Chapter 2
Formal Preliminaries

Abstract This chapter introduces the formal concepts required for subsequent chapters. Some notational conventions are maintained and important terms such as 'relation', 'function', and 'structure' are explicated. The probabilistic concepts relevant for later chapters are illustrated. I also explain the most important graph-theoretical concepts and introduce Bayesian networks. Most of the presented concepts are illustrated by means of simple examples.

2.1 Overview

This chapter introduces the formal concepts required for subsequent chapters. In Sect. 2.2, some notational conventions are maintained and important terms such as 'relation', 'function', and 'structure' are explicated within a set theoretical framework. In Sect. 2.3, the relevant probabilistic concepts as well as some theorems which will be useful in later chapters are illustrated. Statistical variables are introduced in Sect. 2.4; the main differences between statistical variables and predicate constants are highlighted. Next, the notion of a probability distribution over a set of statistical variables is explicated, followed by demonstrating how the in Sect. 2.3 introduced probabilistic concepts and theorems can be applied to statistical variables. In Sect. 2.5 probabilistic dependence/independence relations between statistical variables and some interrelated notions are introduced. In Sect. 2.6 the most important graph-theoretical concepts are explained. Last but not least, Sect. 2.7 introduces Bayesian networks, and thus, finally connects graphs to probability distributions over sets of statistical variables. All theorems and equations presented in Chap. 2 are stated without proof; most of the presented concepts are illustrated by means of simple examples. Readers already familiar with these concepts can just skip the whole chapter.

2.2 Logic and Set Theory

During the following chapters I presuppose a standard first order logic with identity in which the symbols '¬', '∧', and '∨' stand for the standard sentential connectives *negation*, *conjunction*, and *disjunction*, respectively, while '=' stands for *identity*. The symbols '∀' and '∃' stand for *universal* and *existential quantification*, respectively. Because plain arrows are reserved for representing causal relations, the symbols '⇒' and '≡' shall be used for the sentential connectives *implication* and *equivalence*, respectively. Upper-case letters from A to C ('A', 'B', 'C', 'A_1', 'B_1', 'C_1', etc.) are meta-variables for formulae, lower-case letters from a to c ('a', 'b', 'c', 'a_1', 'b_1', 'c_1', etc.) are individual constants, and lower-case letters from u to w ('u', 'v', 'w', 'u_1', 'v_1', 'w_1', etc.) are individual variables. 'Q', 'R', 'Q_1', 'R_1', etc. are used for predicate constants. '±' is used as a meta-symbol so that '±A' stands for 'either A or ¬A', while '$≡_{df}$' and '$=_{df}$' are used as meta-symbols that indicate definitions.[1]

In addition I presuppose a typical set theoretical framework in which '∈' stands for the *element relation*, while '{ }' indicates specific sets. Most of the time, 'M', 'N', 'M_1', 'N_1', etc. will be used for designating sets. 'Ø' stands for the *empty set*. The symbols '⊆' and '⊂' stand for the relations *subset* and *proper subset*, respectively, while '∪', '∩', '⋃', '⋂', '$\mathcal{P}$', '×', '¯', and '\' are constants for the functions *union*, *intersection*, *general union*, *general intersection*, *powerset*, *Cartesian product*, *complement*, and *relative complement*, respectiveley. '⟨ ⟩' indicates *n*-tuples, '[ ]' stands for *closed* and '] [' for *open interval*, and '| |' for *cardinality*.

2.3 Probability Theory

Basics When specifying a probability function, one typically starts by identifying a set of *elementary events* $e_1, \ldots, e_n$. $e_1, \ldots, e_n$ can, for example, be the possible outcomes of an experiment. Next one can choose an algebra over $\{e_1, \ldots, e_n\}$ such as, for instance, $\mathcal{P}(\{e_1, \ldots, e_n\})$.[2] Let us call $A \in \mathcal{P}(\{e_1, \ldots, e_n\})$ an *event*. Then we can define a *probability function P* as a function satisfying the following three axioms of probability calculus:

(1) $0 \leq P(A) \leq 1$

(2) $P(A \vee \neg A) = 1$

(3) $P(A \vee B) = P(A) + P(B)$, *provided A and B are mutually exclusive* (2.1)

---

[1] The symbol '$≡_{df}$' stands for a definition via equivalence, while '$=_{df}$' stands for a definition via identity.

[2] An algebra over a set M is a subset of $\mathcal{P}(M)$ that contains M and is closed under the complement as well as the union operation.

Fig. 2.1
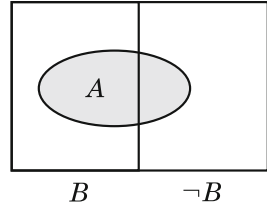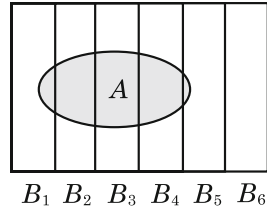$$P(A) = P(A \wedge B) + P(A \wedge \neg B)$$

B $\neg B$

Fig. 2.2
$$P(A) = \sum_{i=1}^{6} P(A \wedge B_i)$$

B_1 B_2 B_3 B_4 B_5 B_6

According to axiom (1), the probability of an event A lies within the interval $[0, 1]$, axiom (2) assures that the probability of the *sure event* equals 1, and axiom (3) tells us how to compute the probability of an event $A \vee B$ on the basis of the probabilities of A and B, provided A and B are mutually exclusive.

Important notions and theorems Let us take a brief look at some interesting stipulations and theorems of probability calculus. Let us begin with the following formula that tells us how to compute the probability of A whenever we know the probabilities of $A \wedge B$ and $A \wedge \neg B$:

$$P(A) = P(A \wedge B) + P(A \wedge \neg B) \tag{2.2}$$

How Equation 2.2 works can be illustrated by means of the diagram in Fig. 2.1. The areas in the diagram correspond to the probabilities assigned by P. The area of the whole diagram corresponds to the sure event which gets probability 1. The left half (area B) of the diagram corresponds to the probability of B (which equals 0.5) and the right half (area $\neg B$) of the diagram corresponds to the probability of $\neg B$ (which also equals 0.5). The part of area A that lies in area B corresponds to the probability of $A \wedge B$, the part of area A that lies in area $\neg B$ corresponds to the probability of $A \wedge \neg B$, and the whole area A corresponds to the probability of A, i.e., the probability of $A \wedge B$ plus the probability of $A \wedge \neg B$.

The basic idea behind determining the probability of A by means of the probabilities of $A \wedge B$ and $A \wedge \neg B$ can be generalized to the so-called *law of total probability* (see Fig. 2.2 for an illustration by means of a diagram). Whenever $\{B_1, \ldots, B_n\}$ is a set of exhaustive and mutually exclusive events, then:

$$P(A) = \sum_{i=1}^{n} P(A \wedge B_i) \tag{2.3}$$

Fig. 2.3 $P(A|B) = \frac{P(A \wedge B)}{P(B)}$
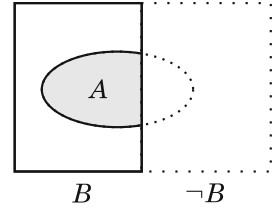
An important probabilistic concept is the concept of *conditional probability*, which can be defined for all cases in which $P(B) > 0$ holds[3]:

$$P(A|B) =_{df} \frac{P(A \wedge B)}{P(B)} \tag{2.4}$$

The main idea behind this definition is that the probability of A conditional on B should equal the probability of A in the light of B, i.e., the probability of A when B is treated as if it were the sure event. This can be illustrated by means of the diagram in Fig. 2.3: Like in Fig. 2.1, the area of the whole diagram (i.e., the area included within the continuous and the dashed lines) corresponds to the sure event, which gets probability 1, while the areas B and $\neg B$ correspond to the probabilities of B and $\neg B$, respectively. When conditionalizing on B, we treat B as if it were the sure event, i.e., as if B would get probability 1. We imagine the area of the whole diagram to be restricted to area B. The probability of A conditional on B then corresponds to the ratio of the parts of area A in B to the whole area B.

The so-called *product rule*, a theorem that allows one to compute the joint probability of two events A and B on the basis of the conditional probability of A given B and the probability of B, is a direct consequence of the definition of conditional probability (Equation 2.4):

$$P(A \wedge B) = P(A|B) \cdot P(B) \tag{2.5}$$

The product rule can be generalized to the so-called *chain rule formula*:

$$P(A_1 \wedge \ldots \wedge A_n) = P(A_1) \cdot P(A_2|A_1) \cdot \ldots \cdot P(A_n|A_1 \wedge \ldots \wedge A_{n-1}) \tag{2.6}$$

Or equivalently:

$$P(A_1 \wedge \ldots \wedge A_n) = \prod_{i=1}^{n} P(A_i|A_1 \wedge \ldots \wedge A_{i-1}) \tag{2.7}$$

---

[3]This restriction is required because division by 0 is undefined.

As we have seen, the product rule (Equation 2.5) allows us to compute the joint probability of $A \wedge B$ on the basis of the conditional probability of A given B and the probability of B. But what if we want to know the probability of A? Well, in that case we just have to sum up the conditional probability of A in the light of B weighted on B's probability and the conditional probability of A in the light of $\neg B$ weighted on $\neg B$'s probability:

$$P(A) = P(A|B) \cdot P(B) + P(A|\neg B) \cdot P(\neg B) \tag{2.8}$$

Equation 2.8 can be illustrated by Fig. 2.1: Since, according to Equation 2.5, $P(A|B) \cdot P(B)$ equals $P(A \wedge B)$ and $P(A|\neg B) \cdot P(\neg B)$ equals $P(A \wedge \neg B)$, the part of area A in B in Fig. 2.1 corresponds to $P(A|B) \cdot P(B)$ and the part of area A in $\neg B$ in Fig. 2.1 corresponds to $P(A|\neg B) \cdot P(\neg B)$. The sum of these two parts of A corresponds to the probability of A.

Equation 2.8 can be generalized to the following *law of total probability*. Whenever $\{B_1, \ldots, B_n\}$ is a set of exhaustive and mutually exclusive events, then:

$$P(A) = \sum_{i=1}^{n} P(A|B_i) \cdot P(B_i) \tag{2.9}$$

An illustration of how Equation 2.9 works (analogously to the one given above for Equation 2.8 by means of Fig. 2.1) can be given for Equation 2.9 by means of Fig. 2.2.

The following equation is called *Bayes' theorem*:

$$P(B|A) = \frac{P(A|B) \cdot P(B)}{P(A)} \tag{2.10}$$

Bayes' theorem is a direct consequence of the definition of conditional probability (Equation 2.4) and the product rule (Equation 2.5).

2.4 Statistical Variables

Basics Statistical variables represent properties at a very abstract level and can be used like predicate constants in first order logic. In more detail, a statistical variable is a function that assigns an element of a specified space of exhaustive and mutually exclusive sets of properties to every individual in a given domain D. The sets of properties a statistical variable X (I use upper-case letters 'X', 'Y', 'Z', 'X_1', 'Y_1', 'Z_1', etc. from the end of the alphabet for statistical variables) can assign to individuals in a given domain D are called *values* of X. In the following, '$val(X)$' will stand for the set of all values a variable X can assign to individuals u in a given domain D.

Statistical variables can be discrete or continuous. While $val(X)$ of a *discrete variable* X is finite, the set of possible values $val(X)$ of a *continuous variable* X is infinite. One special kind of discrete variables are binary variables. *Binary variables* are variables with exactly two possible values (1 and 0, *yes* and *no*, or *on* and *off*, etc.). Whenever continuous quantities (e.g., weight, mass, length, etc.) are considered in subsequent chapters, they will be represented by discrete variables sufficiently fine-grained to match the applied measurement methods. This means that $val(X)$ with $|val(X)| = n+1$ of such a variable X will, for example, be identical to $\{[0 \cdot \varepsilon, 1 \cdot \varepsilon],]1 \cdot \varepsilon, 2 \cdot \varepsilon], \ldots,](n-1) \cdot \varepsilon, n \cdot \varepsilon],]n \cdot \varepsilon, \infty[\}$, where ε corresponds to the given measurement accuracy, i.e., ε is the smallest measurable quantity given a certain measurement method. This procedure avoids measure theory and the use of integrals, which makes the subsequent chapters much more accessible.

Since probabilistic statements containing statistical variables can get very complex and convoluted, I will use the following conventions: Whenever reference to specific individuals is not necessary, then (i) formulae like '$P(X(u) = x)$' can be replaced by '$P(X = x)$', while (ii) formulae like '$P(X = x)$' can be replaced by '$P(x)$'. Instead of the quite long '$\forall x \in val(X)$' and '$\exists x \in val(X)$' it is oftentimes more convenient to write '$\forall x$' and '$\exists x$', respectively, for short.

Probability distributions over sets of variables Given a set V of statistical variables $X_1, \ldots, X_n$, $P(X_1, \ldots, X_n)$ is called a *probability distribution* over V if and only if P assigns a value $r_i \in [0, 1]$ to every event $A \in val(X_1) \times \ldots \times val(X_n)$. Given a probability distribution P over $V = \{X_1, \ldots, X_n\}$, all kinds of probabilities can be computed. The probability of the instantiation of a variable X_i to some value x_i, for example, can be computed as $\sum_A P(A)$, where A is an element of $val(X_1) \times \ldots \times val(X_n)$ in which x_i occurs. The probability of the instantiation $x_{i_1}, \ldots, x_{i_m}$ of more than one variable $X_{i_1}, \ldots, X_{i_m} \in V$ is, accordingly, defined as $\sum_A P(A)$, where A is an element of $val(X_1) \times \ldots \times val(X_n)$ in which instantiation $x_{i_1}, \ldots, x_{i_m}$ occurs, etc. In fact, every probability distribution over V gives rise to a probability function P over $\mathcal{P}(val(X_1) \times \ldots \times val(X_n))$, which is the power set algebra over the elementary events $X_1 = x_1, \ldots, X_n = x_n$.

Given a probability distribution $P(X_1, \ldots, X_n)$, for every sequence of statistical variables $X_{i_1}, \ldots, X_{i_m}$ a new statistical variable M can be defined. This can be done in the following way: If we want to introduce a variable M for a sequence of variables $X_{i_1}, \ldots, X_{i_m}$, then the set of possible values of this newly introduced variable M can be defined as $val(M) = val(X_{i_1}) \times \ldots \times val(X_{i_m})$, and the probabilities of M's value instantiations $m = \langle x_{i_1}, \ldots, x_{i_m} \rangle$ are defined as $P(x_{i_1}, \ldots, x_{i_m})$. In the following I will often loosely refer to variables M for sequences of variables $X_{i_1}, \ldots, X_{i_m}$ as a sequence or set of variables.

Whenever a probability distribution P over a variable set $V = \{X_1, \ldots, X_n\}$ is specified, the corresponding probability distribution P' for a subset $V' = \{X_{i_1}, \ldots, X_{i_m}\}$ of V can be defined as $P'(X_{i_1}, \ldots, X_{i_m}) =_{df} P(X_{i_1}, \ldots, X_{i_m})$. So P' coincides with P over the value space $val(X_{i_1}) \times \ldots \times val(X_{i_m})$. P' is called P's *restriction* to V' and is denoted by '$P \uparrow V'$'.

Sometimes it may be convenient to define a probability distribution that is conditionalized on a certain fixed context $M = m$, where a context is a set of variables tied to certain values. We can define such a distribution $P_m(X)$ as $P_m(X) =_{df} P(X|m)$.

Important notions and theorems The basic axioms of probability calculus as well as the equations introduced in Sect. 2.3 do also hold for probability distributions over sets of statistical variables. I will demonstrate this for the following more important equations and begin with the law of total probability. (The given equations can be motivated in the same way as their counterparts in Sect. 2.3.) 'A' and 'B_i' in Equation 2.3 must be specified to 'x' and 'y', respectively, where 'x' stands for a value instantiation of a variable X and 'y' ranges over the possible values of a variable Y. (Note that X and Y may also be sets of variables.)

Now the following equation holds in any probability distribution P:

$$P(x) = \sum_{y \in val(Y)} P(x, y) \tag{2.11}$$

Whenever $P(y) > 0$ holds, the conditional probability of x given y for statistical variables is defined as follows:

$$P(x|y) =_{df} \frac{P(x, y)}{P(y)} \tag{2.12}$$

The product rule for statistical variables:

$$P(x, y) = P(x|y) \cdot P(y) \tag{2.13}$$

The chain rule formula for statistical variables:

$$P(x_1, \ldots, x_n) = P(x_1) \cdot P(x_2|x_1) \cdot \ldots \cdot P(x_n|x_1, \ldots, x_{n-1}) \tag{2.14}$$

Or equivalently:

$$P(x_1, \ldots, x_n) = \prod_{i=1}^{n} P(x_i|x_1, \ldots, x_{i-1}) \tag{2.15}$$

The law of total probability for statistical variables:

$$P(x) = \sum_{y \in val(Y)} P(x|y) \cdot P(y) \tag{2.16}$$

And last but not least, Bayes' theorem for statistical variables: Whenever $P(x) > 0$, then:

$$P(y|x) = \frac{P(x|y) \cdot P(y)}{P(x)} \tag{2.17}$$

Equations 2.11, 2.13, 2.14, 2.15, and 2.16 can be generalized for contexts $M = m$ as follows: If (i) A in the formula $P(A)$ appearing at the right hand side of the '$=$' in the respective equation does have the form $\ldots|\ldots$, then just add 'm' at the right-hand side of '$|$'. If (ii) A does not already have the form $\ldots|\ldots$, then add '$|m$' between 'A' and the bracket '$)$'. Following this procedure, we get the following conditionalized versions of the theorems presented before:

$$P(x|m) = \sum_{y \in val(Y)} P(x, y|m) \tag{2.18}$$

$$P(x, y|m) = P(x|y, m) \cdot P(y|m) \tag{2.19}$$

$$P(x_1, \ldots, x_n|m) = P(x_1|m) \cdot P(x_2|x_1, m) \cdot \ldots \cdot P(x_n|x_1, \ldots, x_{n-1}, m) \tag{2.20}$$

$$P(x_1, \ldots, x_n|m) = \prod_{i=1}^{n} P(x_i|x_1, \ldots, x_{i-1}, m) \tag{2.21}$$

$$P(x|m) = \sum_{y \in val(Y)} P(x|y, m) \cdot P(y|m) \tag{2.22}$$

2.5 Correlation and Probabilistic Independence

Probabilistic dependence/independence relations Statistical correlation is a relation among statistical variables or sets of variables.[4] Probabilistic dependence can be defined with respect to a given probability distribution in the following way:

Definition 2.1 (probabilistic dependence) If P is a probability distribution over variable set V and $X, Y, Z \in V$, then: $DEP_P(X, Y|Z) \equiv_{df} \exists x \exists y \exists z (P(x|y, z) \neq P(x|z) \wedge P(y, z) > 0)$.

Read '$DEP_P(X, Y|Z)$' as 'X is probabilistically dependent on Y conditional on Z in P' or as 'X and Y are correlated given Z in P'. We will follow the convention to identify unconditional dependence $DEP(X, Y)$ with dependence given the empty set $DEP(X, Y|Z = \emptyset)$.

Probabilistic independence can be defined as the negation of statistical correlation:

---

[4]In the following, variables *X, Y, Z* could also be exchanged by sets of variables X, Y, Z and vice versa, where these sets X, Y, Z have to be treated as new variables as explained in Sect. 2.4.

Definition 2.2 (probabilistic independence) If P is a probability distribution over variable set V and $X, Y, Z \in V$, then: $INDEP_P(X, Y|Z) \equiv_{df} \neg DEP_P(X, Y|Z)$, i.e., $\forall x \forall y \forall z (P(x|y, z) = P(x|z) \vee P(y, z) = 0)$ holds.

Again, we identify unconditional independence $INDEP(X, Y)$ as independence given the empty set $INDEP(X, Y|Z = \emptyset)$.

Properties of probabilistic dependence/independence relations The following properties (which are also called *graphoid axioms*) hold for all probability distributions P (Pearl 2000, p. 11; Dawid 1979; Pearl and Paz 1985):

Symmetry: $INDEP_P(X, Y|Z) \Rightarrow INDEP_P(Y, X|Z)$
Decomposition: $INDEP_P(X, \{Y, W\}|Z) \Rightarrow INDEP_P(X, Y|Z)$
Weak union: $INDEP_P(X, \{Y, W\}|Z) \Rightarrow INDEP_P(X, Y|\{Z, W\})$
Contraction: $INDEP_P(X, Y|Z) \wedge INDEP_P(X, W|\{Z, Y\}) \Rightarrow INDEP_P(X, \{Y, W\}|Z)$

Here is an explanation of the four properties if Z is simply the empty set. In that case the axiom of symmetry states that Y does not depend on X when X does not depend on Y. The axiom of decomposition says that whenever X is independent of both Y and W, then it will also be independent of Y alone. The axiom of weak union tells us that conditionalizing on W does not render X and Y dependent if X is independent of both Y and W. The axiom of contraction finally states that X is independent of both Y and W if X is independent of Y and independent of W when conditionalizing on Y.

2.6 Graph Theory

Graphs are tools for representing diverse kinds of systems and relations among parts of these systems. A *graph G* is an ordered pair $\langle V, E \rangle$, where V is a set consisting of any objects. The elements of V are called the *vertices* of the graph. E is a set of so-called *edges*. The edges of a graph are typically lines and/or arrows (possibly having different kinds of heads and tails) that connect two vertices and capture diverse relations among objects in V. One advantage of graphs is that they can represent the structure of a system in a very vivid way: Whenever the domain of objects we are interested in is finite and the relations among these objects we want to represent are binary, then we can draw a corresponding graph. An example: Suppose we are interested in a population M of five people $a_1, a_2, a_3, a_4,$ and a_5 and in the supervisor relation Q. Suppose further that a_1 is a supervisor of a_3, that a_2 is a supervisor of a_3 and a_4, and that a_3 and a_4 are supervisors of a_5. Thus, the structure of the system is $\langle M, Q \rangle$, where $Q = \{\langle a_1, a_3 \rangle, \langle a_2, a_3 \rangle, \langle a_2, a_4 \rangle, \langle a_3, a_5 \rangle, \langle a_4, a_5 \rangle\}$. To draw a graph $G = \langle V, E \rangle$ capturing this structure this graph's vertex set V should be identical to M. In addition, we need to make some conventions about how E and Q are connected. In our example, let us make the following conventions:

Fig. 2.4 A graph
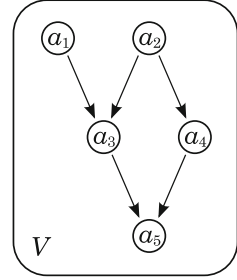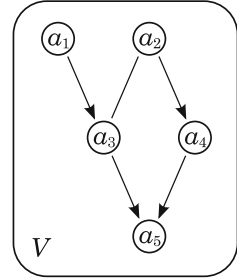representing structure $\langle M, Q \rangle$

Fig. 2.5 A graph
representing a class of
structures, *viz.*
$\{\langle M, Q \rangle, \langle M, Q' \rangle\}$

- $u \text{ --- } v$ in G if and only if $Q(u, v) \vee Q(v, u)$.
- $u \longrightarrow v$ if and only if $Q(u, v) \wedge \neg Q(v, u)$.

Now we can represent the supervisor relation in population M by graph $G = \langle V, E \rangle$, where $V = M$ and $E = \{a_1 \longrightarrow a_3, a_2 \longrightarrow a_3, a_2 \longrightarrow a_4, a_3 \longrightarrow a_5, a_4 \longrightarrow a_5\}$ (see also Fig. 2.4). According to the two conventions about how E and Q are connected, we can uniquely determine G on the basis of $\langle M, Q \rangle$ as well as $\langle M, Q \rangle$ on the basis of G. Another nice feature of graphs is that they can represent more than one structure—they can capture a whole class of structures. The graph in Fig. 2.5, for instance, represents, according to the conventions made above, $\langle M, Q \rangle$ as well as $\langle M, Q' \rangle$ with $Q' = \{\langle a_1, a_3 \rangle, \langle a_3, a_2 \rangle, \langle a_2, a_4 \rangle, \langle a_3, a_5 \rangle, \langle a_4, a_5 \rangle\}$.[5]

After giving some ideas of what can be done by means of graphs, I will give a brief overview of the graph theoretical terminology relevant for the subsequent chapters: A graph $G = \langle V, E \rangle$ is called a *cyclic graph* if there is at least one chain of edges of the form $u \longrightarrow \ldots \longrightarrow u$ in G; otherwise it is called an *acyclic graph*. A graph $G = \langle V, E \rangle$ is called an *undirected graph* if all edges in E have the form $u \text{ --- } v$. A graph $G = \langle V, E \rangle$ including different kinds of edges (e.g., '$\longrightarrow$' and '---') is called a *mixed graph*. A graph $G = \langle V, E \rangle$ whose set of edges E does only contain directed edges ('$\longrightarrow$') is called a *directed graph*. Graphs that are directed as well as acyclic are called *directed acyclic graphs* (or DAGs for short). If V is a set of

---

[5]$\langle M, Q'' \rangle$ with $Q'' = \{\langle a_1, a_3 \rangle, \langle a_2, a_3 \rangle, \langle a_2, a_4 \rangle, \langle a_3, a_2 \rangle, \langle a_3, a_5 \rangle, \langle a_4, a_5 \rangle\}$ is excluded because of the implicit assumption that the supervisor relation is asymmetric.

vertices and all pairs of vertices in V are connected by an edge in graph $G = \langle V, E \rangle$, then G is called a *complete graph* over V.

If two vertices in a graph are connected by an edge, then they are called *adjacent*. A chain π of any kind of edges of the form $u - \ldots - v$ in a graph $G = \langle V, E \rangle$ (where $u, v \in V$) is called a *path* between u and v in G. A path π between u and v that contains a subpath of the form $w_1 \longrightarrow w_2 \longleftarrow w_3$ is called a *collider path* between u and v with w_2 as a *collider* on this path π and is represented by '$u \longrightarrow\longleftarrow v$'. A path π of the form $u \longrightarrow \ldots \longrightarrow v$ is called a *directed path* in G going from u to v and is represented via '$u \longrightarrow\longrightarrow v$'. '$u \overset{\longleftarrow}{\longrightarrow} v$' stands short for a *direct cycle* π of the form $u \longrightarrow v \longrightarrow u$, while '$u \overset{\longleftarrow\longleftarrow}{\longrightarrow\longrightarrow} v$' is short for a (direct or indirect) *cycle* π of the form $u \longrightarrow\longrightarrow v \longrightarrow\longrightarrow u$.

If the set of edges of a graph contains one or more arrows, the following family-terminology can be used for describing several relations among objects in G's vertice set V: Whenever u and v are connected by a directed path π ($u \longrightarrow\longrightarrow v$) in $G = \langle V, E \rangle$, then u is called an *ancestor* of v in G and v is called a *descendant* of u in G. The set of all ancestors of a vertex u shall be referred to via '$Anc_{\langle V,E \rangle}(u)$', while '$Des_{\langle V,E \rangle}(u)$' is used to designate the set of descendants of u. A path π between u and v of the form $u \longleftarrow\longleftarrow w \longrightarrow\longrightarrow v$ such that no vertex on π appears more often than once on π is called a *common ancestor path* between u and v (with w as a *common ancestor* of u and v) and is represented by '$u \longleftarrow\longrightarrow v$'. Whenever $u \longrightarrow v$ holds in G, then u is called a *parent* of v, while v is called a *child* of u. '$Par_{\langle V,E \rangle}(u)$' shall stand for the set of parents of u while '$Chi_{\langle V,E \rangle}(u)$' shall refer to the set of children of u in graph G.

2.7 Bayesian Networks

Markovian parents Bayesian networks were originally developed to compactly represent probability distributions and to simplify probabilistic reasoning (Neapolitan 1990; Pearl 1988). The main idea behind the concept of a Bayesian network is the following: As seen in Sect. 2.4, a probability distribution P over a set of variables V is specified by assigning a value $r_i \in [0, 1]$ to every instantiation $V = v$. Since $|val(V)|$ becomes horribly large even if V contains only a few variables, a lot of space would be required to write the whole probability distribution down, while computing probabilities for specific events $M = m$ (with $M \subseteq V$) can consume a lot of time and resources. So is there any possibility to store probability distributions in a more compact way? The formalism of Bayesian networks provides a positive answer to this question (provided the corresponding graph is sparse; for details see below). According to the chain rule formula (Equation 2.14), $P(x_1, \ldots, x_n) = \prod_{i=1}^{n} P(x_i | x_1, \ldots, x_{i-1})$ holds for arbitrary value instantiations $x_1, \ldots, x_n$ of arbitrary orderings of variables $X_1, \ldots, X_n \in V$, and thus, we can specify a probability distribution also by assigning values $r_i \in [0, 1]$ to every possible value x_i of every variable $X_i \in V$ conditional on all possible combinations of values $x_1, \ldots, x_{i-1}$ of X_i's predecessors $X_1, \ldots, X_{i-1}$ in the given ordering.

Table 2.1 Exemplary
probability distribution P over
$V = \{X, Y, Z\}$

$P(x_1, y_1, z_1) =$	$\frac{9}{32}$
$P(x_1, y_1, z_2) =$	$\frac{3}{32}$
$P(x_1, y_2, z_1) =$	$\frac{2}{32}$
$P(x_1, y_2, z_2) =$	$\frac{2}{32}$
$P(x_2, y_1, z_1) =$	$\frac{6}{32}$
$P(x_2, y_1, z_2) =$	$\frac{2}{32}$
$P(x_2, y_2, z_1) =$	$\frac{4}{32}$
$P(x_2, y_2, z_2) =$	$\frac{4}{32}$

Table 2.2 Application of the
chain rule formula to P

$P(x_1, y_1, z_1) = P(x_1) \cdot P(y_1\|x_1) \cdot P(z_1\|x_1, y_1) =$	$\frac{9}{32}$
$P(x_1, y_1, z_2) = P(x_1) \cdot P(y_1\|x_1) \cdot P(z_2\|x_1, y_1) =$	$\frac{3}{32}$
$P(x_1, y_2, z_1) = P(x_1) \cdot P(y_2\|x_1) \cdot P(z_1\|x_1, y_2) =$	$\frac{2}{32}$
$P(x_1, y_2, z_2) = P(x_1) \cdot P(y_2\|x_1) \cdot P(z_2\|x_1, y_2) =$	$\frac{2}{32}$
$P(x_2, y_1, z_1) = P(x_2) \cdot P(y_1\|x_2) \cdot P(z_1\|x_2, y_1) =$	$\frac{6}{32}$
$P(x_2, y_1, z_2) = P(x_2) \cdot P(y_1\|x_2) \cdot P(z_2\|x_2, y_1) =$	$\frac{2}{32}$
$P(x_2, y_2, z_1) = P(x_2) \cdot P(y_2\|x_2) \cdot P(z_1\|x_2, y_2) =$	$\frac{4}{32}$
$P(x_2, y_2, z_2) = P(x_2) \cdot P(y_2\|x_2) \cdot P(z_2\|x_2, y_2) =$	$\frac{4}{32}$

Table 2.3 P can also be
specified by the conditional
probabilities above

$P(x_1) = \frac{1}{2}$	$P(y_1\|x_1) = \frac{3}{4}$	$P(z_1\|x_1, y_1) = \frac{3}{4}$
	$P(y_1\|x_2) = \frac{1}{2}$	$P(z_1\|x_1, y_2) = \frac{1}{2}$
		$P(z_1\|x_2, y_1) = \frac{3}{4}$
		$P(z_1\|x_2, y_2) = \frac{1}{2}$

Here is an example demonstrating how this procedure can be used to store
probability distributions in a more compact way: Assume X, Y, and Z are binary
variables with $val(X) = \{x_1, x_2\}$, $val(Y) = \{y_1, y_2\}$, and $val(Z) = \{z_1, z_2\}$. Assume
further that P is a probability distribution over $V = \{X, Y, Z\}$ determined by the
equations in Table 2.1. Here we need eight equations, one for each elementary event,
to specify P. Given the ordering X, Y, Z, the equations in Table 2.2 hold due to
the chain rule formula (Equation 2.14). It follows that P can also be specified by
the factors appearing in the equations in Table 2.2, i.e., by the seven equations in
Table 2.3.

We can write down P in an even more compact way. For this purpose, the notion
of the set of the *Markovian parents* (Par^M) of a variable X_i in a given ordering
$X_1, \ldots, X_n$ will be helpful (cf. Pearl 2000, p. 14):

Definition 2.3 (Markovian parents) If P is a probability distribution over variable
set V and $X_1, \ldots, X_n$ is an ordering of the variables in V, then for all $X_i \in V$ and

$M \subseteq V$: M is the set of Markovian parents of X_i if and only if M is the narrowest subset of $\{X_1, \ldots, X_{i-1}\}$ for which $INDEP_P(X_i, \{X_1, \ldots, X_{i-1}\}|M)$ holds.

In other words, the set of Markovian parents of a variable X_i in a given ordering $X_1, \ldots, X_n$ is the narrowest subset of $\{X_1, \ldots, X_{i-1}\}$ such that conditionalizing on $Par^M(X_i)$ makes X_i probabilistically independent of all its predecessors $X_1, \ldots, X_{i-1}$ in this ordering.

Let us now have a closer look at how specifying P in the example discussed above can be simplified by Definition 2.3: Since there are no predecessors of X, $Par^M(X)$ is the empty set. $Par^M(Y)$ cannot be the empty set since $P(y_1) = P(y_1|x_1) \cdot P(x_1) + P(y_1|x_2) \cdot P(x_2) = \frac{3}{4} \cdot \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{2} = \frac{5}{8} \neq \frac{3}{4} = P(y_1|x_1)$ holds, and thus, $DEP_P(X, Y)$. It follows trivially from the definition of conditional probabilistic dependence (Definition 2.2) that Y is probabilistically independent of $\{X\}$ conditional on $\{X\}$, and thus, $Par^M(Y)$ must be $\{X\}$. Since the following equations hold due to Equations 2.16 and 2.17, we get $INDEP_P(Z, \{X, Y\}|\{Y\})$ with Definition 2.2, and hence, $\{Y\}$ seems to be a good candidate for the set of Markovian parents of Z:

$$P(z_1|y_1) = P(z_1|x_1, y_1) \cdot P(x_1|y_1) + P(z_1|x_2, y_1) \cdot P(x_2|y_2)$$

$$= P(z_1|x_1, y_1) \cdot \frac{P(y_1|x_1) \cdot P(x_1)}{P(y_1)} + P(z_1|x_2, y_1) \cdot \frac{P(y_1|x_2) \cdot P(x_2)}{P(y_1)}$$

$$= P(z_1|x_1, y_1) \cdot \frac{P(y_1|x_1) \cdot P(x_1)}{P(y_1|x_1) \cdot P(x_1) + P(y_1|x_2) \cdot P(x_2)}$$

$$+ P(z_1|x_2, y_1) \cdot \frac{P(y_1|x_2) \cdot P(x_2)}{P(y_1|x_1) \cdot P(x_1) + P(y_1|x_2) \cdot P(x_2)}$$

$$= \frac{3}{4} \cdot \frac{\frac{3}{4} \cdot \frac{1}{2}}{\frac{3}{4} \cdot \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{2}} + \frac{3}{4} \cdot \frac{\frac{1}{2} \cdot \frac{1}{2}}{\frac{3}{4} \cdot \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{2}} = \frac{3}{4} = P(z_1|x_1, y_1) = P(z_1|x_2, y_1)$$

$$\tag{2.23}$$

$$P(z_1|y_2) = P(z_1|x_1, y_2) \cdot P(x_1|y_2) + P(z_1|x_2, y_2) \cdot P(x_2|y_2)$$

$$= P(z_1|x_1, y_2) \cdot \frac{P(y_2|x_1) \cdot P(x_1)}{P(y_2)} + P(z_1|x_2, y_2) \cdot \frac{P(y_2|x_2) \cdot P(x_2)}{P(y_2)}$$

$$= P(z_1|x_1, y_2) \cdot \frac{P(y_2|x_1) \cdot P(x_1)}{P(y_2|x_1) \cdot P(x_1) + P(y_2|x_2) \cdot P(x_2)}$$

$$+ P(z_1|x_2, y_2) \cdot \frac{P(y_2|x_2) \cdot P(x_2)}{P(y_2|x_1) \cdot P(x_1) + P(y_2|x_2) \cdot P(x_2)}$$

$$= \frac{1}{2} \cdot \frac{\frac{1}{4} \cdot \frac{1}{2}}{\frac{1}{4} \cdot \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{2}} + \frac{1}{2} \cdot \frac{\frac{1}{2} \cdot \frac{1}{2}}{\frac{1}{4} \cdot \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{2}} = \frac{1}{2} = P(z_1|x_1, y_2) = P(z_1|x_2, y_2)$$

$$\tag{2.24}$$

Table 2.4 The concept of a variable's Markovian parents allows for an even more compact specification of P

$P(x_1) = \frac{1}{2}$	$P(y_1\|x_1) = \frac{3}{4}$	$P(z_1\|y_1) = \frac{3}{4}$
	$P(y_1\|x_2) = \frac{1}{2}$	$P(z_1\|y_2) = \frac{1}{2}$

With help of these equations it is easy to see that $P(z_1) = P(z_1|y_1) \cdot P(y_1) + P(z_1|y_2) \cdot P(y_2) = \frac{3}{4} \cdot \frac{5}{8} + \frac{1}{2} \cdot \frac{3}{8} = \frac{11}{32} \neq \frac{3}{4} = P(z_1|y_1)$, and thus, also $DEP_P(Y, Z)$ holds. So $\emptyset$ cannot be the set of Markovian parents of Z. Therefore, $\{Y\}$ is in fact the narrowest set of predecessors of Z in the given ordering X, Y, Z that screens Z off from all its predecessors, and hence, $\{Y\}$ is the much sought-after set of Markovian parents of Z. So we can store our exemplary probability distribution P via the five (instead of the original eight) equations in Table 2.4 determining the probabilities of the diverse variable values conditional on their Markovian parents.

Summarizing the considerations above, it turns out that the following equation holds for any given ordering of variables $X_1, \ldots, X_n$—this equation provides a simplification of the chain rule formula Equation 2.15 for fixed orderings by means of the notion of a variable's Markovian parents[6]:

$$P(x_1, \ldots, x_n) = \prod_{i=1}^{n} P(x_i|x_1, \ldots, x_{i-1}) = \prod_{i=1}^{n} P(x_i|par^M(X_i)) \qquad (2.25)$$

Markov condition and Markov compatibility Let us come to the notion of a Bayesian network now. Bayesian networks are closely connected to the notion of Markovian parents. A Bayesian network combines a probability distribution P over a set of variables V with a graph over V in such a way that a set of probabilistic independencies that hold in P can be read off this graph's structure. A *Bayesian network* (BN) is an ordered pair $\langle G, P \rangle$, where $G = \langle V, E \rangle$ is a DAG (whose vertex set contains only statistical variables) and P is a probability distribution over this graph's vertex set V that satisfies the so-called *Markov condition* (MC). A DAG and a probability distribution satisfying MC are also said to be *Markov compatible* (cf. Pearl 2000, p. 16).

Definition 2.4 (Markov condition) A graph $G = \langle V, E \rangle$ and a probability distribution P over V satisfy the Markov condition if and only if it holds for all $X \in V$ that $INDEP_P(X, V \backslash Des_G(X)|Par_G(X))$.

So $\langle G, P \rangle$ is a BN if and only if all variables X in G's vertex set V are probabilistically independent of all non-descendants of X conditional on X's parents. Since BNs satisfy the Markov condition, MC and the BN whose graph is depicted

---

[6]While '$Par^M(X)$' denotes the set of X's Markovian parents, '$par^M(X)$' stands for the instantiation of X's Markovian parents $Par^M(X)$ induced by $x_1, \ldots, x_n$ in $P(x_1, \ldots, x_n)$ on the left hand side of Equation 2.25.
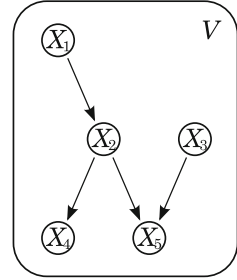
Fig. 2.6 Exemplary DAG

Table 2.5 Independencies implied by MC and the graph depicted in Fig. 2.6

$INDEP_P(X_1, \{X_3\}|\emptyset)$
$INDEP_P(X_2, \{X_3\}|\{X_1\})$
$INDEP_P(X_3, \{X_1, X_2, X_4\}|\emptyset)$
$INDEP_P(X_4, \{X_1, X_3, X_5\}|\{X_2\})$
$INDEP_P(X_5, \{X_1, X_4\}|\{X_2, X_3\})$

in Fig. 2.6 imply, for instance, the probabilistic independence relations in Table 2.5 for the associated probability distribution P over V. (The independencies following trivially from Definition 2.2, e.g., $INDEP(X_1, \{X_1\}|\emptyset)$ or $INDEP(X_2, \{X_1\}|\{X_1\})$, are not mentioned in this list.)

If an ordering $X_1, \ldots, X_n$ of variables in V corresponds to the ordering of these variables in a BN $\langle V, E, P \rangle$ (this means that there is no arrow '$\longrightarrow$' in the BN's graph $G = \langle V, E \rangle$ pointing from a variables X_j to a variable X_i, where X_i is a predecessor of X_j in the given ordering), then the set of Markovian parents of every variable in this ordering is a subset of this variable's set of parents[7]:

$$Par^M(X) \subseteq Par_{\langle V,E \rangle}(X) \tag{2.26}$$

$$P(x_1, \ldots, x_n) = \prod_{i=1}^{n} P(x_i|x_1, \ldots, x_{i-1}) = \prod_{i=1}^{n} P(x_i|par^M(X_i))$$

$$= \prod_{i=1}^{n} P(x_i|par_{\langle V,E \rangle}(X_i)) \tag{2.27}$$

Minimality condition The minimality condition can be defined as follows:

Definition 2.5 (minimality condition) A graph $G = \langle V, E \rangle$ and a probability distribution P over V satisfy the minimality condition if and only if $\langle V, E, P \rangle$ satisfies MC and there is no submodel $\langle V, E', P \rangle$ of $\langle V, E, P \rangle$ with $E' \subset E$ that satisfies MC.

---

[7]Here is an example where the inclusion in Equation 2.26 is strict: Assume our BN has the graph depicted in Fig. 2.6. Assume further that X_5 depends on X_2, but not on X_3. In that case, X_5's only Markov parent in the ordering $X_1, \ldots X_5$ is X_2, while X_5's graphical parents are X_2 and X_3.

Table 2.6 If the graph depicted in Fig. 2.6 is interpreted as the graph of a minimal BN, then the independence and dependence relations below are implied by this BN's topological structure; if interpreted as the graph of a non-minimal BN, on the other hand, only the independencies in the left column are implied. The 'M' on the right hand side of the stroke '|' functions as a proxy for some subset of $V = \{X_1, \ldots X_5\}$ not containing the variables X_i and X_j on the left hand side of the '|'

Independence relations	Dependence relations
$INDEP_P(X_1, \{X_3\}|\emptyset)$	$DEP_P(X_1, \{X_2\}|M)$
$INDEP_P(X_2, \{X_3\}|\{X_2\})$	$DEP_P(X_2, \{X_4\}|M)$
$INDEP_P(X_3, \{X_1, X_2, X_4\}|\emptyset)$	$DEP_P(X_2, \{X_5\}|M)$
$INDEP_P(X_4, \{X_1, X_3, X_5\}|\{X_2\})$	$DEP_P(X_3, \{X_5\}|M)$
$INDEP_P(X_5, \{X_1, X_4\}|\{X_2, X_3\})$	

A BN that satisfies the minimality condition (cf. Spirtes et al. 2000, p. 12) is called a *minimal Bayesian network*. In a minimal BN, every connection between two variables by an arrow $X \longrightarrow Y$ is probabilistically productive. So the graph $\langle V, E \rangle$ of a minimal BN $\langle V, E, P \rangle$ does not only imply some probabilistic independence relations, but also some probabilistic dependence relations that have to hold in any compatible probability distribution P. If the graph depicted in Fig. 2.6, for instance, is the graph of a minimal BN $\langle V, E, P \rangle$, then the dependence/independence relations in Table 2.6 have to hold in P. (Also all dependence and independence relations implied by the relations in Table 2.6 have, of course, to hold in P.)

If a BN $\langle V, E, P \rangle$ does satisyfy the minimality condition, then also the following stronger version of Equation 2.26 holds:

$$Par^M(X) = Par_{\langle V, E \rangle}(X) \tag{2.28}$$

***d*-separation and *d*-connection** If we take a look at any arbitrarily chosen directed acyclic graph $G = \langle V, E \rangle$, then, thanks to MC, we can read off the graph which probabilistic independence relations have to hold in any probability distribution P Markov-compatible with G. The graph depicted in Fig. 2.7, for instance, is compatible with any probability distribution P including the probabilistic independence relations in Table 2.7. So we know a lot about a BN's probability distribution only by looking at the topology of this BN's graph. But is there a way we can learn even more from a BN's graph? MC does often not tell us directly whether two variables of a BN's graph are probabilistically dependent/independent and how their probabilistic dependence/independence would be affected when one conditionalizes on certain variables or sets of variables of this graph. So, for instance, are X_2 and X_4 correlated conditional on X_3 in Fig. 2.7? We could try to use the independencies in Table 2.7 together with the graphoid axioms introduced in Sect. 2.5 to answer this question without tedious probabilistic computation. But maybe there is an even simpler way to answer the question just by looking at the BN's graph?

Fortunately, the answer to this question is an affirmative one: There is a strong connection between the topological properties of a BN's DAG and the diverse
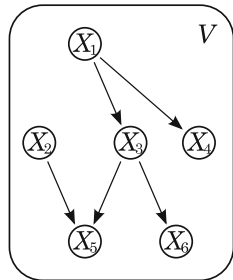
Fig. 2.7 Another exemplary
DAG

Table 2.7 Independencies
implied by MC and the graph
depicted in Fig. 2.7

$INDEP_P(X_1, \{X_2\}\|\emptyset)$
$INDEP_P(X_2, \{X_1, X_3, X_4, X_6\}\|\emptyset)$
$INDEP_P(X_3, \{X_2, X_4\}\|\{X_1\})$
$INDEP_P(X_4, \{X_2, X_3, X_5, X_6\}\|\{X_1\})$
$INDEP_P(X_5, \{X_1, X_4, X_6\}\|\{X_2, X_3\})$
$INDEP_P(X_6, \{X_1, X_2, X_4, X_5\}\|\{X_3\})$

dependence/independence relations which may or may not hold in associated probability distributions P. This connection is established via the notion of *d-separation/d-connection* (cf. Pearl 2000, pp. 16f)[8]:

Definition 2.6 (*d*-separation/*d*-connection) $X, Y \in V$ are *d*-separated by $M \subseteq V \setminus \{X, Y\}$ in directed graph $G = \langle V, E \rangle$ ($SEP^d_{\langle V,E \rangle}(X, Y|M)$) if and only if every path π between X and Y contains a subpath of one of the following forms (where $Z_1, Z_2, Z_3 \in V$):

(a) $Z_1 \longrightarrow Z_2 \longrightarrow Z_3$ with $Z_2 \in M$, or
(b) $Z_1 \longleftarrow Z_2 \longrightarrow Z_3$ with $Z_2 \in M$, or
(c) $Z_1 \longrightarrow Z_2 \longleftarrow Z_3$ with $Z_2 \notin M$, where also no descendant of Z_2 is in M.

X and Y are *d*-connected given M in $G = \langle V, E \rangle$ ($CON^d_{\langle V,E \rangle}(X, Y|M)$) if and only if they are not *d*-separated by M.

When there is a path π between X and Y that goes through M, then we say that M *blocks* this path π if M *d*-separates X and Y, i.e., $SEP^d_{\langle V,E \rangle}(X, Y|M)$ holds. A path π not blocked by M is said to be *activated* by M.

Here are some examples for illustrating how Definition 2.6 works: In the graph in Fig. 2.7, X_1 and X_3 are *d*-connected (given the empty set) because there is a path between X_1 and X_3, *viz.* $X_1 \longrightarrow X_3$, not satisfying one of the conditions (a)–(c) in Definition 2.6. X_4 and X_5 are also *d*-connected (given the empty set); they are connected via path $\pi: X_5 \longleftarrow X_3 \longleftarrow X_1 \longrightarrow X_4$ which contains no subpath of the form $Z_1 \longrightarrow Z_2 \longrightarrow Z_3$ with $Z_2 \in \emptyset$ (thus, (a) is not satisfied), no subpath

---

[8]The term '*d*-connection' is due to the fact that *d*-connection was initially defined for directed graphs; hence, the '*d*' for 'directed'.

of the form $Z_1 \longleftarrow Z_2 \longrightarrow Z_3$ with $Z_2 \in \emptyset$ (thus, (b) is not satisfied), and no subpath of the form $Z_1 \longrightarrow Z_2 \longleftarrow Z_3$ with $Z_2 \notin \emptyset$ and no descendant of Z_2 in $\emptyset$ (thus, (c) is not satisfied). This path π is blocked when one conditionalizes on X_1, X_3, or on $\{X_1, X_3\}$. X_1 and X_2 are d-separated (by the empty set) because X_1 and X_2 are connected only by a collider path where neither the collider X_5 nor one of X_5's descendants (actually, there are none) is an element of the empty set. Though $SEP^d_{\langle V,E \rangle}(X_1, X_2 | \emptyset)$, X_1 and X_2 become d-connected when conditionalizing on X_5. (Conditions (a)–(c) in Definition 2.6 are not satisfied in that case.) If one conditionalizes not only on X_5, but also on X_3 (i.e., on the set $\{X_3, X_5\}$), probability propagation over the path $X_2 \longrightarrow X_5 \longleftarrow X_3 \longleftarrow X_1 \longrightarrow X_4$ is blocked again, because condition (a) of Definition 2.6 is satisfied in that case.

d-separation/d-connection and probabilistic dependence/independence relations of a BN's probability distribution P are connected via the *d-separation criterion* (cf. Pearl 2000, p. 18):

Criterion 2.1 (*d*-separation criterion) *If graph $G = \langle V, E \rangle$ and probability distribution P over V satisfy MC, then $INDEP_P(X, Y | M)$ holds for all $X, Y \in V$ and $M \subseteq V \setminus \{X, Y\}$ whenever $SEP^d_{\langle V,E \rangle}(X, Y | M)$ holds.*

The d-separation criterion identifies all and only the independencies also implied by MC (cf. Pearl 2000, p. 19). Criterion 2.1 finally allows one to read off the independence relations which have to hold in any probability distribution P compatible to a given directed graph G. X_4 and X_5 are, for instance, d-connected in the graph depicted in Fig. 2.7, and thus, X_4 and X_5 may be correlated in the BN's associated probability distribution. (It may also be the case that X_4 and X_5 are independent.) When one conditionalizes on any set of variables lying on the path π: $X_5 \longleftarrow X_3 \longleftarrow X_1 \longrightarrow X_4$, e.g., on X_1, then this path is blocked, and thus, Criterion 2.1 implies $INDEP_P(X_5, X_4 | X_1)$.