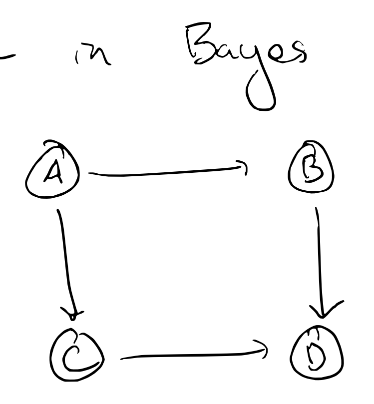
Marginalizing

Let X, Y be sets of variables. Suppose we know the joint distribution over X, Y i.e. we can compute probabilities $P(x,y)$ for outcomes x of X and y of Y.

We can then compute probabilities $P(x)$ by marginalizing out Y:

$$P(x) = \sum_y P(x,y).$$

This is the same idea that we used for doing inference in Bayes nets:



$$P(d) = \sum_{a,b,c} P(a,b,c,d)$$

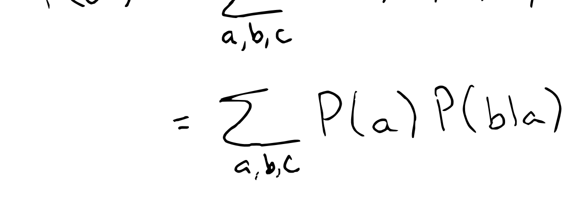
$$= \sum_{a,b,c} P(a)P(b|a)P(c|a)P(d|b,c).$$

$$P(c|d) = \frac{P(c,d)}{P(d)} = \frac{\sum_{a,b} P(a,b,c,d)}{\sum_{a,b,c'} P(a,b,c',d)}$$

$$= \frac{\sum_{a,b} P(a)P(b|a)P(c|a)P(d|b,c)}{\sum_{a,b,c'} P(a)P(b|a)P(c'|a)P(d|b,c')}.$$

Computing these naively can be expensive!

Example



$$P(d) = \sum_{a,b,c} P(a)P(b|a)P(c|b)P(d|c)$$

If there are 5 possible values for each random variable, need to look up probabilities $5^3 \cdot 4 = 500$ times.

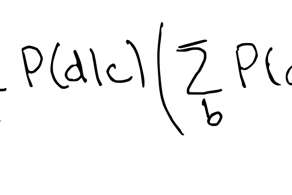
We can do better:

$$P(d) = \sum_c P(d|c)\Big(\sum_b P(c|b)\Big(\underbrace{\sum_a P(b|a)P(a)}_{\text{this term does not depend on } c,d}\Big)\Big)$$

For each value of b, we can compute once and cache the value $P(b) = \sum_a P(b|a)P(a)$. Computing all the $P(b)$'s takes $5^2 \cdot 2 = 50$ lookups.

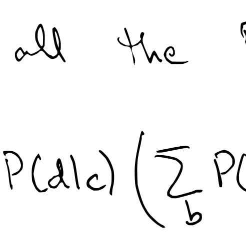
$$P(d) = \sum_c P(d|c)\Big(\sum_b P(c|b)P(b)\Big)$$

After caching all 5 of the $P(b)$'s, we're effectively reduced to the Bayes net



It again takes $5^2 \cdot 2 = 50$ lookups to compute $P(c) = \sum_b P(c|b)P(b)$ for each c.

This reduces the problem to

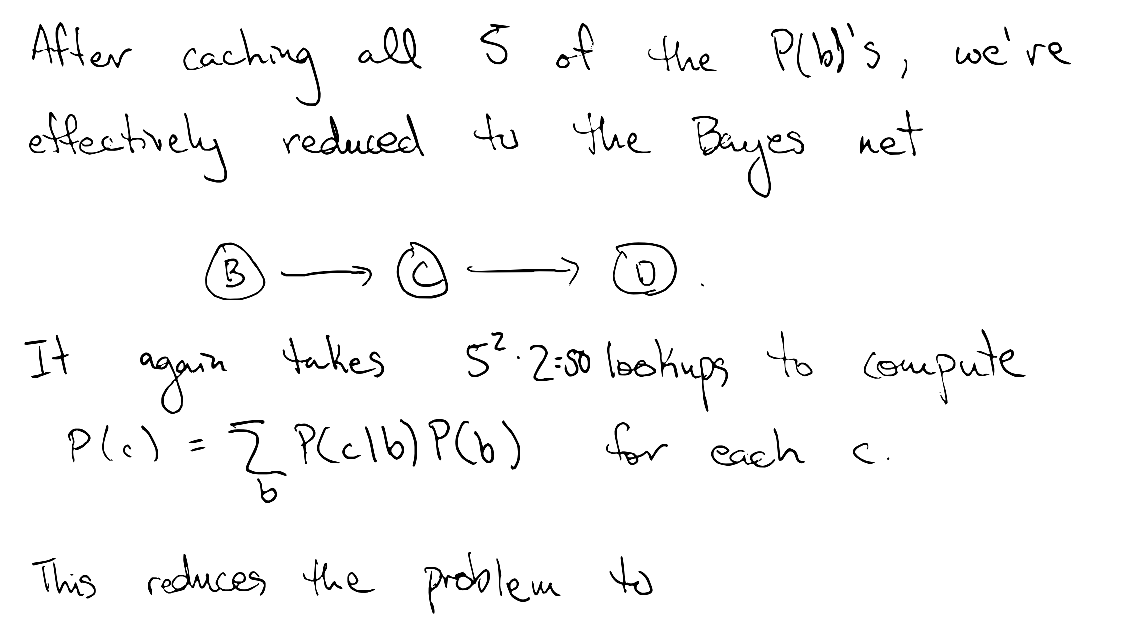


$$P(d) = \sum_c P(d|c)P(c)$$

which takes another $5^2 \cdot 2 = 50$ lookups.

In total, this took 150 lookups.

Another example



If each variable is T or F the naive approach to computing $P(D=T)$ takes $2^6 \cdot 6$ lookups.

$$P(D=T) = \sum_{t,s,a,l,b,x} P(t)P(s)P(l|s)P(b|s)P(a|l,t)P(x|a)P(D=T|a,b)$$

Let's eliminate in the order T, X, S, L, B, A:

$P(a) = \sum_t P(a|t)P(t)$ ($2^2 \cdot 2$ lookups)

$1 = \sum_x P(x|a)$ (2 lookups (or 0 if we realize at the start that we can ignore X))

$P(l,b) = \sum_s P(l|s)P(b|s)P(s)$ ($2 \cdot 3$ lookups)

$P(a,b) = \sum_l P(a|l)P(l,b)$ ($2 \cdot 2$ lookups)

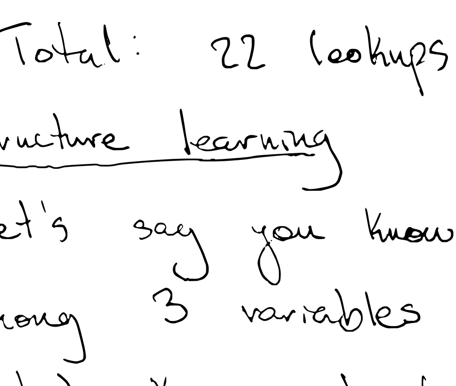
$P(a) = \sum_b P(a,b)$ (2 lookups)

$P(D=T) = \sum_a P(D=T|a)P(a)$ ($2 \cdot 2$ lookups)

Total: 22 lookups vs. $2^6 \cdot 6 = 384$ lookups.

Structure learning

Let's say you know the joint distribution among 3 variables (e.g. by collecting lots of data). You want to distinguish the following 3 possible Bayes nets:



(chain)
(fork)
(collider)

How can you do it?

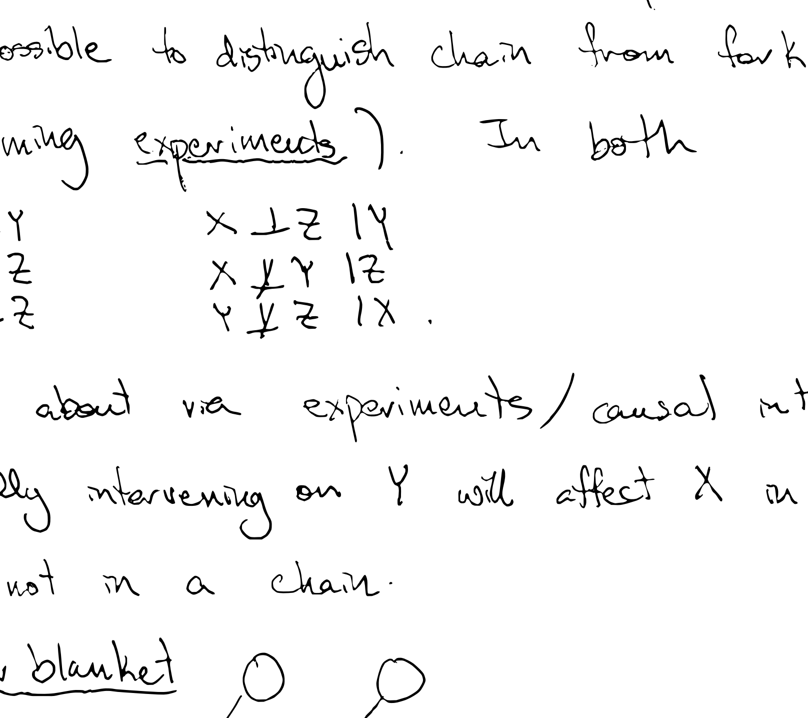
To distinguish collider from other two: check whether X and Z are independent.

Not possible to distinguish chain from fork (without performing experiments). In both

$X \not\perp Y$	$X \perp Z \mid Y$
$X \not\perp Z$	$X \not\perp Y \mid Z$
$Y \not\perp Z$	$Y \not\perp Z \mid X$.

What about via experiments / causal interventions? Causally intervening on Y will affect X in a fork, but not in a chain.

Markov blanket



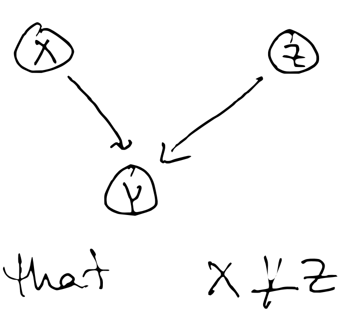
Suppose given a Bayes net and a variable X. Let's say we want to form a guess about the value of X, but we can't measure it directly. We can measure any other variable, but measuring variables is expensive. So we ask:

Q: Among sets of variables that contain as much information about X as possible, what's the smallest such set?

Q': What is the minimal set B of variables not containing X such that $X \perp\!\!\!\perp \text{(rest of the variables)} \mid B$?

A: S is the set consisting of X's parents, X's children, and X's spouses. This is called the Markov blanket of X.

Clearly S needs to include X's parents and children. Why the spouses?



Recall that $X \not\perp Z \mid Y$, so once we've included $Y \in S$, we must include Z as well.