Chapter 5

Foundations of Bayesian Networks

The Reverend Thomas Bayes (1702–1761) developed Bayes' Theorem in the eighteenth century. Since that time the theorem has had a great impact on statistical inference because it enables us to infer the probability of a cause when its effect is observed. In the 1980s, the method was extended to model the probabilistic relationships among many causally related variables. The graphical structures that describe these relationships have come to be known as *Bayesian networks*. This chapter introduces these networks. (Applications of Bayesian networks to bioinformatics appear in Part III.) In Sections 5.1 and 5.2 we define Bayesian networks and discuss their properties. Section 5.3 shows how causal graphs often yield Bayesian networks. In Section 5.4 we discuss doing probabilistic inference using Bayesian networks. Section 5.5 introduces Bayesian networks containing continuous variables.

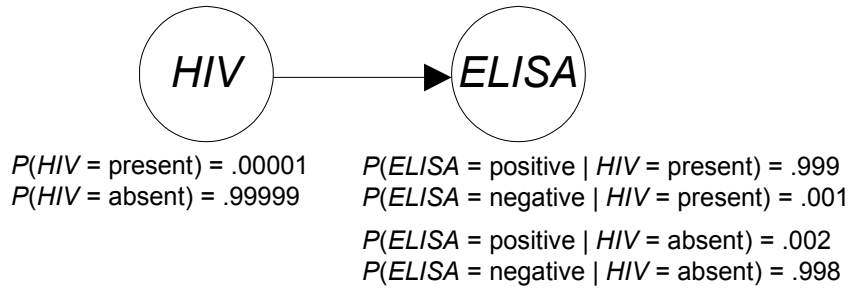

Figure 5.1: A two-node Bayesian network.

5.1 What Is a Bayesian Network?

Recall that in Example 2.30, we computed the probability of Joe having the HIV virus, given that he tested positive for it using the ELISA test. Specifically, we knew that

$$P(ELISA = positive|HIV = present) = .999$$

$$P(ELISA = positive|HIV = absent) = .002,$$

and

$$P(HIV = present) = .00001.$$

We then employed Bayes' Theorem to compute

$$\begin{aligned} & P(present|positive) \\ = \; & \frac{P(positive|present)P(present)}{P(positive|present)P(present) + P(positive|absent)P(absent)} \\ = \; & \frac{(.999)(.00001)}{(.999)(.00001) + (.002)(.99999)} \\ = \; & .00497. \end{aligned}$$

We summarize the information used in this computation in Figure 5.1, which is a two-node/variable Bayesian network. Notice that it represents the random variables HIV and $ELISA$ by nodes in a directed acyclic graph (DAG) and the causal relationship between these variables with an edge from HIV to $ELISA$. That is, the presence of HIV has a causal effect on whether the test result is positive; so there is an edge from HIV to $ELISA$. Besides showing a DAG representing the causal relationships, Figure 5.1 shows the prior probability distribution of HIV and the conditional probability distribution of $ELISA$ given each value of its parent HIV. In general, Bayesian networks consist of a DAG, whose edges represent relationships among random variables that are often (but not always) causal; the prior probability distribution of every variable that is a root in the DAG; and the conditional probability distribution of every

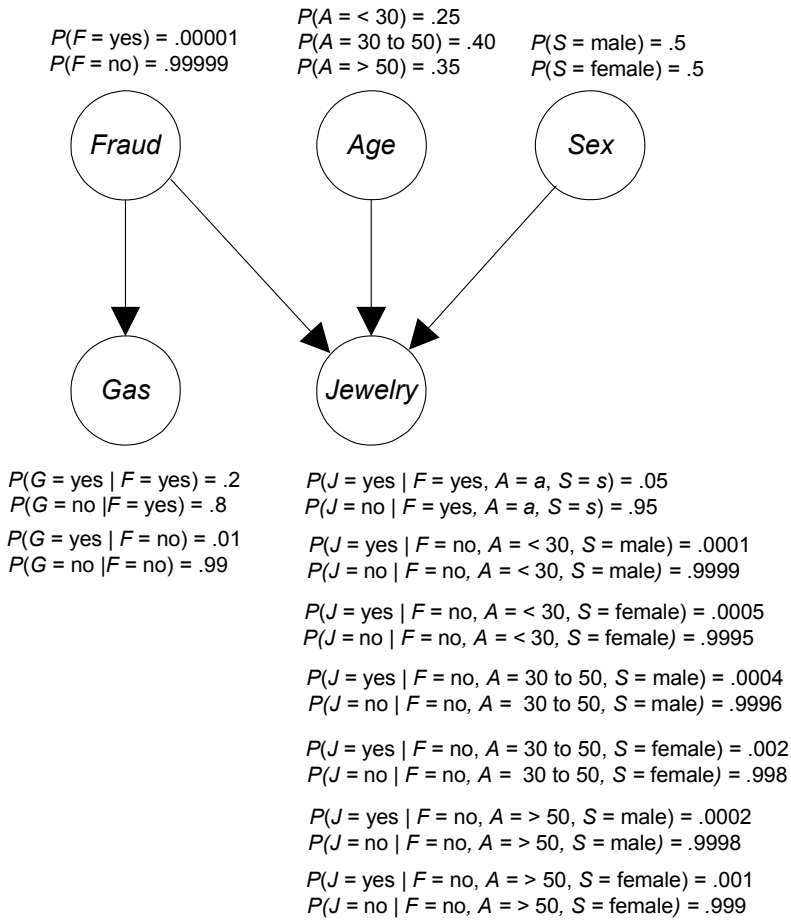

Figure 5.2: Bayesian network for detecting credit card fraud.

non-root variable given each set of values of its parents. We use the terms *node* and *variable* interchangeably in discussing Bayesian networks.

Let's illustrate a more complex Bayesian network by considering the problem of detecting credit card fraud (taken from [Heckerman, 1996]). Suppose that we have identified the following variables as being relevant to the problem:

Variable	What the Variable Represents
Fraud (F)	Whether the current purchase is fraudulent
Gas (G)	Whether gas has been purchased in the last 24 hours
Jewelry (J)	Whether jewelry has been purchased in the last 24 hours
Age (A)	Age of the card holder
Sex (S)	Sex of the card holder

These variables are all causally related. That is, a credit card thief is likely to

buy gas and jewelry, and middle-aged women are most likely to buy jewelry, whereas young men are least likely to buy jewelry. Figure 5.2 shows a DAG representing these causal relationships. Notice that it also shows the conditional probability distribution of every non-root variable given each set of values of its parents. The Jewelry variable has three parents, and there is a conditional probability distribution for every combination of values of those parents. The DAG and the conditional distributions together constitute a Bayesian network.

You could have a few questions concerning this Bayesian network. First you might ask, "What value does it have?" That is, what useful information can we obtain from it? Recall how we used Bayes' Theorem to compute $P(HIV = present|ELISA = positive)$ from the information in the Bayesian network in Figure 5.1. Similarly, we can compute the probability of credit card fraud given values of the other variables in this Bayesian network. For example, we can compute $P(F = yes|G = yes, J = yes, A = < 30, S = female)$. If this probability is sufficiently high, we can deny the current purchase or require additional identification. The computation is not a simple application of Bayes' Theorem as was the case for the two-node Bayesian network in Figure 5.1. Rather it is done using sophisticated algorithms.

Second, you might ask how we obtained the probabilities in the network. They can either be obtained from the subjective judgements of an expert in the area or be learned from data. (In Chapter 8 we discuss techniques for learning them from data.)

Finally, you could ask why we are bothering to include the variables for age and sex in the network when the age and sex of the card holder has nothing to do with whether the card has been stolen (fraud). That is, fraud has no causal effect on the card holder's age or sex, and vice versa. The reason we include these variables is quite subtle. It is because fraud, age, and sex all have a common effect, namely the purchasing of jewelry. So, when we know jewelry has been purchased, the three variables are rendered probabilistically dependent owing to what psychologists call **discounting**. For example, if jewelry has been purchased in the last 24 hours, it increases the likelihood of fraud. However, if the card holder is a middle-aged woman, the likelihood of fraud is lessened (discounted) because such women are prone to buying jewelry. That is, the fact that the card holder is a middle-aged woman explains the jewelry purchase. On the other hand, if the card holder is a young man, the likelihood of fraud is increased because such men are unlikely to purchase jewelry.

We have informally introduced Bayesian networks, their properties, and their usefulness. Next we formally develop their mathematical properties.

5.2 Properties of Bayesian Networks

After defining Bayesian networks, we show how they are ordinarily represented.

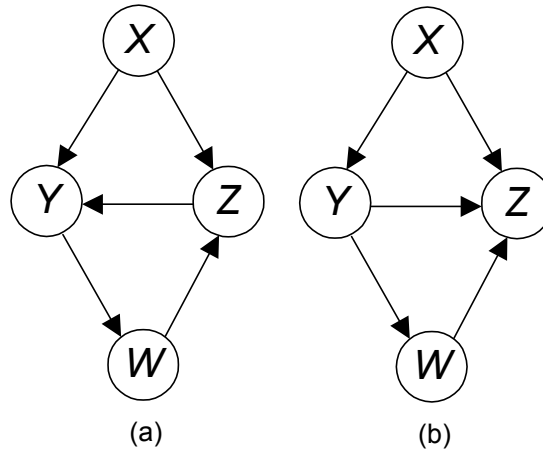

Figure 5.3: Both graphs are directed graphs; only the one in (b) is a directed acyclic graph.

5.2.1 Definition of a Bayesian Network

First, let's review some graph theory. A **directed graph** is a pair (V, E), where V is a finite, nonempty set whose elements are called **nodes** (or vertices), and E is a set of ordered pairs of distinct elements of V. Elements of E are called **directed edges**, and if $(X, Y) \in \mathsf{E}$, we say there is an edge from X to Y. The graph in Figure 5.3 (a) is a directed graph. The set of nodes in that figure is

$$\mathsf{V} = \{X, Y, Z, W\},$$

and the set of edges is

$$\mathsf{E} = \{(X, Y), (X, Z), (Y, W), (W, Z), (Z, Y)\}.$$

A **path** in a directed graph is a sequence of nodes $[X_1, X_2, \ldots, X_k]$ such that $(X_{i-1}, X_i) \in \mathsf{E}$ for $2 \leq i \leq k$. For example, $[X, Y, W, Z]$ is a path in the directed graph in Figure 5.3 (a). A **chain** in a directed graph is a sequence of nodes $[X_1, X_2, \ldots, X_k]$ such that $(X_{i-1}, X_i) \in \mathsf{E}$ or $(X_i, X_{i-1}) \in \mathsf{E}$ for $2 \leq i \leq k$. For example, $[Y, W, Z, X]$ is a chain in the directed graph in Figure 5.3 (b), but it is not a path. A **cycle** in a directed graph is a path from a node to itself. In Figure 5.3 (a) $[Y, W, Z, Y]$ is a cycle from Y to Y. However, in Figure 5.3 (b) $[Y, W, Z, Y]$ is not a cycle because it is not a path. A directed graph $\mathbb{G}$ is called a **directed acyclic graph** (DAG) if it contains no cycles. The directed graph in Figure 5.3 (b) is a DAG, whereas the one in Figure 5.3 (a) is not.

Given a DAG $\mathbb{G} = (\mathsf{V}, \mathsf{E})$ and nodes X and Y in V, Y is called a **parent** of X if there is an edge from Y to X, Y is called a **descendent** of X and X is called an **ancestor** of Y if there is a path from X to Y, and Y is called a

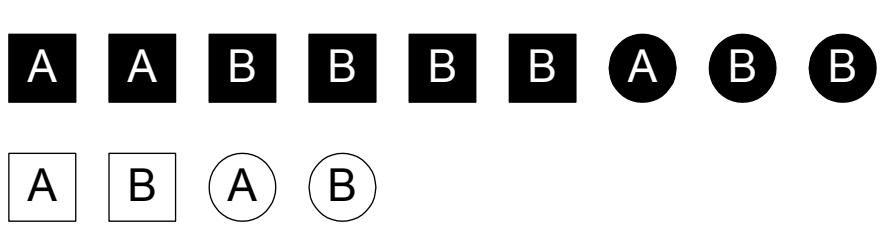

Figure 5.4: The random variables L and S are not independent, but they are conditionally independent given C.

nondescendent of X if Y is not a descendent of X and Y is not a parent of X.[1]

We can now state the following definition.

Definition 5.1 *Suppose we have a joint probability distribution P of the random variables in some set V and a DAG $\mathbb{G} = (\mathsf{V}, \mathsf{E})$. We say that $(\mathbb{G}, P)$ satisfies the **Markov condition** if for each variable $X \in \mathsf{V}$, X is conditionally independent of the set of all its nondescendents given the set of all its parents. Using the notation established in Chapter 2, Section 2.2.2, this means that if we denote the sets of parents and nondescendents of X by PA_X and ND_X, respectively, then*

$$I_P(X, \mathsf{ND}_X | \mathsf{PA}_X).$$

*If $(\mathbb{G}, P)$ satisfies the Markov condition, $(\mathbb{G}, P)$ is called a **Bayesian network**.*

Example 5.1 *Recall Chapter 2, Figure 2.1, which appears again as Figure 5.4. In Chapter 2, Example 2.23 we let P assign 1/13 to each object in the figure, and we defined these random variables on the set containing the objects.*

Variable	Value	Outcomes Mapped to This Value
L	l_1	All objects containing an A
	l_2	All objects containing a B
S	s_1	All square objects
	s_2	All circular objects
C	c_1	All black objects
	c_2	All white objects

We then showed that L and S are conditionally independent given C. That is, using the notation established in Chapter 2, Section 2.2.2, we showed

$$I_P(L, S | C).$$

Consider the DAG $\mathbb{G}$ in Figure 5.5. For that DAG we have the following.

[1] It is not standard to exclude a node's parents from its nondescendents, but this definition better serves our purposes.

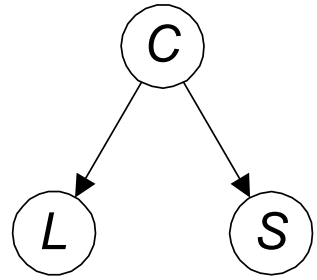

Figure 5.5: The joint probability distribution of L, S, and C constitutes a Bayesian network with this DAG.

Node	Parents	Nondescendents
L	C	S
S	C	L
C	$\varnothing$	$\varnothing$

For $(\mathbb{G}, P)$ to satisfy the Markov condition, we need to have

$$I_P(L, S|C)$$
$$I_P(S, L|C).$$

Note that since C has no nondescendents, we do not have a conditional independency for C. Since independence is symmetric, $I_P(L, S|C)$ implies $I_P(L, S|C)$. Therefore, all the conditional independencies required by the Markov condition are satisfied, and $(\mathbb{G}, P)$ is a Bayesian network.

Next we further illustrate the Markov condition with a more complex DAG.

Example 5.2 *Consider the DAG $\mathbb{G}$ in Figure 5.6. If $(\mathbb{G}, P)$ satisfied the Markov condition with some probability distribution P of X, Y, Z, W, and V, we would have the following conditional independencies.*

Node	Parents	Nondescendents	Conditional Independency
X	$\varnothing$	$\varnothing$	None
Y	X	Z, V	$I_P(Y, \{Z, V\}\|X)$
Z	X	Y	$I_P(Z, Y\|X)$
W	Y, Z	X, V	$I_P(W, \{X, V\}\|\{Y, Z\})$
V	Z	X, Y, W	$I_P(V, \{X, Y, W\}\|Z)$

5.2.2 Representation of a Bayesian Network

A Bayesian network $(\mathbb{G}, P)$, by definition, is a DAG $\mathbb{G}$ and joint probability distribution P that together satisfy the Markov condition. Then why in Figures 5.1 and 5.2 do we show a Bayesian network as a DAG and a set of conditional

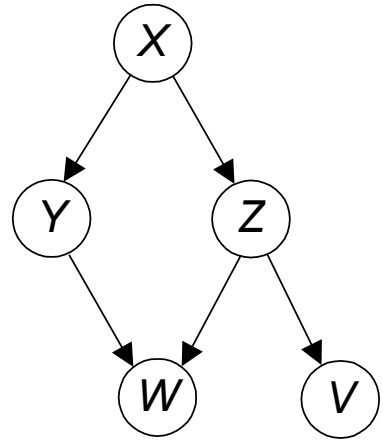

Figure 5.6: A DAG.

probability distributions? The reason is that $(\mathbb{G}, P)$ satisfies the Markov condition if and only if P is equal to the product of its conditional distributions in $\mathbb{G}$. Specifically, we have the following theorem.

Theorem 5.1 *$(\mathbb{G}, P)$ satisfies the Markov condition (and thus is a Bayesian network) if and only if P is equal to the product of its conditional distributions of all nodes given their parents in G, whenever these conditional distributions exist.*

Proof. *The proof can be found in [Neapolitan, 2004].* ■

Example 5.3 *We showed that the joint probability distribution P of the random variables L, S, and C defined on the set of objects in Figure 5.4 constitutes a Bayesian network with the DAG $\mathbb{G}$ in Figure 5.5. Next we illustrate that the preceding theorem is correct by showing that P is equal to the product of its conditional distributions in $\mathbb{G}$. Figure 5.7 shows those conditional distributions. We computed them directly from Figure 5.4. For example, since there are nine black objects (c_1) and six of them are squares (s_1), we compute*

$$P(s_1|c_1) = \frac{6}{9} = \frac{2}{3}.$$

The other conditional distributions are computed in the same way. To show that the joint distribution is the product of the conditional distributions, we need to show for all values of i, j, and k that

$$P(s_i, l_j, c_k) = P(s_i|c_k)P(l_j|c_k)P(c_k).$$

There are a total of eight combinations. We show that the equality holds for one of them. It is left as an exercise to show that it holds for the others. To

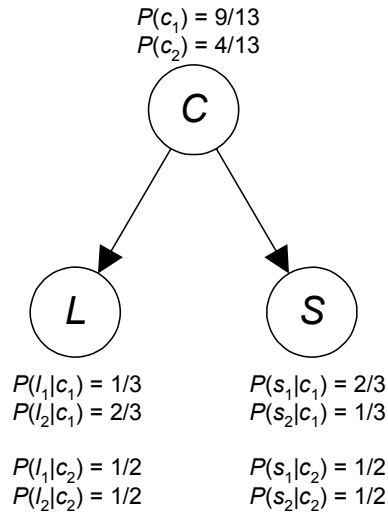

Figure 5.7: A Bayesian network representing the probability distribution P of the random variables L, S, and C defined on the set of objects in Figure 5.4.

that end, we have directly from Figure 5.4 that

$$P(s_1, l_1, c_1) = \frac{2}{13}.$$

From Figure 5.7 we have

$$P(s_1|c_1)P(l_1|c_1)P(c_1) = \frac{2}{3} \times \frac{1}{3} \times \frac{9}{13} = \frac{2}{13}.$$

Owing to Theorem 5.1, we can represent a Bayesian network $(\mathbb{G}, P)$ using the DAG $\mathbb{G}$ and the conditional distributions. We don't need to show every value in the joint distributions. These values can all be computed from the conditional distributions. So we always show a Bayesian network as the DAG and the conditional distributions as we did in Figures 5.1, 5.2, and 5.7. Herein lies the representational power of Bayesian networks. If there is a large number of variables, there are many values in the joint distribution. However, if the DAG is sparse, there are relatively few values in the conditional distributions. For example, suppose all variables are binary, and a joint distribution satisfies the Markov condition with the DAG in Figure 5.8. Then there are $2^{10} = 1024$ values in the joint distribution, but only $2 + 2 + 8 \times 8 = 68$ values in the conditional distributions. Note that we are not even including redundant parameters in this count. For example, in the Bayesian network in Figure 5.7 it is not necessary to show $P(c_2) = 4/13$ because $P(c_2) = 1 - P(c_1)$. So we need only show $P(c_1) = 9/13$. If we eliminate redundant parameters, there are only 34 values in the conditional distributions for the DAG in Figure 5.8 but still 1023 in the joint distribution. We see then that a Bayesian network is a structure for representing a joint probability distribution succinctly.

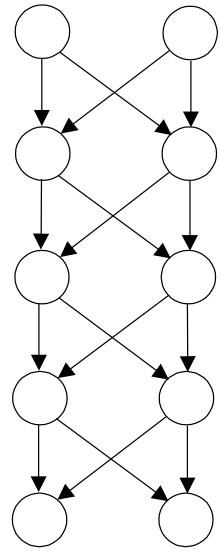

Figure 5.8: If all variables are binary and a joint distribution satisifes the Markov condition with this DAG, there are 1024 values in the joint distribution, but only 68 values in the conditional distributions.

It is important to realize that we can't take just any DAG and expect a joint distribution to equal the product of its conditional distributions in the DAG. This is only true if the Markov condition is satisfied. The next example illustrates this idea.

Example 5.4 *Consider again the joint probability distribution P of the random variables L, S, and C defined on the set of objects in Figure 5.4. Figure 5.9 shows its conditional distributions for the DAG in that figure. Note that we no longer show redundant parameters in our figures. If P satisfied the Markov condition with this DAG, we would have to have $I_P(L, S)$ because L has no parents and S is the sole nondescendent of L. It is left as an exercise to show that this independency does not hold. Furthermore, P is not equal to the product of its conditional distributions in this DAG. For example, we have directly from Figure 5.4 that*

$$P(s_1, l_1, c_1) = \frac{2}{13} = .15385.$$

From Figure 5.9 we have

$$P(c_1|l_1, s_1)P(l_1)P(s_1) = \frac{2}{3} \times \frac{5}{13} \times \frac{8}{13} = .15779.$$

It seems that we are left with a conundrum. That is, our goal is to succinctly represent a joint probability distribution using a DAG and conditional

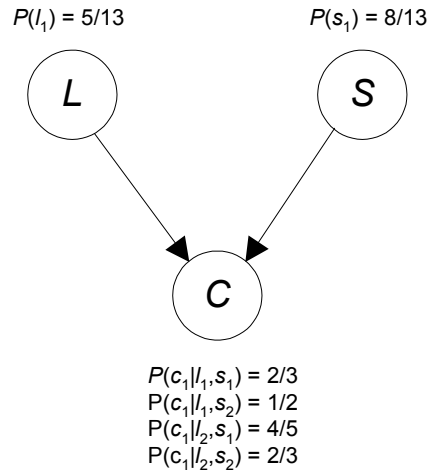

Figure 5.9: The joint probability distribution P of the random variables L, S, and C defined on the set of objects in Figure 5.4 does not satisfy the Markov condition with this DAG.

distributions for the DAG (a Bayesian network) rather than enumerating every value in the joint distribution. However, we don't know which DAG to use until we check whether the Markov condition is satisfied, and, in general, we would need to have the joint distribution to check this. A common way out of this predicament is to construct a **causal DAG**, which is a DAG in which there is an edge from X to Y if X causes Y. The DAGs in Figures 5.1 and 5.2 are causal; other DAGs shown so far in this chapter are not causal.

Next we discuss why a causal DAG should satisfy the Markov condition with the probability distribution of the variables in the DAG. A second way of obtaining the DAG is to learn it from data. This second way is discussed in Chapter 8.

5.3 Causal Networks as Bayesian Networks

Before discussing why a causal DAG should often satisfy the Markov condition with the probability distribution of the variables in the DAG, we formalize the notion of causality.

5.3.1 Causality

After providing an operational definition of a cause, we show a comprehensive example of identifying a cause according to this definition.

An Operational Definition of a Cause

One dictionary definition of a cause is "the one, such as a person, an event, or a condition, that is responsible for an action or a result." Although useful, this definition is certainly not the last word on the concept of causation, which has been investigated for centuries (see, e.g., [Hume, 1748]; [Piaget, 1966]; [Eells, 1991]; [Salmon, 1997]; [Spirtes et al., 1993; 2000]; [Pearl, 2000]). This definition does, however, shed light on an operational method for identifying causal relationships. That is, if the action of making variable X take some value sometimes changes the value taken by variable Y, then we assume X is responsible for sometimes changing Y's value, and we conclude X is a cause[2] of Y. More formally, we say we **manipulate** X when we force X to take some value, and we say X **causes** Y if there is some manipulation of X that leads to a change in the probability distribution of Y. We assume that if manipulating X leads to a change in the probability distribution of Y, then X obtaining a value by any means whatsoever also leads to a change in the probability distribution of Y. So we assume that causes and their effects are statistically correlated. However, as we shall discuss soon, variables can be correlated without one causing the other.

A manipulation consists of a **randomized controlled experiment** (**RCE**) using some specific population of entities (e.g., individuals with chest pain) in some specific context (e.g., they currently receive no chest pain medication and they live in a particular geographical area). The causal relationship discovered is then relative to this population and this context.

Let's discuss how the manipulation proceeds. We first identify the population of entities we want to consider. Our random variables are features of these entities. Next we ascertain the causal relationship we want to investigate. Suppose we are trying to determine if variable X is a cause of variable Y. We then sample a number of entities from the population. For every entity selected, we manipulate the value of X so that each of its possible values is given to the same number of entities (if X is continuous, we choose the values of X according to a uniform distribution). After the value of X is set for a given entity, we measure the value of Y for that entity. The more the resultant data show a dependency between X and Y, the more the data support that X causes Y. The manipulation of X can be represented by a variable M that is external to the system being studied. There is one value m_i of M for each value x_i of X; the probabilities of all values of M are the same; and when M equals m_i, X equals x_i. That is, the relationship between M and X is deterministic. The data support that X causes Y to the extent that the data indicate $P(y_i|m_j) \neq P(y_i|m_k)$ for $j \neq k$. Manipulation is actually a special kind of causal relationship that we assume exists primordially and is within our control so that we can define and discover other causal relationships.

[2] This notion of causality does not pertain to *token* causality, which concerns individual, causal events rather than probabilistic relationships among variables.

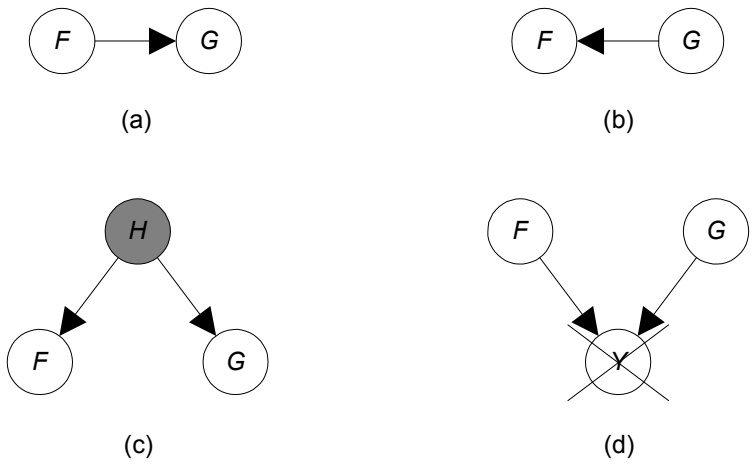

Figure 5.10: The edges in these graphs represent causal influences. All four causal relationships could account for F and G being correlated.

An Illustration of Manipulation

We demonstrate these ideas with a comprehensive example concerning recent headline news. The pharmaceutical company Merck had been marketing its drug finasteride as medication for men with benign prostatic hyperplasia (BPH). Based on anecdotal evidence, it seemed that there was a correlation between use of the drug and regrowth of scalp hair. Let's assume that Merck took a random sample from the population of interest and, based on that sample, determined there is a correlation between finasteride use and hair regrowth. Should Merck conclude that finasteride causes hair regrowth and therefore market it as a cure for baldness? Not necessarily. There are quite a few causal explanations for the correlation of two variables. We discuss these relationships next.

Possible Causal Relationships Let F be a variable representing finasteride use and G be a variable representing scalp hair growth. The actual values of F and G are unimportant to the present discussion. We could use either continuous or discrete values. If F caused G, then indeed they would be statistically correlated, but this would also be the case if G caused F or if they had some hidden common cause H. If we represent a causal influence by a directed edge, Figure 5.10 shows these three possibilities plus one more. Figure 5.10 (a) shows the conjecture that F causes G, which we already suspect might be the case. However, it could be that G causes F (Figure 5.10 (b)). You may argue that, based on domain knowledge, this does not seem reasonable. However, we do not, in general, have domain knowledge when doing a statistical analysis. So, from the correlation alone, the causal relationships in Figures 5.10 (a) and 5.10 (b) are equally reasonable. Even in this domain, G causing F seems possible. A man might have used some other hair regrowth product such as minoxidil that

caused him to regrow hair, become excited about the regrowth, and decide to try other products such as finasteride, which he heard might cause regrowth.

A third possibility, shown in Figure 5.10 (c), is that F and G have some hidden common cause H that accounts for their statistical correlation. For example, a man concerned about hair loss might try both finasteride and minoxidil in his effort to regrow hair. The minoxidil might cause hair regrowth, whereas the finasteride might not. In this case the man's concern is a cause of finasteride use and hair regrowth (indirectly through minoxidil use), whereas the two are not causally related.

A fourth possibility is that our sample (or even our entire population) consists of individuals who have some (possibly hidden) effect of both F and G. For example, suppose finasteride and apprehension about lack of hair regrowth both cause hypertension,[3] and our sample consists of individuals who have hypertension Y. We say a node is **instantiated** when we know its value for the entity currently being modeled. So we are saying the variable Y is instantiated to the same value for every entity in our sample. This situation is depicted in Figure 5.10 (d), where the cross through Y means that the variable is instantiated. Usually, the instantiation of a common effect creates a dependency between its causes because each cause explains the occurrence of the effect, thereby making the other cause less likely. As noted earlier, psychologists call this **discounting**. So, if this were the case, discounting would explain the correlation between F and G. This type of dependency is called **selection bias**.[4]

A final possibility (not shown in Figure 5.10) is that F and G are not causally related at all. A notable example of this situation occurs when our entities are points in time and our random variables are values of properties at these different points in time. Such variables are often correlated without having any apparent causal connection. For example, if our population consists of points in time, J is the Dow Jones Average at a given time, and L is Professor Neapolitan's hairline at a given time, then J and L are correlated.[5] Yet they do not seem to be causally connected. Some argue that there are hidden common causes beyond our ability to measure. We will not discuss this issue further here. We only want to note the difficulty with such correlations. In light of the factors we've discussed, we see then that we cannot deduce a causal relationship between two variables from the mere fact that they are statistically correlated.

Any of the causal relationships shown in Figure 5.10 could occur in combination, resulting in F and G being correlated. For example, it could be both that finasteride causes hair regrowth and that excitement about regrowth may cause the use of finasteride, meaning we could have a causal loop or feedback. So, we could have the causal relationships in both Figures 5.10 (a) and 5.10 (b).

[3] There is no evidence that either finasteride or apprehension about the lack of hair regrowth causes hypertension. This example is only for the sake of illustration.

[4] This could happen if our sample is a **convenience sample**, which is a sample in which the participants are selected at the convenience of the researcher. The researcher makes no attempt to ensure that the sample is an accurate representation of the larger population. In the context of the current example, this might be the case if it is convenient for the researcher to observe males hospitalized for hypertension.

[5] Unfortunately, his hairline did not go back down in fall, 2008.

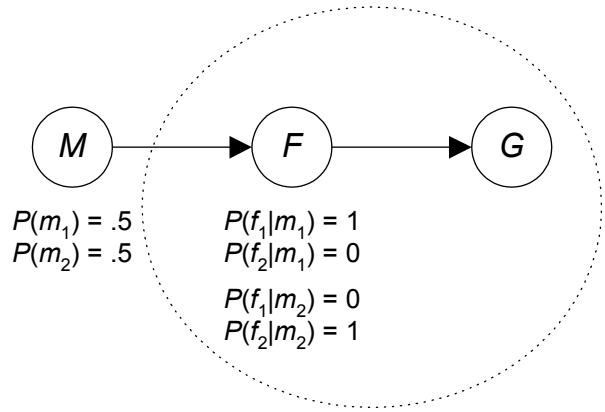

Figure 5.11: An RCE investigating whether F causes G.

It might not be obvious why two variables with a common cause would be correlated. Consider the present example. Suppose H is a common cause of F and G and neither F nor G caused the other. Then H and F are correlated, because H causes F, and H and G are correlated because H causes G, which implies F and G are correlated transitively through H. Here is a more detailed explanation. For the sake of example, suppose h_1 is a value of H that has a causal influence on F taking value f_1 and on G taking value g_1. Then if F had value f_1, each of its causes would become more probable because one of them should be responsible. So, $P(h_1|f_1) > P(h_1)$. Now, since the probability of h_1 has gone up, the probability of g_1 would also go up because h_1 causes g_1. Therefore, $P(g_1|f_1) > P(g_1)$, which means F and G are correlated.

Merck's Manipulation Study Since Merck could not conclude that finasteride causes hair regrowth from their mere correlation alone, they did a manipulation study to test this conjecture. The study was done on 1879 men aged 18 to 41 with mild to moderate hair loss of the vertex and anterior mid-scalp areas. Half of the men were given 1 mg. of finasteride, whereas the other half were given 1 mg. of a placebo. The following table shows the possible values of the variables in the study, including the manipulation variable M.

Variable	Value	When the Variable Takes This Value
F	f_1	Subject takes 1 mg. of finasteride
	f_2	Subject takes 1 mg. of a placebo
G	g_1	Subject has significant hair regrowth
	g_2	Subject does not have significant hair regrowth
M	m_1	Subject is chosen to take 1 mg. of finasteride
	m_2	Subject is chosen to take 1 mg. of a placebo

An RCE used to test the conjecture that F causes G is shown in Figure 5.11. There is an oval around the system being studied (F and G and their

possible causal relationship) to indicate that the manipulation comes from outside the system. The edges in Figure 5.11 represent causal influences. The RCE supports the conjecture that F causes G to the extent that the data support $P(g_1|m_1) \neq P(g_1|m_2)$. Merck decided that "significant hair regrowth" would be judged according to the opinion of independent dermatologists. A panel of independent dermatologists evaluated photos of the men after 24 months of treatment. The panel judged that significant hair regrowth was demonstrated in 66% of men treated with finasteride, compared to 7% of men treated with placebo.

Basing our probability on these results, we have that $P(g_1|m_1) \approx .67$ and $P(g_1|m_2) \approx .07$. In a more analytical analysis, only 17% of men treated with finasteride demonstrated hair loss (defined as any decrease in hair count from baseline). In contrast, 72% of the placebo group lost hair, as measured by hair count. Merck concluded that finasteride does indeed cause hair regrowth and on December 22, 1997, announced that the U.S. Food and Drug Administration granted marketing clearance to Propecia™ (finasteride 1 mg.) for treatment of male pattern hair loss (androgenetic alopecia), for use in men only (see [McClennan and Markham, 1999] for more on this topic).

5.3.2 Causality and the Markov Condition

First we more rigorously define a causal DAG. After that we state the causal Markov assumption and argue why it should be satisfied.

Causal DAGs

We say X is a **cause** of Y if a manipulation of X results in a change in the probability distribution of Y. A **causal graph** is a directed graph containing a set of causally related random variables $\mathbf{V}$ such that for every $X, Y \in \mathbf{V}$ there is an edge from X to Y if and only if X is a cause of Y, and there is no subset of variables $\mathbf{W}_{XY}$ of $\mathbf{V}$ such that if we knew the values of the variables in $\mathbf{W}_{XY}$, a manipulation of X would no longer change the probability distribution of Y. If there is an edge from X to Y, we call X a **direct cause** of Y. Note that whether or not X is a direct cause of Y depends on the variables included in $\mathbf{V}$. A causal graph is a **causal DAG** if the causal graph is acyclic (i.e., there are no causal feedback loops).

Example 5.5 *Testosterone (T) is known to convert to dihydro-testosterone (D), and dihydro-testosterone is believed to be the hormone necessary for erectile function (E). A study in [Lugg et al., 1995] tested the causal relationship among these variables in rats. They manipulated testosterone to low levels and found that both dihydro-testosterone and erectile function declined. They then held dihydro-testosterone fixed at low levels and found that erectile function was low regardless of the manipulated value of testosterone. Finally, they held dihydro-testosterone fixed at high levels and found that erectile function was high regardless of the manipulated value of testosterone. So they learned that, in a causal graph containing only the variables T, D, and E, T is a direct cause*

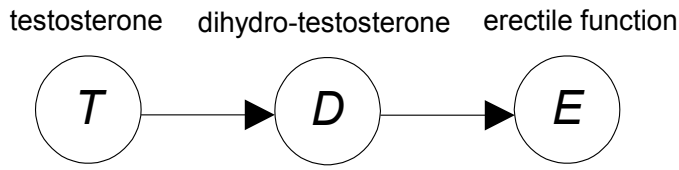

Figure 5.12: A causal DAG.

of D, and D is a direct cause of E, but, although T is a cause of E, it is not a direct cause. So the causal graph (DAG) is the one in Figure 5.12.

Notice that if the variable D were not in the DAG in Figure 5.12, T would be called a direct cause of E, and there would be an edge from T directly into E instead of the directed path through D. In general, our edges always represent only the relationships among the identified variables. It seems we can usually conceive of intermediate, unidentified variables along each edge. Consider the following example taken from [Spirtes et al., 1993; 2000], p. 42:

> If C is the event of striking a match, and A is the event of the match catching on fire, and no other events are considered, then C is a direct cause of A. If, however, we added B, the sulfur on the match tip achieved sufficient heat to combine with the oxygen, then we could no longer say that C directly caused A, but rather C directly caused B and B directly caused A. Accordingly, we say that B is a causal mediary between C and A if C causes B and B causes A.

Note that, in this intuitive explanation, a variable name is used to also stand for a value of the variable. For example, A is a variable whose value is *on-fire* or *not-on-fire*, and A is also used to represent that the match is on fire. Clearly, we can add more causal mediaries. For example, we could add the variable D, representing whether the match tip is abraded by a rough surface. C would then cause D, which would cause B, and so on. We could go much further and describe the chemical reaction that occurs when sulfur combines with oxygen.

Indeed, it seems we can conceive of a continuum of events in any causal description of a process. We see then that the set of observable variables is observer dependent. Apparently an individual, given myriad sensory input, selectively records discernible events and develops cause/effect relationships among them. Therefore, rather than assuming that there is a set of causally related variables out there, it seems more appropriate to only assume that, in a given context or application, we identify certain variables and develop a set of causal relationships among them.

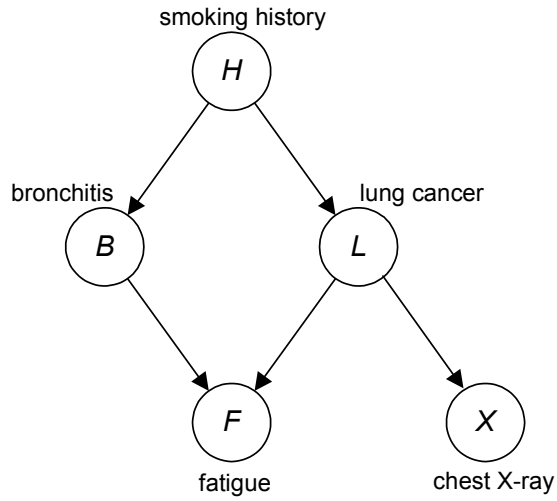

Figure 5.13: A causal DAG.

The Causal Markov Assumption

If we assume that the observed probability distribution P of a set of random variables $\mathbf{V}$ satisfies the Markov condition with the causal DAG $\mathbb{G}$ containing the variables, we say we are making the **causal Markov assumption**, and we call $(\mathbb{G}, P)$ a **causal network**. Why should we make the causal Markov assumption? To answer this question we show several examples.

Example 5.6 *Consider again the situation involving testosterone (T), dihydrotestosterone (D), and erectile function (E). Recall the manipulation study in [Lugg et al., 1995], which we discussed in Example 5.5. This study showed that if we instantiate D, the value of E is independent of the value of T. So there is experimental evidence that the Markov condition is satisfied for a three-variable causal chain.*

Example 5.7 *A history of smoking (H) is known to cause both bronchitis (B) and lung cancer (L). Lung cancer and bronchitis both cause fatigue (F), but only lung cancer can cause a chest X-ray (X) to be positive. There are no other causal relationships among the variables. Figure 5.13 shows a causal DAG containing these variables. The causal Markov assumption for that DAG entails the following conditional independencies.*

Node	Parents	Nondescendents	Conditional Independency
H	$\varnothing$	$\varnothing$	None
B	H	L, X	$I_P(B, \{L, X\}\|H)$
L	H	B	$I_P(L, B\|H)$
F	B, L	H, X	$I_P(F, \{H, X\}\|\{B, L\})$
X	L	H, B, F	$I_P(X, \{H, B, F\}\|L)$

Given the causal relationship in Figure 5.13, we would not expect bronchitis and lung cancer to be independent, because if someone had lung cancer it would make it more probable that the individual smoked (since smoking can cause lung cancer), which would make it more probable that another effect of smoking, namely bronchitis, was present. However, if we knew someone smoked, it would already be more probable that the person had bronchitis. Learning that the individual had lung cancer could no longer increase the probability of smoking (which is now 1), which means it cannot change the probability of bronchitis. That is, the variable H shields B from the influence of L, which is what the causal Markov condition says. Similarly, a positive chest X-ray increases the probability of lung cancer, which in turn increases the probability of smoking, which in turn increases the probability of bronchitis. So, a chest X-ray and bronchitis are not independent. However, if we knew the person had lung cancer, the chest X-ray could not change the probability of lung cancer and thereby change the probability of bronchitis. So B is independent of X conditional on L, which is what the causal Markov condition says.

In summary, if we create a causal graph containing the variables X and Y, if X and Y do not have a hidden common cause (i.e., a cause that is not in our graph), if there are no causal paths from Y back to X (i.e., our graph is a DAG), and if we do not have selection bias (i.e., our probability distribution is not obtained from a population in which a common effect is instantiated to the same value for all members of the population), then we feel X and Y are independent if we condition on a set of variables including at least one variable in each of the causal paths from X to Y. Since the set of all parents of Y is such a set, we feel that the Markov condition holds relative to X and Y. So we conclude that the causal Markov assumption is justified for a causal graph if the following conditions are satisfied:

1. There are no hidden common causes. That is, all common causes are represented in the graph.

2. There are no causal feedback loops. That is, our graph is a DAG.

3. Selection bias is not present.

Note that, for the Markov condition to hold, there must be an edge from X to Y whenever there is a causal path from X to Y besides the ones containing variables in our graph. However, we need not stipulate this requirement because it is entailed by the definition of a causal graph. Recall that in a causal graph there is an edge from X to Y if X is a direct cause of Y.

Perhaps the condition that is most frequently violated is that there can be no hidden common causes. We discuss this condition further with a final example.

Example 5.8 *Suppose we wanted to create a causal DAG containing the variables cold (C), sneezing (S), and runny nose (R). Since a cold can cause both sneezing and a runny nose and neither of these conditions can cause each other,*

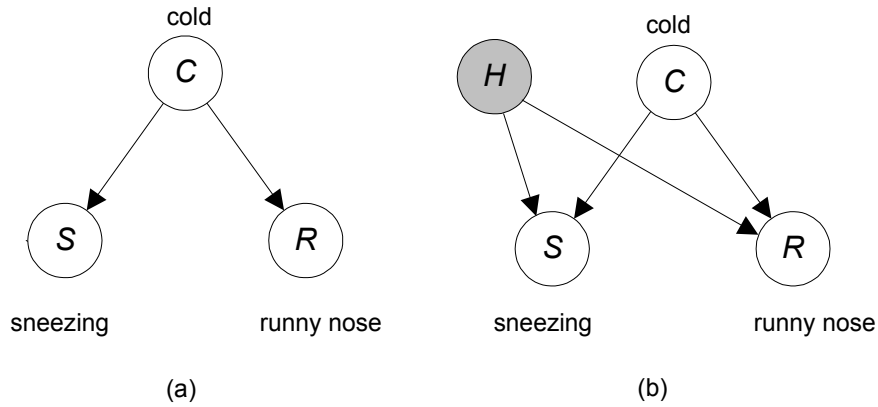

Figure 5.14: The causal Markov assumption would not hold for the DAG in (a) if there is a hidden common cause as depicted in (b).

we would create the DAG in Figure 5.14 (a). The causal Markov condition for that DAG would entail $I_P(S, R|C)$. However, if there were a hidden common cause of S and R as depicted in Figure 5.14 (b), this conditional independency would not hold because even if the value of C were known, S would change the probability of H, which in turn would change the probability of R. Indeed, there is at least one other cause of sneezing and runny nose, namely hay fever. So when making the causal Markov assumption, we must be certain that we have identified all common causes.

5.3.3 The Markov Condition without Causality

We have argued that a causal DAG often satisfies the Markov condition with the joint probability distribution of the random variables in the DAG. This does not mean that the edges in a DAG in a Bayesian network must be causal. That is, a DAG can satisfy the Markov condition with the probability distribution of the variables in the DAG without the edges being causal. For example, we showed that the joint probability distribution P of the random variables L, S, and C defined on the set of objects in Figure 5.4 satisfies the Markov condition with the DAG $\mathbb{G}$ in Figure 5.5. However, we would not argue that the color of the objects causes their shape or the letter that is on them. As another example, if we reversed the edges in the DAG in Figure 5.12 to obtain the DAG $E \rightarrow DHT \rightarrow T$, the new DAG would also satisfy the Markov condition with the probability distribution of the variables, yet the edges would not be causal.

5.4 Inference in Bayesian Networks

As noted previously, a standard application of Bayes' Theorem is inference in a two-node Bayesian network. Larger Bayesian networks address the problem

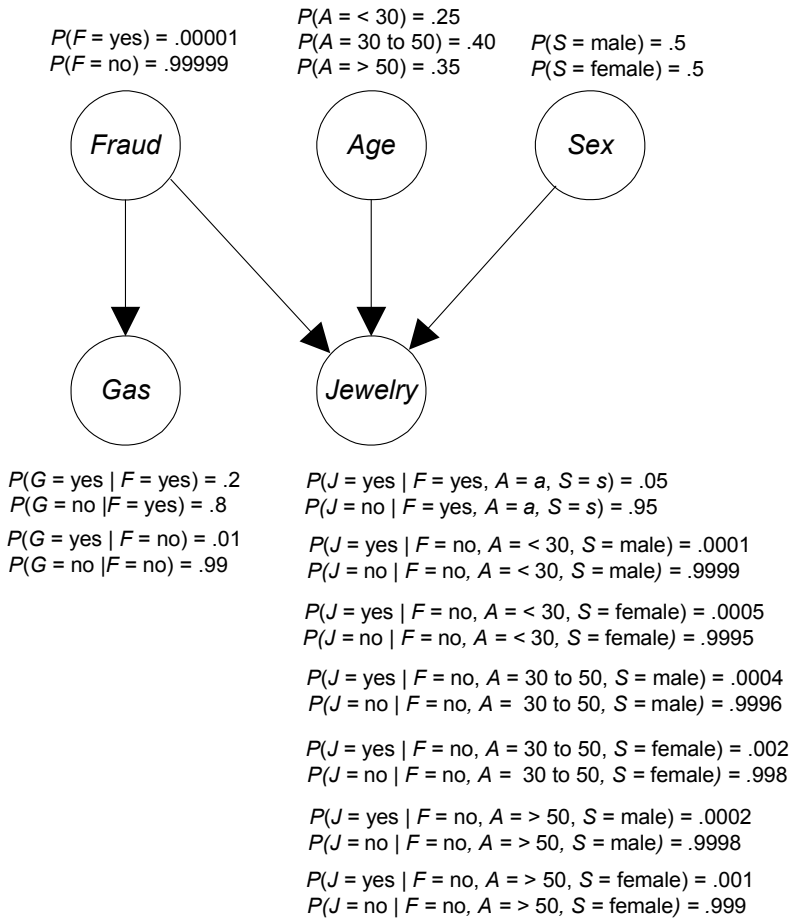

Figure 5.15: Bayesian network for detecting credit card fraud.

of representing the joint probability distribution of a large number of variables. For example, Figure 5.2, which appears again as Figure 5.15, represents the joint probability distribution of variables related to credit card fraud. Inference in this network consists of computing the conditional probability of some variable (or set of variables), given that other variables are instantiated to certain values. For example, we might want to compute the probability of credit card fraud, given that gas has been purchased, jewelry has been purchased, and the card holder is male. To accomplish this inference we need sophisticated algorithms. First, we show simple examples illustrating how one of these algorithms uses the Markov condition and Bayes' Theorem to do inference. Then we reference papers describing some of the algorithms. Finally we show examples using the algorithms to do inference.

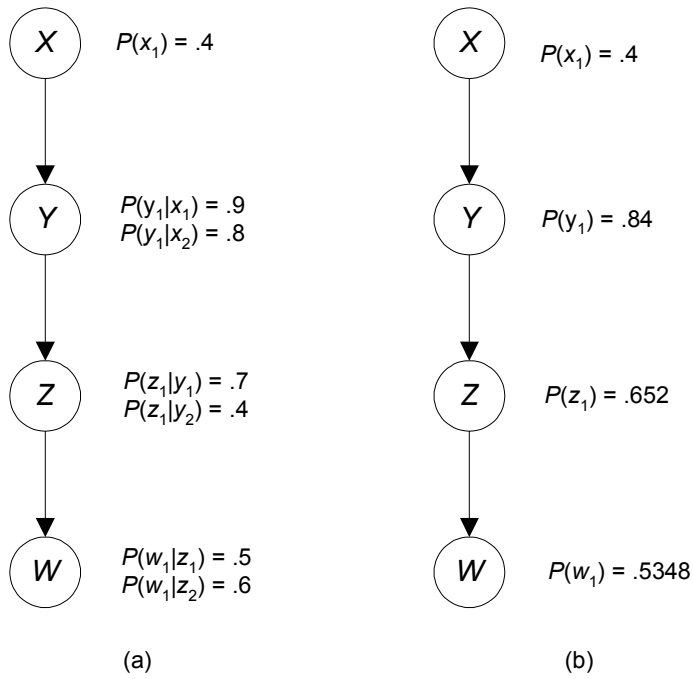

Figure 5.16: A Bayesian network appears in (a), and the prior probabilities of the variables in that network are shown in (b). Each variable has only two values, so only the probability of one is shown in (a).

5.4.1 Examples of Inference

Next we present some examples illustrating how the conditional independencies entailed by the Markov condition can be exploited to accomplish inference in a Bayesian network.

Example 5.9 *Consider the Bayesian network in Figure 5.16 (a). The prior probabilities of all variables can be computed using the Law of Total Probability:*

$$P(y_1) = P(y_1|x_1)P(x_1) + P(y_1|x_2)P(x_2) = (.9)(.4) + (.8)(.6) = .84$$

$$P(z_1) = P(z_1|y_1)P(y_1) + P(z_1|y_2)P(y_2) = (.7)(.84) + (.4)(.16) = .652$$

$$P(w_1) = P(w_1|z_1)P(z_1) + P(w_1|z_2)P(z_2) = (.5)(.652) + (.6)(.348) = .5348.$$

These probabilities are shown in Figure 5.16 (b). Note that the computation for each variable requires information determined for its parent. We can therefore consider this method a message-passing algorithm in which each node passes

its child a message needed to compute the child's probabilities. Clearly, this algorithm applies to an arbitrarily long linked list and to trees.

Example 5.10 *Suppose now that X is instantiated for x_1. Since the Markov condition entails that each variable is conditionally independent of X given its parent, we can compute the conditional probabilities of the remaining variables by again using the Law of Total Probability (however, now with the background information that $X = x_1$) and passing messages down as follows:*

$$P(y_1|x_1) = .9$$

$$\begin{aligned} P(z_1|x_1) &= P(z_1|y_1, x_1)P(y_1|x_1) + P(z_1|y_2, x_1)P(y_2|x_1) \\ &= P(z_1|y_1)P(y_1|x_1) + P(z_1|y_2)P(y_2|x_1) \qquad \text{// Markov condition} \\ &= (.7)(.9) + (.4)(.1) = .67 \end{aligned}$$

$$\begin{aligned} P(w_1|x_1) &= P(w_1|z_1, x_1)P(z_1|x_1) + P(w_1|z_2, x_1)P(z_2|x_1) \\ &= P(w_1|z_1)P(z_1|x_1) + P(w_1|z_2)P(z_2|x_1) \\ &= (.5)(.67) + (.6)(1 - .67) = .533. \end{aligned}$$

Clearly, this algorithm also applies to an arbitrarily long linked list and to trees.

The preceding example shows how we can use downward propagation of messages to compute the conditional probabilities of variables below the instantiated variable. Next we illustrate how to compute conditional probabilities of variables above the instantiated variable.

Example 5.11 *Suppose W is instantiated for w_1 (and no other variable is instantiated). We can use upward propagation of messages to compute the conditional probabilities of the remaining variables. First, we use Bayes' Theorem to compute $P(z_1|w_1)$:*

$$P(z_1|w_1) = \frac{P(w_1|z_1)P(z_1)}{P(w_1)} = \frac{(.5)(.652)}{.5348} = .6096.$$

Then, to compute $P(y_1|w_1)$, we again apply Bayes' Theorem:

$$P(y_1|w_1) = \frac{P(w_1|y_1)P(y_1)}{P(w_1)}.$$

We cannot yet complete this computation because we do not know $P(w_1|y_1)$. We can obtain this value using downward propagation as follows:

$$P(w_1|y_1) = (P(w_1|z_1)P(z_1|y_1) + P(w_1|z_2)P(z_2|y_1).$$

After doing this computation, also computing $P(w_1|y_2)$ (because X will need this value) and then determining $P(y_1|w_1)$, we pass $P(w_1|y_1)$ and $P(w_1|y_2)$ to X. We then compute $P(w_1|x_1)$ and $P(x_1|w_1)$ in sequence:

$$P(w_1|x_1) = (P(w_1|y_1)P(y_1|x_1) + P(w_1|y_2)P(y_2|x_1)$$

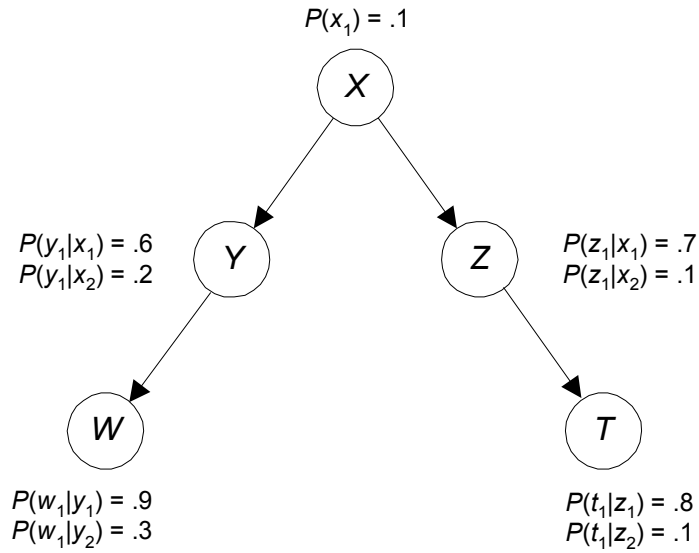

Figure 5.17: A Bayesian network. Each variable has only two possible values, so only the probability of one is shown.

$$P(x_1|w_1) = \frac{P(w_1|x_1)P(x_1)}{P(w_1)}.$$

It is left as an exercise to perform these computations. Clearly, this upward propagation scheme applies to an arbitrarily long linked list.

The next example shows how to turn corners in a tree.

Example 5.12 *Consider the Bayesian network in Figure 5.17. Suppose W is instantiated for w_1. We compute $P(y_1|w_1)$ followed by $P(x_1|w_1)$ using the upward propagation algorithm just described. Then we proceed to compute $P(z_1|w_1)$ followed by $P(t_1|w_1)$ using the downward propagation algorithm. This is left as an exercise.*

5.4.2 Inference Algorithms and Packages

By exploiting local independencies as we did in the previous subsection, Pearl [1986, 1988] developed a message-passing algorithm for inference in Bayesian networks. Based on a method originated in [Lauritzen and Spiegelhalter, 1988], Jensen et al. [1990] developed an inference algorithm that involves the extraction of an undirected triangulated graph from the DAG in a Bayesian network and the creation of a tree, whose vertices are the cliques of this triangulated graph. Such a tree is called a **junction tree**. Conditional probabilities are then computed by passing messages in the junction tree. Li and D'Ambrosio

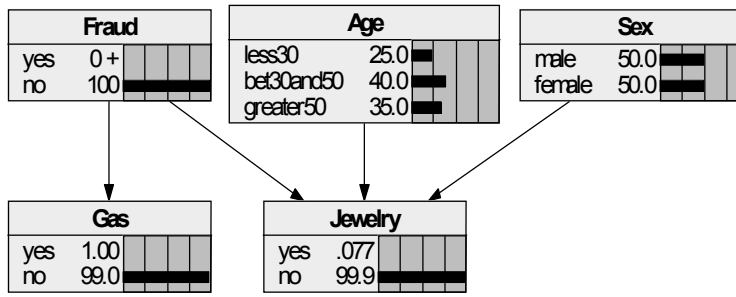

Figure 5.18: The fraud detection Bayesian network in Figure 5.15, implemented using Netica.

[1994] took a different approach. They developed an algorithm that approximates finding the optimal way to compute marginal distributions of interest from the joint probability distribution. They call this **symbolic probabilistic inference** (SPI).

All these algorithms are worst-case nonpolynomial time. This is not surprising, since the problem of inference in Bayesian networks has been shown to be NP-hard [Cooper, 1990]. In light of this result, approximation algorithms for inference in Bayesian networks have been developed. One such algorithm, likelihood weighting, was developed independently in [Fung and Chang, 1990] and [Shachter and Peot, 1990]. It is proven in [Dagum and Luby, 1993] that the problem of approximate inference in Bayesian networks is also NP-hard. However, there are restricted classes of Bayesian networks that are provably amenable to a polynomial-time solution (see [Dagum and Chavez, 1993]). Indeed, a variant of the likelihood weighting algorithm, which is worst-case polynomial time as long as the network does not contain extreme conditional probabilities, appears in [Pradhan and Dagum, 1996].

Practitioners need not concern themselves with all these algorithms since a number of packages for doing inference in Bayesian networks have been developed. A few of them are shown here:

1. Netica (www.norsys.com/)
2. GeNIe (genie.sis.pitt.edu/)
3. HUGIN (/www.hugin.com/)
4. Elvira (www.ia.uned.es/~elvira/)
5. BUGS (www.mrc-bsu.cam.ac.uk/bugs/)

In this book we ordinarily use Netica to illustrate inference. Figure 5.18 shows the fraud detection network in Figure 5.15 implemented using Netica.

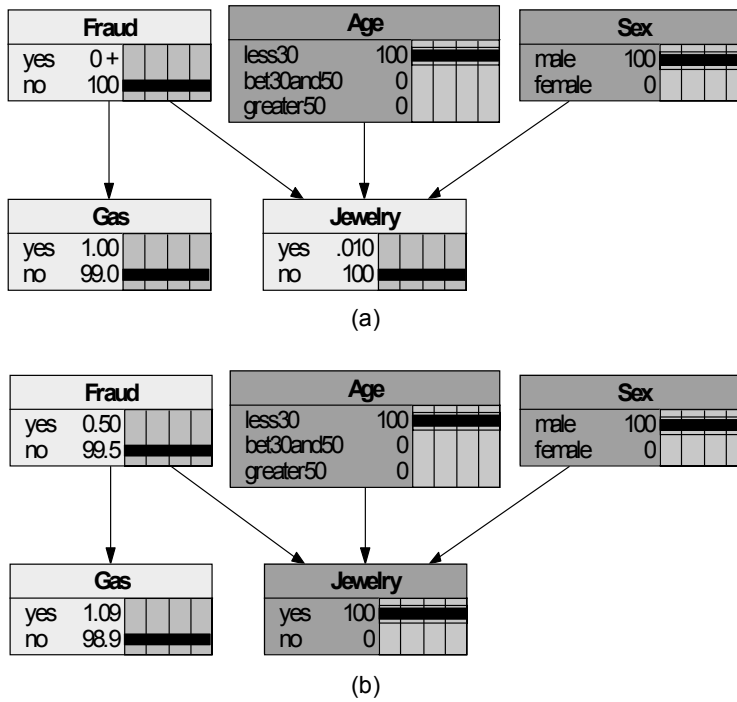

Figure 5.19: In (a) *Age* has been instantiated to *less*30 and *Sex* has been instantiated to *male*. In (b) *Age* has been instantiated to *less*30, *Sex* has been instantiated to *male*, and *Jewelry* has been instantiated to *yes*.

5.4.3 Inference Using Netica

Next we illustrate inference in a Bayesian network using Netica. Notice from Figure 5.18 that Netica computes and shows the prior probabilities of the variables rather than showing the conditional probability distributions. Probabilities are shown as percentages. For example, the fact that there is a .077 next to *yes* in the *Jewelry* node means

$$P(Jewelry = yes) = .00077.$$

This is the prior probability of a jewelry purchase in the past 24 hours being charged to any particular credit card.

After variables are instantiated, Netica shows the conditional probabilities of the other variables given these instantiations. In Figure 5.19 (a) we instantiated *Age* to *less*30 and *Sex* to *male*. So the fact that there is .010 next to *yes* in the *Jewelry* node means

$$P(Jewelry = yes | Age = less30, Sex = male) = .00010.$$

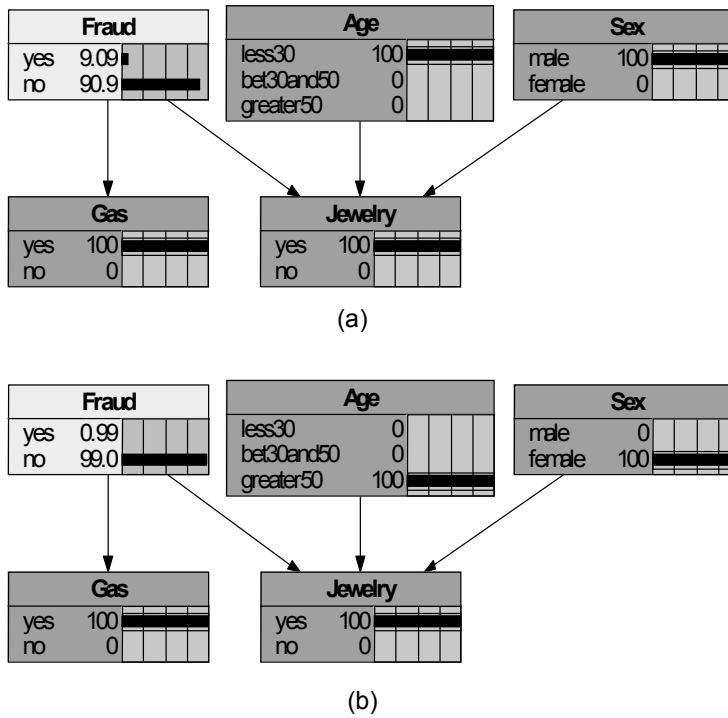

Figure 5.20: *Sex* and *Jewelry* have both been instantiated to *yes* in both (a) and (b). However, in (a) the card holder is a young man, whereas in (b) it is an older woman.

Notice that the probability of *Fraud* has not changed. This is what we would expect. First the Markov condition says that *Fraud* should be independent of *Age* and *Sex*. Second, it seems they should be independent. That is, the fact that the card holder is a young man should not make it more or less likely that the card is being used fraudulently. Figure 5.19 (b) has the same instantiations as Figure 5.19 (a) except that we have also instantiated *Jewelry* to *yes*. Notice that the probability of *Fraud* has now changed. First, the jewelry purchase makes *Fraud* more likely to be *yes*. Second, the fact that the card holder is a young man means it is less likely the card holder would make the purchase, thereby making *Fraud* even more likely to be *yes*.

In Figures 5.20 (a) and 5.20 (b), *Gas* and *Jewelry* have both been instantiated to *yes*. However, in Figure 5.20 (a) the card holder is a young man, whereas in Figure 5.20 (b) it is an older woman. This illustrates discounting of the jewelry purchase. When the card holder is a young man, the probability of *Fraud* being *yes* is high (.0909). However, when it is an older woman, it is still low (.0099) because the fact that the card holder is an older woman explains the jewelry purchase.

5.5 Networks with Continuous Variables

So far in all our Bayesian networks the variables have been discrete. Next we discuss Bayesian networks that contain continuous variables.

5.5.1 Gaussian Bayesian Networks

Gaussian Bayesian networks contain variables that are normally distributed. (The normal distribution is reviewed in Section 3.3.) We motivate such networks with the following example.

Example 5.13 *Suppose you are considering taking a job that pays* \$10 *an hour and you expect to work* 40 *hours per week. However, you are not guaranteed* 40 *hours, and you estimate the number of hours actually worked in a week to be normally distributed with mean* 40 *and standard deviation* 5*. You have not yet fully investigated the benefits such as bonus pay and nontaxable deductions (e.g., contributions to a retirement program). However, you estimate these other influences on your gross taxable weekly income to also be normally distributed with mean* 0 *(that is, you feel they about offset) and standard deviation* 30*. Furthermore, you assume that these other influences are independent of your hours worked.*

We define the following random variables.

Variable	What the Variable Represents
X	Hours worked in the week
Y	Salary obtained in the week

Based on the preceding discussion, X is distributed as follows:

$$\rho(x) = \text{NormalDen}(x; 40, 5^2).$$

A portion of your salary Y is a deterministic function of X. That is, you will receive $10x$ dollars if you work x hours. However, your gross salary may be greater or less than this based on the other influences we discussed. That is,

$$y = 10x + \varepsilon_Y,$$

where

$$\rho(\varepsilon_Y) = \text{NormalDen}(\varepsilon_Y; 0, 30^2).$$

Since the expected value of those other influences is 0*,*

$$E(Y|x) = 10x.$$

and since the variance of those other influences is 30^2*,*

$$V(Y|x) = 30^2.$$

So Y is distributed conditionally as follows:

$$\rho(y|x) = \text{NormalDen}(y; 10x, 30^2).$$

Therefore, the relationship between X and Y is represented by the Bayesian network in Figure 5.21.

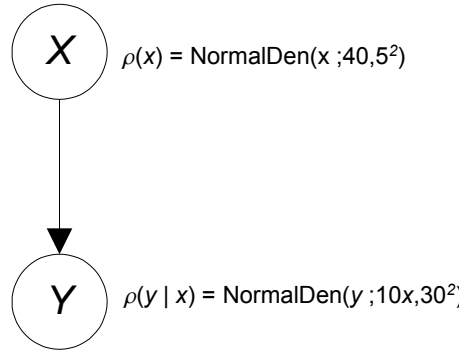

Figure 5.21: A Gaussian Bayesian network.

The Bayesian network we've just developed is an example of a Gaussian Bayesian network. In general, in a Gaussian Bayesian network, the root is normally distributed, and each non-root Y is a linear function of its parents plus an error term ε_Y that is normally distributed with mean 0 and variance σ_Y^2.So if $X_1, X_2, \ldots$ and X_k are the parents of Y, then

$$Y = b_1x_1 + b_2x_2 + \cdots b_kx_k + \varepsilon_Y, \tag{5.1}$$

where

$$\rho(\varepsilon_Y) = \text{NormalDen}(\varepsilon_Y; 0, \sigma_Y^2),$$

and Y is distributed conditionally as follows:

$$\rho(y|x) = \text{NormalDen}(y; b_1x_1 + b_2x_2 + \cdots b_kx_k, \sigma_Y^2).$$

The linear relationship in Equality 5.1 has been used in causal models in economics [Joereskog, 1982], in structural equations in psychology [Bentler, 1980], and in path analysis in sociology and genetics [Kenny, 1979], [Wright, 1921].

Pearl [1988] developed an exact inference algorithm for Gaussian Bayesian networks. It is described in [Neapolitan, 2004]. Most Bayesian network inference algorithms handle Gaussian Bayesian networks. Some use the exact algorithm; others discretize the continuous distribution and then do inference using discrete variables. Next we show two examples.

Example 5.14 *HUGIN (www.hugin.com/) does exact inference in Gaussian Bayesian networks. Figure 5.22 shows the network in Figure 5.21 developed using HUGIN. The prior means and variances are shown under the DAG. Suppose now you just got your paycheck and it is only* \$300. *Your spouse becomes suspicious that you did not work very many hours. So your spouse instantiates* Y *to* 300 *in the network. The updated mean and variance of* X *are shown in Figure 5.22 under the priors. It turns out that the expected value of the hours you worked is only about* 32.64.

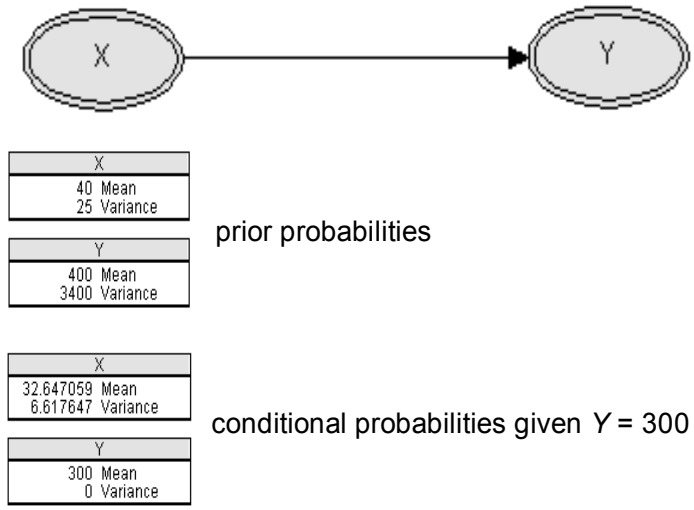

Figure 5.22: The Bayesian network in Figure 5.21, implemented in HUGIN.

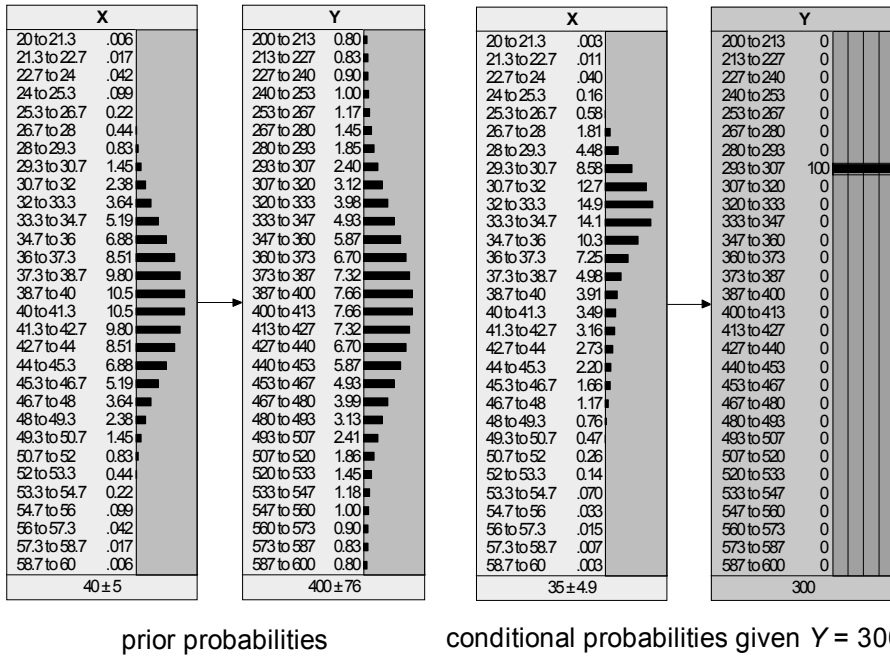

Figure 5.23: The Bayesian network in Figure 5.21, implemented in Netica.

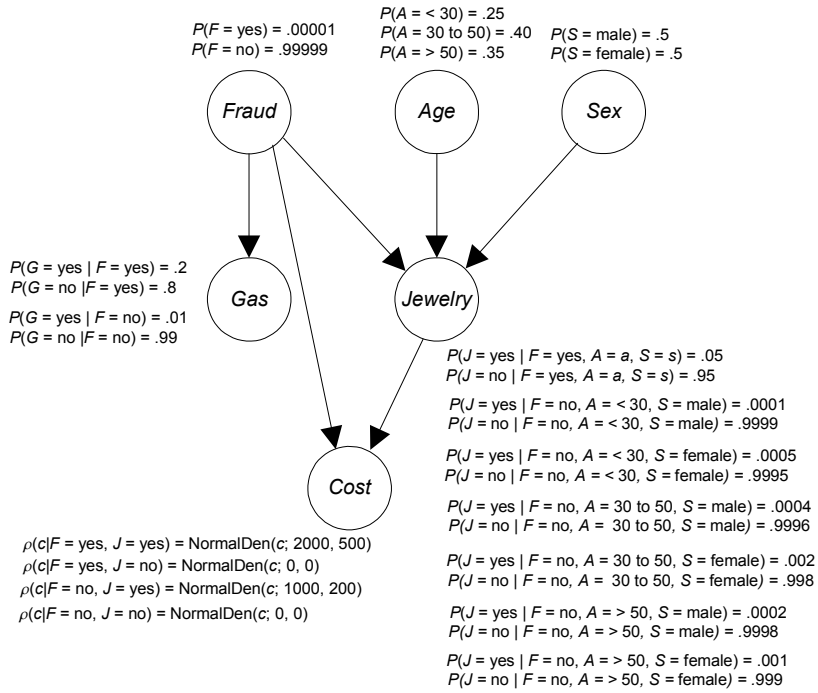

Figure 5.24: A hybrid Bayesian network modeling the situation in which the cost of the jewelry is likely to be higher if the purchase was fraudulent.

Netica (www.norsys.com/) requires that we discretize continuous variables. On the left in Figure 5.23 we show the network in Figure 5.21 with no nodes instantiated; on the right Y is instantiated to 300*. Notice that, owing to the approximation, we do not obtain the value of* 32.64 *for the updated mean of X. Rather, we obtain the value of* 35*. If we had a finer discretization, we should have obtained better results.*

5.5.2 Hybrid Networks

Hybrid Bayesian networks contain both discrete and continuous variables. Figure 5.24 shows a hybrid network, which will be discussed shortly. Methods for exact inference in hybrid Bayesian networks have been developed. For example, Shenoy [2006] develops a method that approximates general hybrid Bayesian networks by a mixture of Gaussian Bayesian networks. However, packages often deal with hybrid networks by discretizing the continuous distributions. HUGIN allows Gaussian variables to have discrete parents while still doing exact inference. It could, therefore, handle the Bayesian network in the following example.

Example 5.15 *Recall the Bayesian network in Figure 5.2, which models fraudulent use of a credit card. Suppose that if jewelry is purchased, the cost of the*

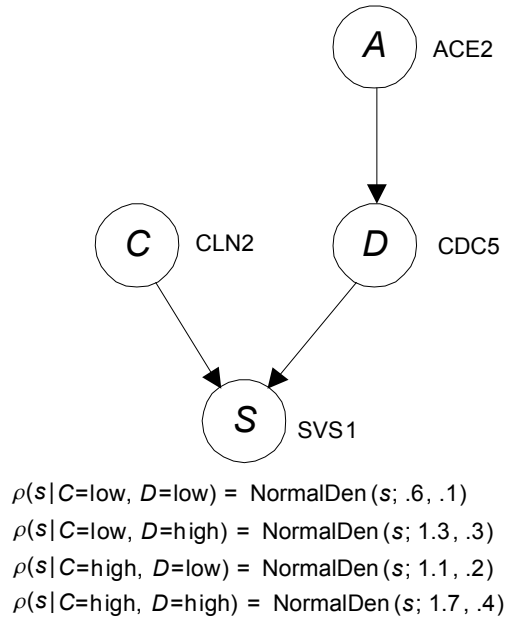

Figure 5.25: A Bayesian network showing possible causal relationships among the expression levels of genes. Only the conditional probability distribution of the leaf is shown.

jewelry is likely to be greater if the purchase was due to fraudulent use. We could model this situation using the hybrid Bayesian network in Figure 5.24. The variable Cost is normally distributed given each set of values of its discrete parents. Note that if $J = no$, the distribution is $\text{NormalDen}(s; 0, 0)$. *This is the same as stating that*

$$P(C = 0|F = yes, J = no) = 0.$$

However, we showed the conditional probability distribution as a normal distribution to be consistent with the other distributions of C.

Example 5.16 *Recall from Section 4.2.2 that the protein transcription factor produced by one gene can have a causal effect on the level of mRNA (called the gene expression level) of another gene. In Chapter 12 we discuss methods that use Bayesian networks to learn these causal effects from data. Gene expression level is often set as the ratio of measured expression to a control level. So, values greater than 1 would indicate a relatively high expression level, whereas values less than 1 would indicate a relatively low expression level. Since gene expression levels are continuous, we could try learning a Gaussian Bayesian network. Another approach taken in [Segal et al., 2005] is to learn a network in*

which each variable is normally distributed given values of its parents. However, each parent has only two values, namely high and low, which determine the conditional distribution of the child. The value high represents all expression levels greater than 1*, and the value low represents all expression levels less than or equal to* 1*. Such a network appears in Figure 5.25. The nodes in the network represent genes. This network is not exactly hybrid, because every variable is continuous. However, the conditional distributions are based on discrete values.*

5.6 How Do We Obtain the Probabilities?

So far we have simply shown the conditional probability distributions in the Bayesian networks we have presented. We have not been concerned with how we obtained them. For example, in the credit card fraud example we simply stated that $P(Age = less30) = .25$. However, how did we obtain this and other probabilities? As mentioned at the beginning of this chapter, they can either be obtained from the subjective judgements of an expert in the area, or they can be learned from data. In Chapter 8 we discuss techniques for learning them from data. Here, we show two techniques for simplifying the process of ascertaining them. The first technique concerns the case where a node has multiple parents, while the second technique concerns nodes that represent continuous random variables.

5.6.1 The Noisy OR-Gate Model

After discussing a problem in obtaining the conditional probabilities when a node has multiple parents, we present models that address this problem.

Difficulty Inherent in Multiple Parents

Suppose lung cancer, bronchitis, and tuberculosis all cause fatigue, and we need to model this relationship as part of a system for medical diagnosis. The portion of the DAG concerning only these four variables appears in Figure 5.26. We need to assess eight conditional probabilities for node F, one for each of the eight combinations of that node's parents. That is, we need to assess the following:

$$P(F = yes|B = no, T = no, L = no)$$

$$P(F = yes|B = no, T = no, L = yes)$$

$$\dots$$

$$P(F = yes|B = yes, T = yes, L = yes).$$

It would be quite difficult to obtain these values either from data or from an expert physician. For example, to obtain the value of $P(F = yes|B = yes, T = yes, L = no)$ directly from data, we would need a sufficiently large population of individuals who are known to have both bronchitis and tuberculosis, but not lung cancer. To obtain this value directly from an expert, the expert would

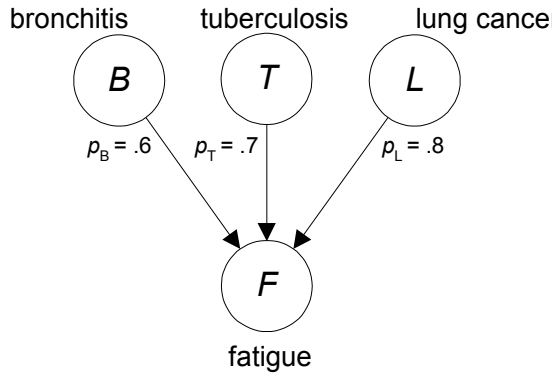

Figure 5.26: We need to assess eight conditional probabilities for node F.

have to be familiar with the likelihood of being fatigued when two diseases are present and the third is not. Next, we show a method for obtaining these conditional probabilities in an indirect way.

The Basic Noisy OR-Gate Model

The noisy OR-gate model concerns the case where the relationships between variables ordinarily represent causal influences, and each variable has only two values. The situation shown in Figure 5.26 is a typical example. Rather than assessing all eight probabilities, we assess the causal strength of each cause for its effect. The **causal strength** is the probability of the cause resulting in the effect whenever the cause is present. In Figure 5.26 we have shown the causal strength p_B of bronchitis for fatigue to be .6. *The assumption is that bronchitis will always result in fatigue unless some unknown mechanism inhibits this from taking place*, and this inhibition takes place 40% of the time. So 60% of the time bronchitis will result in fatigue. Presently, *we assume that all causes of the effect are articulated in the DAG*, and the effect cannot occur unless at least one of its causes is present. In this case, mathematically we have

$$p_B = P(F = yes|B = yes, T = no, L = no).$$

The causal strengths of tuberculosis and lung cancer for fatigue are also shown in Figure 5.26. These three causal strengths should not be as difficult to ascertain as all eight conditional probabilities. For example, to obtain p_B from data we only need a population of individuals who have lung bronchitis and do not have the other diseases. To obtain p_B from an expert, the expert need only ascertain the frequency with which bronchitis gives rise to fatigue.

We can obtain the eight conditional probabilities we need from the three causal strengths if we make one additional assumption. *We need to assume that the mechanisms that inhibit the causes act independently from each other.* For

example, the mechanism that inhibits bronchitis from resulting in fatigue acts independently from the mechanism that inhibits tuberculosis from resulting in fatigue. Mathematically, this assumption is as follows:

$$\begin{aligned} P(F = no|B = yes, T = yes, L = no) &= (1 - p_B)(1 - p_T) \\ &= (1 - .6)(1 - .7) = .12. \end{aligned}$$

Note that in the previous equality we are conditioning on bronchitis and tuberculosis both being present and lung cancer being absent. In this case, fatigue should occur unless the causal effects of bronchitis and tuberculosis are both inhibited. Since we have assumed these inhibitions act independently, the probability that both effects are inhibited is the product of the probabilities that each is inhibited, which is $(1 - p_B)(1 - p_T)$.

In this same way, if all three causes are present, we have

$$\begin{aligned} P(F = no|B = yes, T = yes, L = yes) &= (1 - p_B)(1 - p_T)(1 - p_L) \\ &= (1 - .6)(1 - .7)(1 - .8) = .024. \end{aligned}$$

Notice that when more causes are present, it is less probable that fatigue will be absent. This is what we would expect. In the following example we compute all eight conditional probabilities needed for node F in Figure 5.26.

Example 5.17 *Suppose we make the assumptions in the noisy OR-gate model, and the causal strengths of bronchitis, tuberculosis, and lung cancer for fatigue are the ones shown in Figure 5.26. Then*

$$P(F = no|B = no, T = no, L = no) = 1$$

$$\begin{aligned} P(F = no|B = no, T = no, L = yes) &= (1 - p_L) \\ &= (1 - .8) = .2 \end{aligned}$$

$$\begin{aligned} P(F = no|B = no, T = yes, L = no) &= (1 - p_T) \\ &= (1 - .7) = .3 \end{aligned}$$

$$\begin{aligned} P(F = no|B = no, T = yes, L = yes) &= (1 - p_T)(1 - p_L) \\ &= (1 - .7)(1 - .8) = .06 \end{aligned}$$

$$\begin{aligned} P(F = no|B = yes, T = no, L = no) &= (1 - p_B) \\ &= (1 - .6) = .4 \end{aligned}$$

$$\begin{aligned} P(F = no|B = yes, T = no, L = yes) &= (1 - p_B)(1 - p_L) \\ &= (1 - .6)(1 - .8) = .08 \end{aligned}$$

$$\begin{aligned} P(F = no|B = yes, T = yes, L = no) &= (1 - p_B)(1 - p_T) \\ &= (1 - .6)(1 - .7) = .12 \end{aligned}$$

$$\begin{aligned} P(F = no|B = yes, T = yes, L = yes) &= (1 - p_B)(1 - p_T)(1 - p_L) \\ &= (1 - .6)(1 - .7)(1 - .8) = .024. \end{aligned}$$

Note that since the variables are binary, these are the only values we need to ascertain. The remaining probabilities are uniquely determined by these. For example,

$$P(F = yes|B = yes, T = yes, L = yes) = 1 - .024 = .976.$$

Although we illustrated the model for three causes, it clearly extends to an arbitrary number of causes. We showed the assumptions in the model in italics when we introduced them. Next, we summarize them and show the general formula.

The **noisy OR-gate model** makes the following three assumptions:

1. **Causal inhibition:** This assumption entails that there is some mechanism which inhibits a cause from bringing about its effect, and the presence of the cause results in the presence of the effect if and only if this mechanism is disabled (turned off).

2. **Exception independence:** This assumption entails that the mechanism that inhibits one cause is independent of the mechanism that inhibits other causes.

3. **Accountability:** This assumption entails that an effect can happen only if at least one of its causes is present and is not being inhibited.

The **general formula for the noisy OR-gate model** is as follows: Suppose Y has n causes $X_1, X_2, \ldots, X_n$, all variables are binary, and we assume the noisy OR-gate model. Let p_i be the causal strength of X_i for Y. That is,

$$p_i = P(Y = yes|X_1 = no, X_2 = no, \ldots X_i = yes, \ldots X_n = no).$$

Then if X is a set of nodes that are instantiated to yes,

$$P(Y = no|\mathsf{X}) = \prod_{i \text{ such that } X_i \in \mathsf{X}} (1 - p_i).$$

The Leaky Noisy OR-Gate Model

Of the three assumptions in the noisy OR-gate model, the assumption of accountability seems to be justified least often. For example, in the case of fatigue there are certainly other causes of fatigue such as listening to a lecture by Professor Neapolitan. So the model in Figure 5.26 does not contain all causes of fatigue, and the assumption of accountability is not justified. It seems in many, if not most, situations we would not be certain that we have elaborated all known causes of an effect. Next, we show a version of the model that does not assume accountability. The derivation of the formula for this model is not simple and intuitive like the one for the basic noisy OR-gate model. So we first present the model without deriving it and then show the derivation.

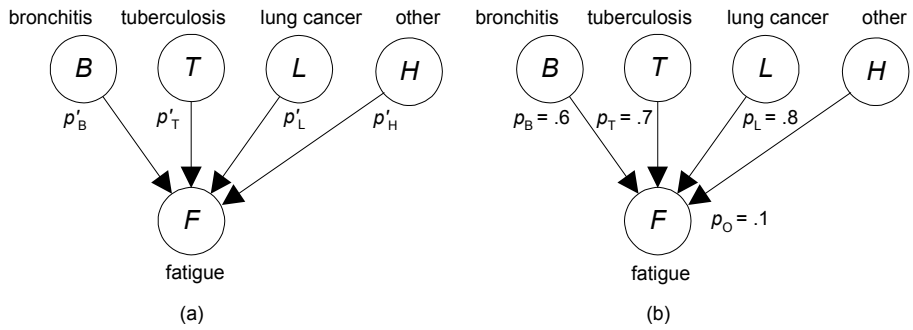

Figure 5.27: The probabilities in (a) are the causal strengths in the noisy OR-gate model. The probabilities in (b) are the ones we ascertain.

The Leaky Noisy OR-Gate Formula The leaky noisy OR-gate model assumes that all causes that have not been articulated can be grouped into one other cause H and that the articulated causes, along with H, satisfy the three assumptions in the noisy OR-gate model. This is illustrated for the fatigue example in Figure 5.27 (a). The probabilities in that figure are the causal strengths in the noisy OR-gate model. For example,

$$p'_B = P(F = yes|B = yes, T = no, L = no, H = no).$$

We could not ascertain these values because we do know whether or not H is present. The probabilities in Figure 5.27 (b) are the ones we could ascertain. For each of the three articulated causes, the probability shown is the probability the effect is present given the remaining two articulated causes are not present. For example,

$$p_B = P(F = yes|B = yes, T = no, L = no).$$

Note the difference in the probabilities p'_B and p_B. The latter one does not condition on a value of H, while the former one does. The probability p_0 is different from the other probabilities. It is the probability that the effect will be present given none of the articulated causes are present. That is,

$$p_0 = P(F = yes|B = no, T = no, L = no).$$

Note again that we are not conditioning on a value of H.

Our goal is to develop conditional probability distributions for a Bayesian network containing the nodes B, T, L, and F from the probabilities ascertained in Figure 5.27 (b). We show an example that realizes this goal after presenting the formula necessary to the task.

The **general formula for the leaky noisy OR-gate model** is as follows (a derivation appears in the next subsection): Suppose Y has n causes $X_1, X_2, \ldots, X_n$, all variables are binary, and we assume the leaky noisy OR-gate model. Let

$$p_i = P(Y = yes|X_1 = no, X_2 = no, \ldots X_i = yes, \ldots X_n = no) \qquad (5.2)$$

$$p_0 = P(Y = yes | X_1 = no, X_2 = no, \ldots X_n = no). \tag{5.3}$$

Then if X is a set of nodes that are instantiated to yes,

$$P(Y = no|\mathsf{X}) = (1 - p_0) \prod_{i \text{ such that } X_i \in \mathsf{X}} \frac{1 - p_i}{1 - p_0}.$$

Example 5.18 *Let's compute the conditional probabilities for a Bayesian network containing the nodes B, T, L, and F from the probabilities ascertained in Figure 5.27 (b). We have*

$$\begin{aligned}
P(F = no|B = no, T = no, L = no) &= 1 - p_0 \\
&= 1 - .1 = .9 \\
\\
P(F = no|B = no, T = no, L = yes) &= (1 - p_0)\frac{1 - p_L}{1 - p_0} \\
&= 1 - .8 = .2 \\
\\
P(F = no|B = no, T = yes, L = no) &= (1 - p_0)\frac{1 - p_T}{1 - p_0} \\
&= 1 - .7 = .3 \\
\\
P(F = no|B = no, T = yes, L = yes) &= (1 - p_0)\frac{1 - p_T}{1 - p_0}\frac{1 - p_L}{1 - p_0} \\
&= \frac{(1 - .7)(1 - .8)}{1 - .1} = .067 \\
\\
P(F = no|B = yes, T = no, L = no) &= (1 - p_0)\frac{1 - p_B}{1 - p_0} \\
&= 1 - .6 = .4 \\
\\
P(F = no|B = yes, T = no, L = yes) &= (1 - p_0)\frac{1 - p_B}{1 - p_0}\frac{1 - p_L}{1 - p_0} \\
&= \frac{(1 - .6)(1 - .8)}{1 - .1} = .089 \\
\\
P(F = no|B = yes, T = yes, L = no) &= (1 - p_0)\frac{1 - p_B}{1 - p_0}\frac{1 - p_T}{1 - p_0} \\
&= \frac{(1 - .6)(1 - .7)}{1 - .1} = .133 \\
\\
P(F = no|B = yes, T = yes, L = yes) &= (1 - p_0)\frac{1 - p_B}{1 - p_0}\frac{1 - p_T}{1 - p_0}\frac{1 - p_L}{1 - p_0} \\
&= \frac{(1 - .6)(1 - 7)(1 - .8)}{(1 - .1)(1 - .1)} = .030.
\end{aligned}$$

A Derivation of the Formula The following lemmas and theorem derive the formula in the leaky noisy OR-gate model.

Lemma 5.1 *Given the assumptions and notation shown above for the leaky noisy OR-gate model,*

$$p_0 = p'_H \times P(H = yes).$$

Proof. *Owing to Equality 5.3, we have*

$$\begin{aligned} p_0 &= P(Y = yes|X_1 = no, X_2 = no, \ldots X_n = no) \\ &= P(Y = yes|X_1 = no, X_2 = no, \ldots X_n = no, H = yes)P(H = yes) + \\ &\quad P(Y = yes|X_1 = no, X_2 = no, \ldots X_n = no, H = no)P(H = no) \\ &= p'_H \times P(H = yes) + 0 \times P(H = no). \end{aligned}$$

This completes the proof. ■

Lemma 5.2 *Given the assumptions and notation shown above for the leaky noisy OR-gate model,*

$$1 - p'_i = \frac{1 - p_i}{1 - p_0}.$$

Proof. *Owing to Equality 5.2, we have* ■

$$\begin{aligned} & 1 - p_i \\ &= P(Y = no|X_1 = no, \ldots X_i = yes, \ldots X_n = no) \\ &= P(Y = no|X_1 = no, \ldots X_i = yes, \ldots X_n = no, H = yes)P(H = yes) + \\ &\quad P(Y = no|X_1 = no, \ldots X_i = yes, \ldots X_n = no, H = no)P(H = no) \\ &= (1 - p'_i)(1 - p'_H)P(H = yes) + (1 - p'_i)P(H = no) \\ &= (1 - p'_i)(1 - p'_H \times P(H)) \\ &= (1 - p'_i)(1 - p_0). \end{aligned}$$

The last equality is due to Lemma 5.1. This completes the proof.

Theorem 5.2 *Given the assumptions and notation shown above for the leaky noisy OR-gate model,*

$$P(Y = no|\mathsf{X}) = (1 - p_0) \prod_{i \text{ such that } X_i \in \mathsf{X}} \frac{1 - p_i}{1 - p_0}.$$

Proof. *We have*

$$\begin{aligned} P(Y = no|\mathsf{X}) &= P(Y = no|\mathsf{X}, H = yes)P(H = yes) + \\ &\quad P(Y = no|\mathsf{X}, H = yes)P(H = yes) \\ &= P(H = yes)(1 - p'_H) \prod_{i \text{ such that } X_i \in \mathsf{X}} (1 - p'_i) + \\ &\quad P(H = no) \prod_{i \text{ such that } X_i \in \mathsf{X}} (1 - p'_i) \\ &= (1 - p_0) \prod_{i \text{ such that } X_i \in \mathsf{X}} (1 - p'_i) \\ &= (1 - p_0) \prod_{i \text{ such that } X_i \in \mathsf{X}} \frac{1 - p_i}{1 - p_0}. \end{aligned}$$

The second to the last equality is due to Lemma 5.1, and the last is due to Lemma 5.2. ■

Further Models

A generalization of the noisy OR-gate model to the case of more than two values appears in [Srinivas, 1993]. Diez and Druzdzel [2002] propose a general framework for canonical models, classifying them into three categories: deterministic, noisy, and leaky. They then analyze the most common families of canonical models, namely the noisy OR/MAX, the noisy AND/MIN, and the noisy XOR. Other models for succinctly representing the conditional distributions use the **sigmoid** function [Neal, 1992] and the **logit** function [McLachlan and Krishnan, 1997]. Another approach to reducing the number of parameter estimates is the use of **embedded Bayesian networks**, which is discussed in [Heckerman and Meek, 1997].

5.6.2 Methods for Discretizing Continuous Variables

Often, a Bayesian network contains both discrete and continuous random variables. For example, the Bayesian network in Figure 5.2 contains four random variables that are discrete and one, namely *Age*, that is continuous.[6] However, notice that a continuous probability density function for node *Age* does not appear in the network. Rather, the possible values of the node are three ranges for ages, and the probability of each of these ranges is specified in the network. This is called **discretizing** the continuous variables. Although many Bayesian network inference packages allow the user to specify both continuous variables and discrete variables in the same network, we can sometimes obtain simpler and better inference results by representing the variables as discrete. One reason for this is that, if we discretize the variables, we do not need to assume any particular continuous probability density function. Next, we present two of the most popular methods for discretizing continuous random variables.

[6] Technically, if we count age only by years it is discrete. However, even in this case, it is usually represented by a continuous distribution because there are so many values.

Bracket Medians Method

In the **Bracket Medians Method** the mass in a continuous probability distribution function $F(x) = P(X \leq x)$ is divided into n equally spaced intervals. The method proceeds as follows ($n = 5$ in this explanation):

1. Determine n equally spaced intervals in the interval $[0, 1]$. If $n = 5$, the intervals are $[0, .2]$, $[.2, .4]$, $[.4, .6]$, $[.6, .8]$, and $[.8, 1.0]$.

2. Determine points x_1, x_2, x_3, x_4, x_5, and x_6 such that

$$\begin{aligned} P(X \leq x_1) &= .0 \\ P(X \leq x_2) &= .2 \\ P(X \leq x_3) &= .4 \\ P(X \leq x_4) &= .6 \\ P(X \leq x_5) &= .8 \\ P(X \leq x_6) &= 1.0, \end{aligned}$$

   where the values on the right in these equalities are the endpoints of the five intervals.

3. For each interval $[x_i, x_{i+1}]$ compute the bracket median d_i, which is the value such that

$$P(x_i \leq X \leq d_i) = P(d_i \leq X \leq x_{i+1}).$$

4. Define the discrete variable D with the following probabilities:

$$\begin{aligned} P(D = d_1) &= .2 \\ P(D = d_2) &= .2 \\ P(D = d_3) &= .2 \\ P(D = d_4) &= .2 \\ P(D = d_5) &= .2. \end{aligned}$$

Example 5.19 *Recall that the normal density function is given by*

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad -\infty < x < \infty,$$

where

$$E(X) = \mu \qquad \text{and} \qquad Var(X) = \sigma^2,$$

and the cumulative distribution function for this density function is given by

$$F(x) = P(X \leq x) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{x} e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt \qquad -\infty < x < \infty.$$

These functions for $\mu = 50$ and $\sigma = 15$ are shown in Figure 5.28. This might be the distribution of age for some particular population. Next we use the Bracket Medians Method to discretize it into three ranges. Then $n = 3$ and our four steps are as follows:

1. *Since there is essentially no mass < 0 or > 100, our three intervals are $[0, .333]$, $[.333, .666]$, and $[.666, 1]$.*

2. *We need to find points x_1, x_2, x_3, and x_4 such that*

$$\begin{aligned} P(X \leq x_1) &= .0 \\ P(X \leq x_2) &= .333 \\ P(X \leq x_3) &= .666 \\ P(X \leq x_4) &= 1. \end{aligned}$$

   Clearly, $x_1 = 0$ and $x_4 = 100$. To determine x_2 we need to determine

$$x_2 = F^{-1}(.333).$$

   Using the mathematics package Maple, we have

$$x_2 = \text{NormalInv}(.333; 50, 15) = 43.5.$$

   Similarly,

$$x_3 = \text{NormalInv}(.666; 50, 15) = 56.4.$$

   In summary, we have

$$x_1 = 0 \qquad x_2 = 43.5 \qquad x_3 = 56.4 \qquad x_4 = 1.$$

3. *Compute the bracket medians. We compute them using Maple by solving the following equations:*

$$\text{NormalDist}(d_1; 50, 15)$$
$$= \text{NormalDist}(43.5; 50, 15) - \text{NormalDist}(d_1; 50, 15)$$

   Solution is $d_1 = 35.5$.

$$\text{NormalDist}(d_2; 50, 15) - \text{NormalDist}(43.5; 50, 15)$$
$$= \text{NormalDist}(56.4; 50, 15) - \text{NormalDist}(d_2; 50, 15)$$

   Solution is $d_2 = 50.0$.

$$\text{NormalDist}(d_3; 50, 15) - \text{NormalDist}(56.4; 50, 15)$$
$$= 1 - \text{NormalDist}(d_3; 50, 15)$$

   Solution is $d_3 = 64.5$.

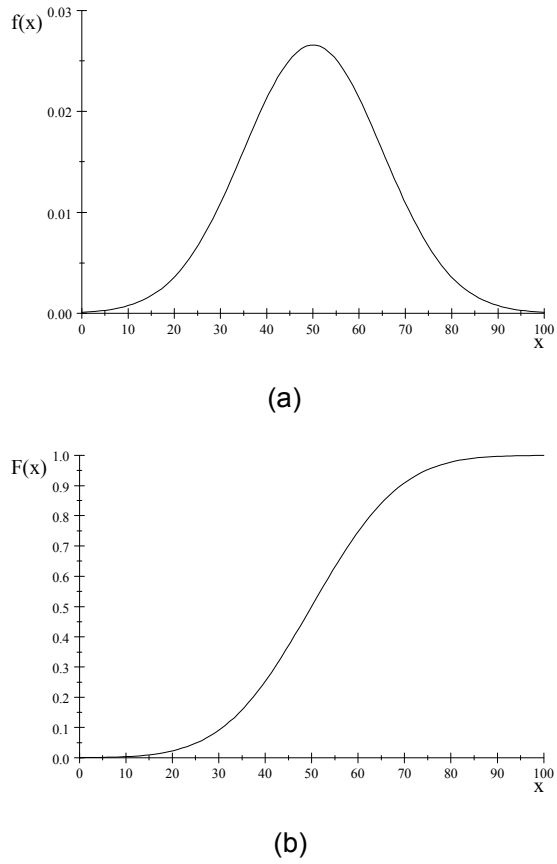

Figure 5.28: The normal density function with $\mu = 50$ and $\sigma = 15$ appears in (a), while the corresponding normal cumulative distribution function appears in (b).

4. *Finally, we set*

$$\begin{aligned} P(D = 35.5) &= .333 \\ P(D = 50.0) &= .333 \\ P(D = 64.5) &= .333. \end{aligned}$$

If, for example, a data item's continuous value is between 0 *and* 43.5, *we assign the data item a discrete value of* 35.5.

The variable D requires a numeric value if we need to perform computations using it. However, if the variable does not require a numeric value for computational purposes, we need not perform Step 3 in the Bracket Medians Method.

Rather, we just show ranges as the values of D. In the previous example, we would set

$$\begin{aligned} P(D = \ < 43.5) &= .333 \\ P(D = 43.5 \text{ to } 56.4) &= .333 \\ P(D = \ > \ 56.4) &= .333. \end{aligned}$$

Recall that this is what we did for *Age* in the Bayesian network in Figure 5.2. In this case, if a data item's continuous value is between 0 and 43.5, we simply assign the data item that range.

Pearson-Tukey Method

In some applications we want to give special attention to the case when a data item falls in the tail of a density function. For example, if we are trying to predict whether a company will go bankrupt, then unusually low cash flow is indicative they will, while unusually high cash flow is indicative they will not [McKee and Lensberg, 2002]. Values in the middle are not indicative one way or the other. In such cases we want to group the values in each tail together. The Bracket Medians Method does not do this. However, the Pearson-Tukey Method [Keefer, 1983], which we describe next, does. Neapolitan and Jiang [2007] discuss the bankruptcy prediction application in detail.

In the **Pearson-Tukey Method** the mass in a continuous probability distribution function $F(x) = P(X \leq x)$ is divided into three intervals. The method proceeds as follows:

1. Determine points x_1, x_2, and x_3 such that

$$\begin{aligned} P(X \leq x_1) &= .05 \\ P(X \leq x_2) &= .50 \\ P(X \leq x_3) &= .95. \end{aligned}$$

2. Define the discrete variable D with the following probabilities:

$$\begin{aligned} P(D = x_1) &= .185 \\ P(D = x_2) &= .63 \\ P(D = x_3) &= .185. \end{aligned}$$

Example 5.20 *Suppose we have the normal distribution discussed in Example 5.19. Next, we apply the Pearson-Tukey Method to that distribution.*

1. *Using Maple, we have*

$$x_1 = \text{NormalInv}(.05; 50, 15) = 25.3$$

$$x_2 = \text{NormalInv}(.50; 50, 15) = 50$$

$$x_3 = \text{NormalInv}(.95; 50, 15) = 74.7.$$

2. *We set*

$$\begin{aligned} P(D = 25.3) &= .185 \\ P(D = 50.0) &= .63 \\ P(D = 74.7) &= .185. \end{aligned}$$

To assign data items discrete values, we need to determine the range of values corresponding to each of the cutoff points. That is, we compute the following:

$$\text{NormalInv}(.185; 50, 15) = 36.6$$

$$\text{NormalInv}(1 - .185; 50, 15) = 63.4.$$

If a data item's continuous value is < 36.6*, we assign the data item the value* 25.3*; if the value is in* $[36.6, 63.4]$*, we assign the value* 50*; and if the value is* > 63.4*, we assign the value* 74.7*.*

Notice that when we used the Pearson-Tukey Method, the middle discrete value represented numbers in the interval $[36.6, 63.4]$, while when we used the Bracket's Median Method, the middle discrete value represented numbers in the interval $[43.5, 56.4]$. The interval for the Pearson-Tukey Method is larger, meaning more numbers in the middle are treated as the same discrete value, and the other two discrete values represent values only in the tails.

If the variable does not require a numeric value for computational purposes, we need not perform Steps 1 and 2, but rather just determine the range of values corresponding to each of the cutoff points and just show ranges as the values of D. In the previous example, we would set

$$\begin{aligned} P(D = \ < 36.6) &= .185 \\ P(D = 36.6 \text{ to } 63.4) &= .63 \\ P(D = \ > \ 63.4) &= .185. \end{aligned}$$

In this case if a data item's continuous value is between 0 and 36.6, we simply assign the data item that range.

EXERCISES

Section 5.2

Exercise 5.1 In Example 5.3 it was left as an exercise to show for all values of s, l, and c that

$$P(s, l, c) = P(s|c)P(l|c)P(c).$$

Show this.

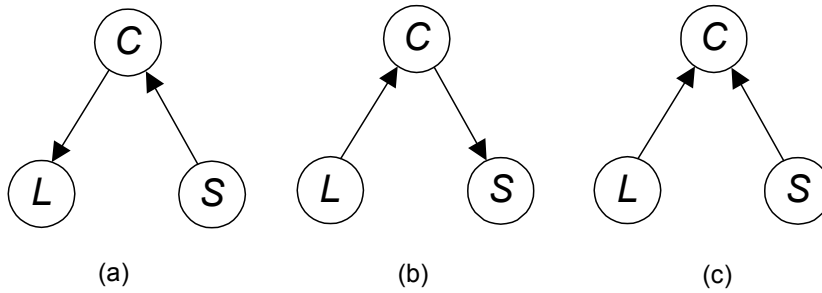

Figure 5.29: The probability distribution discussed in Example 5.1 satisfies the Markov condition with the DAGs in (a) and (b), but not with the DAG in (c).

Exercise 5.2 Consider the joint probability distribution P in Example 5.1.

1. Show that probability distribution P satisfies the Markov condition with the DAG in Figure 5.29 (a) and that P is equal to the product of its conditional distributions in that DAG.
2. Show that P probability distribution satisfies the Markov condition with the DAG in Figure 5.29 (b) and that P is equal to the product of its conditional distributions in that DAG.
3. Show that P probability distribution does not satisfy the Markov condition with the DAG in Figure 5.29 (c) and that P is not equal to the product of its conditional distributions in that DAG.

Section 5.3

Exercise 5.3 Professor Morris investigated gender bias in hiring in the following way. He gave hiring personnel equal numbers of male and female résumés to review, and then he investigated whether their evaluations were correlated with gender. When he submitted a paper summarizing his results to a psychology journal, the reviewers rejected the paper because they said this was an example of fat-hand manipulation. Investigate the concept of fat-hand manipulation, and explain why the journal reviewers might have thought this.

Exercise 5.4 Consider the following piece of medical knowledge: tuberculosis and lung cancer can each cause shortness of breath (dyspnea) and a positive chest X-ray. Bronchitis is another cause of dyspnea. A recent visit to Asia could increase the probability of tuberculosis. Smoking can cause both lung cancer and bronchitis. Create a DAG representing the causal relationships among these variables. Complete the construction of a Bayesian network by

determining values for the conditional probability distributions in this DAG, either based on your own subjective judgement or from data.

Exercise 5.5 Explain why, if we reverse the edges in the DAG in Figure 5.12 to obtain the DAG $E \rightarrow D \rightarrow T$, the new DAG also satisfies the Markov condition with the probability distribution of the variables.

Section 5.4

Exercise 5.6 Compute $P(x_1|w_1)$, assuming the Bayesian network in Figure 5.16.

Exercise 5.7 Compute $P(t_1|w_1)$, assuming the Bayesian network in Figure 5.17.

Exercise 5.8 Compute $P(x_1|t_2, w_1)$, assuming the Bayesian network in Figure 5.17.

Exercise 5.9 Using Netica, develop the Bayesian network in Figure 5.2, and use that network to determine the following conditional probabilities.

1. $P(F = yes|Sex = male)$. Is this conditional probability different from $P(F = yes)$? Explain why it is or is not.
2. $P(F = yes|J = yes)$. Is this conditional probability different from $P(F = yes)$? Explain why it is or is not.
3. $P(F = yes|Sex = male, J = yes)$. Is this conditional probability different from $P(F = yes|J = yes)$? Explain why it is or is not.
4. $P(G = yes|F = yes)$. Is this conditional probability different from $P(G = yes)$? Explain why it is or is not.
5. $P(G = yes|J = yes)$. Is this conditional probability different from $P(G = yes)$? Explain why it is or is not.
6. $P(G = yes|J = yes, F = yes)$. Is this conditional probability different from $P(G = yes|F = yes)$? Explain why it is or is not.
7. $P(G = yes|A = < 30)$. Is this conditional probability different from $P(G = yes)$? Explain why it is or is not.
8. $P(G = yes|A = < 30, J = yes)$. Is this conditional probability different from $P(G = yes|J = yes)$? Explain why it is or is not.

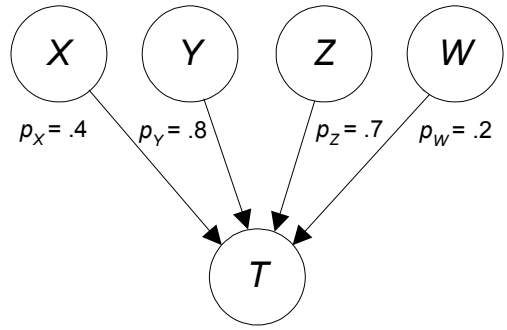

Figure 5.30: The noisy OR-gate model is assumed.

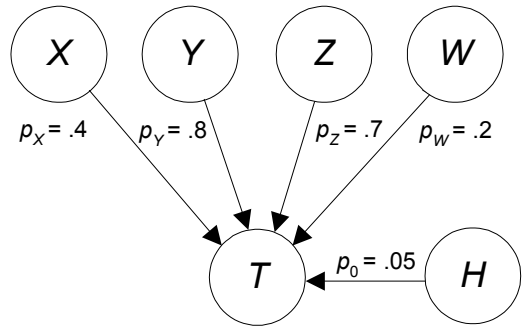

Figure 5.31: The leaky noisy OR-gate model is assumed.

Section 5.5

Exercise 5.10 Using Netica, HUGIN, or some other Bayesian network software package, implement the Bayesian network in Figure 5.24. Using that network compute the conditional probabilities in Exercise 5.9. Compare your answers to those obtained in Exercise 5.9.

Section 5.6

Exercise 5.11 Assume the noisy OR-gate model and the causal strengths are those shown in Figure 5.30. Compute the probability $T = yes$ for all combinations of values of the parents.

Exercise 5.12 Assume the leaky noisy OR-gate model and the relevant probabilities are those shown in Figure 5.31. Compute the probability $T = yes$ for

all combinations of values of the parents. Compare the results to those obtained in Exercise 5.11.

Exercise 5.13 Suppose we have the normal density function with $\mu = 100$ and $\sigma = 20$.

1. Discretize this function into four ranges using the Brackets Median Method.
2. Discretize this function using the Pearson-Tukey Method.