variable in the DAG has no nondescendents. (Recall that in this book we do not consider parents nondescendents.) So each variable is trivially independent of its nondescendents given its parents, and the Markov condition is satisfied. Another way to look at this is to notice that the chain rule (see Chapter 2, Section 2.2.1) says that for all values $x$, $y$, $z$, and $w$ of $X$, $Y$, $Z$, and $W$

$$P(x,y,z,w) = P(w|z,y,x)P(z|y,x)P(x).$$

So $P$ is equal to the product of its conditional distributions in the DAG in Figure 4.7, which means, owing to Chapter 3, Theorem 3.1, $P$ satisfies the Markov condition with that DAG.

Recall that our goal with a Bayesian network is to represent a probability distribution succinctly. A complete DAG does not accomplish this because, if there are $n$ binomial variables, the last variable in a complete DAG would require $2^{n-1}$ conditional distributions. To represent a distribution $P$ succinctly, we need to find a sparse DAG (one containing few edges) that satisfies the Markov condition with $P$. The next two sections present two methods for doing this.

## 4.3 Score-Based Structure Learning★

After illustrating score-based structure learning using the Bayesian score, we discuss model averaging.

### 4.3.1 Learning Structure Using the Bayesian Score

We say a DAG **includes** a probability distribution $P$ if the DAG does not entail any conditional independencies that are not in $P$. In **score-based structure learning**, we assign a score to each DAG based on the data such that in the limit (of the size of the data set) we have the following: (1) DAGs that include $P$ score higher than DAGs that do not include $P$, and (2) smaller DAGs that include $P$ score higher than larger DAGs that include $P$. After scoring the DAGs, we use the score, possibly along with prior probabilities, to learn a DAG.

The most straightforward score, called the **Bayesian score**, is the probability of the data $\mathsf{D}$ given the DAG. That is,

$$score_{Bayesian}(\mathbb{G}) = P(\mathsf{D}|\mathbb{G}).$$

We present this score shortly. However, first we need to discuss the probability of data.

#### Probability of Data

Suppose we are going to toss the same coin two times in a row. Let $X_1$ be a random variable whose value is the result of the first toss, and let $X_2$ be a random variable whose value is the result of the second toss. If we know that the probability of heads for this coin is .5 and make the usual assumption that the outcomes of the two tosses are independent, we have

$$P(X_1 = heads, X_2 = heads) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}.$$

This is a standard result. Suppose now that we are going to toss a thumbtack two times in a row, and we represent prior ignorance of the probability of heads by taking $a = b = 1$. If the outcome of the first toss is heads, using the notation developed in Section 4.1, our updated probability of heads is

$$P(heads|1,0) = \frac{a+1}{a+b+1} = \frac{1+1}{1+1+1} = \frac{2}{3}.$$

Heads is more probable for the second toss because our belief has changed owing to heads occurring on the first toss. So using our current notation in which we have articulated two random variables, we have that

$$\begin{aligned} P(X_1 = heads, X_2 = heads) &= P(X_2 = heads|X_1 = heads)P(X_1 = heads) \\ &= \frac{2}{3} \times \frac{1}{2} = \frac{1}{3}. \end{aligned}$$

In the same way

$$\begin{aligned} P(X_1 = heads, X_2 = tails) &= P(X_2 = tails|X_1 = heads)P(X_1 = heads) \\ &= \frac{1}{3} \times \frac{1}{2} = \frac{1}{6} \end{aligned}$$

$$\begin{aligned} P(X_1 = tails, X_2 = heads) &= P(X_2 = heads|X_1 = tails)P(X_1 = tails) \\ &= \frac{1}{3} \times \frac{1}{2} = \frac{1}{6} \end{aligned}$$

$$\begin{aligned} P(X_1 = tails, X_2 = tails) &= P(X_2 = tails|X_1 = tails)P(X_1 = tails) \\ &= \frac{2}{3} \times \frac{1}{2} = \frac{1}{3}. \end{aligned}$$

It may seem odd to you that the four outcomes do not have the same probability. However, recall that we do not know the probability of heads. Therefore, we learn something about the probability of heads from the result of the first toss. If heads occurs on the first toss, that probability goes up, while if tails occurs, it goes down. So two consecutive heads or two consecutive tails are more probable *a priori* than a head followed by a tail or a tail followed by a head.

The above result extends readily to a sequence of tosses. For example, suppose we toss the thumbtack three times. Then owing to the chain rule,

$$\begin{aligned} & P(X_1 = heads, X_2 = tails, X_3 = tails) \\ &= P(X_3 = tails|X_2 = tails, X_1 = heads)P(X_2 = tails|X_1 = heads) \\ &\quad P(X_1 = heads) \\ &= \frac{b+1}{a+b+2} \times \frac{b}{a+b+1} \times \frac{a}{a+b} \\ &= \frac{1+1}{1+1+2} \times \frac{1}{1+1+1} \times \frac{1}{1+1} = .0833. \end{aligned}$$

We get the same probability regardless of the order of the outcomes as long as the number of heads and tails is the same. For example,

$$\begin{aligned}
&P(X_1 = tails, X_2 = tails, X_3 = heads) \\
&= P(X_3 = heads|X_2 = tails, X_1 = tails)P(X_2 = tails|X_1 = tails) \\
&\quad P(X_1 = tails) \\
&= \frac{a}{a+b+2} \times \frac{b+1}{a+b+1} \times \frac{b}{a+b} \\
&= \frac{1}{1+1+2} \times \frac{2}{1+1+1} \times \frac{1}{1+1} = .0833.
\end{aligned}$$

This result is actually the assumption of exchangeability, which was discussed in Section 4.1.1.

We now have the following theorem:

**Theorem 4.2** *Suppose we are about to repeatedly toss a thumbtack (or perform any repeatable experiment with two outcomes), and our prior experience is equivalent to having seen $a$ heads and $b$ tails in $m$ trials, where $a$ and $b$ are positive integers. Let $D$ be data that consist of $s$ heads and $t$ tails in $n$ trials.*

$$P(D) = \frac{(m-1)!}{(m+n-1)!} \times \frac{(a+s-1)!(b+t-1)!}{(a-1)!(b-1)!}.$$

***Proof.*** *Since the probability is the same regardless of the order in which the heads and tails occur, we can assume all the heads occur first. We therefore have (as before the notation $s,t$ on the right side of the conditioning bar means we have seen $s$ heads and $t$ tails)*

$$\begin{aligned}
&P(D) \\
&= P(X_{s+t} = tails|s, t-1) \cdots P(X_{s+2} = tails|s, 1)P(X_{s+1} = tails|s, 0) \\
&\quad P(X_s = heads|s-1, 0) \cdots P(X_2 = heads|1, 0)P(X_1 = heads) \\
&= \frac{b+t-1}{a+b+s+t-1} \times \cdots \frac{b+1}{a+b+s+1} \times \frac{b}{a+b+s} \times \\
&\quad \frac{a+s-1}{a+b+s-1} \times \cdots \frac{a+1}{a+b+1} \times \frac{a}{a+b} \\
&= \frac{(a+b-1)!}{(a+b+s+t-1)!} \times \frac{(a+s-1)!}{(a-1)!} \times \frac{(b+t-1)!}{(b-1)!} \\
&= \frac{(m-1)!}{(m+n-1)!} \times \frac{(a+s-1)!(b+t-1)!}{(a-1)!(b-1)!}.
\end{aligned}$$

*This completes the proof.* ■

**Example 4.12** *Suppose, before tossing a thumbtack, we assign $a = 3$ and $b = 5$ to model the slight belief that tails is more probable than heads. We then* *toss the coin* 10 *times and obtain* 4 *heads and* 6 *tails. Owing to the preceding theorem, the probability of obtaining these data $D$ is given by*

$$\begin{aligned} P(D) &= \frac{(m-1)!}{(m+n-1)!} \times \frac{(a+s-1)!(b+t-1)!}{(a-1)!(b-1)!} \\ &= \frac{(8-1)!}{(8+10-1)!} \times \frac{(3+4-1)!(5+6-1)!}{(3-1)!(5-1)!} = .00077. \end{aligned}$$

*Note that the probability of these data is very small. This is because there are many possible outcomes (namely $2^{10}$) of tossing a thumbtack* 10 *times.*

Theorem 4.2 holds even if $a$ and $b$ are not integers. The proof can be found in [Neapolitan, 2004]. We merely state the result here.

**Theorem 4.3** *Suppose we are about to repeatedly toss a thumbtack (or perform any repeatable experiment with two outcomes), and we represent our prior experience concerning the probabilities of heads and tails using positive real numbers $a$ and $b$, where $m = a + b$. Let $D$ be data that consist of $s$ heads and $t$ tails in $n$ trials. Then*

$$P(D) = \frac{\Gamma(m)}{\Gamma(m+n)} \times \frac{\Gamma(a+s)\Gamma(b+t)}{\Gamma(a)\Gamma(b)}. \qquad (4.3)$$

In the preceding theorem $\Gamma$ denotes the gamma function. When $n$ is an integer $\geq 1$, we have that

$$\Gamma(n) = (n-1)!$$

So the preceding theorem generalizes Theorem 4.2.

**Example 4.13** *Recall that in Example 4.7 we set $a = 1/360$ and $b = 19/360$. Then after seeing* 3 *windows containing NiS, our updated values of $a$ and $b$ became as follows:*

$$\begin{aligned} a &= \frac{1}{360} + 3 = \frac{1081}{360} \\ b &= \frac{19}{360} + 0 = \frac{19}{360}. \end{aligned}$$

*We then wanted the probability that the next* 36 *windows would contain NiS. This is the probability of obtaining data with $s = 36$ and $t = 0$. Owing to the previous theorem, the probability of these data $D$ is given by*

$$\begin{aligned} P(D) &= \frac{\Gamma(m)}{\Gamma(m+n)} \times \frac{\Gamma(a+s)\Gamma(b+t)}{\Gamma(a)\Gamma(b)} \\ &= \frac{\Gamma(\frac{1100}{360})}{\Gamma(\frac{1100}{360}+36)} \times \frac{\Gamma(\frac{1081}{360}+36)\Gamma(\frac{19}{360}+0)}{\Gamma(\frac{1081}{360})\Gamma(\frac{19}{360})} \\ &= .866. \end{aligned}$$

*Recall that we obtained this same result by direct computation at the end of Example 4.7.*

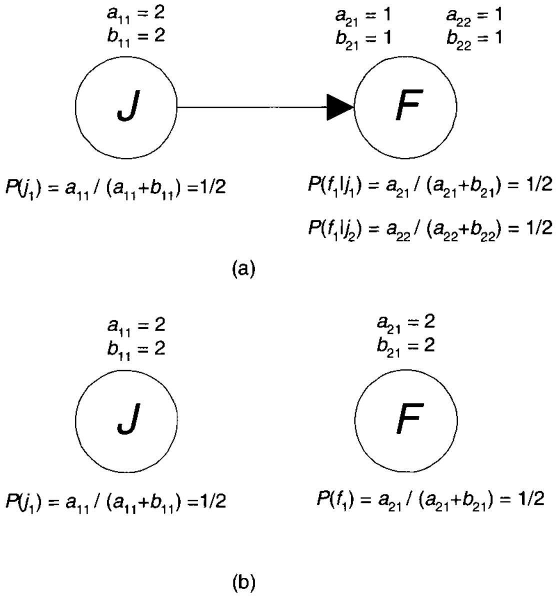
$a_{11} = 2$
$b_{11} = 2$
$a_{21} = 1$
$b_{21} = 1$
$a_{22} = 1$
$b_{22} = 1$
J
F
$P(j_1) = a_{11} / (a_{11}+b_{11}) = 1/2$
$P(f_1|j_1) = a_{21} / (a_{21}+b_{21}) = 1/2$
$P(f_1|j_2) = a_{22} / (a_{22}+b_{22}) = 1/2$
(a)
$a_{11} = 2$
$b_{11} = 2$
$a_{21} = 2$
$b_{21} = 2$
J
F
$P(j_1) = a_{11} / (a_{11}+b_{11}) = 1/2$
$P(f_1) = a_{21} / (a_{21}+b_{21}) = 1/2$
(b)

Figure 4.8: The network in (a) models that $J$ has a causal effect on $F$, while the network in (b) models that neither variable causes the other.

We developed the method for computing the probability of data for the case of binomial variables. It extends readily to multinomial variables. See [Neapolitan, 2004] for that extension.

## Probability of DAG Models and DAG Learning

Next, we show how we score a DAG model by computing the probability of the data given the model and then how we use that score to learn a DAG.

**Models with Two Nodes** First, we show how to learn two-node networks from data. Then we discuss using the learned network to do inference.

**Learning Two-Node Networks** Suppose we have a Bayesian network for learning as discussed in Section 4.1.2. For example, we may have the network in Figure 4.8 (a). Here, we call such a network a **DAG model**. We can score the DAG model $\mathbb{G}$ based on data $\mathsf{D}$ by determining how probable the data are given the DAG model. That is, we compute $P(\mathsf{D}|\mathbb{G})$. The formula for this probability is the same as that developed in the previous subsection, except there is a term of the form in Equality 4.3 for each probability in the network. So the probability of data $\mathsf{D}$ given the DAG model $\mathbb{G}_1$ in Figure 4.8 (a) is given by

$$\begin{aligned} P(\mathsf{D}|\mathbb{G}_1) = {} & \frac{\Gamma(m_{11})}{\Gamma(m_{11}+n_{11})} \times \frac{\Gamma(a_{11}+s_{11})\Gamma(b_{11}+t_{11})}{\Gamma(a_{11})\Gamma(b_{11})} \times \\ & \frac{\Gamma(m_{21})}{\Gamma(m_{21}+n_{21})} \times \frac{\Gamma(a_{21}+s_{21})\Gamma(b_{21}+t_{21})}{\Gamma(a_{21})\Gamma(b_{21})} \times \\ & \frac{\Gamma(m_{22})}{\Gamma(m_{22}+n_{22})} \times \frac{\Gamma(a_{22}+s_{22})\Gamma(b_{22}+t_{22})}{\Gamma(a_{22})\Gamma(b_{22})}. \end{aligned} \tag{4.4}$$

The data used in each term include only the data relevant to the conditional probability the term represents. This is exactly the same scheme that was used to learn parameters in Section 4.1.2. For example, the value of $s_{21}$ is the number of cases that have $J$ equal to $j_2$ and $F$ equal to $f_1$.

Similarly, the probability of data D given the DAG model $\mathbb{G}_2$ in Figure 4.8 (b) is given by

$$\begin{aligned} P(\mathsf{D}|\mathbb{G}_2) = {} & \frac{\Gamma(m_{11})}{\Gamma(m_{11}+n_{11})} \times \frac{\Gamma(a_{11}+s_{11})\Gamma(b_{11}+t_{11})}{\Gamma(a_{11})\Gamma(b_{11})} \times \\ & \frac{\Gamma(m_{21})}{\Gamma(m_{21}+n_{21})} \times \frac{\Gamma(a_{21}+s_{21})\Gamma(b_{21}+t_{21})}{\Gamma(a_{21})\Gamma(b_{21})}. \end{aligned} \tag{4.5}$$

Note that the values of $a_{11}, s_{11}$, etc. in Equality 4.5 are the ones relevant to $\mathbb{G}_2$ and are not the same values as those in Equality 4.4. We have not explicitly shown their dependence on the DAG model for the sake of notational simplicity.

**Example 4.14** *Suppose we wish to determine whether job status ($J$) has a causal effect on whether someone defaults on a loan ($F$). Furthermore, we articulate just two values for each variable as follows:*

| Variable | Value | When the Variable Takes This Value |
|---|---|---|
| $J$ | $j_1$ | Individual is a white collar worker. |
| | $j_2$ | Individual is a blue collar worker. |
| $F$ | $f_1$ | Individual has defaulted on a loan at least once. |
| | $f_2$ | Individual has never defaulted on a loan. |

*We represent the assumption that $J$ has a causal effect on $F$ with the DAG model $\mathbb{G}_1$ in Figure 4.8 (a) and the assumption that neither variable has a causal effect on the other with the DAG model $\mathbb{G}_2$ in Figure 4.8 (b). We assume that $F$ does not have a causal effect on $J$; so we do not model this situation. Note that in both models we used a prior equivalent sample size of* 4 *and we represented the prior belief that all probabilities are .5. In general, we can use whatever prior sample size and prior belief that best model our prior knowledge. The only requirement is that both DAG models have the same prior equivalent sample size.*

*Suppose that next we obtain the data* D *in the following table:*

| Case | $J$ | $F$ |
|---|---|---|
| 1 | $j_1$ | $f_1$ |
| 2 | $j_1$ | $f_1$ |
| 3 | $j_1$ | $f_1$ |
| 4 | $j_1$ | $f_1$ |
| 5 | $j_1$ | $f_2$ |
| 6 | $j_2$ | $f_1$ |
| 7 | $j_2$ | $f_2$ |
| 8 | $j_2$ | $f_2$ |

*Then, owing to Equality 4.4,*

$$\begin{aligned}
P(\mathsf{D}|\mathbb{G}_1) &= \frac{\Gamma(m_{11})}{\Gamma(m_{11}+n_{11})} \times \frac{\Gamma(a_{11}+s_{11})\Gamma(b_{11}+t_{11})}{\Gamma(a_{11})\Gamma(b_{11})} \times \\
&\quad \frac{\Gamma(m_{21})}{\Gamma(m_{21}+n_{21})} \times \frac{\Gamma(a_{21}+s_{21})\Gamma(b_{21}+t_{21})}{\Gamma(a_{21})\Gamma(b_{21})} \times \\
&\quad \frac{\Gamma(m_{22})}{\Gamma(m_{22}+n_{22})} \times \frac{\Gamma(a_{22}+s_{22})\Gamma(b_{22}+t_{22})}{\Gamma(a_{22})\Gamma(b_{22})} \\
&= \frac{\Gamma(4)}{\Gamma(4+8)} \times \frac{\Gamma(2+5)\Gamma(2+3)}{\Gamma(2)\Gamma(2)} \times \\
&\quad \frac{\Gamma(2)}{\Gamma(2+5)} \times \frac{\Gamma(1+4)\Gamma(1+1)}{\Gamma(1)\Gamma(1)} \times \\
&\quad \frac{\Gamma(2)}{\Gamma(2+3)} \times \frac{\Gamma(1+1)\Gamma(1+2)}{\Gamma(1)\Gamma(1)} \\
&= 7.2150 \times 10^{-6}.
\end{aligned}$$

*Furthermore,*

$$\begin{aligned}
P(\mathsf{D}|\mathbb{G}_2) &= \frac{\Gamma(m_{11})}{\Gamma(m_{11}+n_{11})} \times \frac{\Gamma(a_{11}+s_{11})\Gamma(b_{11}+t_{11})}{\Gamma(a_{11})\Gamma(b_{11})} \times \\
&\quad \frac{\Gamma(m_{21})}{\Gamma(m_{21}+n_{21})} \times \frac{\Gamma(a_{21}+s_{21})\Gamma(b_{21}+t_{21})}{\Gamma(a_{21})\Gamma(b_{21})} \times \\
&= \frac{\Gamma(4)}{\Gamma(4+8)} \times \frac{\Gamma(2+5)\Gamma(2+3)}{\Gamma(2)\Gamma(2)} \times \\
&\quad \frac{\Gamma(4)}{\Gamma(4+8)} \times \frac{\Gamma(2+5)\Gamma(2+3)}{\Gamma(2)\Gamma(2)} \\
&= 6.7465 \times 10^{-6}.
\end{aligned}$$

*If our prior belief is that neither model is more probable than the other, we assign*

$$P(\mathbb{G}_1) = P(\mathbb{G}_2) = .5.$$

*Then, owing to Bayes' Theorem,*

$$\begin{aligned} P(\mathbb{G}_1|\mathsf{D}) &= \frac{P(\mathsf{D}|\mathbb{G}_1)P(\mathbb{G}_1)}{P(\mathsf{D}|\mathbb{G}_1)P(\mathbb{G}_1)+P(\mathsf{D}|\mathbb{G}_2)P(\mathbb{G}_2)} \\ &= \frac{7.2150\times 10^{-6}\times .5}{7.2150\times 10^{-6}\times .5+6.7465\times 10^{-6}\times .5} \\ &= .517 \end{aligned}$$

*and*

$$\begin{aligned} P(\mathbb{G}_2|\mathsf{D}) &= \frac{P(\mathsf{D}|\mathbb{G}_2)P(\mathbb{G}_2)}{P(\mathsf{D})} \\ &= \frac{6.7465\times 10^{-6}(.5)}{7.2150\times 10^{-6}\times .5+6.7465\times 10^{-6}\times .5} \\ &= .483. \end{aligned}$$

*We select (learn) DAG $\mathbb{G}_1$ and conclude that it is more probable that job status does have a causal effect on whether someone defaults on a loan.*

**Example 4.15** *Suppose we are doing the same study as in Example 4.14, and we obtain the data in the following table:*

| Case | $J$ | $F$ |
|---|---|---|
| 1 | $j_1$ | $f_1$ |
| 2 | $j_1$ | $f_1$ |
| 3 | $j_1$ | $f_1$ |
| 4 | $j_1$ | $f_1$ |
| 5 | $j_2$ | $f_2$ |
| 6 | $j_2$ | $f_2$ |
| 7 | $j_2$ | $f_2$ |
| 8 | $j_2$ | $f_2$ |

*Then it is left as an exercise to show*

$$P(\mathsf{D}|\mathbb{G}_1) = 8.6580\times 10^{-5}$$

$$P(\mathsf{D}|\mathbb{G}_2) = 4.6851\times 10^{-6},$$

*and if we assign the same prior probability to both DAG models,*

$$\begin{aligned} P(\mathbb{G}_1|\mathsf{D}) &= .949 \\ P(\mathbb{G}_2|\mathsf{D}) &= .051. \end{aligned}$$

*We select (learn) DAG $\mathbb{G}_1$. Notice that the causal model is substantially more probable. This makes sense because even though there is not much data, it exhibits complete dependence.*

**Example 4.16** *Suppose we are doing the same study as in Example 4.14, and we obtain the data $\mathsf{D}$ in the following table:*

| Case | $J$ | $F$ |
|---|---|---|
| 1 | $j_1$ | $f_1$ |
| 2 | $j_1$ | $f_1$ |
| 3 | $j_1$ | $f_2$ |
| 4 | $j_1$ | $f_2$ |
| 5 | $j_2$ | $f_1$ |
| 6 | $j_2$ | $f_1$ |
| 7 | $j_2$ | $f_2$ |
| 8 | $j_2$ | $f_2$ |

*Then it is left as an exercise to show*

$$P(\mathsf{D}|\mathbb{G}_1) = 2.4050 \times 10^{-6}$$

$$P(\mathsf{D}|\mathbb{G}_2) = 4.6851 \times 10^{-6},$$

*and if we assign the same prior probability to both DAG models,*

$$\begin{aligned} P(\mathbb{G}_1|\mathsf{D}) &= .339 \\ P(\mathbb{G}_2|\mathsf{D}) &= .661. \end{aligned}$$

*We select (learn) DAG $\mathbb{G}_2$. Notice that it is somewhat more probable that the two variables are independent. This makes sense since the data exhibit complete independence.*

**Using the Learned Network to Do Inference** Once we learn a DAG from data, we can then learn the parameters. The result will be a Bayesian network which we can use to do inference for the next case. The next example illustrates the technique.

**Example 4.17** *Suppose we have the situation and data in Example 4.14. That is, we have the data $\mathsf{D}$ in the following table:*

| Case | $J$ | $F$ |
|---|---|---|
| 1 | $j_1$ | $f_1$ |
| 2 | $j_1$ | $f_1$ |
| 3 | $j_1$ | $f_1$ |
| 4 | $j_1$ | $f_1$ |
| 5 | $j_1$ | $f_2$ |
| 6 | $j_2$ | $f_1$ |
| 7 | $j_2$ | $f_2$ |
| 8 | $j_2$ | $f_2$ |

*Then, as shown in Example 4.14, we would learn the DAG in Figure 4.8 (a). Next, we can update the conditional probabilities in the Bayesian network for learning in Figure 4.8 (a) using the above data and the parameter learning technique discussed in Section 4.1.2. The result is the Bayesian network in Figure 4.9. Suppose now that we find out that Sam has $F = f_2$. That is, Sam*

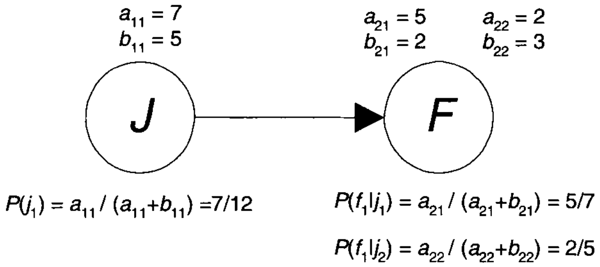
$a_{11} = 7$
$b_{11} = 5$
$a_{21} = 5$
$b_{21} = 2$
$a_{22} = 2$
$b_{22} = 3$
J
F
$P(j_1) = a_{11} / (a_{11}+b_{11}) = 7/12$
$P(f_1|j_1) = a_{21} / (a_{21}+b_{21}) = 5/7$
$P(f_1|j_2) = a_{22} / (a_{22}+b_{22}) = 2/5$

Figure 4.9: An updated Bayesian network for learning based on the data in Example 4.17.

*has never defaulted on a loan. We can use the Bayesian network to compute the probability that Sam is a white collar worker. For this simple network we can just use Bayes' Theorem as follows*

$$\begin{aligned} P(j_1|f_1) &= \frac{P(f_1|j_1)P(j_1)}{P(f_1|j_1)P(j_1) + Pf_1|j_2)P(j_2)} \\ &= \frac{(5/7)\,(7/12)}{(5/7)\,(7/12) + (2/5)\,(5/12)} = .714. \end{aligned}$$

*The probabilities in the previous calculation are all conditional on the data* D *and the DAG that we select. However, once we select a DAG and learn the parameters, we do not bother to show this dependence.*

**Bigger DAG Models** We illustrated scoring using only two variables. Ordinarily, we want to learn structure when there are many more variables. When there are only a few variables, we can exhaustively score all possible DAGs as was done in the previous examples. We then select the DAG(s) with the highest score. However, when the number of variables is not small, to find the maximizing DAGs by exhaustively considering all DAG patterns is computationally unfeasible. That is, Robinson [1977] showed that the number of DAGs containing $n$ nodes is given by the following recurrence:

$$\begin{aligned} f(n) &= \sum_{i=1}^{n} (-1)^{i+1} \binom{n}{i} 2^{i(n-i)} f(n-i) \qquad n > 2 \\ f(0) &= 1 \\ f(1) &= 1. \end{aligned}$$

It is left as an exercise to show $f(2) = 3$, $f(3) = 25$, $f(5) = 29{,}000$, and $f(10) = 4.2 \times 10^{18}$. Furthermore, Chickering [1996] proved that for certain classes of prior distributions, the problem of finding a highest scoring DAG is NP-hard. So researchers developed heuristic DAG search algorithms. It is beyond the scope of this book to discuss these algorithms. See [Neapolitan, 2004] for a detailed discussion of them.

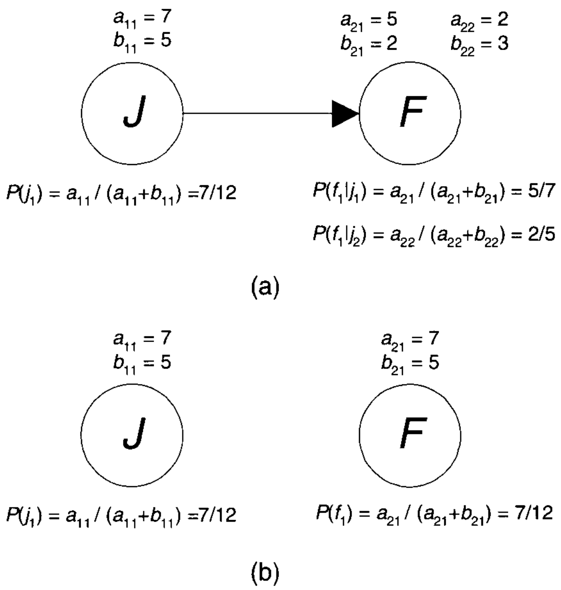
$a_{11} = 7$
$b_{11} = 5$
$a_{21} = 5$
$b_{21} = 2$
$a_{22} = 2$
$b_{22} = 3$
J
F
$P(j_1) = a_{11} / (a_{11}+b_{11}) = 7/12$
$P(f_1|j_1) = a_{21} / (a_{21}+b_{21}) = 5/7$
$P(f_1|j_2) = a_{22} / (a_{22}+b_{22}) = 2/5$
(a)
$a_{11} = 7$
$b_{11} = 5$
$a_{21} = 7$
$b_{21} = 5$
J
F
$P(j_1) = a_{11} / (a_{11}+b_{11}) = 7/12$
$P(f_1) = a_{21} / (a_{21}+b_{21}) = 7/12$
(b)

Figure 4.10: Updated Bayesian network for learning based on the data in Examples 4.14 and 4.17.

## 4.3.2 Model Averaging

Heckerman et al. [1999] illustrate that when the number of variables is small and the amount of data is large, one structure can be orders of magnitude more likely than any other. In such cases, model selection yields good results. However, recall that in Example 4.14 we had little data, we obtained $P(\mathbb{G}_1|\mathsf{D}) = .517$ and $P(\mathbb{G}_2|\mathsf{D}) = .483$, and we chose (learned) DAG $\mathbb{G}_1$ because it was most probable. Then in Example 4.17 we used a Bayesian network containing DAG $\mathbb{G}_1$ to do inference for Sam. Since the probabilities of the two models are so close, it seems somewhat arbitrary to choose $\mathbb{G}_1$. So model selection does not seem appropriate. Next, we describe another approach.

Instead of choosing a single DAG and then using it to do inference, we could use the law of total probability to do the inference as follows: we perform the inference using each DAG and multiply the result (a probability value) by the posterior probably of the DAG. This is called **model averaging**.

**Example 4.18** *Recall that based on the data in Example 4.14 we learned that*

$$P(\mathbb{G}_1|\mathsf{D}) = .517$$

*and*

$$P(\mathbb{G}_2|\mathsf{D}) = .483.$$

*In Example 4.17 we updated a Bayesian network containing $\mathbb{G}_1$ based on the data to obtain the Bayesian network in Figure 4.10 (a). If in the same way we*