Uncertain Information Processing in Expert Systems

Petr Hájek
Tomáš Havránek
Radim Jiroušek

Uncertain
Infori ocessing
in
Expert Systems

Authors

Petr Hájek
Institute of Computer and Information Science
Prague, Czechoslovakia

Tomáš Hav[illegible]
Institute of Computer and Information Science
Prague, Czechoslovakia

Radim Jiroušek
Institute of Information Theory and Automation
Prague, Czechoslovakia

CRC Press
Boca Raton Ann Arbor London Tokyo

Library of Congress Cataloging-in-Publication Data

Uncertain information processing in expert systems/Petr Hájek,
Tomáš Havránek, Radim Jiroušek.
p. cm.
Includes bibliographical references and index.
ISBN 0-8493-6368-3
1. Expert systems (Computer science) 2. Uncertainty (Information theory) I. Hájek, Petr. II. Havránek, Tomáš, III. Jiroušek, Radim.
QA76.76.E95U52 1992
006.3'3—dc20 92-45031
CIP

Direct all inquiries to CRC Press, Inc., 2000 Corporate Blvd., N.W., Boca Raton, Florida 33431.

International Standard Book Number 0-8493-6368-3

Library of Congress Card Number 92-45031
Printed in the United States of America 1 2 3 4 5 6 7 8 9 0

Printed on acid-free paper

IN MEMORIAM

Tomáš Havránek died on May 17, 1991, at age 43. The following is the translation of Dr. Hájek's speech at his funeral.

Parting from Professor Tomáš Havránek makes us extremely sad. Paraphrasing David from the Old Testament: "I grieve over you, my brother Tomáš." We knew that he felt sick, but the end was so abrupt that the message about his death was shocking for all who knew him.

Tomáš's father was world renown linguist Bohuslav Havránek and his parents influenced him considerably. He did not continue the work of his father, instead he became a mathematical statistician. He was one of the last students of the famous professor Jaroslav Hájek, of whom he tragically reminds us: both died untimely. But Tomáš by far was not only a statistician. His tendency toward interdisciplinary thinking manifested itself while he was a student. He wrote his masters thesis about probabilistic automata, a topic on the borderline of probability theory, mathematical logic, and computer science.

He spent his whole academic career in the Czechoslovak Academy of Sciences; from 1970 to 1984 in the Center of Biomathematics of the Physiological Institute and from 1985 on in the Institute of Computer Science. Both institutions changed their names during the time, but this is unimportant. What is important is that Tomáš influenced them both greatly and was finally elected director of the latter.

Let me say some words about his scientific work. He had a remarkable ability of broad and at the same time deep inquiry and interest in finding relations among seemingly disparate parts of mathematics. In particular his interests were probability theory and mathematical statistics, mathematical logic, computability theory, graph theory, and mathematical foundations of artificial intelligence. Havránek was the main pioneer of computational statistics in Czechoslovakia.

His interest was not confined to strictly mathematical problems, his sense of concrete applications was extremely deep. At his first place of employment he was in constant contact with physicians and became an expert in biomathematics. He liked to say that he fingered many sets of data, meaning that he completed their thorough analysis. He was never a routine practitioner and his work with data led him to

problems of inference and deduction, thus to logical questions, questions of hypothesis generation and generation of probabilistic models. He published his papers in such recognized journals as *Biometrics*, *Biometrika*, *Journal of American Statistical Association*, *International Journal for Man-Machine Studies*, *Synthese*, and *Theory and Decision*. He wrote four books, some of them alone, some with co-authors. Writing books with Tomáš was an adventure, a toil, and a pleasure.

Havránek paid much attention to the organization of big and small seminars and conferences. For decades he was a person of prominence regarding our seminar of applied mathematical logic and all activities related to it, notably the GUHA-circle. He became one of the leading personalities of COMPSTAT conferences. He was a member of the organizing committee of five COMPSTATs and on the last COMPSTAT he was an invited speaker. Furthermore, he was a member of the Bernoulli Society and the International Association for Statistical Computing, repeatedly he was one of their officers.

He liked teaching students and others, e.g. physicians. In the time after the velvet revolution he had several important functions and it is admirable that he was still able to continue his scientific work.

Let me now say a few words about why I liked him, as so many other people did. He was extremely reliable and kind-hearted. He impresssed everyone by his broad knowledge. He was a very good husband, father, and friend. The sudden death of the man so near to us, the man about whom we were sure was at the pinnacle of his strength, may lead us to contemplate temporal and eternal values. The heavier the loss of Tomáš Havránek is, the more we should be grateful for the way he enriched and unforgettably influenced our lives.

PREFACE

The importance of reasoning under uncertainty for expert systems has been stressed since the beginnings of knowledge engineering (Feigenbaum, 1977; Buchanan, 1982). Two early approaches turned out to be very influential: that of MYCIN (Shortliffe, 1976; Buchanan and Shortliffe, 1984) and that of PROSPECTOR (Duda et al., 1976; Duda et al., 1978). Even if their motivations were different (MYCIN's formalism was presented as nonprobabilistic whereas PROSPECTOR's formalism had probabilistic —or at least pseudoprobabilistic—foundations) the machinery of both systems is rather similar and isomorphic in a certain sense. Both systems are *rule based*, that is, their knowledge consists of rules of the form "if (assumption) then (conclusion) with some degree of belief (weight)," and *compositional*, that is, (roughly), they combine effects of particular rules using a binary combining function to compute their joint effect. Rule-based compositional systems have become extremely popular and in fact most expert system shells still deal with uncertainty in this (rule-based, compositional) way.

On the other hand, it soon turned out that the compositional approach suffers from inherent inadequacies. This becomes clear if one tries to interpret it probabilistically, without simple-minded and unjustified assumptions. (In short, probability is not compositional.) This brings us to one of the central questions of this book. What is the real meaning of probability theory for expert systems?

It is interesting to note that the attitude of the AI community toward probability theory has changed. As Shafer remarks (Shafer, 1987b), in the early stages of development, symbolic manipulations were stressed as typical for AI programs, in contrast to "number crunching" and since probabilities are real numbers it was easy to classify probabilistic computations as number crunching. Later the focus of attention shifted to *knowledge processing*; clearly, knowledge can be both numerical as well as nonnumerical. This shift, together with the invention of *local probabilistic computations* (Pearl, 1982, 1988; Perez, 1983; Perez and Jiroušek, 1985; Lauritzen and Spiegelhalter, 1988), made the advent of purely probabilistic systems possible. Local computations mean the following: in spite of the fact that a probabilistic description of a large system of variates (propositions, etc.) is given, in general employing an unbearable amount of numbers, the knowledge of some dependence structure allows us to represent probability in a comprehensible way and to compute probabilities concerning some few variates locally, that is, from the part of the representations concerning not too many other variates. It is remarkable that graph theory can be utilized in a nontrivial and elegant manner to make this precise.

We present here three different techniques for probabilistic expert systems: Lauritzen and Spiegelhalter's method of local computations, Shachter's method of influential diagrams, and Perez's method based on the notion of simplification of the dependence structure. This gives a partial answer to the above question on the role of probability for expert systems: *probabilistic systems are possible*, that is, theoretically well founded and computationally feasible. This is not the whole answer; we may still ask if alternative approaches to uncertainty are relevant for expert systems and what their relation is to probability. This book investigates two candidates: first, we analyze the "old-fashioned" rule-based compositional systems, elucidate their structure, and determine that the methods of local computations can be used to exhibit a tricky use of compositional rule-based systems guaranteeing probabilistic consistency at least

partially and in a weak sense. Second, we introduce the reader to the beginnings of Dempster-Shafer theory and its relation to probabilistic systems (and local computations) as well as to compositional systems. Methods of fuzzy and possibilistic logic are not discussed; the reader may consult, e.g., Dubois and Prade (1991) for detailed information. We also call the reader's attention to extremely interesting work by Paris and Vencovská (1987, 1989a–c, 1990) concerning alternative approaches to a calculus of beliefs, reasons for maximal entropy and questions of computational complexity.

Three books are related to ours: those by Pearl (1988a), Neapolitano (1990), and Weichselberger and Pöhlmann (1990).* They all concentrate on the use of probability theory in expert systems; the first two present, in addition, the method of Lauritzen and Spiegelhalter. While these books discuss several topics not covered in the present book (let us mention, e.g., axiomatization of various notions of independence [Pearl, 1988]), the following are specific to ours: (1) a detailed investigation of the problem of approximation of an unknown probability distribution (as opposed to Shachter's and Lauritzen-Spiegelhalter's framework, in which the joint probability is assumed to be fully known, i.e., Perez's approach), and (2) detailed reconsiderations of compositional systems and their algebraic and probabilistic analysis (the algebraic analysis is by Hájek and Valdes, the probabilistic by Hájek). The latter should contribute to an unprejudiced evaluation of them. We shall develop a method of "guarded use" of a certain kind of compositional system and criticize the "nihilistic" treatment of compositional systems presented by Weichselberger and Pöhlmann (1990). Further related books appeared or reached us while the book was in print: Lopez de Mantaras (1990), Kruse, Schwecke, and Heimsohn (1991), Smets, Mamdani, Dubois, and Prade (eds., 1988, collection of papers), and Kruse and Siegel (eds., 1991, proceedings). All of them are recommended to the reader; but let us note that our specific topics are not treated in detail in any of them.

The reader is expected to be able to follow a mathematical text; no specific knowledge is assumed, but some acquaintaince with probability theory, graph theory, mathematical logic (propositional calculus), and group theory (or at least of some domains from this list) would be helpful. We hope that the book will be useful for expert system developers, theorists, and possibly also users; as well as for researchers interested in the relationship between inference, probability, and logic. The material presented in this book was the subject of a one-week intensive course organized by the Czechoslovak Technical Society in Alšovice, Czechoslovakia, in September 1988.

* Due to the courtesy of Professor Pearl, his book was at our disposal immediately after it appeared; the other two books are more recent and came into our hands only after detailed plans for the present book were finished.

ACKNOWLEDGMENTS

We have profited very much from our scientific contact with some of the leading personalities working in this field, including A. P. Dempster, S. L. Lauritzen, J. B. Paris, A. Perez, G. Shafer, D. J. Spiegelhalter, J. Whittaker, and others. For years (even for decades) we have enjoyed the stimulating and creative environment of the Seminar of Applied Mathematical Logic at the Mathematical Institute of the Czechoslovak Academy of Sciences; this seminar was originally primarily oriented to the GUHA method of automated hypothesis formation (see Hájek and Havránek, 1978) but increasingly has focused its main interest more and more on the foundations of expert systems.

Our thanks to William Eldridge for a detailed revision of the book from an English standpoint and to Mrs. D. Berková, Mrs. K. Falušová, and Mrs. I. Marešová for the preparation of the manuscript on computer. Also thanks to M. Daniel who caught many errors, both serious and minor. Finally we owe a great debt to our families for their support, patience, and understanding.

Prague, February 1991.

TABLE OF CONTENTS

PRELIMINARIES: BASIC MATHEMATICAL NOTIONS

In this section we present some important mathematical notions used throughout the book. Notions of particular importance are those of a *variate* and of an *array*. They help us to present logical and probabilistic notions in a unified manner.

1. VARIATES AND ARRAYS

A *variate* X consists of a *name* N and a *frame*. A name can be symbolic (p, q, $p \rightarrow q$, etc.) or verbal (color, temperature, etc.); a frame is a set, thought of as a set of possible values (of a variate). In this book we shall use almost exclusively *finite* frames.

Examples of variates are as follows:

Sex	Temperature	Color	Smoker	p
Male	Low	Red	Yes	Yes
Female	Medium	Blue	No	No
	High	White		

Variates with such frames as {yes, no}, {TRUE, FALSE}, or $\{1, 0\}$ are called *Boolean variates* (more precisely, one-dimensional Boolean variates; see below). The two values in question are called *truth values*; the two-element set of truth values is called the one-dimensional Boolean frame.

Let I be a finite set and let $(X_i)_{i \in I}$ be a system of variates with pairwise disjoint names: $X_i = (N_i, V_i)$, $N_i \neq N_j$ for $i \neq j$. We associate with $(X_i)_{i \in I}$ a new variate $\mathbf{X} = F((X_i)_{i \in I})$, the name of which is the system $(N_i)_{i \in I}$ of names and the frame of which is the cartesian product $\mathbf{V} = \prod_{i \in I} V_i$, that is, the set of all tuples $(v_i)_{i \in I}$ in which $v_i \in V_i$ for each i. The system $(X_i)_{i \in I}$ of variates together with the corresponding variate $F((X_i)_{i \in I})$ is called the *field* of variates given by $(X_i)_{i \in I}$; it is denoted $(X_i)_{i \in I}$ if there is no danger of misunderstanding. For example, if $I = \{1, 2\}$ and X_1, X_2 are the first two variates listed above, then (X_1, X_2) is the field

Sex	Temperature
Male	Low
Male	Medium
Male	High
Female	Low
Female	Medium
Female	High

If I has n elements, we say that $(X_i)_{i \in I}$ is an n-dimensional field. If all X_i are Boolean (one-dimensional) variates then $(X_i)_{i \in I}$ is an n-dimensional Boolean field; its frame consists of all n-tuples of truth values (the Boolean n-dimensional frame).

Let V be a frame; an *array* on V is just a mapping, the domain of which is V. An array A maps V into some set, for example, into real numbers, integers, into a frame, and so on. If the domain of A is the Boolean n-dimensional frame ($n = 1, 2, \ldots$) and its range is included in $\{0, 1\}$ then it is called a *Boolean array*. An example of a

two-dimensional Boolean arrays is as follows:

11	0
10	1
01	1
00	0

It is defined on the set of all pairs of 0's and 1's and its value for a pair (u, v) is 1 if and only if (iff) the pair contains exactly one 1 and exactly one 0.]

2. PROPOSITIONAL CALCULUS

Let $X_1, \ldots, X_n$ be n Boolean variates with names $p_1, \ldots, p_n$, respectively. In this section, $p_1, \ldots, p_n$ will be called *atomic formulas* (*atoms*) or *propositional variables* (*propositions*). We can construct other formulas (names of Boolean variates) using *logical connectives*: negation ($\neg$, unary), conjunction (&, binary), disjunction ($\vee$, binary), implication ($\rightarrow$, binary), and possibly others. "Unary" means that the connective is applied to one formula; "binary" means it is applied to two.

The inductive definition of a *formula* reads as follows.

1. Each propositional variable is a formula.
2. If Φ and Ψ are formulas, then so are $\neg\Phi$, $(\Phi \,\&\, \Psi)$, $(\Phi \vee \Psi)$, and $(\Phi \rightarrow \Psi)$.
3. Each formula results from a set of propositional variables by finitely many applications of the formulas in part 2.

Examples of formulas are as follows.

$$p_1, \quad \neg p_2, \quad (p_1 \,\&\, \neg p_2), \quad p_3, \quad ((p_1 \,\&\, \neg p_2) \vee p_3),$$

$$p_4, \quad (((p_1 \,\&\, \neg p_2) \vee p_3) \rightarrow p_4),$$

and so on. Outermost parentheses are often deleted; for example, $((p_1 \,\&\, \neg p_2) \vee p_3)$ is written as $(p_1 \,\&\, \neg p_2) \vee p_3$. Formulas are symbolic propositions that can be true or false. In other words, with each formula Φ we associate a one-dimensional Boolean variate $(\Phi, \{1, 0\})$. Furthermore, we associate with Φ an n-dimensional Boolean array showing how the truth value of Φ depends on the truth values of the atoms $p_1, \ldots, p_n$.

Presented here are four particular arrays called *truth tables of connectives*.

	$T_\neg$
1	0
0	1

	$T_\&$
11	1
10	0
01	0
00	0

	$T_\vee$
11	1
10	1
01	1
00	0

	$T_\rightarrow$
11	1
10	0
01	1
00	1

These arrays can be interpreted intuitively: the negation of a proposition is true iff the proposition is false; the conjunction of two propositions is true iff both propositions are true; the disjunction of two propositions is true iff at least one of them is true; the implication of two propositions is true if it is not the case that the first one (the *antecedent*) is true and the second one (the *succedent*) is false.

The n-dimensional array, called the *truth table*, of a formula Φ is defined inductively according to the definition of a formula:

$$\text{If } \Phi \text{ is an atom, } p_i, \text{ then } T_\Phi(u_1, \ldots, u_n) = 1 \text{ iff } u_i = 1.$$

Furthermore,

$$T_{\Phi \& \Psi}(u_1, \ldots, u_n) = T_\&(T_\Phi(u_1, \ldots, u_n), T_\Psi(u_1, \ldots, u_n)),$$

$$T_{\Phi \vee \Psi}(u_1, \ldots, u_n) = T_\vee(T_\Phi(u_1, \ldots, u_n), T_\Psi(u_1, \ldots, u_n)),$$

$$T_{\Phi \to \Psi}(u_1, \ldots, u_n) = T_\to(T_\Phi(u_1, \ldots, u_n), T_\Psi(u_1, \ldots, u_n)).$$

As an example, let us successively compute the truth table of the formula $(p_1 \,\&\, \neg\, p_2) \vee p_3$:

p_1	p_2	p_3	$\neg\, p_2$	$p_1 \,\&\, \neg\, p_2$	$(p_1 \,\&\, \neg\, p_2) \vee p_3$
1	1	1	0	0	1
1	1	0	0	0	0
1	0	1	1	1	1
1	0	0	1	1	1
0	1	1	0	0	1
0	1	0	0	0	0
0	0	1	1	0	1
0	0	0	1	0	0

Recall that the set $\{1, 0\}^n$ of all n-tuples of zeros and ones is the n-dimensional Boolean frame. Each n-tuple $(\varepsilon_1, \ldots, \varepsilon_n)$ is a combination of truth values and may be called a *possible world* (one possibility of how truth values may be assigned to the propositions $p_1, \ldots, p_n$). Observe that each formula Φ built from $p_1, \ldots, p_n$ divides possible worlds into two sets: the set of all worlds in which Φ is true ($T_\Phi(u_1, \ldots, u_n) = 1$) and the set of all worlds in which Φ is false. A *tautology* is a formula true in each world; for example, $(p_1 \vee \neg\, p_1)$ is a tautology.

A formula Φ is a *logical consequence* of a set of formulas $\Psi_1, \ldots, \Psi_k$ if Φ is true in each world in which all the formulas $\Psi_1, \ldots, \Psi_k$ are true. For example, Φ is a consequence of Ψ and $(\Psi \to \Phi)$; in other words, if the implication $(\Psi \to \Phi)$ is true in a world and its first member (antecedent) Ψ is true in that world then the second member Φ (succedent) is also true in that world. Thus we may infer Φ from $(\Psi \to \Phi)$ and Ψ; knowing that Ψ and $\Psi \to \Phi$ are true we know that Φ is also true. The rule "infer Φ from Ψ and $(\Psi \to \Phi)$" is called *modus ponens*.

Each set K of propositional formulas may play the role of a *Boolean knowledge base* (on some domain). K is consistent if at least one world exists in which all elements of K are true; otherwise it is inconsistent (contradictory). In artificial intelligence (AI), one often works with particular formulas having the form $(\Psi \to \Phi)$, where Ψ and Φ are rather simple pairwise disjoint formulas, for example, $p_1 \,\&\, \neg\, p_3 \,\&\, p_5 \to \neg\, p_7$ [more explicitly, $((p_1 \,\&\, \neg\, p_3) \,\&\, p_5) \to \neg\, p_7$]; these are called *rules* and a knowledge base consisting of rules is a *rule base*.

If K is a rule base, then the task to be solved by a logical expert system may be formulated as follows: given truth values of some propositions, decide (using K) the truth value of another proposition (if possible). This will be discussed below in detail; for now a trivial example is as follows: if K contains just the rule above and we know

that p_1 and p_5 are true but p_3 is false, then we know that p_7 is false, that is, we may infer $\neg p_7$ from p_1, $\neg p_3$, and p_5 using the rule base. If we know that p_1, p_3, and p_5 are true we can infer neither p_7 nor $\neg p_7$.

We shall also introduce *three-valued* propositional calculi (with the truth values "yes, no, unknown") and shall, more generally, introduce *many-valued* propositional calculi and discuss their relevance for expert systems.

3. PROBABILITY

Let V be a frame; a *probability distribution* on V is an array P on V, assigning to each element $v \in V$ a real number $0 \leq P(v) \leq 1$ and such that $\Sigma_{v \in V} P(v) = 1$. More generally, a *potential* is an array A on V such that each $A(v)$ is a nonnegative real number and $\Sigma_{v \in V} A(v)$ is positive (i.e., at least one value $A(v)$ is positive).

Clearly, each potential can be transformed to a probability distribution through *normalization*, i.e., by setting $P(v) = A(v)/(\Sigma_{w \in V} A(w))$.

A probability distribution P on a frame V uniquely defines the probability of each subset U of V as follows: $P(U) = \Sigma_{v \in U} P(v)$. This is the classical definition of a probability on a finite set. We denote both the probability distribution and the corresponding probability by P if there is no danger of misunderstanding.

A *random variate* is a variate (N, V) (where N is the name and V the frame of the variate) together with a proability distribution P on V. (Recall that we consider only finite frames.) An example of a random variate is as follows.

	Color		**Color**		
Probability distribution	Red	0.6	Red	120	
	Blue	0.3	Blue	60	Potential
	White	0.1	White	20	

A *random field* consists of a field of variates $\mathbf{X} = (X_i)_{i \in I}$ and of a probability distribution P on its frame $\mathbf{V} = \Pi_{i \in I} V_i$. For example, if $(X_i)_{i=1,2}$ is as above, then the following is a corresponding random field:

Sex	**Temperature**	
Male	Low	0.2
Male	Medium	0.2
Male	High	0.1
Female	Low	0.1
Female	Medium	0.2
Female	High	0.2

Generalizing the above, each tuple $\mathbf{v} \in \mathbf{V} = \Pi_{i \in I} V_i$ can be called a *possible world*; a random variate determines the probability of each possible world. The *probability of a set* U of worlds is then defined as the sum of probabilities of the elements of U, that is, $P(U) = \Sigma_{\mathbf{v} \in U} P(\mathbf{v})$.

A set of worlds may be defined by various conditions: for example if $I = \{1, 2, \ldots, n\}$ then U may be the set of all tuples the second coordinate of which is c, that is, $U = \{(u_1, \ldots, u_n)\colon u_2 = c\}$. Because the name of the second variate is N_2 we can write $P(U)$ (for this particular U) as $P(N_2 = c)$ or $P((c)N_2)$ or even $P(X_2 = c)$, and so on, just replacing the set in question by its definition. The meaning of $P(X_2 = c \,\&\, X_3 = d)$ should then be clear.

Let $(\mathbf{X}, P)$ be a random field with the index set I (i.e., $\mathbf{X}$ is a field $(X_i)_{i \in I}$ and P is a probability distribution on $\mathbf{V} = \prod_{i \in I} V_i$). Let $\emptyset \neq J \subseteq I$ and consider the field $\mathbf{X}_J = (X_i)_{i \in J}$. The *marginalization* P_J of P to $\mathbf{X}_J$ is the probability distribution on the frame $\mathbf{V}_J = \prod_{i \in J} V_i$ defined as follows: for $\mathbf{u} \in \mathbf{V}_J$,

$$P_J(\mathbf{u}) = \sum [P(\mathbf{v}) : \mathbf{v} \in \mathbf{V} \text{ and } \mathbf{v}_J = \mathbf{u}]$$

where $\mathbf{v}_J = (v_i)_{i \in J}$; thus $P_J(\mathbf{u})$ is the sum of $P(\mathbf{v})$ for all $\mathbf{v} = (v_i)_{i \in I} \in \mathbf{V}$ whose restrictions to J is $\mathbf{u}$.

For example, recall the probability $P(N_2 = c)$ above; this is the probability of the set U of all n-tuples $\mathbf{u}$ such that $u_2 = c$, hence $\Sigma\{P(\mathbf{u}) : u_2 = c\}$.

The last sum is equal to the marginal probability $P_{\{2\}}(c)$ (the value of the probability distribution defined in V_2). The definition of marginalization readily generalizes for potentials.

CONDITIONAL PROBABILITY. The well-known general definition is as follows: let V be a frame, P a probability distribution on V, and let $U \subseteq V$ be such that $P(U) > 0$. Then the conditional probability $P(W \mid U)$ is defined for all $W \subseteq V$ as $P(W \mid U) = P(W \cap U)/P(U)$. Observe that this is a probability given by the following probability distribution on V:

$$P^{\mid U}(v) = \begin{cases} P(v)/P(U) & \text{for } v \in U \\ 0 & \text{otherwise.} \end{cases}$$

Example. Consider three Boolean variates X_1, X_2, X_3 and let P be a probability distribution on $\{1, 0\}^3$. What is the probability distribution for X_1, X_3 given the condition $X_2 = 1$? First, compute $P(X_2 = 1) = \Sigma\{P(u_1, 1, u_3) : u_1, u_3 \in \{1, 0\}\}$ and denote this constant by α. Then we have conditional probability on $\{0, 1\}^3$:

$$p^{\mid X_2 = 1}(u_1, u_2, u_3) = \begin{cases} P(u_1, u_2, u_3)/\alpha & \text{if } u_2 = 1 \\ 0 & \text{if } u_2 = 0. \end{cases}$$

What we want is the marginalization of this probability to X_1, X_3, that is,

$$P^{\mid X_2 = 1}_{\{1,3\}}(u_1, u_3) = \sum_{u_2} P^{X_2 = 1}(u_1, u_2, u_3) = P(u_1, 1, u_3)/\alpha.$$

This probability distribution is often denoted by $P(X_1, X_3 \mid X_2 = 1)$ if there is no danger of misunderstanding. This notation describes our conditioning as well as marginalization but does not allow one to visualize the arguments. $P(X_1, X_3 \mid X_2)$ signifies, if used, the collection of probability distributions $P(X_1, X_3 \mid X_2 = u)$ for all u from the frame X_2, here for $u = 1$ and $u = 0$.

PROBABILISTIC INFERENCE. Assume the variates $X_1, \ldots, X_n$ as given and consider the field $(X_i)_{i \in I} = \mathbf{X}$. Let P denote an unknown probability distribution over $\mathbf{X}$. Any condition of the form $P_Y(u) = \alpha$ and/or $P^{\mid Z = \beta}_Y(u) = \alpha$ (telling values of some marginalization and/or conditioning of P for some arguments) can be considered as a piece of knowledge about an unknown P. A set K of such conditions is a *stochastic (probabilistic) knowledge base*. It is stochastically consistent if there is at least one probability distribution P on the frame satisfying all conditions from K.

Having such a stochastic knowledge base, the task to be solved by a probabilistic expert system may be formulated as follows: given marginal probabilities of some variates (input information I), determine (or estimate) on the basis of K some other marginal probabilities conditioned by I. It is to be expected that our knowledge base does not determine P uniquely; this problem will be discussed in detail in Chapter 5.

4. GRAPHS

An *oriented graph* **G** consists of a set I of *vertices* and a set of oriented *edges*; each edge goes from a unique vertex α into a unique vertex β different from α. We identify edges with ordered pairs of vertices, thus $E \subseteq I \times I$. It is customary to depict graphs using small circles as vertices and arrows as edges, for example, Figure 1 is a representation of the oriented graph (I, E) with $I = \{a, b, c, d\}$ and $E = \{\langle a, b\rangle, \langle a, c\rangle, \langle b, d\rangle\}$.

An *unoriented graph* is defined similarly but now E consists of unordered pairs of vertices. Unoriented edges are depicted as lines without any orientation; thus Figure 2 represents (I, E) where I is as above and $E = \{\{a, b\}, \{a, c\}, \{b, d\}\}$. Each oriented graph determines uniquely the oriented graph resulting by "forgetting the orientation," that is, replacing each oriented edge $\langle x, y\rangle$ by an unoriented edge $\{x, y\}$; on the other hand, unoriented graphs are in one-to-one correspondence with oriented graphs whose set E of edges is a symmetric relation, that is, $\langle x, y\rangle \in E$ implies $\langle y, x\rangle \in E$ for each $x, y \in I$ (see Figure 3). Note that it is also common to call oriented graphs *directed graphs* and unoriented graphs *undirected graphs*.

Consider the notation of oriented graphs: if an oriented graph $\mathbf{G} = (I, E)$ is fixed and $\alpha, \beta \in I$, then $\alpha \to \beta$ means $\langle \alpha, \beta\rangle \in E$; thus, there is an edge from α to β.

A sequence $\alpha_1, \ldots, \alpha_n$ of vertices is a *path* if $\alpha_1 \to \alpha_2 \to \cdots \to \alpha_n$; a *cycle* is a path $\alpha_1, \ldots, \alpha_n$ for which $\alpha_1 = \alpha_n$. An oriented graph is *acyclic* if it contains no cycles. It is usual to speak of DAGs (directed acyclic graphs).

Let $\mathbf{G} = (I, E)$ be a finite DAG and let $\alpha \in I$. The *height* of α is the maximum length of all paths in **G** the last element of which is α. Clearly, since each path in a DAG is a sequence without repetitions and **G** is finite, each vertex has a uniquely determined finite height denoted by $\mathrm{hg}_G(\alpha)$.

For each $\alpha \in \mathbf{G}$, let $\mathrm{pa}(\alpha)$ be the set of all parents of α:

$$\mathrm{pa}(\alpha) = \{\beta \in I : \beta \to \alpha\}.$$

Clearly, $\mathrm{hg}_G(\alpha) = 1$ if $\mathrm{pa}(\alpha) = \emptyset$; otherwise

$$\mathrm{hg}_G(\alpha) = 1 + \max_G\{\mathrm{hg}_G\, \beta \mid \beta \in \mathrm{pa}(\alpha)\}.$$

We may prove assertions on elements of a DAG by induction on height; the exact formulation follows.

Lemma. *Let* **G** *be a DAG and* $A \subseteq I$. *Assume the following.*

1. *All vertices of height* 1 *belong to* A.
2. *For all* $\alpha \in I$, *if* $\mathrm{pa}(\alpha) \subseteq A$ *then* $\alpha \in A$.

Then $A = I$.

This lemma is proved by the usual induction on k that all vertices of height k are in A.