*DAACS* is a system for software debugging (Burnell & Horvitz 1995).

### Information processing

*VISTA* is a system used by NASA when launching space shuttles. Its purpose is to filter and display information on the propulsion system (Horvitz & Barry 1995).

Bruza & van der Gaag (1993) developed a language for constructing Bayesian networks for information retrieval, and Fung & Favero (1995) describe another system for information retrieval.

### Medicine

*Child* helps in diagnosing congenital heart diseases (Franklin et al. 1989, Lauritzen et al. 1994). The system is described in Section 3.5.

*MUNIN* is a system for obtaining a preliminary diagnosis of neuromuscular diseases on the basis of electromyografic findings (Andreassen et al. 1989).

*Painulim* diagnoses neuromuscular diseases (Xiang et al. 1993).

*Pathfinder* is of assistance to community pathologists with the diagnosis of lymph-node pathology (Heckerman et al. 1992, Heckerman & Nathwani 1992a,b). The system is described in Section 5.6. Pathfinder has been integrated with videodiscs to the commercial system *Intellipath* (Nathwani et al. 1990).

*SWAN* is a system for insulin dose adjustment of diabetes patients (Andreassen et al. 1991, Hejlesen et al. 1993).

### Miscellaneous

*Hailfinder* was developed for forecasting severe weather in the plane of northeastern Colorado (Abramson et al. 1996).

*FRAIL* is an automatic Bayesian network construction system (Goldman & Charniak 1993). It has been developed for building Bayesian networks for interpretation of written prose (Charniak & Goldman 1991).

# Chapter 2

# Causal and Bayesian networks

This chapter introduces *causal networks* as graphical representations of causal relations in a domain. Through several examples, basic rules for chained reasoning about certainty are introduced. These rules are formalized in the concept of *d-separation.*

In Section 2.3 we present the probability calculus used in this book, and we define the concept of a *Bayesian network.* In Section 2.4 the introductory examples are modelled as Bayesian networks and the reasoning is performed through probability calculations.

Finally we describe the BOBLO system.

## 2.1 Examples

In this section we give three examples. They illustrate crucial points to consider when reasoning about certainty has to be formalized.

### 2.1.1 Icy roads

Police Inspector Smith is impatiently awaiting the arrival of Mr Holmes and Dr Watson; they are late and Inspector Smith has another important appointment (lunch). Looking out of the window he wonders whether the roads are icy. Both are notoriously bad drivers, so if the roads are icy they may crash.

His secretary enters and tells him that Dr Watson has had a car accident. "Watson? OK. It could be worse ... icy roads! Then Holmes has most probably crashed too. I'll go for lunch now."

"Icy roads?", the secretary replies, "It is far from being that cold, and furthermore all the roads are salted." Inspector Smith is relieved. "Bad luck for Watson. Let us give Holmes ten minutes more."

To formalize the story, let the events be represented by variables with two states, *yes* and *no*. Suppose also that to each event is associated a *certainty*, which is a real number. So, we have the three variables: *icy roads (I), Holmes crashes (H)* and *Watson crashes (W)*. $I$ has the effect of increasing the certainty of both $H$ and

$W$. We may think of the impact as an increasing function from the certainty of the cause to the certainty of the effect. The situation is illustrated in Figure 2.1.

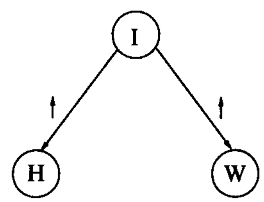


**Figure 2.1** A network model of *icy roads*. The arrows on the links model the causal impact, and the small arrows attached to the links indicate the direction of the impact on the certainty.

When Inspector Smith is told that Watson has had a car accident, he is doing a reasoning in the opposite direction to the causal arrows. Since the impact function pointing at $W$ is increasing, the inverse function is also increasing. Hence, he gets an increased certainty of $I$. The increased certainty of $I$ in turn creates a new expectation, namely an increased certainty of $H$.

Next, when his secretary tells him that the roads cannot possibly be icy, the fact that Watson has crashed cannot change his expectation concerning road conditions and, consequently, Watson's crash has no influence on $H$.

This is an example of how dependence/independence changes with the information at hand. When nothing is known about the condition of the roads, then $H$ and $W$ are *dependent*: information on either event affects the certainty of the other. However, when the condition of the roads is known for certain, then they are *independent*: information on $W$ has no effect on the certainty of $H$ and vice versa. This phenomenon is called *conditional independence.*

### 2.1.2 Wet grass

Mr Holmes now lives in Los Angeles. One morning when Holmes leaves his house, he realizes that his grass is wet. Is it due to rain ($R$), or has he forgotten to turn off the sprinkler ($S$)? His belief in both events increases.

Next he notices that the grass of his neighbour, Dr Watson, is also wet. Elementary: Holmes is almost certain that it has been raining.

A formalization of the situation is shown in Figure 2.2.

When Holmes notices his own wet grass, he is doing a reasoning in the opposite direction to the causal arrows. Since both impact functions pointing at $H$ are increasing, his certainty of both $R$ and $S$ increases. The increased certainty of $R$ in turn creates an increased certainty of $W$.

Therefore Holmes checks Watson's grass, and when he discovers that it is also wet, he immediately increases the certainty of $R$ drastically.

The next step in the reasoning is hard for machines, but natural for human beings, namely *explaining away*: Holmes' wet grass has been explained and thus there is



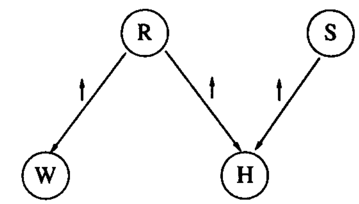


**Figure 2.2** A network model for the *wet grass* example. *Rain* and *sprinkler* are causes of *Holmes'* grass being wet. Only *rain* can cause *Watson's* grass to be wet.

no longer any reason to believe that the sprinkler has been on. Hence, the certainty of $S$ is reduced to its initial size.

Explaining away is another example of dependence changing with the information available. In the initial state, when nothing is known, $R$ and $S$ are independent. However, when we have information on Holmes' grass, then $R$ and $S$ become dependent.

### 2.1.3 Causation and reasoning

A possible source of confusion should be sorted out at this point. The graphs in Figures 2.1 and 2.2 were presented as models for impacts between events, but the reasoning based on the graphs is concerned with how our certainty of the various events is affected by new certainty of other events.

Actually, the models are guidelines for ways of reasoning about unknown events. When reasoning in the direction of the links, the statement in the model is:

*The event A causes with certainty x the event B.*

From this we reason:

*If we know that A has taken place, then B has taken place with certainty x.*

Reasoning in the opposite direction to the links is more delicate. So far we have only said that the certainty of the cause $A$ increases when the consequence $B$ has taken place. If you want to get a quantitative statement, your certainty calculus must have a way of inverting the causal statements. In Section 2.4 we show that for probability calculus, Bayes' rule is used for the inversion.

Some scientists take the point of view that the networks are not causal models, but models for how information may propagate between events. This is, from a foundational point of view, perfectly valid as long as you do not model interfering actions in your network. We shall expand on this in Chapter 6.

### 2.1.4 Earthquake or burglary

Mr Holmes is working at his office when he receives a telephone call from Watson, who tells him that Holmes' burglar alarm ($A$) has gone off. Convinced that a burglar

(*B*) has broken into his house, Holmes rushes to his car and heads for home. O[n] his way he listens to the radio (*R*), and in the news it is reported that there has be[en] a small earthquake (*E*) in the area. Knowing that earthquakes have a tendency [to] turn the burglar alarm on, he returns to his work leaving his neighbours the pleasu[re] of the noise. Figure 2.3 gives a model for the reasoning.

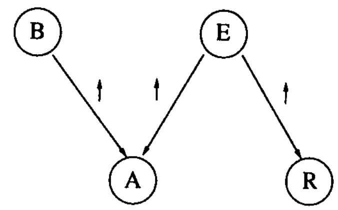


**Figure 2.3** A model for the *earthquake* example. Notice that the structure is similar to Figure 2.2.

### 2.1.5 Prior certainties

It has been typical of the reasoning in the examples of this section that if some event is known, then the certainty of other events must be changed. If, in a certainty calculus, the actual certainty of a specific event has to be calculated, then knowledge of certainties prior to any information is also needed. In particular, prior certainties are required for the events which are not effects of causes in the network.

Take for instance the *wet grass* example. Given that Holmes' grass is wet, the certainty of *R* is still dependent on whether rain at night is a rare event (as in Los Angeles) or very common (as in London).

The same goes for the earthquake in Section 2.1.4. Though *E* may have a stronger effect on *A* than *B* has, and therefore information on *A* will increase the certainty of earthquake more than on burglary, the resulting certainty on *E* should still be lower than the certainty on *B*. To be able to do this reasoning, prior certainties on *E* and *B* are required.

## 2.2 Causal networks and d-separation

The models in Section 2.1 are examples of *causal networks*. A causal network consists of a set of *variables* and a set of *directed links* between variables. Mathematically the structure is called a directed graph. When talking about the relations in a directed graph we use the wording of family relations: if there is a link from *A* to *B* we say that *B* is a *child* of *A*, and *A* is a *parent* of *B*.

The variables represent events (propositions). In Section 2.1, each variable had the states *yes* and *no* reflecting whether a certain event had taken place or not. In general, a variable can have any number of states. A variable may, for example, be the colour of a car (states *blue*, *green*, *red*, *brown*), the number of children in a family (states 0, 1, 2, 3, 4, 5, 6, > 6), or a disease (states *bronchitis*, *tuberculosis*,



*lung cancer*). Variables may have a countable or a continuous state-set, but in this book we solely consider variables with a finite number of states.

In a causal network a variable represents a set of possible states of affairs. A variable is in exactly one of its states; which one may be unknown to us.

Reasoning about uncertainty also has a quantitative part, namely calculation and combination of certainty numbers. The considerations in this section are independent of the particular uncertainty calculus. Whatever calculus is used, it must obey the rules illustrated in Section 2.1 that we formalize in this section.

#### Serial connections

Consider the situation in Figure 2.4. *A* has an influence on *B* which in turn has influence on *C*. Obviously, evidence on *A* will influence the certainty of *B* which then influences the certainty of *C*. Similarly, evidence on *C* will influence the certainty on *A* through *B*. On the other hand, if the state of *B* is known, then the channel is blocked, and *A* and *C* become independent. We say that *A* and *C* are *d-separated given B*, and when the state of a variable is known we say that it is *instantiated*.

*We conclude that evidence may be transmitted through a serial connection unless the state of the variable in the connection is known.*

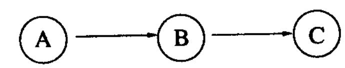


**Figure 2.4** Serial connection. When *B* is instantiated it blocks communication between *A* and *C*.

#### Diverging connections

The situation in Figure 2.5 is a generalization of the *icy roads* example. Influence can pass between all the children of *A* unless the state of *A* is known. We say that *B*, *C*, ..., *E* are d-separated given *A*.

*So, evidence may be transmitted through a diverging connection unless it is instantiated.*

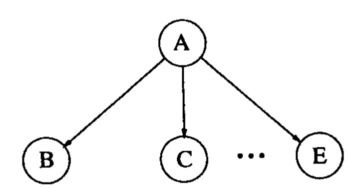


**Figure 2.5** Diverging connection. If *A* is instantiated, it blocks communication between its children.

### Converging connections

A description of the situation in Figure 2.6 requires a little more care. If nothing is known about $A$ except what may be inferred from knowledge of its parents $B, \ldots, E$, then the parents are independent: evidence on one of them has no influence on the certainty of the others.

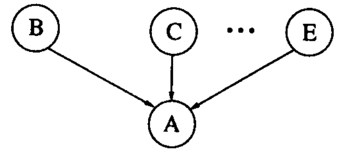


**Figure 2.6** Converging connection. If $A$ changes certainty, it opens communication between its parents.

Now, if any other kind of evidence influences the certainty of $A$, then the parents become dependent due to the principle of explaining away. The evidence may be direct evidence on $A$, or it may be evidence from a child. This phenomenon is called *conditional dependence*. In Figure 2.7 some illustrating examples are listed.

*The conclusion is that evidence may only be transmitted through a converging connection if either the variable in the connection or one of its descendants has received evidence.*

**Remark.** Evidence on a variable is a statement of the certainties of its states. If the statement gives the exact state of the variable we call it *hard* evidence, otherwise it is called *soft*. Hard evidence is also called *instantiation*. Blocking in the case of serial and diverging connections requires hard evidence, while opening in the case of converging connections holds for all kinds of evidence.

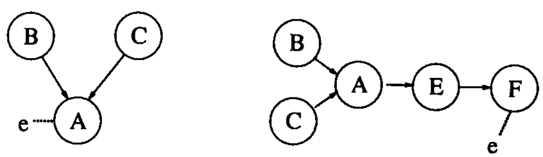


**Figure 2.7** Examples where the parents of $A$ are dependent. The dotted lines indicate insertion of evidence.

### 2.2.1 d-separation

The three cases given above cover all the ways in which evidence may be transmitted through a variable, and following the rules it is possible to decide for any pair of variables in a causal network whether they are dependent given the evidence entered into the network. The rules are formulated in the following.

**Definition** (d-separation). Two variables $A$ and $B$ in a causal network are *d-separated* if for all paths between $A$ and $B$ there is an intermediate variable $V$ such that either



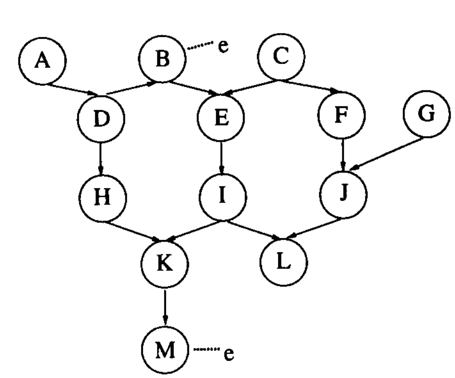


**Figure 2.8** A causal network with $B$ and $M$ instantiated. $A$ is d-separated from $G$ only.

– the connection is serial or diverging and the state of $V$ is known

or

– the connection is converging and neither $V$ nor any of $V$s descendants have received evidence.

If $A$ and $B$ are not d-separated we call them *d-connected*.

Figure 2.8 gives an example of a larger network. The evidence entered at $B$ and $M$ represents instantiation. If evidence is entered at $A$ it may be transmitted to $D$. The variable $B$ is blocked, so the evidence cannot pass through $B$ to $E$. However, it may be passed to $H$ and $K$. Since the child $M$ of $K$ has received evidence, evidence from $H$ may pass to $I$ and further to $E, C, F, J$ and $L$. So, the path $A - D - H - K - I - E - C - F - J - L$ is a d-connecting path.

Figure 2.9 gives two illustrating examples.

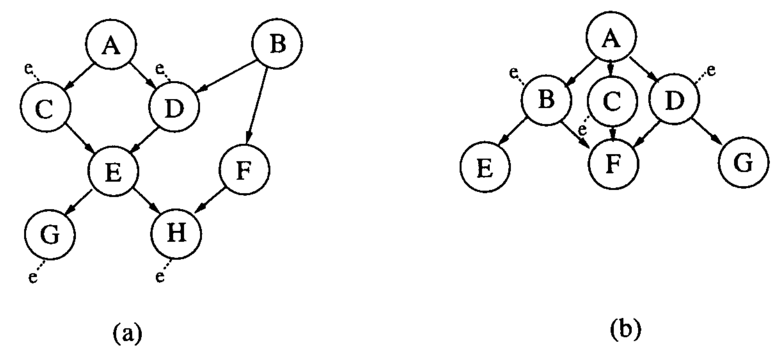


**Figure 2.9** Causal networks with hard evidence entered (the variables are instantiated). (a) Although all neighbours of $E$ are instantiated it is d-connected to $F$, $B$ and $A$. (b) $F$ is d-separated from the remaining un-instantiated variables.

Note that although $A$ and $B$ are d-connected, changes in the belief in $A$ need not change the belief in $B$.

You may wonder why we have introduced d-separation as a definition rather than as a theorem. A theorem should be as follows.

**Claim.** If $A$ and $B$ are d-separated, then changes in the certainty of $A$ have no impact on the certainty on $B$.

However, the claim cannot be established as a theorem without a more precise description of the concept of "certainty". You can take d-separation as a property of human reasoning and require that any certainty calculus must comply with the claim.

## 2.3 Bayesian networks

So far nothing has been said about the quantitative part of certainty assessment. Various certainty calculi exist, but in this book we only treat the so called Bayesian calculus, which is *classical probability calculus*.

### 2.3.1 Basic axioms

The probability $P(A)$ of an event $A$ is a number in the unit interval [0, 1]. Probabilities obey the following basic axioms.

(i) $P(A) = 1$ if and only if $A$ is certain.

(ii) If $A$ and $B$ are mutually exclusive, then

$$P(A \vee B) = P(A) + P(B).$$

### 2.3.2 Conditional probabilities

The basic concept in the Bayesian treatment of certainties in causal networks is *conditional probability*. Whenever a statement of the probability, $P(A)$, of an event $A$ is given, then it is given conditioned by other known factors. A statement like "The probability of the die turning up 6 is $\frac{1}{6}$" usually has the unsaid prerequisite that it is a fair die – or rather, as long as I know nothing of it, I assume it to be a fair die. This means that the statement should be "Given that it is a fair die, the probability ...". In this way, any statement on probabilities is a statement conditioned on what else is known.

A conditional probability statement is of the following kind:

*Given the event $B$, the probability of the event $A$ is $x$.*

The notation for the statement above is $P(A \mid B) = x$.

It should be stressed that $P(A \mid B) = x$ does not mean that whenever $B$ is true then the probability for $A$ is $x$. It means that if $B$ is true, and *everything else known is irrelevant for $A$*, then $P(A) = x$.



The *fundamental rule* for probability calculus is the following:

$$P(A \mid B)P(B) = P(A, B), \tag{2.1}$$

where $P(A, B)$ is the probability of the joint event $A \wedge B$. Remembering that probabilities should always be conditioned by a context $C$, the formula should read

$$P(A \mid B, C)P(B \mid C) = P(A, B \mid C). \tag{2.2}$$

From 2.1 it follows that $P(A \mid B)P(B) = P(B \mid A)P(A)$ and this yields the well known *Bayes' rule*:

$$P(B \mid A) = \frac{P(A \mid B)P(B)}{P(A)}. \tag{2.3}$$

Bayes' rule conditioned on $C$ reads

$$P(B \mid A, C) = \frac{P(A \mid B, C)P(B \mid C)}{P(A \mid C)}. \tag{2.4}$$

Formula (2.2) should be considered an axiom for probability calculus rather than a theorem. A justification for the formula can be found by counting frequencies: suppose we have $m$ cats ($C$) of which $n$ are brown ($B$), and $i$ of the brown cats are Abyssinians ($A$). Then the frequency of $A$s given $B$ among the cats, $f(A \mid B, C)$, is $\frac{i}{n}$, the frequency of $B$s, $f(B \mid C)$, is $\frac{n}{m}$, and the frequency of brown Abyssinian cats, $f(A, B \mid C)$ is $\frac{i}{m}$. Hence,

$$f(A \mid B, C)f(B \mid C) = f(A, B \mid C).$$

#### Likelihood

Sometimes $P(A \mid B)$ is called the *likelihood of $B$ given $A$*, and it is denoted $L(B \mid A)$.

The reason for this is the following. Assume $B_1, \ldots, B_n$ are possible scenarios with an effect on the event $A$, and we know $A$. Then $P(A \mid B_i)$ is a measure of how likely it is that $B_i$ is the cause. In particular, if all $B_i$s have the same prior probability, Bayes' rule yields

$$P(B_i \mid A) = \frac{P(A \mid B_i)P(B_i)}{P(A)} = kP(A \mid B_i),$$

where $k$ is independent of $i$.

### 2.3.3 Subjective probabilities

The justification in the previous section for the fundamental rule was based on frequencies. This does not mean that we only consider probabilities based on frequencies. Probabilities may also be completely subjective estimates of the certainty of an event.

A subjective probability may, for example, be my personal assessment of the chances of selling more than 2,000 copies of this book in 1997.

A way to assess this probability could be the following. I am given the choice between two gambles:

(1) if more than 2,000 copies are sold in 1997 I will receive $100;

(2) I will by the end of 1997 be allowed to draw a ball from an urn with $n$ red balls and $100 - n$ white balls. If my ball is red I will get $100.

Now, if all balls in the urn are red I will prefer (2), and if all balls are white I will prefer (1). There is a number $n$ for which the two gambles are equally attractive, and for this $n$, $\frac{n}{100}$ is my estimate of the probability of selling more than 2,000 copies of this book in 1997 (I shall not disclose the $n$ to the reader).

For subjective probabilities defined through such ball drawing gambles the fundamental rule can also be proved.

### 2.3.4 Probability calculus for variables

As stated in Section 2.2, the nodes in a causal network are *variables* with a *finite number of mutually exclusive states*.

If $A$ is a variable with states $a_1, \ldots, a_n$, then $P(A)$ is a probability distribution over these states:

$$P(A) = (x_1, \ldots, x_n) \qquad x_i \geq 0 \qquad \sum_{i=1}^{n} x_i = 1,$$

where $x_i$ is the probability of $A$ being in state $a_i$.

**Notation**. The probability of $A$ being in state $a_i$ is denoted $P(A = a_i)$ and denoted $P(a_i)$ if the variable is obvious from the context.

If the variable $B$ has states $b_1, \ldots, b_m$, then $P(A \mid B)$ is an $n \times m$ table containing numbers $P(a_i \mid b_j)$ (see Table 2.1).

$P(A, B)$, the joint probability for the variables $A$ and $B$, is also an $n \times m$ table. It consists of a probability for each configuration $(a_i, b_j)$ (see Table 2.2).

When the fundamental rule (2.1) is used on variables $A$ and $B$, then the procedure is to apply the rule to the $n \cdot m$ configurations $(a_i, b_j)$:

$$P(a_i \mid b_j)P(b_j) = P(a_i, b_j).$$

This means that in the table $P(A \mid B)$, for each $j$ the column for $b_j$ is multiplied by $P(b_j)$ to obtain the table $P(A, B)$. If $P(B) = (0.4, 0.4, 0.2)$ then Table 2.2 is the result of using the fundamental rule on Table 2.1. When applied to variables, we use the same notation for the fundamental rule:

$$P(A \mid B)P(B) = P(A, B).$$

From a table $P(A, B)$ the probability distribution $P(A)$ can be calculated. Let $a_i$ be a state of $A$. There are exactly $m$ different events for which $A$ is in state $a_i$, namely the mutually exclusive events $(a_i, b_1), \ldots, (a_i, b_m)$. Therefore, by axiom (ii)

$$P(a_i) = \sum_{j=1}^{m} P(a_i, b_j).$$



**Table 2.1** An example of $P(A \mid B)$. Note that the columns sum to one.

| | $b_1$ | $b_2$ | $b_3$ |
|---|---|---|---|
| $a_1$ | 0.4 | 0.3 | 0.6 |
| $a_2$ | 0.6 | 0.7 | 0.4 |

**Table 2.2** An example of $P(A, B)$. Note that the sum of all entries is one.

| | $b_1$ | $b_2$ | $b_3$ |
|---|---|---|---|
| $a_1$ | 0.16 | 0.12 | 0.12 |
| $a_2$ | 0.24 | 0.28 | 0.08 |

This calculation is called *marginalization* and we say that the variable $B$ is marginalized out of $P(A, B)$ (resulting in $P(A)$). The notation is

$$P(A) = \sum_{B} P(A, B). \tag{2.5}$$

By marginalizing $B$ out of Table 2.2 we get $P(A) = (0.4, 0.6)$.

The division in Bayes' rule (2.3) is treated in the same way as the multiplication in the fundamental rule (see Table 2.3).

### 2.3.5 Conditional independence

The blocking of transmission of evidence as described in Section 2.2.1 is, in the Bayesian calculus, reflected in the concept of *conditional independence*.The variables $A$ and $C$ are *independent given the variable B* if

$$P(A \mid B) = P(A \mid B, C). \tag{2.6}$$

This means that if the state of $B$ is known then no knowledge of $C$ will alter the probability of $A$.

**Table 2.3** $P(B \mid A)$ as a result of applying Bayes' rule to Table 2.1 and $P(B) = (0.4, 0.4, 0.2)$.

| | $a_1$ | $a_2$ |
|---|---|---|
| $b_1$ | 0.4 | 0.4 |
| $b_2$ | 0.3 | 0.47 |
| $b_3$ | 0.3 | 0.13 |

**Remark.** If condition $B$ is empty, we simply say that $A$ and $C$ are independent.

Conditional independence appears in the cases of serial and diverging connections (see Figure 2.10).

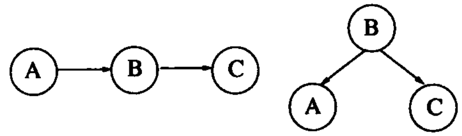


**Figure 2.10** Examples where A and C are conditionally independent given B.

Definition (2.6) may look asymmetric; however, if (2.6) holds, then – by the conditioned Bayes' rule (2.4) – we get

$$P(C \mid B, A) = \frac{P(A \mid C, B)P(C \mid B)}{P(A \mid B)} = \frac{P(A \mid B)P(C \mid B)}{P(A \mid B)} = P(C \mid B).$$

The proof requires that $P(A \mid B) > 0$. That is, for states $a, b$ with $P(A = a \mid B = b) = 0$ the calculation is not valid. However, for our considerations it does not matter; if $B$ is in state $b$ then the evidence $A = a$ is impossible and will not appear. So, why bother with the transmission of it?

### 2.3.6 Definition of Bayesian networks

Causal relations also have a quantitative side, namely their *strength*. This is expressed by attaching numbers to the links.

Let $A$ be a parent of $B$. Using probability calculus it would be natural to let $P(B \mid A)$ be the strength of the link. However, if $C$ is also a parent of $B$, then the two conditional probabilities $P(B \mid A)$ and $P(B \mid C)$ alone do not give any clue on how the impacts from $A$ and $B$ interact. They may co-operate or counteract in various ways. So, we need a specification of $P(B \mid A, C)$.

It may happen that the domain to be modelled contains feed-back cycles (see Fig. 2.11).

Feed-back cycles are difficult to model quantitatively (this is, for example, what differential equations are all about); for causal networks no calculus has been developed that can cope with feed-back cycles Therefore we require that the network does not contain cycles.

A Bayesian network consists of the following.

A set of *variables* and a set of *directed edges* between variables.

Each variable has a finite set of mutually exclusive states.

The variables together with the directed edges form a *directed acyclic graph* (DAG). (A directed graph is *acyclic* if there is no directed path $A_1 \rightarrow \cdots \rightarrow A_n$ such that $A_1 = A_n$.)

To each variable $A$ with parents $B_1, \ldots, B_n$ there is attached a conditional probability table $P(A \mid B_1, \ldots, B_n)$.



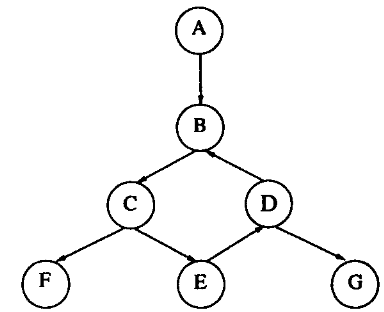


**Figure 2.11** A directed graph with a feed-back cycle. This is not allowed in Bayesian networks.

Note that if $A$ has no parents then the table reduces to unconditional probabilities $P(A)$. For the DAG in Figure 2.12 the prior probabilitiess $P(A)$ and $P(B)$ must be specified. It has been claimed that prior probabilities are an unwanted introduction of bias to the model, and calculi have been invented in order to avoid it. However, as discussed in Section 2.1.5, prior probabilities are necessary – not for mathematical reasons – but because prior certainty assessments are an integral part of human reasoning about certainty.

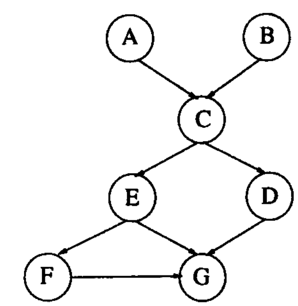


**Figure 2.12** A directed acyclic graph (DAG). The probabilities to specify are $P(A)$, $P(B)$, $P(C \mid A, B)$, $P(E \mid C)$, $P(D \mid C)$, $P(F \mid E)$ and $P(G \mid D, E, F)$.

One of the advantages of Bayesian networks is that they *admit d-separation:* if $A$ and $B$ are d-separated in a Bayesian network with evidence $e$ entered, then $P(A \mid B, e) = P(A \mid e)$. This means that you can use d-separation to read-off conditional independencies. We will use this fact without proof.

### 2.3.7 The chain rule

Let $U = (A_1, \ldots, A_n)$ be a universe of variables. If we have access to the joint probability table $P(U) = P(A_1, \ldots, A_n)$, then we can also calculate $P(A_i)$ as well as $P(A_i \mid e)$, where $e$ is evidence (see Section 4.2). However, $P(U)$ grows exponentially with the number of variables, and $U$ need not be very large before the table becomes intractably large. Therefore, we look for a more compact *representation* of $P(U)$: a way of storing information from which $P(U)$ can be calculated if needed.

A Bayesian network over $U$ is such a representation. If the conditional independencies in the Bayesian network hold for $U$, then $P(U)$ can be calculated from the conditional probabilities specified in the network.

**Theorem 2.1** (The chain rule.) *Let* BN *be a Bayesian network over*

$$U = \{A_1, \ldots, A_m\}.$$

*Then the joint probability distribution $P(U)$ is the product of all conditional probabilities specified in* BN:

$$P(U) = \prod_i P(A_i \mid pa(A_i))$$

*where $pa(A_i)$ is the parent set of $A_i$.*

*Proof.* (Induction in the number of variables in the universe $U$.)

If $U$ consists of one variable then the theorem is trivial.

Assume the chain rule to be true for all networks consisting of $n-1$ variables, and let $U$ be the universe for a DAG with $n$ variables. Since the network is acyclic there is at least one variable $A$ without children. Consider the DAG with $A$ removed.

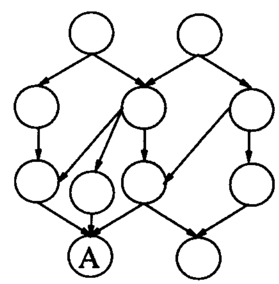


**Figure 2.13** A DAG with $n$ variables. If the variable $A$ is removed, the induction hypothesis can be applied.

From the induction hypothesis we have that $P(U \setminus \{A\})$ is the product of all specified probabilities – except $P(A \mid pa(A))$.

By the fundamental rule we have

$$P(U) = P(A \mid U \setminus \{A\})P(U \setminus \{A\}).$$

Since $A$ is independent of $U \setminus (\{A\} \cup pa(A))$ given $pa(A)$ (see Fig. 2.13), we get

$$P(U) = P(A \mid U \setminus \{A\})P(U \setminus \{A\}) = P(A \mid pa(A))P(U \setminus \{A\}).$$

The righthand side above is the product of all specified probabilities.



**Table 2.4** Conditional probabilities for $H$ and $W$.

| | $I=y$ | $I=n$ |
|---|---|---|
| $H=y$ | 0.8 | 0.1 |
| $H=n$ | 0.2 | 0.9 |
| | $P(H \mid I)$ | |

| | $I=y$ | $I=n$ |
|---|---|---|
| $W=y$ | 0.8 | 0.1 |
| $W=n$ | 0.2 | 0.9 |
| | $P(W \mid I)$ | |

**Table 2.5** Joint probability table for $P(W, I)$ and $P(H, I)$.

| | $I=y$ | $I=n$ |
|---|---|---|
| $y$ | 0.56 | 0.03 |
| $n$ | 0.14 | 0.27 |

.07

## 2.4 The examples revisited

In this section we apply the rules of probability calculus on the introductory examples. This is done to illustrate that probability calculus can be used to perform the reasoning in the examples – in particular explaining away. In Chapter 4 we give a general algorithm for probability updating in Bayesian networks. This algorithm makes the calculations considerably easier than those in this section.

### 2.4.1 Icy roads

(See Fig. 2.1.) For the quantitative modelling we need three probability assessments: $P(H \mid I)$, $P(W \mid I)$ and $P(I)$. The model in Figure 2.1 reflects that only knowledge of icy roads is relevant for $H$ and $W$. We should then attach a certainty to $I$ based on whatever knowledge may be available. In this case the police inspector has been looking out of the window and wondering whether the roads were icy. We let the probability for icy roads be 0.7.

Since both Holmes and Watson are bad drivers, we put the probability of a crash in the case of icy roads to 0.8, and the probability of a crash if the roads are not icy we put to 0.1 (they *are* bad drivers). An overview of the conditional probabilities is given in Table 2.4.

To calculate the initial probabilities for $H$ and $W$ we first use the fundamental rule (2.1) to calculate $P(W, I)$ and $P(H, I)$:

$$P(W=y, I=y) = P(W=y \mid I=y)P(I=y) = 0.8 \cdot 0.7 = 0.56.$$

Table 2.5 gives all four probabilities.

In order to get the probabilities for $W$ and $H$ we marginalize $I$ out of Table 2.5 and get

$$P(W) = P(H) = (0.59, 0.41).$$

The information that Watson has crashed is now used to update the probability of

$I$. For this, Bayes' rule is used:

$$P(I \mid W = y) = \frac{P(W = y \mid I)P(I)}{P(W = y)} = \frac{1}{0.59}(0.8 \cdot 0.7, 0.1 \cdot 0.3) = (0.95, 0.05).$$

To update the probability of $H$, first we use the fundamental rule (2.1) to calcula $P(H, I)$ as shown in Table 2.6.

**Table 2.6** Tables showing the calculation of $P(H, I)$.

| | $I = y$ | $I = n$ | | | $I = y$ | $I = n$ |
|---|---|---|---|---|---|---|
| $H = y$ | $0.8 \cdot 0.95$ | $0.1 \cdot 0.05$ | = | $H = y$ | 0.76 | 0.005 |
| $H = n$ | $0.2 \cdot 0.95$ | $0.9 \cdot 0.05$ | | $H = n$ | 0.19 | 0.045 |

Finally, calculate $P(H)$ by marginalizing $I$ out of $P(H, I)$. The result is

$$P(H) = (0.765, 0.235).$$

This is the quantitative effect of the information that Watson has crashed.

At last, when Inspector Smith is convinced that the roads are not icy, then $P(H \mid I = n) = (0.1, 0.9)$.

The calculation can be considered in a different way. First we calculate $P(H, I)$ and $P(W, I)$ (Table 2.5), and we have two joint probability tables with the variable $I$ in common.

If evidence on $W$ now arrives in the form of $P^*(W) = (0, 1)$, then

$$P^*(W, I) = P(I \mid W)P^*(W) = \frac{P(W, I)}{P(W)} P^*(W).$$

This means that the joint probability table for $W$ and $I$ is updated by multiplying by the new distribution and dividing by the old one. The multiplication consists of annihilating all entries with $W = n$. The division by $P(W)$ only has an effect on entries with $W = y$, so therefore the division is by $P(W = y)$.

Next, calculate $P^*(I)$ from $P^*(W, I)$ by marginalization, and use $P^*(I)$ to update $P(H, I)$

$$P^*(H, I) = \frac{P(H, I)}{P(I)} \cdot P^*(I)$$

and finally $P^*(H)$ is calculated by marginalizing $P^*(H, I)$.

### 2.4.2 Wet grass

(See Fig. 2.2.) Let the prior probabilities for $R$ and $S$ be $P(R) = (0.2, 0.8)$ and $P(S) = (0.1, 0.9)$. The remaining probabilities are listed in Table 2.7. First, calculate the prior probabilities for $W$ and $H$ by formulae (2.1) and (2.5). That is, first



calculate $P(W, R)$ and then marginalize $R$ out. The result is $P(W) = (0.36, 0.64)$.

**Table 2.7** The probabilities for the *wet grass* example. The vectors $(\alpha, \beta)$ in the righthand table represent $(H = y, H = n)$.

| | $R = y$ | $R = n$ |
|---|---|---|
| $W = y$ | 1 | 0.2 |
| $W = n$ | 0 | 0.8 |
| $P(W \mid R)$ | | |

| | $R = y$ | $R = n$ |
|---|---|---|
| $S = y$ | (1, 0) | (0.9, 0.1) |
| $S = n$ | (1, 0) | (0, 1) |
| $P(H \mid R, S)$ | | |

The calculation of $P(H, R, S)$ follows the same scheme, only the product is

$$P(H, R, S) = P(H \mid R, S)P(R, S).$$

Since $R$ and $S$ are independent (see Fig. 2.2) we have (see Exercise 2.9)

$$P(H, R, S) = P(H \mid R, S)P(R)P(S).$$

The result is given in Table 2.8. Marginalizing $R$ and $S$ out of $P(H, R, S)$ yields $P(H) = (0.272, 0.728)$. We shall use the approach outlined at the end of Section 2.4.1. We have established joint probability tables for two of the clusters, $(W, R)$ and $(H, R, S)$, with the variable $R$ in common.

**Table 2.8** The prior probability table for $P(H, R, S)$. The vectors $(\alpha, \beta)$ in the table represent $(H = y, H = n)$.

| | $R = y$ | $R = n$ |
|---|---|---|
| $S = y$ | (0.02, 0) | (0.072, 0.008) |
| $S = n$ | (0.18, 0) | (0, 0.72) |

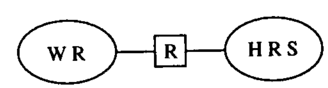


**Figure 2.14** The clusters for the *wet grass* example. They communicate through the variable $R$.

The evidence $H = y$ is used to update $P(H, R, S)$ by annihilating all entries with $H = n$ and dividing by $P(H = y)$. Since the result shall be a probability table with all entries summing to one we need not calculate $P(H)$. After all entries with $H = n$ have been annihilated (Table 2.9), we simply normalize the table by dividing by the sum of the remaining entries (see Table 2.10).

The distributions $P^*(R)$ and $P^*(S)$ are calculated through marginalization of

$$P^*(H, R, S).$$

**Table 2.9** $P(H, R, S)$ with all entries with $H = n$ annihilated.

| | $R = y$ | $R = n$ |
|---|---|---|
| $S = y$ | (0.02, 0) | (0.072, 0) |
| $S = n$ | (0.18, 0) | (0, 0) |

**Table 2.10** The calculation of $P^*(H, R, S) = P(H, R, S \mid H = y)$.

| | $R = y$ | $R = n$ | | | $R = y$ | $R = n$ |
|---|---|---|---|---|---|---|
| $S = y$ | $\frac{1}{0.272}(0.02, 0)$ | $\frac{1}{0.272}(0.072, 0)$ | $\approx$ | $S = y$ | (0.074, 0) | (0.264, 0) |
| $S = n$ | $\frac{1}{0.272}(0.18, 0)$ | $\frac{1}{0.272}(0, 0)$ | | $S = n$ | (0.662, 0) | (0, 0) |

We get $P^*(R = y) = 0.736$ and $P^*(S = y) = 0.339$.

Use $P^*(R)$ to update $P(W, R)$ (see Table 2.11):

$$P^*(W, R) = P(W \mid R)P^*(R) = P(W, R)\frac{P^*(R)}{P(R)}.$$

**Table 2.11** Calculation of $P^*(W, R) = P(W, R)\frac{P^*(R)}{P(R)}$.

| | $R = y$ | $R = n$ | | | $R = y$ | $R = n$ |
|---|---|---|---|---|---|---|
| $W = y$ | $0.2 \cdot \frac{0.736}{0.2}$ | $0.16 \cdot \frac{0.264}{0.8}$ | $=$ | $W = y$ | 0.736 | 0.0528 |
| $W = n$ | 0 | $0.64 \cdot \frac{0.264}{0.8}$ | | $W = n$ | 0 | 0.2112 |

Now use $W = y$ to update the distribution for $(W, R)$ (see Table 2.12). We get $P^{**}(R = y) = 0.93$.

We still have to calculate $P^{**}(S) = P(S \mid W = y, H = y)$. The result must reflect the explaining away effect; since the wet grass is explained by rain, the probability for $S = y$ should decrease to its initial value.

The calculation follows the same pattern. A message on $P^{**}(R)$ is sent from $(W, R)$ to $(H, R, S)$ (see Fig. 2.14),

$$P^{**}(H, R, S) = P^*(H, R, S)\frac{P^{**}(R)}{P^*(R)}.$$

By marginalizing we get $P^{**}(S = y) = 0.161$.



**Table 2.12** $P^{**}(W, R) = P(W, R \mid W = y, H = y)$.

| | $R = y$ | $R = n$ |
|---|---|---|
| $W = y$ | $\frac{0.736}{0.7888}$ | $\frac{0.0528}{0.7888}$ |
| $W = n$ | 0 | 0 |

**Table 2.13** $P^{**}(R, S) = P(R, S \mid H = y, W = y)$.

| | $R = y$ | $R = n$ |
|---|---|---|
| $S = y$ | 0.094 | 0.067 |
| $S = n$ | 0.839 | 0 |

The reason why the probability for sprinkler does not drop to the prior probability of 0.1 is that Dr Watson is a forgetful fellow who may have forgotten his sprinkler, and an explanation may be that both sprinklers have been forgotten. This is reflected in the probability $P(W = y \mid R = n) = 0.2$.

## 2.5 BOBLO

BOBLO is a system which helps in the verification of parentage for Jersey cattle through blood-type identification. The introduction of embryo transplantation technology and the increasing trade of semen and embryos have stressed the importance of proper pedigree registration, and therefore there is a need for sophisticated methods for individual identification and parentage control of cattle.

Heredity is determined by *genes* which are placed in chromosomes (see Fig. 2.15).

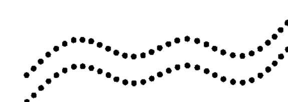

**Figure 2.15** A pair of chromosomes. The pearls in the strings are loci.

Except for the sex chromosomes, chromosomes go in structurally identical pairs – one chromosome inherited from each parent. A chromosome may be considered as a string of genes. The places where the genes are positioned are called *loci*. Each gene has a particular locus of position and genes which can be placed at a particular locus are called *allels*. The pair of allels at a locus (one from each chromosome) is called a *genotype*, and the property determined by a genotype is called the *phenotype*.

For the blood group determination of cattle, ten different independent blood-group systems are used. These systems control 52 different blood-group *factors* which can be measured in a laboratory. In eight of these systems the blood-group determination is relatively simple (controlling from one to four blood-group factors only). However, the systems B- and C- are rather complicated, controlling respectively 2 and 10 of the above-mentioned 52 blood-group factors.

Heredity of blood type follows the normal genetic rules, however, the blood groups are attached to sets of loci rather than to single loci, and instead of allels the term *phenogroup* is used. So, for each blood group, a Bayesian network for inheritance will be as in Figure 2.16.

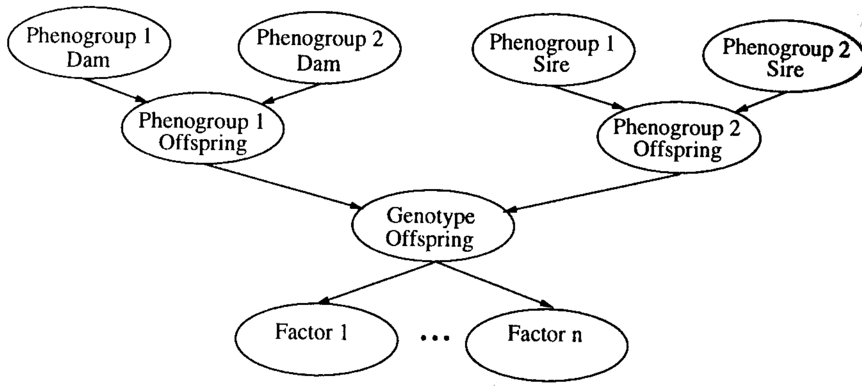


**Figure 2.16** Heredity of blood type. From each parent one out of two phenotypes are chosen. This constitutes the genotype of the offspring, and the genotype determines a set of factors measurable in a laboratory (the phenotype).

If nothing is known of the phenogroups of the parents they are given a prior probability equal to the frequencies of the various phenogroups. Let us, for the example, suppose that there are three phenogroups $f_1, f_2, f_3$ with frequencies $(0.58, 0.1, 0.32)$ (this is the situation for the so-called *F-system*).

When a calf is registered, the parents are stated and their phenogroups are already registered. If the stated parents are the true parents we have no problems, but what if they are not so? Then we will say that the phenogroups of the true parents are distributed as the prior probabilities, that is $(0.58, 0.1, 0.32)$.

So, for modelling the part concerning possible parental errors, we can introduce a node *parental error* with states *both, sire, dam* and *no*, and with prior probabilities to be the frequency of parental errors. This leads to the Bayesian network in Figure 2.17.

The network model in BOBLO also has a part that models the risks of mistakes in the laboratory procedures (see Exercise 3.6). For now, assume that evidence on factors are entered directly to the nodes *factor*. It is assumed that the stated parents are so well known that their genotypes are known, and therefore the state of the variables *phenogroup stated d/s* is known.

Note how the impact of evidence flows from the *factor* nodes to the node *parental error*: it first flows to *phenogroup true d/s* (serial connections). Since evidence has



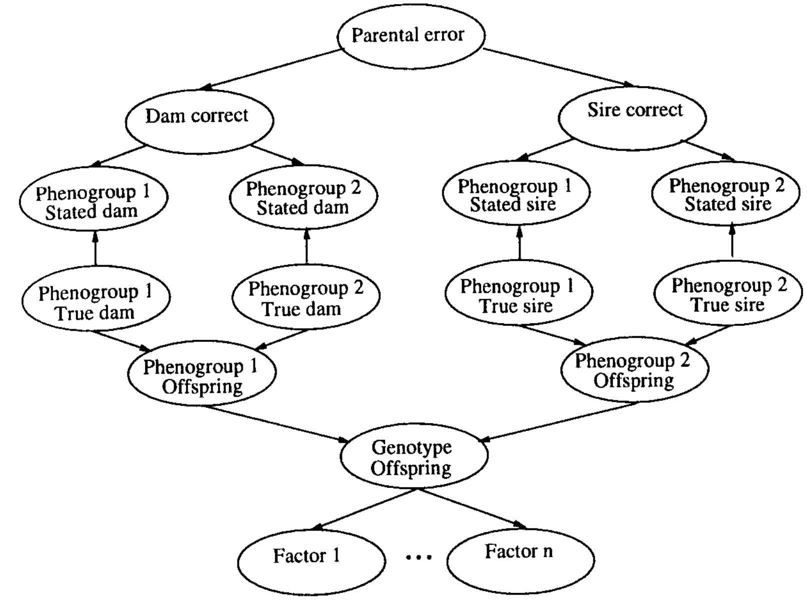


**Figure 2.17** The part of BOBLO modelling parental error. Evidence is entered into the variables *factor* and *phenogroup stated d/s*. Evidence from *factor* is transmitted to *parental error* because *phenogroup stated* has received evidence.

been entered to *phenogroup stated d/s* the evidence is transmitted further to *dam correct* and *sire correct* (converging connections) to end in *parental error*.

BOBLO is an acronym for BOvine BLOod typing, and it has been in use at the Danish Blood Type Laboratory improving the accuracy of detecting parental errors (tests quantifying the improvement have not been finished).

## 2.6 Summary

### d-separation in causal networks

Two variables $A$ and $B$ in a causal network are d-separated if for all paths between $A$ and $B$ there is an intermediate variable $V$ such that either

- the connection is serial or diverging and the state of $V$ is known or
- the connection is converging, and neither $V$ nor any of $V$s descendants have received evidence.

### The fundamental rule for probability calculus

$$P(A \mid B, C)P(B \mid C) = P(A, B \mid C)$$

**Bayes' rule**

$$P(B \mid A, C) = \frac{P(A \mid B, C)P(B \mid C)}{P(A \mid C)}$$

**Marginalization**

$$P(A) = \sum_i P(A, b_i) = P(A, b_1) + \cdots + P(A, b_n)$$

**Conditional independence**

$A$ and $C$ are independent given $B$ if $P(A \mid B) = P(A \mid B, C)$.

**Definition of Bayesian networks**

A Bayesian network consists of the following.

A set of *variables* and a set of *directed edges* between variables.

Each variable has a finite set of states.

The variables together with the directed edges form a *directed acyclic graph* (DAG).

To each variable $A$ with parents $B_1, \ldots, B_n$ there is attached a conditional probability table $P(A \mid B_1, \ldots, B_n)$.

**Admittance of d-separation in Bayesian networks**

If $A$ and $B$ are d-separated in a Bayesian network with evidence $e$ entered, then $P(A \mid B, e) = P(A \mid e)$.

**The chain rule**

Let $BN$ be a Bayesian network over $U = \{A_1, \ldots, A_m\}$. Then the joint probability distribution $P(U)$ is the product of all conditional probabilities specified in $BN$:

$$P(U) = \prod_i P(A_i \mid pa(A_i)),$$

where $pa(A_i)$ is the parent set of $A_i$.



## 2.7 Bibliographical notes

The two Examples 2.1.2 and 2.1.4 are inspired by Pearl (1988). The concepts of causal network, d-connection, and the definition in Section 2.2.1 are due to Pearl (1986b) and Verma (1987). A proof that Bayesian networks admit d-separation can be found in Pearl (1988) or in Lauritzen (1996). Bayesian networks have a long history in statistics, and in the first half of the 1980s they were introduced to the field of expert systems through work by Pearl (1982) and Spiegelhalter & knill-Jones (1984). BOBLO is documented in Rasmussen (1995a,b).

## Exercises

**Exercise 2.1** Show that d-connectedness is *symmetric* (if $A$ is d-connected to $B$, then $B$ is d-connected to $A$).

Give an example proving that d-connectedness is not *transitive* ($A$ d-connected to $B$ and $B$ d-connected to $C$, but $A$ and $C$ are not d-connected).

**Exercise 2.2** In the graphs below determine which variables are d-connected to $A$.

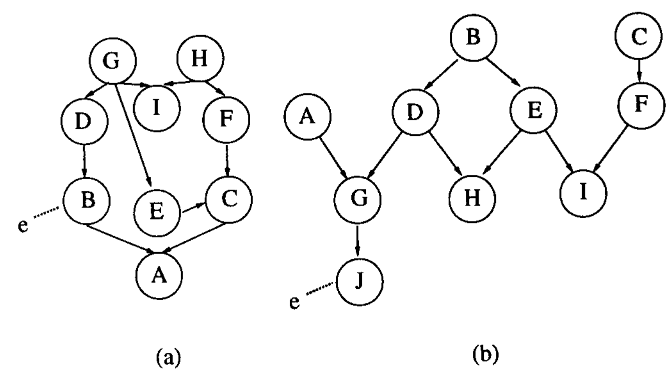


Figure for Exercise 2.2

**Exercise 2.3** Let $A$ be a variable in a DAG. Assume that the following variables are instantiated: the parents of $A$, the children of $A$, the spouses of $A$ (variables that share a child with $A$).

Show that $A$ is d-separated from the remaining uninstantiated variables.

**Exercise 2.4** Let $D_1$ and $D_2$ be DAGs over the same variables. $D_1$ is an *I-submap* of $D_2$ if all d-separation properties of $D_1$ also hold for $D_2$. If, also, $D_2$ is an I-submap of $D_2$, they are sid to be *I-equivalent*.

Which of the four DAGs in the figure below are I-equivalent?

**Table 2.14** Table for Exercise 2.5.

| | $b_1$ | $b_2$ | $b_3$ |
|---|---|---|---|
| $a_1$ | 0.05 | 0.10 | 0.05 |
| $a_2$ | 0.15 | 0.00 | 0.25 |
| $a_3$ | 0.10 | 0.20 | 0.10 |

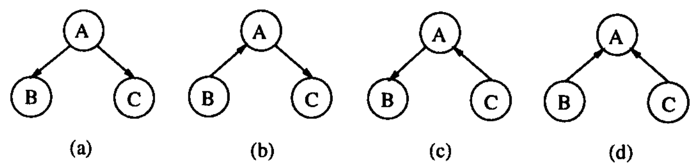


Figure for Exercise 2.4.

**Exercise 2.5** Calculate $P(A)$, $P(B)$, $P(A \mid B)$, and $P(B \mid A)$ from Table 2.14.

**Table 2.15** $P(A, B, C)$ for Exercise 2.6.

| | $b_1$ | $b_2$ |
|---|---|---|
| $a_1$ | (0.006, 0.054) | (0.048, 0.432) |
| $a_2$ | (0.014, 0.126) | (0.032, 0.288) |

**Table 2.16** Conditional probability tables for Exercise 2.7.

| | $a_1$ | $a_2$ | | $a_1$ | $a_2$ |
|---|---|---|---|---|---|
| $b_1$ | 0.2 | 0.3 | $c_1$ | 0.5 | 0.6 |
| $b_2$ | 0.8 | 0.7 | $c_2$ | 0.5 | 0.4 |
| $P(B \mid A)$ | | | $P(C \mid A)$ | | |

**Exercise 2.6** In Table 2.15, a joint probability table for the binary variables $A$, $B$, and $C$ is given.

(i) Calculate $P(B, C)$ and $P(B)$.

(ii) Are $A$ and $C$ independent given $B$?

**Exercise 2.7** The DAG (a) in Exercise 2.4 has $P(A) = (0.1, 0.9)$ and the conditional probability given in Table 2.16.

Calculate $P(A, B, C)$.

**Exercise 2.8** Perform a Bayesian calculation of the reasoning in Section 2.1.4 (earthquake or burglary). Use the probabilities in Table 2.17 and $P(B) = (0.01, 0.99)$, $P(E) = (0.001, 0.999)$.



**Table 2.17** Tables for Exercise 2.8. Probabilities for radio and alarm.

| | $E = y$ | $E = n$ |
|---|---|---|
| $R = y$ | 0.95 | 0.01 |
| $R = n$ | 0.05 | 0.99 |
| $P(R \mid E)$ | | |

| | $B = y$ | $B = n$ |
|---|---|---|
| $E = y$ | (0.98, 0.02) | (0.95, 0.05) |
| $E = n$ | (0.95, 0.05) | (0.03, 0.97) |
| $P(A \mid B, E)$ | | |

**Exercise 2.9** Let $P(c_i \mid b_j) \neq 0$ for all $i, j$. Prove that $A$ and $C$ are independent given $B$ if and only if $P(A, C \mid B) = P(A \mid B)P(C \mid B)$.