independence graphs. In this case to each maximal clique with underlying attribute set X a marginal relation with indicator function

$$R_X(\omega_X) = \begin{cases} 1, & \text{if } \exists \omega \in \Omega : \text{proj}_X^V(\omega) = \omega_X \wedge R(\omega) = 1 \\ 0, & \text{otherwise} \end{cases}$$

is assigned. Inferences are drawn by carrying out sequences of extension and projection operations on these marginal distributions. For example, if the actual values of a subset $Y \subset X$ of attributes are observed, this information can be represented as a (one tuple) relation R_Y. This relation is used to restrict the marginal relation R_X by removing from it all tuples ω_X, for which $R_Y(\text{proj}_Y^X(\omega_X)) = 0$. From the obtained restricted relation R'_X the possible values of the attributes of a subset $Z \subset X$ can then be inferred by projecting the tuples in R_X to Z, yielding a relation R_Z. If $Z \subset W$, where W is the underlying set of attributes of another maximal clique of the independence graph, the constraints obtained for the possible values of the attributes in Z can be propagated by the same steps to other attributes not contained in X (Kruse and Schwecke 1990, Kruse *et al* 1994).

Since the relational inference mechanism is idempotent, cycles in the hypergraph corresponding to a given undirected independence graph can be tolerated. Nevertheless, cycles should be avoided, because they can increase the running time of the inference process, since it may be that they have to be traversed several times to yield the conclusion (Kruse *et al* 1994).

Learning relational networks from data. It can be shown that the natural join of all the maximal clique relations of an independence graph G is the original relation R. Hence relational graphical models represent a *lossless join decomposition* (Maier 1983, Ullman 1988) of the underlying relation R, i.e.

$$R(\omega) = \min_{X \in \text{cliques}(G)} R_X(\omega_X), \tag{F1.2.2}$$

where cliques(G) is the set of all maximal cliques, each of which is represented by the set of attributes whose corresponding nodes are contained in it. Hence the quality of a given (hyper)graph can always be assessed by testing whether the natural join of its associated marginal relations yields the original relation, and if it does not, by computing the number of additional tuples the join contains (Dechter 1990).

An obstacle for such an induction method is its complexity: If we are given a hypergraph H and a relation R, then only in cases where H is tractable (for instance, where H is a hypertree) one can decide in reasonable time whether H yields a lossless join decomposition of R. To select from a *class* of hypergraphs a hypergraph yielding a lossless join decomposition of a given relation, turns out to be an even harder problem, which is presumably intractable even in cases where each individual member of the class is tractable (Dechter and Pearl 1992).

Hence, for efficiency, heuristic approaches have to be tolerated. For example, the marginal distributions used by the hypergraph can be evaluated separately by computing the number of occurring tuples in the subspace relative to the size of the subspace. This evaluation measure is plausible, because it describes the fraction of Ω still considered possible by the information contained in the marginal relation. Since the natural join of the marginal relations intersects these fractions, and since a natural join of marginal projections can produce only additional tuples, a learning algorithm should strive at minimizing this measure. Optimum weight spanning tree construction can be used as a search method (Dechter 1990).

F1.2.5 Probabilistic graphical models

In purely probabilistic approaches quantitative knowledge about the dependencies between the attributes in V is described by a probability distribution P on Ω. $P(\omega) = p \in [0, 1]$ means that the combination of attribute values in ω has the probability p.

A conditional probability distribution is defined in the usual way, i.e. as

$$P(\omega_X \mid \omega_Y) = \frac{P(\omega_{X \cup Y})}{P(\omega_Y)}.$$

Conditional independence is defined in accordance with the usual notion of stochastic independence as follows: Let X, Y, and Z be three disjoint subsets of attributes in V. X is called *conditionally independent* of Y given Z w.r.t. P, abbreviated $X \perp\!\!\!\perp_P Y \mid Z$, iff

$$\forall \omega \in \Omega: \quad P(\omega_{X \cup Y} \mid \omega_Z) = P(\omega_X \mid \omega_Z) \cdot P(\omega_Y \mid \omega_Z) \tag{F1.2.3}$$

whenever $P(\omega_Z) > 0$.

There is a large variety of probabilistic graphical models based on this definition, for example influence diagrams used to represent decision processes (Smith 1989, Shachter 1990, Heckerman 1991), *Bayesian networks* (Pearl 1986, Pearl 1988) and *Markov networks* (Lauritzen and Spiegelhalter 1988, Pearl 1988). We restrict our discussion to the latter two, which are an advanced and widely discussed framework for knowledge representation and propagation in probabilistic expert systems.

Bayesian networks. The most popular kind of probabilistic graphical models in artificial intelligence is the *Bayesian network*, also called *belief network* (Pearl 1986, Pearl 1988). A Bayesian network consists of a directed acyclic graph and a set of conditional probability distributions $P(\omega_A \mid \omega_{\text{parents}(A)})$, $A \in V$, where parents(A) is the set of attributes corresponding to the parents of the node corresponding to attribute A.

A Bayesian network specifies a decomposition of the joint probability distribution P on Ω into a set of conditional probability distributions: A strictly positive probability distribution P on Ω *factorizes* w.r.t. a directed acyclic graph, if

$$P(\omega) = \prod_{A \in V} P(\omega_A \mid \omega_{\text{parents}(A)}). \tag{F1.2.4}$$

In this case P satisfies the *global Markov property* (cf. section F1.2.3). It follows, that a Bayesian network can also be seen as a graphical representation of a Markov chain.

Since a Bayesian network is a directed graph, it is well-suited to represent direct causal dependencies between variables. In many cases this is quite natural for knowledge representation, e.g. in expert systems designed for diagnostic reasoning (abductive inference) in medical applications.

Markov networks. An alternative type of probabilistic graphical models uses undirected graphs and is called a *Markov network* (Pearl 1988, Lauritzen and Spiegelhalter 1988). It represents Markov random fields, which are used, for instance, in imaging and spatial reasoning (Besag *et al* 1991) and in stochastic models for neural networks (Hertz *et al* 1991).

Similar to Bayesian networks it describes a decomposition of the joint probability distribution P on Ω, but it uses a *potential representation*: A strictly positive probability distribution P on Ω *factorizes* w.r.t. an undirected graph, if

$$\forall X \in \text{cliques}(G) : \exists \phi_X : \quad P(\omega) = \prod_{X \in \text{cliques}(G)} \phi_X(\omega_X), \tag{F1.2.5}$$

where cliques(G) is the set of all maximal cliques, each of which is represented by the set of attributes whose corresponding nodes are contained in it. The ϕ_X are strictly positive functions defined on Ω_X, $X \subseteq V$.

Reasoning in probabilistic networks. Oriented at the way the human mind reasons, (Pearl 1986) developed a local propagation algorithm that works in singly connected Bayesian networks, the details of which we cannot present here. (Lauritzen and Spiegelhalter 1988) approached the same problem from a purely mathematical point of view. Their method consists in transforming a given directed acyclic graph into a triangulated undirected graph and creating from it a tree whose vertices are the cliques of this triangulated graph. To propagate evidence, probabilities in the original Bayesian network are updated by message passing between the vertices of this clique tree. Again we cannot describe the method in detail. Flexible software tools using this method are, for example, HUGIN (Andersen *et al* 1989, Jensen and Liang 1994) and BAIES (Cowell 1992).

A technique for local computations in hypertrees, which refers to the more general framework of valuation-based systems (Shenoy 1989) (see section F1.2.3), was proposed in (Shenoy and Shafer 1990, Shafer and Shenoy 1988) and is implemented in PULCINELLA (Saffiotti and Umkehrer 1991).

Learning probabilistic networks from data. Because of the additional task of finding the probabilities in the conditional or marginal distributions needed in the network, we have to distinguish between quantitative and qualitative network induction — in contrast to learning relational networks, where we only needed to consider qualitative network induction.

Quantitative network induction for a given network structure consists in estimating the joint probability distribution P, where P is selected from a family of parameterized probability distributions. A lot of approaches have been developed in this field, using methods such as maximum likelihood, maximum penalized likelihood, or fully Bayesian approaches, which involve different computational techniques of probabilistic inference such as the expectation maximization (EM) algorithm, Gibbs sampling, Laplace approximation, and Monte Carlo methods. For an overview, see (Buntine 1994, Spiegelhalter *et al* 1993).

Qualitative network induction consists in learning a network structure from a database of sample cases. In principle one could use the factorization property of a probabilistic network to evaluate its quality by comparing for each $\omega \in \Omega$ the probability computed from the network with the relative frequency found in the database to learn from. But just as for relational networks this approach is usually much too costly.

Other methods are based on linearity and normality assumptions (Pearl and Wermuth 1993), rely on the extensive testing of conditional independences (CI tests) (Verma and Pearl 1992), or use a Bayesian approach (Cooper and Herskovits 1992, Lauritzen *et al* 1993). Unfortunately, the first group is fairly restrictive, CI tests tend to be unreliable unless the volume of data is enormous, and with an increasing number of vertices they soon become computationally intractable. Bayesian learning requires debatable prior assumptions (for example, default uniform priors on distributions, uniform priors on the possible graphs) and also tends to be inefficient unless greedy search methods are used.

Nevertheless, several network induction algorithms have successfully been applied. The oldest example is an algorithm to decompose a multi-variate probability distribution into a tree of two-

dimensional distributions (Chow and Liu 1968). It uses mutual information as the evaluation measure and optimum weight spanning tree construction as the search method.

Another example is the $K2$ algorithm (Cooper and Herskovits 1992), which uses a greedy parent search and a Bayesian evaluation measure. Its drawback, which consists in the fact that it needs a topological order of the attributes, can be overcome by a hybrid algorithm (Singh and Valtorta 1993), which combines CI tests (to find a topological order) and $K2$ (to construct the Bayesian network with respect to this topological order). Unfortunately, $K2$ can deal only with complete and precise data. The treatment of missing values and hidden variables is clear only from a theoretical point of view (Cooper and Herskovits 1992).

A third algorithm, which uses a backward search strategy, is described in (Højsgaard and Thiesson 1994). Several evaluation measures, which can be used with optimum weight spanning tree construction and greedy parent search as well as other search methods, are surveyed in (Borgelt and Kruse 1997).

F1.2.6 Possibilistic graphical models

Possibility distributions. A *possibility distribution* π on a universe of discourse Ω is a mapping from Ω into the unit interval, i.e. $\pi : \Omega \to [0,1]$ (Zadeh 1978). From an intuitive point of view, $\pi(\omega)$ quantifies the degree of possibility that $\omega = \omega_0$ is true, where ω_0 is the actual state of the world (cf. section F1.2.2): $\pi(\omega) = 0$ means that $\omega = \omega_0$ is impossible, $\pi(\omega) = 1$ means that $\omega = \omega_0$ is possible without any restrictions, and $\pi(\omega) \in (0,1)$ means that $\omega = \omega_0$ is possible only with restrictions, i.e. that there is evidence that supports $\omega = \omega_0$ as well as evidence that contradicts $\omega = \omega_0$.

Several suggestions have been made for semantics of a *theory of possibility* as a framework for reasoning with uncertain and imprecise data. Among the numerical approaches, we like to mention possibility distributions as epistemic interpretations of fuzzy sets (Zadeh 1978), the axiomatic view of possibility theory based on the concept of a possibility measure (Dubois and Prade 1988), Spohn's theory of epistemic states (Spohn 1990), possibility distributions as one-point coverages of random sets (Nguyen 1978, Hestir *et al* 1991), contour functions of consonant belief functions (Shafer 1976), falling shadows in set-valued statistics (Wang 1983), and possibility theory based on likelihoods (Dubois *et al* 1993).

The view of a possibility distribution as an *information-compressed* representation of uncertain *and* imprecise knowledge about a state ω_0 of the world can be clarified in a random set framework that generalizes traditional approaches like (Strassen 1964, Dempster 1967, Kampé de Fériet 1982). Let $(C, 2^C, P)$, $C = \{c_1, c_2, \ldots, c_m\}$, be a finite probability space and $\gamma : C \to 2^\Omega$ a set-valued mapping. C is seen as a set of contexts that have to be distinguished for a set-valued specification of ω_0. The contexts are supposed to describe different physical and observation-related frame conditions. $P(\{c\})$ is the (subjective) probability of the (occurence or selection of the) context c.

A set $\gamma(c)$ is assumed to be the *most specific correct set-valued specification* of ω_0, which is implied by the frame conditions that characterize the context c. By 'most specific set-valued specification' we mean that $\omega_0 \in \gamma(c)$ is guaranteed to be true for $\gamma(c)$, but is not guaranteed for any proper subset of $\gamma(c)$. The resulting *random set* $\Gamma = (\gamma, P)$ is an imperfect (i.e. imprecise *and* uncertain) specification of ω_0. Let π_Γ denote the *one-point coverage of* Γ (the *possibility distribution induced by* Γ*)*, which is defined as

$$\pi_\Gamma : \Omega \to [0,1], \qquad \pi_\Gamma(\omega) = P\left(\{c \in C \mid \omega \in \gamma(c)\}\right).$$

In a complete modeling, the contexts in C must be specified in detail, so that the relationships between all contexts c_j and their corresponding specifications $\gamma(c_j)$ are made explicit. But if the contexts are unknown or ignored, then $\pi_\Gamma(\omega)$ is the total mass of all contexts c that provide a specification $\gamma(c)$ in which ω_0 is contained, and this quantifies the *possibility of truth* of the statement "$\omega = \omega_0$" (Gebhardt and Kruse 1993b, Gebhardt and Kruse 1996a).

That in this interpretation a possibility distribution represents uncertain *and* imprecise knowledge can be understood best by comparing it to a probability distribution and to a relation. A probability distribution covers *uncertain*, but *precise* knowledge. This becomes obvious, if one notices that a possibility distribution in the interpretation described above reduces to a probability distribution, if $\forall c_j \in C : |\gamma(c_j)| = 1$, i.e. if for all contexts the specification of ω_0 is precise. On the other hand, a relation represents *imprecise*, but *certain* knowledge about dependencies between

attributes. Thus, not surprisingly, a relation can also be seen as a special case of a possibility distribution in the interpretation given above, namely if there is only one context. Hence the context-dependent specifications are responsible for the imprecision, the contexts for the uncertainty in the imperfect knowledge expressed by a possibility distribution.

From this distinction it is clear, that the uncertainty part can be excluded from the (information-compressed) imperfect knowledge represented by π_Γ by assuming *α-correctness* of Γ w.r.t. ω_0. α-correctness means that there exists a subset $C' \subseteq C$ of contexts such that $P(C') \geq \alpha$ and $\forall c' \in C' : \omega_0 \in \gamma(c')$. In this case the *$\alpha$-cut* $[\pi_\Gamma]_\alpha = \{\omega \mid \pi_\Gamma(\omega) \geq \alpha\}$ of the possibility distribution π_Γ turns out to be the most specific correct set-valued specification of ω_0 (Gebhardt and Kruse 1993a).

Operations on possibility distributions can also be performed within a pure random set background (Kampé de Fériet 1982, Hestir *et al* 1991), but this has the disadvantage that it is not in accordance with the *extension principle* (Zadeh 1975). From a semantical point of view this principle has been claimed to be the adequate way of generalizing operations from crisp or imprecise (multi-valued) data to the possibilistic setting (Dubois and Prade 1991, Kruse *et al* 1994). Therefore it seems to be more appropriate to base the interpretation of possibility degrees on the above mentioned concepts of α-correctness and maximum specificity. For an detailed discussion of this view of possibility theory, we refer to (Gebhardt and Kruse 1993a, Gebhardt and Kruse 1995). These papers justify the extension principle as a theorem in the underlying formal and semantical framework. Special aspects of possibility measures for decision making are considered in (Gebhardt and Kruse 1994).

Possibilistic networks. Although well-known for a couple of years (Hisdal 1978), a unique concept of possibilistic independence has not been fixed yet. For some recent discussions, see (Farinas del Cerro and Herzig 1994, Fonck 1994). In our opinion, the main problem is that possibility theory is a calculus for uncertain *and* imprecise reasoning, the first of which is related to probability theory, the latter to relational theory (see above). But due to their reference to different types of imperfect knowledge, relational and probability theory lead to different concepts of independence, namely lossless join decomposability and stochastic independence, respectively. Stochastic independence is an *uncertainty-based* type of independence, whereas lossless join decomposability is an *imprecision-based* type of independence. Since possibility theory addresses both kinds of imperfect knowledge, notions of possibilistic independence can be uncertainty-based or imprecision-based. Hence there are at least two ways of introducing and justifying them.

A more fundamental justification of the distinction between imprecision and uncertainty starts from considering two levels of reasoning, namely the *credal level*, on which all operations on pieces of knowledge take place, and the *pignistic level*, on which the final step of decision making follows (Smets and Kennes 1994). Imprecision-based possibilistic independence is strongly oriented at the credal level, applying the extension principle as the basic concept of operating on possibility distributions. On this level normalization of possibility distributions is avoided, since it would change absolute to relative degrees of possibility. In contrast to this, an uncertainty-based approach to possibilistic independence refers to the pignistic level. Here decision making aspects are taken into account and thus relative degrees of possibility of events are used. On this level, the need for normalization is obvious.

With respect to this consideration in (de Campos *et al* 1995) two definitions of possibilistic independence have been justified, namely uncertainty-based possibilistic independence, which is derived from *Dempster's rule of conditioning* (Shafer 1976) adapted to possibility measures, and imprecision-based possibilistic independence, which coincides with the well-known concept of *possibilistic non-interactivity* (Dubois and Prade 1988). The latter can be seen as a generalization of lossless join decomposability to the possibilistic setting, since it treats each α-cut of a possibility distribution like a relation.

Because of its consistency with the extension principle, we confine to possibilistic non-interactivity. As a concept of possibilistic independence it can be defined as follows: Let X, Y, and Z be three disjoint subsets of variables in V. Then X is called *conditionally independent* of Y given Z w.r.t. π, abbreviated $X \perp\!\!\!\perp_\pi Y \mid Z$, iff

$$\forall \omega \in \Omega : \quad \pi(\omega_{X \cup Y} \mid \omega_Z) = \min\{\pi(\omega_X \mid \omega_Z), \pi(\omega_Y \mid \omega_Z)\} \tag{F1.2.6}$$

whenever $\pi(\omega_Z) > 0$, where $\pi(\cdot \mid \cdot)$ is a non-normalized conditional possibility distribution, i.e.

$$\pi(\omega_X \mid \omega_Z) = \max\{\pi(\omega') \mid \omega' \in \Omega \wedge \text{proj}_X^V(\omega) = \omega_X \wedge \text{proj}_Z^V(\omega) = \omega_Z\}.$$

Both mentioned types of possibilistic independence satisfy the *semi-graphoid axioms* (see section F1.2.3). Possibilistic independence based on Dempster's rule in addition satisfies the intersection axiom (Fonck 1994). But note that the intersection axiom is related to uncertainty-based independence. Relational independence does not satisfy this axiom, and therefore it cannot be satisfied by possibilistic non-interactivity as a more general type of imprecision-based independence.

Similar to probabilistic networks, a possibilistic network can be seen as a decomposition of a multi-variate possibility distribution. The factorization formulae can be derived from the corresponding probabilistic factorization formulae (for Markov networks) by replacing the product by the minimum.

Reasoning in possibilistic networks In section F1.2.3 we already mentioned valuation-based systems (VBS), which can handle imperfect knowledge w.r.t. various uncertainty calculi, possibility theory among them. Possibilistic independence in VBSs corresponds to uncertainty-based independence based on Dempster's rule. Hence, is this type of conditional independence is chosen, the VBSs propagation algorithm (Shafer and Shenoy 1988, Shenoy and Shafer 1990) can be used to draw inferences.

But if possibilistic non-interactivity is chosen as the notion of conditional independence in order to be consistent with the extension principle, the VBS approach has to be slightly modified, since no normalization takes place. The related local propagation algorithms for hypertree structures, which we cannot describe in detail here, are presented in (Kruse *et al* 1994) and have been implemented in the tool POSSINFER (Kruse *et al* 1994, Gebhardt and Kruse 1996a). They are closely related to the projection-extension-mechanism described in section F1.2.4. Since this type of possibilistic inference propagation is idempotent (just like relational mechanism), cycles in the underlying hypergraph can be tolerated.

Learning possibilistic networks from data. Just as for relational and probabilistic networks, it is possible in principle to estimate the quality of a given possibilistic network by exploiting its factorization property. For each $\omega \in \Omega$ the degree of possibility computed from the network is compared to the degree of possibility derived from the database to learn from. But again this approach can be costly.

Contrary to probabilistic networks, the induction of possibilistic networks from data has been studied much less extensively. A first result, which consists in an algorithm that is closely related to the $K2$ algorithm for the induction of Bayesian networks, was presented in (Gebhardt and Kruse 1995). Instead of the Bayesian evaluation measure used in $K2$, it relies on a measure derived from the *nonspecificity* of a possibility distribution. Roughly speaking, the notion of nonspecificity plays the same role in possibility theory that the notion of *entropy* plays in probability theory. Based on the connection of the imprecision part of a possibility distribution to relations, the nonspecificity of a possibility distribution can also be seen as a generalization of *Hartley information* (Hartley 1928) to the possibilistic setting.

In (Gebhardt and Kruse 1996b) a rigid foundation of a learning algorithm for possibilistic networks is given. It starts from a comparison of the nonspecificity of a given multi-variate possibility distribution to the distribution represented by a possibilistic network, thus measuring the loss of specificity, if the multi-variate possibility distribution is represented by the network. In order to arrive at an efficient algorithm, an approximation for this loss of specificity is derived, which can be computed locally on the hyperedges of the network. As the search method a generalization of the optimum weight spanning tree algorithm to hypergraphs is used.

Several other heuristic local evaluation measures, which can be used with different search methods, are presented in (Borgelt and Kruse 1997).

It should be emphasized, that, as already discussed above, an essential advantage of possibilistic networks over probabilistic ones is their ability to deal with imprecision, i.e. multi-valued, information. When learning possibilistic networks from data, this leads to the convenient situation that missing values in an observation or a set of values for an attribute, all of which have to be considered possible, do not pose any problems.

References

Andersen S K, Olesen K G, Jensen F V, and Jensen F 1989 HUGIN — a shell for building Bayesian belief universes for expert systems *Proc. 11th International Joint Conference on Artificial Intelligence* pp 1080–1085

Besag J, York J, and Mollie A 1991 Bayesian image restauration with two applications in spatial statistics *Ann. Inst. Statist. Math.* **43**(1) 1–59

Borgelt C and Kruse R 1997 Evaluation Measures for Learning Probabilistic and Possibilistic Networks. *Proc. 6th IEEE Int. Conf. on Fuzzy Systems* (Barcelona, Spain) 669–676

Buntine W 1994 Operations for learning with graphical models *Journal of Artificial Intelligence Research* **2** 159–225

de Campos L M, Gebhardt J, and Kruse R 1995 *Syntactic and semantic approaches to possibilistic independence* Technical report, University of Granada and University of Braunschweig

Cheeseman P and Oldford R W, eds. 1994 *Selecting Models from Data (Lecture Notes in Statistics 89)* (New York, NY: Springer)

Chow C K and Liu C N 1968 Approximating Discrete Probability Distributions with Dependence Trees *IEEE Trans. on Information Theory* **14**(3) 462–467

Cooper G and Herskovits E 1992 A Bayesian method for the induction of probabilistic networks from data *Machine Learning* **9** 309–347

Cowell R 1992 BAIES — a probabilistic expert system shell with qualitative and quantitative learning *Bayesian Statistics 4* eds. J Bernardo, J Berger, A Dawid, and A Smith (Oxford, England: Oxford University Press) pp 595–600

Dawid A 1979 Conditional independence in statistical theory *SIAM Journal on Computing* **41** 1–31

Dechter R 1990 Decomposing a relation into a tree of binary relations *Journal of Computer and Systems Sciences* **41** 2–24

Dechter R and Pearl J 1992 Structure identification in relational data *Artificial Intelligence* **58** 237–270

Dempster A P 1967 Upper and lower probabilities induced by a multivalued mapping *Ann. Math. Stat.* **38** 325–339

Dubois D and Prade H 1988 *Possibility Theory* (New York, NY: Plenum Press)

Dubois D and Prade H 1991 Fuzzy sets in approximate reasoning, part 1: inference with possibility distributions *Fuzzy Sets and Systems* **40** 143–202

Dubois D, Moral S, and Prade H 1993 *A semantics for possibility theory based on likelihoods* Annual report, CEC–ESPRIT III BRA 6156 DRUMS II

Farinas del Cerro L and Herzig A 1994 Possibility theory and independence *Proc. of the Fifth IPMU Conference* pp 820–825

Fonck P 1994 Conditional independence in possibility theory *Proc. 10th Conf. on Uncertainty in Artificial Intelligence* eds. R López de Mántaras and D Poole (San Mateo, CA: Morgan Kaufmann) pp 221–226

Frydenberg M 1990 The chain graph Markov property *Scandinavian Journal of Statistics* **17** 333–353

Gebhardt J and Kruse R 1993a A new approach to semantic aspects of possibilistic reasoning *Symbolic and Quantitative Approaches to Reasoning and Uncertainty, Lecture Notes in Computer Science 747* eds. M Clarke, R Kruse, and S Moral (Berlin, Germany: Springer) pp 151–160

Gebhardt J and Kruse R 1993b The context model — an integrating view of vagueness and uncertainty *Int. Journal of Approximate Reasoning* **9** 283–314

Gebhardt J and Kruse R 1994 On an information compression view of possibility theory *Proc. 3rd IEEE Int. Conf. on Fuzzy Systems* (Orlando) pp 1285–1288

Gebhardt J and Kruse R 1995 Learning possibilistic networks from data *Proc. 5th Int. Workshop on Artificial Intelligence and Statistics* (Fort Lauderdale, FL) pp 233–244

Gebhardt J and Kruse R 1996 POSSINFER — A Software Tool for Possibilistic Inference *Fuzzy Set Methods in Information Engineering: A Guided Tour of Applications* eds. D Dubois, H Prade, and R Yager (New York, NY: Wiley) 407–418

Gebhardt J and Kruse R 1996 Tightest Hypertree Decompositions of Multivariate Possibility Distributions *Proc. Int. Conf. on Information Processing and Management of Uncertainty in Knowledge-based Systems (IPMU'96)* (Granada, Spain) 923–927

Gebhardt J and Kruse R 1997 Parallel Combination of Information Sources *Handbook of Defeasible Reasoning and Uncertainty Management Systems, Vol.1: Updating Uncertain Information* eds. D Gabbay and P Smets (Dordrecht, Netherlands: Kluwer)

Hartley R V L 1928 Transmission of information *The Bell Systems Technical Journal* **7** 535–563

Heckerman D 1991 *Probabilistic Similarity Networks* (Cambridge, MA: MIT Press)

Hertz J, Krogh A, and Palmer R 1991 *Introduction to the Theory of Neural Computation* (Reading, MA: Addison Wesley)

Hestir K, Nguyen H T, and Rogers G S 1991 A random set formalism for evidential reasoning *Conditional Logic in Expert Systems* eds. I R Goodman, M M Gupta, H T Nguyen, and G S Rogers (Amsterdam, Netherlands: North Holland) pp 209–344

Hisdal E 1978 Conditional possibilities, independence, and noninteraction *Fuzzy Sets and Systems* **1** 283–297

Højsgaard S and Thiesson B 1994 BIFROST — block rekursive models induced from relevant knowledge, observations, and statistical techniques *Computational Statistics and Data Analysis*

Jensen F V and Liang J 1994 drHUGIN — a system for value of information in Bayesian networks *Proc. 5th Int. Conf. on Information Processing and Management of Uncertainty in knowledge-based Systems*

Kampé de Fériet J 1982 Interpretation of membership functions of fuzzy sets in terms of plausibility and belief *Fuzzy Information and Decision Processes* eds. M M Gupta and E Sanchez (Amsterdam, Netherlands: North Holland) pp 13–98

Kruse R and Schwecke E 1990 Fuzzy Reasoning in a Multidimensional Space of Hypotheses. *Int. Journal of Approximate Reasoning* **4** 47–68

Kruse R, Gebhardt J, and Klawonn F 1994 *Foundations of Fuzzy Systems* (Chichester, England: Wiley) Translation of the book: *Fuzzy Systeme (Series: Leitfäden und Monographien der Informatik)* (Stuttgart, Germany: Teubner)

Lauritzen S L and Spiegelhalter D J 1988 Local computations with probabilities on graphical structures and their application to expert systems *Journal of the Royal Stat. Soc., Series B* **2**(50) 157–224

Lauritzen S, Dawid A, Larsen B, and Leimer H G 1990 Independence properties of directed markov fields *Networks* **20** 491–505

Lauritzen S L, Thiesson B, and Spiegelhalter D 1993 Diagnostic systems created by model selection methods — a case study. *Proc. 4th Int. Workshop on Artificial Intelligence and Statistics* (Fort Lauderdale, FL) pp 93–105

Maier D 1983 *The Theory of Relational Databases* (Rockville, MD: Computer Science Press)

Nguyen H T 1978 On random sets and belief functions *Journal of Mathematical Analysis and Applications* **65** 531–542

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

Pearl J and Paz A 1987 Graphoids: a graph based logic for reasoning about relevance relations *Advances in Artificial Intelligence 2* eds. B D Boulay et al (Amsterdam, Netherlands: North Holland) pp 357–363

Pearl J 1988 *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference (2nd edition)* (San Mateo, CA: Morgan Kaufmann)

Pearl J and Wermuth N 1993 When can association graphs admit a causal interpretation? *Proc. 4th Int. Workshop on Artificial Intelligence and Statistics* (Fort Lauderdale, FL) pp 141–150

Saffiotti A and Umkehrer E 1991 PULCINELLA: a general tool for propagating uncertainty in valuation networks *Proc. 7th Conf. on Uncertainty in Artificial Intelligence* eds. B D'Ambrosio, P Smets, and P P Bonisonne (San Mateo, CA: Morgan Kaufmann) pp 323–331

Shachter R 1990 An ordered examination of influence diagrams *Networks* **20** 535–563

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

Shafer G and Shenoy P P 1988 *Local computation in hypertrees* Working paper 201 (Lawrence, KS: School of Business, University of Kansas)

Shenoy P P 1989 A valuation–based language for expert systems *Int. Journal of Approximate Reasoning* **3** 383–411

Shenoy P P and Shafer G R 1990 Axioms for probability and belief-function propagation *Uncertainty in Artificial Intelligence (4)* eds. R D Shachter, T S Levitt, L N Kanal, and J F Lemmer (Amsterdam, Netherlands: North Holland) pp 169–198

Shenoy P P 1991 *Conditional independence in valuation-based systems* Working Paper 236 (Lawrence, KS: School of Business, University of Kansas)

Shenoy P P 1992a Valuation-based systems: a framework for managing uncertainty in expert systems *Fuzzy Logic for the Management of Uncertainty* eds. L A Zadeh and J Kacprzyk (New York, NY: Wiley) pp 83–104

Shenoy P P 1992b Conditional independence in uncertainty theories *Proc. 8th Conf. on Uncertainty in Artificial Intelligence* eds. D Dubois, M P Wellman, B D'Ambrosio, and P Smets (San Mateo, CA: Morgan Kaufmann) pp 284–291

Singh M and Valtorta M 1993 An algorithm for the construction of Bayesian network structures from data *Proc. 9th Conf. on Uncertainty in Artificial Intelligence* (Washington) pp 259–265

Smets P and Kennes R 1994 The transferable belief model *Artificial Intelligence* **66** 191–234

Smith J Q 1989 Influence diagrams for statistical modeling *Annals of Statistics* **17**(2) 654–672

Spiegelhalter D, Dawid A, Lauritzen S, and Cowell R 1993 Bayesian analysis in expert systems *Statistical Science* **8**(3) 219–283

Spirtes P, Glymour C, and Scheines R 1993 *Causation, Prediction, and Search* (Lecture Notes in Statistics 81) (New York, NY: Springer)

Spohn W 1980 Stochastic independence, causal independence, and shieldability *Journal of Philosophical Logic* **9** 73–99

Spohn W 1990 A general non-probabilistic theory of inductive reasoning *Uncertainty in Artificial Intelligence* eds. R D Shachter, T S Levitt, L N Kanal, and J F Lemmer (Amsterdam, Netherlands: North Holland) pp 149–158

Strassen V 1964 Meßfehler und Information *Zeitschrift Wahrscheinlichkeitstheorie und verwandte Gebiete* **2** 273–305

Ullman J D 1988 *Principles of Database and Knowledge-Base Systems*, Volume 1 (Rockville, Maryland: Computer Science Press Inc)

Verma T S and Pearl J 1990 Causal networks: semantics and expressiveness *Uncertainty in Artificial Intelligence* eds. R D Shachter, T S Levitt, L N Kanal, and J F Lemmer (Amsterdam, Netherlands: North Holland) pp 69–76

Verma T S and Pearl J 1992 An algorithm for deciding if a set of observed independencies has a causal explanation *Proc. 8th Conf. on Uncertainty in Artificial Intelligence* pp 323–330

Wang P Z 1983 From the fuzzy statistics to the falling random subsets *Advances in Fuzzy Sets, Possibility and Applications* ed. P P Wang (New York, NY: Plenum Press) pp 81–96

Whittaker J 1990 *Graphical Models in Applied Multivariate Statistics* (Chichester, England: Wiley and Sons)

Zadeh L A 1975 The concept of a linguistic variable and its application to approximate reasoning *Information Sciences* **9** 43–80

Zadeh L A 1978 Fuzzy sets as a basis for a theory of possibility *Fuzzy Sets and Systems* **1** 3–28

