CS 181 Spring 2021 Section 10
Solution

1 Bayesian Networks

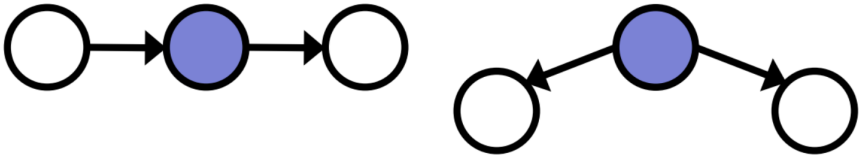
A Bayesian network is a graphical model that represents random variables and their dependencies using a directed acyclic graph. Bayesian networks are useful because they allow us to efficiently model joint distributions over many variables by taking advantage of the local dependencies. With Bayesian networks, we can easily reason about conditional independence and perform inference on large joint distributions.

1.1 D-separation rules

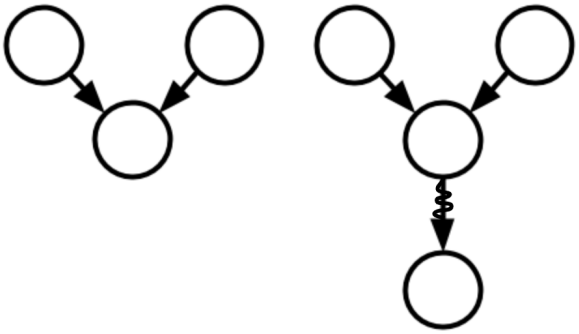
Let X_A and X_B denote sets of variables that we are interested in reasoning about. X_A and X_B are *d-separated* by a set of evidence X_E if **every** undirected path from X_A to X_B is "blocked" by X_E. A path is blocked by evidence X_E if EITHER:

1. There is a node Z with non-converging arrows on the path, and $Z \in X_E$.

   The shaded node indicates an evidence node.



2. There is a node Z with converging arrows on the path, and neither Z nor its descendants are in X_E.



Make sure to check **every** undirected path from X_A to X_B. Within each path, only one node Z needs to fall under one of the two cases described above for the whole path to be blocked.

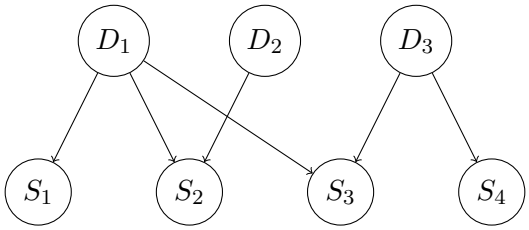
If X_A and X_B are d-separated by X_E (i.e., all paths are blocked), then X_A and X_B are conditionally independent given X_E ($X_A \perp X_B \,|\, X_E$).

2 Network Basics

A patient goes to the doctor for a medical condition, and the doctor suspects 3 diseases as the cause of the condition. The 3 diseases are D_1, D_2, and D_3, and they are independent from each other (given no other observations). There are 4 symptoms S_1, S_2, S_3, and S_4, and the doctor wants to check for presence in order to find the most probable cause. S_1 can be caused by D_1, S_2 can be caused by D_1 and D_2, S_3 can be caused by D_1 and D_3, and S_4 can be caused by D_3. Assume all random variables are Bernoulli, i.e. the patient has the disease/symptom or not.

- **Q:** Draw a Bayesian network for this problem with the variable ordering $D_1, D_2, D_3, S_1, S_2, S_3, S_4$.

  A: Note that there are many valid networks (depending on the chosen variable ordering), some more efficient (i.e. requiring fewer parameters) than others. Here is a compact representation that comes from variable ordering $D_1, D_2, D_3, S_1, S_2, S_3, S_4$. (Recall that all dependencies to earlier variables need to be indicated with edges).



- **Q:** Write down the expression for the joint probability distribution given this network.

  A: $p(D_1, D_2, D_3, S_1, S_2, S_3, S_4)$

  $= p(D_1)p(D_2)p(D_3)p(S_1|D_1)p(S_2|D_1, D_2)p(S_3|D_1, D_3)p(S_4|D_3)$

- **Q:** How many parameters are required to describe this joint distribution?

  A:

Conditional Probability Table	Number of Parameters
$p(D_1)$	1
$p(D_2)$	1
$p(D_3)$	1
$p(S_1\|D_1)$	2
$p(S_2\|D_1, D_2)$	4
$p(S_3\|D_1, D_3)$	4
$p(S_4\|D_3)$	2
Total Number of Parameters	15

- **Q:** How many parameters would be required to represent the CPTs in a Bayesian network if there were no conditional independences between variables?

  A: The network would be structured as a clique, and considering order $D_1, D_2, D_3, S_1, S_2, S_3, S_4$, the number of parameters for the CPTs would be $1 + 2 + 4 + 8 + 16 + 32 + 64 = 127$.

Conditional Probability Table	Number of Parameters
$p(D_1)$	1
$p(D_2\|D_1)$	2
$p(D_3\|D_1, D_2)$	4
$p(S_1\|D_1, D_2, D_3)$	8
$p(S_2\|D_1, D_2, D_3, S_1)$	16
$p(S_3\|D_1, D_2, D_3, S_1, S_2)$	32
$p(S_4\|D_1, D_2, D_3, S_1, S_2, S_3)$	64
Total Number of Parameters	127

(We can see there is no saving relative to specifying the joint probability distribution directly, which would require $2^7 - 1 = 127$ numbers.)

- **Q:** What diseases do we gain information about when observing the fourth symptom ($S_4 = true$)?

  A: We have independence relations $I(D_1, S_4)$ (since the path is blocked without observing S_3 and $I(D_2, S_4)$ (since the path is blocked at both S_2 and S_3). What is left is dependence between D_3 and S_4. Thus, we only learn information about D_3.

- **Q:** Suppose we know that the third symptom is present ($S_3 = true$). What does observing the fourth symptom ($S_4 = true$) tell us now?

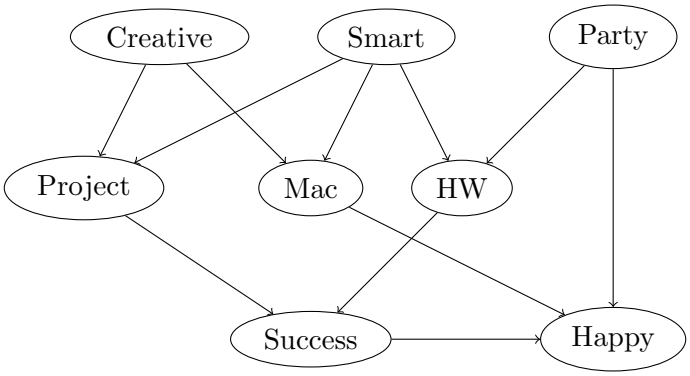
  A: With $S_3 = true$, observing $S_4 = true$ now also gives us information about D_1 (via 'explaining away', or using d-separation, because the D_1 to S_4 path is no longer blocked at S_3). We still don't learn any information about D_2 because the D_2 to S_4 path remains blocked at S_2.

3 D-Separation

As part of a comprehensive study of the role of CS 181 on people's happiness, we have been collecting important data from students. In an entirely optional survey that all students are required to complete, we ask the following highly objective questions:

Do you party frequently [Party: Yes/No]?
Are you smart [Smart: Yes/No]?
Are you creative [Creative: Yes/No]?
Did you do well on all your homework assignments? [HW: Yes/No]
Do you use a Mac? [Mac: Yes/No]
Did your last major project succeed? [Project: Yes/No]
Did you succeed in your most important class? [Success: Yes/No]
Are you currently Happy? [Happy: Yes/No]

After consulting behavioral psychologists we build the following model:



- **Q:** True or False: $Party$ is independent of $Success$ given HW.

  A: False; there is a path that is not blocked: $Party - HW - Smart - Project - Success$ has neither a converging arrows not in the set of evidence or a non-converging arrows in the set.

- **Q:** True or False: $Creative$ is independent of $Happy$ given Mac.

  A: False; there is a path that is not blocked: $Creative - Project - Success - Happy$

- **Q:** True or False: $Party$ is independent of $Smart$ given $Success$.

  A: False; there is a path that is not blocked between $Party$ and $Smart$: the path $Party - HW - Success$ is not blocked because the converging arrows node at HW has a descendant ($Success$) in the evidence.

- **Q:** True or False: $Party$ is independent of $Creative$ given $Happy$.

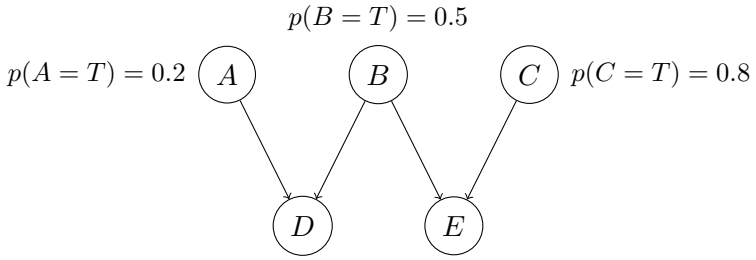
A: False; there is a path that is not blocked between $Party$ and $Creative$ through the converging arrows at $Happy$. There are actually multiple not-blocked paths – can you find them?

- **Q:** True or False: $Party$ is independent of $Creative$ given $Success$, $Project$ and $Smart$.

  A: True! All paths between $Party$ and $Creative$ are blocked. Working from $Party$, the paths that come through $Happy$ are blocked there (converging arrows, no evidence). Those that come through HW and $Smart$ are blocked at $Smart$. Those that come through $HW, Success, Project$ are blocked at $Project$.

4 Inference

Consider the following Bayesian network, where all variables are Bernoulli.



A	B	$p(D = T\|A, B)$
F	F	0.9
F	T	0.6
T	F	0.5
T	T	0.1

B	C	$p(E = T\|B, C)$
F	F	0.2
F	T	0.4
T	F	0.8
T	T	0.3

- **Q:** What is the probability that all five variables are simultaneously false (F)?

  A:

$$p(A = F, B = F, C = F, D = F, E = F) =$$

$$p(A = F)p(B = F)p(C = F)p(D = F|A = F, B = F)p(E = F|B = F, C = F)$$

$$= (0.8)(0.5)(0.2)(0.1)(0.8)$$
$$= 0.0064$$

- **Q:** What is the probability that A is false given that the remaining variables are all known to be true (T)?

  A: For this part, we need to calculate $p(A = F|B = T, C = T, D = T, E = T)$.

  By the definition of conditional probability,

$$p(A = F|B = T, C = T, D = T, E = T)$$

$$= \frac{p(A = F, B = T, C = T, D = T, E = T)}{P(B = T, C = T, D = T, E = T)}$$

$$= \frac{p(A = F, B = T, C = T, D = T, E = T)}{P(A = F, B = T, C = T, D = T, E = T) + P(A = T, B = T, C = T, D = T, E = T)}$$

The joint probabilities $p(A = F, B = T, C = T, D = T, E = T)$ and $p(A = T, B = T, C = T, D = T, E = T)$ can be computed as:

$$p(A = F, B = T, C = T, D = T, E = T)$$

$$= p(A = F)p(B = T)p(C = T)p(D = T|A = F, B = T)p(E = T|B = T, C = T)$$
$$= (0.8)(0.5)(0.8)(0.6)(0.3)$$
$$= (0.05760)$$

$$p(A = T, B = T, C = T, D = T, E = T)$$

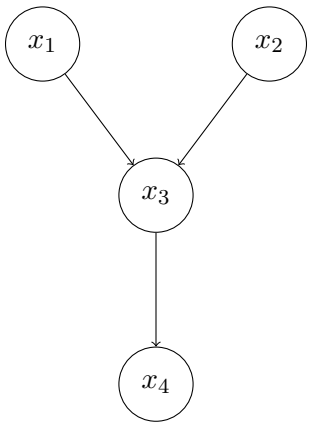
$$= p(A = T)p(B = T)p(C = T)p(D = T|A = T, B = T)p(E = T|B = T, C = T)$$
$$= (0.2)(0.5)(0.8)(0.1)(0.3)$$
$$= (0.00240)$$

Finally, we can plug this in to get:

$$p(A = T|B = F, C = F, D = F, E = F) = \frac{.05760}{.05760 + .00240} = .96$$

5 Variable Elimination in Bayesian Networks

We apply an inference algorithm called variable elimination to the following Bayesian network:



Assume that all of the random variables are Bernoulli, meaning their domain is $\{0, 1\}$ with domain size $k = 2$. In this network, we can encode the joint distribution as

$$p(x_1, x_2, x_3, x_4) = p(x_1)p(x_2)p(x_3|x_1, x_2)p(x_4|x_3)$$

If we wanted to calculate the marginal distribution of x_4 that is, have x_4 be our query without any evidence (conditioned on variables), we could naively marginalize out all other variables:

$$\begin{aligned} p(x_4) &= \sum_{x_1}\sum_{x_2}\sum_{x_3} p(x_1, x_2, x_3, x_4) \\ &= \sum_{x_1}\sum_{x_2}\sum_{x_3} p(x_1)p(x_2)p(x_3|x_1, x_2)p(x_4|x_3) \end{aligned}$$

To calculate these sums we would need to multiply two k-dimensional vectors for each of the $k^3 = 8$ possible combinations of x_1, x_2, x_3. In general, the number of combinations grows exponentially in the number of variables ($O(k^n)$ if you're familiar with big-O notation).

Note that Bayesian nets encode dependencies between variables, which we can use to calculate the marginal distribution more efficiently. By reordering the sums and eliminating one variable at a time, we derive the variable elimination procedure:

$$\begin{aligned} p(x_4) &= \sum_{x_1}\sum_{x_2}\sum_{x_3} p(x_1)p(x_2)p(x_3|x_1, x_2)p(x_4|x_3) \\ &= \sum_{x_3} p(x_4|x_3) \sum_{x_2} p(x_2) \sum_{x_1} p(x_3|x_1, x_2)p(x_1) \\ &= \sum_{x_3} p(x_4|x_3) \sum_{x_2} p(x_2)p(x_3|x_2) \\ &= \sum_{x_3} p(x_4|x_3)p(x_3) \\ &= p(x_4) \end{aligned}$$

Here, we eliminate x_1 using a k by k matrix $g_1(x_3, x_2)$, because we have to sum over x_1 for each possible value of x_2 and x_3. Then we eliminate x_2 with a K-dimensional vector $g_2(x_3)$, likewise because we sum over x_2 for each possible value of x_3. Lastly, we eliminate x_3, which results in a final K-dimensional vector of probabilities for x_4. Notice that we have a poly-tree, and we're eliminating leaves first and working towards our query variable, x_4.

In this way, we can perform the same computation in $O(k^3)$ time, because the longest elimination step has to do k^2 sum-product calculations for each element in the matrix $g_1(x_3, x_2)$, and each sum-product calculation takes $O(k)$ time (since we are summing over x_1). Compare this polynomial-time complexity (where we add the time taken for each elimination step, so our total time complexity only depends on the longest step) with the exponential-time complexity (based on the number of variables) of the naive approach.

Alternatively, we could have eliminated variables in a different order:

$$\begin{aligned}
p(x_4) &= \sum_{x_1}\sum_{x_2}\sum_{x_3} p(x_1)p(x_2)p(x_3|x_1,x_2)p(x_4|x_3) \\
&= \sum_{x_1} p(x_1) \sum_{x_2} p(x_2) \sum_{x_3} p(x_3|x_1,x_2)p(x_4|x_3) \\
&= \sum_{x_1} p(x_1) \sum_{x_2} p(x_2)p(x_4|x_1,x_2) \\
&= \sum_{x_1} p(x_1)p(x_4|x_1) \\
&= p(x_4)
\end{aligned}$$

Here, we eliminate x_3, then x_2, then x_1. Notice that the ordering matters: eliminating x_3 first results in a $k \times k \times k$ object $g(x_1, x_2, x_4)$, so our overall algorithm will run in $O(k^4)$ time. (Again, note that we have to account for both the k^3 sum-product calculations and the $O(k)$ time to do each calculation.)

In general, the computational cost of variable elimination depends on the number of variables in these intermediate factors, in particular the largest object computed ('tree-width').

5.1 Exercise: Variable Elimination

Consider the Bayesian network described in above, and assume the following Conditional Probability Table (CPT). Let $x_i \in \{0, 1\}$ denote the values that variable X_i can take. Our goal is to find $p(x_4)$.

x_1	$p(x_1)$
0	0.3
1	0.7

x_2	$p(x_2)$
0	0.6
1	0.4

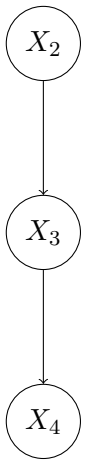
x_3	x_1	x_2	$p(x_3\|x_1, x_2)$
0	0	0	0.5
0	0	1	0.2
0	1	0	0.9
0	1	1	0.5
1	0	0	0.5
1	0	1	0.8
1	1	0	0.1
1	1	1	0.5

x_4	x_3	$p(x_4\|x_3)$
0	0	0.7
0	1	0.1
1	0	0.3
1	1	0.9

1. Eliminate X_1 first. Draw the resulting Bayesian network and compute the CPT.
2. Eliminate X_3 first. Draw the resulting Bayesian network and compute the CPT.
3. How many sum-product calculations do each of these variable elimination orders require? Which one is preferable?

Solution

1. The resulting network is:



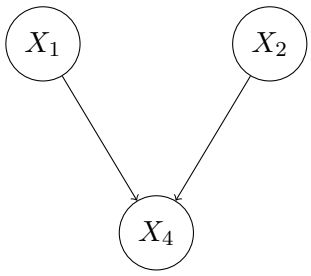
The variable elimination process eliminates X_1 by marginalizing out X_1: $p(x_3|x_2) = \sum_{x_1} p(x_3|x_1, x_2)p(x_1)$. For example:

$$\begin{aligned} p(X_3 = 0|X_2 = 0) &= \sum_{x_1 \in \{0,1\}} p(X_3 = 0|X_1 = x_1, X_2 = 0)p(X_1 = x_1) \\ &= 0.5 \cdot 0.3 + 0.9 \cdot 0.7 \\ &= 0.78 \end{aligned}$$

This is a sum-product calculation, and we need to do one for each value of X_2 and X_3. Thus, there are four sum-product calculations in total. The resulting CPT is:

x_3	x_2	$p(x_3 \mid x_2)$
0	0	0.78
0	1	0.41
1	0	0.22
1	1	0.59

2. The resulting network is



The variable elimination process eliminates X_3 by marginalizing out X_3: $p(x_4|x_1, x_2) = \sum_{x_3} p(x_4|x_3)p(x_3 \mid x_1, x_2)$. This would be the first intermediate term. For example:

$$\begin{aligned} p(X_4 = 0|X_1 = 0, X_2 = 0) &= \sum_{x_3 \in \{0,1\}} p(X_4 = 0|X_3 = x_3)p(X_3 = x_3|X_1 = 0, X_2 = 0) \\ &= 0.7 \cdot 0.5 + 0.1 \cdot 0.5 \\ &= 0.40 \end{aligned}$$

We need to do this for each combination of values for X_1, X_2 and X_4. Thus, there are eight sum-product calculations in total. The resulting CPT is:

x_4	x_1	x_2	$p(x_4|x_1, x_2)$
0	0	0	0.40
0	0	1	0.22
0	1	0	0.64
0	1	1	0.40
1	0	0	0.60
1	0	1	0.78
1	1	0	0.36
1	1	1	0.60

3. In these variable elimination operations, we need to compute intermediate terms. The cost of computing these depends on the number of variables that they mention, since each variable increases the number of required sum-product calculations by a factor of $k = 2$.

   For the first ordering, the intermediate terms are:

   - $p(x_3 \mid x_2)$: mentions x_2 and x_3, and thus requires four sum-product calculations (for each row in the original CPT)
   - $p(x_3)$: mentions x_3 and thus requires two sum-product calculations
   - $p(x_4)$: mentions x_4 and thus requires two sum-product calculations

We have a total of $4 + 2 + 2 = 8$ sum-product calculations.

For the second ordering, the intermediate terms are:

- $p(x_4 \mid x_1, x_2)$: mentions x_1, x_2 and x_4, and thus requires eight sum-product calculations (for each row in the original CPT)
- $p(x_4 \mid x_1)$: mentions x_1 and x_4, and thus requires four sum-product calculations
- $p(x_4)$: mentions x_4 and thus requires two sum-product calculations

We have a total of $8 + 4 + 2 = 14$ sum-product calculations.

Thus, we see that the first ordering is preferable since it requires fewer computational steps.

End Solution