CS 181: Bayesian Networks

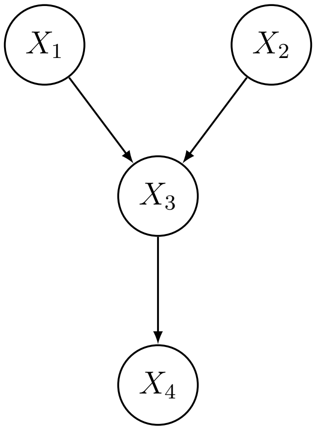
Week of: April 10, 2017
Harvard University

1 Section Objectives

- To understand key concepts including independence assumptions, d-separation, and inference.
- To understand model design choices related to building Bayesian networks.

2 Introduction

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 between variables. With Bayesian networks, we can easily reason about conditional independence and perform inference on large joint distributions.



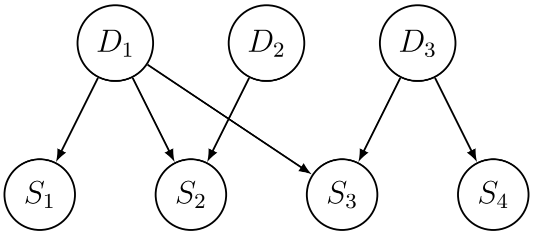
Modeling the joint distribution $p(X_1, X_2, X_3, X_4)$ using the dependencies between X_1, X_3 and X_2, X_3 and X_3, X_4.

3 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.

  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 $1 + 2 + 4 + 8 + 16 + 32 + 64 = 127$. (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 is an example of the 'explaining away' phenomenon in the compact Bayesian Network?

  A: S_3 depends on D_1 and D_3. When we know S_3, then conditioned on this D_1 and D_3 are not independent, and if we observe D_1 then D_3 is less likely to be a cause ("D_1 explains away D_3").

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

  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 ($S_3 = \mathit{true}$). What does observing the fourth symptom ($S_4 = \mathit{true}$) tell us now?

  A: With $S_3 = \mathit{true}$, observing $S_4 = \mathit{true}$ now also gives us informaion 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 abhout D_2 because the D_2 to S_4 path remains blocked at S_2.

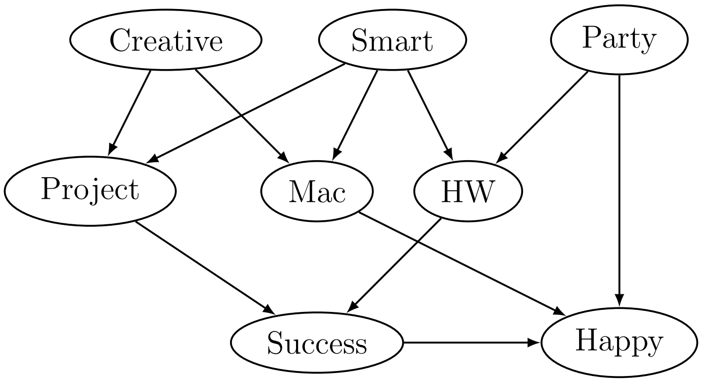
4 D-Separation

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]? (Please only answer Yes or 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*.

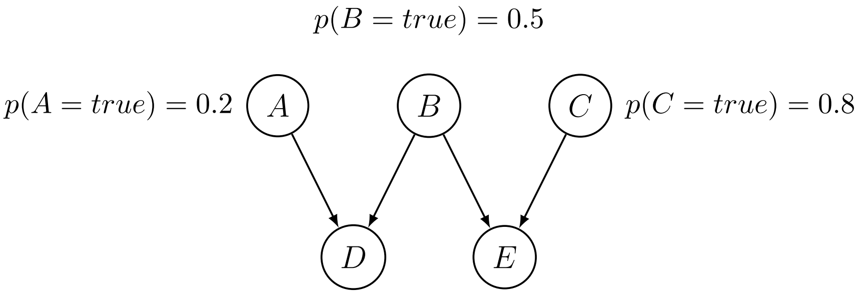
  A: False; there is a path that is not blocked between *Party* and *Creative* through the converging arrows at *Happy*. There are actually multiple not-blocked paths – can you find them?

- **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*.

5 Inference

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



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

B	C	$p(E = true \mid B, C)$
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 $false$?

  A:

$$\begin{aligned} p(\neg A, \neg B, \neg C, \neg D, \neg E) &= p(\neg A)p(\neg B)p(\neg C)p(\neg D|\neg A, \neg B)p(\neg E|\neg B, \neg C) \\ &= (0.8)(0.5)(0.2)(0.1)(0.8) \\ &= 0.0064 \end{aligned}$$

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

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

  We know that $p(\neg A|B, C, D, E) \propto p(\neg A, B, C, D, E)$. The joint probabilities $p(\neg A, B, C, D, E)$ and $p(A, B, C, D, E)$ can be computed as:

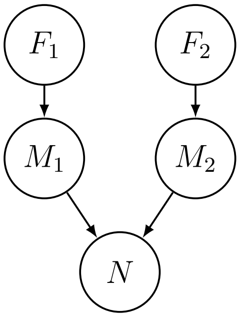
$$\begin{aligned} p(\neg A, B, C, D, E) &= p(\neg A)p(B)p(C)p(D|\neg A, B)p(E|B, C) \\ &= (0.8)(0.5)(0.8)(0.6)(0.3) \\ &= (0.05760) \\ p(A, B, C, D, E) &= p(A)p(B)p(C)p(D|A, B)p(E|B, C) \\ &= (0.2)(0.5)(0.8)(0.1)(0.3) \\ &= (0.00240) \end{aligned}$$

Finally, by normalization we have:

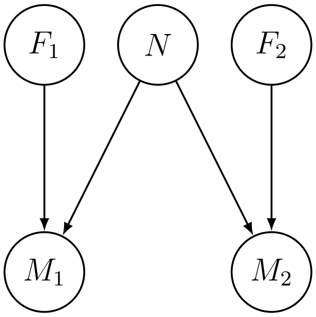
$$p(\neg A|B,C,D,E) = \frac{.05760}{.05760 + .00240} = .96$$

6 Reasoning about the Correctness of Networks

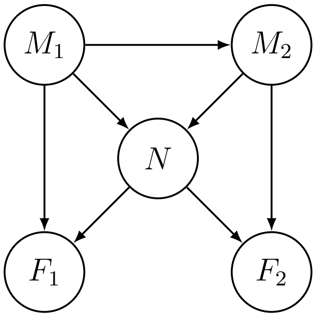
Two astronomers in different parts of the world make measurements M_1 and M_2 of the number of stars N in some small region of the sky. Each telescope may be badly out of focus (events F_1 and F_2). Consider the following Bayesian networks.



(i)



(ii)



(iii)

- **Q:** Which of these Bayesian networks correctly represents the distribution?

  A: Network (i) is incorrect. It states, for example, that $I(N, F_1 \mid M_1, M_2)$, whereas we know that F_1 still provides information about N conditioned on the two measurements (what if telescope 1 is faulty?).

  Network (ii) is correct. Consider ordering F_1, N, F_2, M_1, M_2. We can check that N is independent of F_1, F_2 is independent of N and F_1, M_1 depends on F_1 and N but not F_2, and M_2 depends on F_2 and N but not F_1, and not M_1 when conditioned on N.

  Network (iii) is correct. Consider ordering M_1, M_2, N, F_1, F_2. The only edges omitted when constructing the network are

  – from M_2 to F_1, but conditioned on N and M_1 then M_2 provides no information about F_1.

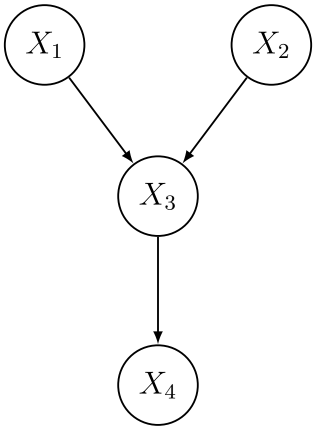
  – from M_1 to F_2 (same argument) and from F_1 to F_2, and F_1 provides no information about F_2 when conditioned on N and M_2.

- **Q:** Of the correct networks, why might one be preferred?

  A: Network (ii) may be preferable because it represents all of the necessary dependences while using fewer edges and, hence, fewer parameters. That is, network (ii) expresses the distribution more efficiently.

7 Variable Elimination

In this section, we discuss an exact inference algorithm called variable elimination. Consider the Bayesian network we saw in lecture:



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

$$p(x_1, x_2, x_3, x_4) = p(x_3|x_1, x_2)p(x_4|x_3)p(x_1)p(x_2) \tag{1}$$

If we wanted to calculate the marginal distribution of X_4, we could naively marginalize out the other variables, giving us

$$p(x_4) = \sum_{x_1}\sum_{x_2}\sum_{x_3} p(x_3|x_1, x_2)p(x_4|x_3)p(x_1)p(x_2) \tag{2}$$

Calculating this naively requires multiplying 4 values for each of the 8 possible combinations of x_1, x_2, x_3. In general, if there were many variables then the number of combinations would grow exponentially in the number of variables!

However, note that because we have a compact, Bayesian net representation, we can calculate the marginal distribution more efficiently. By reordering the sums and eliminating one variable at a time, we derive the variable elimination procedure. For example, we can calculate the joint distribution as:

$$
\begin{aligned}
p(x_4) &= \sum_{x_1,x_2,x_3} p(x_3|x_1,x_2)p(x_4|x_3)p(x_1)p(x_2) && (3)\\
&= \sum_{x_2,x_3} p(x_4|x_3)p(x_2) \sum_{x_1} p(x_3|x_1,x_2)p(x_1) && (4)\\
&= \sum_{x_3} p(x_4|x_3) \sum_{x_2} p(x_2)p(x_3|x_2) && (5)\\
&= \sum_{x_3} p(x_4|x_3)p(x_3) && (6)\\
&= p(x_4) && (7)
\end{aligned}
$$

Here, we eliminate x_1, then x_2, then x_3. This is working in 'leaves first' order towards the query, x_4. Alternatively, we could have eliminated variables in a different order, as follows:

$$
\begin{aligned}
p(x_4) &= \sum_{x_1,x_2,x_3} p(x_3|x_1,x_2)p(x_4|x_3)p(x_1)p(x_2) && (8)\\
&= \sum_{x_1,x_2} p(x_1)p(x_2) \sum_{x_3} p(x_3|x_1,x_2)p(x_4|x_3) && (9)\\
&= \sum_{x_1} p(x_1) \sum_{x_2} p(x_2)p(x_4|x_1,x_2) && (10)\\
&= \sum_{x_1} p(x_1)p(x_4|x_1) && (11)\\
&= p(x_4) && (12)
\end{aligned}
$$

Here, we eliminate x_3 then x_3 then x_1. For the following questions, assume the following CPTs:

x_1	$p(x_1)$
0	0.3
1	0.7

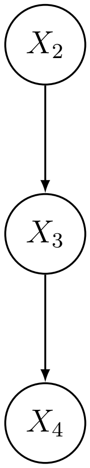
x_2	$p(x_2)$
0	0.6
1	0.4

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

- **Q:** Following the first ordering, first use variable elimination on X_1 to compute the CPT for $p(X_3|X_2)$. This represents the first intermediate term. Draw the resulting Bayesian network.

  A: 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:

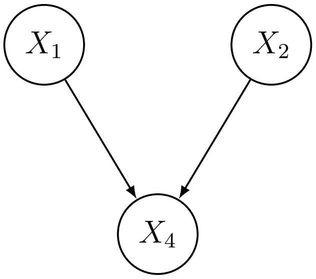
$$\begin{aligned} p(X_3 = 0|X_2 = 0) &= \sum_{x_1 \in \{0,1\}} p(X_3 = 0|X_1 = x_1, X_2 = 0)p(X_1 = x_1) \\ &= 0.5 \cdot 0.3 + 0.9 \cdot 0.7 \\ &= 0.78 \end{aligned}$$

This is a 'sum-product calculation'. We need to do this for each value of X_2 and X_3. Thus, there are four sum-product calculations to perform. The resulting CPT is:

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

- **Q:** Using the second ordering, we would first use variable elimination on X_3 to compute the probability table for $p(X_4|X_1, X_2)$. Compute the CPT and draw the resulting Bayesian network.

  A: 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 value of X_1, X_2 and X_4. Thus, there are eight sum-product calculations to perform. 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

- **Q:** How many sum-product calculations do each of these variable elimination orders require? Which one is preferable?

  A: In these variable elimination operations, we need to compute intermediate terms. The cost of computing these intermediate terms depends on the number of variables that they mention, because we have a sum product calculation for each mention.

  For the first ordering, the intermediate terms are:

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

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

For the second ordering, the intermediate terms are:

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

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

Thus, we see that the first ordering requires fewer computational steps.