Chapter 3

Bayesian Networks

The Reverend Thomas Bayes (1702-1761) developed Bayes' Theorem in the 18th 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. We introduce these networks next. Applications of Bayesian networks to finance and marketing appear in Parts II and III. In Sections 3.1 and 3.2 we define Bayesian networks and discuss their properties. Section 3.3 shows how causal graphs often yield Bayesian networks. In Section 3.4 we discuss doing probabilistic inference using Bayesian networks. Section 3.5 concerns obtaining the conditional probabilities necessary to a Bayesian network. Finally, Section 3.6 shows the conditional independencies entailed by a Bayesian network.

3.1 What Is a Bayesian Network?

Recall that in Chapter 2, Example 2.29, 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

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

and 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 3.1, 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 3.1 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 non-root variable given each set of values of its parents. We use the terms "node" and "variable" interchangeably when discussing Bayesian networks.

Let's illustrate a more complex Bayesian network by considering the problem of detecting credit card fraud (taken from [Heckerman, 1996]). Let's say 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,

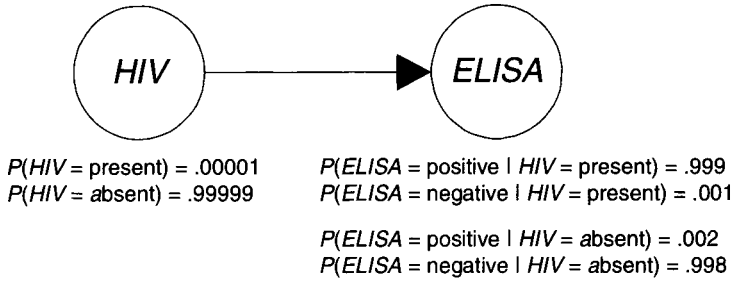

Figure 3.1: A two-node Bayesian network.

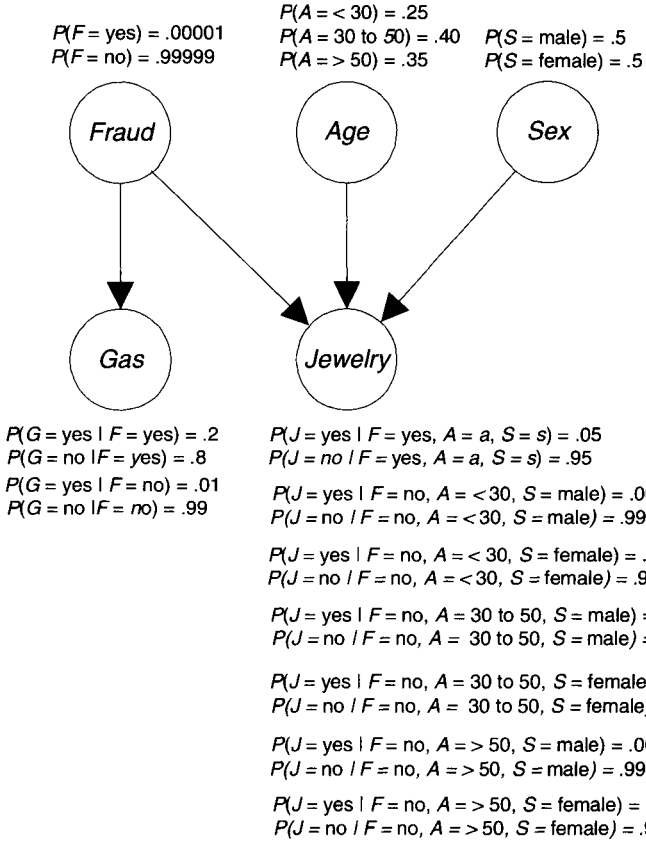

Figure 3.2: A Bayesian network for detecting credit card fraud.

while young men are least likely to buy jewelry. Figure 3.2 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 may have a few questions concerning this Bayesian network. First, you may 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 3.1. 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 3.1. Rather it is done using sophisticated algorithms. Second, you may 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 4 we discuss techniques for learning them from data, while in Parts II and III we show examples of obtaining them from experts and learning them from data. Finally, you may 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 buy jewelry. That is, the fact that the card holder is a middle-aged woman explains away 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.

3.2 Properties of Bayesian Networks

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

3.2.1 Definition of a Bayesian Network

First, let's review the definition of a DAG. A **directed graph** is a pair (V, E), where V is a finite, nonempty set whose elements are called **nodes** (or vertices),

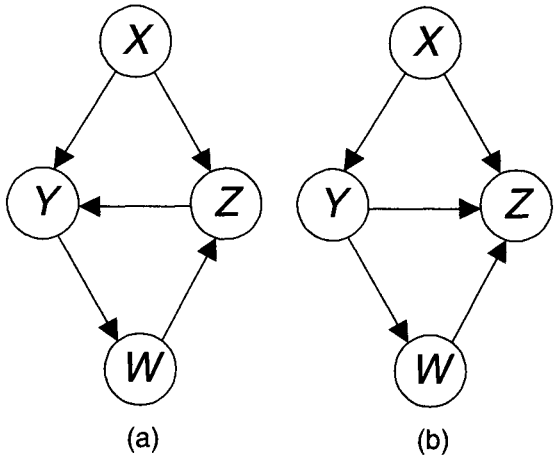

Figure 3.3: Both graphs are directed graphs; only the one in (b) is a directed acyclic graph.

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. Figure 3.3 (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 3.3 (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 3.3 (b), but it is not a path. A **cycle** in a directed graph is a path from a node to itself. In Figure 3.3 (a) $[Y, W, Z, Y]$ is a cycle from Y to Y. However, in Figure 3.3 (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 3.3 (b) is a DAG, while the one in Figure 3.3 (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 **nondescendent** of X if Y is not a descendent of X and Y is not a parent of X.[1]

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

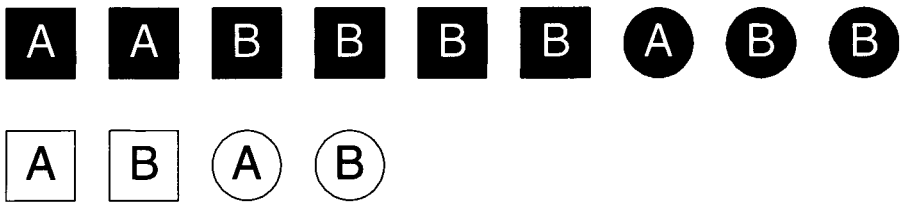

Figure 3.4: The random variables L and S are not independent, but they are conditionally independent given C.

We can now state the following definition:

Definition 3.1 *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 and ND, respectively, then*

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

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

Example 3.1 *Recall Chapter 2, Figure 2.1, which appears again as Figure 3.4. In Chapter 2, Example 2.21 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 3.5. For that DAG we have the following:

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

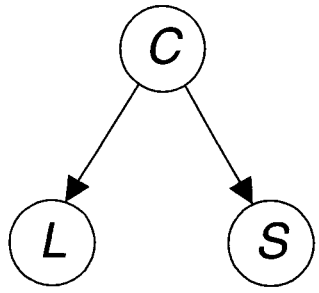

Figure 3.5: The joint probability distribution of L, S, and C constitutes a Bayesian network with this DAG.

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 3.2 *Consider the DAG $\mathbb{G}$ in Figure 3.6. 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, W	$I_P(W, \{X, W\}\|\{Y, Z\})$
V	Z	X, Y, W	$I_P(V, \{X, Y, W\}\|Z)$

3.2.2 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 3.1 and 3.2 do we show a Bayesian network as a DAG and a set of conditional 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 this theorem.

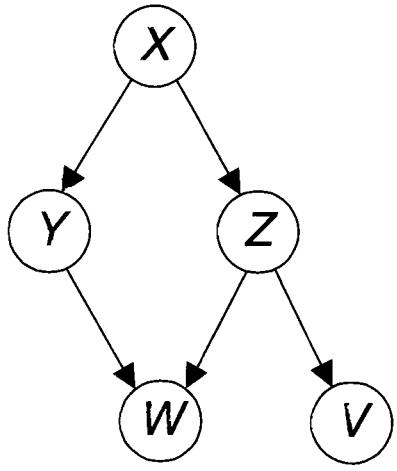

Figure 3.6: A DAG illustrating the Markov condition.

Theorem 3.1 *$(\mathbb{G}, P)$ 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 3.3 *We showed that the joint probability distribution P of the random variables L, S, and C defined on the set of objects in Figure 3.4 constitutes a Bayesian network with the DAG $\mathbb{G}$ in Figure 3.5. 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 3.7 shows those conditional distributions. We computed them directly from Figure 3.4. For example, since there are 9 black objects (c_1) and 6 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 l, s, and c that

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

There are a total of eight combinations of the three variables. 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 that end, we have directly from Figure 3.4 that

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

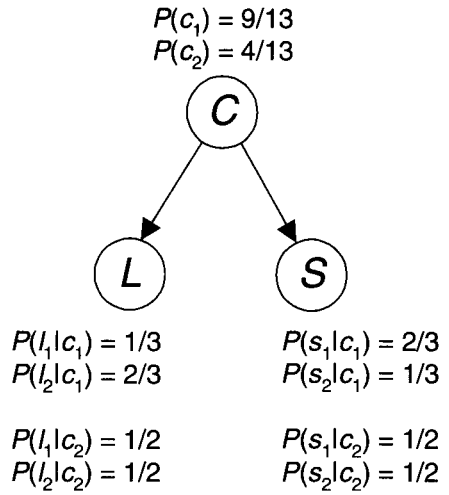

Figure 3.7: A Bayesian network representing the probability distribution P of the random variables L, S, and C defined on the set of objects in Figure 3.4.

From Figure 3.7 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 3.1, 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 is done in Figures 3.1, 3.2, and 3.7. Herein lies the representational power of Bayesian networks. If there are 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 3.8. 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 3.7 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 3.8, but still 1023 in the joint distribution. We see then that a Bayesian network is a structure for representing a joint probability distribution succinctly.

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.

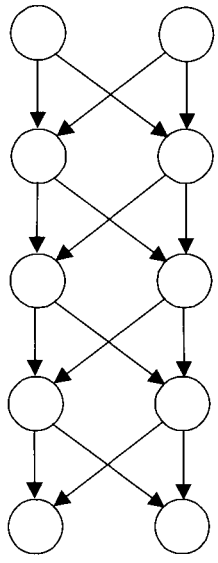

Figure 3.8: If all variables are binary, and a joint distribution satisifes the Markov condition with this DAG, there are 1024 values in the joint distribution, but only 68 values in the conditional distributions.

Example 3.4 *Consider again the joint probability distribution P of the random variables L, S, and C defined on the set of objects in Figure 3.4. Figure 3.9 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 3.4 that*

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

From Figure 3.9 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 we are left with a conundrum. That is, our goal is to represent a joint probability distribution succinctly using a DAG and conditional 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

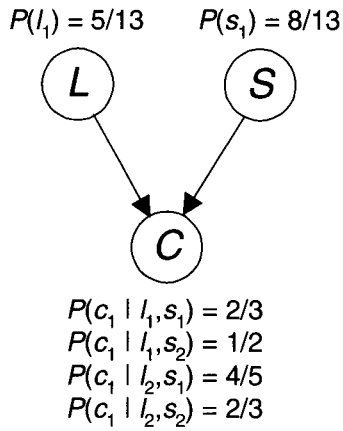

Figure 3.9: The joint probability distribution P of the random variables L, S, and C defined on the set of objects in Figure 3.4 does not satisfy the Markov condition with this DAG.

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 3.1 and 3.2 are causal, while 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 4.

3.3 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.

3.3.1 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 this definition is useful, it 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 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 wish to consider. Our random variables are features of these entities. Next, we ascertain the causal relationship we wish 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.

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 they conclude 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 next.

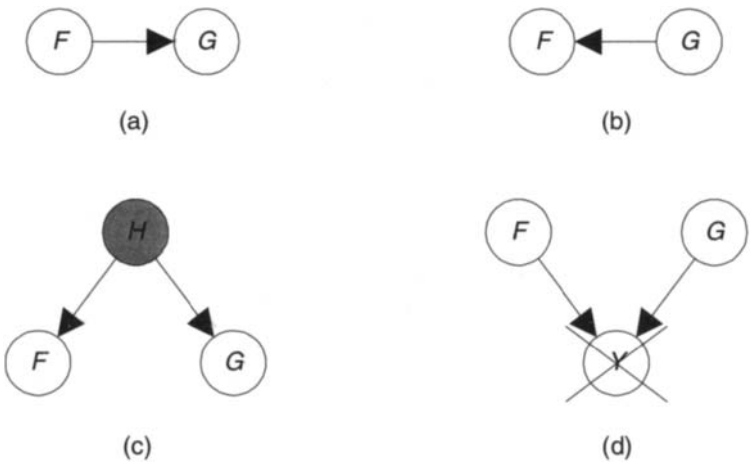

Figure 3.10: The edges in this graphs represent causal influences. All four causal relationships could account for F and G being correlated.

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 3.10 shows these three possibilities plus one more. Figure 3.10 (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 3.10 (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 Figure 3.10 (a) and (b) are equally reasonable. Even in this domain, G causing F seems possible. A man may have used some other hair regrowth product such as minoxidil which caused him to regrow hair, became excited about the regrowth, and decided to try other products such as finasteride which he heard might cause regrowth. A third possibility, shown in Figure 3.10 (c), is that F and G have some hidden common cause H which 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 may cause hair regrowth, while the finasteride may not. In this case the man's concern is a cause of finasteride use and hair regrowth (indirectly through minoxidil use), while the latter 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 are both causes of hypertension,[2] 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 3.10 (d), where the cross through Y means the variable is instantiated. Ordinarily, the instantiation of a common effect creates a dependency between its causes because each cause explains away 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**.[3] A final possibility (not depicted in Figure 3.10) is that F and G are not causally related at all. The most notable example of this situation is when our entities are points in time, and our random variables are values of properties at these different points in time. Such random 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.[4] 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 wish to note the difficulty with such correlations. In light of all of the above, we see then that we cannot deduce the causal relationship between two variables from the mere fact that they are statistically correlated.

Note that any of the four causal relationships shown in Figure 3.10 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 use of finasteride, meaning we could have a causal loop or feedback. Therefore, we would have the causal relationships in both Figure 3.10 (a) and Figure 3.10 (b).

It may 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.

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

[3] This could happen if our sample is a **convenience sample**, which is a sample where the participants are selected at the convenience of the researcher. The researcher makes no attempt to insure 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.

[4] Unfortunately, his hairline did not go back down in 2003.

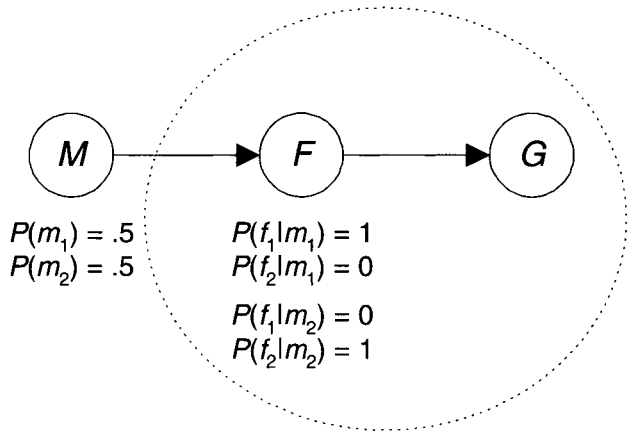

Figure 3.11: An RCE investigating whether F causes G.

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 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, while 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 1mg of finasteride.
	m_2	Subject is chosen to take 1mg of a placebo.

An RCE used to test the conjecture that F causes G is shown in Figure 3.11. 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 3.11 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 $P(g_1|m_1) \approx .67$ and

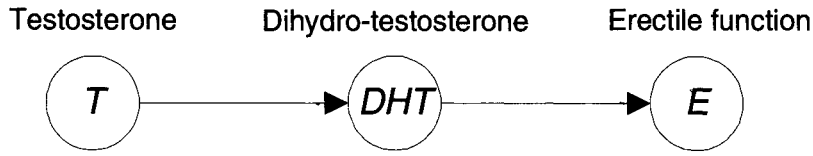

Figure 3.12: A causal DAG.

$P(g_1|m_2) \approx .07$. 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 Dec. 22, 1997, announced that the U.S. Food and Drug Administration granted marketing clearance to Propecia(TM) (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).

3.3.2 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

A **causal graph** is a directed graph containing a set of causally related random variables V such that for every $X, Y \in \mathsf{V}$ there is an edge from X to Y if and only if X is a direct cause of Y. By a **direct cause** we mean a manipulation of X results in a change in the probability distribution of Y, and there is no subset of variables W of the set of variables in the graph such that if we knew the values of the variables in W, a manipulation of X would no longer change the probability distribution of Y. A causal graph is a **causal DAG** if the directed graph is acyclic (i.e., there are no causal feedback loops).

Example 3.5 *Testosterone (T) is known to convert to dihydro-testosterone (DHT), and DHT 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 DHT and erectile function declined. They then held DHT fixed at low levels and found that erectile function was low regardless of the manipulated value of testosterone. Finally, they held DHT 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, DHT, and E, T is a direct cause of DHT, DHT 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 3.12.*

Notice that if the variable DHT were not in the DAG in Figure 3.12, there would be an edge from T directly into E instead of the directed path through DHT. 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 stand also 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, etc. 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 a myriad of 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.

The Causal Markov Assumption

If we assume 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 3.6 *Consider again the situation involving testosterone (T), DHT, and erectile function (E). The manipulation study in [Lugg et al., 1995] showed that if we instantiate DHT, 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 3.7 *A history of smoking (H) is known to cause both bronchitis (B) and lung cancer (L). Lung cancer and bronchitis both cause fatigue (F),*

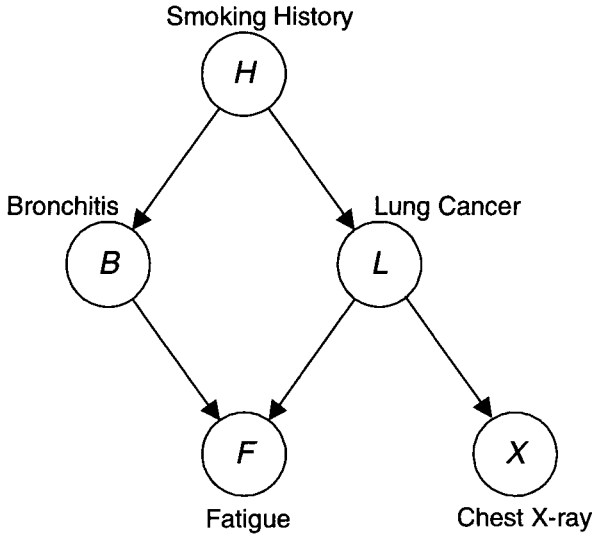

Figure 3.13: A causal DAG.

while only lung cancer can cause a chest X-ray (X) to be positive. There are no other causal relationships among the variables. Figure 3.13 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 3.13, we would not expect bronchitis and lung cancer to be independent because if someone had lung cancer it would make it more probable that they 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 they 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 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 3.8 *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, we would create the DAG in Figure 3.14 (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 3.14 (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.*

3.3.3 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

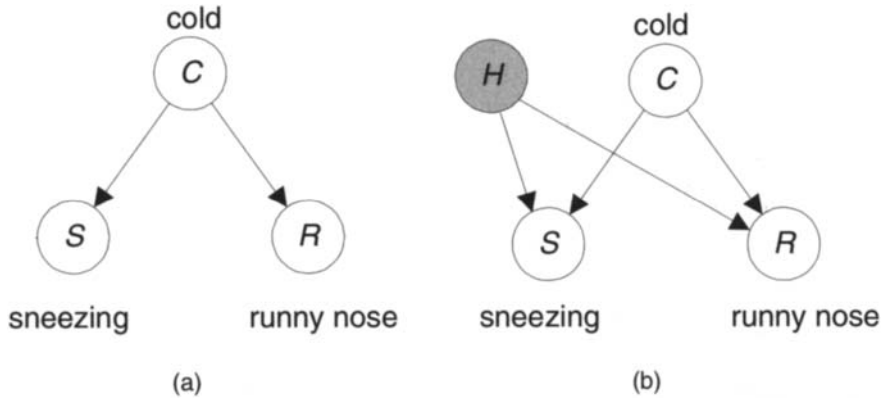

Figure 3.14: The causal Markov assumption would not hold for the DAG in (a) if there is a hidden common cause as depicted in (b).

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 3.4 satisfies the Markov condition with the DAG $\mathbb{G}$ in Figure 3.5. 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 3.12 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.

3.4 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 of representing the joint probability distribution of a large number of variables. For example, Figure 3.2, which appears again as Figure 3.15, 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 may want to compute the probability of credit card fraud given 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 of using the algorithms to do inference.

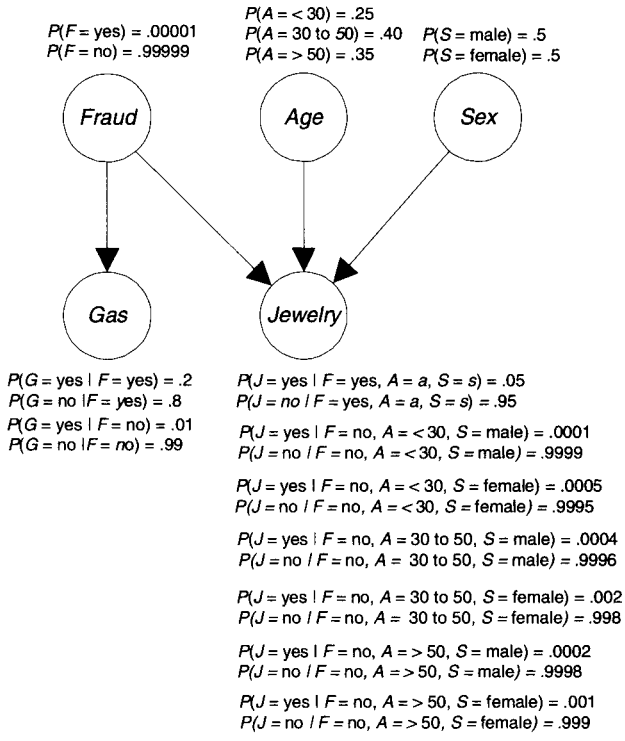

Figure 3.15: A Bayesian network for detecting credit card fraud.

3.4.1 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 3.9 *Consider the Bayesian network in Figure 3.16 (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 3.16 (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.

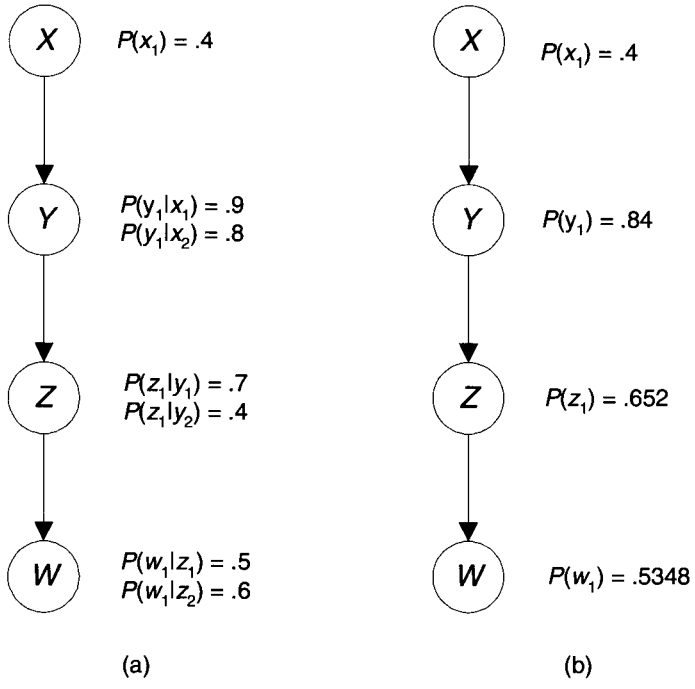

Figure 3.16: A Bayesian network appears in (a), and the prior probabilities of the variables in that network are shown in (b). Each variable only has two values, so only the probability of one is shown in (a).

Example 3.10 *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:*

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

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

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

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 3.11 *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:

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

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

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 3.12 *Consider the Bayesian network in Figure 3.17. 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. It is left as an exercise to do this.*

3.4.2 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

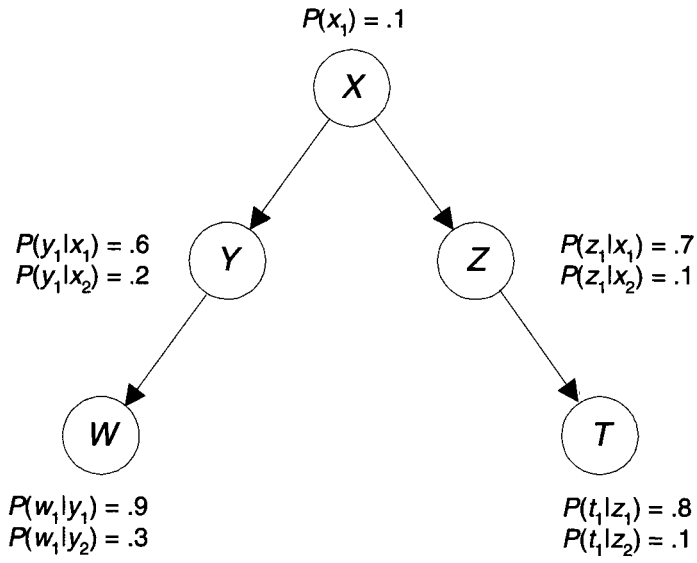

Figure 3.17: A Bayesian network. Each variable only has two possible values; so only the probability of one is shown.

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 [1994] took a different approach. They developed an algorithm which 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 which 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 [Pradham and Dagum, 1996].

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

1. Netica (www.norsys.com)

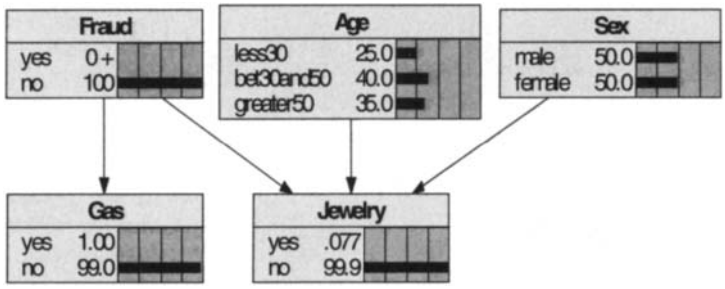

Figure 3.18: The fraud detection Bayesian network in Figure 3.15 implemented using Netica.

2. GeNIe (http://genie.sis.pitt.edu/)
3. HUGIN (http://www.hugin.dk)
4. Elvira (http://www.ia.uned.es/~elvira/index-en.html)
5. BUGS (http://www.mrc-bsu.cam.ac.uk/bugs/)

In this book we use Netica to do inference. You can download a free version that allows up to 13 nodes in a network.

3.4.3 Inference Using Netica

Next, we illustrate inference in a Bayesian network using Netica. Figure 3.18 shows the fraud detection network in Figure 3.15 implemented using Netica. Note 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 3.19 (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

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

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

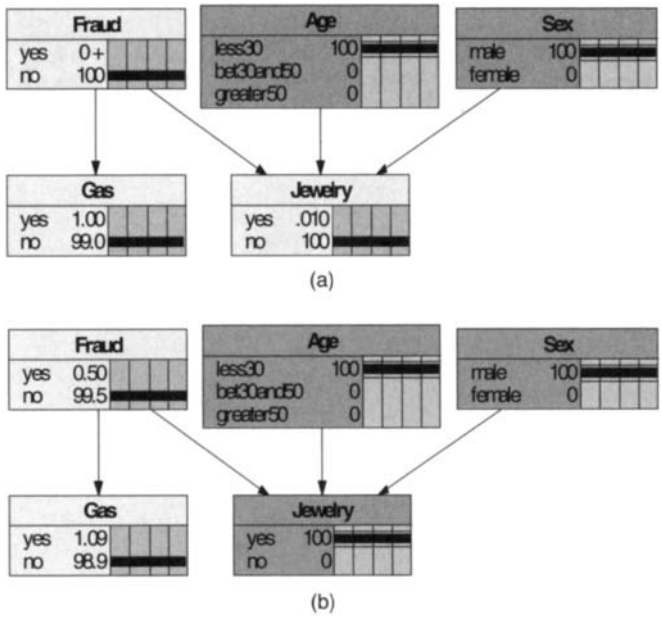

Figure 3.19: In (a) *Age* has been instantiated to *less*30 and *Sex* has been instantiated to *male*. In (b) *Age* has been instantiated to *less*30, *Sex* has been instantiated to *male*, and *Jewelry* has been instantiated to *yes*.

that the card holder is a young man should not make it more or less likely that the card is being used fraudulently. Figure 3.19 (b) has the same instantiations as Figure 3.19 (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 3.20 (a) and (b) *Gas* and *Jewelry* have both been instantiated to *yes*. However, in Figure 3.20 (a) the card holder is a young man, while in Figure 3.20 (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 away the jewelry purchase.

3.5 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

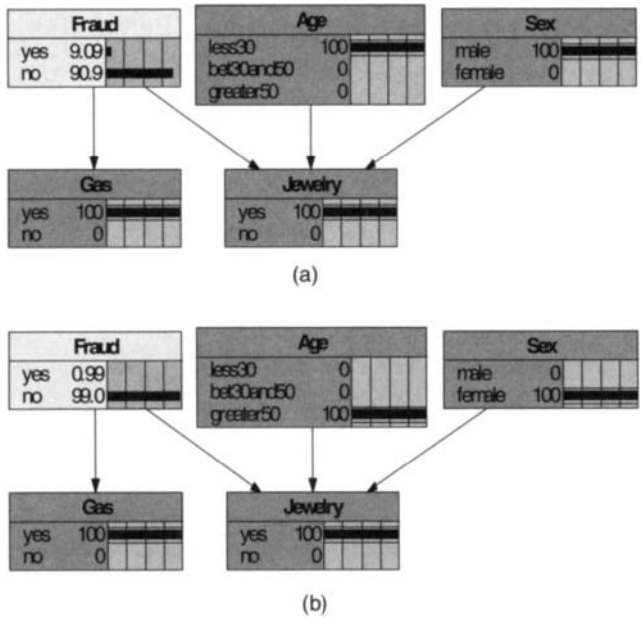

Figure 3.20: *Sex* and *Jewelry* have both been instantiated to *yes* in both (a) and (b). However, in (a) the card holder is a young man, while in (b) it is an older woman.

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 4 we discuss techniques for learning them from data, while in Parts II and III we show examples of obtaining them from experts and of 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.

3.5.1 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 3.21.

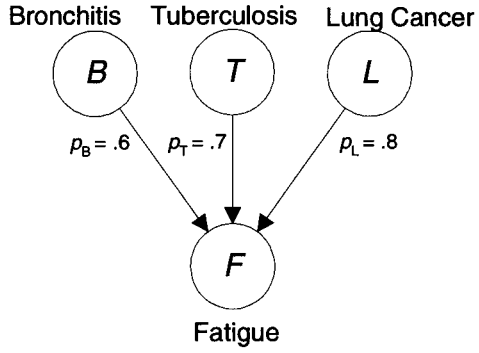

Figure 3.21: We need to assess eight conditional probabilities for node F.

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:

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

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

$$\ldots$$

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

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 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 3.21 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 3.21 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 3.21. 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 3.21.

Example 3.13 *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 3.21. 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,

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

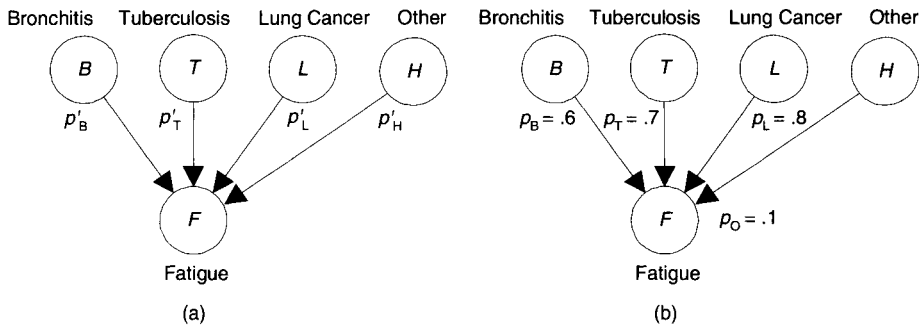

Figure 3.22: The probabilities in (a) are the causal strengths in the noisy OR-gate model. The probabilities in (b) are the ones we ascertain.

The Leaky Noisy OR-Gate 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 3.21 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.

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 3.22 (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 3.22 (b) are the ones we actually 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.

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

$$p_i = P(Y = yes|X_1 = no, X_2 = no, \ldots X_i = yes, \ldots X_n = no) \tag{3.1}$$

$$p_0 = P(Y = yes|X_1 = no, X_2 = no, \ldots X_n = no). \tag{3.2}$$

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 3.14 *Let's compute the conditional probabilities for the network in Figure 3.22 (b). We have*

$$\begin{aligned} P(F = no|B = no, T = no, L = no) &= 1 - p_0 \\ &= 1 - .1 = .9 \end{aligned}$$

$$\begin{aligned} P(F = no|B = no, T = no, L = yes) &= (1 - p_0)\frac{1 - p_L}{1 - p_0} \\ &= 1 - .8 = .2 \end{aligned}$$

$$\begin{aligned} P(F = no|B = no, T = yes, L = no) &= (1 - p_0)\frac{1 - p_T}{1 - p_0} \\ &= 1 - .7 = .3 \end{aligned}$$

$$\begin{aligned} P(F = no|B = no, T = yes, L = yes) &= (1 - p_0)\frac{1 - p_T}{1 - p_0}\frac{1 - p_L}{1 - p_0} \\ &= \frac{(1 - .7)(1 - .8)}{1 - .1} = .067 \end{aligned}$$

$$\begin{aligned} P(F = no|B = yes, T = no, L = no) &= (1 - p_0)\frac{1 - p_B}{1 - p_0} \\ &= 1 - .6 = .4 \end{aligned}$$

$$\begin{aligned} P(F = no|B = yes, T = no, L = yes) &= (1 - p_0)\frac{1 - p_B}{1 - p_0}\frac{1 - p_L}{1 - p_0} \\ &= \frac{(1 - .6)(1 - .8)}{1 - .1} = .089 \end{aligned}$$

$$\begin{aligned} P(F = no|B = yes, T = yes, L = no) &= (1 - p_0)\frac{1 - p_B}{1 - p_0}\frac{1 - p_T}{1 - p_0} \\ &= \frac{(1 - .6)(1 - .7)}{1 - .1} = .133 \end{aligned}$$

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

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

Lemma 3.1 *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 3.2, 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 3.2 *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 3.1, 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 3.1. This completes the proof. ■

Theorem 3.2 *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 3.1, and the last is due to Lemma 3.2. ■

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 [2006] 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].

3.5.2 Methods for Discretizing Continuous Variables★

Often, a Bayesian network contains both discrete and continuous random variables. For example, the Bayesian network in Figure 3.2 contains four random variables that are discrete and one, namely *Age*, that is continuous.[5] 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

[5] 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.

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. Examples of this appear in Chapter 4, Section 4.7.1, and Chapter 10. Next, we present two of the most popular methods for discretizing continuous random variables.

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 3.15 *Recall that the normal density function is given by*

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

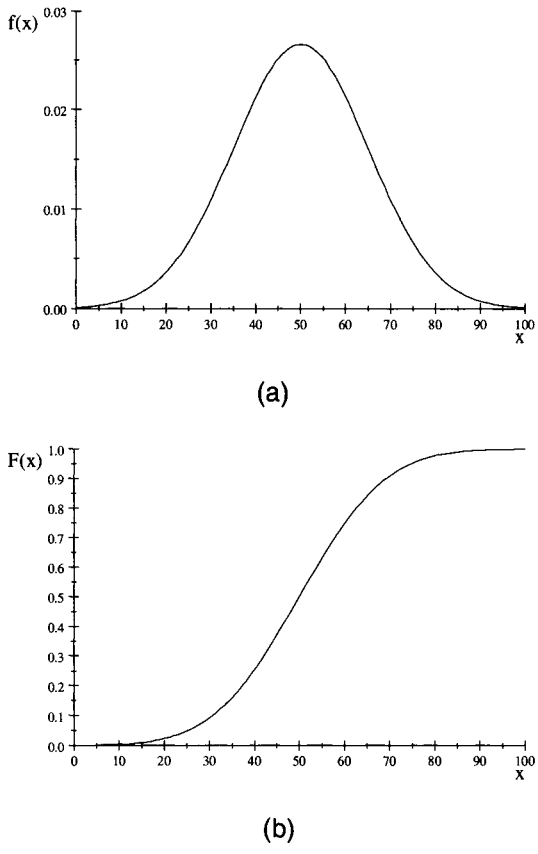

Figure 3.23: The normal density function with $\mu = 50$ and $\sigma = 15$ appears in (a), while the corresponding normal cumulative distribution function appears in (b).

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 3.23. 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 $d_3 = 64.5$.

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. An example would be if the variable were used in a decision analysis application (Chapter 5). 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

$$P(D = \ < 43.5) = .333$$

$$P(D = 43.5 \text{ to } 56.4) = .333$$

$$P(D = \ > \ 56.4) = .333.$$

Recall that this is what we did for *Age* in the Bayesian network in Figure 3.2. 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. In Chapter 10 we will 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

$$P(X \leq x_1) = .05$$
$$P(X \leq x_2) = .50$$
$$P(X \leq x_3) = .95.$$

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

$$P(D = x_1) = .185$$
$$P(D = x_2) = .63$$
$$P(D = x_3) = .185.$$

Example 3.16 *Suppose we have the normal distribution discussed in Example 3.15. 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

$$P(D = \ < 36.6) = .185$$

$$P(D = 36.6 \text{ to } 63.4) = .63$$

$$P(D = \ > \ 63.4) = .185.$$

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

3.6 Entailed Conditional Independencies★

If $(\mathbb{G}, P)$ satisfies the Markov condition, then each node in $\mathbb{G}$ is conditionally independent of the set of all its nondescendents given its parents. Do these conditional independencies entail any other conditional independencies? That is, if $(\mathbb{G}, P)$ satisfies the Markov condition, are there any other conditional independencies which P must satisfy other than the one based on a node's parents? The answer is yes. Such conditional independencies are called entailed conditional independencies. Specifically, we say a DAG **entails a conditional independency** if every probability distribution, which satisfies the Markov condition with the DAG, must have the conditional independency. Before explicitly showing all entailed conditional independencies, we illustrate that one would expect them.

3.6.1 Examples of Entailed Conditional Independencies

Suppose some distribution P satisfies the Markov condition with the DAG in Figure 3.24 (a). Then we know $I_P(C, \{F, G\}|, B)$ because B is the parent of C, and F and G are nondescendents of C. Furthermore, we know $I_P(B, G|F)$ because F is the parent of B, and G is a nondescendent of B. These are the only conditional independencies according to the statement of the Markov condition. However, can any other conditional independencies be deduced from them? For example, can we conclude $I_P(C, G|F)$? Let's first give the variables meaning and the DAG a causal interpretation to see if we would expect this conditional independency.

Suppose we are investigating how professors obtain citations, and the variables represent the following:

G:	Graduate Program Quality
F:	First Job Quality
B:	Number of Publications
C:	Number of Citations.

Further suppose the DAG in Figure 3.24 (a) represents the causal relationships among these variables, and there are no hidden common causes.[6] Then it is reasonable to make the causal Markov assumption, and we would feel that the probability distribution of the variables satisfies the Markov condition with the DAG. Suppose we learn that Professor La Budde attended a graduate program (G) of high quality. We would now expect his first job (F) may well be of high quality, which means that he should have a large number of publications (B), which in turn implies he should have a large number of citations (C). Therefore, we would not expect $I_P(C, G)$.

Suppose we next learn that Professor Pellegrini's first job (F) was of high quality. That is, we instantiate F to "high quality." The cross through the node F in Figure 3.24 (b) indicates it is instantiated. We would now expect his

[6] We make no claim that this model accurately represents the causal relationships among the variables. See [Spirtes et al., 1993, 2000] for a detailed discussion of this problem.

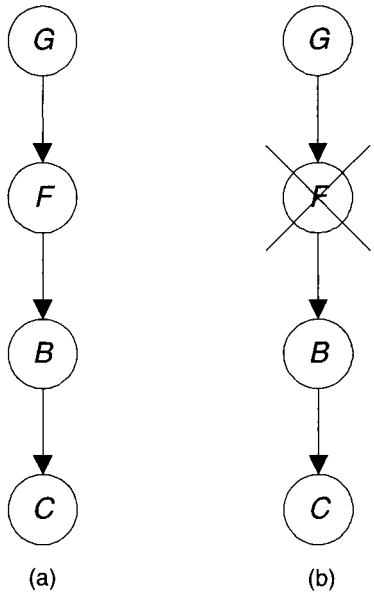

Figure 3.24: A causal DAG is shown in (a). The variable F is instantiated in (b).

number of publications (B) to be large and, in turn, his number of citations (C) to be large. If Professor Pellegrini then tells us he attended a graduate program (G) of high quality, would we expect the number of citations to be even higher than we previously thought? It seems not. The graduate program's high quality implies the number of citations is probably large because it implies the first job is probably of high quality. Once we already know the first job is of high quality, the information concerning the graduate program should be irrelevant to our beliefs concerning the number of citations. Therefore, we would expect C to not only be conditionally independent of G given its parent B, but also its grandparent F. Either one seems to block the dependency between G and C that exists through the chain $[G, F, B, C]$. So we would expect $I_P(C, G|F)$.

It is straightforward to show that the Markov condition does indeed entail $I_P(C, G|F)$ for the DAG $\mathbb{G}$ in Figure 3.24. We show this next. If $(\mathbb{G}, P)$ satisfies the Markov condition,

$$\begin{aligned} P(c|g,f) &= \sum_b P(c|b,g,f)P(b|g,f) \\ &= \sum_b P(c|b,f)P(b|f) \\ &= P(c|f). \end{aligned}$$

The first equality is due to the law of total probability (in the background space that we know the values of g and f), the second equality is due to the Markov

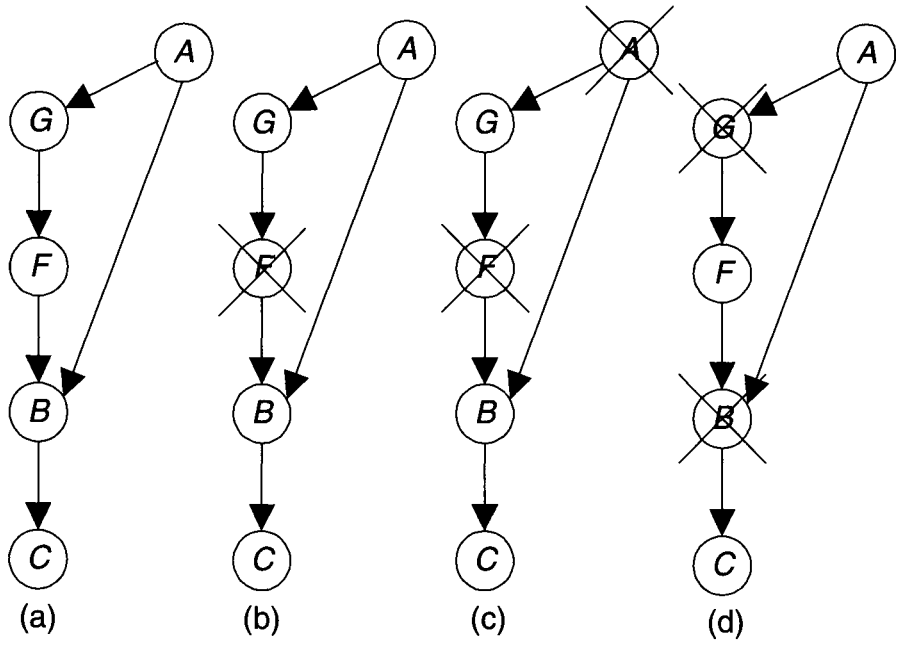

Figure 3.25: A causal DAG is shown in (a). The variable F is instantiated in (b). The variables A and F are instantiated in (c). The variables B and G are instantiated in (d).

condition, and the last equality is again due to the law of total probability.

If we have an arbitrarily long directed linked list of variables, and P satisfies the Markov condition with that list, in the same way as above, we can show that for any variable in the list, the set of variables above it is conditionally independent of the set of variables below it given that variable. That is, the variable blocks the dependency transmitted through the chain.

Suppose now that P does not satisfy the Markov condition with the DAG in Figure 3.24 (a) because there is a common cause A of G and B. For the sake of illustration, let's say A represents the following in the current example:

A: Ability.

Further, suppose there are no other hidden common causes, so we would now expect P to satisfy the Markov condition with the DAG in Figure 3.25 (a). Would we still expect $I_P(C, G|F)$? It seems not. For example, as before, suppose that we initially learn Professor Pellegrini's first job (F) was of high quality. This instantiation is shown in Figure 3.25 (b). We learn next that his graduate program (G) was of high quality. Given the current model, the fact that G is of high quality is indicative of his having high ability (A), which can directly affect his publication rate (B) and therefore his citation rate (C). So we now would feel his citation rate (C) may be even higher than what we

thought when we only knew his first job (F) was of high quality. This means we would not feel $I_P(C, G|F)$ as we did with the previous model. Suppose next that we know Professor Pellegrini's first job (F) was of high quality and that he has high ability (A). These instantiations are shown in Figure 3.25 (c). In this case, his attendance at a high-quality graduate program (G) can no longer be indicative of his ability (A), and therefore, it cannot affect our belief concerning his citation rate (C) through the chain $[G, A, B, C]$. That is, this chain is blocked at A. So we would expect $I_P(C, G|\{A, F\})$. Indeed, it is possible to prove that the Markov condition does entail $I_P(C, G|\{A, F\})$.

Finally, consider the conditional independency $I_P(F, A|G)$. This independency is obtained directly by applying the Markov condition to the DAG in Figure 3.25 (a). So we will not offer an intuitive explanation for it. Rather, we discuss whether we would expect the conditional independency to still exist if we also knew the state of B. Suppose we first learn that Professor Georgakis has a high publication rate (B) and attended a high-quality graduate program (G). These instantiations are shown in Figure 3.25 (d). We later learn she also has high ability (A). In this case, her high ability (A) could explain away her high publication rate (B), thereby making it less probable she had a high-quality first job (F). As mentioned in Section 3.1, psychologists call this discounting. So the chain $[A, B, F]$ is opened by instantiating B, and we would not expect $I_P(F, A|\{B, G\})$. Indeed, the Markov condition does not entail $I_P(F, A|\{B, G\})$. Note that the instantiation of C should also open the chain $[A, B, F]$. That is, if we know the citation rate (C) is high, then it is probable the publication rate (B) is high, and each of the causes of B can explain away this high probability. Indeed, the Markov condition does not entail $I_P(F, A|\{C, G\})$ either.

3.6.2 d-Separation

Figure 3.26 shows the chains that can transmit a dependency between variables X and Y in a Bayesian network. To discuss these dependencies intuitively, we give the edges in that figure causal interpretations as follows:

1. The chain $[X, B, C, D, Y]$ is a causal path from X to Y. In general, there is a dependency between X and Y on this chain, and the instantiation of any intermediate cause on the chain blocks the dependency.

2. The chain $[X, F, H, I, Y]$ is a chain in which H is a common cause of X and Y. In general, there is a dependency between any X and Y on this chain, and the instantiation of the common cause H or either of the intermediate causes F and I blocks the dependency.

3. The chain $[X, J, K, L, Y]$ is a chain in which X and Y both cause K. There is no dependency between X and Y on this chain. However, if we instantiate K or M, in general a dependency would be created. We would then need to also instantiate J or L.

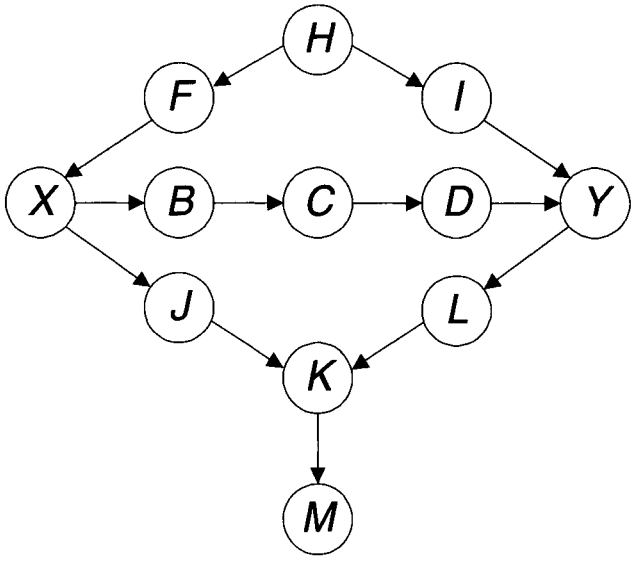

Figure 3.26: This DAG illustrates the chains that can transmit a dependency between X and Y.

To render X and Y conditionally independent we need to instantiate at least one variable on all the chains that transmit a dependency between X and Y. So we would need to instantiate at least one variable on the chain $[X, B, C, D, Y]$, at least one variable on the chain $[X, F, H, I, Y]$, and, if K or M are instantiated, at least one other variable on the chain $[X, J, K, L, Y]$.

Now that we've discussed intuitively how dependencies can be transmitted and blocked in a DAG, we show precisely what conditional independencies are entailed by the Markov condition. To do this, we need the notion of d-separation, which we define shortly. First, we present some preliminary concepts. We say there is a **head-to-head meeting** at X on a chain if the edges incident to X both have their arrows into X. For example, the chain $Y \leftarrow W \rightarrow X \leftarrow V$ has a head-to-head meeting at X. We say there is **head-to-tail meeting** at X on a chain if precisely one of the edges incident to X has its arrows into X. For example, the chain $Y \leftarrow W \leftarrow X \leftarrow V$ has a head-to-tail meeting at X. We say there is **tail-to-tail meeting** at X on a chain if neither of the edges incident to X has its arrows into X. For example, the chain $Y \leftarrow W \leftarrow X \rightarrow V$ has a tail-to-tail meeting at X. We now have the following definition:

Definition 3.2 *Suppose we have a DAG* $\mathbb{G} = (\mathsf{V}, \mathsf{E})$, *a chain* ρ *in the DAG connecting two nodes* X *and* Y, *and a subset of nodes* $\mathsf{W} \subseteq \mathsf{V}$. *We say that the chain* ρ *is* ***blocked*** *by* W *if at least one of the following is true:*

1. *There is a node* $Z \in \mathsf{W}$ *that has a head-to-tail meeting on* ρ.

2. *There is a node* $Z \in \mathsf{W}$ *that has a tail-to-tail meeting on* ρ.

3. *There is a node* Z, *such that* Z *and all* Z*'s descendents are not in* W, *that has a head-to-head meeting on* ρ.

Example 3.17 *For the DAG in Figure 3.26, the following are some examples of chains that are blocked and that are not blocked:*

1. *The chain* $[X, B, C, D, Y]$ *is blocked by* $W = \{C\}$ *because there is a head-to-tail meeting at* C.

2. *The chain* $[X, B, C, D, Y]$ *is blocked by* $W = \{C, H\}$ *because there is a head-to-tail meeting at* C.

3. *The chain* $[X, F, H, I, Y]$ *is blocked by* $W = \{C, H\}$ *because there is a tail-to-tail meeting at* H.

4. *The chain* $[X, J, K, L, Y]$ *is blocked by* $W = \{C, H\}$ *because there is a head-to-head meeting at* K, *and* K *and* M *are both not in* W.

5. *The chain* $[X, J, K, L, Y]$ *is not blocked by* $W = \{C, H, K\}$ *because there is a head-to-head meeting at* K, *and* K *is not in* W.

6. *The chain* $[X, J, K, L, Y]$ *is blocked by* $W = \{C, H, K, L\}$ *because there is a head-to-tail meeting at* L.

We can now define d-separation.

Definition 3.3 *Suppose we have a DAG* $\mathbb{G} = (\mathsf{V}, \mathsf{E})$ *and a subset of nodes* $\mathsf{W} \subseteq \mathsf{V}$. *Then* X *and* Y *are* ***d-separated*** *by* W *if every chain between* X *and* Y *is blocked by* W.

Definition 3.4 *Suppose we have a DAG* $\mathbb{G} = (\mathsf{V}, \mathsf{E})$ *and three subsets of nodes* $\mathsf{X}, \mathsf{Y} \subseteq \mathsf{V}$, *and* W. *We say* X *and* Y *are d-separated by* W *if for every* $X \in \mathsf{X}$ *and* $Y \in \mathsf{Y}$, X *and* Y *are d-separated by* W.

As you may have already suspected, d-separation recognizes all the conditional independencies entailed by the Markov condition. Specifically, we have the following theorem:

Theorem 3.3 *Suppose we have a DAG* $\mathbb{G} = (\mathsf{V}, \mathsf{E})$ *and three subsets of nodes* X, Y, *and* $\mathsf{W} \subseteq \mathsf{V}$. *Then* $\mathbb{G}$ *entails the conditional independency* $I_P(\mathsf{X}, \mathsf{Y}|\mathsf{W})$ *if and only if* X *and* Y *are d-separated by* W.
Proof. *The proof can be found in [Neapolitan, 1990].* ■

We stated the theorem for sets of variables, but it also applies to single variables. That is, if X contains a single variable X and Y contains a single variable Y, then $\mathrm{I}_P(\mathsf{X}, \mathsf{Y}|\mathsf{W})$ is the same as $\mathrm{I}_P(X, Y|\mathsf{W})$. We show examples of this simpler case next and investigate more complex sets in the exercises.

Example 3.18 *Owing to Theorem 3.3, the following are some conditional independencies the Markov condition entails for the DAG in Figure 3.26:*

Conditional Independency	Reason Conditional Independency Is Entailed
$I_P(X, Y\|\{H, C\})$	$[X, F, H, I, Y]$ is blocked at H. $[X, B, C, D, Y]$ is blocked at C. $[X, J, K, L, Y]$ is blocked at K.
$I_P(X, Y\|\{F, D\})$	$[X, F, H, I, Y]$ is blocked at F. $[X, B, C, D, Y]$ is blocked at D. $[X, J, K, L, Y]$ is blocked at K.
$I_P(X, Y\|\{H, C, K, L\})$	$[X, F, H, I, Y]$ is blocked at H. $[X, B, C, D, Y]$ is blocked at C. $[X, J, K, L, Y]$ is blocked at L.
$I_P(X, Y\|\{H, C, M, L\})$	$[X, F, H, I, Y]$ is blocked at H. $[X, B, C, D, Y]$ is blocked at C. $[X, J, K, L, Y]$ is blocked at L.

In the third row it is necessary to include L to obtain the independency because there is a head-to-head meeting at K on the chain $[X, J, K, L, Y]$, and $K \in \{H, C, K, L\}$. Similarly, in the fourth row, it is necessary to include L to obtain the independency because there is a head-to-head meeting at K on the chain $[X, J, K, L, Y]$, M is a descendent of K, and $M \in \{H, C, M, L\}$.

Example 3.19 *Owing to Theorem 3.3, the following are some conditional independencies the Markov condition does not entail for the DAG in Figure 3.26:*

Conditional Independency	Reason Conditional Independency Is Not Entailed
$I_P(X, Y\|H)$	$[X, B, C, D, Y]$ is not blocked.
$I_P(X, Y\|, D)$	$[X, F, H, I, Y]$ is not blocked.
$I_P(X, Y\|\{H, C, K\})$	$[X, J, K, L, Y]$ is not blocked.
$I_P(X, Y\|\{H, C, M\})$	$[X, J, K, L, Y]$ is not blocked.

Example 3.20 *Owing to Theorem 3.3, the Markov condition entails the following conditional independency for the DAG in Figure 3.27:*

Conditional Independency	Reason Conditional Independency Is Entailed
$I_P(W, X)$	$[W, Y, X]$ is blocked at Y. $[W, Y, R, Z, X]$ is blocked at R. $[W, Y, R, Z, S, X]$ is blocked at S.

Note that $I_P(W, X)$ is the same as $I_P(W, X|\varnothing)$, where $\varnothing$ is the empty set, and Y, R, S, and T are all not in $\varnothing$.

It is left as an exercise to show that the Markov condition also entails the following conditional independencies for the DAG in Figure 3.27:

1. $I_P(X, R|\{Y, Z\})$

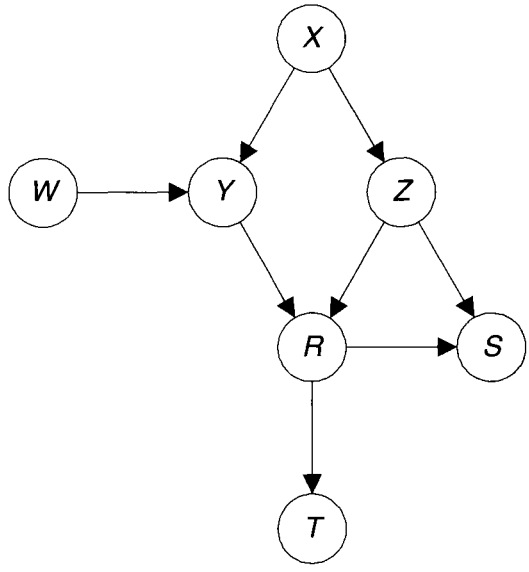

Figure 3.27: The Markov condition entails $I_P(W, X)$ for this DAG.

2. $I_P(X, T|\{Y, Z\})$
3. $I_P(W, T|R)$
4. $I_P(Y, Z|X)$
5. $I_P(W, S|\{R, Z\})$
6. $I_P(W, S|\{Y, Z\})$
7. $I_P(W, S|\{Y, X\})$.

It is also left as an exercise to show that the Markov condition does not entail the following conditional independencies for the DAG in Figure 3.27:

1. $I_P(W, X|Y)$
2. $I_P(W, T|Y)$.

3.6.3 Faithful and Unfaithful Probability Distributions

Recall that a DAG entails a conditional independency if every probability distribution, which satisfies the Markov condition with the DAG, must have the conditional independency. Theorem 3.3 states that all and only d-separations are entailed conditional independencies. Do not misinterpret this result. It does not say that if some particular probability distribution P satisfies the Markov condition with a DAG $\mathbb{G}$, then P cannot have conditional independencies that

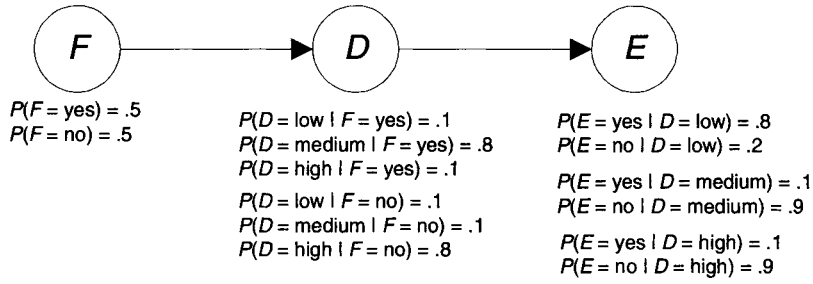

Figure 3.28: For the distribution P in this Bayesian network we have $I_P(E, F)$, but the Markov condition does not entail this conditional independency.

$\mathbb{G}$ does not entail. Rather, it only says that P must have all the conditional independencies that are entailed. We illustrate the difference next.

An Unfaithful Distribution

The following example shows that, even when we obtain a distribution by assigning conditional probabilities in a DAG, we can end up with a distribution that has a conditional independency that is not entailed by the Markov condition.

Example 3.21 *Consider the Bayesian network in Figure 3.28. The only conditional independency entailed by the Markov condition for the DAG $\mathbb{G}$ in that figure is $I_P(E, F|D)$. So Theorem 3.3 says all probability distributions that satisfy the Markov condition with $\mathbb{G}$ must have $I_P(E, F|D)$, which means the probability distribution P in the Bayesian network in Figure 3.28 must have $I_P(E, F|D)$. However, the theorem does not say that P cannot have other independencies. Indeed, it is left as an exercise to show that we have $I_P(E, F)$ for the distribution P in that Bayesian network.*

We purposefully assigned values to the conditional distributions in the network in Figure 3.28 to achieve $I_P(E, F)$. If we randomly assigned values, we would be almost certain to obtain a probability distribution that does not have this independency. That is, Meek [1995] proved that almost all assignments of values to the conditional distributions in a Bayesian network will result in a probability distribution that only has conditional independencies entailed by the Markov condition.

Could actual phenomena in nature result in a distribution like that in Figure 3.28? Although we made up the numbers in the network in that figure, we patterned them after something that actually occurred in nature. Let the variables in the figure represent the following:

Variable	What the Variable Represents
F	Whether the subject takes finasteride
D	Subject's dihydro-testosterone level
E	Whether the subject suffers erectile dysfunction

As shown in Example 3.5, dihydro-testosterone seems to be the hormone necessary for erectile function. Recall from Section 3.3.1 that Merck performed a study indicating that finasteride has a positive causal effect on hair growth. Finasteride accomplishes this by inhibiting the conversion of testosterone to dihydro-testosterone, and dihydro-testosterone is the hormone responsible for hair loss. Given this, Merck feared that dihydro-testosterone would cause erectile dysfunction. That is, ordinarily if X has a causal influence on Y and Y has a causal influence on Z, then X has a causal influence on Z through Y. However, in a manipulation study Merck found that F does not appear to have a causal influence on E. That is, they learned $I_P(E, F)$. The explanation for this is that finasteride does not lower dihydro-testosterone levels beneath some threshold level, and that threshold level is all that is necessary for erectile function. The numbers we assigned in the Bayesian network in Figure 3.28 reflect this. The value of F has no effect on whether D is *low*, and D must be *low* to make the probability that E is *yes* high.

Faithfulness

The probability distribution in the Bayesian network in Figure 3.28 is said to be unfaithful to the DAG in that figure because it contains a conditional independency that is not entailed by the Markov condition. We have the following definition:

Definition 3.5 *Suppose we have a joint probability distribution P of the random variables in some set $\mathbf{V}$ and a DAG $\mathbb{G} = (\mathbf{V}, \mathbf{E})$. We say that $(\mathbb{G}, P)$ satisfies the* ***faithfulness condition*** *if all and only the conditional independencies in P are entailed by $\mathbb{G}$. Furthermore, we say that P and $\mathbb{G}$ are faithful to each other.*

Notice that the faithfulness condition includes the Markov condition because *only* conditional independencies in P can be entailed by $\mathbb{G}$. However, it requires more; that is, it requires that *all* conditional independencies in P must be entailed by $\mathbb{G}$. As noted previously, almost all assignments of values to the conditional distributions will result in a faithful distribution. For example, it is left as an exercise to show that the probability distribution P in the Bayesian network in Figure 3.29 is faithful to the DAG in that figure. We arbitrarily assigned values to the conditional distributions in the figure. However, owing to the result in [Meek, 1995] that almost all assignments will lead to a faithful distribution, we were willing to bet the farm that this assignment would.

Is there some DAG faithful to the probability distribution in the Bayesian network in Figure 3.28? The answer is "no," but it is beyond the scope of this book to show this. See [Neapolitan, 2004] for a proof of this fact. Intuitively,

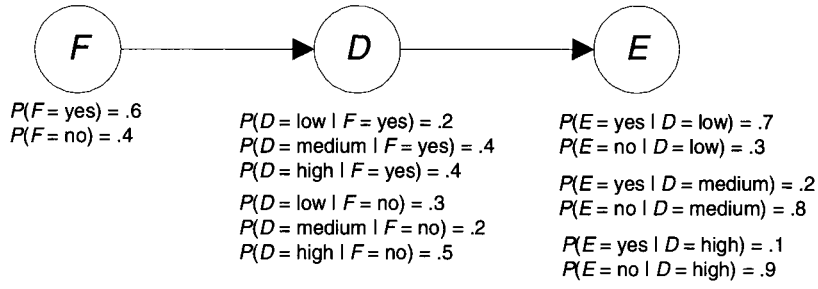

Figure 3.29: The probability distribution in this Bayesian network is faithful to the DAG in the network.

the DAG in Figure 3.28 represents the causal relationships among the variables, which means we should not be able to find a DAG which better represents the probability distribution.

3.6.4 Markov Blankets and Boundaries

A Bayesian network can have a large number of nodes, and the probability of a given node can be affected by instantiating a distant node. However, it turns out that the instantiation of a set of close nodes can shield a node from the effect of all other nodes. The next definition and theorem show this.

Definition 3.6 *Let* V *be a set of random variables,* P *be their joint probability distribution, and* $X \in \mathsf{V}$. *Then a* ***Markov blanket*** M *of* X *is any set of variables such that* X *is conditionally independent of all the other variables given* M. *That is,*

$$I_P(X, \mathsf{V} - (\mathsf{M} \cup \{X\}) | \mathsf{M}).$$

Theorem 3.4 *Suppose* $(\mathbb{G}, P)$ *satisfies the Markov condition. Then for each variable* X, *the set of all parents of* X, *children of* X, *and parents of children of* X *is a Markov blanket of* X.

Proof. *It is straightforward that this set d-separates* X *from the set of all other nodes in* V. *The proof therefore follows from Theorem 3.3.* ■

Example 3.22 *Suppose* $(\mathbb{G}, P)$ *satisfies the Markov condition where* $\mathbb{G}$ *is the DAG in Figure 3.30. Then due to Theorem 3.4,* $\{T, Y, Z\}$ *is a Markov blanket of* X. *So we have*

$$I_P(X, \{S, W\} | \{T, Y, Z\}).$$

Example 3.23 *Suppose* $(\mathbb{G}, P)$ *satisfies the Markov condition where* $\mathbb{G}$ *is the DAG in Figure 3.30, and* P *has the following conditional independency:*

$$I_P(X, \{S, T, W\} | \{Y, Z\}).$$

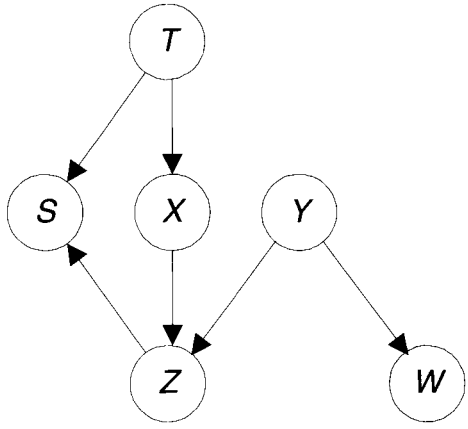

Figure 3.30: If P satisfies the Markov condition with this DAG, then $\{T, Y, Z\}$ is a Markov blanket of X.

Then the Markov blanket $\{T, Y, Z\}$ is not minimal in the sense that its subset $\{Y, Z\}$ is also a Markov blanket of X.

The last example motivates the following definition:

Definition 3.7 *Let V be a set of random variables, P be their joint probability distribution, and $X \in \mathsf{V}$. Then a **Markov boundary** of X is any Markov blanket such that none of its proper subsets is a Markov blanket of X.*

We have the following theorem:

Theorem 3.5 *Suppose $(\mathbb{G}, P)$ satisfies the faithfulness condition. Then for each variable X, the set of all parents of X, children of X, and parents of children of X is the unique Markov boundary of X.*

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

Example 3.24 *Suppose $(\mathbb{G}, P)$ satisfies the faithfulness condition where $\mathbb{G}$ is the DAG in Figure 3.30. Then due to Theorem 3.5, $\{T, Y, Z\}$ is the unique Markov boundary of X.*

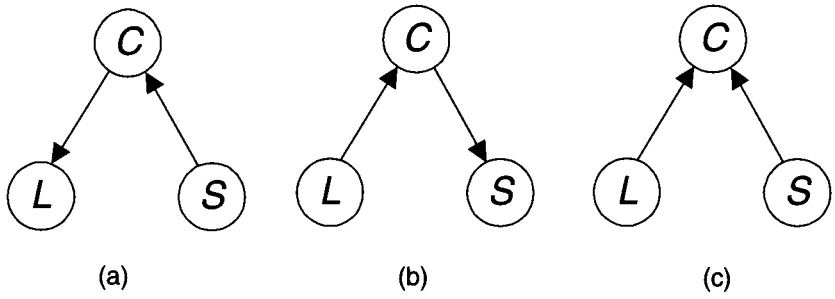

Figure 3.31: The probability distribution discussed in Example 3.1 satisfies the Markov condition with the DAGs in (a) and (b), but not with the DAG in (c).

EXERCISES

Section 3.1

Exercise 3.1 In Example 3.3 it was left as an exercise to show for all value of s, l, and c that

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

Show this.

Exercise 3.2 Consider the joint probability distribution P in Example 3.1.

1. Show that P satisfies the Markov condition with the DAG in Figure 3.31 (a) and that P is equal to the product of its conditional distributions in that DAG.

2. Show that P satisfies the Markov condition with the DAG in Figure 3.31 (b) and that P is equal to the product of its conditional distributions in that DAG.

3. Show that P does not satisfy the Markov condition with the DAG in Figure 3.31 (c) and that P is not equal to the product of its conditional distributions in that DAG.

Exercise 3.3 Create an arrangement of objects similar to the one in Figure 3.4, but with a different distribution of letters, shapes, and colors, so that, if random variables L, S, and C are defined as in Example 3.1, then the only independency or conditional independency among the variables is $I_P(L,S)$. Does this distribution satisfy the Markov condition with any of the DAGs in Figure 3.31? If so, which one(s)?

Exercise 3.4 Consider the joint probability distribution of the random variables defined in Example 3.1 relative to the objects in Figure 3.4. Suppose we compute that distribution's conditional distributions for the DAG in Figure 3.31 (c), and we take their product. Theorem 3.1 says this product is a joint probability distribution that constitutes a Bayesian network with that DAG. Is this the actual joint probability distribution of the random variables?

Section 3.2

Exercise 3.5 Professor Morris investigated gender bias in hiring in the following way. He gave hiring personnel equal numbers of male and female resumes 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 they might have thought this.

Exercise 3.6 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 3.7 Explain why, if we reverse the edges in the DAG in Figure 3.12 to obtain the DAG $E \rightarrow DHT \rightarrow T$, the new DAG also satisfies the Markov condition with the probability distribution of the variables.

Section 3.3

Exercise 3.8 Compute $P(x_1|w_1)$ assuming the Bayesian network in Figure 3.16.

Exercise 3.9 Compute $P(t_1|w_1)$ assuming the Bayesian network in Figure 3.17.

Exercise 3.10 Compute $P(x_1|t_2, w_1)$ assuming the Bayesian network in Figure 3.17.

Exercise 3.11 Using Netica develop the Bayesian network in Figure 3.2, and use that network to determine the following conditional probabilities:

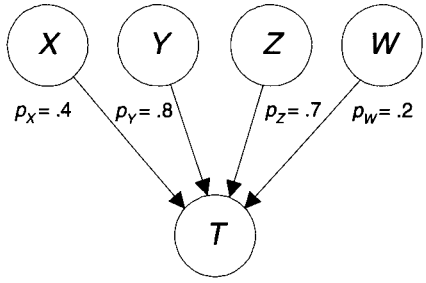

Figure 3.32: The noisy OR-gate model is assumed.

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.

Section 3.4

Exercise 3.12 Assume the noisy OR-gate model and the causal strengths are those shown in Figure 3.32. Compute the probability $T = yes$ for all combinations of values of the parents.

Exercise 3.13 Assume the leaky noisy OR-gate model and the relevant probabilities are those shown in Figure 3.33. Compute the probability $T = yes$ for all combinations of values of the parents. Compare the results to those obtained in Exercise 3.12.

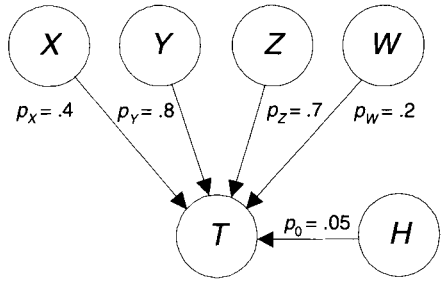

Figure 3.33: The leaky noisy OR-gate model is assumed.

Exercise 3.14 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.

Section 3.5

Exercise 3.15 Consider the DAG $\mathbb{G}$ in Figure 3.25 (a). Prove that the Markov condition entails $I_P(C, G|\{A, F\})$ for $\mathbb{G}$.

Exercise 3.16 Suppose we add another variable R, an edge from F to R, and an edge from R to C to the DAG $\mathbb{G}$ in Figure 3.25 (a). The variable R might represent the professor's initial reputation. State which of the following conditional independencies you would feel are entailed by the Markov condition for $\mathbb{G}$. For each that you feel is entailed, try to prove it actually is.

1. $I_P(R, A)$
2. $I_P(R, A|F)$
3. $I_P(R, A|\{F, C\})$.

Exercise 3.17 Show that the Markov condition entails the following conditional independencies for the DAG in Figure 3.27:

1. $I_P(X, R|\{Y, Z\})$
2. $I_P(X, T|\{Y, Z\})$
3. $I_P(W, T|R)$

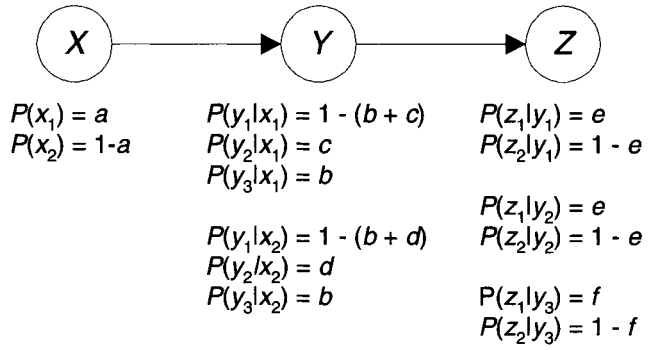

Figure 3.34: Any probability distribution P obtained by assigning values to the parameters in this network is not faithful to the DAG in the network because we have $I_P(X, Z)$.

4. $I_P(Y, Z|X)$
5. $I_P(W, S|\{R, Z\})$
6. $I_P(W, S|\{Y, Z\})$
7. $I_P(W, S|\{Y, X\})$.

Exercise 3.18 Show that the Markov condition does not entail the following conditional independencies for the DAG in Figure 3.27:

1. $I_P(W, X|Y)$
2. $I_P(W, T|Y)$.

Exercise 3.19 State which of the following conditional independencies are entailed by the Markov condition for the DAG in Figure 3.27:

1. $I_P(W, S|\{R, X\})$
2. $I_P(\{W, X\}, \{S, T\}|\{R, Z\})$
3. $I_P(\{Y, Z\}, T|\{R, S\})$
4. $I_P(\{X, S\}, \{W, T\}|\{R, Z\})$
5. $I_P(\{X, S, Z\}, \{W, T\}|R)$
6. $I_P(\{X, Z\}, W)$
7. $I_P(\{X, S, Z\}, W)$.

Does the Markov condition entail $I_P(\{X, S, Z\}, W|\mathrm{U})$ for any subset of variables U in that DAG?

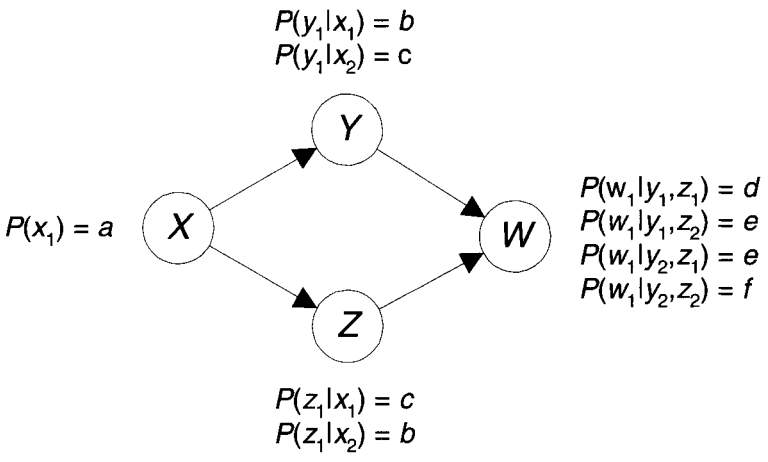

Figure 3.35: Any probability distribution P obtained by assigning values to the parameters in this network is not faithful to the DAG in the network because we have $I_P(X, W)$.

Exercise 3.20 Show $I_P(F, E)$ for the distribution P in the Bayesian network in Figure 3.28.

Exercise 3.21 Consider the Bayesian network in Figure 3.34. Show that for all assignments of values to a, b, c, d, e, and f that yield a probability distribution P, we have $I_P(X, Z)$. Such probability distributions are not faithful to the DAG in that figure because X and Z are not d-separated by the empty set. Note that the probability distribution in Figure 3.28 is a member of this family of distributions.

Exercise 3.22 Assign arbitrary values to the conditional distributions for the DAG in Figure 3.34, and see if the resultant distribution is faithful to the DAG. Try to find an unfaithful distribution besides ones in the family shown in that figure.

Exercise 3.23 Consider the Bayesian network in Figure 3.35. Show that for all assignments of values to a, b, c, d, e, and f that yield a probability distribution P, we have $I_P(X, W)$. Such probability distributions are not faithful to the DAG in that figure because X and W are not d-separated by the empty set.

If the edges in the DAG in Figure 3.35 represent causal influences, X and W would be independent if the causal effect of X on W through Y negated the causal effect of X on W through Z. If X represents an individual's age, W represents the individual's typing ability, Y represent's the individual's experience, and Z represents the individual's manual dexterity, do you feel X and W might be independent for this reason?

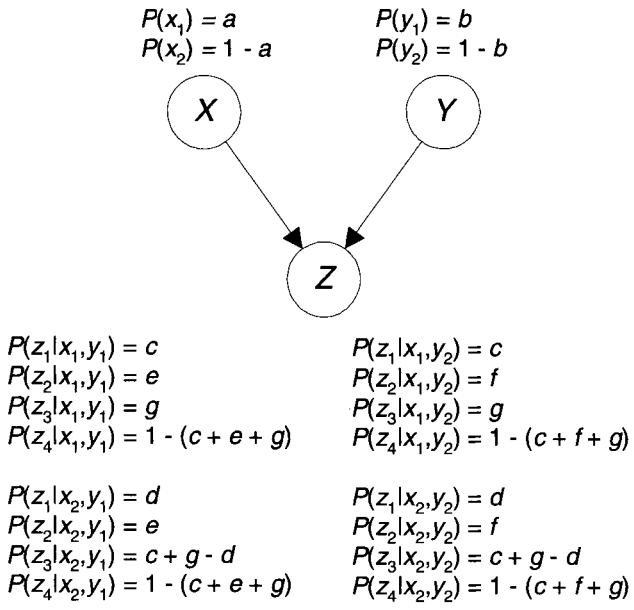

Figure 3.36: Any probability distribution P obtained by assigning values to the parameters in this network is not faithful to the DAG in the network because we have $I_P(X, Y|Z)$.

Exercise 3.24 Consider the Bayesian network in Figure 3.36. Show that for all assignments of values to a, b, c, d, e, f, and g that yield a probability distribution P, we have $I_P(X, Y|Z)$. Such probability distributions are not faithful to the DAG in that figure because X and Y are not d-separated by Z.

If the edges in the DAG in Figure 3.36 represent causal influences, X and Y would be independent given Z if no discounting occurred. Try to find some causal influences that might behave like this.

Exercise 3.25 Apply Theorem 3.4 to find a Markov blanket for each node in the DAG in Figure 3.30.

Exercise 3.26 Apply Theorem 3.4 to find a Markov blanket for each node in the DAG in Figure 3.27.