2

Causal and Bayesian Networks

In this chapter we introduce causal networks, which are the basic graphical feature for (almost) everything in this book. We give rules for reasoning about relevance in causal networks; is knowledge of A relevant for my belief about B? These sections deal with reasoning under uncertainty in general. Next, Bayesian networks are defined as causal networks with the strength of the causal links represented as conditional probabilities. Finally, the chain rule for Bayesian networks is presented. The chain rule is the property that makes Bayesian networks a very powerful tool for representing domains with inherent uncertainty. The sections on Bayesian networks assume knowledge of probability calculus as laid out in Sections 1.1–1.4.

2.1 Reasoning Under Uncertainty

2.1.1 Car Start Problem

The following is an example of the type of reasoning that humans do daily.

"In the morning, my car will not start. I can hear the starter turn, but nothing happens. There may be several reasons for my problem. I can hear the starter roll, so there must be power from the battery. Therefore, the most-probable causes are that the fuel has been stolen overnight or that the spark plugs are dirty. It may also be due to dirt in the carburetor, a loose connection in the ignition system, or something more serious. To find out, I first look at the fuel meter. It shows half full, so I decide to clean the spark plugs."

To have a computer do the same kind of reasoning, we need answers to questions such as, "What made me conclude that among the probable causes "stolen fuel", and "dirty spark plugs" are the two most-probable causes?" or "What made me decide to look at the fuel meter, and how can an observation concerning fuel make me conclude on the seemingly unrelated spark plugs?" To be more precise, we need ways of representing the problem and ways of performing inference in this representation such that a computer can simulate this kind of reasoning and perhaps do it better and faster than humans.

For propositional logic, Boolean logic is the representation framework, and various derived structures, such as truth tables and binary decision diagrams, have been invented together with efficient algorithms for inference.

In logical reasoning, we use four kinds of logical connectives: conjunction, disjunction, implication, and negation. In other words, simple logical statements are of the kind, "if it rains, then the lawn is wet," "both John and Mary have caught the flu," "either they stay at home or they go to the cinema," or "the lawn is not wet." From a set of logical statements, we can deduce new statements. From the two statements "if it rains, then the lawn is wet" and "the lawn is not wet," we can infer that it is not raining.

When we are dealing with uncertain events, it would be nice if we could use similar connectives with certainties rather than truth values attached, so we may extend the truth values of propositional logic to "certainties," which are numbers between 0 and 1. A certainty 0 means "certainly not true," and the higher the number, the higher the certainty. Certainty 1 means "certainly true."

We could then work with statements such as, "if I take a cup of coffee while on break, I will with certainty 0.5 stay awake during the next lecture"or "if I take a short walk during the break, I will with certainty 0.8 stay awake during the next lecture." Now, suppose I take a walk as well as have a cup of coffee. How certain can I be to stay awake? To answer this, I need a rule for how to *combine* certainties. In other words, I need a function that takes the two certainties 0.5 and 0.8 and returns a number, which should be the certainty resulting from combining the certainty from the two statements.

The same is needed for chaining: "if a then b with certainty x," and "if b then c with certainty y." I know a, so what is the certainty of c?

It has turned out that any function for combination and chaining will in some situations lead to wrong conclusions.

Another problem, which is also a problem for logical reasoning, is abduction: I have the rule "a woman has long hair with certainty 0.7." I see a long-haired person. What can I infer about the person's sex?

2.1.2 A Causal Perspective on the Car Start Problem

A way of structuring a situation for reasoning under uncertainty is to construct a graph representing causal relations between events.

Example 2.1 (A reduced Car Start Problem).

To simplify the situation, assume that we have the events $\{yes, no\}$ for *Fuel?*, $\{yes, no\}$ for *Clean Spark Plugs?*, $\{full, \frac{1}{2}, empty\}$ for *Fuel Meter*, and $\{yes,\ no\}$ for *Start?*. In other words, the events are clustered around variables, each with a set of outcomes, also called *states*. We know that the state of *Fuel?* and the state of *Clean Spark Plugs?* have a causal impact on the state of *Start?*. Also, the state of *Fuel?* has an impact on the state of *Fuel Meter Standing*. This is represented by the graph in Figure 2.1.

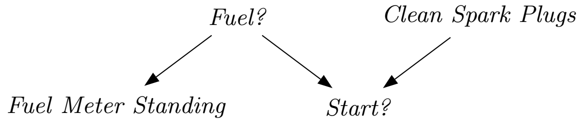

Fig. 2.1. A causal network for the reduced Car Start Problem.

If we add a direction from *no* to *yes* inside each variable (and from *empty* to *full*), we can also represent directions of the impact. For the present situation, we can say that all the impacts are positive (with the direction); that is, the more the certainty of the cause is moved in a positive direction, the more the certainty of the affected variable will also be moved in a positive direction. To indicate this, we can label the links with the sign "+" as is done in Figure 2.2.

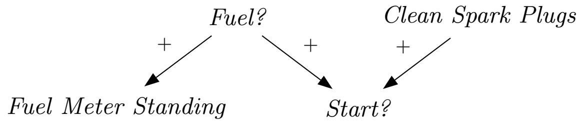

Fig. 2.2. A causal network for the reduced Car Start Problem with a sign indicating direction of impact.

We can use the graph in Figure 2.2 to perform some reasoning. Obviously, if I know that the spark plugs are not clean, then the certainty for no start will increase. However, my situation is the opposite. I realize that I have a start problem. As my certainty on *Start?* is moved in a negative direction, I find the possible causes (*Clean Spark Plugs?* and *Fuel?*) for such a move more certain; that is, the sign "+" is valid for both directions. Now, because the certainty on for *Fuel?* = *no* has increased, I will have a higher expectation that *Fuel Meter Standing* is in state *empty*.

The movement of the certainty for *Fuel Meter Standing* tells me that by reading the fuel meter I will get information related to the start problem. I read the fuel meter, it says $\frac{1}{2}$, and reasoning backward yields that the certainty on *Fuel?* is moved in a negative direction.

So far, the reasoning has been governed by simple rules that can easily be formalized. The conclusion is harder: "Lack of fuel does not seem to be the reason for my start problem, so most probably the spark plugs are not clean." Is there a formalized rule that allows this kind of reasoning on a causal network to be computerized? We will return to this problem in Section 2.2.

Note: The reasoning has focused on changes of certainty. In certainty calculus, if the actual certainty of a specific event must be calculated, then knowledge of certainties prior to any information is also needed. In particular, prior certainties are required for the events that are not effects of causes in the network. If, for example, my car cannot start, the actual certainty that the fuel has been stolen depends on my neighborhood.

2.2 Causal Networks and d-Separation

A causal network consists of a set of *variables* and a set of *directed links* (also called *arcs*) 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 propositions (or sample spaces), see also Section 1.3. A variable can have any number of states (or outcomes). A variable may, for example, be the color of a car (states *blue*, *green*, *red*, *brown*), the number of children in a specific 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 we consider only variables with a finite number of states (we shall return to the issue of continuous state spaces in Section 3.3.8).

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.

As illustrated in Section 2.1.2, causal networks can be used to follow how a change of certainty in one variable may change the certainty for other variables. We present in this section a set of rules for that kind of reasoning. The rules are independent of the particular calculus for uncertainty.

Serial Connections

Consider the situation in Figure 2.3. Here A has an influence on B, which in turn has an influence on C. Obviously, evidence about A will influence the certainty of B, which then influences the certainty of C. Similarly, evidence about C will influence the certainty of 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. When the state of a variable is known, we say that the variable is *instantiated.*

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

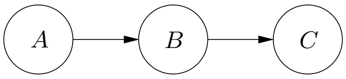

Fig. 2.3. Serial connection. When B is instantiated, it blocks communication between A and C.

Example 2.2. Figure 2.4 shows a causal model for the relations between *Rainfall* (*no*, *light*, *medium*, *heavy*), *Water level* (*low*, *medium*, *high*), and *Flooding* (*yes*, *no*). If I have not observed the water level, then knowing that there has been a flooding will increase my belief that the water level is high, which in turn will tell me something about the rainfall. The same line of reasoning holds in the other direction. On the other hand, if I already know the water level, then knowing that there has been flooding will not tell me anything new about rainfall.

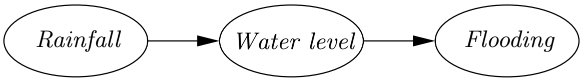

Fig. 2.4. A causal model for *Rainfall*, *Water level*, and *Flooding.*

Diverging Connections

The situation in Figure 2.5 is called a *diverging* connection. Influence can pass between all the children of A unless the state of A is known. That is, $B, C, \ldots, E$ are d-separated given A.

Evidence may be transmitted through a diverging connection unless it is instantiated.

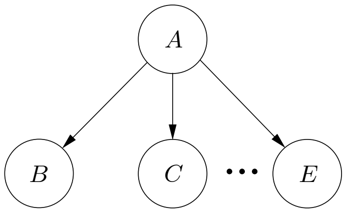

Fig. 2.5. Diverging connection. If A is instantiated, it blocks communication between its children.

Example 2.3. Figure 2.6 shows the causal relations between *Sex* (*male*, *female*), *length of hair* (*long*, *short*), and *stature* (<168 cm, ≥168 cm).

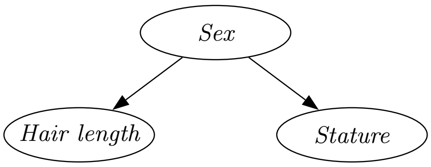

Fig. 2.6. Sex has an impact on length of hair as well as stature.

If we do not know the sex of a person, seeing the length of his/her hair will tell us more about the sex, and this in turn will focus our belief on his/her stature. On the other hand, if we know that the person is a man, then the length of his hair gives us no extra clue on his stature.

Converging Connections

A description of the situation in Figure 2.7 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 about one of them cannot influence the certainties of the others through A. Knowledge of one possible cause of an event does not tell us anything about the other possible causes. However, if anything is known about the consequences, then information on one possible cause may tell us something about the other causes. This is the *explaining away* effect illustrated in the car start problem: the car cannot start, and the potential causes include dirty spark plugs and an empty fuel tank. If we now get the information that there is fuel in the tank, then our certainty in the spark plugs being dirty will increase (since this will explain why the car cannot start). Conversely, if we get the information that there is no fuel on the car, then our certainty in the spark plugs being dirty will decrease (since the lack of fuel explains why the car cannot start). In Figure 2.8, two examples are shown. Observe that in the second example we observe only A indirectly through information about F; knowing the state of F tells us something about the state of E, which in turn tells us something about A.

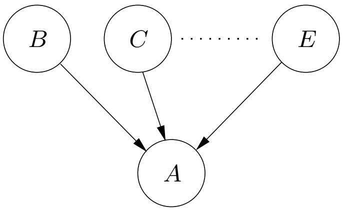

Fig. 2.7. Converging connection. If A changes certainty, it opens communication between its parents.

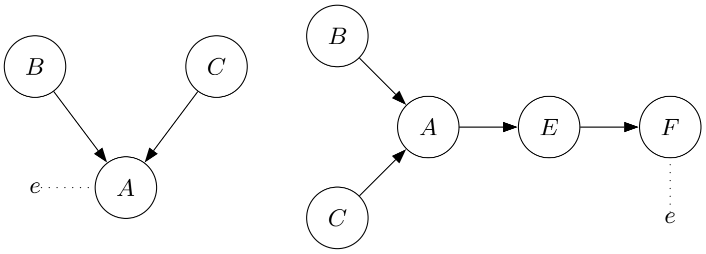

Fig. 2.8. Examples in which the parents of A are dependent. The dotted lines indicate insertion of evidence.

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

Remark: Evidence about a variable is a statement of the certainties of its states. If the variable is instantiated, we call it *hard* evidence; otherwise, it is called *soft*. In the example above, we can say that hard evidence about the variable F provides soft evidence about the variable A. Blocking in the case of serial and diverging connections requires hard evidence, whereas opening in the case of converging connections holds for all kinds of evidence.

Example 2.4. Figure 2.9 shows the causal relations among *Salmonella* infection, *flu*, *nausea*, and *pallor*.

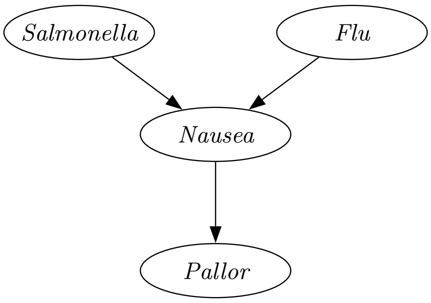

Fig. 2.9. Salmonella and flu may cause nausea, which in turn causes pallor.

If we know nothing of nausea or pallor, then the information on whether the person has a *Salmonella* infection will not tell us anything about flu. However, if we have noticed that the person is pale, then the information that he/she does not have a *Salmonella* infection will make us more ready to believe that he/she has the flu.

2.2.1 d-separation

The three preceding cases cover all 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 independent given the evidence entered into the network. The rules are formulated in the following definition.

Definition 2.1 (d-separation). *Two distinct variables A and B in a causal network are* d-separated *("d" for "directed graph") if for all paths between A and B, there is an intermediate variable V (distinct from A and B) such that either*

- *the connection is serial or diverging and V is instantiated*
  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 2.10 gives an example of a larger network. The evidence entered at B and M represents instantiations. 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 2.11 gives two other examples.

Note that although A and B are d-connected, changes in the belief in A will not necessarily change the belief in B. To stress this difference, we will sometimes say that A and B are *structurally independent* if they are d-separated (see also Exercise 2.23).

In connection to d-separation, a special set of nodes for a node A is the so-called *Markov blanket* for A:

Definition 2.2. *The* Markov blanket *of a variable A is the set consisting of the parents of A, the children of A, and the variables sharing a child with A.*

The Markov blanket has the property that when instantiated, A is d-separated from the rest of the network (see Figure 2.12).

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 of 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 should comply with the claim.

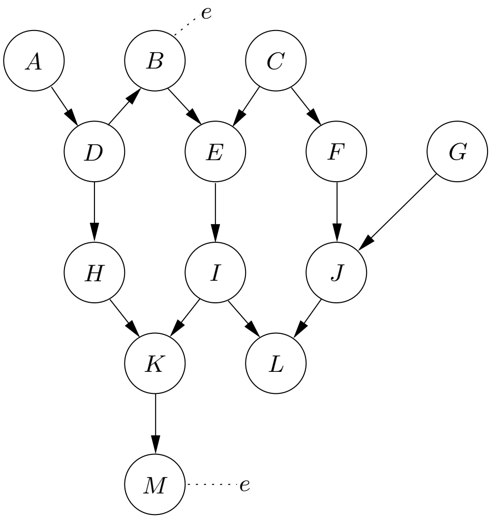

Fig. 2.10. A causal network with M and B instantiated. The node A is d-separated from G only.

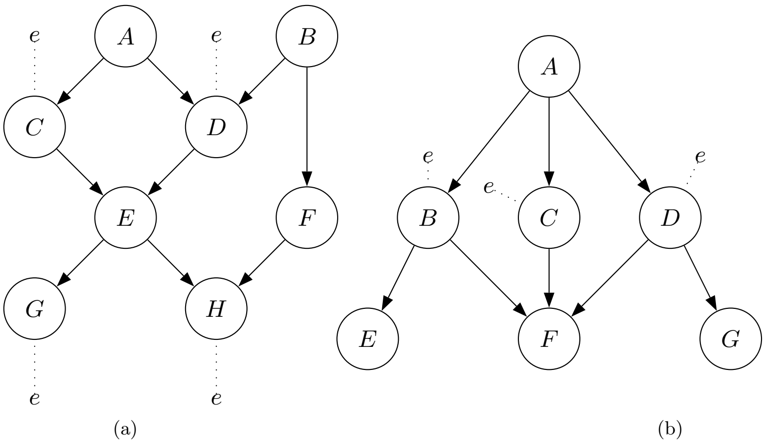

Fig. 2.11. Causal networks with hard evidence entered (the variables are instantiated). (a) Although all neighbors of E are instantiated, it is d-connected to F, B, and A. (b) F is d-separated from the remaining uninstantiated variables.

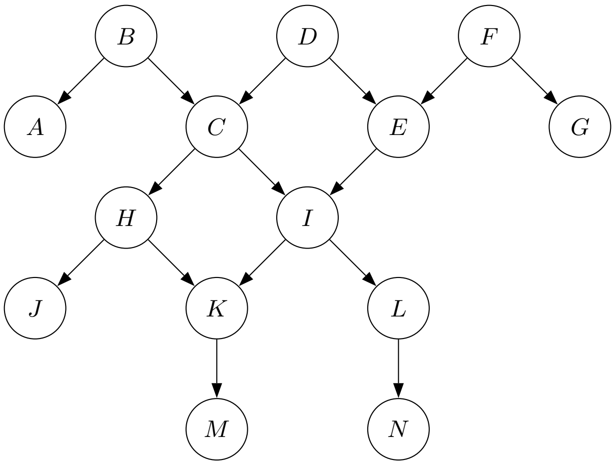

Fig. 2.12. The Markov blanket for I is $\{C, E, H, K, L\}$. Note that if only I's neighbors are instantiated, then J is not d-separated from I.

From the definition of d-separation we see that in order to test whether two variables, say A and B, are d-separated given hard evidence on a set of variables $\mathcal{C}$ you would have to check whether all paths connecting A and B are d-separating paths. An easier way of performing this test, without having to consider the various types of connections, is as follows: First you construct the so-called *ancestral graph* consisting of A, B, and $\mathcal{C}$ together with all nodes from which there is a directed path to either A, B, or $\mathcal{C}$ (see Figure 2.13(a)). Next, you insert an undirected link between each pair of nodes with a common child and then you make all links undirected. The resulting graph (see Figure 2.13(b)) is known as the *moral graph* for Figure 2.13(a). The moral graph can now be used to check whether A and B are d-separated given $\mathcal{C}$: if all paths connecting A and B intersect $\mathcal{C}$, then A and B are d-separated given $\mathcal{C}$.

The above procedure generalizes straightforwardly to the case in which we work with sets of variables rather than single variables: you just construct the ancestral graph using these sets of variables and perform the same steps as above: $\mathcal{A}$ and $\mathcal{B}$ are then d-separated given $\mathcal{C}$ if all paths connecting a variable in $\mathcal{A}$ with a variable in $\mathcal{B}$ intersect a variable in $\mathcal{C}$.

2.3 Bayesian Networks

2.3.1 Definition of Bayesian Networks

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

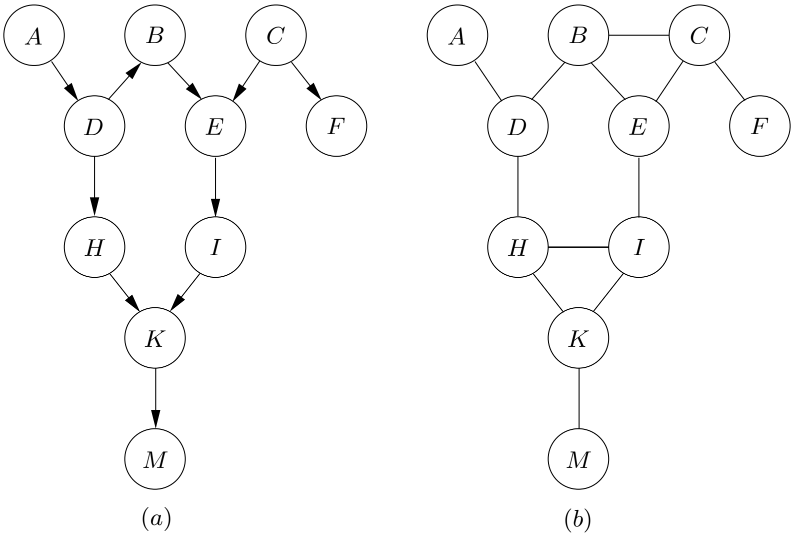

Fig. 2.13. To test whether A is d-separated from F given evidence on B and M in Figure 2.10, we first construct the ancestral graph for $\{A, B, F, M\}$ (figure (a)). Next we add an undirected link between pairs of nodes with a common child and then the direction is dropped on all links (figure (b)). In the resulting graph we have that the path $A - D - H - K - I - E - C - F$ does not intersect B and M, hence A and F are d-connected given B and M.

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 about how the impacts from A and C interact. They may cooperate or counteract in various ways, so we need a specification of $P(B \mid A, C)$.

It may happen that the domain to be modeled contains causal feedback cycles (see Figure 2.14).

Feedback cycles are difficult to model quantitatively. For causal networks, no calculus has been developed that can cope with feedback cycles, but certain noncausal models have been proposed to deal with this issue. For Bayesian networks we require that the network does not contain cycles.

Definition 2.3. *A* Bayesian network *consists of the following:*

- *A set of* variables[1] *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 an* acyclic directed graph *(traditionally abbreviated DAG); a directed graph is* acyclic *if there is no directed path $A_1 \rightarrow \cdots \rightarrow A_n$ so that $A_1 = A_n$.*

[1] When we wish to emphasize that this kind of variable represents a sample space we call it a *chance variable*.

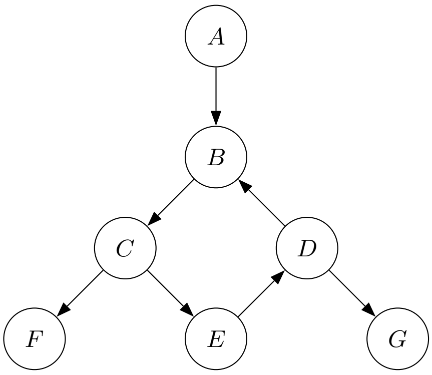

Fig. 2.14. A directed graph with a feedback cycle. This is not allowed in Bayesian networks.

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

Note that if A has no parents, then the table reduces to the unconditional probability table $P(A)$. For the DAG in Figure 2.15, the prior probabilities $P(A)$ 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 2.1.2, prior probabilities are necessary not for mathematical reasons but because prior certainty assessments are an integral part of human reasoning about certainty (see also Exercise 1.12).

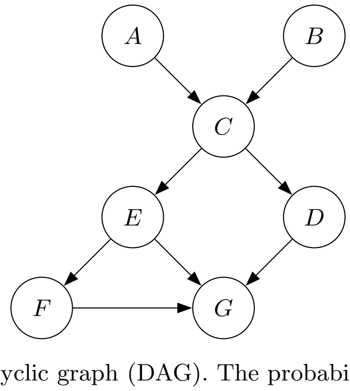

Fig. 2.15. A directed acyclic graph (DAG). The probabilities to specify are $P(A)$, $P(B)$, $P(C \,|\, A, B)$, $P(E \,|\, C)$, $P(D \,|\, C)$, $P(F \,|\, E)$, and $P(G \,|\, D, E, F)$.

The definition of Bayesian networks does not refer to causality, and there is no requirement that the links represent causal impact. That is, when building the structure of a Bayesian network model, we need not insist on having the links go in a causal direction. However, we then need to check the model's d-separation properties and ensure that they correspond to our perception of the world's conditional independence properties. The model should not include conditional independences that do not hold in the real world.

This also means that if A and B are d-separated given evidence e, then the probability calculus used for Bayesian networks must yield $P(A \mid e) = P(A \mid B, e)$ (see Section 2.3.2).

Example 2.5 (A Bayesian network for the Car Start Problem).

The Bayesian network for the reduced Car Start Problem is the one in Figure 2.16.

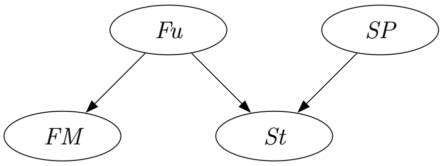

Fig. 2.16. The causal network for the reduced car start problem. We have used the abbreviations *Fu (Fuel?), SP (Clean Spark Plugs?), St (Start?)*, and *FM (Fuel Meter Standing).*

For the quantitative modeling, we need the probability assessments $P(\mathit{Fu})$, $P(\mathit{SP})$, $P(\mathit{St} \mid \mathit{Fu}, \mathit{SP})$, $P(\mathit{FM} \mid \mathit{Fu})$. To avoid having to deal with numbers that are too small, let $P(\mathit{Fu}) = (0.98, 0.02)$ and $P(\mathit{SP}) = (0.96, 0.04)$. The remaining tables are given in Table 2.1. Note that the table for $P(\mathit{FM} \mid \mathit{Fu})$ reflects the fact that the fuel meter may be malfunctioning, and the table for $P(\mathit{St} \mid \mathit{Fu}, \mathit{SP})$ leaves room for causes other than no fuel and dirty spark plugs by assigning $P(\mathit{St} = \mathit{no} \mid \mathit{Fu} = \mathit{yes}, \mathit{SP} = \mathit{yes}) > 0$.

2.3.2 The Chain Rule for Bayesian Networks

Let $\mathcal{U} = \{A_1, \ldots, A_n\}$ be a universe of variables. If we have access to the joint probability table $P(\mathcal{U}) = P(A_1, \ldots, A_n)$, then we can also calculate $P(A_i)$ as well as $P(A_i \mid e)$, where e is evidence about some of the variables in the Bayesian network (see, e.g., Section 1.3.1). However, $P(\mathcal{U})$ grows exponentially with the number of variables, and $\mathcal{U}$ need not be very large before the table becomes intractably large. Therefore, we look for a more compact *representation* of $P(\mathcal{U})$, i.e., a way of storing information from which $P(\mathcal{U})$ can be calculated if needed.

Let BN be a Bayesian network over $\mathcal{U}$, and let $P(\mathcal{U})$ be a probability distribution reflecting the properties specified by BN: (i) the conditional probabilities for a variable given its parents in $P(\mathcal{U})$ must be as specified in BN, and (ii) if the variables A and B are d-separated in BN given the set $\mathcal{C}$, then A and B are independent given $\mathcal{C}$ in $P(\mathcal{U})$.

	$Fu = yes$	$Fu = no$
$FM = full$	0.39	0.001
$FM = \frac{1}{2}$	0.60	0.001
$FM = empty$	0.01	0.998

$P(FM \mid Fu)$

	$Fu = yes$	$Fu = no$
$Sp = yes$	(0.99, 0.01)	(0,1)
$Sp = no$	(0.01, 0.99)	(0,1)

$P(St \mid Fu, Sp)$

Table 2.1. Conditional probabilities for the model in Figure 2.16. The numbers (x, y) in the lower table represent ($St = yes$, $St = no$).

Based on these two properties, what other properties can be deduced about $P(\mathcal{U})$? If the universe consists of only one variable A, then BN specifies $P(A)$, and $P(\mathcal{U})$ is uniquely determined. We shall show that this holds in general.

For probability distributions over sets of variables, we have an equation called *the chain rule*. For Bayesian networks this equation has a special form. First we state the general chain rule:

Proposition 2.1 (The general chain rule). *Let $\mathcal{U} = \{A_1, \ldots, A_n\}$ be a set of variables. Then for any probability distribution $P(\mathcal{U})$ we have*

$$P(\mathcal{U}) = P(A_n \mid A_1, \ldots, A_{n-1})P(A_{n-1} \mid A_1, \ldots, A_{n-2}) \ldots P(A_2 \mid A_1)P(A_1).$$

Proof. Iterative use of the fundamental rule:

$$\begin{aligned} P(\mathcal{U}) &= P(A_n \mid A_1, \ldots, A_{n-1})P(A_1, \ldots, A_{n-1}), \\ P(A_1, \ldots, A_{n-1}) &= P(A_{n-1} \mid A_1, \ldots, A_{n-2})P(A_1, \ldots, A_{n-2}), \\ &\vdots \\ P(A_1, A_2) &= P(A_2 \mid A_1)P(A_1). \end{aligned}$$

□

Theorem 2.1 (The chain rule for Bayesian networks). *Let BN be a Bayesian network over $\mathcal{U} = \{A_1, \ldots, A_n\}$. Then BN specifies a unique joint probability distribution $P(\mathcal{U})$ given by the product of all conditional probability tables specified in BN:*

$$P(\mathcal{U}) = \prod_{i=1}^{n} P(A_i \mid \mathrm{pa}(A_i)),$$

where $\mathrm{pa}(A_i)$ *are the parents of* A_i *in* BN*, and* $P(\mathcal{U})$ *reflects the properties of* BN*.*

Proof. First we should show that $P(\mathcal{U})$ is indeed a probability distribution. That is, we need to show that Axioms 1–3 hold. This is left as an exercise (see Exercise 2.15).

Next we prove that the specification of BN is consistent, so that $P(\mathcal{U})$ reflects the properties of BN. It is not hard to prove that the probability distribution specified by the product in the chain rule reflects the conditional probabilities from BN (see Exercise 2.16). We also need to prove that the product reflects the d-separation properties. This is done through induction in the number of variables in BN.

When BN has one variable, it is obvious that the d-separation properties specified by BN hold for the product of all specified conditional probabilities.

Assume that for any Bayesian network with $n-1$ variables and a distribution $P(\mathcal{U})$ specified as the product of all conditional probabilities, it holds that if A and B are d-separated given $\mathcal{C}$, then $P(A \,|\, B, \mathcal{C}) = P(A \,|\, \mathcal{C})$. Let BN be a Bayesian network with n variables $\{A_1, \ldots, A_n\}$. Assume that A_n has no children and let BN' be the result of removing A_n from BN. Clearly BN' is a Bayesian network with the same conditional probability distributions as BN (except for A_n) and with the same d-separation properties over $\{A_1, \ldots, A_{n-1}\}$ as BN. Moreover,

$$\begin{aligned}
P(\mathcal{U} \setminus \{A_n\}) &= \sum_{A_n} P(\mathcal{U}) = \sum_{A_n} \prod_{i=1}^{n} P(A_i \mid \mathrm{pa}(A_i)) \\
&= \prod_{i=1}^{n-1} P(A_i \mid \mathrm{pa}(A_i)) \sum_{A_n} P(A_n \mid \mathrm{pa}(A_n)) \\
&= \prod_{i=1}^{n-1} P(A_i \mid \mathrm{pa}(A_i)) \mathbf{1} = \prod_{i=1}^{n-1} P(A_i \mid \mathrm{pa}(A_i)),
\end{aligned}$$

and by the induction hypothesis $P(\mathcal{U} \setminus \{A_n\})$ reflects the properties of BN'. Now, if A and B are d-separated given $\mathcal{C}$ in BN, then they are also d-separated in BN', and therefore $P(A \,|\, B, \mathcal{C}) = P(A \,|\, \mathcal{C})$. To prove that it also holds for d-separation properties involving A_n, we consider the case in which $A_n \in \mathcal{C}$ and the case in which $A = A_n$. For the first case we have that since A_n participates only in a converging connection, it holds that if A and B are d-separated given $\mathcal{C}$, then they are also d-separated given $\mathcal{C} \setminus \{A_n\}$ and we get the situation above. For the second case, we first note that

$$P(A_n \,|\, B, \mathcal{C}) = \sum_{\mathrm{pa}(A_n)} P(A_n \,|\, B, \mathcal{C}, \mathrm{pa}(A_n)) P(\mathrm{pa}(A_n) \,|\, B, \mathcal{C}).$$

Now, if A_n and B are d-separated given $\mathcal{C}$, then $\mathrm{pa}(A_n)$ and B are also d-separated given $\mathcal{C}$, and since A_n is not involved, we have $P(\mathrm{pa}(A_n) \,|\, B, \mathcal{C}) =$

$P(\mathrm{pa}(A_n)\,|\,\mathcal{C})$. So we need to prove only that $P(A_n\,|\,B, \mathcal{C}, \mathrm{pa}(A_n)) = P(A_n \,|\, \mathrm{pa}(A_n))$. Using the fundamental rule and the chain rule, we get

$$\begin{aligned}
P(A_n\,|\,B,\mathcal{C},\mathrm{pa}(A_n)) &= \frac{P(A_n, B, \mathcal{C}, \mathrm{pa}(A_n))}{P(B, \mathcal{C}, \mathrm{pa}(A_n))} = \frac{\sum_{\mathcal{U}\setminus\{A_n,B,\mathcal{C},\mathrm{pa}(A_n)\}} P(\mathcal{U})}{\sum_{\mathcal{U}\setminus\{B,\mathcal{C},\mathrm{pa}(A_n)\}} P(\mathcal{U})} \\
&= \frac{\sum_{\mathcal{U}\setminus\{A_n,B,\mathcal{C},\mathrm{pa}(A_n)\}} \prod_{i=1}^{n} P(A_i \mid \mathrm{pa}(A_i))}{\sum_{\mathcal{U}\setminus\{B,\mathcal{C},\mathrm{pa}(A_n)\}} \prod_{i=1}^{n} P(A_i \mid \mathrm{pa}(A_i))} \\
&= \frac{P(A_n \mid \mathrm{pa}(A_n)) \sum_{\mathcal{U}\setminus\{A_n,B,\mathcal{C},\mathrm{pa}(A_n)\}} \prod_{i=1}^{n-1} P(A_i \mid \mathrm{pa}(A_i))}{\sum_{\mathcal{U}\setminus\{A_n,B,\mathcal{C},\mathrm{pa}(A_n)\}} \prod_{i=1}^{n-1} P(A_i \mid \mathrm{pa}(A_i)) \sum_{A_n} P(A_n \mid \mathrm{pa}(A_n))} \\
&= \frac{P(A_n \mid \mathrm{pa}(A_n)) \sum_{\mathcal{U}\setminus\{A_n,B,\mathcal{C},\mathrm{pa}(A_n)\}} \prod_{i=1}^{n-1} P(A_i \mid \mathrm{pa}(A_i))}{\sum_{\mathcal{U}\setminus\{A_n,B,\mathcal{C},\mathrm{pa}(A_n)\}} \prod_{i=1}^{n-1} P(A_i \mid \mathrm{pa}(A_i))\mathbf{1}} \\
&= P(A_n \mid \mathrm{pa}(A_n)).
\end{aligned}$$

To prove uniqueness, let $\{A_1, \ldots, A_n\}$ be a topological ordering of the variables. Then, for each variable A_i with parents $\mathrm{pa}(A_i)$ we have that A_i is d-separated from $\{A_1, \ldots, A_{i-1}\} \setminus \mathrm{pa}(A_i)$ given $\mathrm{pa}(A_i)$ (see Exercise 2.11). This means that for any distribution P reflecting the specifications by BN we must have $P(A_i\,|\,A_1, \ldots, A_{i-1}) = P(A_i \mid \mathrm{pa}(A_i)$. Substituting this in the general chain rule yields that any distribution reflecting the specifications by BN must be the product of the conditional probabilities specified in BN. □

The chain rule yields that a Bayesian network is a compact representation of a joint probability distribution. The following example illustrates how to exploit that for reasoning under uncertainty.

Example 2.6 (The Car Start Problem revisited).

In this example, we apply the rules of probability calculus to the Car Start Problem. This is done to illustrate that probability calculus can be used to perform the reasoning in the example, in particular, explaining away. In Chapter 4, we give general algorithms for probability updating in Bayesian networks. We will use the Bayesian network from Example 2.5 to perform the reasoning in Section 2.1.1.

We will use the joint probability table for the reasoning. The joint probability table is calculated from the chain rule for Bayesian networks,

$$P(Fu, FM, SP, St) = P(Fu)P(SP)P(FM\,|\,Fu)P(St\,|\,Fu, SP).$$

The result is given in Tables 2.2 and 2.3.

The evidence $St = no$ tells us that we are in the context of Table 2.3. By marginalizing FM and Fu out of Table 2.3 (summing each row), we get

$$P(SP, St = no) = (0.02864, 0.03965).$$

	$FM = \mathit{full}$	$FM = \frac{1}{2}$	$FM = \mathit{empty}$
$Sp = \mathit{yes}$	$(0.363, 0)$	$(0.559, 0)$	$(0.0093, 0)$
$Sp = \mathit{no}$	$(0.00015, 0)$	$(0.00024, 0)$	$(3.9 \cdot 10^{-6}, 0)$

Table 2.2. The joint probability table for $P(\mathit{Fu}, \mathit{FM}, \mathit{SP}, \mathit{St} = \mathit{yes})$.

	$FM = \mathit{full}$	$FM = \frac{1}{2}$	$FM = \mathit{empty}$
$Sp = \mathit{yes}$	$(0.00367, 1.9 \cdot 10^{-5})$	$(0.00564, 1.9 \cdot 10^{-5})$	$(9.4 \cdot 10^{-5}, 0.0192)$
$Sp = \mathit{no}$	$(0.01514, 8 \cdot 10^{-7})$	$(0.0233, 8 \cdot 10^{-7})$	$(0.000388, 0.000798)$

Table 2.3. The joint probability table for $P(\mathit{Fu}, \mathit{FM}, \mathit{SP}, \mathit{St} = \mathit{no})$. The numbers (x, y) in the table represent $(\mathit{Fu} = \mathit{yes}, \mathit{Fu} = \mathit{no})$.

We get the conditional probability $P(\mathit{SP} \mid \mathit{St} = \mathit{no})$ by dividing by $P(\mathit{St} = \mathit{no})$. This is easy, since $P(\mathit{St} = \mathit{no}) = P(\mathit{SP} = \mathit{yes}, \mathit{St} = \mathit{no}) + P(\mathit{SP} = \mathit{no}, \mathit{St} = \mathit{no}) = 0.02864 + 0.03965 = 0.06829$, and we get

$$P(\mathit{SP} \mid \mathit{St} = \mathit{no}) = \left(\frac{0.02864}{0.06829}, \frac{0.03965}{0.06829} \right) = (0.42, 0.58).$$

Another way of saying this is that the distribution we end up with will be a set of numbers that sum to 1. If they do not, normalize by dividing by the sum.

In the same way, we get $P(\mathit{Fu} \mid \mathit{St} = \mathit{no}) = (0.71, 0.29)$.

Next, we get the information that $\mathit{FM} = \frac{1}{2}$, and the context for calculation is limited to the part with $\mathit{FM} = \frac{1}{2}$ and $\mathit{St} = \mathit{no}$. The numbers are given in Table 2.4.

	$\mathit{Fu} = \mathit{yes}$	$\mathit{Fu} = \mathit{no}$
$Sp = \mathit{yes}$	0.00564	$1.9 \cdot 10^{-5}$
$Sp = \mathit{no}$	0.0233	$8 \cdot 10^{-7}$

Table 2.4. $P(\mathit{Fu}, \mathit{SP}, \mathit{St} = \mathit{no}, \mathit{FM} = \frac{1}{2})$.

By marginalizing Sp out and normalizing, we get $P(\mathit{Fu} | \mathit{St} = \mathit{no}, \mathit{FM} = \frac{1}{2}) = (0.999, 0.001)$, and by marginalizing Fu out and normalizing we get $P(\mathit{SP} | \mathit{St} = \mathit{no}, \mathit{FM} = \frac{1}{2}) = (0.196, 0.804)$. The probability of $\mathit{SP} = \mathit{yes}$ increased by observing $\mathit{FM} = \frac{1}{2}$, so the calculus did catch the explaining away effect.

2.3.3 Inserting Evidence

Bayesian networks are used for calculating new probabilities when you get new information. The information so far has been of the type "$A = a$," where A is a variable and a is a state of A. Let A have n states with $P(A) = (x_1, \ldots, x_n)$, and assume that we get the information e that A can be only in state i or j. This statement expresses that all states except i and j are impossible, and we have the probability distribution $P(A, e) = (0, \ldots, 0, x_i, 0, \ldots, 0, x_j, 0, \ldots, 0)$. Note that $P(e)$, the prior probability of e, is obtained by marginalizing A out of $P(A, e)$. Note also that $P(A, e)$ is the result of multiplying $P(A)$ by $(0, \ldots, 0, 1, 0, \ldots, 0, 1, 0, \ldots, 0)$, where the 1's are at the i'th and j'th places.

Definition 2.4. *Let A be a variable with n states. A* finding *on A is an n-dimensional table of zeros and ones.*

To distinguish between the statement e, "A is in either state i or j," and the corresponding 0/1-finding vector, we sometimes use the boldface notation $\mathbf{e}$ for the finding. Semantically, a finding is a statement that certain states of A are impossible.

Now, assume that you have a joint probability table, $P(\mathcal{U})$, and let $\mathbf{e}$ be the preceding finding. The joint probability table $P(\mathcal{U}, e)$ is the table obtained from $P(\mathcal{U})$ by replacing all entries with A not in state i or j by the value zero and leaving the other entries unchanged. This is the same as multiplying $P(\mathcal{U})$ by $\mathbf{e}$,

$$P(\mathcal{U}, e) = P(\mathcal{U}) \cdot \mathbf{e}.$$

Note that $P(e) = \sum_U P(\mathcal{U}, e) = \sum_U (P(\mathcal{U}) \cdot \mathbf{e})$. Using the chain rule for Bayesian networks, we have the following theorem.

Theorem 2.2. *Let BN be a Bayesian network over the universe $\mathcal{U}$, and let $\mathbf{e}_1, \ldots, \mathbf{e}_m$ be findings. Then*

$$P(\mathcal{U}, e) = \prod_{A \in \mathcal{U}} P(A \,|\, pa(A)) \cdot \prod_{i=1}^{m} \mathbf{e}_i,$$

and for $A \in \mathcal{U}$ we have

$$P(A \,|\, e) = \frac{\sum_{\mathcal{U} \setminus \{A\}} P(\mathcal{U}, e)}{P(e)}.$$

Some types of evidence cannot be represented as findings. You may, for example, receive a statement from someone that the chance of A being in state a_1 is twice as high as for a_2. This type of evidence is called *likelihood evidence*. It is possible to treat this kind of evidence in Bayesian networks. The preceding statement is then represented by the distribution $(0.67, 0.33)$, and Theorem 2.2 still holds. However, because it is unclear what it means that a likelihood statement is true, $P(e)$ cannot be interpreted as the probability of the evidence, and $P(\mathcal{U}, e)$ therefore has an unclear semantics. We will not deal further with likelihood evidence.

2.3.4 Calculating Probabilities in Practice

As described in Section 2.3.3 and illustrated in Example 2.6, probability updating in Bayesian networks can be performed using the chain rule to calculate $P(\mathcal{U})$, the joint probability table of the universe. However, $\mathcal{U}$ need not be large before $P(\mathcal{U})$ becomes intractably large. In this section, we illustrate how the calculations can be performed without having to deal with the full joint table. In Chapter 4, we give a detailed treatment of algorithms for probability updating.

Consider the Bayesian network in Figure 2.17, and assume that all variables have ten states. Assume that we have the evidence $e = \{D = d, F = f\}$, and we wish to calculate $P(A \mid e)$.

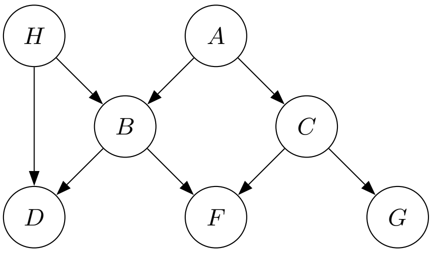

Fig. 2.17. A Bayesian network.

From the chain rule we have

$$\begin{aligned} P(\mathcal{U}, e) &= P(A, B, C, d, f, G, H) \\ &= P(A)P(H)P(B \mid A, H)P(C \mid A)P(d \mid B, H)P(f \mid B, C)P(G \mid C), \end{aligned}$$

where for example $P(d \mid B, H)$ denotes the table over B and H resulting from fixing the D-entry to the state d. We say that the conditional probability table has been *instantiated* to $D = d$. Notice that we need not calculate the full table $P(\mathcal{U})$ with 10^7 entries. If we wait until evidence is entered, we will in this case need to work with a table with only 10^5 entries. Later, we see that we need not work with tables larger than 1000 entries.

To calculate $P(A, e)$, we marginalize the variables B, C, G, and H out of $P(A, B, C, d, f, G, H)$. The order in which we marginalize does not affect the result (Section 1.4), so let us start with G; that is, we wish to calculate

$$\begin{aligned} &\sum_G P(A, B, C, d, f, G, H) \\ &\qquad = \sum_G P(A)P(H)P(B \mid A, H)P(C \mid A)P(d \mid B, H)P(f \mid B, C)P(G \mid C). \end{aligned}$$

In the right-hand product, only the last table contains G in its domain, and due to the distributive law (Section 1.4) we have

$$\begin{aligned}&\sum_G P(A,B,C,d,f,G,H)\\&\qquad = P(A)P(H)P(B \mid A,H)P(C \mid A)P(d \mid B,H)P(f \mid B,C)\sum_G P(G \mid C),\end{aligned}$$

and we need only calculate $\sum_G P(G \mid C)$. Actually, for each state c of C, we have $\sum_G P(G \mid c) = 1$; hence no calculations are necessary. We therefore get

$$\begin{aligned}P(A,B,C,d,f,H) &= \sum_G P(A,B,C,d,f,G,H)\\&= P(A)P(H)P(B \mid A,H)P(C \mid A)P(d \mid B,H)P(f \mid B,C).\end{aligned}$$

Next, we marginalize H out. Using the distributive law again, we get

$$\begin{aligned}&\sum_H P(A,B,C,d,f,H)\\&\qquad = P(A)P(C \mid A)P(f \mid B,C)\sum_H P(H)P(B \mid A,H)P(d \mid B,H).\end{aligned}$$

We multiply the three tables $P(H), P(B \mid A,H)$, and $P(d \mid B,H)$, and we marginalize H out of the product. The result is a table $T(d,B,A)$, and we have

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

Finally, we calculate this product and marginalize B and C out of it.

Notice that we never work with a table of more than three variables (the table produced by multiplying $P(H)$, $P(B \mid A,H)$, and $P(d \mid B,H)$) compared to the five variables in $P(A,B,C,d,f,G,H)$.

The method we just used is called *variable elimination* and can be described in the following way: we start with a set $\mathcal{T}$ of tables, and whenever we wish to marginalize a variable X, we take from $\mathcal{T}$ all tables with X in their domains, calculate the product of them, marginalize X out of it, and place the resulting table in $\mathcal{T}$.

2.4 Graphical Models – Formal Languages for Model Specification

From a mathematical point of view, the basic property of Bayesian networks is the chain rule: a Bayesian network is a compact representation of the joint probability table over its universe. In this respect, a Bayesian network is one type of compact representation among many others. However, there is more to it than this: From a knowledge engineering point of view, a Bayesian network is a type of *graphical model*. The structure of the network is formulated in a graphical communication language for which the language features have a very simple semantics, namely causality. This does not mean that "causality" is an easy concept. It may be very difficult to experience causality, and philosophically the concept is not fully understood. However, most often humans can communicate sensibly about causal relations in a knowledge domain. Furthermore, the graphical specification also specifies the requirements for the quantitative part of the model (the conditional probabilities). In Chapter 3, we extend the modeling language, and in Part II we present other types of graphical models.

As mentioned, graphical models are communication languages. They consist of a qualitative part, where features from graph theory are used, and a quantitative part consisting of *potentials*, which are real-valued functions over sets of nodes from the graph; in Bayesian networks the potentials are conditional probability tables. The graphical part specifies the kind of potentials and their domains.

Graphical models can be used for interpersonal communication: The graphical specification is easy for humans to read, and it helps focus attention, for example in a group working jointly on building a model. For interpersonal communication, the semantics of the various graph-theoretic features must be rather welldefined if misunderstandings are to be avoided.

The next step in the use of graphical models has to do with communication to a computer. You wish to communicate a graphical model to a computer, and the computer should be able to process the model and give answers to various queries. In order to achieve this, the specification language must be formally defined with a well-defined syntax and semantics.

The first concern in constructing a graphical modeling language is to ensure that it is sufficiently welldefined so that it can be communicated to a computer. This covers the graphical part as well as the specification of potentials. The next concern is the scope of the language: what is the range of domains and tasks that you will be able to model with this language? The final concern is tractability: do you have algorithms such that in reasonable time the computer can process a model and query to provide answers?

The Bayesian network is a sufficiently welldefined language, and behind the graphical specification in the user interface, the computer systems for processing Bayesian networks have an alphanumeric specification language, which for some systems is open to the user. Actually, the language for Bayesian networks is a context-free language with a single context-sensitive aspect (no directed cycles).

The scope of the Bayesian network language is hard to define, but the examples in the next chapter show that it has a very broad scope.

Tractability is not a yes or no issue. As described in Chapter 4, there are algorithms for probability updating in Bayesian networks, but basically probability updating is NP-hard. This means that some models have an updating time exponential in the number of nodes.

On the other hand, the running times of the algorithms can be easily calculated without actually running them. In Chapter 4 and Part II, we treat complexity issues for the various graphical languages presented.

2.5 Summary

d-Separation in Causal Networks

Two distinct 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 (distinct from A and B) such that either

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

Definition of Bayesian Networks

A Bayesian network consists of the following:

- There is 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 an *acyclic directed graph* (DAG).
- To each variable A with parents $B_1, \ldots, B_n$ there is attached a conditional probability table $P(A \mid B_1, \ldots, B_n)$.

The Chain Rule for Bayesian Networks

Let BN be a Bayesian network over $\mathcal{U} = \{A_1, \ldots, A_n\}$. Then BN specifies a unique joint probability distribution $P(\mathcal{U})$ given by the product of all conditional probability tables specified in BN:

$$P(\mathcal{U}) = \prod_{i=1}^{n} P(A_i \mid \mathrm{pa}(A_i)),$$

where $\mathrm{pa}(A_i)$ are the parents of A_i in BN, and $P(\mathcal{U})$ reflects the properties of BN.

Admittance of d-Separation in Bayesian Networks

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

Inserting Evidence

Let $\mathbf{e}_1, \ldots, \mathbf{e}_m$ be findings, and then

$$P(\mathcal{U}, e) = \prod_{i=1}^{n} P(A_i \mid \mathrm{pa}(A_i)) \prod_{j=1}^{m} \mathbf{e}_j$$

and

$$P(A \,|\, e) = \frac{\sum_{\mathcal{U} \setminus \{A\}} P(\mathcal{U}, e)}{P(e)}.$$

2.6 Bibliographical Notes

The connection between causation and conditional independence was studied by Spohn (1980), and later investigated with special focus on Bayesian networks in (Pearl, 2000). The concepts of causal network, d-connection, and the definition in Section 2.2.1 are due to Pearl (1986) and Verma (1987). A proof that Bayesian networks admit d-separation can be found in (Pearl, 1988) or in (Lauritzen, 1996). Geiger and Pearl (1988) proved that d-separation is the correct criterion for directed graphical models, in the sense that for any DAG, a probability distribution can be found for which the d-separation criterion is sound and complete. Meek (1995) furthermore proved that for a given DAG, the set of discrete probability distributions for which the d-separation criterion is not complete has measure zero. That is, given a random Bayesian network, there is almost no chance that it contains conditionally independent variables that cannot be read off the graph by d-separation. The method for discovering d-separation properties using ancestral graphs was first presented in (Lauritzen *et al.*, 1990).

Bayesian networks have a long history in statistics, and can be traced back at least to the work in (Minsky, 1963). In the first half of the 1980s they were introduced to the field of expert systems through work by Pearl (1982) and Spiegelhalter and Knill-Jones (1984). Some of the first real-world applications of Bayesian networks were Munin (Andreassen *et al.*, 1989, 1992) and Pathfinder (Heckerman *et al.*, 1992). The basis for the inference method presented in Section 2.3.4 originates from (D'Ambrosio, 1991) and was modified to the presented variable elimination in (Dechter, 1996). The fact that inference is NP-hard was proved in (Cooper, 1987).

2.7 Exercises

Exercise 2.1. To illustrate that simple rules cannot cope with uncertainty reasoning, consider the following two cases:

(i) I have an urn with a red ball and a white ball in it. If I add a red ball and shake it, what is the certainty of drawing a red ball in one draw? If I add a white ball instead, what is the certainty of drawing a red ball? If I combine the two actions, what is the certainty of drawing a red ball?
(ii) When shooting, I am more certain to hit the target if I close the left eye. I am also more certain to hit the target if I close the right eye. What is the combined certainty if I do both?

Exercise 2.2. Construct a causal network and follow the reasoning in the following story. Mr. Holmes is working in his office when he receives a phone call from his neighbor, who tells him that Holmes' burglar alarm has gone off. Convinced that a burglar has broken into his house, Holmes rushes to his car and heads for home. On his way, he listens to the radio, and in the news it is reported that there has been a small earthquake in the area. Knowing that earthquakes have a tendency to turn on burglar alarms, he returns to work.

Exercise 2.3. Consider the Car Start Problem in Section 2.1.1 with the causal network in Figure 2.1, and the following twist on the story: "I distinctly remember visiting the pump last night, so the fuel meter should be reading *full*. Since this is not the case, either there must be a leak in the tank, someone has stolen gasoline during the night, or the fuel meter is malfunctioning. Sniffing the air I smell no gasoline, so I conclude that a thief has been visiting last night or that the fuel meter is malfunctioning." Alter the causal network in Figure 2.1 to incorporate the above twist on the story.

Exercise 2.4. In the graphs in Figures 2.18 and 2.19, determine which variables are d-separated from A.

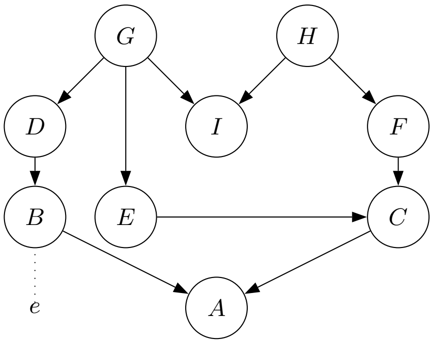

Fig. 2.18. Figure for Exercise 2.4.

Exercise 2.5. For each pair of variables in the causal network in Figure 2.1, state whether the variables can be d-separated, and if so which set(s) of variables that allow this.

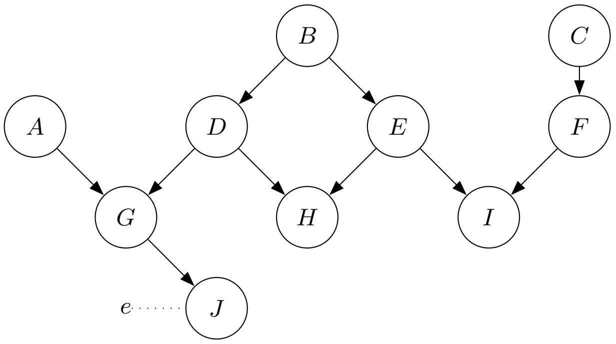

Fig. 2.19. Figure for Exercise 2.4.

Exercise 2.6. Consider the network in Figure 2.20. What are the minimal set(s) of variables required to d-separate C and E (that is, sets of variables for which no proper subset d-separates C and E)? What are the minimal set(s) of variables required to d-separate A and B? What are the maximal set(s) of variables that d-separate C and E (that is, sets of variables for which no proper superset d-separates C and E)? What are the maximal set(s) of variables that d-separate A and B?

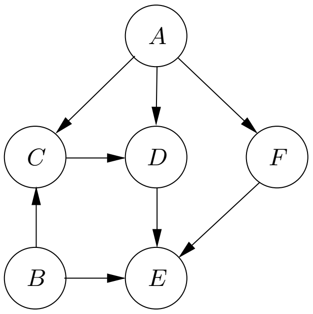

Fig. 2.20. A causal network for Exercise 2.6.

Exercise 2.7. Consider the network in Figure 2.20. What is the Markov blanket of each variable?

Exercise 2.8. Let A be a variable in a DAG. Assume that all variables in A's Markov blanket are instantiated. Show that A is d-separated from the remaining uninstantiated variables.

Exercise 2.9. Apply the procedure using the ancestral graph given in Section 2.2.1 to determine whether A is d-separated from C given B in the network in Figure 2.19.

Exercise 2.10. Let D_1 and D_2 be DAGs over the same variables. The graph D_1 is an *I-submap* of D_2 if all d-separation properties of D_1 also hold for D_2. If D_2 is also an I-submap of D_1, they are said to be *I-equivalent*. Which of the four DAGs in Figure 2.21 are I-equivalent?

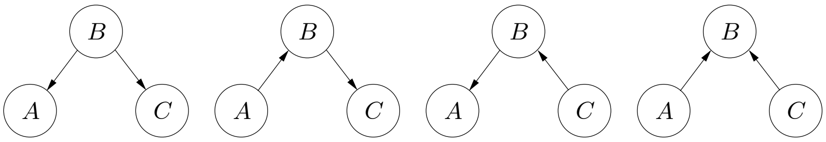

Fig. 2.21. Figure for Exercise 2.10.

Exercise 2.11. Let $\{A_1, \ldots, A_n\}$ be a topological ordering of the variables in a Bayesian network, and consider variable A_i with parents $\text{pa}(A_i)$. Prove that A_i is d-separated from $\{A_1, \ldots, A_{i-1}\} \setminus \text{pa}(A_i)$ given $pa(A_i)$.

Exercise 2.12. Consider the network in Figure 2.20. Which conditional probability tables must be specified to turn the graph into a Bayesian network?

Exercise 2.13. In Figure 2.22 the structure of a simple Bayesian network is shown. The accompanying conditional probability tables are shown in Tables 2.5 and 2.6, and the prior probabilities for A are 0.9 and 0.1. Are A and C d-separated given B? Are A and C conditionally independent given B?

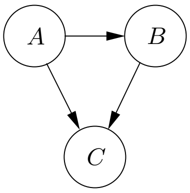

Fig. 2.22. A simple Bayesian network for Exercise 2.13.

	$A = a_1$	$A = a_2$
$B = b_1$	0.3	0.6
$B = b_2$	0.7	0.4

Table 2.5. $P(B \mid A)$.

	$A = a_1$	$A = a_2$
$B = b_1$	(0.1 ; 0.9)	(0.1 ; 0.9)
$B = b_2$	(0.2 ; 0.8)	(0.2 ; 0.8)

Table 2.6. $P(C \mid A, B)$.

Exercise 2.14. Consider the network in Figure 2.20. Using the chain rule, establish an expression for the joint distribution over the universe $\{A, B, C, D, E, F\}$. Use this expression to show that B and D are conditionally independent given A and C.

Exercise 2.15. Prove that the probability distribution $P(\mathcal{U})$ defined by the chain rule for Bayesian networks is indeed a probability distribution.

Exercise 2.16. Prove that the probability distribution $P(\mathcal{U})$ defined by the chain rule for a Bayesian network BN reflects the conditional probabilities specified in BN.

Exercise 2.17. Consider the Bayesian network from Exercise 2.13 and the finding $e = (0, 1)$ over A. What is $P(B, C, e)$?

Exercise 2.18. What steps would be taken if variable elimination were used to calculate the probability table $P(F \mid C = c_1)$ for the network in Figure 2.20? Assuming that each variable has ten states, what is the maximum size of a table during the procedure?

Exercise 2.19. Consider the DAG (a) in Exercise 2.10.

- Show that $P(B \mid A, C) = P(B \mid A)$.
- We have $P(A) = (0.1, 0.9)$ and the conditional probability tables in Table 2.7. Calculate $P(A, B, C)$.

	a_1	a_2
b_1	0.2	0.3
b_2	0.8	0.7

$P(B \mid A)$

	a_1	a_2
c_1	0.5	0.6
c_2	0.5	0.4

$P(C \mid A)$

Table 2.7. Conditional probability tables for Exercise 2.19.

Exercise 2.20. ^{E} Install an editor for Bayesian networks (a reference to a list of systems can be found in the preface).

Exercise 2.21. ^{E} Construct a Bayesian network for Exercise 1.12.

Exercise 2.22. ^{E} Construct a Bayesian network to follow the reasoning from Exercise 2.2. Use your own estimates of probabilities for the network.

Exercise 2.23. ^{E} Consider the Bayesian network in Figure 2.23 with conditional probabilities given in Table 2.8. Use your system to investigate whether A and C are independent.

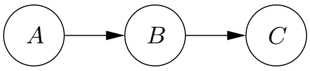

Fig. 2.23. Figure for Exercise 2.23.

	$A = yes$	$A = no$
b_1	0.6	0.2
b_2	0.1	0.5
b_3	0.2	0.1
b_4	0.1	0.2

$P(B \mid A)$

	b_1	b_2	b_3	b_4
$C = yes$	0.8	0.8	0.2	0.2
$C = no$	0.2	0.2	0.8	0.8

$P(C \mid B)$

Table 2.8. Tables for Exercise 2.23.

Exercise 2.24. ^{E} Use your system and Section 2.5 to perform the reasoning in Section 2.1.2.

Springer

http://www.springer.com/978-0-387-68281-5

Bayesian Networks and Decision Graphs
Nielsen, Th.D.; VERNER JENSEN, F.
2007, XVI, 448 p., Hardcover
ISBN: 978-0-387-68281-5