EXERCISES 13

PROBABILISTIC REASONING

13.1 Representing Knowledge in an Uncertain Domain

Exercise 13.MRFS

In this chapter we present Bayesian networks, originally called belief networks, for the representation of uncertainty but other types of graphical models may also be used. Consider, **Markov networks**, also called Markov random fields, cited in the bibliographical and historical notes. What differentiates Markov and Bayesian networks? What are advantages and disadvantages of each? What is a domain in which each network might be more suitable than the other?

Bayesian networks are directed acyclic graphs; Markov networks are undirected cyclic graphs. These formulations necessarily admit different independence assumptions; d-separation is necessary to determine the conditional independence of nodes in a Bayesian network while only separation between two nodes (through any path) is necessary in a Markov network. Additionally, values in Markov networks do not represent probabilities but arbitrary-valued potentials which, in Bayesian networks, must be conditional probabilities. Because of these differences finding the joint probability of a Markov networks is NP-hard while it is considerably easier in a Bayesian network (one must compute the product of all nodes in a Markov network as opposed to successive products of directed subsets in a Bayes net). Medical diagnosis in which nodes (such as symptoms and diseases) represent causal relationships would better suit Bayesian networks while image processing in which nodes (such as pixels in an image) do not represent causality would better suit Markov networks.

13.2 The Semantics of Bayesian Networks

Exercise 13.DAGS

Recall that any directed acyclic graph G has an associated family of probability distributions, which consists of all probability distributions that can be represented by a Bayes net with structure G. Consider the the six directed acyclic graphs in Figure S13.1.

a. Assume all we know about the joint distribution $P(A, B, C)$ is that it can be represented by the product $P(A \mid B, C)P(B \mid C)P(C)$. Which of the six graphs are guaranteed to be able to represent $P(A, B, C)$?

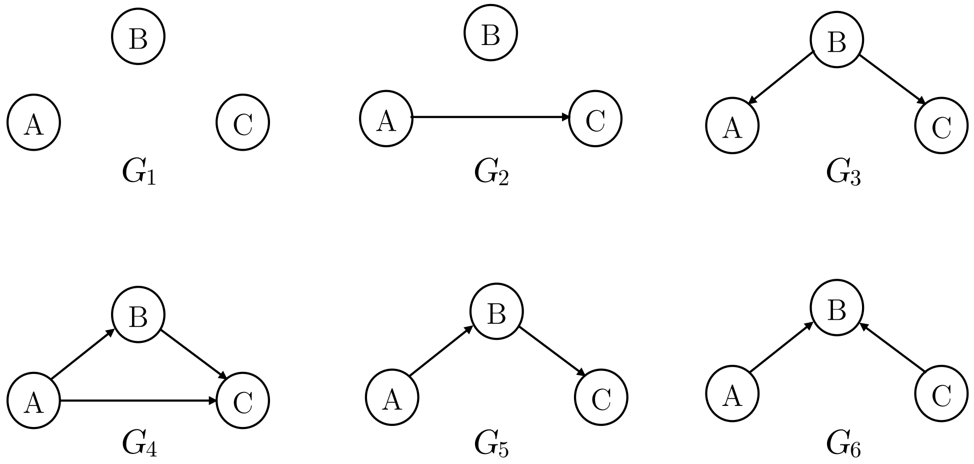


Figure S13.1 Six directed acyclic graphics.

b. Now assume all we know about the joint distribution $P(A,B,C)$ is that it can be represented by the product $P(C \mid B)P(B \mid A)P(A)$. Which of the six graphs are guaranteed to be able to represent $P(A,B,C)$?

a. In this case, the decomposition of $P(A,B,C)$ is just that given by the chain rule, so no conditional independence is implied. G_4 is fully connected, and is therefore able to represent any joint distribution. The others assert conditional independences and so may not be able to represent $P(A,B,C)$.

b. G_3, G_4, and G_5 can represent the distribution. G_1 assumes all variables are independent; G_2 asserts that B is independent of the others; and G_6 asserts $A \perp\!\!\!\perp C$.

Exercise 13.ABDE

Consider a Bayes net over the random variables A, B, C, D, E with the structure shown in Figure S13.2, with full joint distribution $P(A,B,C,D,E)$.

a. Consider the marginal distribution $P(A,B,D,E) = \sum_c P(A,B,c,D,E)$, where C was eliminated. Draw the minimal Bayes net that is guaranteed to be able to represent this distribution.

b. Assume we are given an observation: $A=a$. Draw the minimal Bayes net that is guaranteed to be able to represent the distribution $P'(B,C,D,E) = P(B,C,D,E \mid A=a)$.

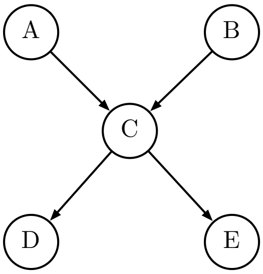
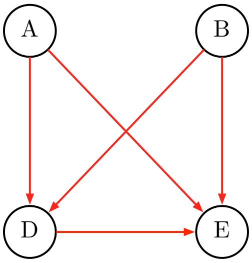


Figure S13.2 A Bayes net over the random variables A, B, C, D, E.

c. Assume we are given two observations: $D = d, E = e$. Draw the minimal Bayes net that is guaranteed to be able to represent the distribution $P''(A, B, C) = P(A, B, C \mid D = d, E = e)$.

a. For $P(A, B, D, E)$ we have



The high level overview for these types of problems is that the resultant graph must be able to encode the same conditional independence assumptions from the initial Bayes net. In the BN above, we have the following independence assumptions:

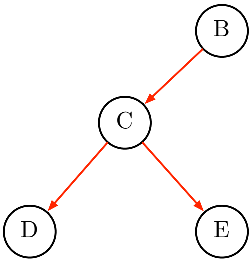
- $A \perp\!\!\!\perp B$
- $A \perp\!\!\!\perp D \mid C$
- $B \perp\!\!\!\perp D \mid C$
- $A \perp\!\!\!\perp E \mid C$
- $B \perp\!\!\!\perp E \mid C$
- $D \perp\!\!\!\perp E \mid C$

When we marginalize out C, we remove C from the graph. The conditional independence assumptions involving C no longer matter, so we just need to preserve:

$$A \perp\!\!\!\perp B\,.$$

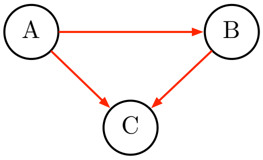
The graph shown guarantees this. The arrow between D and E could go either way.

b. For $P'(B, C, D, E) = P(B, C, D, E \mid A = a)$, we have

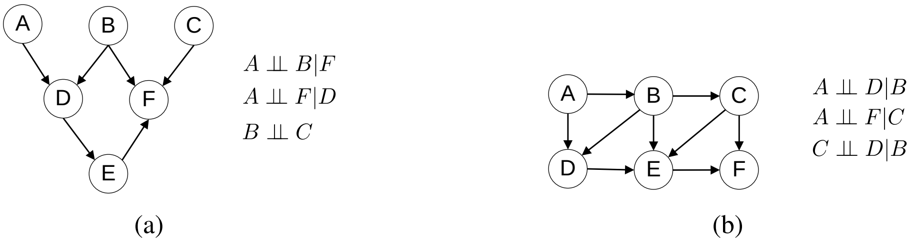


Observing A fixes its value and removes it from the Bayes net. By d-separation no further dependence is introduced.

c. For $P''(A, B, C) = P(A, B, C \mid D = d, E = e)$ we have



Observing D and E makes an active path to their parent C, which in turn activates the common-effect triple ABC and renders A and B dependent.



$A \perp\!\!\!\perp B | F$
$A \perp\!\!\!\perp F | D$
$B \perp\!\!\!\perp C$

$A \perp\!\!\!\perp D | B$
$A \perp\!\!\!\perp F | C$
$C \perp\!\!\!\perp D | B$

(a) (b)

Figure S13.3 Figure for Exercise 13.RMVE.

Exercise 13.RMVE

For the graphs in Figure S13.3, what is the minimal set of edges that must be removed such that the corresponding independence relations are guaranteed to be true?

a. Remove AD.

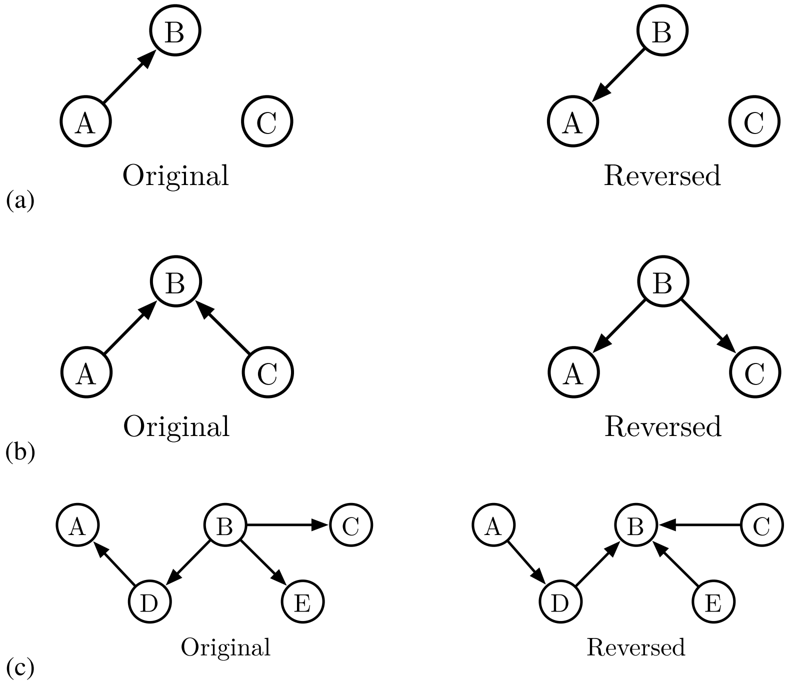


Figure S13.4 Figure for Exercise 13.REVB.

b. Remove AD and either EF or AB.

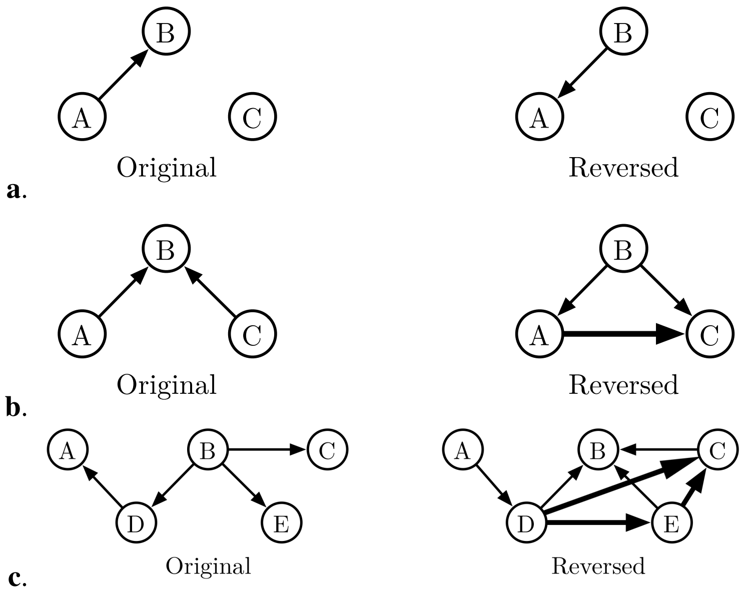
Exercise 13.REVB

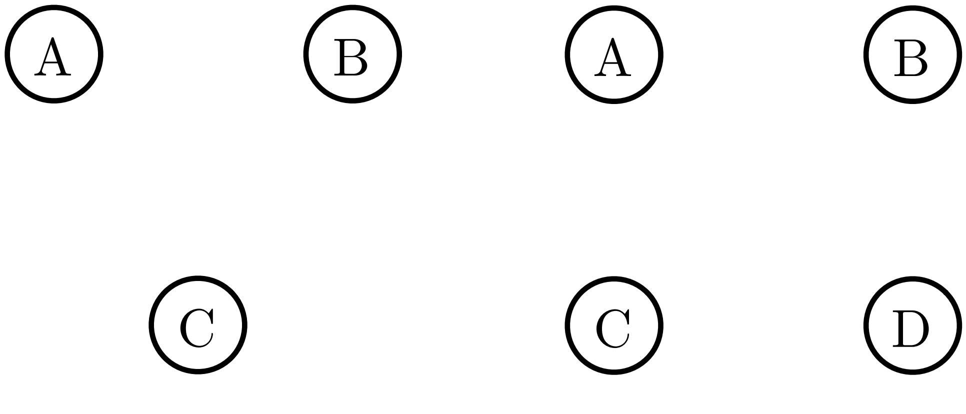
Figure S13.3 shows pairs of Bayes nets. In each, the **original** network is shown on the left. The **reversed** network, shown on the right, has all the arrows reversed. Therefore, the reversed network may not be able to represent all of the distributions that the original network is able to represent.

For each pair, add the *minimal* number of arrows to the **reversed** network such that it is guaranteed to be able to represent all the distributions that the **original** network is able to represent.

The reversal of common effects and common causes requires more arrows so that the dependence of children (with parents unobserved) and the dependence of parents (with a common child observed) can be captured.

To guarantee that a Bayes net B' is able to represent all of the distributions of an original Bayes net B, B' must only make a subset of the independence assumptions in B. If independences are in B' that are not in B, then the distributions of B' have constraints that could prevent it from representing a distribution of B.





(a) (b)

Figure S13.5 Figure for Exercise 13.MINB.

Exercise 13.MINB

For the following Bayes nets, add the *minimal* number of arrows such that the resulting structure is able to represent all distributions that satisfy the stated independence and non-independence constraints (note these are constraints on the Bayes net *structure*, so each non-independence constraint should be interpreted as disallowing all structures that make the given independence assumption). If no arrows are needed, write “none needed.” If no such Bayes net is possible, write “not possible.”

a. See Figure S13.5 (a). **Constraints:**

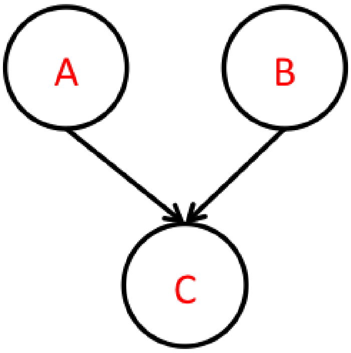
- $A \perp\!\!\!\perp B$
- not $A \perp\!\!\!\perp B \,|\, \{C\}$

b. See Figure S13.5 (b). **Constraints:**

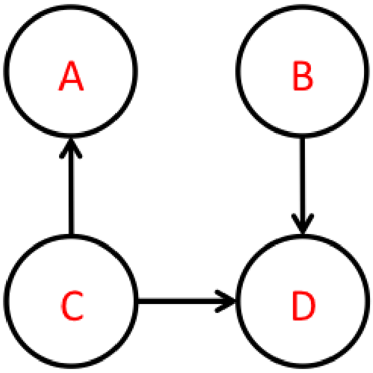
- $A \perp\!\!\!\perp D \,|\, \{C\}$
- $A \perp\!\!\!\perp B$
- $B \perp\!\!\!\perp C$
- not $A \perp\!\!\!\perp B \,|\, \{D\}$

c. See Figure S13.5 (b). **Constraints:**

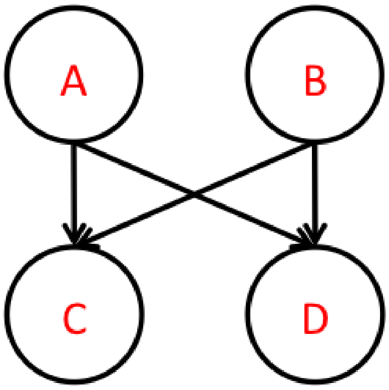
- $A \perp\!\!\!\perp B$
- $C \perp\!\!\!\perp D \,|\, \{A, B\}$
- not $C \perp\!\!\!\perp D \,|\, \{A\}$
- not $C \perp\!\!\!\perp D \,|\, \{B\}$



a.



b.



c.

Exercise 13.CPTE

Equation (13.1) on page 433 defines the joint distribution represented by a Bayesian network in terms of the parameters $\theta(X_i \,|\, Parents(X_i))$. This exercise asks you to derive the equivalence between the parameters and the conditional probabilities $\mathbf{P}(X_i \mid Parents(X_i))$ from this definition.

a. Consider a simple network $X \rightarrow Y \rightarrow Z$ with three Boolean variables. Use Equations (12.3) and (12.7) (pages 390 and 396) to express the conditional probability $P(z \,|\, y)$ as the ratio of two sums, each over entries in the joint distribution $\mathbf{P}(X, Y, Z)$.

b. Now use Equation (13.1) to write this expression in terms of the network parameters $\theta(X)$, $\theta(Y \,|\, X)$, and $\theta(Z \,|\, Y)$.

c. Next, expand out the summations in your expression from part (b), writing out explicitly the terms for the true and false values of each summed variable. Assuming that all network parameters satisfy the constraint $\sum_{x_i} \theta(x_i \,|\, parents(X_i)) = 1$, show that the resulting expression reduces to $\theta(z \,|\, y)$.

d. Generalize this derivation to show that $\theta(X_i \,|\, Parents(X_i)) = \mathbf{P}(X_i \,|\, Parents(X_i))$ for any Bayesian network.

This question is quite tricky and students may require additional guidance, particularly on the last part. It does, however, help them become comfortable with operating on complex sum-of-product expressions, which are at the heart of graphical models.

a. By Equations (12.3) and (12.7), we have

$$P(z \,|\, y) = \frac{P(y, z)}{P(y)} = \frac{\sum_x P(x, y, z)}{\sum_{x,z'} P(x, y, z')} .$$

b. By Equation (13.1), this can be written as

$$P(z \,|\, y) = \frac{\sum_x \theta(x)\theta(y \,|\, x)\theta(z \,|\, y)}{\sum_{x,z'} \theta(x)\theta(y \,|\, x)\theta(z' \,|\, y)} .$$

c. For students who are not familiar with direct manipulation of summation expressions, the expanding-out step makes it a bit easier to see how to simplify the expressions. Expanding out the sums, collecting terms, using the sum-to-1 property of the parameters, and finally cancelling, we have

$$\begin{aligned}
P(z\,|\,y) &= \frac{\theta(x)\theta(y\,|\,x)\theta(z\,|\,y) + \theta(\neg x)\theta(y\,|\,\neg x)\theta(z\,|\,y)}{\theta(x)\theta(y\,|\,x)\theta(z\,|\,y)+\theta(x)\theta(y\,|\,x)\theta(\neg z\,|\,y)+\theta(\neg x)\theta(y\,|\,\neg x)\theta(z\,|\,y)+\theta(\neg x)\theta(y\,|\,\neg x)\theta(\neg z\,|\,y)}\\
&= \frac{\theta(z\,|\,y)\,\left[\theta(x)\theta(y\,|\,x) + \theta(\neg x)\theta(y\,|\,\neg x)\right]}{\left[\theta(x)\theta(y\,|\,x) + \theta(\neg x)\theta(y\,|\,\neg x)\right]\,\left[\theta(z\,|\,y) + \theta(\neg z\,|\,y)\right]}\\
&= \frac{\theta(z\,|\,y)\,\left[\theta(x)\theta(y\,|\,x) + \theta(\neg x)\theta(y\,|\,\neg x)\right]}{\left[\theta(x)\theta(y\,|\,x) + \theta(\neg x)\theta(y\,|\,\neg x)\right]}\\
&= \theta(z\,|\,y)\ .
\end{aligned}$$

If, instead, students are prepared to work on the summations directly, the key step is moving the sum over z' inwards::

$$\begin{aligned}
P(z\,|\,y) &= \frac{\theta(z\,|\,y)\ \sum_x \theta(x)\theta(y\,|\,x)}{\sum_x \theta(x)\theta(y\,|\,x)\ \sum_{z'} \theta(z'\,|\,y)}\\
&= \frac{\theta(z\,|\,y)\ \sum_x \theta(x)\theta(y\,|\,x)}{\sum_x \theta(x)\theta(y\,|\,x)}\\
&= \theta(z\,|\,y)\ .
\end{aligned}$$

d. The general case is a bit more difficult—the key to a simple proof is figuring out how to split up all the variables. First, however, we need a little lemma: for any set of variables **V**, we have

$$\sum_{\mathbf{v}} \prod_i \theta(v_i\,|\,pa(V_i)) = 1\ .$$

This generalizes the sum-to-1 rule for a single variable, and is easily proved by induction given any topological ordering for the variables in **V**.

One of the principal rules for manipulating nested summations is that a particular summation can be pushed to the right as long as all occurrences of that variable remain to the right of the summation. For this reason, the *descendants* of Z, which we will call **U**, are a very useful subset of the variables in the network. In particular, they have the property that they cannot be parents of any other variable in the network. (If there was such a variable, it would be a descendant of Z by definition!) We will divide the variables into Z, **Y** (the parents of Z), **U** (the descendants of Z), and **X** (all other variables). We know that variables in $\mathbf{X}$ and $\mathbf{Y}$ have no parents in Z and $\mathbf{U}$. So we have

$$
\begin{aligned}
P(z \,|\, \mathbf{y}) &= \frac{\sum_{\mathbf{x},\mathbf{u}} P(\mathbf{x},\mathbf{y},z,\mathbf{u})}{\sum_{\mathbf{x},z',\mathbf{u}} P(\mathbf{x},\mathbf{y},z',\mathbf{u})} \\
&= \frac{\sum_{\mathbf{x},\mathbf{u}} \prod_i \theta(x_i \,|\, pa(X_i)) \prod_j \theta(y_j \,|\, pa(Y_j)) \theta(z \,|\, \mathbf{y}) \prod_k \theta(u_k \,|\, pa(U_k))}{\sum_{\mathbf{x},z',\mathbf{u}} \prod_i \theta(x_i \,|\, pa(X_i)) \prod_j \theta(y_j \,|\, pa(Y_j)) \theta(z' \,|\, \mathbf{y}) \prod_k \theta(u_k \,|\, pa(U_k))} \\
&= \frac{\sum_{\mathbf{x}} \prod_i \theta(x_i \,|\, pa(X_i)) \prod_j \theta(y_j \,|\, pa(Y_j)) \theta(z \,|\, \mathbf{y}) \sum_{\mathbf{u}} \prod_k \theta(u_k \,|\, pa(U_k))}{\sum_{\mathbf{x}} \prod_i \theta(x_i \,|\, pa(X_i)) \prod_j \theta(y_j \,|\, pa(Y_j)) \sum_{z'} \theta(z' \,|\, \mathbf{y}) \sum_{\mathbf{u}} \prod_k \theta(u_k \,|\, pa(U_k))} \\
&\qquad \text{(moving the sums in as far as possible)} \\
&= \frac{\sum_{\mathbf{x}} \prod_i \theta(x_i \,|\, pa(X_i)) \prod_j \theta(y_j \,|\, pa(Y_j)) \theta(z \,|\, \mathbf{y})}{\sum_{\mathbf{x}} \prod_i \theta(x_i \,|\, pa(X_i)) \prod_j \theta(y_j \,|\, pa(Y_j)) \sum_{z'} \theta(z' \,|\, \mathbf{y})} \\
&\qquad \text{(using the generalized sum-to-1 rule for } \mathbf{u}\text{)} \\
&= \frac{\sum_{\mathbf{x}} \prod_i \theta(x_i \,|\, pa(X_i)) \prod_j \theta(y_j \,|\, pa(Y_j)) \theta(z \,|\, \mathbf{y})}{\sum_{\mathbf{x}} \prod_i \theta(x_i \,|\, pa(X_i)) \prod_j \theta(y_j \,|\, pa(Y_j))} \\
&\qquad \text{(using the sum-to-1 rule for } z'\text{)} \\
&= \theta(z \,|\, \mathbf{y}) \,.
\end{aligned}
$$

Exercise 13.ARCR

The operation of **arc reversal** in a Bayesian network allows us to change the direction of an arc $X \to Y$ while preserving the joint probability distribution that the network represents (Shachter, 1986). Arc reversal may require introducing new arcs: all the parents of X also become parents of Y, and all parents of Y also become parents of X.

a. Assume that X and Y start with m and n parents, respectively, and that all variables have k values. By calculating the change in size for the CPTs of X and Y, show that the total number of parameters in the network cannot decrease during arc reversal. (*Hint*: the parents of X and Y need not be disjoint.)

b. Under what circumstances can the total number remain constant?

c. Let the parents of X be $\mathbf{U} \cup \mathbf{V}$ and the parents of Y be $\mathbf{V} \cup \mathbf{W}$, where $\mathbf{U}$ and $\mathbf{W}$ are disjoint. The formulas for the new CPTs after arc reversal are as follows:

$$
\begin{aligned}
\mathbf{P}(Y \,|\, \mathbf{U},\mathbf{V},\mathbf{W}) &= \sum_x \mathbf{P}(Y \,|\, \mathbf{V},\mathbf{W},x) \mathbf{P}(x \,|\, \mathbf{U},\mathbf{V}) \\
\mathbf{P}(X \,|\, \mathbf{U},\mathbf{V},\mathbf{W},Y) &= \mathbf{P}(Y \,|\, X,\mathbf{V},\mathbf{W}) \mathbf{P}(X \,|\, \mathbf{U},\mathbf{V}) / \mathbf{P}(Y \,|\, \mathbf{U},\mathbf{V},\mathbf{W}) \,.
\end{aligned}
$$

Prove that the new network expresses the same joint distribution over all variables as the original network.

a. Suppose that X and Y share l parents. After the reversal Y will gain $m-l$ new parents, the $m-l$ original parents of X that it does not share with Y, and loses one parent: X. After the reversal X will gain $n-l$ new parents, the $n-l-1$ original parents of Y that it does not share with X and isn't X itself, and plus Y. So, after the reversal Y will have $n+(m-l-1) = m+(n-l-1)$ parents, and X will have $m+(n-l) = n+(m-l)$

parents.

Observe that $m - l \geq 0$, since this is the number of original parents of X not shared with Y, and that $n - l - 1 \geq 0$, since this is the number of original parents of Y not shared with X and not equal to X. This shows the number of parameters can only increase: before we had $k^m + k^n$, after we have $k^{m+(n-l-1)} + k^{n+(m-l)}$.

(As a sanity check on our counting above, if are reversing a single arc without any extra parents, we have $l = 0$, $m = 0$, and $n = 1$; the previous formulas say we will have $m' = 0$ and $n' = 1$ afterwards, which is correct.)

b. For the number of parameters to remain constant, assuming that $k > 1$, requires by our previous calculation that $m - l = 0$ and $n - l - 1 = 0$. This holds exactly when X and Y share all their parents originally (except Y also has X as a parent).

c. For clarity, denote by $P'(Y|U,V,W)$ and $P'(X|U,V,W,Y)$ the new CPTs, and note that the set of variables $V \cup W$ does not include X. It suffices to show that

$$P'(X,Y|U,V,W) = P(X,Y|U,V,W)$$

To see this, let D denote the variables, outside of $\{X,Y\} \cup U \cup V \cup W$, which have either X or Y as ancestor in the original network, and $\overline{D}$ those which don't. Since the arc reversed graph only adds or removes arcs incoming to X or Y, it cannot change which variables lie in D or $\overline{D}$. We then have

$$\begin{aligned} P(D,\overline{D},X,Y,U,V,W) &= P(\overline{D},U,V,W)P(X,Y|U,V,W)P(D|X,Y,U,V,W) \\ &= P'(\overline{D},U,V,W)P(X,Y|U,V,W)P(D|X,Y,U,V,W) \\ &= P'(\overline{D},U,V,W)P(X,Y|U,V,W)P'(D|X,Y,U,V,W) \\ &= P'(\overline{D},U,V,W)P'(X,Y|U,V,W)P'(D|X,Y,U,V,W) \\ &= P'(D,\overline{D},X,Y,U,V,W) \end{aligned}$$

the second as arc reversal does not change the CPTs of variables in $\overline{D},U,V,W$ contains all its parents, the third as if we condition on X,Y,U,V,W the original and arc-reversed Bayesian networks are the same, and the fourth by hypothesis.

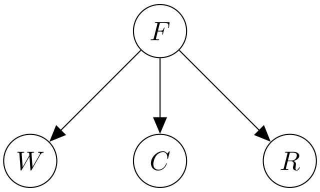


Figure S13.6 A naive Bayes model for fishing.

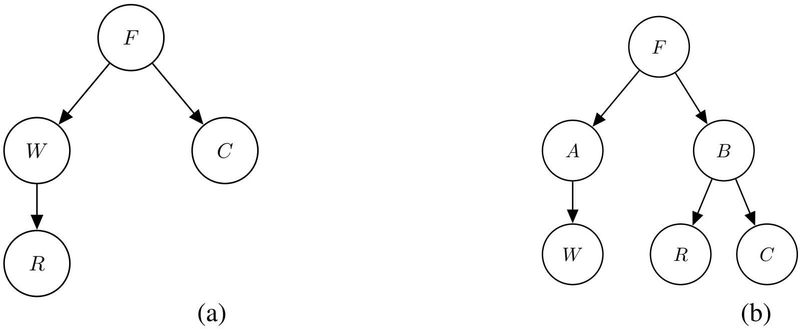


Figure S13.7 Two true distributions for Exercise 13.BFSH.

Then, calculating:

$$\begin{aligned}
&P'(X,Y|U,V,W)\\
&= P'(Y|U,V,W)P'(X|U,V,W,Y)\\
&= \left(\sum_x P(Y|V,W,x)P(x|U,V)\right) P(Y|X,V,W)P(X|U,V)\\
&= \left(\sum_x P(Y|U,V,W,x)P(x|U,V)/P(Y|U,V,W)\right) P(Y|X,V,W)P(X|U,V)\\
&= \left(\sum_x \frac{P(Y,U,V,W,x)P(x,U,V)P(U,V,W)}{P(U,V,W,x)P(U,V)P(Y,U,V,W)}\right) P(Y|X,V,W)P(X|U,V)\\
&= \left(\sum_x P(x|Y,U,V,W)P(x|U,V)/P(x|U,V,W)\right) P(Y|X,V,W)P(X|U,V)\\
&= \left(\sum_x P(x|Y,U,V,W)\right) P(Y|X,V,W)P(X|U,V)\\
&= P(Y|X,V,W)P(X|U,V)
\end{aligned}$$

where the third step follows as V, W, x is the parent set of Y it's conditionally independent of U, and the second to last step follows as U, V is the parent set of X it's conditionally independent of x.

Exercise 13.BFSH

Exercise Exercise 12.FISH introduced the naive Bayes model in Figure S13.6. (F) is true iff today was a good day of fishing. There are three features: whether it rained (R), how many fish were caught (C) with values $\{none, some, lots\}$, and whether it was windy (W).

For each of the true distributions in Figure S13.7, determine whether the Naive Bayes modeling assumption holds and whether the naive Bayes model is guaranteed to be able to represent the true conditional probability, $\mathbf{P}(F \mid W, C, R)$.

a. The naive Bayes assumption does not hold as W is not independent of R given the class F. However, based on the true distribution model, we can conclude that $P(F|W, R, C) = P(F|W, C)$. In other words, R is not required to model the conditional distribution. If we set $P(R|F) = constant$ (same for both values of F), the naive Bayes model will also have $P(F|W, R, C) = P(F|W, C)$ and can exactly model the conditional distribution for F implied by this Bayes net.

b. The naive Bayes assumption does not hold as R is not independent of C given the class F and it cannot represent the true distribution.

Exercise 13.COIN

We have a bag of three biased coins a, b, and c with probabilities of coming up heads of 30%, 60%, and 75%, respectively. One coin is drawn randomly from the bag (with equal likelihood of drawing each of the three coins), and then the coin is flipped three times to generate the outcomes X_1, X_2, and X_3.

a. Draw the Bayesian network corresponding to this setup and define the necessary CPTs.

b. Calculate which coin was most likely to have been drawn from the bag if the observed flips come out heads twice and tails once.

a. With the random variable C denoting which coin $\{a, b, c\}$ we drew, the network has C at the root and X_1, X_2, and X_3 as children.

The CPT for C is:

C	$P(C)$
a	$1/3$
b	$1/3$
c	$1/3$

The CPT for X_i given C are the same, and equal to:

C	X_1	$P(C)$
a	$heads$	0.3
b	$heads$	0.6
c	$heads$	0.75

b. The coin most likely to have been drawn from the bag given this sequence is the value of C with greatest posterior probability $P(C|\text{2 heads, 1 tails})$. Now,

$$\begin{aligned}
&P(C|\text{2 heads, 1 tails})\\
&= P(\text{2 heads, 1 tails}|C)P(C)/P(\text{2 heads, 1 tails})\\
&\propto P(\text{2 heads, 1 tails}|C)P(C)\\
&\propto P(\text{2 heads, 1 tails}|C)
\end{aligned}$$

where in the second line we observe that the constant of proportionality $1/P(\text{2 heads, 1 tails})$ is independent of C, and in the last we observe that $P(C)$ is also independent of the value of C since it is, by hypothesis, equal to $1/3$.

From the Bayesian network we can see that X_1, X_2, and X_3 are conditionally independent given C, so for example

$$\begin{aligned}
&P(X_1 = tails, X_2 = heads, X_3 = heads|C = a)\\
&= P(X_1 = tails|C = a)P(X_2 = heads|C = a)P(X_3 = heads|C = a)\\
&= 0.7 \times 0.3 \times 0.3 = 0.063
\end{aligned}$$

Note that since the CPTs for each coin are the same, we would get the same probability above for any ordering of 2 heads and 1 tails. Since there are three such orderings, we have

$$P(2heads, 1tails|C = a) = 3 \times 0.063 = 0.189.$$

Similar calculations to the above find that

$$\begin{aligned}
P(2heads, 1tails|C = b) &= 0.432\\
P(2heads, 1tails|C = c) &= 0.422
\end{aligned}$$

showing that coin b is most likely to have been drawn.

Alternatively, one could directly compute the value of $P(C|\text{2 heads, 1 tails})$.

Exercise 13.BURG

Consider the Bayesian network in Figure 13.2.

a. If no evidence is observed, are *Burglary* and *Earthquake* independent? Prove this from the numerical semantics and from the topological semantics.

b. If we observe $Alarm = true$, are *Burglary* and *Earthquake* independent? Justify your answer by calculating whether the probabilities involved satisfy the definition of conditional independence.

a. Yes. From the numerical semantics, we have

$$\begin{aligned}
P(B,E) &= \sum_{a,j,m} P(B,E,a,j,m) = \sum_{a,j,m} P(B)P(E)P(a|B,E)P(j|a)P(m|a) \\
&= P(B)P(E)\sum_{a,j,m} P(a|B,E)P(j|a)P(m|a) \\
&= P(B)P(E)\sum_{a} P(a|B,E)\left(\sum_{j} P(j|a)\right)\left(\sum_{m} P(m|a)\right) \\
&= P(B)P(E)
\end{aligned}$$

using the sum-to-1 constraint for the conditional distributions. Topologically, E is independent of its non-descendants (i.e., B) given its parents (i.e., the empty set), so B and E are absolutely independent.

b. We check whether $P(B,E|a) = P(B|a)P(E|a)$. First computing $P(B,E|a)$

$$\begin{aligned}
P(B,E|a) &= \alpha P(a|B,E)P(B,E) \\
&= \alpha \begin{cases} .95 \times 0.001 \times 0.002 & \text{if } B=b \text{ and } E=e \\ .94 \times 0.001 \times 0.998 & \text{if } B=b \text{ and } E=\neg e \\ .29 \times 0.999 \times 0.002 & \text{if } B=\neg b \text{ and } E=e \\ .001 \times 0.999 \times 0.998 & \text{if } B=\neg b \text{ and } E=\neg e \end{cases} \\
&= \alpha \begin{cases} 0.0008 & \text{if } B=b \text{ and } E=e \\ 0.3728 & \text{if } B=b \text{ and } E=\neg e \\ 0.2303 & \text{if } B=\neg b \text{ and } E=e \\ 0.3962 & \text{if } B=\neg b \text{ and } E=\neg e \end{cases}
\end{aligned}$$

where α is a normalization constant. Checking whether $P(b,e|a) = P(b|a)P(e|a)$ we find

$$P(b,e|a) = 0.0008 \neq 0.0863 = 0.3736 \times 0.2311 = P(b|a)P(e|a)$$

showing that B and E are not conditionally independent given A.

Exercise 13.MRBL

Suppose that in a Bayesian network containing an unobserved variable Y, all the variables in the Markov blanket $MB(Y)$ have been observed. Suppose we now remove the node Y from the network. Y's parents lose a child, and Y's children lose a parent. The CPTs for Y's children given their now-reduced sets of parents are reset to *arbitrary* values. Prove that the posterior distribution $Q(X_i|mb(Y))$ for any other unobserved variable X_i in the new network is identical to $P(X_i|mb(Y))$ in the original network.

Let $\mathbf{X}$ be the set of all variables in the original Bayesian network except for Y and $MB(Y)$, with $X_i \in \mathbf{X}$, and let $MB(Y)$ be comprised of $\mathbf{W} = children(Y)$ and the remaining variables $\mathbf{U}$. The key point here is that in the new network defined by Q, the parents of variables in W are all variables in $MB(W)$, so they are all *observed* variables; and the only differences in the CPTS for variables in Q compared to P are for the variables in W.

Hence, we have

$$\begin{aligned}
Q(\mathbf{X}|mb(Y)) &= \alpha Q(\mathbf{X}, mb(Y)) \\
&= = \alpha \prod_i Q(X_i|Pa(X_i)) \prod_j Q(U_j|Pa(U_j)) \prod_k Q(W_k|pa(W_k)) \\
&= = \alpha' \prod_i Q(X_i|Pa(X_i)) \prod_j Q(U_j|Pa(U_j)) \\
&= = \alpha' \prod_i P(X_i|Pa(X_i)) \prod_j P(U_j|Pa(U_j)) = \alpha' P(\mathbf{X}, mb(Y)) = P(\mathbf{X}|mb(Y)) \,.
\end{aligned}$$

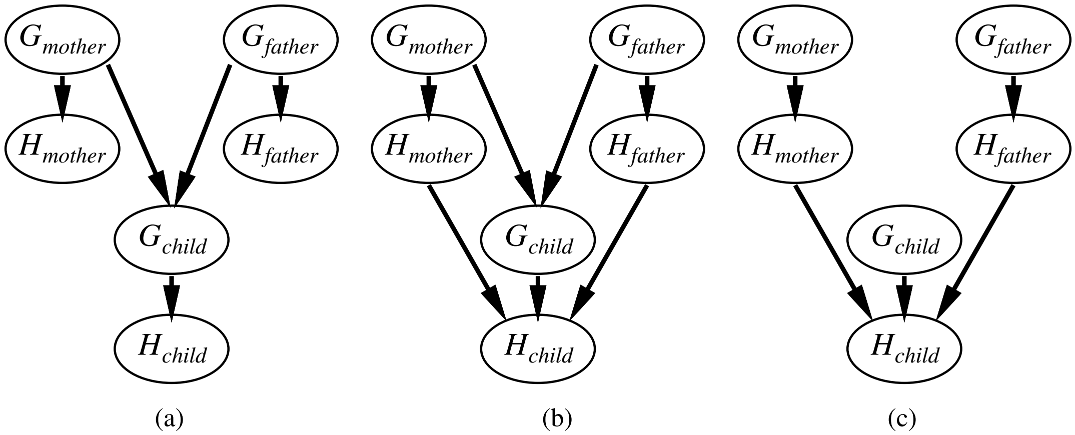


Figure S13.8 Figure for Exercise 13.HAND.

Exercise 13.HAND

Let H_x be a random variable denoting the handedness of an individual x, with possible values l or r. A common hypothesis is that left- or right-handedness is inherited by a simple mechanism; that is, perhaps there is a gene G_x, also with values l or r, and perhaps actual handedness turns out mostly the same (with some probability s) as the gene an individual possesses. Furthermore, perhaps the gene itself is equally likely to be inherited from either of an individual's parents, with a small nonzero probability m of a random mutation flipping the handedness.

a. Which of the three networks in Figure S13.8 claim that $\mathbf{P}(G_{father}, G_{mother}, G_{child}) = \mathbf{P}(G_{father})\mathbf{P}(G_{mother})\mathbf{P}(G_{child})$?

b. Which of the three networks make independence claims that are consistent with the hypothesis about the inheritance of handedness?

c. Which of the three networks is the best description of the hypothesis?

d. Write down the CPT for the G_{child} node in network (a), in terms of s and m.

e. Suppose that $P(G_{father}=l) = P(G_{mother}=l) = q$. In network (a), derive an expression for $P(G_{child}=l)$ in terms of m and q only, by conditioning on its parent nodes.

f. Under conditions of genetic equilibrium, we expect the distribution of genes to be the same across generations. Use this to calculate the value of q, and, given what you know about handedness in humans, explain why the hypothesis described at the beginning of this question must be wrong.

a. (c) matches the equation. The equation describes absolute independence of the three genes, which requires no links among them.

b. (a) and (b). The *assertions* are the *absent* links; the extra links in (b) may be unnecessary but they do not assert an actual dependence. (c) asserts independence of genes which contradicts the inheritance scenario.

c. (a) is best. (b) has spurious links among the H variables, which are not directly causally connected in the scenario described. (In reality, handedness may also be passed down by example/training.)

d. Notice that the $l \rightarrow r$ and $r \rightarrow l$ mutations cancel when the parents have different genes, so we still get 0.5.

G_{mother}	G_{father}	$P(G_{child}=l\mid\ldots)$	$P(G_{child}=r\mid\ldots)$
l	l	$1-m$	m
l	r	0.5	0.5
r	l	0.5	0.5
r	r	m	$1-m$

e. This is a straightforward application of conditioning:

$$\begin{aligned}
P(G_{child}=l) &= \sum_{g_m,g_f} P(G_{child}=l|g_m,g_f)P(g_m,g_f)\\
&= \sum_{g_m,g_f} P(G_{child}=l|g_m,g_f)P(g_m)P(g_f)\\
&= (1-m)q^2 + 0.5q(1-q) + 0.5(1-q)q + m(1-q)^2\\
&= q^2 - mq^2 + q - q^2 + m - 2mq + mq^2\\
&= q + m - 2mq
\end{aligned}$$

f. Equilibrium means that $P(G_{child}=l)$ (the prior, with no parent information) must equal $P(G_{mother}=l)$ and $P(G_{father}=l)$, i.e.,

$$q + m - 2mq = q, \text{ hence } q = 0.5.$$

But few humans are left-handed ($x \approx 0.08$ in fact), so something is wrong with the symmetric model of inheritance and/or manifestation. The "high-school" explanation is that the "right-hand gene is dominant," i.e., preferentially inherited, but current studies suggest also that handedness is not the result of a single gene and may also involve cultural factors. An entire journal (*Laterality*) is devoted to this topic.

Exercise 13.MARB

The **Markov blanket** of a variable is defined on page 437. Prove that a variable is independent of all other variables in the network, given its Markov blanket and derive Equation (13.10) (page 461).

These proofs are tricky for those not accustomed to manipulating probability expressions, and students may require some hints.

a. There are several ways to prove this. Probably the simplest is to work directly from the global semantics. First, we rewrite the required probability in terms of the full joint:

$$\begin{aligned}
P(x_i|x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_n) &= \frac{P(x_1,\ldots,x_n)}{P(x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_n)}\\
&= \frac{P(x_1,\ldots,x_n)}{\sum_{x_i} P(x_1,\ldots,x_n)}\\
&= \frac{\prod_{j=1}^{n} P(x_j|parentsX_j)}{\sum_{x_i}\prod_{j=1}^{n} P(x_j|parentsX_j)}
\end{aligned}$$

Now, all terms in the product in the denominator that do not contain x_i can be moved outside the summation, and then cancel with the corresponding terms in the numerator. This just leaves us with the terms that do mention x_i, i.e., those in which X_i is a child or a parent. Hence, $P(x_i|x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_n)$ is equal to

$$\frac{P(x_i|parentsX_i)\prod_{Y_j\in Children(X_i)} P(y_j|parents(Y_j))}{\sum_{x_i} P(x_i|parentsX_i)\prod_{Y_j\in Children(X_i)} P(y_j|parents(Y_j))}$$

Now, by reversing the argument in part (b), we obtain the desired result.

b. This is a relatively straightforward application of Bayes' rule. Let $\mathbf{Y}=Y_1,\ldots,y_\ell$ be the children of X_i and let $\mathbf{Z}_j$ be the parents of Y_j other than X_i. Then we have

$$\begin{aligned}
&\mathbf{P}(X_i|MB(X_i))\\
&= \mathbf{P}(X_i|Parents(X_i),\mathbf{Y},\mathbf{Z}_1,\ldots,\mathbf{Z}_\ell)\\
&= \alpha\mathbf{P}(X_i|Parents(X_i),\mathbf{Z}_1,\ldots,\mathbf{Z}_\ell)\mathbf{P}(\mathbf{Y}|Parents(X_i),X_i,\mathbf{Z}_1,\ldots,\mathbf{Z}_\ell)\\
&= \alpha\mathbf{P}(X_i|Parents(X_i))\mathbf{P}(\mathbf{Y}|X_i,\mathbf{Z}_1,\ldots,\mathbf{Z}_\ell)\\
&= \alpha\mathbf{P}(X_i|Parents(X_i))\prod_{Y_j\in Children(X_i)} P(Y_j|Parents(Y_j))
\end{aligned}$$

where the derivation of the third line from the second relies on the fact that a node is independent of its nondescendants given its parents.

Exercise 13.DCAR

Consider the network for car diagnosis shown in Figure S13.9.

a. Extend the network with the Boolean variables *IcyWeather* and *StarterMotor*.

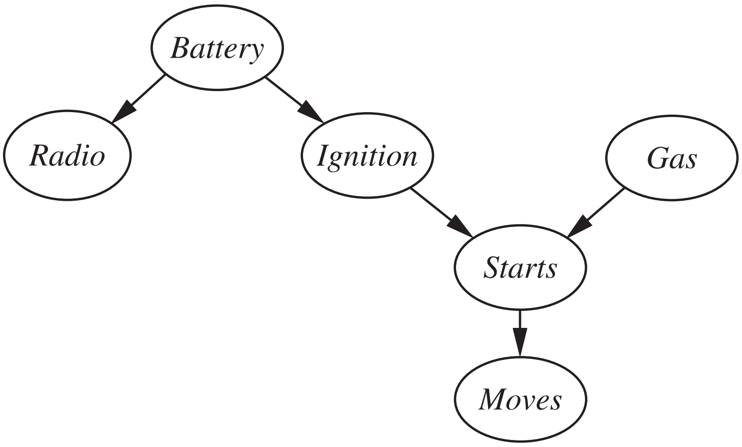


Figure S13.9 Car network for Exercise 13.DCAR.

b. Give reasonable conditional probability tables for all the nodes.

c. How many independent values are contained in the joint probability distribution for eight Boolean nodes, assuming that no conditional independence relations are known to hold among them?

d. How many independent probability values do your network tables contain?

e. The conditional distribution for $Starts$ could be described as a **noisy-AND** distribution. Define this family in general and relate it to the noisy-OR distribution.

Adding variables to an existing net can be done in two ways. Formally speaking, one should insert the variables into the variable ordering and rerun the network construction process from the point where the first new variable appears. Informally speaking, one never really builds a network by a strict ordering. Instead, one asks what variables are direct causes or influences on what other ones, and builds local parent/child graphs that way. It is usually easy to identify where in such a structure the new variable goes, but one must be very careful to check for possible induced dependencies downstream.

a. $IcyWeather$ is not caused by any of the car-related variables, so needs no parents. It directly affects the battery and the starter motor. $StarterMotor$ is an additional precondition for $Starts$. The new network is shown in Figure S13.10.

b. Reasonable probabilities may vary a lot depending on the kind of car and perhaps the personal experience of the assessor. The following values indicate the general order of magnitude and relative values that make sense:

- A reasonable prior for IcyWeather might be 0.05 (perhaps depending on location and season).
- $P(Battery|IcyWeather) = 0.95$, $P(Battery|\neg IcyWeather) = 0.997$.
- $P(StarterMotor|IcyWeather) = 0.98$, $P(Battery|\neg IcyWeather) = 0.999$.

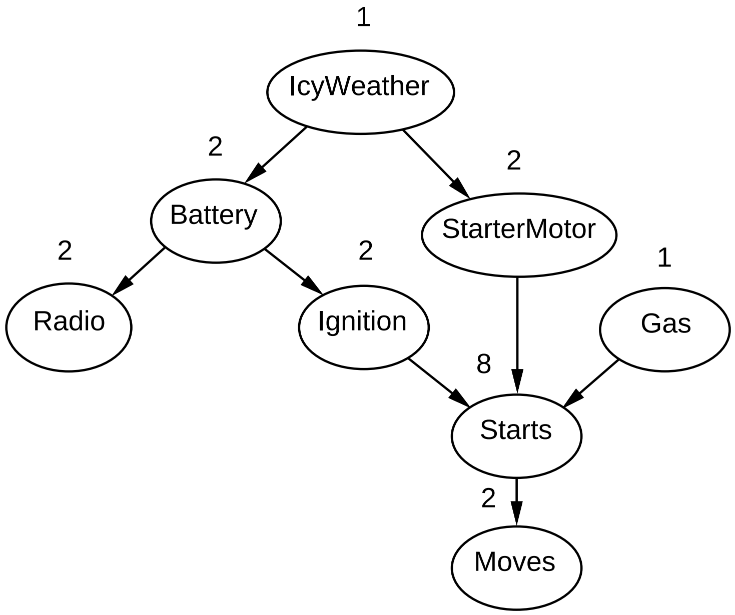


Figure S13.10 Car network amended to include $IcyWeather$ and $StarterMotorWorking$ (SMW).

- $P(Radio|Battery) = 0.9999$, $P(Radio|\neg Battery) = 0.05$.
- $P(Ignition|Battery) = 0.998$, $P(Ignition|\neg Battery) = 0.01$.
- $P(Gas) = 0.995$.
- $P(Starts|Ignition, StarterMotor, Gas) = 0.9999$, other entries 0.0.
- $P(Moves|Starts) = 0.998$.

c. With 8 Boolean variables, the joint has $2^8 - 1$ = 255 independent entries.

d. Given the topology shown in Figure S13.10, the total number of independent CPT entries is 1+2+2+2+2+1+8+2= 20.

e. The CPT for $Starts$ describes a set of nearly necessary conditions that are together almost sufficient. That is, all the entries are nearly zero except for the entry where all the conditions are true. That entry will be not quite 1 (because there is always some other possible fault that we didn't think of), but as we add more conditions it gets closer to 1. If we add a $Leak$ node as an extra parent, then the probability is exactly 1 when all parents are true. We can relate noisy-AND to noisy-OR using de Morgan's rule: $A \wedge B \equiv \neg(\neg A \vee \neg B)$. That is, noisy-AND is the same as noisy-OR except that the polarities of the parent and child variables are reversed. In the noisy-OR case, we have

$$P(Y = true|x_1, \ldots, x_k) = 1 - \prod_{\{i:x_i = true\}} q_i$$

where q_i is the probability that the *presence* of the ith parent *fails* to cause the child to be *true*. In the noisy-AND case, we can write

$$P(Y = true | x_1, \ldots, x_k) = \prod_{\{i : x_i = false\}} r_i$$

where r_i is the probability that the *absence* of the ith parent *fails* to cause the child to be *false* (e.g., it is magically bypassed by some other mechanism).

Exercise 13.FLUN

Consider a simple Bayesian network with root variables $Cold$, Flu, and $Malaria$ and child variable $Fever$, with a noisy-OR conditional distribution for $Fever$ as described in Section 13.2.2. By adding appropriate auxiliary variables for inhibition events and fever-inducing events, construct an equivalent Bayesian network whose CPTs (except for root variables) are deterministic. Define the CPTs and prove equivalence.

Recall that the noisy-OR CPT for a Boolean variable X with Boolean parents $U_1, \ldots, U_k$ is given by

$$P(X = true \,|\, u_1, \ldots, u_k)) = 1 - \prod_{\{j : u_j = true\}} q_j \; .$$

We would like to construct a deterministic CPT for X with additional parent variables $\mathbf{Z}$, such that the original conditional distribution is recovered when those variables are summed out:

$$\sum_{\mathbf{z}} P(X = true \,|\, u_1, \ldots, u_k, \mathbf{z})) = 1 - \prod_{\{j : u_j = true\}} q_j \; .$$

Let Z_j be a Boolean inhibitory variable for parent U_j; it will be a root variable with prior probability q_j. The basic idea is that the effect occurs if any cause occurs and the cause is not inhibited; equivalently, the effect does not occur if all the occuring causes are inhibited. So the the required CPT implements the following deterministic relationship:

$$X = (U_1 \wedge \neg Z_1) \vee \cdots \vee (U_k \wedge \neg Z_k) \; .$$

Since the inhibition variables are independent, the probability of all occurring causes being inhibited is exactly $\prod_{\{j : u_j = true\}} q_j$, as required.

Exercise 13.LGEX

Consider the family of linear Gaussian networks, as defined on page 440.

a. In a two-variable network, let X_1 be the parent of X_2, let X_1 have a Gaussian prior, and let $\mathbf{P}(X_2 \,|\, X_1)$ be a linear Gaussian distribution. Show that the joint distribution $P(X_1, X_2)$ is a multivariate Gaussian, and calculate its covariance matrix.

b. Prove by induction that the joint distribution for a general linear Gaussian network on $X_1, \ldots, X_n$ is also a multivariate Gaussian.

This exercise is a little tricky and will appeal to more mathematically oriented students.

a. The basic idea is to multiply the two densities, match the result to the standard form for a multivariate Gaussian, and hence identify the entries in the inverse covariance matrix. Let's begin by looking at the multivariate Gaussian. From page 1000 in Appendix A we have

$$P(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^n|\mathbf{\Sigma}|}} e^{-\frac{1}{2}\left((\mathbf{x}-\mu)^\top \mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)\right)} ,$$

where μ is the mean vector and $\mathbf{\Sigma}$ is the covariance matrix. In our case, **x** is $(x_1\ x_2)^\top$ and let the (as yet) unknown μ be $(m_1\ m_2)^\top$. Suppose the inverse covariance matrix is

$$\mathbf{\Sigma}^{-1} = \begin{pmatrix} c & d \\ d & e \end{pmatrix}$$

Then, if we multiply out the exponent, we obtain

$$\begin{aligned} &-\tfrac{1}{2}\left((\mathbf{x}-\mu)^\top \mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)\right) = \\ &\qquad -\tfrac{1}{2} \cdot c(x_1-m_1)^2 + 2d(x_1-m_1)(x_2-m_2) + f(x_2-m_2)^2 \end{aligned}$$

Looking at the distributions themselves, we have

$$P(x_1) = \frac{1}{\sigma_1\sqrt{2\pi}} e^{-(x_1-\mu_1)^2/(2\sigma_1^2)}$$

and

$$P(x_2|x_1) = \frac{1}{\sigma_2\sqrt{2\pi}} e^{-(x_2-(ax_1+b))^2/(2\sigma_2^2)}$$

hence

$$P(x_1, x_2) = \frac{1}{\sigma_1\sigma_2(2\pi)} e^{-(\sigma_2^2(x_2-(ax_1+b))^2+\sigma_1^2(x_2-(ax_1+b))^2)/(2\sigma_1^2\sigma_2^2)}$$

We can obtain equations for c, d, and e by picking out the coefficients of x_1^2, x_1x_2, and x_2^2:

$$\begin{aligned} c &= (\sigma_2^2 + a^2\sigma_1^2)/\sigma_1^2\sigma_2^2 \\ 2d &= -2a/\sigma_2^2 \\ e &= 1/\sigma_2^2 \end{aligned}$$

We can check these by comparing the normalizing constants.

$$\frac{1}{\sigma_1\sigma_2(2\pi)} = \frac{1}{\sqrt{(2\pi)^n|\mathbf{\Sigma}|}} = \frac{1}{(2\pi)\sqrt{1/|\mathbf{\Sigma}^{-1}|}} = \frac{1}{(2\pi)\sqrt{1/(ce-d^2)}}$$

from which we obtain the constraint

$$ce - d^2 = 1/\sigma_1^2\sigma_2^2$$

which is easily confirmed. Similar calculations yield m_1 and m_2, and plugging the results back shows that $P(x_1, x_2)$ is indeed multivariate Gaussian. The covariance matrix is

$$\mathbf{\Sigma} = \begin{pmatrix} c & d \\ d & e \end{pmatrix}^{-1} = \frac{1}{ce-d^2}\begin{pmatrix} e & -d \\ -d & c \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & a\sigma_1^2 \\ a\sigma_1^2 & \sigma_2^2 + a^2\sigma_1^2 \end{pmatrix}$$

b. The induction is on n, the number of variables. The base case for $n=1$ is trivial. The inductive step asks us to show that if any $P(x_1, \ldots, x_n)$ constructed with linear–Gaussian conditional densities is multivariate Gaussian, then any $P(x_1, \ldots, x_n, x_{n+1})$ constructed with linear–Gaussian conditional densities is also multivariate Gaussian. Without loss of generality, we can assume that X_{n+1} is a leaf variable added to a network defined in the first n variables. By the product rule we have

$$\begin{aligned} P(x_1, \ldots, x_n, x_{n+1}) &= P(x_{n+1}|x_1, \ldots, x_n)P(x_1, \ldots, x_n) \\ &= P(x_{n+1}|parents(X_{n+1}))P(x_1, \ldots, x_n) \end{aligned}$$

which, by the inductive hypothesis, is the product of a linear Gaussian with a multivariate Gaussian. Extending the argument of part (a), this is in turn a multivariate Gaussian of one higher dimension.

Exercise 13.PROD

The probit distribution defined on page 442 describes the probability distribution for a Boolean child, given a single continuous parent.

a. How might the definition be extended to cover multiple continuous parents?

b. How might it be extended to handle a *multivalued* child variable? Consider both cases where the child's values are ordered (as in selecting a gear while driving, depending on speed, slope, desired acceleration, etc.) and cases where they are unordered (as in selecting bus, train, or car to get to work). (*Hint*: Consider ways to divide the possible values into two sets, to mimic a Boolean variable.)

a. With multiple continuous parents, we must find a way to map the parent value vector to a single threshold value. The simplest way to do this is to take a linear combination of the parent values.

b. For ordered values $y_1 < y_2 < \cdots < y_d$, we assume some unobserved continuous dependent variable y^* that is normally distributed conditioned on the parent variables, and define cutpoints c_j such that $Y = y_j$ iff $c_{j-1} \leq y^* \leq c_j$. The probability of this event is given by subtracting the cumulative distributions at the adjacent cutpoints.

The unordered case is not obviously meaningful if we insist that the relationship between parents and child be mediated by a single, real-valued, normally distributed variable.

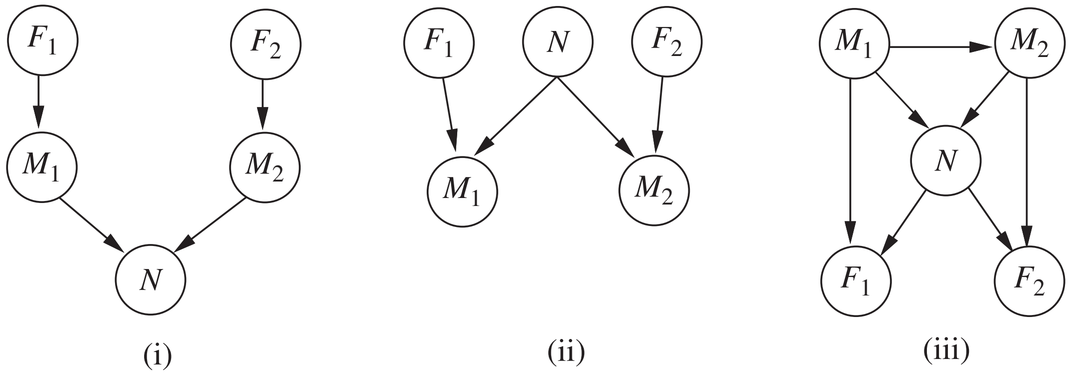


Figure S13.11 A Bayesian network for Exercise 13.TELC.

Exercise 13.TELE

Two astronomers in different parts of the world make measurements M_1 and M_2 of the number of stars N in some small region of the sky, using their telescopes. Normally, there is a small possibility e of error by up to one star in each direction. Each telescope can also (with a much smaller probability f) be badly out of focus (events F_1 and F_2), in which case the scientist will undercount by three or more stars (or if N is less than 3, fail to detect any stars at all). Consider the three networks shown in Figure S13.11.

a. Which of these Bayesian networks are correct (but not necessarily efficient) representations of the preceding information?

b. Which is the best network? Explain.

c. Write out a conditional distribution for $\mathbf{P}(M_1 \mid N)$, for the case where $N \in \{1, 2, 3\}$ and $M_1 \in \{0, 1, 2, 3, 4\}$. Each entry in the conditional distribution should be expressed as a function of the parameters e and/or f.

d. Suppose $M_1 = 1$ and $M_2 = 3$. What are the *possible* numbers of stars if you assume no prior constraint on the values of N?

e. What is the *most likely* number of stars, given these observations? Explain how to compute this, or if it is not possible to compute, explain what additional information is needed and how it would affect the result.

a. Although (i) in some sense depicts the "flow of information" during calculation, it is clearly incorrect as a network, since it says that given the measurements M_1 and M_2, the number of stars is independent of the focus. (ii) correctly represents the causal structure: each measurement is influenced by the actual number of stars and the focus, and the two telescopes are independent of each other. (iii) shows a correct but more complicated network—the one obtained by ordering the nodes M_1, M_2, N, F_1, F_2. If you order M_2 before M_1 you would get the same network except with the arrow from M_1 to M_2 reversed.

b. (ii) requires fewer parameters and is therefore better than (iii).

c. To compute $\mathbf{P}(M_1|N)$, we will need to condition on F_1 (that is, consider both possible cases for F_1, weighted by their probabilities).

$$\begin{aligned}\mathbf{P}(M_1|N) &= \mathbf{P}(M_1|N,F_1)\mathbf{P}(F_1|N) + \mathbf{P}(M_1|N,\neg F_1)\mathbf{P}(\neg F_1|N)\\ &= \mathbf{P}(M_1|N,F_1)\mathbf{P}(F_1) + \mathbf{P}(M_1|N,\neg F_1)\mathbf{P}(\neg F_1)\end{aligned}$$

Let f be the probability that the telescope is out of focus. The exercise states that this will cause an "undercount of three or more stars," but if N = 3 or less the count will be 0 if the telescope is out of focus. If it is in focus, then we will assume there is a probability of e of counting one too few, and e of counting one too many. The rest of the time $(1-2e)$, the count will be accurate. Then the table is as follows:

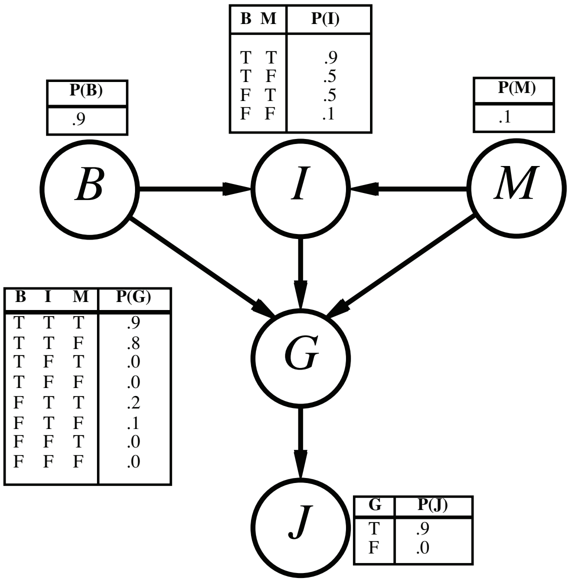
	$N=1$	$N=2$	$N=3$
$M_1=0$	f + e(1-f)	f	f
$M_1=1$	(1-2e)(1-f)	e(1-f)	0.0
$M_1=2$	e(1-f)	(1-2e)(1-f)	e(1-f)
$M_1=3$	0.0	e(1-f)	(1-2e)(1-f)
$M_1=4$	0.0	0.0	e(1-f)

Notice that each column has to add up to 1. Reasonable values for e and f might be 0.05 and 0.002.

d. This question causes a surprising amount of difficulty, so it is important to make sure students understand the reasoning behind an answer. One approach uses the fact that it is easy to reason in the forward direction, that is, try each possible number of stars N and see whether measurements $M_1=1$ and $M_2=3$ are possible. (This is a sort of mental simulation of the physical process.) An alternative approach is to enumerate the possible focus states and deduce the value of N for each. Either way, the solutions are $N=2$, 4, or ≥ 6.

e. We cannot calculate the most likely number of stars without knowing the prior distribution $P(N)$. Let the priors be p_2, p_4, and $p_{\geq 6}$. The posterior for $N=2$ is $p_2e^2(1-f)^2$; for $N=4$ it is at most p_4ef (at most, because with $N=4$ the out-of-focus telescope could measure 0 instead of 1); for $N\geq 6$ it is at most $p_{\geq 6}f^2$. If we assume that the priors are roughly comparable, then $N=2$ is most likely because we are told that f is much smaller than e.

For follow-up or alternate questions, it is easy to come up with endless variations on the same theme involving sensors, failure nodes, hidden state. One can also add in complex mechanisms, as for the $Starts$ variable in Exercise 13.DCAR.



P(B)
.9

B	M	P(I)
T	T	.9
T	F	.5
F	T	.5
F	F	.1

P(M)
.1

B	I	M	P(G)
T	T	T	.9
T	T	F	.8
T	F	T	.0
T	F	F	.0
F	T	T	.2
F	T	F	.1
F	F	T	.0
F	F	F	.0

G	P(J)
T	.9
F	.0

Figure S13.12 A Bayesian network for Exercise 13.BATF.

Exercise 13.BATF

Consider the Bayes net shown in Figure S13.12.

a. Which of the following are asserted by the network *structure*?

(i) $\mathbf{P}(B, I, M) = \mathbf{P}(B)\mathbf{P}(I)\mathbf{P}(M)$.

(ii) $\mathbf{P}(J \mid G) = \mathbf{P}(J \mid G, I)$.

(iii) $\mathbf{P}(M \mid G, B, I) = \mathbf{P}(M \mid G, B, I, J)$.

b. Calculate the value of $P(b, i, m, \neg g, j)$.

c. Calculate the value of $P(b, i, \neg m, g, j)$.

d. Calculate the probability that someone goes to jail given that they broke the law, have been indicted, and face a politically motivated prosecutor.

e. A **context-specific independence** (see page 438) allows a variable to be independent of some of its parents given certain values of others. In addition to the usual conditional independences given by the graph structure, what context-specific independences exist in the Bayes net in Figure S13.12?

f. Suppose we want to add the variable $P = PresidentialPardon$ to the network; draw the new network and briefly explain any links you add.

a. The network asserts (ii) and (iii). (For (iii), consider the Markov blanket of M.)

b. $P(b, i, m, \neg g, j) = P(b)P(m)P(i|b, m)P(\neg g|b, i, m)P(j|\neg g)$
$= .9 \times .1 \times .9 \times .1 \times 0 = 0.$
It would be reasonable to avoid the unnecessary work and point out that the model

precludes going to jail if found not guilty so the event must have probability 0.

c. $P(b, i, \neg m, g, j) = P(b)P(\neg m)P(i|b, \neg m)P(g|b, i, \neg m)P(j|g)$
$= .9 \times .9 \times .5 \times .8 \times .9 = .2916$

d. Since B, I, M are fixed true in the evidence, we can treat G as having a prior of 0.9 and just look at the submodel with G, J:
$\mathbf{P}(J|b, i, m) = \alpha \sum_g \mathbf{P}(J, g) = \alpha[\mathbf{P}(J, g) + \mathbf{P}(J, \neg g)]$
$= \alpha[\langle P(j, g), P(\neg j, g)\rangle + \langle P(j, \neg g), P(\neg j, \neg g)\rangle$
$= \alpha[\langle .81, .09\rangle + \langle 0, 0.1\rangle] = \langle .81, .19\rangle$
That is, the probability of going to jail is 0.81.

e. Intuitively, a person cannot be found guilty if not indicted, regardless of whether they broke the law and regardless of the prosecutor. This is what the CPT for G says; so G is context-specifically independent of B and M given $I = false$.

f. A pardon is unnecessary if the person is not indicted or not found guilty; so I and G are parents of P. One could also add B and M as parents of P, since a pardon is more likely if the person is actually innocent and if the prosecutor is politically motivated. (There are other causes of $Pardon$, such as $LargeDonationToPresidentsParty$, but such variables are not currently in the model.) The pardon (presumably) is a get-out-of-jail-free card, so P is a parent of J.

Exercise 13.BNTF

Which of the following are true, and which are false?

a. Bayes nets are organized into layers with connections only between adjacent layers.

b. The topology of a Bayes net can assert that some variables are *not* conditionally independent.

c. Some Bayes nets require as many parameters as the explicitly tabulated full joint distribution.

d. In any Bayes net, the parents of a single child are always conditionally independent of each other given the child.

a. False. Any connections are allowed as long as the network remains acyclic.

b. False. The topology asserts conditional *indpendence* relationships, but cannot assert *dependence*. Even if a link exists, the associated CPT can still deny that there is any dependence.

c. True. This is the case when every variable has all of its predecessors as parents.

d. False. The parents will usually be conditionally *dependent* given the child, as one cause explains away another.

Exercise 13.BNCI

In the Bayes net in Figure S13.13, state whether each of the following assertions is necessarily true, necessarily false, or undetermined.

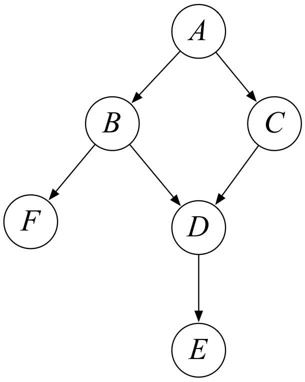


Figure S13.13 A Bayesian network for Exercise 13.BNCI.

a. A is absolutely independent of E.
b. B is conditionally independent of C given A.
c. F is conditionally independent of C given A.
d. B is conditionally independent of C given A and E.

Remember that the topology of a Bayes net cannot assert that a given independence does not hold.

a. Undetermined.
b. True.
c. True.
d. Undetermined.

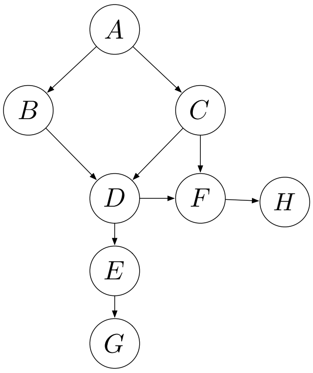


Figure S13.14 A Bayesian network for Exercise 13.CIBN.

Exercise 13.CIBN

In the Bayes net in Figure S13.14, state whether each of the following assertions is necessarily true, necessarily false, or undetermined.

a. B is absolutely independent of C.

b. B is conditionally independent of C given G.

c. B is conditionally independent of C given H.

d. A is conditionally independent of D given G.

e. A is conditionally independent of D given H.

f. B is conditionally independent of C given A, F.

g. F is conditionally independent of B given D, A.

h. F is conditionally independent of B given D, C.

Remember that the topology of a Bayes net cannot assert that a given independence does not hold.

a. Undetermined.

b. Undetermined.

c. Undetermined.

d. Undetermined.

e. Undetermined.

f. Undetermined.

g. Undetermined.

h. True.

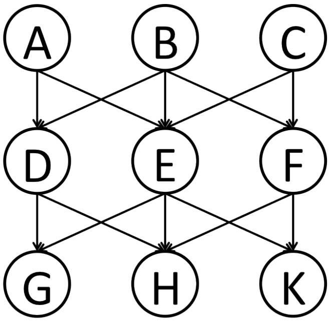


Figure S13.15 A Bayesian network for Exercise 13.TFBN.

Exercise 13.TFBN

In the Bayes net in Figure S13.15, state whether each of the following assertions is necessarily true, necessarily false, or undetermined.

a. A is absolutely independent of C.

b. A is conditionally independent of C given E.

c. A is conditionally independent of C given G.

d. A is absolutely independent of K.

e. A is conditionally independent of G given D, E, F.

f. A is conditionally independent of B given D, E, F.

g. A is conditionally independent of C given D, F, K.

h. A is conditionally independent of G given D.

Remember that the topology of a Bayes net cannot assert that a given independence does not hold.

a. True.

b. Undetermined.

c. Undetermined.

d. Undetermined.

e. True.

f. Undetermined.

g. Undetermined.

h. Undetermined.

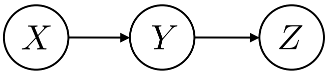


Figure S13.16 A Bayesian network for Exercise 13.BNNT.

Exercise 13.BNNT

In the Bayes net in Figure S13.16, which of the following are *necessarily* true?

a. $P(X, Y, Z) = P(X)P(Y|X)P(Z|X, Y)$.

b. $P(X, Y, Z) = P(X)P(Y|X)P(Z|Y)$.

c. $P(X, Y, Z) = P(X)P(Y|Z)P(Z|X)$.

d. $P(X, Y, Z) = P(Z)P(Y|Z)P(X|Y, Z)$.

e. $P(X, Y, Z) = P(Z)P(Y|Z)P(X|Y)$.

a. True.

b. True.

c. False.

d. True.

e. True.

Exercise 13.PRRV

Let P be a probability distribution over random variables A, B, C. Let Q be another probability distribution over the same variables, defined by a Bayes net in which B and C are conditionally independent given A. We are given that $Q(A) = P(A)$, $Q(B|A) = P(B|A)$, and $Q(C|A) = P(C|A)$. Which if the following can be deduced? Explain.

a. $Q(B) = P(B)$.

b. $Q(C) = P(C)$.

c. $Q(A, B, C) = P(A, B, C)$.

a. True. $Q(B) = \sum_a Q(B|a)Q(a) = \sum_a P(B|a)P(a) = P(B)$.

b. True, by the same argument.

c. False. In P, it is not necessarily the case that B and C are conditionally independent given A. We have

$$P(A, B, C) = P(A)P(B|A)P(C|A, B) = Q(A)Q(B|A)P(C|A, B)$$

and

$$Q(A, B, C) = Q(A)Q(B|A)Q(C|A) = Q(A)Q(B|A)P(C|A)$$

but we cannot establish that $P(C|A, B) = P(C|A)$.

Exercise 13.CDBV

You are given the following conditional distributions that connect the binary variables W, X, Y, Z:

X	$P(X)$
0	0.75
1	0.25

X	W	$P(W\|X)$
0	0	0.4
0	1	0.6
1	0	0.4
1	1	0.6

X	Y	$P(Y\|X)$
0	0	0.3
0	1	0.7
1	0	0.1
1	1	0.9

Z	W	$P(W\|Z)$
0	0	0.2
0	1	0.8
1	0	0.8
1	1	0.2

Which of the Bayes nets in Figure S13.17 can represent a joint distribution that is consistent with these conditional distributions? Of those, which are minimal, in the sense that no edge can be removed while retaining the property?

Remember that the topology of a Bayes net makes claims of conditional independence through the absence of arrows. The first three networks assert that W and Z are absolutely

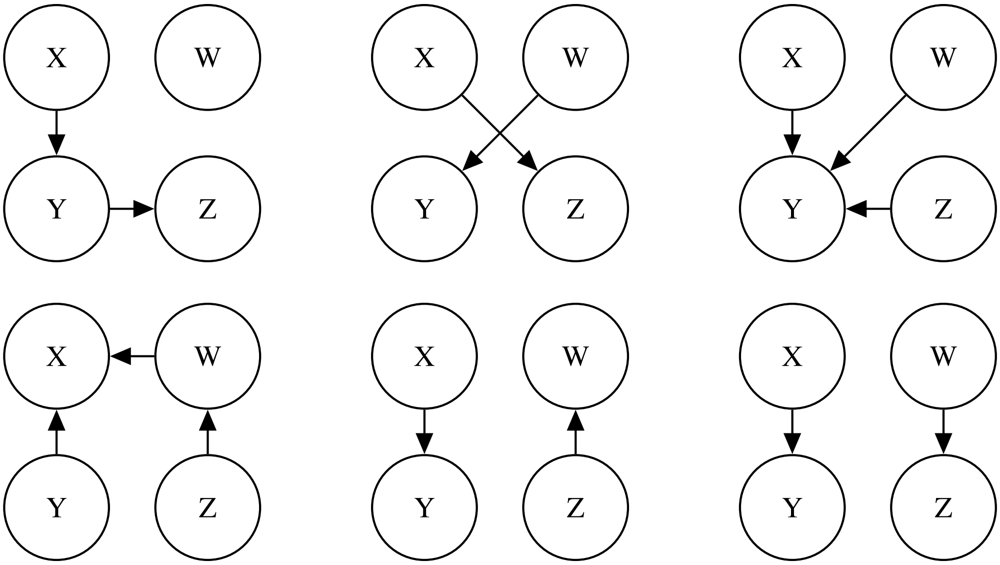


Figure S13.17 Bayes nets for Exercise 13.CDBV

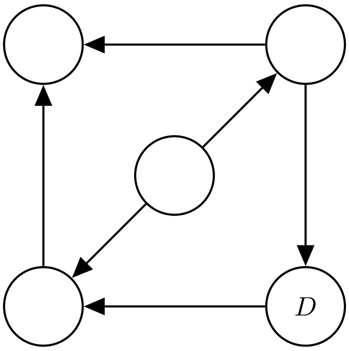


Figure S13.18 Bayes net for Exercise 13.ABCE.

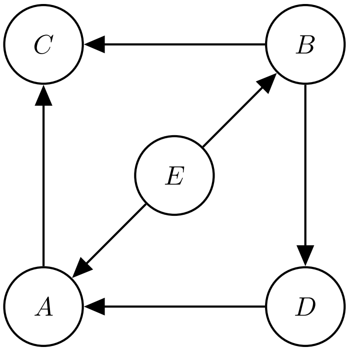
independent, which is not the case. Examining the distribution $P(W|X)$, we see that the probability of W does not actually depend on X, so W and X are absolutely independent. The only dependencies that are required are between X and Y and between W and Z. The three networks in the second row capture these dependencies. (The direction of the arrow doesn't matter.) The fourth network has a superfluous arrow from W to X, so the fifth and sixth networks are minimal.

Exercise 13.ABCE

Label the blank nodes in the Bayes net below with the variables $\{A, B, C, E\}$ such that the following independence assertions are true:

- A is conditionally independent of B given D, E;
- E is conditionally independent of D given B;
- E is conditionally independent of C given A, B;
- C is conditionally independent of D given A, B;

There is only one solution, as follows:



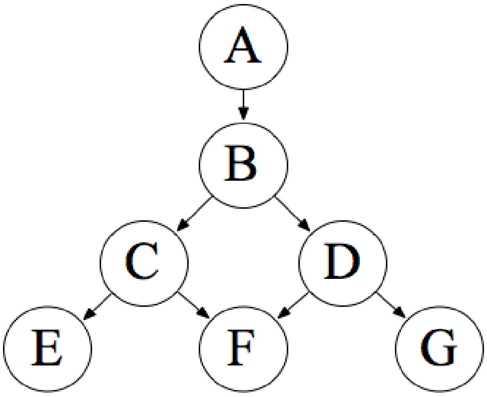


Figure S13.19 Figure for Exercise 13.DESP.

Exercise 13.DSEP

a. Consider the Bayes net in Figure S13.19.

(i) Given B, what variable(s) is E guaranteed to be independent of?

(ii) Given B and F, what variable(s) is G guaranteed to be independent of?

b. Now we'd like to formulate d-separation as a search problem. Specifically, you're given a variable X and a variable Y, a Bayes net G, and a set of observed variables E. You're also given E^+, which is the set of variables that are the parents or ancestors of evidence variables. Given this information, define a search problem that finds Y if X and Y are not d-separated, and does not find a goal otherwise. You may find the notation $W \to U \in G$ meaning "an arc from W to U is in the Bayes net" helpful. A full credit solution will have a minimal state space.

c. Give a non-trivial consistent heuristic for this problem.

a. (i) Given B, E is guaranteed to be independent of A, D, G.

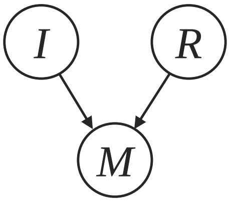
(ii) Given B and F, G is guaranteed to be independent of A.

b.
- **State Space:** (n, d) where n is a node and $d \in \{$in,out,either$\}$.
- **Initial State:** $(X,$ either$)$.
- **Successor Function:** The successor function extends the current state using active triples (from the Bayes' ball algorithm). Every node and edge direction that can be appended to the current state to yield a valid active triple in the graph is returned. The direction 'either' can be interpreted as either 'in' or 'out' for extension using an active triple.
- **Goal Test:** Is equal to $(Y,$ in$)$ or $(Y$out$)$.

c. The length of the shortest path from the node n in the current state to Y, where edge direction is ignored.

I	*P(I)*
+*i*	0.8
-*i*	0.2

R	*P(R)*
+*r*	0.4
-*r*	0.6



I	*R*	*M*	*P(M|I,R)*
+*i*	+*r*	+*m*	0
+*i*	+*r*	-*m*	1.0
+*i*	-*r*	+*m*	0
+*i*	-*r*	-*m*	1.0
-*i*	+*r*	+*m*	0.7
-*i*	+*r*	-*m*	0.3
-*i*	-*r*	+*m*	0.2
-*i*	-*r*	-*m*	0.8

Figure S13.20 Figure for Exercise 13.MUMP.

Exercise 13.MUMP

There has been an outbreak of mumps in your college. You feel fine, but you're worried that you might already be infected. You decide to use Bayes nets to analyze the probability that you've contracted the mumps.

You first think about the following two factors:

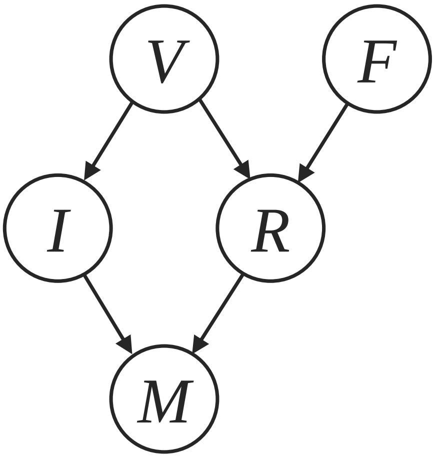


Figure S13.21 Figure for Exercise 13.MUMP.

- You think you have immunity from the mumps ($+i$) due to being vaccinated recently, but the vaccine is not completely effective, so you might not be immune ($-i$).
- Your roommate didn't feel well yesterday, and though you aren't sure yet, you suspect they might have the mumps ($+r$).

Denote these random variables by I and R. Let the random variable M take the value $+m$ if you have the mumps, and $-m$ if you do not. You write down the Bayes net in Figure S13.20 to describe your chances of being sick:

a. Fill in the following table with the joint distribution over I, M, and R, $P(I, M, R)$.

I	R	M	$P(I, R, M)$
$+i$	$+r$	$+m$	0
$+i$	$+r$	$-m$	
$+i$	$-r$	$+m$	0
$+i$	$-r$	$-m$	
$-i$	$+r$	$+m$	0.056
$-i$	$+r$	$-m$	0.024
$-i$	$-r$	$+m$	0.024
$-i$	$-r$	$-m$	0.096

b. What is the marginal probability $P(+m)$ that you have the mumps?

c. Assuming you do have the mumps, you're concerned that your roommate may have the disease as well. What is the probability $P(+r \mid +m)$ that your roommate has the mumps given that you have the mumps? Note that you still don't know whether or not you have immunity.

You're still not sure if you have enough information about your chances of having the mumps, so you decide to include two new variables in the Bayes net. Your roommate went to a party over the weekend, and there's some chance another person at the party had the mumps ($+f$). Furthermore, both you and your roommate were vaccinated at a clinic that reported a vaccine mix-up. Whether or not you got the right vaccine ($+v$ or $-v$) has ramifications for both your immunity (I) and the probability that your roommate has since contracted the disease (R). Accounting for these, you draw the modified Bayes net shown in Figure S13.21:

d. Which of the following statements are *guaranteed* to be true for this Bayes net?

(i) $V \perp\!\!\!\perp M \mid I,\ R$
(ii) $V \perp\!\!\!\perp M \mid R$
(iii) $M \perp\!\!\!\perp F \mid R$
(iv) $V \perp\!\!\!\perp F$
(v) $V \perp\!\!\!\perp F \mid M$
(vi) $V \perp\!\!\!\perp F \mid I$

a.

I	R	M	$P(I,R,M)$
$+i$	$+r$	$+m$	0
$+i$	$+r$	$-m$	**0.32**
$+i$	$-r$	$+m$	0
$+i$	$-r$	$-m$	**0.48**
$-i$	$+r$	$+m$	0.056
$-i$	$+r$	$-m$	0.024
$-i$	$-r$	$+m$	0.024
$-i$	$-r$	$-m$	0.096

b.

$$\begin{aligned} P(+m) &= \sum_{i,r} P(i,r,+m) \\ &= P(+i,+r,+m) + P(+i,-r,+m) + P(-i,+r,+m) + P(-i,-r,+m) \\ &= 0 + 0 + 0.056 + 0.024 = 0.08 = \frac{8}{100} \end{aligned}$$

c.

$$P(+r \mid +m) = \frac{P(+r,+m)}{P(+m)} = \frac{\sum_i P(i,+r,+m)}{P(+m)} = \frac{0+0.056}{0.080} = 0.143 = \frac{7}{10}$$

d. Assertions (i), (iv), and (vi) are guaranteed to be true.

Exercise 13.NINE

Consider the Bayes net below in Figure S13.22 with 9 variables:

a. Which random variables are independent of $X_{3,1}$?

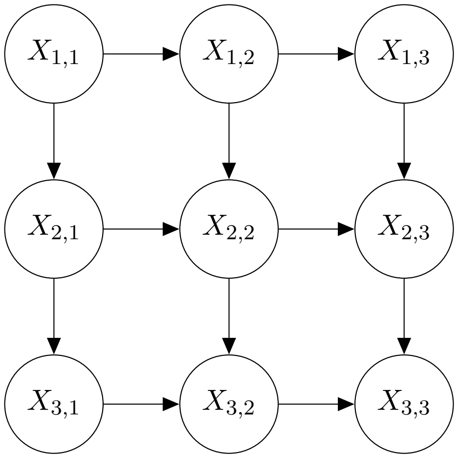


Figure S13.22 Caption

b. Which random variables are conditionally independent of $X_{3,1}$ *given* $X_{1,1}$?

c. Which random variables are conditionally independent of $X_{3,1}$ *given* $X_{1,1}$ and $X_{3,3}$?

a. None—there is at least one active path between $X_{3,1}$ and every other node.

b. $X_{1,2}$ and $X_{1,3}$. $X_{1,1}$ blocks the only active paths to both $X_{1,2}$ and $X_{1,3}$, so both of those become independent of $X_{3,1}$ given $X_{1,1}$.

c. None. The path from a node down to $X_{3,3}$ and up to another node is an active path. [[check]]

Exercise 13.EDGD

For the Bayes net structures in Figure S**??** and Figure S**??** that are missing a direction on their edges, assign a direction to each edge such that the Bayes net structure implies the stated conditional independences and does not imply the conditional independences stated not to hold.

a. **Constraints:**

- $D \perp\!\!\!\perp G$
- not $D \perp\!\!\!\perp A$
- $D \perp\!\!\!\perp E$
- $H \perp\!\!\!\perp F$

b. **Constraints:**

- $D \perp\!\!\!\perp F$
- not $D \perp\!\!\!\perp G$

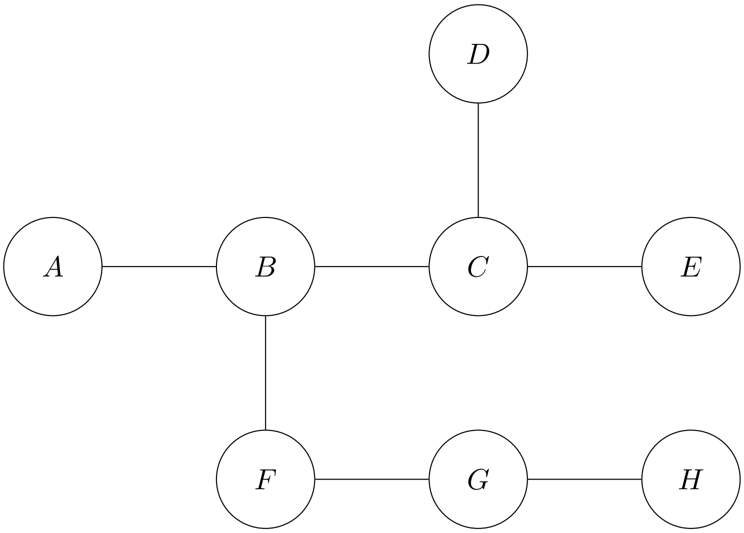


Figure S13.23 Figure (a) for Exercise 13.EDGD.

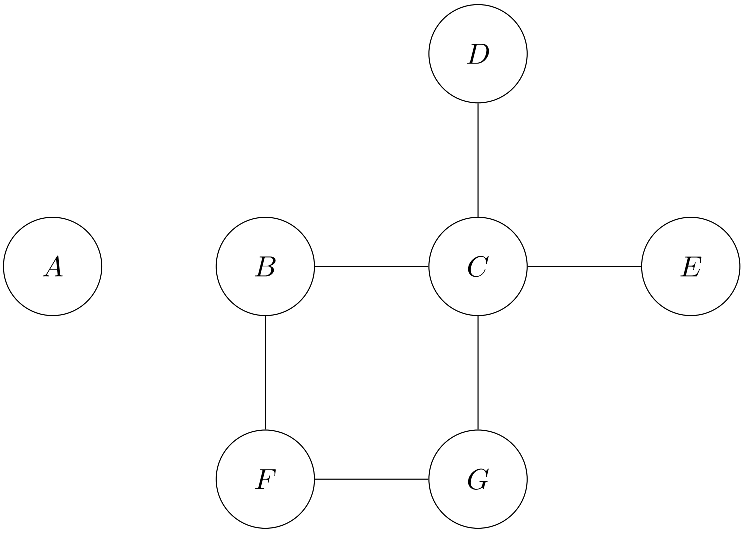


Figure S13.24 Figure (b) for Exercise 13.EDGD.

- $D \perp\!\!\!\perp E$
- Bayes net has no directed cycles

a. The following edge directions are required:

$B \rightarrow A$
$C \rightarrow B$
$D \rightarrow C$
$E \rightarrow C$
$F \rightarrow B$
$F \rightarrow G$
$H \rightarrow G$

b. The following edge directions are required:

$C \rightarrow B$
$F \rightarrow B$
$F \rightarrow G$
$C \rightarrow G$
$D \rightarrow C$
$E \rightarrow C$

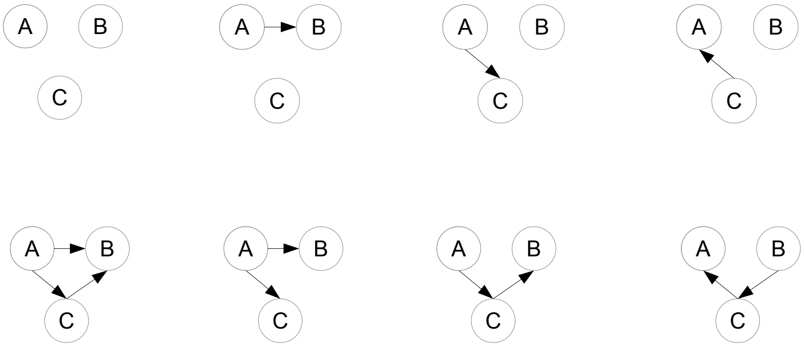


Figure S13.25 A Bayesian network for Exercise 13.PABC.

Exercise 13.PABC

In which of the Bayes nets in Figure S13.25 does the equation $P(A,B)P(C) = P(A)P(B,C)$ *necessarily* hold?

It holds in the first, third, and fourth networks. We can rewrite the equation as follows:

$$\begin{aligned} & P(A,B)P(C) = P(A)P(B,C) \\ \Rightarrow\ & P(A,B)/P(A) = P(B,C)/P(C) \\ \Rightarrow\ & P(B \mid A) = P(B \mid C)\,. \end{aligned}$$

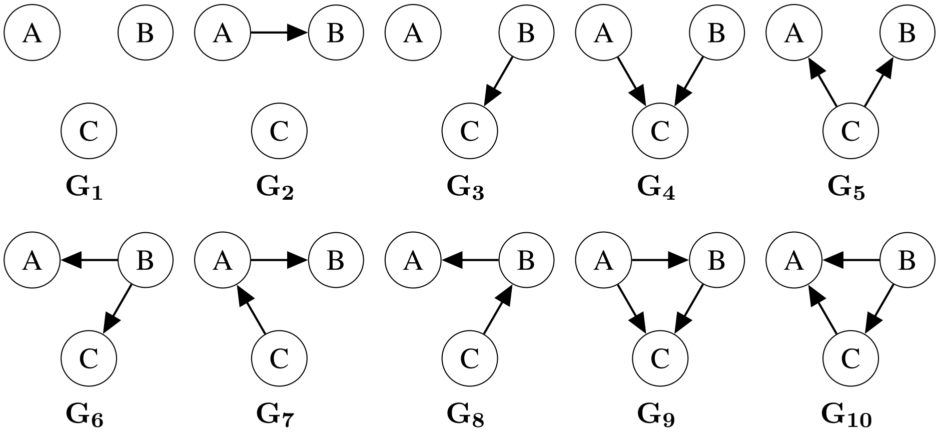


Figure S13.26 Ten Bayes nets for Exercise 13.GTEN.

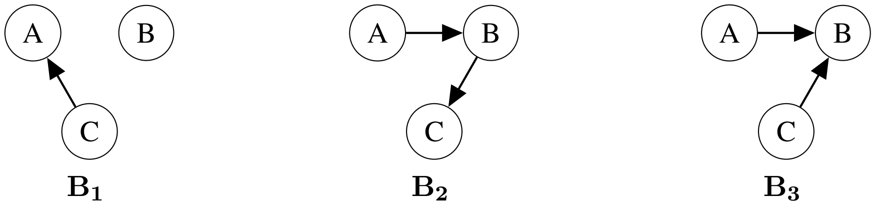


Figure S13.27 Three Bayes nets for Exercise 13.GTEN.

The only way this can *necessarily* hold (i.e., without numerical coincidences) is if B is absolutely independent of A and C.

Exercise 13.GTEN

Assume we are given the ten Bayes nets in Figure S13.26, labeled $\mathbf{G_1}$ to $\mathbf{G_{10}}$. Assume we are also given the three Bayes nets in Figure S13.27, labeled $\mathbf{B_1}$ to $\mathbf{B_3}$.

a. Assume we know that a joint distribution $\mathbf{d_1}$ (over $\mathbf{A}, \mathbf{B}, \mathbf{C}$) can be represented by Bayes net $\mathbf{B_1}$. Which of $\mathbf{G_1}$ through $\mathbf{G_{10}}$ are guaranteed to be able to represent $\mathbf{d_1}$?

b. Assume we know that a joint distribution $\mathbf{d_2}$ (over $\mathbf{A}, \mathbf{B}, \mathbf{C}$) can be represented by Bayes net $\mathbf{B_2}$. Which of $\mathbf{G_1}$ through $\mathbf{G_{10}}$ are guaranteed to be able to represent $\mathbf{d_2}$?

c. Assume we know that a joint distribution $\mathbf{d_3}$ (over $\mathbf{A}, \mathbf{B}, \mathbf{C}$) ***cannot*** be represented by Bayes net $\mathbf{B_3}$. Which of $\mathbf{G_1}$ through $\mathbf{G_{10}}$ are guaranteed to be able to represent $\mathbf{d_3}$?

d. Assume we know that a joint distribution $\mathbf{d_4}$ (over $\mathbf{A}, \mathbf{B}, \mathbf{C}$) can be represented by Bayes nets $\mathbf{B_1}$, $\mathbf{B_2}$, and $\mathbf{B_3}$. Which of $\mathbf{G_1}$ through $\mathbf{G_{10}}$ are guaranteed to be able to represent $\mathbf{d_4}$?

It helps to enumerate all of the (conditional) independence assumptions encoded in the Bayes nets. They are:

- $\mathbf{G_1}$: $A \perp\!\!\!\perp B; A \perp\!\!\!\perp B \,|\, C; A \perp\!\!\!\perp C; A \perp\!\!\!\perp C \,|\, B; B \perp\!\!\!\perp C; B \perp\!\!\!\perp C \,|\, A$
- $\mathbf{G_2}$: $A \perp\!\!\!\perp C; A \perp\!\!\!\perp C \,|\, B; B \perp\!\!\!\perp C; B \perp\!\!\!\perp C \,|\, A$
- $\mathbf{G_3}$: $A \perp\!\!\!\perp B; A \perp\!\!\!\perp B \,|\, C; A \perp\!\!\!\perp C; A \perp\!\!\!\perp C \,|\, B$
- $\mathbf{G_4}$: $A \perp\!\!\!\perp B$
- $\mathbf{G_5}$: $A \perp\!\!\!\perp B \,|\, C$
- $\mathbf{G_6}$: $A \perp\!\!\!\perp C \,|\, B$
- $\mathbf{G_7}$: $B \perp\!\!\!\perp C \,|\, A$
- $\mathbf{G_8}$: $A \perp\!\!\!\perp C \,|\, B$
- $\mathbf{G_9}$: $\{\}$
- $\mathbf{G_{10}}$: $\{\}$
- $\mathbf{B_1}$: $A \perp\!\!\!\perp B; A \perp\!\!\!\perp B \,|\, C; B \perp\!\!\!\perp C; B \perp\!\!\!\perp C \,|\, A$
- $\mathbf{B_2}$: $A \perp\!\!\!\perp C \,|\, B$
- $\mathbf{B_3}$: $A \perp\!\!\!\perp C$

a. $\mathbf{G_4}$, $\mathbf{G_5}$, $\mathbf{G_7}$, $\mathbf{G_9}$, $\mathbf{G_{10}}$.
Since $\mathbf{B_1}$ can represent $\mathbf{d_1}$, we know that $\mathbf{d_1}$ must satisfy the assumptions that $\mathbf{B_1}$ follows, which are:
$A \perp\!\!\!\perp B; A \perp\!\!\!\perp B \,|\, C; B \perp\!\!\!\perp C; B \perp\!\!\!\perp C \,|\, A$. We cannot assume that $\mathbf{d_1}$ satisfies the other two assumptions, which are $A \perp\!\!\!\perp C$ and $A \perp\!\!\!\perp C \,|\, B$, and so a Bayes net that makes at least one of these two extra assumptions will not be guaranteed to be able to represent $\mathbf{d_1}$. This eliminates the choices $\mathbf{G_1}, \mathbf{G_2}, \mathbf{G_3}, \mathbf{G_6}, \mathbf{G_8}$. The other choices $\mathbf{G_4}, \mathbf{G_5}, \mathbf{G_7}, \mathbf{G_9}, \mathbf{G_{10}}$ are guaranteed to be able to represent $\mathbf{d_1}$ because they do not make any additional independence assumptions that $\mathbf{B_1}$ makes.

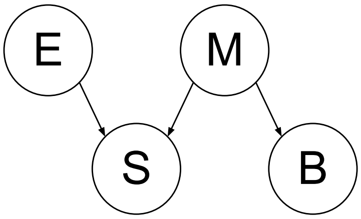
b. $\mathbf{G_6}$, $\mathbf{G_8}$, $\mathbf{G_9}$, $\mathbf{G_{10}}$.
Since $\mathbf{B_2}$ can represent $\mathbf{d_2}$, we know that $\mathbf{d_2}$ must satisfy the assumptions that $\mathbf{B_2}$ follows, which is just: $A \perp\!\!\!\perp C \,|\, B$. We cannot assume that $\mathbf{d_2}$ satisfies any other assumptions, and so a Bayes net that makes at least one other extra assumptions will not be guaranteed to be able to represent $\mathbf{d_2}$. This eliminates the choices $\mathbf{G_1}, \mathbf{G_2}, \mathbf{G_3}, \mathbf{G_4}, \mathbf{G_5}, \mathbf{G_7}$. The other choices $\mathbf{G_6}, \mathbf{G_8}, \mathbf{G_9}, \mathbf{G_{10}}$ are guaranteed to be able to represent $\mathbf{d_2}$ because they do not make any additional independence assumptions that $\mathbf{B_2}$ makes.

c. $\mathbf{G_9}$, $\mathbf{G_{10}}$.
Since $\mathbf{B_3}$ cannot represent $\mathbf{d_3}$, we know that $\mathbf{d_3}$ is unable to satisfy at least one of the assumptions that $\mathbf{B_3}$ follows. Since $\mathbf{B_3}$ only makes one independence assumption, which is $A \perp\!\!\!\perp C$, we know that $\mathbf{d_3}$ does not satisfy $A \perp\!\!\!\perp C$. However, we can't claim anything about whether or not $\mathbf{d_3}$ makes any of the other independence assumptions. $\mathbf{d_3}$ might not make any (conditional) independence assumptions at all, and so only the Bayes nets that don't make any assumptions will be guaranteed to be able to represent $\mathbf{d_3}$. Hence, the answers are the fully connected Bayes nets, which are $\mathbf{G_9}, \mathbf{G_{10}}$.

d. All ten networks can represent d_4.
Since $\mathbf{B_1}, \mathbf{B_2}, \mathbf{B_3}$ can represent $\mathbf{d_4}$, we know that $\mathbf{d_4}$ must satisfy the assumptions that $\mathbf{B_1}, \mathbf{B_2}, \mathbf{B_3}$ make. The union of assumptions made by these Bayes nets are: $A \perp$

$P(E)$	
$+e$	0.4
$-e$	0.6

$P(S \mid E, M)$			
$+e$	$+m$	$+s$	1.0
$+e$	$+m$	$-s$	0.0
$+e$	$-m$	$+s$	0.8
$+e$	$-m$	$-s$	0.2
$-e$	$+m$	$+s$	0.3
$-e$	$+m$	$-s$	0.7
$-e$	$-m$	$+s$	0.1
$-e$	$-m$	$-s$	0.9



$P(M)$	
$+m$	0.1
$-m$	0.9

$P(B \mid M)$		
$+m$	$+b$	1.0
$+m$	$-b$	0.0
$-m$	$+b$	0.1
$-m$	$-b$	0.9

Figure S13.28 A Bayes net for the end of the world.

$\perp B; A \perp\!\!\!\perp B \mid C; B \perp\!\!\!\perp C; B \perp\!\!\!\perp C \mid A, A \perp\!\!\!\perp C, A \perp\!\!\!\perp C \mid B$. Note that this set of assumptions encompasses all the possible assumptions that you can make with 3 random variables, so any Bayes net over $\mathbf{A}, \mathbf{B}, \mathbf{C}$ will be able to represent $\mathbf{d_4}$.

13.3 Exact Inference in Bayesian Networks

Exercise 13.DOOM

A smell of sulphur (S) can be caused either by rotten eggs (E) or as a sign of the doom brought by the Mayan Apocalypse (M). The Mayan Apocalypse also causes the oceans to boil (B). The Bayesian network and corresponding conditional probability tables for this situation are shown in Figure S13.28.

a. Compute the joint probability $P(-e, -s, -m, -b)$.

b. What is the probability that the oceans boil?

c. What is the probability that the Mayan Apocalypse is occurring, given that the oceans are boiling?

d. What is the probability that the Mayan Apocalypse is occurring, given that there is a smell of sulphur, the oceans are boiling, and there are rotten eggs?

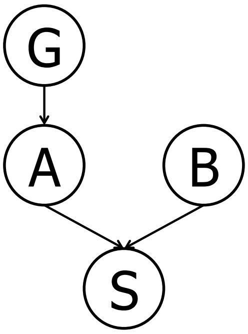
e. What is the probability that rotten eggs are present, given that the Mayan Apocalypse is occurring?

a. $P(-e, -s, -m, -b) = P(-e)P(-m)P(-s \mid -e, -m)P(-b \mid -m) = (0.6)(0.9)(0.9)(0.9) = 0.4374$.

b. $P(+b) = P(+b \mid +m)P(+m) + P(+b \mid -m)P(-m) = (1.0)(0.1) + (0.1)(0.9) = 0.19$.

$\mathbf{P}(G)$	
$+g$	0.1
$-g$	0.9

$\mathbf{P}(A \mid G)$		
$+g$	$+a$	1.0
$+g$	$-a$	0.0
$-g$	$+a$	0.1
$-g$	$-a$	0.9



$\mathbf{P}(B)$	
$+b$	0.4
$-b$	0.6

$\mathbf{P}(S \mid A, B)$			
$+a$	$+b$	$+s$	1.0
$+a$	$+b$	$-s$	0.0
$+a$	$-b$	$+s$	0.9
$+a$	$-b$	$-s$	0.1
$-a$	$+b$	$+s$	0.8
$-a$	$+b$	$-s$	0.2
$-a$	$-b$	$+s$	0.1
$-a$	$-b$	$-s$	0.9

Figure S13.29 A Bayes net for genetic disease.

c. $P(+m \mid +b) = \frac{P(+b \mid +m)P(+m)}{P(+b)} = \frac{(1.0)(0.1)}{0.19} \approx .5263.$

d.

$$\begin{aligned} P(+m \mid +s, +b, +e) &= \frac{P(+m, +s, +b, +e)}{\sum_m P(m, +s, +b, +e)} \\ &= \frac{P(+e)P(+m)P(+s \mid +e, +m)P(+b \mid +m))}{\sum_m P(+e)P(m)P(+s \mid +e, m)P(+b \mid m)} \\ &= \frac{(0.4)(0.1)(1.0)(1.0)}{(0.4)(0.1)(1.0)(1.0) + (0.4)(0.9)(0.8)(0.1)} \\ &= \frac{0.04}{0.04 + 0.0288} \approx .5814 \end{aligned}$$

e. $P(+e \mid +m) = P(+e) = 0.4$ since E is independent of M.

Exercise 13.BGEN

Suppose that a patient can have a symptom (S) that can be caused by two different diseases (A and B). It is known that the variation of gene G plays a big role in the occurrence of disease A. The Bayes net and corresponding conditional probability tables for this situation are in Figure S13.29.

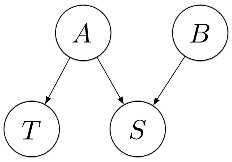
a. Compute the joint probability $\mathbf{P}(+g, +a, +b, +s)$.

b. What is the probability that a patient has disease A?

c. What is the probability that a patient has disease A given that they have disease B?

d. What is the probability that a patient has disease A given that they have symptom S and disease B?

$\mathbf{P}(A)$	
$+a$	0.1
$-a$	0.9



$\mathbf{P}(B)$	
$+b$	0.5
$-b$	0.5

$\mathbf{P}(S \mid A, B)$			
$+a$	$+b$	$+s$	1
$+a$	$+b$	$-s$	0
$+a$	$-b$	$+s$	0.80
$+a$	$-b$	$-s$	0.20
$-a$	$+b$	$+s$	1
$-a$	$+b$	$-s$	0
$-a$	$-b$	$+s$	0
$-a$	$-b$	$-s$	1

$\mathbf{P}(T \mid A)$		
$+a$	$+t$	1
$+a$	$-t$	0
$-a$	$+t$	0.2
$-a$	$-t$	0.8

Figure S13.30 A Bayes net with a test for disease.

e. What is the probability that a patient has the disease-carrying gene variation G given that they have disease A?

f. What is the probability that a patient has the disease-carrying gene variation G given that they have disease B?

a. $P(+g,+a,+b,+s) = P(+g)P(+a \mid +g)P(+b)P(+s \mid +b,+a) = (0.1)(1.0)(0.4)(1.0) = 0.04$.

b. $P(+a) = P(+a \mid +g)P(+g) + P(+a \mid -g)P(-g) = (1.0)(0.1) + (0.1)(0.9) = 0.19$.

c. $P(+a \mid +b) = P(+a) = 0.19$ because A abd B are absolutely independent.

d. $P(+a \mid +s,+b) = \frac{P(+a,+b,+s)}{P(+a,+b,+s)+P(-a,+b,+s)} = \frac{P(+a)P(+b)P(+s \mid +a,+b)}{P(+a)P(+b)P(+s \mid +a,+b)+P(-a)P(+b)P(+s \mid -a,+b)}$
$= \frac{(0.19)(0.4)(1.0)}{(0.19)(0.4)(1.0)+(0.81)(0.4)(0.8)} = \frac{0.076}{0.076+0.2592} \approx 0.2267$.

e. $P(+g \mid +a) = \frac{P(+g)P(+a \mid +g)}{P(+g)P(+a \mid +g)+P(-g)P(+a \mid -g)} = \frac{(0.1)(1.0)}{(0.1)(1.0)+(0.9)(0.1)} = \frac{0.1}{0.1+0.09} = 0.5263$.

f. $P(+g \mid +b) = P(+g) = 0.1$, since $G \perp\!\!\!\perp B$.

Exercise 13.BTST

Suppose that a patient can have a symptom (S) that can be caused by two different diseases (A and B). Disease A is much rarer, but there is a test T that tests for the presence of A. The Bayes net and corresponding conditional probability tables for this situation are shown in Figure S13.30.

a. Compute the entry $\mathbf{P}(-a,-t,+b,+s)$ from the joint distribution.

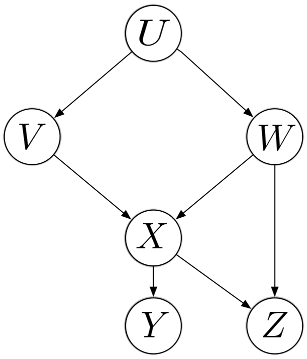


Figure S13.31 A simple Bayes net.

b. What is the probability that a patient has disease A given that they have disease B?

c. What is the probability that a patient has disease A given that they have symptom S, disease B, and test T returns positive?

d. What is the probability that a patient has disease A given that they have symptom S and test T returns positive?

e. Suppose that both diseases A and B become more likely as a person ages. Add any necessary variables and/or arcs to the Bayes net to represent this change. For any variables you add, *briefly* (one sentence or less) state what they represent. Also, state one independence or conditional independence assertion that is *removed* due to your changes.

f. Based only on the structure of the (new) Bayes net given in Figure S13.31, decide whether the following conditional independence assertions are guaranteed to be true, guaranteed to be false, or cannot be determined by the structure alone.

(i) $V \perp\!\!\!\perp W$
(ii) $V \perp\!\!\!\perp W \mid U$
(iii) $V \perp\!\!\!\perp W \mid U, Y$
(iv) $V \perp\!\!\!\perp Z \mid U, X$
(v) $X \perp\!\!\!\perp Z \mid W$

a. $P(-a, -t, +b, +s) = P(-t \mid -a)P(-a)P(+s \mid +b, -a)P(+b) = (0.8)(0.9)(1)(0.5) = 0.36$.

b. The network asserts absolute independence of A and B, so $P(+a \mid +b) = P(+a) = 0.1$.

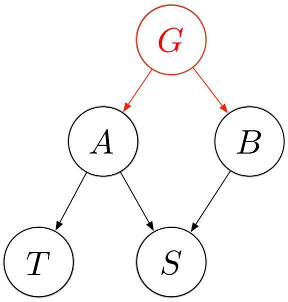
c.

$$\begin{aligned} P(+a \mid +t,+s,+b) &= \frac{\mathbf{P}(+a,+t,+s,+b)}{\mathbf{P}(+t,+s,+b)} = \frac{\mathbf{P}(+a)\mathbf{P}(+t \mid +a)\mathbf{P}(+s \mid +a,+b)\mathbf{P}(+b)}{\sum_{a\in\{+a,-a\}} \mathbf{P}(a,+t,+s,+b)} \\ &= \frac{(0.1)(1)(1)(0.5)}{(0.1)(1)(1)(0.5)+(0.9)(0.2)(1)(0.5)} = \frac{.05}{.05+.09} = \frac{5}{14} \approx 0.35. \end{aligned}$$

d. $\mathbf{P}(+a \mid +t,+s) = \frac{\mathbf{P}(+a,+t,+s)}{\mathbf{P}(+t,+s)} \frac{\sum_b \mathbf{P}(+a)\mathbf{P}(+t \mid +a)\mathbf{P}(+s \mid +a,b)\mathbf{P}(b)}{\sum_a \sum_b \mathbf{P}(a)\mathbf{P}(+t \mid a)\mathbf{P}(+s \mid a,b)\mathbf{P}(b)} = 0.5.$

e. **New variable(s) (and brief definition)**: G: Age in years, or as a Boolean for old or not, etc.

Removed conditional independence assertion: There are a few. The simplest is $A \perp\!\!\!\perp B$ is no longer guaranteed. Another is $B \perp\!\!\!\perp T$.



f.

(i)	$V \perp\!\!\!\perp W$	Cannot be determined.
(ii)	$V \perp\!\!\!\perp W \mid U$	Guaranteed true.
(iii)	$V \perp\!\!\!\perp W \mid U, Y$	Cannot be determined.
(iv)	$V \perp\!\!\!\perp Z \mid U, X$	Cannot be determined.
(v)	$X \perp\!\!\!\perp Z \mid W$	Cannot be determined.

Exercise 13.BJNT

a. Express the joint probability distribution $\mathbf{P}(A,B,C)$ induced by the Bayes net in Figure S13.32 in terms of the associated conditional probability distributions.

Compute the values of the following probabilities:

b. $P(C = true)$.

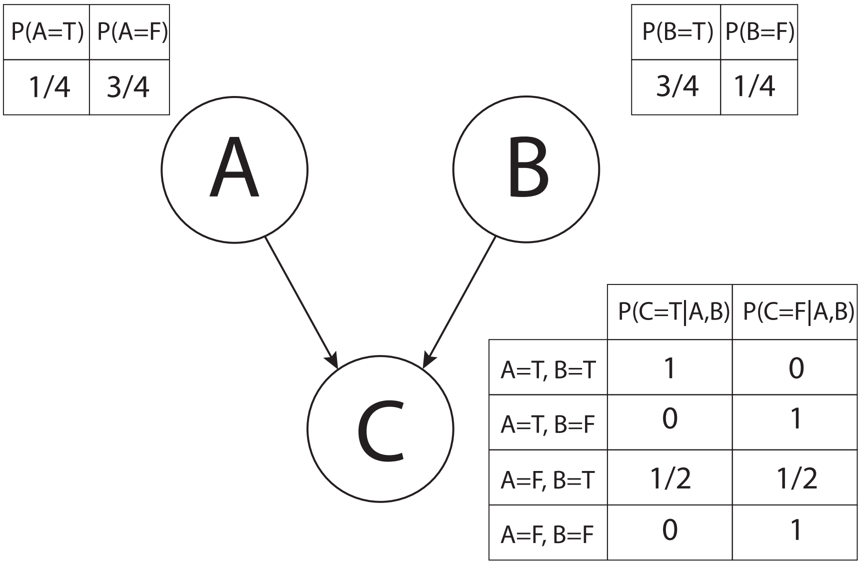
c. $P(A = true, B = true)$.

d. $(A = true, B = true \mid C = true)$.

a. The joint probability is the product of the conditional probability if each variable given its parents. A and B have no parents, so we have $\mathbf{P}(A,B,C) = P(A)P(B)P(C|A,B)$.

b. $P(C = true) = \sum_{a,b} P(a)P(b)P(C = true \mid a,b) = \frac{1}{4}\frac{3}{4}1 + \frac{1}{4}\frac{1}{4}0 + \frac{3}{4}\frac{3}{4}\frac{1}{2} + \frac{3}{4}\frac{1}{4}0 = \frac{15}{32}$.

c. $P(A = true, B = true) = P(A = true)P(B = true) = \frac{3}{16}$, by absolute independence of A and B.



P(A=T)	P(A=F)
1/4	3/4

P(B=T)	P(B=F)
3/4	1/4

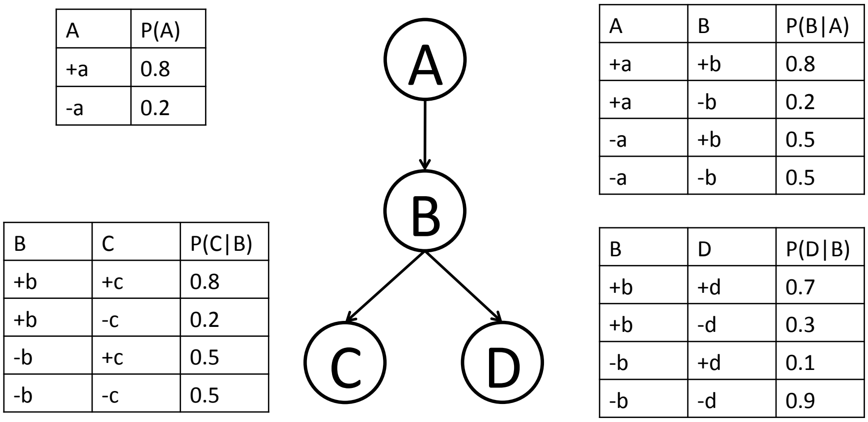
	P(C=T|A,B)	P(C=F|A,B)
A=T, B=T	1	0
A=T, B=F	0	1
A=F, B=T	1/2	1/2
A=F, B=F	0	1

Figure S13.32 A Bayesian network for Exercise 13.BJNT.

d.

$$\begin{aligned}(A=true, B=true \mid C=true) &= \frac{P(A=true, B=true, C=true)}{P(C=true)}\\ &= \frac{P(A=true)P(B=true)P(C=true|A=true, B=true)}{P(C=true)}\\ &= (\frac{1}{4}\frac{3}{4}1)/\frac{15}{32} = \frac{2}{5}\end{aligned}$$

.



A	P(A)
+a	0.8
-a	0.2

A	B	P(B|A)
+a	+b	0.8
+a	-b	0.2
-a	+b	0.5
-a	-b	0.5

B	C	P(C|B)
+b	+c	0.8
+b	-c	0.2
-b	+c	0.5
-b	-c	0.5

B	D	P(D|B)
+b	+d	0.7
+b	-d	0.3
-b	+d	0.1
-b	-d	0.9

Figure S13.33 A Bayesian network for Exercise 13.BJTW.

Exercise 13.BJTW

Consider the joint distribution $P(A,B,C,D)$ defined by the Bayes net in Figure S13.33. Compute the values of the following quantities:

a. $P(A = true)$.

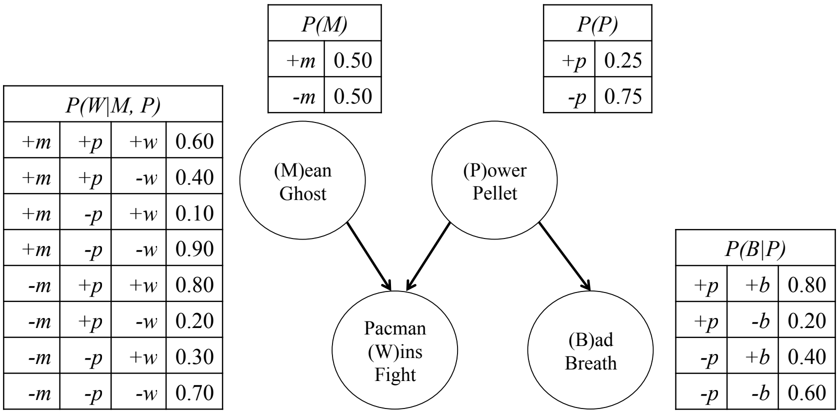
b. $P(A = true, B = false, C = false, D = true)$.

c. $P(A = true, B = false, C = false, D = true)$.

a. 0.8.

b. $0.8 \times 0.2 \times 0.1 \times 0.5 = 0.008$.

c. $\frac{0.8 \times 0.2 \times 0.1 \times 0.5}{0.8 \times 0.2 \times 0.1 \times 0.5 + 0.2 \times 0.5 \times 0.1 \times 0.5} = 0.615$



P(M)	
+m	0.50
-m	0.50

P(P)	
+p	0.25
-p	0.75

P(W|M, P)			
+m	+p	+w	0.60
+m	+p	-w	0.40
+m	-p	+w	0.10
+m	-p	-w	0.90
-m	+p	+w	0.80
-m	+p	-w	0.20
-m	-p	+w	0.30
-m	-p	-w	0.70

P(B|P)		
+p	+b	0.80
+p	-b	0.20
-p	+b	0.40
-p	-b	0.60

Figure S13.34 A Bayesian network for Exercise 13.PACL.

$P(M,P,W,B)$				
$+m$	$+p$	$+w$	$+b$	0.0800
$+m$	$+p$	$+w$	$-b$	0.0150
$+m$	$+p$	$-w$	$+b$	0.0400
$+m$	$+p$	$-w$	$-b$	0.0100
$+m$	$-p$	$+w$	$+b$	0.0150
$+m$	$-p$	$+w$	$-b$	0.0225
$+m$	$-p$	$-w$	$+b$	0.1350
$+m$	$-p$	$-w$	$-b$	0.2025

Figure S13.35 A table for Exercise 13.PACL.

Exercise 13.PACL

PacLabs has just created a new type of mini power pellet that is small enough for Pacman to carry around with him when he's running around mazes. Unfortunately, these mini-pellets don't guarantee that Pacman will win all his fights with ghosts, and they look just like the regular dots Pacman carries around to snack on.

Pacman just ate a snack (P), which was either a mini-pellet ($+p$), or a regular dot ($-p$), and is about to get into a fight (W), which he can win ($+w$) or lose ($-w$). Both these variables are unknown, but fortunately, Pacman is a master of probability. He knows that his bag of snacks has 5 mini-pellets and 15 regular dots. He also knows that if he ate a mini-pellet, he has a 70% chance of winning, but if he ate a regular dot, he only has a 20% chance.

a. What is $P(+w)$, the marginal probability that Pacman will win?

b. Pacman won! Hooray! What is the conditional probability $P(+p \mid +w)$ that the food he ate was a mini-pellet, given that he won?
Pacman can make better probability estimates if he takes more information into account. First, Pacman's breath, B, can be bad ($+b$) or fresh ($-b$). Second, there are two types of ghost (M): mean ($+m$) and nice ($-m$). Pacman has encoded his knowledge about the situation in the Bayes net in Figure S13.34.

c. What is the probability of the atomic event $(-m, +p, +w, -b)$, where Pacman eats a mini-pellet and has fresh breath before winning a fight against a nice ghost?

d. Which of the following conditional independence statements are guaranteed to be true by the Bayes net graph structure?

(i) $W \perp\!\!\!\perp B$

(ii) $W \perp\!\!\!\perp B \mid P$

(iii) $M \perp\!\!\!\perp P$

(iv) $M \perp\!\!\!\perp P \mid W$

(v) $M \perp\!\!\!\perp B$

(vi) $M \perp\!\!\!\perp B \mid P$

(vii) $M \perp\!\!\!\perp B \mid W$

For the remainder of this question, use the half of the joint probability table that has been computed for you in Figure S13.35.

e. What is the marginal probability, $P(+m, +b)$ that Pacman encounters a mean ghost and has bad breath?

f. Pacman observes that he has bad breath and that the ghost he's facing is mean. What is the conditional probability, $P(+w \mid +m, +b)$, that he will win the fight, given his observations?

g. Pacman's utility is +10 for winning a fight, -5 for losing a fight, and -1 for running away from a fight. Pacman wants to maximize his expected utility. Given that he has bad breath and is facing a mean ghost, should he stay and fight, or run away? Justify your answer numerically!

a.

$$\begin{aligned} P(+w) &= P(+w,+p)+P(+w,-p) = P(+w|+p)P(+p)+P(+w|-p)P(-p) \\ &= \frac{7}{10}\times\frac{1}{4}+\frac{2}{10}\times\frac{3}{4} = \frac{13}{40} = 0.325 \end{aligned}$$

b.

$$\begin{aligned} P(+p|+w) &= \frac{P(+w,+p)}{P(+w)} = \frac{P(+w|+p)P(+p)}{P(+w)} \\ &= \frac{\frac{7}{10}\times\frac{1}{4}}{\frac{13}{40}} = \frac{7}{13} \approx 0.538 \end{aligned}$$

c. $P(-m,+p,+w,-b) = P(-m)P(+p)P(+w|-m,+p)P(-b|+p) = \frac{1}{2}\times\frac{1}{4}\times\frac{4}{5}\times\frac{1}{5} = \frac{1}{50} = 0.02$.

d. Conditional independence assertions ii, iii, v, and vi are correct.

e. $P(+m,+b) = 0.08+0.04+0.015+0.135 = 0.27$.

f. $P(+w|+m,+b) = \frac{P(+w,+m,+b)}{P(+m,+b)} = \frac{0.08+0.015}{0.27} = \frac{19}{54} \approx 0.352$.

g. This question is a simple preview of material from Chapter 15, but we have already seen the basic idea in the context of backgammon and other game of chance in Chapter **??**. Let U_f be the utility of fighting and U_r be the utility of running.

$$\begin{aligned} E(U_f|+m,+b) &= 10\times P(+w|+m,+b)+(-5)\times P(-w|+m,+b) \\ &\approx 10\times 0.352 - 5\times 0.648 \\ &= 0.28 > -1 = U_r \end{aligned}$$

Since $E(U_f|+m,+b) > E(U_r|+m,+b)$, Pacman should stay and fight.

Exercise 13.VETF

Which of the following are true assertions about the variable elimination algorithm, and which are false?

a. When changing a Bayes net by removing a parent from a variable, the maximum factor size (where size is the number of non-fixed variables involved in the factor) generated during the optimally ordered variable elimination is reduced by at most 1.

b. The ordering of variables in variable elimination affects the maximum factor size generated by at most a factor of two.

c. The size of factors generated during variable elimination is upper-bounded by twice the size of the largest conditional probability table in the original Bayes net.

d. The size of factors generated during variable elimination is the same if we exactly reverse the elimination ordering.

e. During variable elimination, the ordering of elimination does not affect the final answer.

a. False.

b. False.

c. False.

d. False.

e. True. The algorithm is always sound.

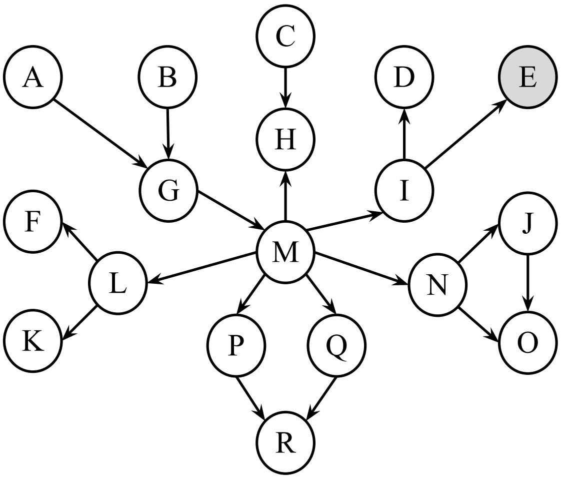


Figure S13.36 A Bayesian network for Exercise 13.ELIM.

Exercise 13.ELIM

Consider the Bayes net in Figure S13.36 with the query $P(Q \,|\, +e)$.

a. Which variables can be ignored when computing the answer to the query?

b. Prove a general result concerning the irrelevance of variables in computing the posterior distribution of set of query variables given a set of evidence variables.

As stated in the chapter, variables that are not in the union of the ancestors of the query and ancestors of the evidence can be pruned.

a. In computing $P(Q \,|\, +e)$, we can ignore C, D, F, H, J, K, L, N, O, P, R.

b. The proof of irrelevance not shown.

Exercise 13.ELMU

a. For the Bayes net in Figure S13.37, we are given the query $P(Z \,|\, +y)$. All variables are Boolean. Assume we run variable elimination to compute the answer to this query, with the following variable elimination ordering: U, V, W, T, X.

After inserting evidence, we have the following factors to start out with:

$$P(U), P(V), P(W \,|\, U, V), P(X \,|\, V), P(T \,|\, V), P(+y \,|\, W, X), P(Z \,|\, T)\,.$$

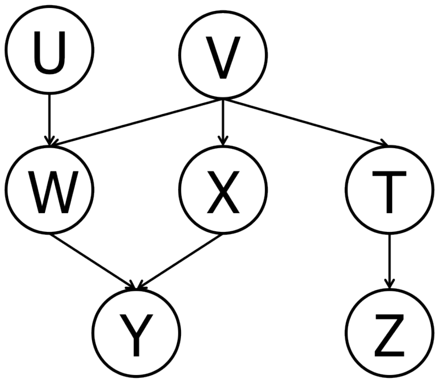


Figure S13.37 A Bayesian network for Exercise 13.ELMU.

When eliminating U we generate a new factor f_1 as follows:

$$f_1(V, W) = \sum_u P(u)P(W \mid u, V) .$$

This leaves us with the factors:

$$P(V), P(X \mid V), P(T \mid V), P(+y \mid W, X), P(Z \mid T), f_1(V, W) .$$

Now complete the remaining steps:

(i) When eliminating V we generate a new factor f_2 as follows:
(ii) This leaves us with the factors:
(iii) When eliminating W we generate a new factor f_3 as follows:
(iv) This leaves us with the factors:
(v) When eliminating T we generate a new factor f_4 as follows:
(vi) This leaves us with the factor:
(vii) When eliminating X we generate a new factor f_5 as follows:
(viii) This leaves us with the factor:

b. Briefly explain how $P(Z \mid +y)$ can be computed from f_5.

c. Amongst $f_1, f_2, \ldots, f_5$, which is the largest factor generated? How large is this factor?

d. Find a variable elimination ordering for the same query, i.e., for $P(Z \mid y)$, for which the maximum size factor generated along the way is smallest. Hint: the maximum size factor generated in your solution should have only 2 variables, for a size of $2^2 = 4$ entries.

a. (i) When eliminating V we generate a new factor f_2 as follows:

$$f_2(T,W,X) = \sum_v P(v)P(X \mid v)P(T \mid v)f_1(v,W).$$

(ii) This leaves us with the factors:

$$P(+y \mid W,X), P(Z \mid T), f_2(T,W,X).$$

(iii) When eliminating W we generate a new factor f_3 as follows:

$$f_3(T,X,+y) = \sum_w P(+y \mid w,X)f_2(T,w,X).$$

(iv) This leaves us with the factors:

$$P(Z \mid T), f_3(T,X,+y).$$

(v) When eliminating T we generate a new factor f_4 as follows:

$$f_4(X,+y,Z) = \sum_t P(Z \mid t)f_3(t,X,+y).$$

(vi) This leaves us with the factor:

$$f_4(X,+y,Z).$$

(vii) When eliminating X we generate a new factor f_5 as follows:

$$f_5(+y,Z) = \sum_x f_4(x,+y,Z).$$

(viii) This leaves us with the factor:

$$f_5(+y,Z)$$

.

b. Simply renormalize f_5 to obtain $P(Z \mid +y)$. Concretely, $P(z \mid +y) = \frac{f_5(z,+y)}{\sum_{z'} f_5(z',y)}$.

c. $f_2(T,W,X)$ is the largest factor generated. It has 3 variables, hence $2^3 = 8$ entries.

d. One possible ordering is T, X, W, U, V.

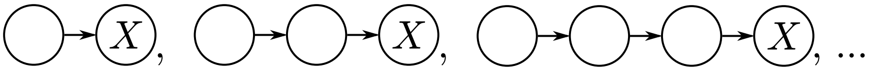


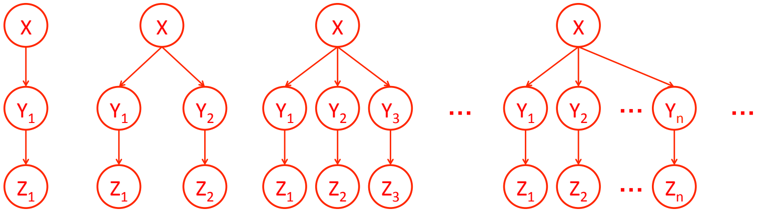
Figure S13.38 A Bayesian network for Exercise 13.SEQE.

Exercise 13.SEQE

Consider the sequence of graphs in Figure S13.38 and the application of variable elimination to each in order to compute $\mathbf{P}(X)$. For each, regardless of the elimination ordering, the largest factor produced in finding will have a table with 2^2 entries, assuming that all variables are Boolean.

Now draw a sequence of graphs such that, if you used the best elimination ordering for each graph, the largest factor table produced in variable elimination would have a constant number of entries, but if you used the worst elimination ordering for each graph, the number of entries in the largest factor table would grow exponentially as you move down the sequence. Provide (i) the sequence of graphs, (ii) the sequence of queries for which variable elimination is done, (iii) the best ordering, (iv) the worst ordering.

There are many solutions, here is one family of Bayes net graphs that will do, with query $P(Y_n \mid Z_1, \ldots, Z_n)$, a best ordering: $Y_1, Y_2, \ldots, Y_{n-1}, X$, a worst ordering: $X, Y_1, Y_2, \ldots, Y_{n-1}$.



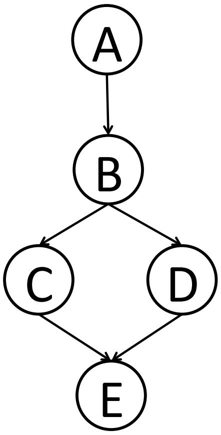


Figure S13.39 A Bayesian network for Exercise 13.ELMO.

Exercise 13.ELMO

In this question, we consider the efficiency of variable elimination as a function of the elimination ordering. For any variable, X, let $|X|$ denote the size of X's range (the number of values it can take). At any stage in the variable elimination process, there will be a set of factors $\mathcal{F} = \{F_1, \ldots F_n\}$. Each factor F_i can be written as $P(L_i \mid R_i)$, where L_i and R_i are both sets of variables (for simplicity, assume that there is no evidence). For any variable X, we define $I(X)$ to be the set of indices of factors that include X: $I(X) = \{i \mid X \in (L_i \cup R_i)\}$.

When eliminating a specific variable, Y, we start by joining the appropriate factors from $\mathcal{F}$ to create a single joined factor, called $join(Y, \mathcal{F})$. Note that $join(Y, \mathcal{F})$ is created *before* performing the elimination step, so it still includes Y.

a. For a given factor, F, we use $size(F)$ to denote the size of the table needed to represent F. Give a precise general expression for $size(join(Y, \mathcal{F}))$, using the notation above.

b. Consider a generic Bayes net with variables $X \in V$ that encodes a set of initial factors $\mathcal{F}_0$, where the query is to compute the marginal $\mathbf{P}(Q)$. Formulate a standard search problem (as in Chapter 3) that determines the optimal variable elimination ordering, where cost is measured by the sum of the sizes of the tables created during elimination. Note that for this problem we are only concerned with the sum of the sizes of the tables created by the *join* step, not the smaller tables that are subsequently created by the *eliminate* step. (You may also find it helpful to use the notation $eliminate(X, F)$ to denote the result of summing over values of a variable X to eliminate X from factor F.)

c. Let W be the set of variables remaining to be eliminated and let $parents(X)$ and $children(X)$ denote the sets containing all of X's parents/children in the Bayes net. Which of the following are admissible heuristics for the search problem you defined in part (b)?

(i) $|W|$

(ii) $|\mathcal{F}|$

(iii) $\sum_{\mathrm{X} \in W} |\mathrm{X}|$

(iv) $\prod_{\mathrm{X} \in W} |\mathrm{X}|$

(v) $\sum_{\mathrm{X} \in W} \left(|\mathrm{X}| \prod_{\mathrm{Y} \in \mathrm{parents(X)}} |\mathrm{Y}| \right)$

(vi) $\sum_{\mathrm{X} \in W} \left(|\mathrm{X}| \prod_{\mathrm{Y} \in \mathrm{children(X)}} |\mathrm{Y}| \right)$

d. Consider the query $P(D \mid +e, +c)$ on the Bayes net given in Figure S13.39. Draw the complete search tree for finding the optimal elimination order for this problem. Annotate nodes with states, and annotate costs and actions on the edges. Hint: the start state is $(\{P(A), P(B \mid A), P(+c \mid B), P(D \mid B), P(+e \mid +c, D)\}, \{A, B\})$, and your tree should have five nodes in all. What is the optimal plan and what is its cost?

a. The size is $\prod_{I(X_i) \in \{Y, \mathcal{F}\}} |X_i|$

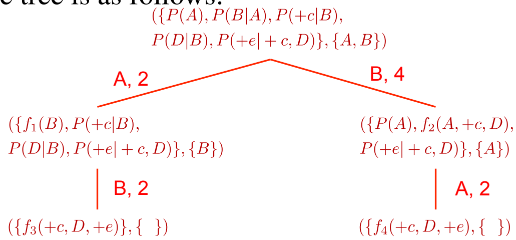
b. The problem formulation is as follows:

- State space: a state consists of the current set of factors and a set of variables still to be eliminated.
- Initial state: the set of conditional probability tables from the Bayes net and the set of non-evidence, non-query variables.
- Goal states: any state where the set of variables still to be eliminated is empty.

- Actions: eliminate any variable in the set of variables still to be eliminated.
- Transition model: the new state contains the new factor description that would result from summing out the variable to be eliminated (without actually doing the computation!); and the eliminated variable is removed from the set of variables still to be eliminated.
- Action cost function: the number of entries in the table representation of the new factor.

c. The admissible heuristics are (i), (ii), and (iii). The rest might overestimate the size of the table in networks that can efficiently eliminate variable.

d. The tree is as follows:


$(\{P(A), P(B|A), P(+c|B),$
$P(D|B), P(+e|+c, D)\}, \{A, B\})$
A, 2
B, 4
$(\{f_1(B), P(+c|B),$
$P(D|B), P(+e|+c, D)\}, \{B\})$
$(\{P(A), f_2(A, +c, D),$
$P(+e|+c, D)\}, \{A\})$
B, 2
A, 2
$(\{f_3(+c, D, +e)\}, \{\ \})$
$(\{f_4(+c, D, +e), \{\ \})$

There is one optimal plan: the left branch, which first eliminates A and then B. Its cost is $2 + 2 = 4$.

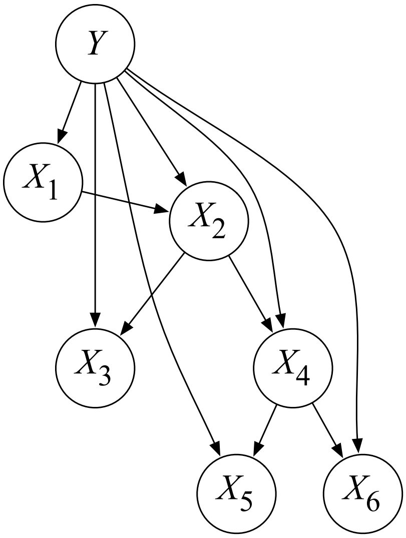


Figure S13.40 A TANBnetwork for Exercise 13.TANB.

Exercise 13.TANB

A tree-augmented Naive Bayes model (TANB) is identical to a Naive Bayes model, except the features are no longer assumed conditionally independent given the class Y. Specifically, if $(X_1, X_2, \ldots, X_n)$ are the observable features, a TANB allows $X_1, \ldots, X_n$ to be in a tree-structured Bayes net in addition to having Y as a parent. An example is given in Figure S13.40:

a. Suppose we observe no variables as evidence in the TANB above. What is the classification rule for the TANB? Write the formula in terms of the conditional distributions in the TANB.

b. Assume we observe all the variables $X_1 = x_1, X_2 = x_2, \ldots, X_6 = x_6$ in the TANB above. Write the classification rule for the TANBin terms of the conditional distributions.

c. Specify an elimination order that is efficient for variable elimination applied to the query $\mathbf{P}(Y \mid X_5 = x_5)$ in the TANB above (the query variable Y should not be included your ordering). How many variables are in the biggest factor (there may be more than one; if so, list only one of the largest) induced by variable elimination with your ordering? Which variables are they?

d. Specify an elimination order that is efficient for the query $P(X_3 \mid X_5 = x_5)$ in the TANB above (including X_3 in your ordering). How many variables are in the biggest factor (there may be more than one; if so, list only one of the largest) induced by variable elimination with your ordering? Which variables are they?

e. Does it make sense to run Gibbs sampling to do inference in a TANB? In two or fewer sentences, justify your answer.

a. The solution is simply the maximal y according to the prior probabilities of y: $\operatorname{argmax}_y\ P(y)$.

b. We want the most probable y given the variables $X_1, \ldots, X_6$, which is (using the same reasoning as for Naive Bayes)

$$\begin{aligned}&\underset{y}{\operatorname{argmax}}\ P(y \mid x_1, \ldots, x_6)\\ &= \underset{y}{\operatorname{argmax}}\ P(y, x_1, \ldots, x_6)\\ &= \underset{y}{\operatorname{argmax}}\ P(y)P(x_1 \mid y)P(x_2 \mid x_1, y)P(x_3 \mid x_2, y)P(x_4 \mid x_2, y)\\ &P(x_5 \mid x_4, y)P(x_6 \mid x_4, y)\ .\end{aligned}$$

c. We can ignore the variables X_3 and X_6, since when we marginalize them in the elimination order, they will sum to 1. Thus any elimination order including X_3 or X_6 is incorrect. Otherwise, we essentially just walk up the tree:

$$P(Y \mid x_5) \propto P(Y, x_5) = \sum_{x_1,x_2,x_3,x_4,x_6} P(Y, x_1, \ldots, x_6)$$

$$= P(Y) \sum_{x_1,\ldots,x_4,x_6} P(x_1 \mid Y)P(x_2 \mid x_1, Y)P(x_3 \mid x_2, Y)P(x_4 \mid x_2, Y)$$

$$P(x_5 \mid x_4, Y)P(x_6 \mid x_4, Y)$$

$$= P(Y) \sum_{x_1,x_2,x_4} P(x_1 \mid Y)P(x_2 \mid x_1, Y)P(x_4 \mid x_2, Y)P(x_5 \mid x_4, Y)$$

$$\underbrace{\sum_{x_3} P(x_3 \mid x_1, Y) \sum_{x_6} P(x_6 \mid x_4, Y)}_{=1}$$

$$= P(Y) \sum_{x_1} P(x_1 \mid Y) \sum_{x_2} P(x_2 \mid x_1, Y) \sum_{x_4} P(x_4 \mid x_2, Y)P(x_5 \mid x_4, Y)$$

So possible orders include $X_4 \prec X_1 \prec X_2$, $X_4 \prec X_2 \prec X_1$, $X_1 \prec X_2 \prec X_4$, and $X_1 \prec X_4 \prec X_2$. Any order must start with one of X_4 or X_1, then so long as X_2 follows one of them it is correct.

d. Given that X_6 is a child node of X_4, it marginalizes to 1 and has no effect on inference. So any elimination ordering including X_6 is incorrect. Other than that, we can explicitly compute the elimination by marginalizing, and we get

$$P(X_3 \mid x_5) \propto P(X_3, x_5)$$

$$= \sum_{x_1,x_2,x_4,y} P(y)P(x_1 \mid y)P(x_2 \mid x_1, y)P(X_3 \mid x_2, y)P(x_5 \mid x_4, y)P(x_4 \mid x_2, y)$$

$$= \sum_{y} P(y) \sum_{x_2} P(X_3 \mid x_2, y) \sum_{x_1} P(x_1 \mid y)P(x_2 \mid x_1, y) \sum_{x_4} P(x_5 \mid x_4, y)P(x_4 \mid x_2, y).$$

The maximum factor size is 3; one such example above is (X_4, X_2, Y). So one possible ordering is $X_4 \prec X_1 \prec X_2 \prec Y \prec X_3$. The possible orders must have one of X_4 and X_1 first, which yields a factor over (X_2, Y) as well as factors (X_2, Y, X_3) and (Y). Then any ordering of X_2, Y will be fine. X_3 must clearly be last. Note that eliminating Y early or X_2 before eliminating one of X_4 and X_1 will yield factors of size 4, so that is incorrect.

e. No, it does not really make sense to perform Gibbs sampling. Inference is always efficient—factors are of size at most 3—because everything is a tree given the node Y.

Exercise 13.NUKE

In your local nuclear power station, there is an alarm that senses when a temperature gauge exceeds a given threshold. The gauge measures the temperature of the core. Consider the Boolean variables A (alarm sounds), F_A (alarm is faulty), and F_G (gauge is faulty) and the multivalued nodes G (gauge reading) and T (actual core temperature).

a. Draw a Bayesian network for this domain, given that the gauge is more likely to fail when the core temperature gets too high.

b. Is your network a polytree? Why or why not?

c. Suppose there are just two possible actual and measured temperatures, normal and high; the probability that the gauge gives the correct temperature is x when it is working, but y when it is faulty. Give the conditional probability table associated with G.

d. Suppose the alarm works correctly unless it is faulty, in which case it never sounds. Give the conditional probability table associated with A.

e. Suppose the alarm and gauge are working and the alarm sounds. Calculate an expression for the probability that the temperature of the core is too high, in terms of the various conditional probabilities in the network.

This question exercises many aspects of the student's understanding of Bayesian networks and uncertainty.

a. A suitable network is shown in Figure S13.41. The key aspects are: the failure nodes are parents of the sensor nodes, and the temperature node is a parent of both the gauge and the gauge failure node. It is exactly this kind of correlation that makes it difficult for humans to understand what is happening in complex systems with unreliable sensors.

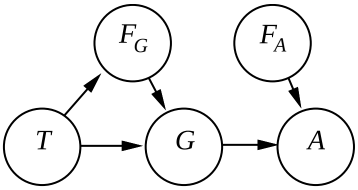


Figure S13.41 A Bayesian network for the nuclear alarm problem.

b. No matter which way the student draws the network, it should not be a polytree because of the fact that the temperature influences the gauge in two ways.

c. The CPT for G is shown below. Students should pay careful attention to the semantics of F_G, which is true when the gauge is *faulty*, i.e., *not* working.

	$T = Normal$		$T = High$	
	F_G	$\neg F_G$	F_G	$\neg F_G$
$G = Normal$	y	x	$1-y$	$1-x$
$G = High$	$1-y$	$1-x$	y	x

d. The CPT for A is as follows:

	$G = Normal$		$G = High$	
	F_A	$\neg F_A$	F_A	$\neg F_A$
A	0	0	0	1
$\neg A$	1	1	1	0

e. This part actually asks the student to do something usually done by Bayesian network algorithms. The great thing is that doing the calculation without a software package makes it easy to see the nature of the calculations that the algorithms are systematizing. It illustrates the magnitude of the achievement involved in creating complete and correct algorithms.

Abbreviating $T = High$ and $G = High$ by T and G, the probability of interest here is $P(T|A, \neg F_G, \neg F_A)$. Because the alarm's behavior is deterministic, we can reason that if the alarm is working and sounds, G must be $High$. Because F_A and A are d-separated from T, we need only calculate $P(T|\neg F_G, G)$.

There are several ways to go about doing this. The "opportunistic" way is to notice that the CPT entries give us $P(G|T, \neg F_G)$, which suggests using the generalized Bayes' Rule to switch G and T with $\neg F_G$ as background:

$$P(T|\neg F_G, G) \propto P(G|T, \neg F_G) P(T|\neg F_G)$$

We then use Bayes' Rule again on the last term:

$$P(T|\neg F_G, G) \propto P(G|T, \neg F_G) P(\neg F_G|T) P(T)$$

A similar relationship holds for $\neg T$:

$$P(\neg T|\neg F_G, G) \propto P(G|\neg T, \neg F_G) P(\neg F_G|\neg T) P(\neg T)$$

Normalizing, we obtain

$$P(T|\neg F_G, G) = \frac{P(G|T,\neg F_G) P(\neg F_G|T) P(T)}{P(G|T,\neg F_G) P(\neg F_G|T) P(T) + P(G|\neg T,\neg F_G) P(\neg F_G|\neg T) P(\neg T)}$$

The "systematic" way to do it is to revert to joint entries (noticing that the subgraph of T, G, and F_G is completely connected so no loss of efficiency is entailed). We have

$$P(T|\neg F_G, G) = \frac{P(T, \neg F_G, G)}{P(G, \neg F_G)} = \frac{P(T, \neg F_G, G)}{P(T, G, \neg F_G) + P(T, G, \neg F_G)}$$

Now we use the chain rule to rewrite the joint entries as CPT entries:

$$P(T|\neg F_G, G) = \frac{P(T) P(\neg F_G|T) P(G|T,\neg F_G)}{P(T) P(\neg F_G|T) P(G|T,\neg F_G) + P(\neg T) P(\neg F_G|\neg T) P(G|\neg T,\neg F_G)}$$

which of course is the same as the expression arrived at above. Letting $P(T) = p$,

$P(F_G|T) = g$, and $P(F_G|\neg T) = h$, we get

$$P(T|\neg F_G, G) = \frac{p(1-g)(1-x)}{p(1-g)(1-x) + (1-p)(1-h)x}$$

Exercise 13.CHEA

Cheating dealers have become a serious problem at the mini-Blackjack tables. A mini-Blackjack deck has 3 card types (5,10,11) and an honest dealer is equally likely to deal each of the 3 cards. When a player holds 11, cheating dealers deal a 5 with probability $\frac{1}{4}$, 10 with probability $\frac{1}{2}$, and 11 with probability $\frac{1}{4}$. You estimate that $\frac{4}{5}$ of the dealers in your casino are honest ($H = \mathit{true}$) while $\frac{1}{5}$ are cheating ($H = \mathit{false}$).

a. You see a dealer deal an 11 to a player holding 11. What is the probability that the dealer is cheating?

The casino has decided to install a camera to observe its dealers. Cheating dealers are observed doing suspicious things on camera ($C = \mathit{true}$) $\frac{4}{5}$ of the time, while honest dealers are observed doing suspicious things $\frac{1}{4}$ of the time.

b. Draw a Bayes net with the variables H (honest dealer), D (card dealt to a player holding 11), and C (suspicious behavior on camera). Write the conditional probability tables.

c. List all conditional independence assertions made by your Bayes net.

d. What is the probability that a dealer is honest given that he deals a 10 to a player holding 11 and is observed doing something suspicious?

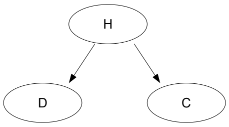
You can either arrest dealers or let them continue working. If you arrest a dealer and he turns out to be cheating, you will earn a \$4 bonus. However, if you arrest the dealer and he turns out to be innocent, he will sue you for -\$10. Allowing the cheater to continue working will cost you -\$2, while allowing an honest dealer to continue working will get you \$1. Assume a linear utility function $U(x) = x$.

e. You observe a dealer doing something suspicious (C) and also observe that he deals a 10 to a player holding 11. Should you arrest the dealer?

f. A private investigator approaches you and offers to investigate the dealer from the previous part. If you hire him, he will tell you with 100% certainty whether the dealer is cheating or honest, and you can then make a decision about whether to arrest him or not. How much would you be willing to pay for this information?

a.

$$\begin{aligned} P(\neg H \mid D=11) &= \tfrac{P(\neg H, D=11)}{P(D=11)} \\ &= \tfrac{P(D=11 \mid \neg H)P(\neg H)}{P(D=11 \mid \neg H)P(\neg H) + P(D=11 \mid H)P(H)} \\ &= \tfrac{(1/4)(2/10)}{(1/4)(2/10)+(1/3)(8/10)} = 3/19 \end{aligned}$$



b.

H	$P(H)$
true	0.8
false	0.2

H	D	$P(D \mid H)$
true	5	1/3
true	10	1/3
true	11	1/3
false	5	1/4
false	10	1/2
false	11	1/4

H	C	$P(C \mid H)$
true	true	1/4
true	false	3/4
false	true	4/5
false	false	1/5

c. The Bayes net asserts only $D \perp\!\!\!\perp C \mid H$.

d.

$$\begin{aligned} P(h \mid D{=}10, c) &= \frac{P(h, D{=}10, c)}{P(D{=}10, c)} \\ &= \frac{P(h)P(D{=}10 \mid h)p(C \mid H)}{P(h)P(D{=}10 \mid h)P(c \mid h) + P(\neg h)P(D{=}10 \mid \neg h)P(c \mid \neg h)} \\ &= \frac{(4/5)(1/3)(1/4)}{(4/5)(1/3)(1/4) + (1/5)(1/2)(4/5)} = \frac{5}{11} \end{aligned}$$

e. This question is a simple preview of material from Chapter 15, but we have already seen the basic idea in the context of backgammon and other game of chance in Chapter **??**.

Arresting the dealer yields an expected payoff of

$4{*}P(\neg H \mid D = 10, C){+}(-10){*}P(H \mid D = 10, C) = 4(6/11){+}(-10)(5/11) = -26/11$ /

Letting him continue working yields a payoff of

$(-2){*}P(\neg H \mid D = 10, C){+}1{*}P(H \mid D = 10, C) = (-2)(6/11){+}(1)(5/11) = -7/11$.

Therefore, you should let the dealer continue working.

f. This question addresses the theory of information value from Chapter 15. If used prior to students studying that material, it might be a good idea to provide a hint.
If you hire the private investigator, if the dealer is a cheater you can arrest him for a payoff of $4. If he is an honest dealer you can let him continue working for a payoff of

\$1. The benefit from hiring the investigator is therefore

$$(4) * P(\neg h \,|\, D\!=\!10, c) + 1 * P(h \,|\, D\!=\!10, c) = 4(6/11) + (1)(5/11) = 29/11 \,.$$

If you do not hire the investigator, your best course of action is to let the dealer continue working for an expected payoff of $-7/11$. Therefore, you are willing to pay up to $29/11 - (-7/11) = 36/11$ to hire the investigator.

Exercise 13.TELC

Consider the network shown in Figure S13.11(ii), and assume that the two telescopes work identically. $N \in \{1,2,3\}$ and $M_1, M_2 \in \{0,1,2,3,4\}$, with the symbolic CPTs as described in Exercise TELESCOPE-EXERCISE. Using the enumeration algorithm (Figure 13.11 on page 447), calculate the probability distribution $\mathbf{P}(N \,|\, M_1\!=\!2, M_2\!=\!2)$.

The symbolic expression evaluated by the enumeration algorithm is

$$\begin{aligned}\mathbf{P}(N \,|\, M_1\!=\!2, M_2\!=\!2) &= \alpha \sum_{f_1,f_2} \mathbf{P}(f_1, f_2, N, M_1\!=\!2, M_2\!=\!2) \\ &= \alpha \sum_{f_1,f_2} P(f_1)P(f_2)\mathbf{P}(N)P(M_1\!=\!2 \,|\, f_1, N)P(M_2\!=\!2 \,|\, f_2, N) \,.\end{aligned}$$

Because an out-of-focus telescope cannot report 2 stars in the given circumstances, the only non-zero term in the summation is for $F_1 = F_2 = \mathit{false}$, so the answer is

$$\begin{aligned}\mathbf{P}(N \,|\, M_1\!=\!2, M_2\!=\!2) &= \alpha(1-f)(1-f)\langle p_1, p_2, p_3\rangle\langle e, (1-2e), e\rangle\langle e, (1-2e), e\rangle \\ &= \alpha'\langle p_1 e^2, p_2(1-2e)^2, p_3 e^2\rangle \,.\end{aligned}$$

Exercise 13.VEEX

Consider the variable elimination algorithm in Figure 13.13 (page 450).

a. Section 13.3 applies variable elimination to the query

$$\mathbf{P}(\mathit{Burglary} \,|\, \mathit{JohnCalls} = \mathit{true}, \mathit{MaryCalls} = \mathit{true}) \,.$$

Perform the calculations indicated and check that the answer is correct.

b. Count the number of arithmetic operations performed, and compare it with the number performed by the enumeration algorithm.

c. Suppose a network has the form of a *chain*: a sequence of Boolean variables $X_1, \ldots, X_n$ where $\mathit{Parents}(X_i) = \{X_{i-1}\}$ for $i = 2, \ldots, n$. What is the complexity of computing $\mathbf{P}(X_1 \,|\, X_n = \mathit{true})$ using enumeration? Using variable elimination?

d. Prove that the complexity of running variable elimination on a polytree network is linear in the size of the tree for any variable ordering consistent with the network structure.

This question definitely helps students get a solid feel for variable elimination. Students may need some help with the last part if they are to do it properly.

a.

$$
\begin{aligned}
& P(B|j,m) \\
&= \alpha P(B) \sum_e P(e) \sum_a P(a|b,e)P(j|a)P(m|a) \\
&= \alpha P(B) \sum_e P(e) \left[.9 \times .7 \times \begin{pmatrix} .95 & .29 \\ .94 & .001 \end{pmatrix} + .05 \times .01 \times \begin{pmatrix} .05 & .71 \\ .06 & .999 \end{pmatrix} \right] \\
&= \alpha P(B) \sum_e P(e) \begin{pmatrix} .598525 & .183055 \\ .59223 & .0011295 \end{pmatrix} \\
&= \alpha P(B) \left[.002 \times \begin{pmatrix} .598525 \\ .183055 \end{pmatrix} + .998 \times \begin{pmatrix} .59223 \\ .0011295 \end{pmatrix} \right] \\
&= \alpha \begin{pmatrix} .001 \\ .999 \end{pmatrix} \times \begin{pmatrix} .59224259 \\ .001493351 \end{pmatrix} \\
&= \alpha \begin{pmatrix} .00059224259 \\ .0014918576 \end{pmatrix} \\
&\approx \langle .284, .716 \rangle
\end{aligned}
$$

b. Including the normalization step, there are 7 additions, 16 multiplications, and 2 divisions. The enumeration algorithm has two extra multiplications.

c. To compute $\mathbf{P}(X_1|X_n = true)$ using enumeration, we have to evaluate two complete binary trees (one for each value of X_1), each of depth $n-2$, so the total work is $O(2^n)$. Using variable elimination, the factors never grow beyond two variables. For example, the first step is

$$
\begin{aligned}
& \mathbf{P}(X_1|X_n = true) \\
&= \alpha \mathbf{P}(X_1) \ldots \sum_{x_{n-2}} P(x_{n-2}|x_{n-3}) \sum_{x_{n-1}} P(x_{n-1}|x_{n-2}) P(X_n = true|x_{n-1}) \\
&= \alpha \mathbf{P}(X_1) \ldots \sum_{x_{n-2}} P(x_{n-2}|x_{n-3}) \sum_{x_{n-1}} \mathbf{f}_{X_{n-1}}(x_{n-1}, x_{n-2}) \mathbf{f}_{X_n}(x_{n-1}) \\
&= \alpha \mathbf{P}(X_1) \ldots \sum_{x_{n-2}} P(x_{n-2}|x_{n-3}) \mathbf{f}_{\overline{X_{n-1}}X_n}(x_{n-2})
\end{aligned}
$$

The last line is isomorphic to the problem with $n-1$ variables instead of n; the work done on the first step is a constant independent of n, hence (by induction on n, if you want to be formal) the total work is $O(n)$.

d. Here we can perform an induction on the number of nodes in the polytree. The base case is trivial. For the inductive hypothesis, assume that any polytree with n nodes can be evaluated in time proportional to the size of the polytree (i.e., the sum of the CPT sizes). Now, consider a polytree with $n+1$ nodes. Any node ordering consistent with the topology will eliminate first some leaf node from this polytree. To eliminate any leaf node, we have to do work proportional to the size of its CPT. Then, *because the network is a polytree*, we are left with *independent* subproblems, one for each parent. Each subproblem takes total work proportional to the sum of its CPT sizes, so the total work for $n+1$ nodes is proportional to the sum of CPT sizes.

Exercise 13.VELT

Assume we are running variable elimination, and we currently have the following three factors:

A	B	$f_1(A,B)$
true	*true*	0.1
true	*false*	0.5
false	*true*	0.2
false	*false*	0.5

A	C	D	$f_2(A,C,D)$
true	*true*	*true*	0.2
true	*true*	*false*	0.1
true	*false*	*true*	0.5
true	*false*	*false*	0.1
false	*true*	*true*	0.5
false	*true*	*false*	0.2
false	*false*	*true*	0.5
false	*false*	*false*	0.2

B	D	$f_3(B,D)$
true	*true*	0.2
true	*false*	0.2
false	*true*	0.5
false	*false*	0.1

The next step in the variable elimination is to eliminate B.

a. Which factors will participate in the elimination process of B?

b. Perform the join over the factors that participate in the elimination of B. Your answer should be a table similar to the tables above, it is your job to figure out which variables participate and what the numerical entries are.

c. Now perform the summation over B for the factor you obtained from the join and show the factor that results.

a. f_1, f_3

b.

A	B	D	$f_4'(A,B,D)$
true	*true*	*true*	$0.1 \times 0.2 = 0.02$
true	*true*	*false*	$0.1 \times 0.2 = 0.02$
true	*false*	*true*	$0.5 \times 0.5 = 0.25$
true	*false*	*false*	$0.5 \times 0.1 = 0.05$
false	*true*	*true*	$0.2 \times 0.2 = 0.04$
false	*true*	*false*	$0.2 \times 0.2 = 0.04$
false	*false*	*true*	$0.5 \times 0.5 = 0.25$
false	*false*	*false*	$0.5 \times 0.1 = 0.05$

c.

A	D	$f_4(A,D)$
true	*true*	$0.02 + 0.25 = 0.27$
true	*false*	$0.02 + 0.05 = 0.07$
false	*true*	$0.04 + 0.25 = 0.29$
false	*false*	$0.04 + 0.05 = 0.09$

Exercise 13.BPAH

For the Bayes net shown in Figure S13.42, consider the query $P(A|h)$, and the variable elimination ordering B, E, C, F, D.

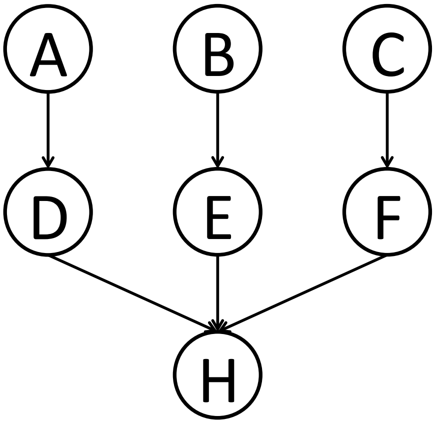


Figure S13.42 A Bayes network for Exercise 13.BPAH.

a. In the table below fill in the factor generated at each step—we did the first row for you.

Variable Eliminated	Factor Generated	Current Factors
(none)	(none)	$P(A), P(B), P(C), P(D\|A), P(E\|B), P(F\|C),$ $P(h\|D, E, F)$
B	$f_1(E)$	$P(A), P(C), P(D\|A), P(F\|C), P(h\|D, E, F), f_1(E)$
E		
C		
F		
D		

b. Which is the largest factor generated? Assuming all variables are binary, how many entries does the corresponding table have?

Variable Eliminated	Factor Generated	Current Factors
(none)	(none)	$P(A), P(B), P(C), P(D\|A), P(E\|B), P(F\|C),$ $P(h\|D, E, F)$
B	$f_1(E)$	$P(A), P(C), P(D\|A), P(F\|C), P(h\|D, E, F), f_1(E)$
E	$f_2(h, D, F)$	$P(A), P(C), P(D\|A), P(F\|C), f_2(h, D, F)$
C	$f_3(F)$	$P(A), P(D\|A), f_2(h, D, F), f_3(F)$
F	$f_4(h, D)$	$P(A), P(D\|A), f_4(h, D)$
D	$f_5(h, A)$	$P(A), f_5(h, A)$

a.

b. The largest factor is $f_2(h, D, F)$, which has $2^2 = 4$ entries.

Exercise 13.BNCM

Investigate the complexity of exact inference in general Bayesian networks:

a. Prove that any 3-SAT problem can be reduced to exact inference in a Bayesian network constructed to represent the particular problem and hence that exact inference is NP-hard. (*Hint*: Consider a network with one variable for each proposition symbol, one for each clause, and one for the conjunction of clauses.)

b. The problem of counting the number of satisfying assignments for a 3-SAT problem is #P-complete. Show that exact inference is at least as hard as this.

a. Consider a 3-CNF formula $C_1 \wedge \ldots C_n$ with n clauses where each clause is a disjunct $C_i = (\ell_{i1} \vee \ell_{i2} \vee \ell_{i3})$ of literals i.e., each ℓ_{ij} is either P_k or $\neg P_k$ for some atomic proposition $P_1, \ldots, P_m$.

Construct a Bayesian network with a (boolean) variable S for the whole formula, C_i for each clause, and P_k for each atomic proposition. We will define parents and CPTs such that for any assignment to the atomic propositions, S is true if and only if the 3-CNF formula is true.

Atomic propositions have no parents, and are true with probability 0.5. Each clause C_i has as its parents the atomic propositions corresponding to the literals ℓ_{i1}, ℓ_{i2}, and ℓ_{i3}). The clause variable is true iff one of its literals is true. Note that this is a deterministic CPT. Finally, S has all the clause variables C_i as its parents, and if true if any only if all clause variables are true.

Notice that $P(S = True) > 0$ if and only if the formula is satisfiable, and exact inference will answer this question.

b. Using the same network as in part (a), notice that $P(S = True) = s2^{-m}$ where s is the number of satisfying assignments to the atomic propositions $P_1, \ldots, P_m$.

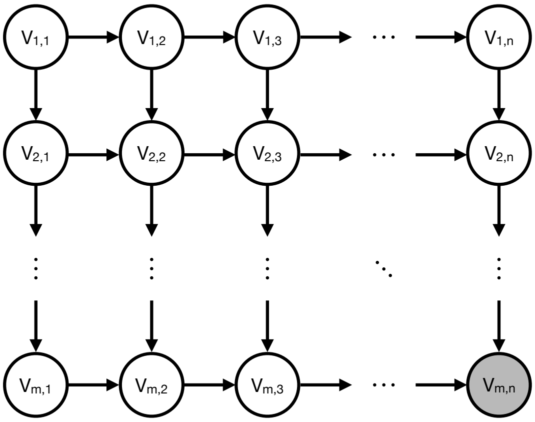


Figure S13.43 A Bayes network for Exercise 13.LATT.

Exercise 13.LATT

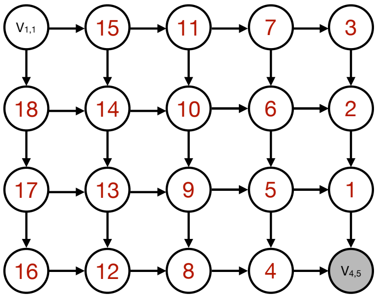
Consider doing inference in an m x n lattice Bayes net, as shown in Figure S13.43. The network consists of mn binary variables $V_{i,j}$, and you have observed that $V_{m,n} = +v_{m,n}$.

You wish to calculate $P(V_{1,1} \mid +v_{m,n})$ using variable elimination. To maximize computational efficiency, you wish to use a variable elimination ordering for which the size of the largest generated factor is as small as possible.

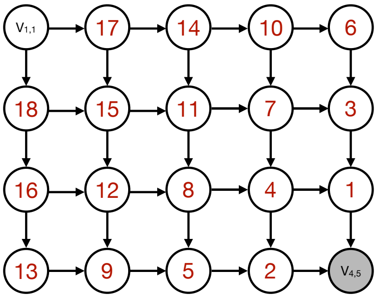
a. First consider the special case where $m = 4$ and $n = 5$. What is the optimal elimination order?

b. Now consider the general case (assume $m > 2$ and $n > 2$). What is the size of the largest factor generated under the most efficient elimination ordering?

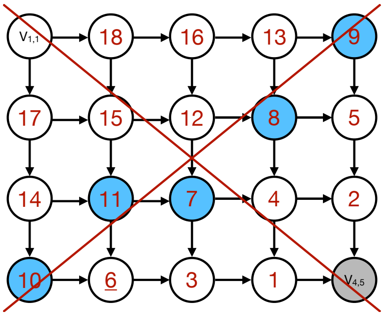
a. There are several possible orderings, here are two:



(or)



Minor variations on the orderings given above are possible. However, it's important to start near the same corner as the evidence variable and to never create a factor that involves more than 4 non-evidence variables. For example, the ordering shown below is suboptimal (eliminating node 6 will create a size 2^5 factor involving the five nodes highlighted in blue):



b. The largest factor should have $\min(m, n)$ variables and $2^{\min(m,n)}$ table entries.

Exercise 13.CUTS

a. Consider answering $P(H \mid +f)$ by variable elimination in the Bayes nets N and N' shown in Figure S13.44, where the elimination order is alphabetical and all variables

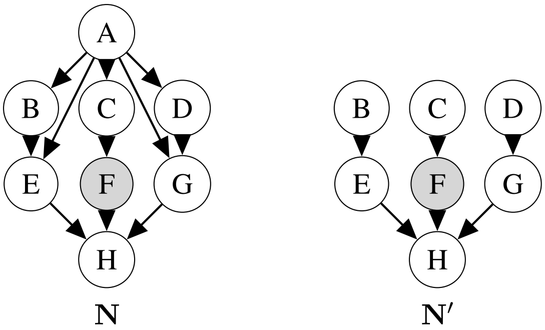


Figure S13.44 Bayes nets for Exercise 13.CUTS.

are binary. How large are the largest factors made during variable elimination for N and N'?

b. Borrowing an idea from **cutset conditioning** in Chapter 5, we may be able to simplify variable elimination in N by instantiating some variables. Let's pick an **instantiation set** to pretend to observe, and then do variable elimination with these additional instantiations.

Consider the original query $P(H \mid +f)$, but let A be the instantiation set so $A = a$ is observed. Now the query is H with observations $F = +f, A = a$.

(i) What is the size of the largest factor made during variable elimination with the $A = a$ instantiation?

(ii) Given a Bayes net over n binary variables with k variables chosen for the instantiation set, how many instantiations of the set are there?

c. Let's answer $P(H \mid +f)$ by variable elimination with the instantiations of A.

(i) What quantity does variable elimination for $P(H \mid +f)$ with the $A = +a$ instantiation compute *without normalization*? That is, what is the last factor made by elimination?

(ii) Let $I_+(H) = F(H, +a, +f)$ and $I_-(H) = F(H, -a, +f)$ be the last factors made by variable elimination with instantiations $A = +a$ and $A = -a$. How do we calculate the desired answer $P(+h \mid +f)$?

d. What is the time complexity of instantiated elimination, expressed in terms of n (the number of variables), k (the instantiation set size), f (the dimension of the largest factor made by elimination without instantiation), and i (the dimension of the largest factor made by elimination with instantiation)?

a. (i) The largest factor made during variable elimination for N is $f(B,C,D,E,G)$ after eliminating A. It has $2^5 = 32$ entries.

(ii) The largest factor made during variable elimination for N' is $f(G,H,+f)$ after eliminating E. It has $2^2 = 4$ entries.

b. (i) The largest factor made during variable elimination with the $A = a$ instantiation is

$f(G,H,+f)$ as in N' in the previous question.

(ii) For a selected instantiation set of k binary variables, there are 2^k total instantiations of these variables.

c. (i) The last factor from variable elimination for $P(H \mid +f)$ with $A=+a$ is $P(H,+a,+f)$. At the end of variable elimination, the last factor is equal to the equivalent entries of the joint distribution with the eliminated variables summed out and the selected values of the evidence variables.

(ii) $P(+h \mid +f) = \frac{I_+(+h)+I_-(+h)}{\sum_h I_+(h)+I_-(h)}$. The last factors are the entries from the corresponding joint $I_+(+h)=p(+h,+a,+f)$ and $I_-(+h)=p(+h,-a,+f)$ and so on. By the law of total probability $P(+h,+f)=P(+h,+a,+f)+P(+h,-a,+f)$, so the joint of the original query and evidence can be computed from the instantiated elimination factors. For the conditional $p(+h \mid +f)=P(+h,+f)/P(+f)$, normalize by the sum over the query $\sum_h p(h,+f) = \sum_h \sum_a p(h,a,+f) = f(+h,+a,+f)+f(+h,-a,+f)+f(-h,+a,+f)+f(-h,-a,+f)$ where the joint over the query and evidence is again computed from the law of total probability over A.

d. The time complexity of instantiated elimination is $O(n\exp(i+k))$. To carry out instantiated elimination, we have to do variable elimination $\exp(k)$ times for all the settings of the instantiation set. Each of these eliminations takes time bounded by $n\exp(i)$ as the largest factor is the most expensive to eliminate and there are n variables to eliminate. If the instantiation set is not too large and the size of the factors made by instantiation elimination are small enough this method can be exponentially faster than regular elimination. The catch is how to select the instantiation set.

Exercise 13.WMCX

The idea behind weighted model counting (WMC) is to use the machinery of SAT-solvers to perform probabilistic inference. As noted in Equation (12.9), the answer to a probabilistic query $\mathbf{P}(X \mid \mathbf{e}) = \alpha\mathbf{P}(X,\mathbf{e})$ is obtained by normalizing a sum of joint probabilities of the form $P(x,\mathbf{e},\mathbf{y})$ where $\mathbf{y}$ denotes values of the hidden variables. Because X, $\mathbf{E}$, and $\mathbf{Y}$ together constitute all the variables in the Bayes net, the probabilities $P(x,\mathbf{e},\mathbf{y})$ for the possible worlds $x,\mathbf{e},\mathbf{y}$ are just products of conditional probabilities from the network.

A WMC algorithm is a particular kind of SAT-solver that, given a CNF sentence S with weights $W(l)$ associated with each literal l, computes the sum of the weights of all satisfying assignments of S, where the weight of an assignment is the product of the weights of its constituent literals.

a. Suppose $S=[(A \lor B) \land (\neg B \lor \neg C) \land (\neg B \lor C)]$, $W(A)=W(B)=W(C)=0.7$, and $W(\neg A)=W(\neg B)=W(\neg C)=0.3$. Show that a WMC solver, given S and W as inputs, returns 0.21.

b. In order to use the WMC algorithm for inference, you will need to develop a translation from a Bayes net model to a weighted CNF representation. For the purposes of this exercise, you may assume that all variables in the Bayes net are Boolean, so they can be mapped directly into propositional variables in the CNF sentence. In addition, for every variable V with parents $\mathbf{U}$, introduce a Boolean variable $R_{v|\mathbf{u}}$ for each possible

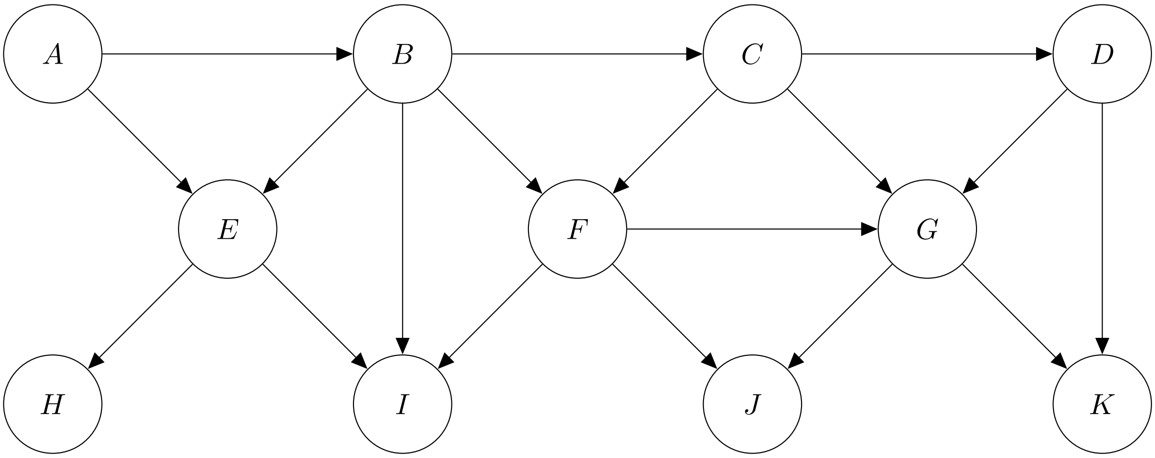


Figure S13.45 Bayes net for Exercise 13.BLNK.

value v of V and every combination of values $\mathbf{u}$ for $\mathbf{U}$. Now, given a query variable X and evidence $\mathbf{e}$, show how to define a CNF sentence S and weights $W(l)$ such that the WMC solver, given S and W, returns a number equal to $P(X = true, \mathbf{e})$. Prove that your definition yields the correct answer.

c. Implementation project:

(i) Identify a suitable Bayes net input format from one of the standard Bayes net packages, and find an efficient WMC solver.

(ii) Implement the translation algorithm to convert any given Boolean-variable Bayes net, evidence set, and query into weighted CNF.

(iii) Compare the computation time and memory requirements of exact inference in the Bayes net package versus running the WMC solver, for a range of networks, evidence sets, and queries. Comment on your results.

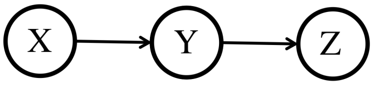
The discovery of a suitable implementation is left for interested readers.

13.4 Approximate Inference for Bayesian Networks

Exercise 13.BLNK

In the network in Figure S13.45, identify the Markov blanket of each variable.

A : B, E; **B** : A, E, F, C; **C** : B, F, G, D; **D** : C, G, K; **E** : A, B, H, I; **F** : B, C, G, I, J; **G** : C, D, F, J, K; **H** : E; **I** : B, E, F; **J** : F, G; **K** : D, G



X	$\mathbf{P}(X)$
$+x$	$2/5$
$-x$	$3/5$

Y	X	$\mathbf{P}(Y \mid X)$
$+y$	$+x$	2/3
$-y$	$+x$	1/3
$+y$	$-x$	3/4
$-y$	$-x$	1/4

Z	Y	$\mathbf{P}(Z \mid Y)$
$+z$	$+y$	1/3
$-z$	$+y$	2/3
$+z$	$-y$	1/5
$-z$	$-y$	4/5

Figure S13.46 Bayes net and distributions for Exercise 13.XYZS.

Exercise 13.XYZS

Assume the Bayes net, and the corresponding distributions over the variables in the Bayes net from Figure S13.46.

a. Your task is now to estimate $\mathbf{P}(+y \mid +x, +z)$ using rejection sampling. Below are some samples that have been produced by prior sampling (that is, the rejection stage in rejection sampling hasn't happened yet). Which of the following samples would be rejected by rejection sampling?

(i) $+x$, $+y$, $+z$
(ii) $-x$, $+y$, $+z$
(iii) $-x$, $-y$, $+z$
(iv) $+x$, $-y$, $-z$
(v) $+x$, $-y$, $+z$

b. Using rejection sampling, give an estimate of $\mathbf{P}(+y \mid +x, +z)$ from the five prior samples, or state why it cannot be computed.

c. Using the following samples (which were generated using likelihood weighting), estimate
$\mathbf{P}(+y \mid +x, +z)$ using likelihood weighting, or state why it cannot be computed.

(i) $+x$, $+y$, $+z$
(ii) $+x$, $-y$, $+z$
(iii) $+x$, $+y$, $+z$

d. Given a sequence of samples, each specifying a value for all the variables in the network, how can one tell if the sequence *could* have been generated by Gibbs sampling?

a. Samples ii, iii, iv would be rejected.

b. Of the two remaining samples, one has $y+$, so the estimate is $1/2$.

c. The three weights are as follows: $w_1 = w_3 = P(+x) * P(+z \mid +y) = 2/5 * 1/3 = 2/15$ and $w_2 = P(+x) * P(+z \mid -y) = 2/5 * 1/5 = 2/25$. Hence $\mathbf{P}(+y \mid +x, +z) \approx (2w_1)/(2w_1 + w_2) = 10/13$.

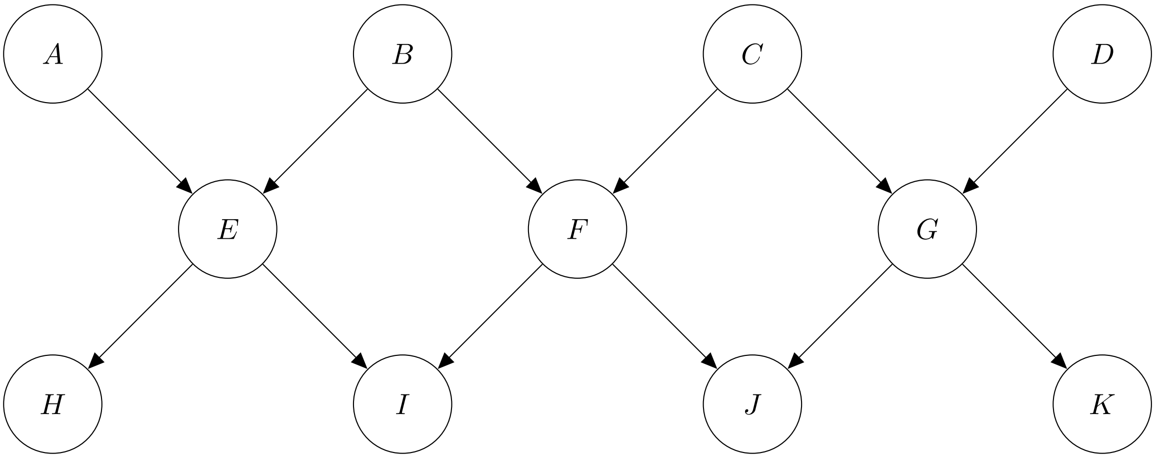


Figure S13.47 Bayes net for Exercise 13.HIJK

d. $\mathbf{P}(Z \mid X)$ is better, because evidence influences the choice of downstream variables, but not upstream variables, and X is a (grand)parent of Z.

e. First, each sample can differ from the preceding sample in the value of only one variable. Second, each transition has to correspond to a non-zero value for the Gibbs distribution. This can be checked by seeing which variable changed (if any) and computing the Gibbs distribution for that variable given its Markov blanket.

Exercise 13.HIJK

We are running Gibbs sampling in the Bayes net shown in Figure S13.47 for the query $P(B,C \mid +h,+i,+j)$. The current state is $+a,+b,+c,+d,+e,+f,+g,+h,+i,+j,+k$. Write out an expression for the Gibbs sampling distribution for each of A, F, and K in terms of conditional probabilities available in the network.

This question requires identifying the Markov blanket and then a straightforward application of Equation (13.10). For A, we have

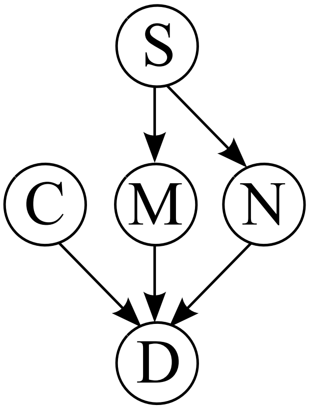
$$P(A \mid +b,+e) \propto P(+e \mid A,+b)P(+b)\ .$$

For F, we have

$$P(F \mid +b,+c,+e,+g,+i,+j) \propto P(F \mid +b,+c)P(+i \mid +e,F)P(+j \mid F,+g)\ .$$

For K we have just $P(K \mid +g)$.

Exercise 13.DANC



S	C	M	N	D
$-s$	$+c$	$+m$	$-n$	$+d$
$+s$	$+c$	$-m$	$-n$	$-d$
$+s$	$-c$	$-m$	$+n$	$-d$
$+s$	$+c$	$+m$	$-n$	$+d$
$+s$	$+c$	$-m$	$+n$	$+d$
$+s$	$-c$	$-m$	$+n$	$-d$
$+s$	$-c$	$-m$	$-n$	$-d$
$+s$	$+c$	$+m$	$+n$	$+d$
$+s$	$+c$	$-m$	$+n$	$-d$
$-s$	$-c$	$-m$	$-n$	$-d$

S	C	M	N	D	weight
$-s$	$+c$	$+m$	$+n$	$-d$	
$-s$	$+c$	$-m$	$-n$	$-d$	
$-s$	$-c$	$-m$	$+n$	$-d$	
$-s$	$+c$	$+m$	$+n$	$-d$	

Figure S13.48 A Bayes net, CPT, and samples for Exercise 13.DANC.

You are an exobiologist, studying the wide range of life in the universe. You are also an avid dancer and have an excellent model of the way species invent dancing. The key variables are:

Sound sensing (S): Whether or not a species has the ability to sense sound
Cold climate (C): Whether or not the native planet of the species has cold weather
Music (M): Whether or not the species invented music
Non-verbal communication (N): Whether or not the species has any form of non-verbal communication

You model the relationships between these variables and **dancing (D)** using the Bayes net specified in Figure S13.48.

You want to know how likely it is for a dancing, sound-sensing species to invent music, according to this Bayes net. You decide to do inference via sampling. You use prior sampling to draw the samples in Figure S13.48.

a. Based on rejection sampling using the samples above, what is the answer to your query $P(-m \mid +d, +s)$?

While your sampling method has worked fairly well in many cases, for rare cases (like species that can't sense sound) your results are less accurate as rejection sampling rejects almost all of the data. You decide to use likelihood weighting instead. The conditional probabilities of the Bayes net are listed in Figure S13.49.

You now wish to compute the probability that a species that has no sound-sensing $(-s)$ or dancing $(-d)$ nonetheless has music $(+m)$, using likelihood weighting. I.e., you want $P(+m \mid -s, -d)$.

b. You draw the samples in Figure S13.48, using likelihood weighting. For each of these samples, indicate its weight.

C	M	N	D	P(D | C , M , N)
$+c$	$+m$	$+n$	$+d$	0.9
$+c$	$+m$	$+n$	$-d$	0.1
$+c$	$+m$	$-n$	$+d$	0.8
$+c$	$+m$	$-n$	$-d$	0.2
$+c$	$-m$	$+n$	$+d$	0.8
$+c$	$-m$	$+n$	$-d$	0.2
$+c$	$-m$	$-n$	$+d$	0.2
$+c$	$-m$	$-n$	$-d$	0.8
$-c$	$+m$	$+n$	$+d$	0.8
$-c$	$+m$	$+n$	$-d$	0.2
$-c$	$+m$	$-n$	$+d$	0.5
$-c$	$+m$	$-n$	$-d$	0.5
$-c$	$-m$	$+n$	$+d$	0.6
$-c$	$-m$	$+n$	$-d$	0.4
$-c$	$-m$	$-n$	$+d$	0.1
$-c$	$-m$	$-n$	$-d$	0.9

S	M	P(M | S)
$+s$	$+m$	0.8
$+s$	$-m$	0.2
$-s$	$+m$	0.1
$-s$	$-m$	0.9

S	P(S)
$+s$	0.9
$-s$	0.1

S	N	P(N | S)
$+s$	$+n$	0.7
$+s$	$-n$	0.3
$-s$	$+n$	0.9
$-s$	$-n$	0.1

C	P(C)
$+c$	0.5
$-c$	0.5

Figure S13.49 CPTs for dancing aliens.

c. Compute the answer to your query, $P(+m \mid -s, -d)$, using likelihood weighting with these samples.

a. The probability estimate is 1/3. Simply find all the rows in the table above with $+d$ and $+s$, and count the number of times $-m$ occurs divided by the total number of such rows.

b.

S	C	M	N	D	weight
$-s$	$+c$	$+m$	$+n$	$-d$	**0.01**
$-s$	$+c$	$-m$	$-n$	$-d$	**0.08**
$-s$	$-c$	$-m$	$+n$	$-d$	**0.04**
$-s$	$+c$	$+m$	$+n$	$-d$	**0.01**

c. $P(+m \mid -s, -d) \approx 1/7$; divide the sum of the weights corresponding to the $+m$ rows by the total sum of the weights.

Exercise 13.MONY

Alice, Bob, Carol, and Dave are being given some money, but they have to share it in a very particular way:

- First, Alice will be given an integer number of dollars A, chosen uniformly at random between 1 and 100 (inclusive).

- Then Bob receives from Alice an integer number of dollars B, chosen uniformly at random between 1 and A (inclusive).
- Then Carol receives from Bob an integer number of dollars C, chosen uniformly at random between 1 and B (inclusive).
- Then Dave receives from Carol an integer number of dollars D, chosen uniformly at random between 1 and C (inclusive).

a. Draw the Bayes net with variables A, B, C, D that corresponds to this process.

b. Write down an exact formula for $P(B=b)$, the probability that Bob receives b dollars. (Your answer may contain a summation.)

c. Show that $P(A=a \mid B=b)$ is proportional to $1/a$.

d. Suppose Dave receives \$5 from Carol. You would like to estimate the probability $P(A > 50 \mid D=5)$. Your first attempt uses rejection prior sampling, generating the following samples from the prior:

$$(A=52, B=21, C=10, D=5)$$
$$(A=34, B=21, C=6, D=3)$$
$$(A=96, B=48, C=12, D=2)$$
$$(A=13, B=12, C=10, D=1)$$
$$(A=54, B=12, C=11, D=6)$$
$$(A=91, B=32, C=31, D=29)$$

What is the estimated probability for the query based on these samples?

e. Your next attempt uses likelihood weighting, generating the following samples:

$$(S = 52, J = 21, B = 10, X = 5)$$
$$(S = 34, J = 21, B = 4, X = 5)$$
$$(S = 87, J = 12, B = 10, X = 5)$$
$$(S = 41, J = 12, B = 5, X = 5)$$
$$(S = 91, J = 32, B = 4, X = 5)$$

Calculate the weights for each sample and find the estimated value for the query.

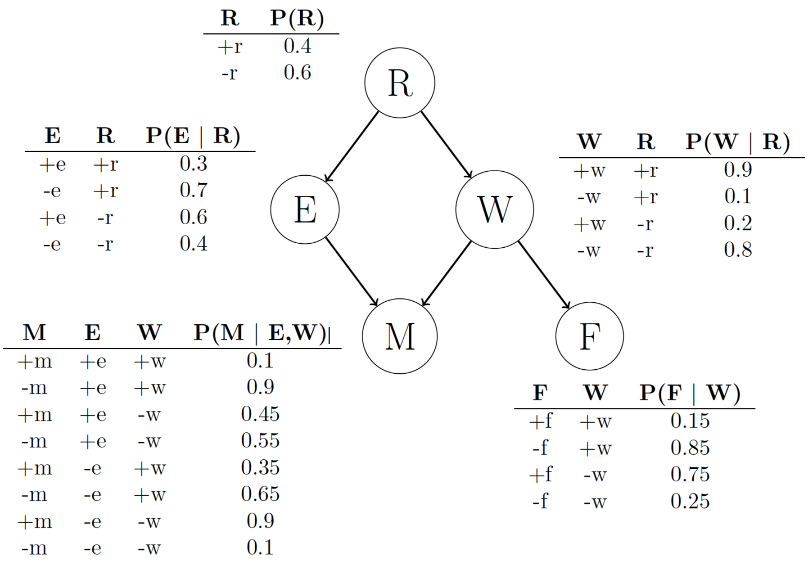
a. The Bayes net is $A \rightarrow B \rightarrow C \rightarrow D$.

b. $P(B=b) = \sum_{a=1}^{100} P(b \mid a)P(a) = \frac{1}{100}\sum_{a=b}^{100} P(b \mid a) = \frac{1}{100}\sum_{a=b}^{100} \frac{1}{a}$. For the mathematically inlcined, this summation has a closed-form solution, giving $\frac{1}{100}(\psi(101) - \psi(b))$ where ψ is the digamma function.

c. Using Bayes' rule, $P(a \mid b) = \alpha P(b \mid a)P(a) = \alpha\frac{1}{100a}$ for $b \leq a \leq 100$ and 0 otherwise.

d. Only the first sample has $D=5$. All other samples are rejected. In that sample, $A > 50$, so the estimated probability is 1.

e. The weights are 0.1, 0.0, 0.1, 0.2, 0.0. The total weight of samples with $A > 50$ is 0.2, out of a total weight for all samples of 0.4, so the estimate is 0.5.



R	P(R)
+r	0.4
-r	0.6

E	R	P(E | R)
+e	+r	0.3
-e	+r	0.7
+e	-r	0.6
-e	-r	0.4

W	R	P(W | R)
+w	+r	0.9
-w	+r	0.1
+w	-r	0.2
-w	-r	0.8

M	E	W	P(M | E,W)|
+m	+e	+w	0.1
-m	+e	+w	0.9
+m	+e	-w	0.45
-m	+e	-w	0.55
+m	-e	+w	0.35
-m	-e	+w	0.65
+m	-e	-w	0.9
-m	-e	-w	0.1

F	W	P(F | W)
+f	+w	0.15
-f	+w	0.85
+f	-w	0.75
-f	-w	0.25

Figure S13.50 A Bayes net and CPTs for sampling.

Exercise 13.SMPL

Consider the Bayes net and corresponding probability tables shown in Figure S13.50.

Fill in the following table with the probabilities of drawing each respective sample given that we are using each of the following sampling techniques. For rejection sampling, we say that a sample has been drawn only if it is not rejected. You may leave your answer in the form of an expression such as $.8 \cdot .4$ without multiplying it out. (Hint: $P(f, m) = .181$)

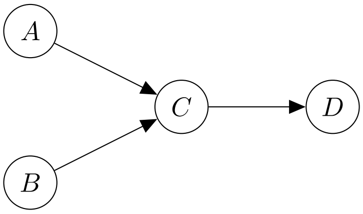
$P(sample\|method)$	$(+r, +e, -w, +m, +f)$	$(+r, -e, +w, -m, +f)$
prior sampling		
rejection sampling		
likelihood weighting		

$P(sample\|method)$	$(+r, +e, -w, +m, +f)$	$(+r, -e, +w, -m, +f)$
prior sampling	$.4 \times .3 \times .1 \times .45 \times .75$ $= .00405$	$.4 \times .7 \times .9 \times .65 \times .15 = 0.02457$
rejection sampling	$\frac{P(+r,+e,-w,+m,+f)}{P(+m,+f)}$ $= \frac{.00405}{.181} = .0224$	0
likelihood weighting	$P(+r)P(+e\|+r)P(-w\|+r)$ $= .4 \times .3 \times .1 = .012$	0

Exercise 13.RMVY

Exercise 13.MRBL askes you to prove that removing an observed variable Y from a Bayes net has no effect on the posterior disribution of any variable X that is outside Y's Markov blanket, provided that every variable in Y's markov blanket is observed. In this question, we consider the effect of removing Y on inference algorithms that estimate the posterior for X from samples. What is the effect of removing Y on (i) rejection sampling and (ii) likelihood weighting? Consider both correctness and efficiency.

Because the posterior is unchanged, and both algorithms are consistent, they will still return the correct answers in the limit of infinitely many samples. Because Y is unobserved, it has no effect on which samples are rejected by rejection sampling, so for any finite sample size the answers will be identical. For importance sampling, we have to consider whether sampling Y affects the answer with a finite sample size. The answer is yes: suppose, for example, that we observe $W = true$, where W is a child of Y, and we know $P(W = true \mid Y = false) = 0$. This means that samples with $Y = false$ get zero weight. It's possible (albeit unlikely) that all N samples happen to choose $Y = false$, so we'd get no useful information about the query at all. Thus, if Y is provably irrelevant to the query, as it is here, importance sampling should *avoid* sampling it because it can only increase the variance of the estimate.



$P(A)$	
+a	0.1
-a	0.9

$P(B)$	
+b	.7
-b	.3

$P(C \mid A, B)$			
+c	+a	+b	.25
-c	+a	+b	.75
+c	-a	+b	.6
-c	-a	+b	.4
+c	+a	-b	.5
-c	+a	-b	.5
+c	-a	-b	.2
-c	-a	-b	.8

$P(D \mid C)$		
+d	+c	.5
-d	+c	.5
+d	-c	.8
-d	-c	.2

Figure S13.51 A Bayes net for sampling

Exercise 13.SMP2

Assume you are given the Bayes net and the corresponding CPTs shown in Figure S13.51.

a. Assume we receive evidence that $A = +a$. If we were to draw samples using rejection sampling, what is the expected fraction of samples that will be **rejected**?

b. Next, assume we observed both $A = +a$ and $D = +d$. What are the weights for the following samples under likelihood weighting sampling?

(i) $(+a, -b, +c, +d)$
(ii) $(+a, -b, -c, +d)$
(iii) $(+a, +b, -c, +d)$

c. Given the samples in the previous question, estimate $P(-b \mid +a, +d)$.

d. Assume we need to (approximately) answer two different inference queries for this graph: $P(C \mid +a)$ and $P(C \mid +d)$. You are required to answer one query using likelihood weighting and one query using Gibbs sampling. Which query would you answer with which algorithm? Justify your answer.

a. Since $P(+a) = \frac{1}{10}$, we would expect that only 10% of the samples could be saved. Therefore, expected 90% of the samples will be rejected.

b. The weights are as follows:

(i) $(+a, -b, +c, +d)$: $P(+a) \cdot P(+d \mid +c) = 0.1 * 0.5 = 0.05$
(ii) $(+a, -b, -c, +d)$: $P(+a) \cdot P(+d \mid -c) = 0.1 * 0.8 = 0.08$
(iii) $(+a, +b, -c, +d)$: $P(+a) \cdot P(+d \mid -c) = 0.1 * 0.8 = 0.08$

c. $P(-b \mid +a, +d) = \frac{P(+a) \cdot P(+d \mid +c) + P(+a) \cdot P(+d \mid -c)}{P(+a) \cdot P(+d \mid +c) + 2 \cdot P(+a) \cdot P(+d \mid -c)} = \frac{0.05+0.08}{0.05+2 \cdot 0.08} = \frac{13}{21}$.

d. Use likelihood weighting for $P(C \mid +a)$ and Gibbs sampling for $P(C \mid +d)$. This is because likelihood weighting only takes upstream evidence into account when sampling. Therefore, Gibbs, which utilizes both upstream and downstream evidence, is more suited to the query $P(C \mid +d)$ which has downstream evidence.

Exercise 13.PRSA

Consider the problem of generating a random sample from a specified distribution on a single variable. Assume you have a random number generator that returns a random number uniformly distributed between 0 and 1.

a. Let X be a discrete variable with $P(X = x_i) = p_i$ for $i \in \{1, \ldots, k\}$. The **cumulative distribution** of X gives the probability that $X \in \{x_1, \ldots, x_j\}$ for each possible j. (See also Appendix A.1.) Explain how to calculate the cumulative distribution in $O(k)$ time and how to generate a single sample of X from it. Can the latter be done in less than $O(k)$ time?

b. Now suppose we want to generate N samples of X, where $N \gg k$. Explain how to do this with an expected run time per sample that is *constant* (i.e., independent of k).

c. Now consider a continuous-valued variable with a parameterized distribution (e.g., Gaussian). How can samples be generated from such a distribution?

d. Suppose you want to query a continuous-valued variable and you are using a sampling algorithm such as LIKELIHOODWEIGHTING to do the inference. How would you have to modify the query-answering process?

a. To calculate the cumulative distribution of a discrete variable, we start from a vector representation p of the original distribution and a vector P of the same dimension. Then, we loop through i, adding up the p_i values as we go along and setting P_i to the running sum, $\sum_{j=i}^{i} p_j$. To sample from the distribution, we generate a random number r uniformly in $[0, 1]$, and then return x_i for the smallest i such that $P_i \geq r$. A naive way to find this is to loop through i starting at 1 until $P_i \geq r$. This takes $O(k)$ time. A more efficient solution is binary search: start with the full range $[1, k]$, choose i at the midpoint of the range. If $P_i < r$, set the range from i to the upper bound, otherwise set the range from the lower bound to i. After $O(\log k)$ iterations, we terminate when the bounds are identical or differ by 1.

b. If we are generating $N \gg k$ samples, we can afford to preprocess the cumulative distribution. The basic insight required is that *if* the original distribution were uniform, it would be possible to sample in $O(1)$ time by returning $\lceil kr \rceil$. That is, we can index directly into the correct part of the range (analog random access, one might say) instead of searching for it. Now, suppose we divide the range $[0, 1]$ into k equal parts and construct a k-element vector, each of whose entries is a list of all those i for which P_i is in the corresponding part of the range. The i we want is in the list with index $\lceil kr \rceil$. We retrieve this list in $O(1)$ time and search through it in order (as in the naive implementation). Let n_j be the number of elements in list j. Then the expected runtime is given by

$$\sum_{j=1}^{k} n_j \cdot 1/k = 1/k \cdot \sum_{j=1}^{k} n_j = 1/k \cdot O(k) = O(1)$$

The variance of the runtime can be reduced by further subdividing any part of the range whose list contains more than some small constant number of elements.

c. One way to generate a sample from a univariate Gaussian is to compute the discretized cumulative distribution (e.g., integrating by Taylor's rule) and use the algorithm described above. We can compute the table once and for all for the standard Gaussian (mean 0, variance 1) and then scale each sampled value z to $\sigma z + \mu$. If we had a closed-form, invertible expression for the cumulative distribution $F(x)$, we could sample exactly, simply by returning $F^{-1}(r)$. Unfortunately the Gaussian density is not exactly integrable. Now, the density $\alpha x e^{-x^2/2}$ *is* exactly integrable, and there are cute schemes for using two samples and this density to obtain an exact Gaussian sample. We leave the details to the interested instructor.

d. When querying a continuous variable using Monte carlo inference, an exact closed-form posterior cannot be obtained. Instead, one typically defines discrete ranges, returning a histogram distribution simply by counting the (weighted) number of samples in each range.

Exercise 13.RAIN

Consider the query $\mathbf{P}(\mathit{Rain} \mid \mathit{Sprinkler} = \mathit{true}, \mathit{WetGrass} = \mathit{true})$ in Figure 13.15(a) (page 453) and how Gibbs sampling can answer it.

a. How many states does the Markov chain have?

b. Calculate the **transition matrix Q** containing $k(\mathbf{y} \rightarrow \mathbf{y}')$ for all $\mathbf{y}$, $\mathbf{y}'$.

c. What does $\mathbf{Q}^2$, the square of the transition matrix, represent?

d. What about $\mathbf{Q}^n$ as $n \rightarrow \infty$?

e. Explain how to do probabilistic inference in Bayesian networks, assuming that $\mathbf{Q}^n$ is available. Is this a practical way to do inference?

a. There are two uninstantiated Boolean variables ($Cloudy$ and $Rain$) and therefore four possible states.

b. First, we compute the sampling distribution for each variable, conditioned on its Markov blanket.

$$\begin{aligned}
\mathbf{P}(C|r,s) &= \alpha\mathbf{P}(C)\mathbf{P}(s|C)\mathbf{P}(r|C)\\
&= \alpha\langle 0.5,0.5\rangle\langle 0.1,0.5\rangle\langle 0.8,0.2\rangle = \alpha\langle 0.04,0.05\rangle = \langle 4/9,5/9\rangle\\
\mathbf{P}(C|\neg r,s) &= \alpha\mathbf{P}(C)\mathbf{P}(s|C)\mathbf{P}(\neg r|C)\\
&= \alpha\langle 0.5,0.5\rangle\langle 0.1,0.5\rangle\langle 0.2,0.8\rangle = \alpha\langle 0.01,0.20\rangle = \langle 1/21,20/21\rangle\\
\mathbf{P}(R|c,s,w) &= \alpha\mathbf{P}(R|c)\mathbf{P}(w|s,R)\\
&= \alpha\langle 0.8,0.2\rangle\langle 0.99,0.90\rangle = \alpha\langle 0.792,0.180\rangle = \langle 22/27,5/27\rangle\\
\mathbf{P}(R|\neg c,s,w) &= \alpha\mathbf{P}(R|\neg c)\mathbf{P}(w|s,R)\\
&= \alpha\langle 0.2,0.8\rangle\langle 0.99,0.90\rangle = \alpha\langle 0.198,0.720\rangle = \langle 11/51,40/51\rangle
\end{aligned}$$

Strictly speaking, the transition matrix is only well-defined for the variant of MCMC in which the variable to be sampled is chosen randomly. (In the variant where the variables are chosen in a fixed order, the transition probabilities depend on where we are in the ordering.) Now consider the transition matrix.

- Entries on the diagonal correspond to self-loops. Such transitions can occur by sampling *either* variable. For example,

$$k((c,r) \rightarrow (c,r)) = 0.5P(c|r,s) + 0.5P(r|c,s,w) = 17/27$$

- Entries where one variable is changed must sample that variable. For example,

$$k((c,r) \rightarrow (c,\neg r)) = 0.5P(\neg r|c,s,w) = 5/54$$

- Entries where both variables change cannot occur. For example,

$$k((c,r) \rightarrow (\neg c,\neg r)) = 0$$

This gives us the following transition matrix, where the transition is from the state given by the row label to the state given by the column label:

$$\begin{array}{c|cccc} & (c,r) & (c,\neg r) & (\neg c,r) & (\neg c,\neg r) \\ \hline (c,r) & 17/27 & 5/54 & 5/18 & 0 \\ (c,\neg r) & 11/27 & 22/189 & 0 & 10/21 \\ (\neg c,r) & 2/9 & 0 & 59/153 & 20/51 \\ (\neg c,\neg r) & 0 & 1/42 & 11/102 & 310/357 \end{array}$$

c. $\mathbf{Q}^2$ represents the probability of going from each state to each state in two steps.

d. $\mathbf{Q}^n$ (as $n \to \infty$) represents the long-term probability of being in each state starting in each state; for ergodic $\mathbf{Q}$ these probabilities are independent of the starting state, so every row of $\mathbf{Q}$ is the same and represents the posterior distribution over states given the evidence.

e. We can produce very large powers of $\mathbf{Q}$ with very few matrix multiplications. For example, we can get $\mathbf{Q}^2$ with one multiplication, $\mathbf{Q}^4$ with two, and $\mathbf{Q}^{2^k}$ with k. Unfortunately, in a network with n Boolean variables, the matrix is of size $2^n \times 2^n$, so each multiplication takes $O(2^{3n})$ operations.

Exercise 13.GIBP

This exercise explores the stationary distribution for Gibbs sampling methods.

a. The convex composition $[\alpha, q_1; 1-\alpha, q_2]$ of q_1 and q_2 is a transition probability distribution that first chooses one of q_1 and q_2 with probabilities α and $1-\alpha$, respectively, and then applies whichever is chosen. Prove that if q_1 and q_2 are in detailed balance with π, then their convex composition is also in detailed balance with π. (*Note*: this result justifies a variant of GIBBS-ASK in which variables are chosen at random rather than sampled in a fixed sequence.)

b. Prove that if each of q_1 and q_2 has π as its stationary distribution, then the sequential composition $q = q_1 \circ q_2$ also has π as its stationary distribution.

a. Supposing that q_1 and q_2 are in detailed balance we have:

$$\begin{aligned} & \pi(\mathbf{x})(\alpha q_1(\mathbf{x} \to \mathbf{x}') + (1-\alpha) q_2(\mathbf{x} \to \mathbf{x}')) \\ = \; & \alpha \pi(\mathbf{x}) q_1(\mathbf{x} \to \mathbf{x}') + (1-\alpha)\pi(\mathbf{x}) q_2(\mathbf{x} \to \mathbf{x}') \\ = \; & \alpha \pi(\mathbf{x}) q_1(\mathbf{x}' \to \mathbf{x}) + (1-\alpha)\pi(\mathbf{x}) q_2(\mathbf{x}' \to \mathbf{x}) \\ = \; & \pi(\mathbf{x})(\alpha q_1(\mathbf{x}' \to \mathbf{x}) + (1-\alpha) q_2(\mathbf{x}' \to \mathbf{x})) \end{aligned}$$

b. The sequential composition is defined by

$$(q_1 \circ q_2)(\mathbf{x} \to \mathbf{x}') = \sum_{\mathbf{x}''} q_1(\mathbf{x} \to \mathbf{x}'') q_2(\mathbf{x}'' \to \mathbf{x}').$$

If q_1 and q_2 both have π as their stationary distribution, then:

$$\begin{aligned}\sum_{\mathbf{x}} \pi(\mathbf{x})(q_1 \circ q_2)(\mathbf{x} \to \mathbf{x}') &= \sum_{\mathbf{x}} \pi(\mathbf{x}) \sum_{\mathbf{x}''} q_1(\mathbf{x} \to \mathbf{x}'')q_2(\mathbf{x}'' \to \mathbf{x}') \\ &= \sum_{\mathbf{x}''} q_2(\mathbf{x}'' \to \mathbf{x}') \sum_{\mathbf{x}} \pi(\mathbf{x})q_1(\mathbf{x} \to \mathbf{x}'') \\ &= \sum_{\mathbf{x}''} q_2(\mathbf{x}'' \to \mathbf{x}')\pi(\mathbf{x}'') \\ &= \pi(\mathbf{x}')\end{aligned}$$

Exercise 13.MEHA

The **Metropolis–Hastings** algorithm is a member of the MCMC family; as such, it is designed to generate samples $\mathbf{x}$ (eventually) according to target probabilities $\pi(\mathbf{x})$. (Typically we are interested in sampling from $\pi(\mathbf{x}) = P(\mathbf{x} \mid \mathbf{e})$.) Like simulated annealing, Metropolis–Hastings operates in two stages. First, it samples a new state $\mathbf{x}'$ from a **proposal distribution** $q(\mathbf{x}' \mid \mathbf{x})$, given the current state $\mathbf{x}$. Then, it probabilistically accepts or rejects $\mathbf{x}'$ according to the **acceptance probability**

$$\alpha(\mathbf{x}' \mid \mathbf{x}) = \min\left(1, \frac{\pi(\mathbf{x}')q(\mathbf{x} \mid \mathbf{x}')}{\pi(\mathbf{x})q(\mathbf{x}' \mid \mathbf{x})}\right) .$$

If the proposal is rejected, the state remains at $\mathbf{x}$.

a. Consider an ordinary Gibbs sampling step for a specific variable X_i. Show that this step, considered as a proposal, is guaranteed to be accepted by Metropolis–Hastings. (Hence, Gibbs sampling is a special case of Metropolis–Hastings.)

b. Show that the two-step process above, viewed as a transition probability distribution, is in detailed balance with π.

a. Because a Gibbs transition step is in detailed balance with π, we have that the acceptance probability is one:

$$\begin{aligned}\alpha(\mathbf{x}' \mid \mathbf{x}) &= \min\left(1, \frac{\pi(\mathbf{x}')q(\mathbf{x} \mid \mathbf{x}')}{\pi(\mathbf{x})q(\mathbf{x}' \mid \mathbf{x})}\right) \\ &= 1\end{aligned}$$

since by definition of detailed balance we have

$$\pi(\mathbf{x}')q(\mathbf{x} \mid \mathbf{x}') = \pi(\mathbf{x})q(\mathbf{x}' \mid \mathbf{x}) .$$

b. Two prove this in two stages. For $\mathbf{x} \neq \mathbf{x}'$ the transition probability distribution is

$q(x' \mid x)\alpha(\mathbf{x}' \mid \mathbf{x})$ and we have:

$$\begin{aligned}
\pi(\mathbf{x})q(x' \mid x)\alpha(\mathbf{x}' \mid \mathbf{x}) &= \pi(\mathbf{x})q(x' \mid x)\min\left(1, \frac{\pi(\mathbf{x}')q(\mathbf{x} \mid \mathbf{x}')}{\pi(\mathbf{x})q(\mathbf{x}' \mid \mathbf{x})}\right) \\
&= \min\left(\pi(\mathbf{x})q(x' \mid x), \pi(\mathbf{x}')q(\mathbf{x} \mid \mathbf{x}')\right) \\
&= \pi(\mathbf{x}')q(x \mid x')\min\left(\frac{\pi(\mathbf{x})q(\mathbf{x}' \mid \mathbf{x})}{\pi(\mathbf{x}')q(\mathbf{x} \mid \mathbf{x}')}, 1\right)
\end{aligned}$$

$$\pi(\mathbf{x}')q(x \mid x')\alpha(\mathbf{x} \mid \mathbf{x}')$$

For $\mathbf{x} = \mathbf{x}'$ the transition probability is some $q'(x \mid x)$ which always satisfies the equation for detailed balance:

$$\pi(\mathbf{x})q'(x \mid x) = \pi(\mathbf{x})q'(x \mid x).$$

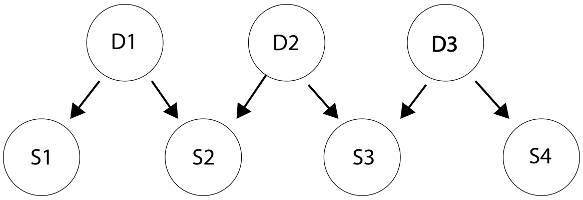


Figure S13.52 A 3-zigzag Bayes net.

Exercise 13.ZGZG

A k-*zigzag* network has k Boolean root variables and $k+1$ Boolean leaf variables, where root i is connected to leaves i and $i+1$. Figure S13.52 shows an example for $k=3$, where each D_i represents a Boolean disease variable and each S_j is a Boolean symptom variable.

a. Does having symptom S_4 affect the probability of disease D_1? Why or why not?

b. Using only conditional probabilities from the model, express the probability of having symptom S_1 but not symptom S_2, given disease D_1.

c. Can exact inference in a k-zigzag net can be done in time $O(k)$? Explain.

d. Suppose the values of all the symptom variables have been observed, and you would like to do Gibbs sampling on the disease variables (i.e., sample each variable given its Markov blanket). What is the largest number of non-evidence variables that have to be considered when sampling any particular disease variable? Explain your answer.

e. Suppose $k=50$. You would like to run Gibbs sampling for 1 million samples. Is it a good idea to precompute all the sampling distributions, so that when generating each individual sample no arithmetic operations are needed? Explain.

f. A k-zigzag++ network is a k-zigzag network with two extra variables: one is a root connected to all the leaves and one is a leaf to which all the roots are connected. You would like to run Gibbs sampling for 1 million samples in a 50-zigzag++ network. Is it a good idea to precompute all the sampling distributions? Explain.

g. Returning to the k-zigzag network, let us assume that in every case the values of all symptom and disease variables are observed. This means that we can consider training a neural net to predict diseases from symptoms directly (rather than using a Bayes net and probabilistic inference), where the symptoms are inputs to the neural net and the diseases are the outputs. The issue is the internal connectivity of the neural net: to which diseases should the symptoms be connected? A neural-net expert opines that we can simply reverse the arrows in the figure, so that D_1 is predicted only as a function $f_1(S_1, S_2)$ and so on (The functions f_i will be learned from data, but the representation of f_i and the learning algorithm is not relevant here.) The reasoning is that argues that because D_1 doesn't affect S_3 and S_4, they are not needed for predicting D_1. Is the expert right? Explain.

a. No. D_1 is independent of its nondescendants (including S_4) given its parents (empty set), i.e., D_1 is absolutely independent of S_4.

b. $P(s_1 \wedge \neg s_2 \,|\, D_1) = P(s_1 \,|\, D_1)P(\neg s_2 \,|\, D_1) = P(s_1 \,|\, D_1) \sum_{d_2} P(d_2)P(\neg s_2 \,|\, D_1, d_2)$.

c. This is correct: the simplest explanation is that the net is a polytree and the CPT sizes are independent of k.

d. The Markov blanket of a disease variable includes its parents, children, children's other parents; for a disease variable, the non-evidence variables in the Markov blanket are just the two neighboring disease variables.

e. Yes. Precomputation requires space for 200 numbers and saves tens of millions of computations.

f. No. The Markov blanket of every disease now includes all other diseases, and a table of 2^{50} numbers is impossible to compute or store.

g. No. While D_1 is absolutely independent of S_3, it is not independent given S_1 and S_2. The reason is that the value of S_3 affects the probability of D_2; D_2 in turn can "explain away" S_2, thereby changing the probability of D_1. The same argument applies, by extension, to all symptoms. Thus, the neural network would need to be completely connected, leading to the need to train $O(k)$ k-ary functions.

13.5 Causal Networks

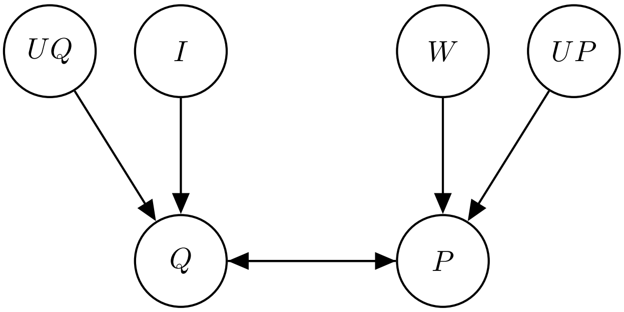
Exercise 13.DMND

In this problem, we will play economist. Consider four variables, price (P), demand (Q), income (I), and wages (W). More specifically, where Q is the quantity of household demand for a product A, P is the unit price of product A, I is household income, W is the wage rate for producing product A. Assume that income and wages are modeled separately (are independent). This example comes from Pearl (2000).

a. Using your background knowledge of which variables influence each other when changed, express this problem as a causal network. Include terms for unmodeled variables.

b. Express this problem a system of structural equations.

c. Suppose we now change the price of the item, we $do(p = x)$. What is the new joint distribution, $P(I, Q, W \mid do(p = x))$?



a.

b.

$$
\begin{aligned}
I &= f_I(UI)\\
W &= f_W(UW)\\
Q &= f_Q(I, P, UQ)\\
P &= f_P(W, Q, UP)
\end{aligned}
$$

c. $P(I, Q, W \mid do(p = x)) = P(I)P(Q \mid I, = x)P(W)$

Exercise 13.EFCT

Prove that if there is no direct path between two variables, X and Y in a causal network, that there is no causal effect between them.

The proof follows from the Markov assumption of causal networks. For example in the sprinkler example in Figure 13.23 the action $do(Sprinkler = true)$ has no effect on the variable $Cloudy$. Equation (13.20) shows that any effect that the sprinkler might have had is removed by the act of intervening—the variables are d-separated. That is, $X \perp\!\!\!\perp Y \mid do(X)$.

Exercise 13.DOOR

Scientific inquiry investigates the causal relation of variables; scientists study the effects of interventions. Causal networks provide a framework for such analysis by exactly specifying variables' putative relations. These networks, drawing on the work of Cinelli et. al (2021), can show us which treatments (the variable, X) we can expect to measure the effect of (on the variable, Y) when controlling for others. In particular, we're interested in whether a variable (here, Z) would serve as a good, bad, or neutral control over the influence of X on Y in a observational setting (we have not intervened, as in a randomized control trial, to

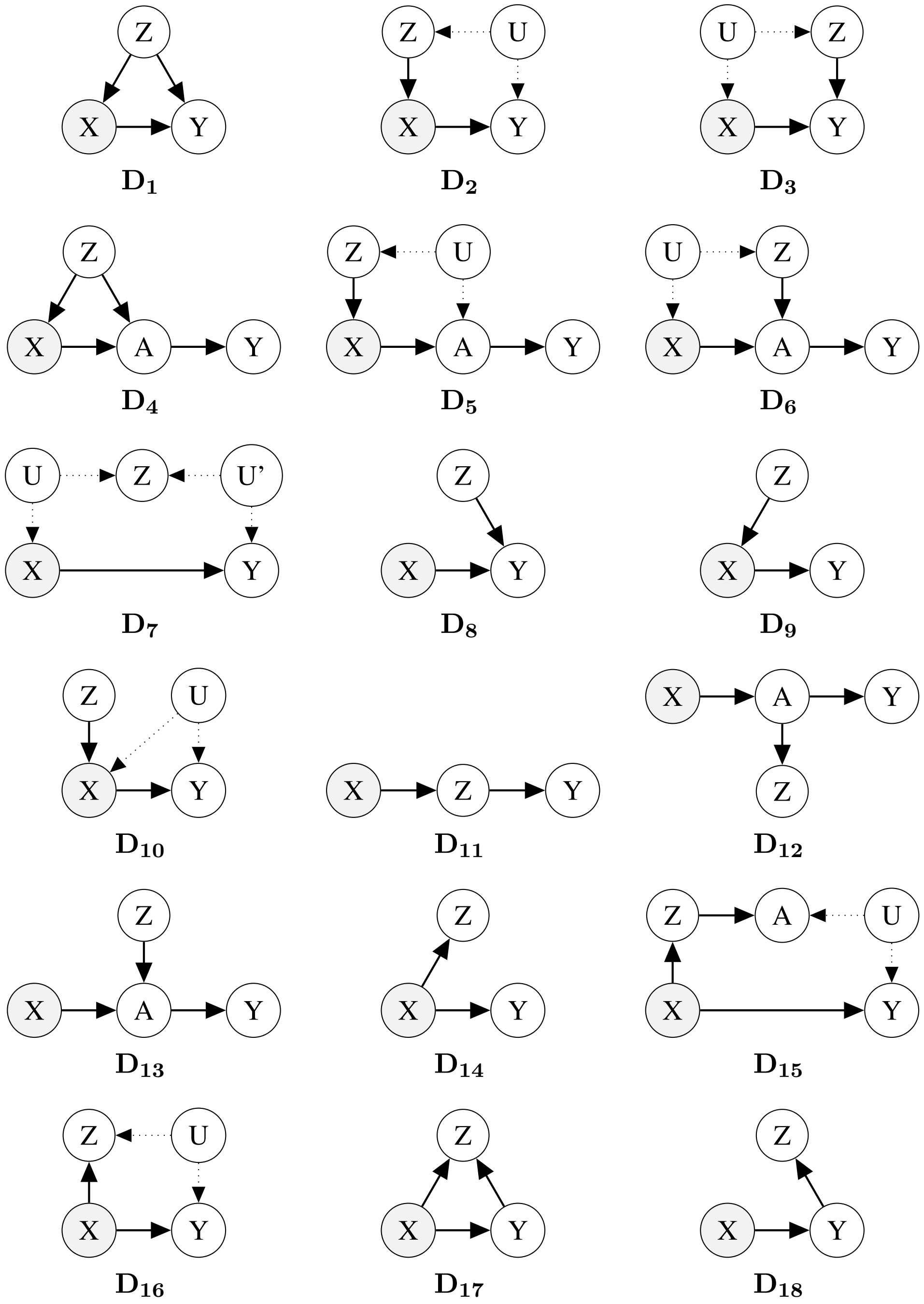


Figure S13.53 Causal networks for Exercise 13.DOOR. The grey circled nodes are observed treatments. Unmodeled variables, U, are shown to influence others with dotted lines.

$do(X)$); whether its inclusion in a regression analysis, for example, would reduce bias (close a confounding path; a good control), increase bias (open a confounding path; a bad control), or keep bias the same (neither open nor close a path; a neutral control).

That is, we want to measure the effect of X on Y. Thus, in which cases ($\mathbf{D_1}$ through $\mathbf{D_{18}}$) would keeping track of Z as well tell us more (good), less (bad), or nothing (neutral) about their relation? One way to measure this is the back-door criterion: does Z d-separate X from Y? (Pearl (2000) goes more in depth on the specifics of the back-door criterion.)

a. For each of the networks in Figure S13.53, state whether Z would serve as a good, bad, or neutral control. If a good or bad control state the reason (which confounding path was opened or closed or otherwise).

b. Now consider the networks in Figure S13.53 in an experimental setting, in which we $do(X)$. Is Z a good, bad, or neutral control? Why?

a. $\mathbf{D_1}$: Good; closes $X \quad Z \rightarrow Y$.

$\mathbf{D_2}$: Good; closes $X \quad Z \quad U \rightarrow Y$.

$\mathbf{D_3}$: Good; closes $X \quad U \quad Z \rightarrow Y$.

$\mathbf{D_4}$: Good; closes $X \quad Z \quad A \rightarrow Y$.

$\mathbf{D_5}$: Good; closes $X \quad Z \quad U \rightarrow A \rightarrow Y$.

$\mathbf{D_6}$: Good; closes $X \quad U \rightarrow Z \rightarrow A \rightarrow Y$.

$\mathbf{D_7}$: Bad; opens $X \quad U \rightarrow Z \leftarrow U' \rightarrow Y$.

$\mathbf{D_8}$: Neutral.

$\mathbf{D_9}$: Neutral.

$\mathbf{D_{10}}$: Bad; opened $Z \rightarrow X \rightarrow Y$.

$\mathbf{D_{11}}$: Bad; closes $X \rightarrow Z \rightarrow Y$, what we are trying to measure.

$\mathbf{D_{12}}$: Bad; controls $X \rightarrow A \rightarrow Y$, what we are trying to measure.

$\mathbf{D_{13}}$: Neutral.

$\mathbf{D_{14}}$: Neutral.

$\mathbf{D_{15}}$: Neutral.

$\mathbf{D_{16}}$: Bad; opens $X \rightarrow Z \quad U \rightarrow Y$.

$\mathbf{D_{17}}$: Bad; opens $X \rightarrow Z \quad Y$.

$\mathbf{D_{18}}$: Bad; opens a (virtual path) from the unmodeled variable which influences Y, $X \rightarrow Y \quad U_y$.

b. In all of the networks except $\mathbf{D_{16,17,18}}$ and $\mathbf{D_{11,12}}$ (in which Z is an effect of X and follow the same reason as in (a)), Z becomes a neutral control because by Equation (13.20) the back-door paths are removed by intervening.