Causal Structure Learning with Bayesian Networks

Week 11

April 8, 2024

Today's Outline

- Building and interpreting Bayes Nets
- Inference of conditional dependencies with Bayes Nets
- Learning Bayes Net Parameters and Structure from Data
- Causal structure learning with the PC algorithm
- Assumptions, extensions, and variant
- Causal Orientation methods
- Bayes Nets with Python

Building a Bayes Net

Small Bayes Nets are easy to build by hand, assuming that we understand the relationships between variables and are able to estimate their conditional probabilities.

Large Bayes Nets may require many person-hours to build, but they can also be **learned** automatically from data.

For example, let's assume that we want to build a Bayes Net to determine whether a terrorist anthrax attack has occurred.

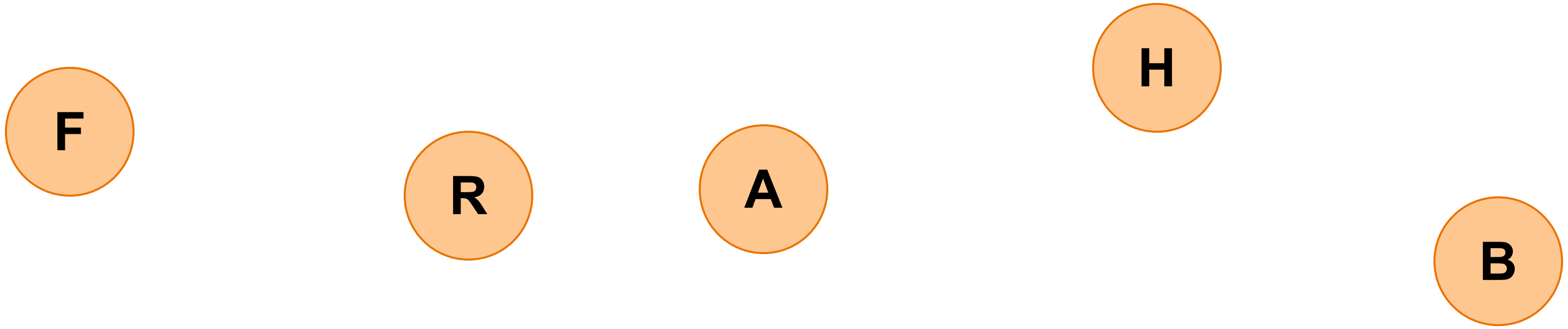
1. An anthrax attack is likely to increase the level of respiratory illness.
2. Seasonal influenza is also likely to cause an increase in respiratory illness.
3. The CDC has a **hospital surveillance system** that alerts when the number of ED visits is abnormally high and has additionally deployed **bio-sensors** for airborne anthrax detection.
4. The hospital surveillance system and the bio-sensors are not perfect: both false alarms and missed outbreaks are possible.

Define the following variables:

F: Flu season
A: Anthrax attack has occurred
R: Respiratory illness increased
B: Bio-sensors detect anthrax
H: Hospital surveillance alert

Building a Bayes Net

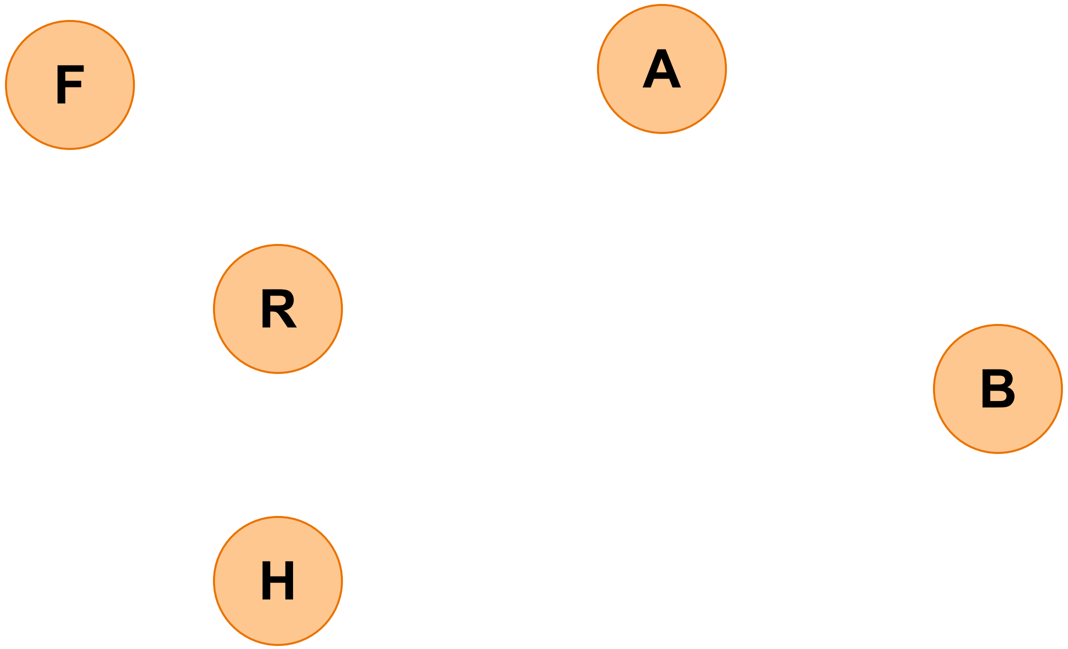
F: Flu season
A: Anthrax attack has occurred
R: Respiratory illness increased
B: Bio-sensors detect anthrax
H: Hospital surveillance alert



Step 1: Choose a set of relevant variables and represent each variable by a node.

Building a Bayes Net

F: Flu season
A: Anthrax attack has occurred
R: Respiratory illness increased
B: Bio-sensors detect anthrax
H: Hospital surveillance alert



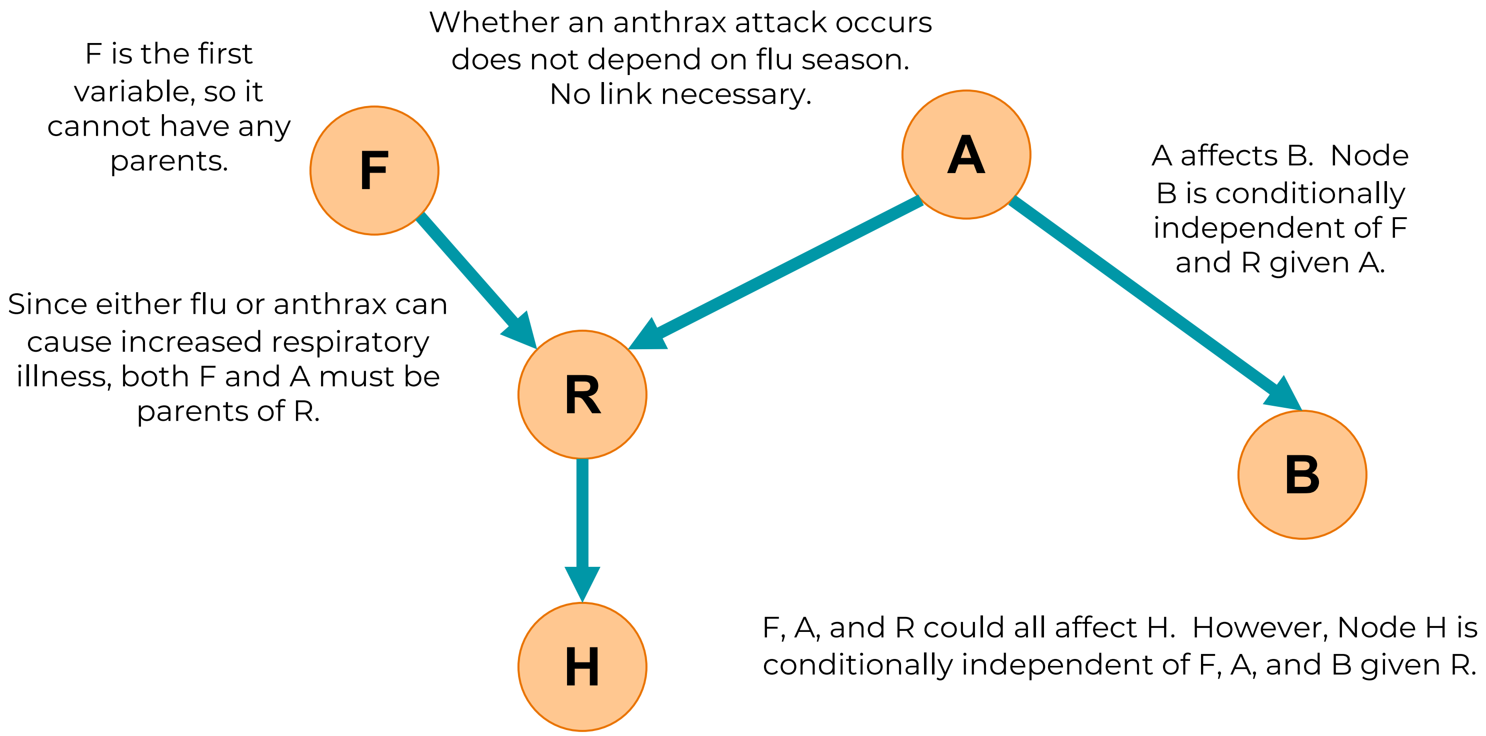
Step 2: Choose an ordering for the variables $X_1..X_M$, such that if X_i influences X_j, then $i < j$.

Hint: put environmental and event variables first, then latent variables, then observations.

Any ordering will produce a valid Bayes Net structure, but using the causal information will produce more compact (fewer links) and more interpretable structures.

Building a Bayes Net

F: Flu season
A: Anthrax attack has occurred
R: Respiratory illness increased
B: Bio-sensors detect anthrax
H: Hospital surveillance alert



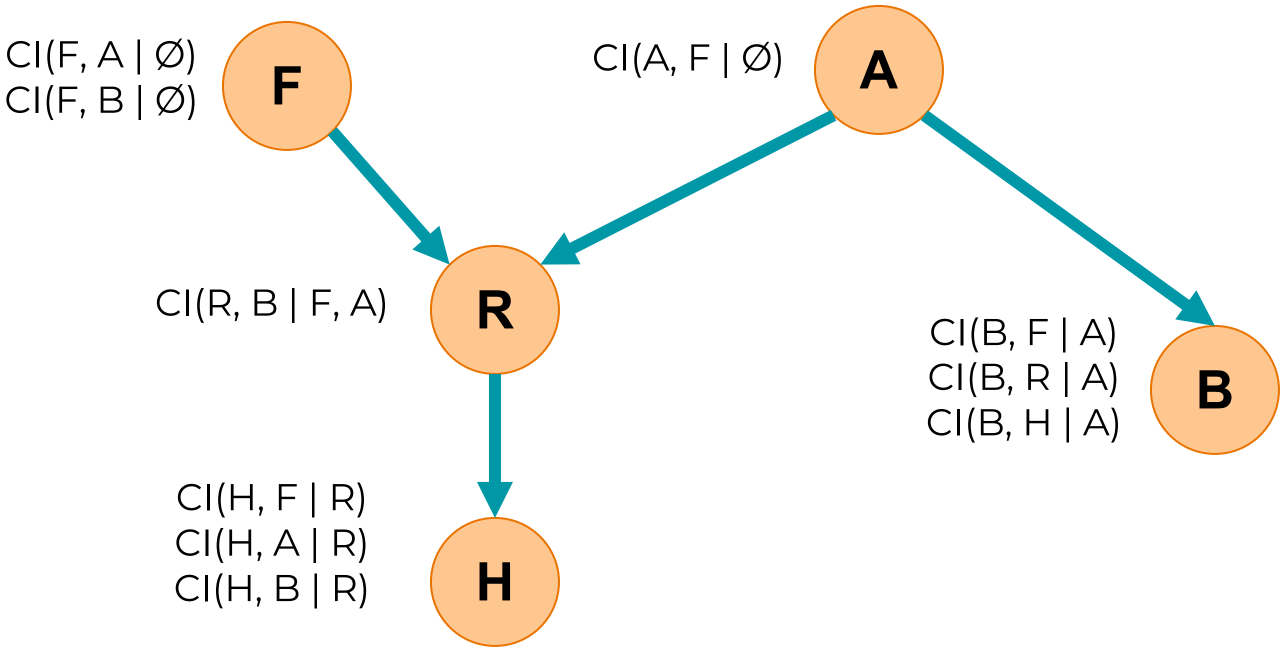
<u>Step 3</u>: Add links.

- For each variable X_i (for i = 2…M), choose a minimal subset of parents from $X_1..X_{i-1}$, such that X_i is conditionally independent of the rest of $X_1..X_{i-1}$ given its parents.

NYU

Interpreting a Bayes Net Structure

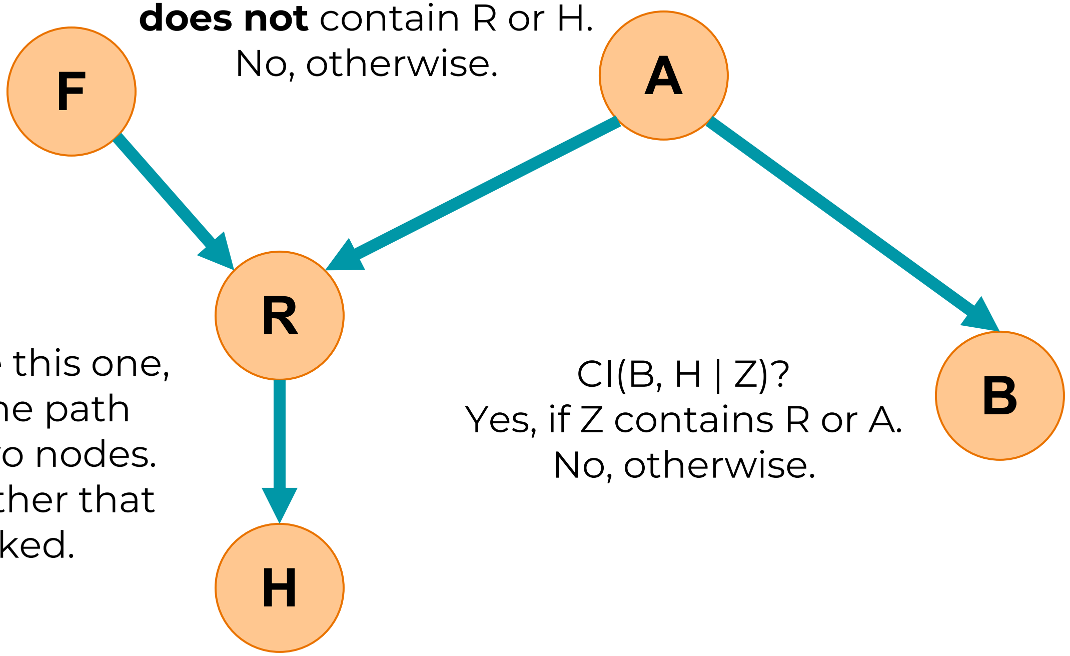
F: Flu season
A: Anthrax attack has occurred
R: Respiratory illness increased
B: Bio-sensors detect anthrax
H: Hospital surveillance alert



- Key property: each node is conditionally independent of all its non-descendants in the tree, given its parents.
- Two unconnected variables may still be correlated.
- Whether any two variables are conditionally independent can be deduced from a Bayes Net using “d-separation.”

D-separation

CI(F, B | Z)?
Yes, if Z contains A, or **does not** contain R or H.
No, otherwise.



In a **polytree** like this one, there is only one path between any two nodes. Just check whether that path is blocked.

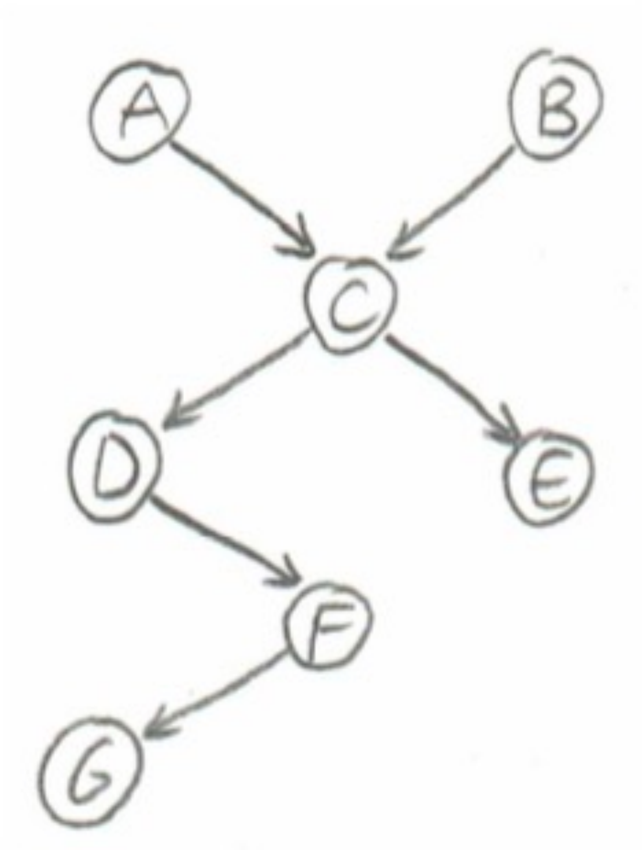
CI(B, H | Z)?
Yes, if Z contains R or A.
No, otherwise.

F: Flu season
A: Anthrax attack has occurred
R: Respiratory illness increased
B: Bio-sensors detect anthrax
H: Hospital surveillance alert

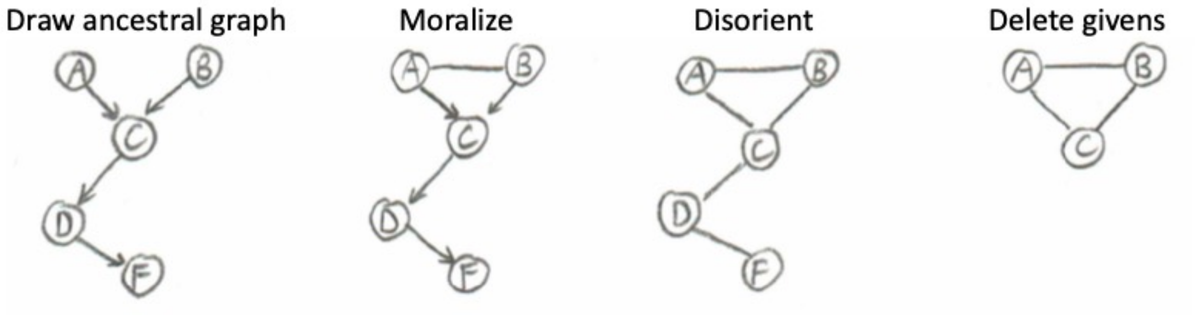
- Nodes X_i and X_j are conditionally independent given a set of nodes Z if every undirected path between X_i and X_j is "blocked" by at least one of the following: $(Z_i \in Z, W_i \notin Z)$ (and no descendent of W_i is in Z)

NYU

D-separation



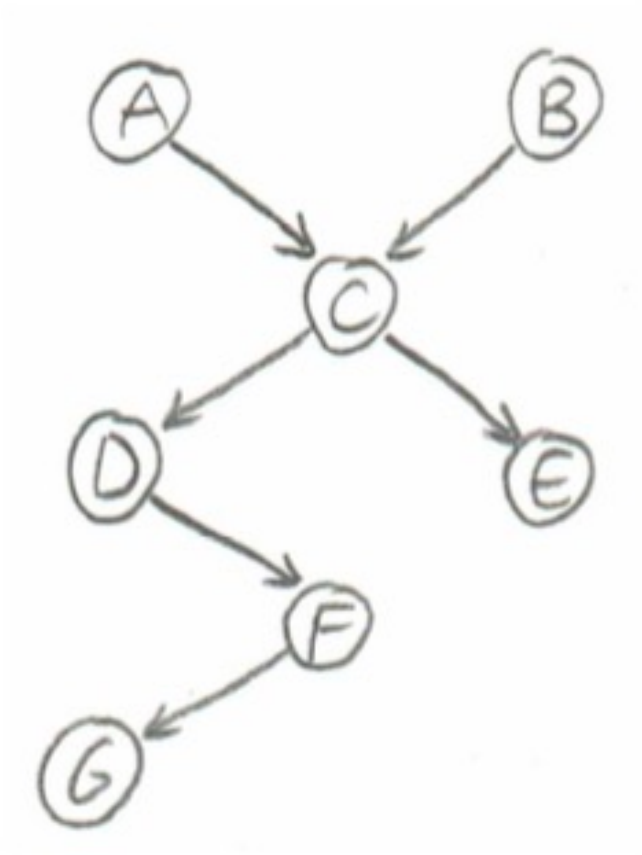
1. Are A and B conditionally independent, given D and F?



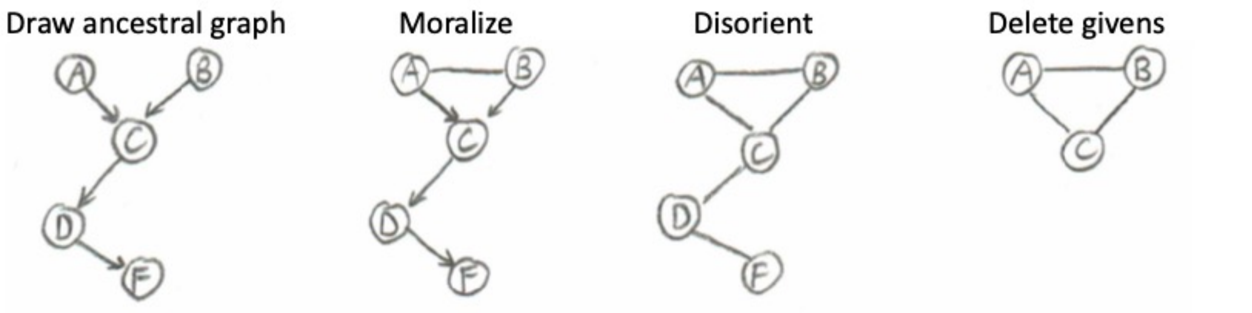
Answer: No, A and B are connected, so they are not required to be conditionally independent given D and F.

Example is from: http://web.mit.edu/jmn/www/6.034/d-separation.pdf

D-separation

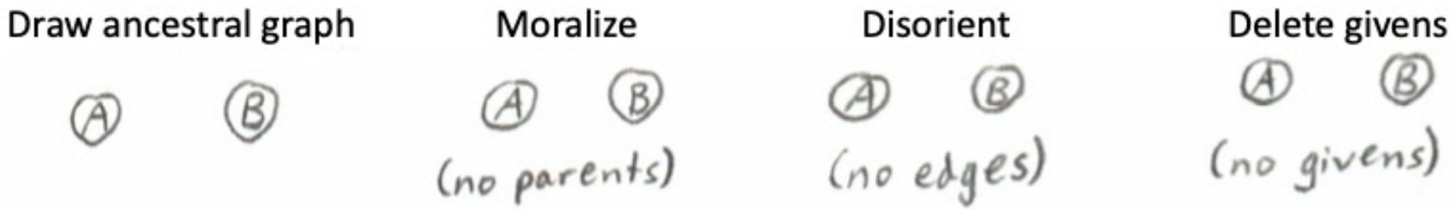


1. Are A and B conditionally independent, given D and F?



Answer: No, A and B are connected, so they are not required to be conditionally independent given D and F.

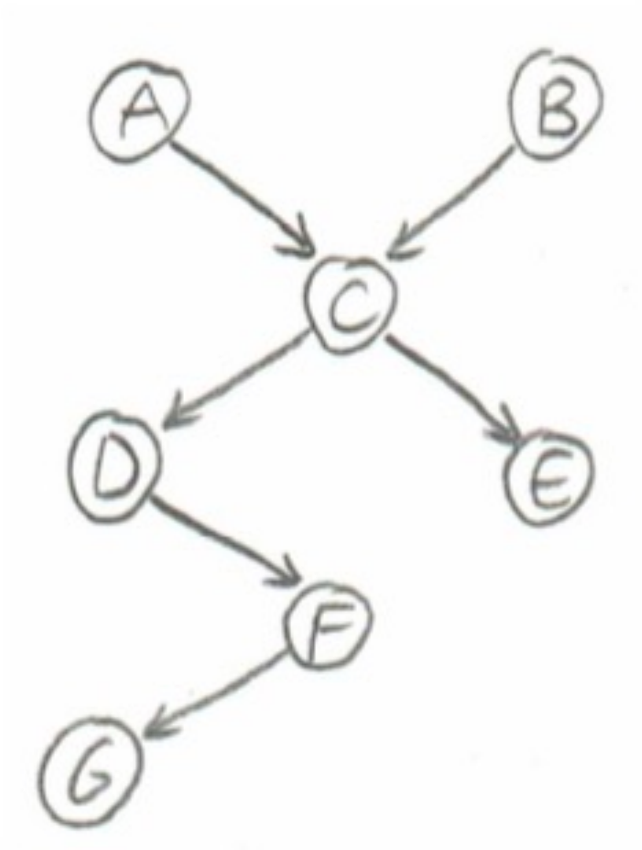
2. Are A and B marginally independent?



Answer: Yes, A and B are not connected, so they are marginally independent.

Example is from: http://web.mit.edu/jmn/www/6.034/d-separation.pdf

D-separation

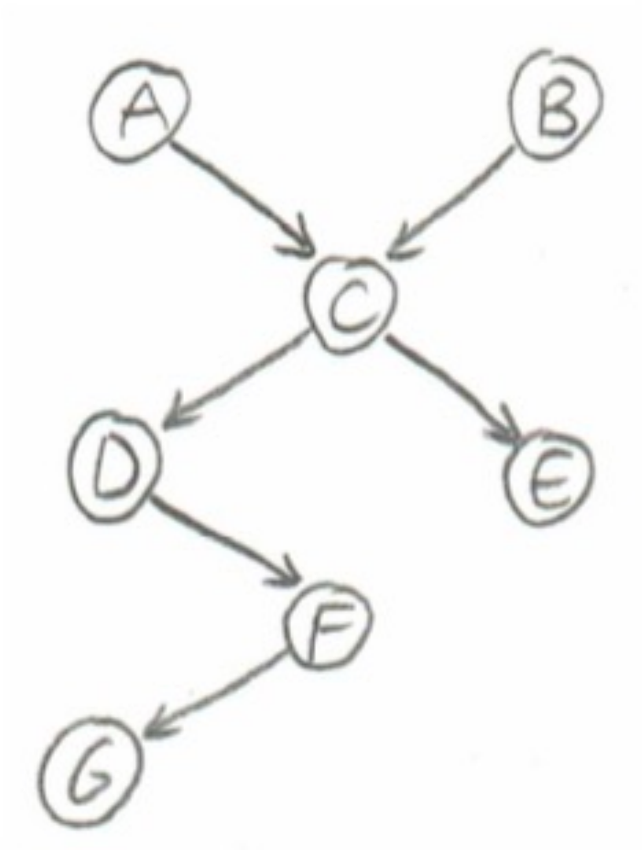


3. Are A and B conditionally independent, given C?

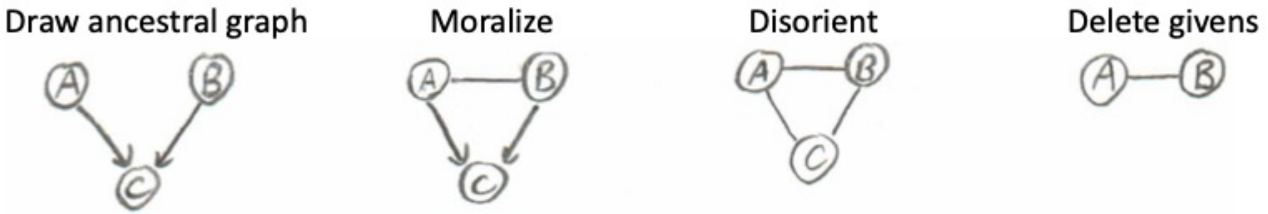
Answer: No, A and B are connected, so they are not required to be conditionally independent given C.

Example is from: http://web.mit.edu/jmn/www/6.034/d-separation.pdf

D-separation

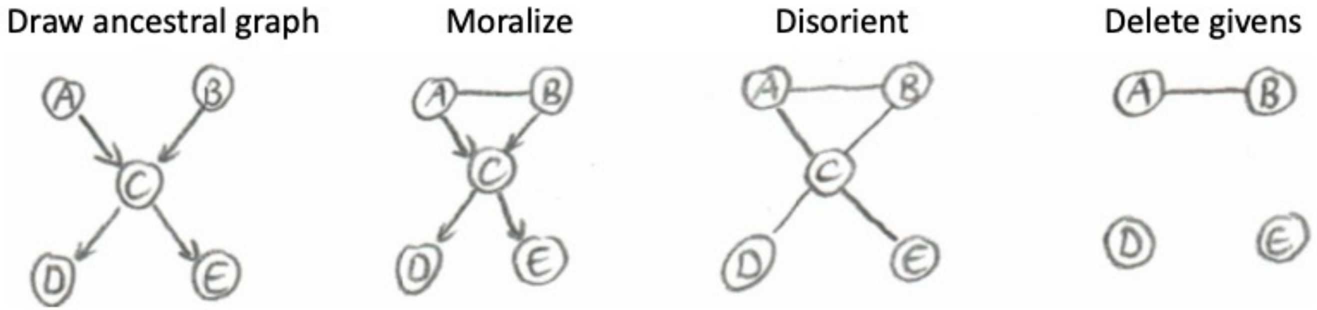


3. Are A and B conditionally independent, given C?



Answer: No, A and B are connected, so they are not required to be conditionally independent given C.

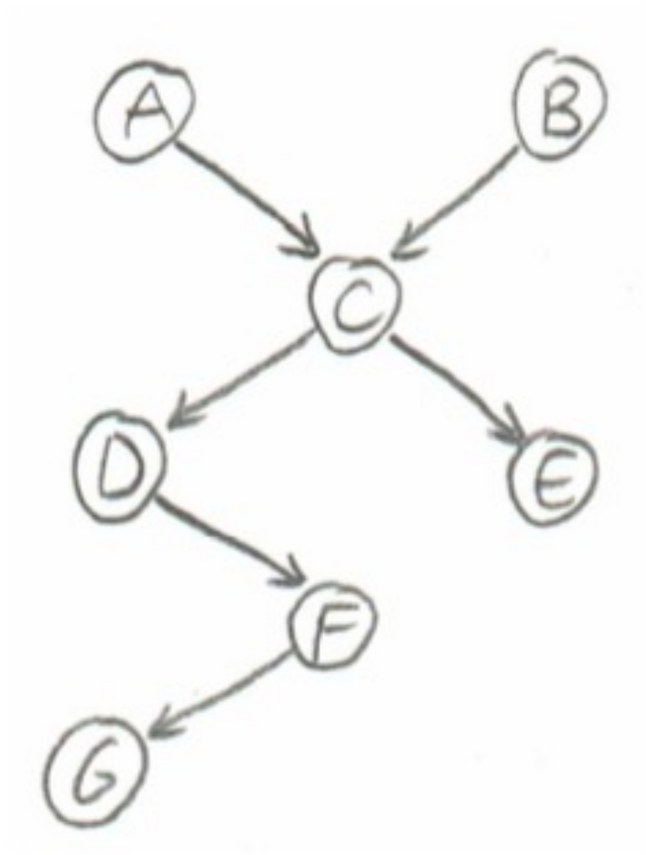
4. Are D and E conditionally independent, given C?



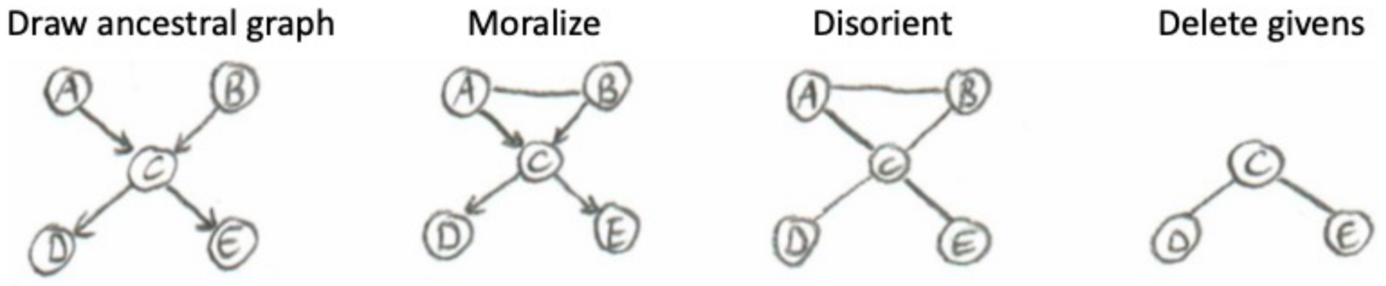
Answer: Yes, D and E are not connected, so they are conditionally independent given C.

Example is from: http://web.mit.edu/jmn/www/6.034/d-separation.pdf

D-separation



5. Are D and E conditionally independent, given A and B?

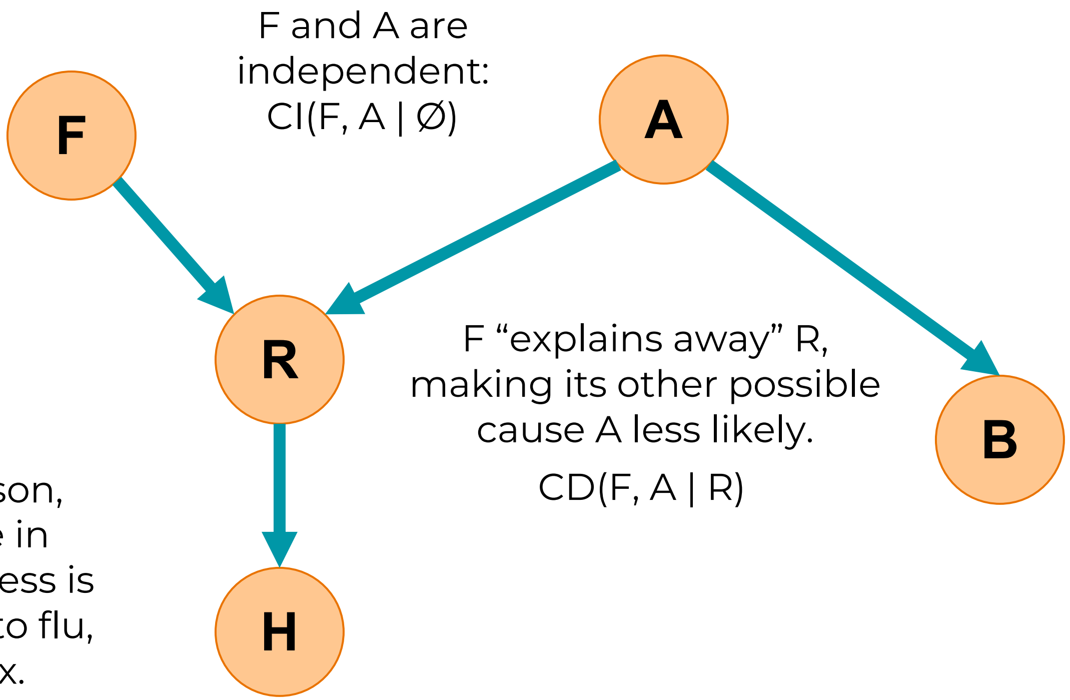


Answer: No, D and E are connected (via a path through C), so they are not required to be conditionally independent given A and B.

Example is from: http://web.mit.edu/jmn/www/6.034/d-separation.pdf

“Explaining away”

What if we know that respiratory illness has increased and want to know whether an anthrax attack has occurred?



If it is flu season, the increase in respiratory illness is probably due to flu, not anthrax.

F: Flu season
A: Anthrax attack has occurred
R: Respiratory illness increased
B: Bio-sensors detect anthrax
H: Hospital surveillance alert

- Nodes X_i and X_j are conditionally independent given a set of nodes Z if every undirected path between X_i and X_j is “blocked” by at least one of the following: $(Z_i \in Z, W_i \notin Z)$ (and no descendent of W_i is in Z)

NYU

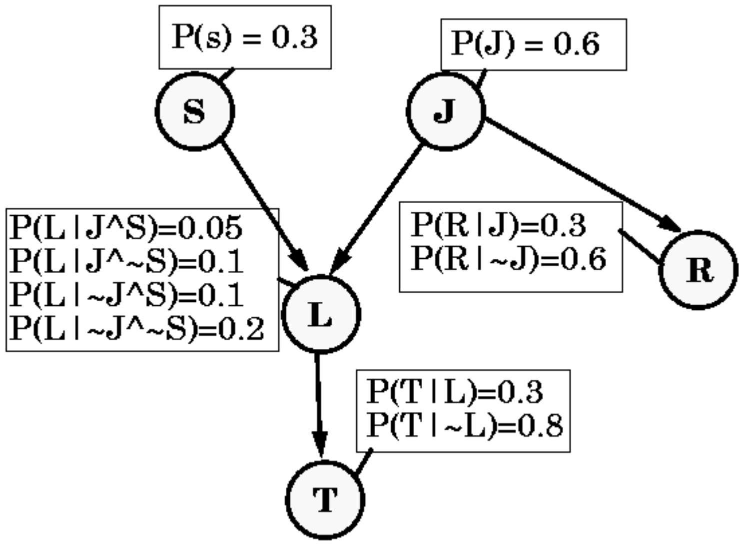
Where Are We Now?

- We can build a Bayesian network by hand, specifying the structure and the conditional probability tables.
- The network structure represents the conditional dependencies and independencies between variables and can also have a causal interpretation.
- Bayes Nets are a more compact representation of the joint probability distribution: we only need to store a number of probabilities exponential in the number of parents per node, not the total number of nodes.

- We will now answer two main questions:
    - How can we use Bayes Nets for probabilistic **inference**? ("What is the probability of an anthrax attack, given that the hospital surveillance system and bio-sensors both alerted?")
    - How can we **learn** the Bayes Net structure and parameters automatically using a large training dataset?

Bayes Net Inference

Question 1: How do joint probabilities be computed using Bayes Net?



Compute Pr(S, ~J, L, ~R, T)

Answer: use conditional independence.

$\Pr(X_1..X_M) = \prod_{i=1..M} \Pr(X_i \mid \text{Parents}(X_i))$

Pr(S, ~J, L, ~R, T) =

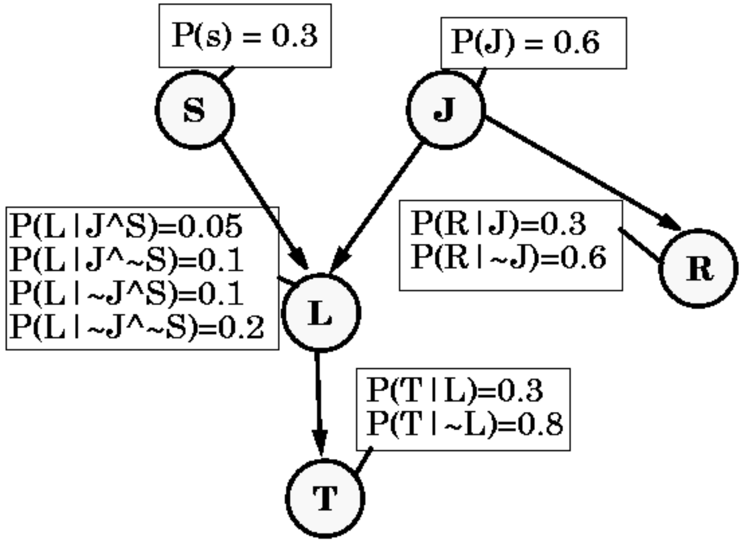
Pr(S) Pr(~J) Pr(L | ~J, S) Pr(~R | ~J) Pr(T | L) =

(0.3) (0.4) (0.1) (0.4) (0.3) = 0.00144.

We can efficiently compute the joint probability of any given assignment of values to variables.

Bayes Net Inference

Question 2: How to compute arbitrary conditional probabilities?



Compute Pr(S, T | ~J, L)

Express the conditional probability as a ratio:

Pr(S, T | ~J, L) = Pr(S, ~J, L, T) / Pr (~J, L)

Express numerator and denominator as sums of joint probabilities:

Pr(S, ~J, L, T) = Pr(S, ~J, L, R, T) + Pr(S, ~J, L, ~R, T)

Pr(~J, L) = Pr(S, ~J, L, R, T) + Pr(S, ~J, L, R, ~T) + Pr(~S, ~J, L, R, T) + Pr(~S, ~J, L, R, ~T) + Pr(S, ~J, L, ~R, T) + Pr(S, ~J, L, ~R, ~T) + Pr(~S, ~J, L, ~R, T) + Pr(~S, ~J, L, ~R, ~T)

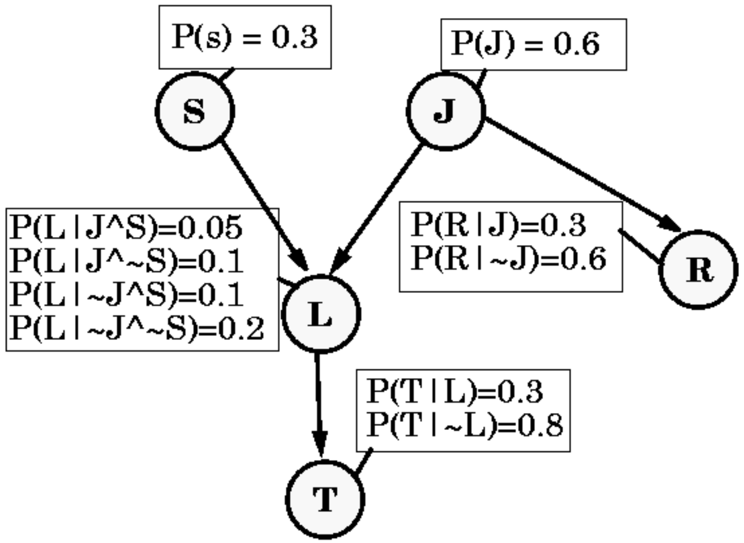
$$\Pr(X \mid Y) = \frac{\sum_{\text{joint entries matching X and Y}} \Pr(\text{joint entry})}{\sum_{\text{joint entries matching Y}} \Pr(\text{joint entry})}$$

You have m binary variables in your Bayes Net, and expression Y uses k variables. How many rows of the joint do you have to calculate?

NYU

Bayes Net Inference

Question 2: How to compute arbitrary conditional probabilities?



Compute Pr(S, T | ~J, L)

Express the conditional probability as a ratio:

Pr(S, T | ~J, L) = Pr(S, ~J, L, T) / Pr (~J, L)

Good news: we can sometimes simplify the probability calculations.

Pr(S, ~J, L, T) = Pr(S) Pr(~J) Pr(L | S, ~J) Pr(T | L)

Pr(~J, L) = Pr(~J) Pr(L | ~J)

Pr(L | ~J) = Pr(L | ~J, S) Pr(S) + Pr(L | ~J, ~S) Pr(~S)

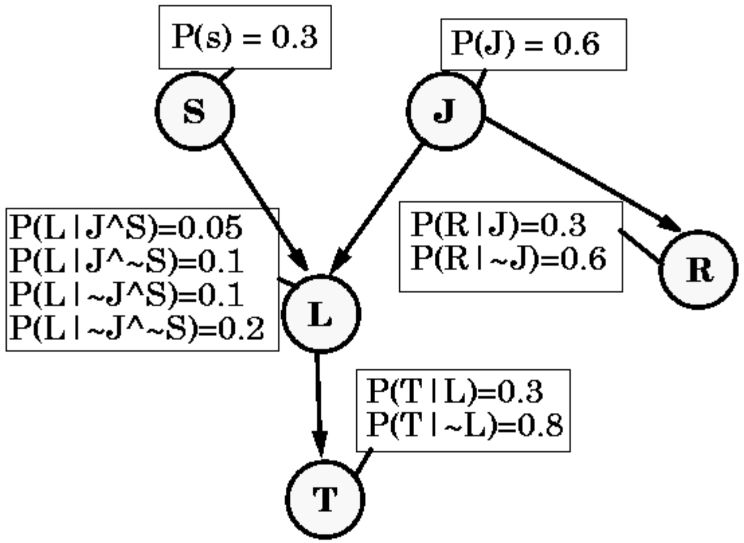
$$\Pr(X \mid Y) = \frac{\sum_{\text{joint entries matching X and Y}} \Pr(\text{joint entry})}{\sum_{\text{joint entries matching Y}} \Pr(\text{joint entry})}$$

You have m binary variables in your Bayes Net, and expression Y uses k variables. How many rows of the joint do you have to calculate?

NYU

Bayes Net Inference

Question 2: How to compute arbitrary conditional probabilities?



Express the conditional probability as a ratio:

Pr(S, T | ~J, L) = Pr(S, ~J, L, T) / Pr (~J, L)

Good news: we can sometimes simplify the probability calculations.

Pr(S, ~J, L, T) = Pr(S) Pr(~J) Pr(L | S, ~J) Pr(T | L)

Pr(~J, L) = Pr(~J) Pr(L | ~J)

Pr(L | ~J) = Pr(L | ~J, S) Pr(S) + Pr(L | ~J, ~S) Pr(~S)

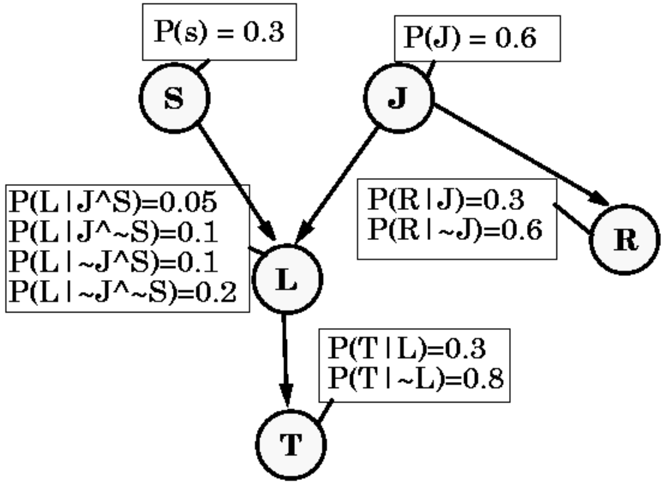
$$\Pr(X \mid Y) = \frac{\sum_{\text{joint entries matching X and Y}} \Pr(\text{joint entry})}{\sum_{\text{joint entries matching Y}} \Pr(\text{joint entry})}$$

Bad news: doing exact inference for Bayes Nets is computationally hard.

But it's tractable in some special cases (e.g. , trees). We can also do efficient **approximate** inference.

You have m binary variables in your Bayes Net, and expression Y uses k variables. How many rows of the joint do you have to calculate?

Approximate Inference



Compute Pr(S, T | ~J, L)

<u>To sample from the joint distribution of S, J, L, R, T</u>

1. Randomly choose S (True with probability 0.3)
2. Randomly choose J (True with probability 0.6)
3. Randomly choose L. The probability that L is true depends on the assignments of S and J. If steps 1 and 2 produced S = True, J = False, then probability that L is true is 0.1.
4. Randomly choose R (Probability depends on J)
5. Randomly choose T (Probability depends on L)

<u>To estimate any conditional probability Pr(X | Y)</u>

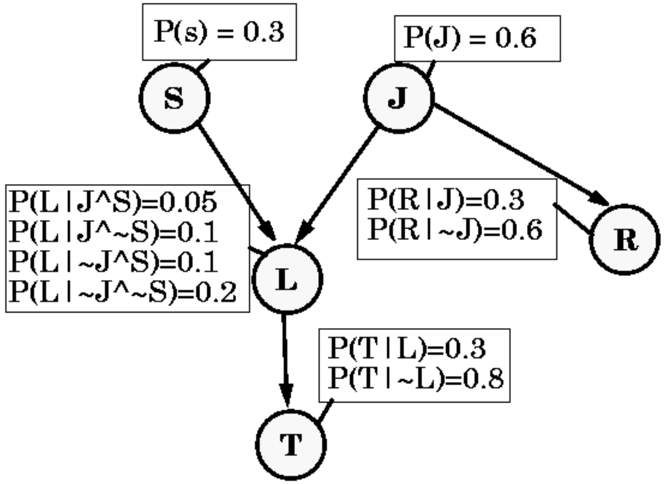
1. Draw N samples from the joint distribution
2. Count N_Y = number of samples where Y is true
3. Count N_{XY} = number of samples where both X and Y are true.
4. Calculate Pr(X | Y) = N_{XY} / N_Y

<u>Problem</u>: many of these N samples are wasted because Y is false.

<u>Solution</u>: only generate samples where Y is true, but weight them so that this property still holds.

For large N, the ratio of N_{XY} to N_Y converges to the true probability Pr(X | Y).

Likelihood Weighted Sampling



Compute Pr(S, T | ~J, L)

<u>Problem:</u> many of these N samples are wasted because Y is false.

<u>Solution</u>: only generate samples where Y is true, but weight them so that this property still holds.

<u>Choosing a sample from the joint, subject to ~J, L</u>:

0. Set initial weight w = 1.
1. Randomly choose S (True with probability 0.3)
2. Multiply w by Pr(J = False) = 0.4. Set J = False.
3. Multiply w by Pr(L = True) given the current assignments of S and J. For example, if steps 1-2 produced S = True and J = False then multiply w by 0.1. Set L = True.
4. Randomly choose R (True with probability 0.6)
5. Randomly choose T (True with probability 0.3)

<u>To estimate any conditional probability Pr(X | Y)</u>

1. Draw N samples from the joint distribution, subject to the constraint Y.
2. For each sample and its weight w ≤ 1:
   Increment N_Y by w.
   If X is true, increment N_{XY} by w.
3. Calculate Pr(X | Y) = N_{XY} / N_Y

Bayes Net Inference

Now, we know how to perform inference with a Bayes Net. This is great if we already have the network structure and parameters specified by an expert... but what if we want to **learn** the Bayes Net from data?

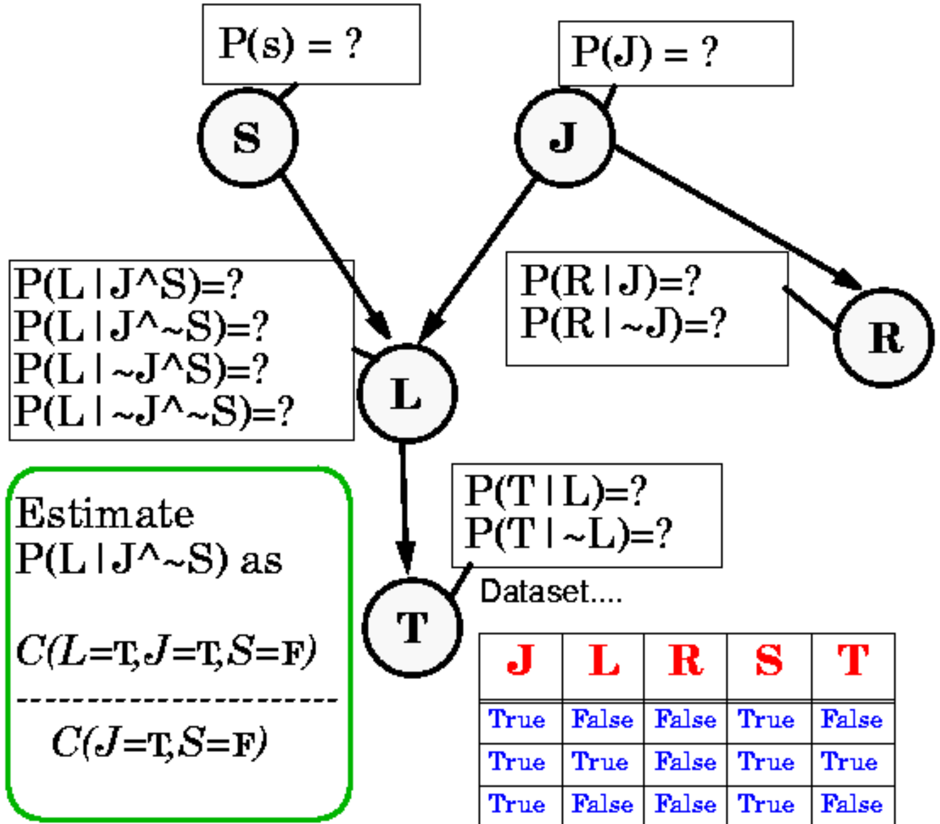
1. Parameter search
2. Structure search

Bayes Net Parameter Learning

Given a Bayesian network structure and a training dataset, we can learn the parameters of each node by maximum likelihood.

Given node X with parent nodes $Q_1..Q_m$, we learn the conditional distribution of X for each distinct combination of parent values.

For example, if $Q_1..Q_m$ are all binary, there are 2^m distributions to learn.



Dataset....

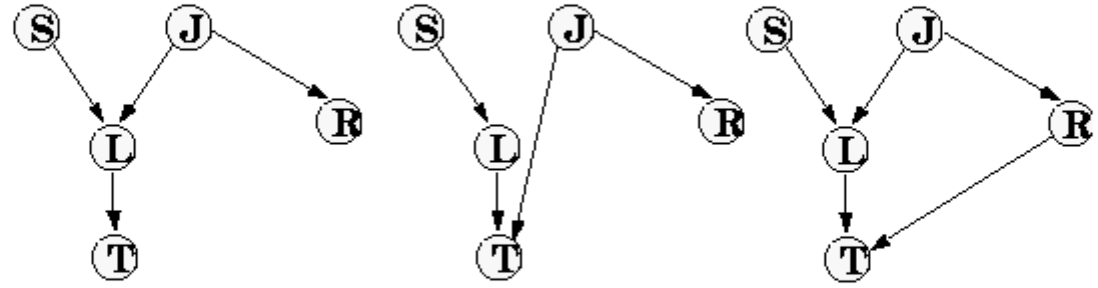
J	L	R	S	T
True	False	False	True	False
True	True	False	True	True
True	False	False	True	False
False	True	True	False	False
True	True	False	True	True
:	:	:	:	:

We learn the distribution $\Pr(X \mid Q_1 = v_1, Q_2 = v_2, ..., Q_m = v_m)$ by looking at the subset of training records with the given parent values and computing the proportion with each X value.

Bayes Net Structure Search

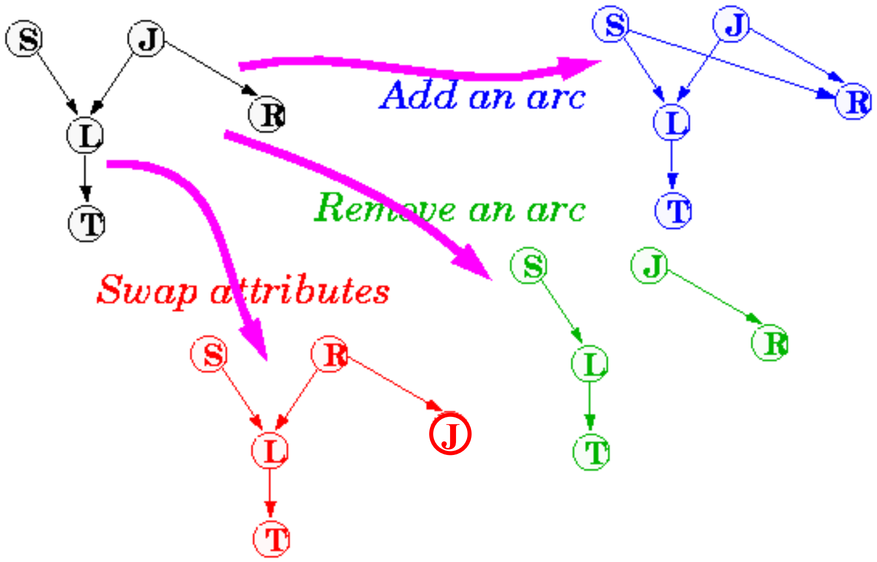
How to automatically find the Bayesian network structure that best fits the data?

This is a hard **state-space search** problem: use hill-climbing or simulated annealing with restarts.



What moveset to use?
How to score a structure?

Here's one possible moveset:



To score a structure, learn all parameters from the training data by maximum likelihood.

Then compute the log-likelihood of the training dataset given the structure and parameters.

Score = log-likelihood – λk, where λ is a constant and k is the total number of parameters.

Bayes Net Structure Search

This "score-based" learning approach can find Bayes Net structures that accurately capture the conditional independence relationships in the data. But what if we want to be able to interpret edges **causally**?

One option: incorporate prior knowledge. The "K2" structure learning algorithm is a hill-climbing approach that relies on a **causal partial ordering** of the variables and only allows edges from X_i to X_j for $i < j$.

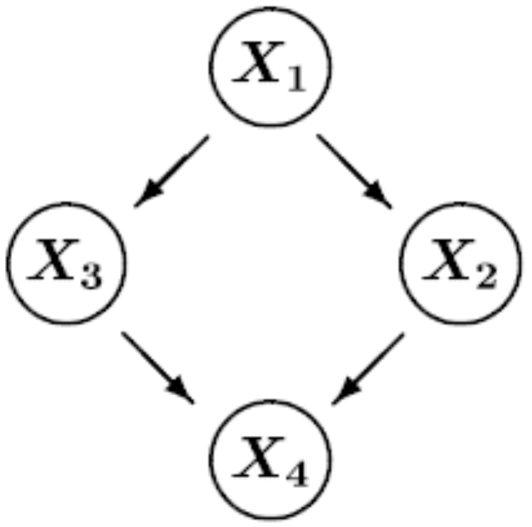
NYU

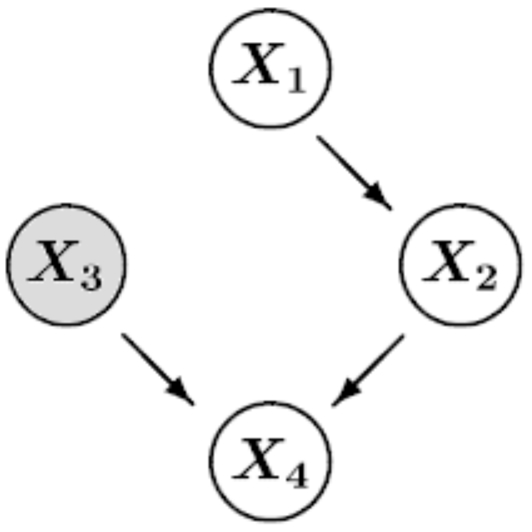
Causal Bayes Nets

- CBN = BN, where edge X → Y is assumed to indicate that X is a direct cause of Y.
- <u>Markov condition</u>: given its parents (causes), each variable is conditionally independent of its non-descendants (non-effects).
  - $\Pr(X_1..X_N) = \prod \Pr(X_i \mid \text{Parents}(X_i))$
  - All this is just like a regular Bayes Net.
- We can also reason about <u>interventions</u>:

  $\Pr(X_i \mid \text{Parents}(X_i), do(X = x)) = 1\{X_i = x_i\}$ for intervened variables ($X_i = x_i$ in X)

  $\Pr(X_i \mid \text{Parents}(X_i), do(X = x)) = \Pr(X_i \mid \text{Parents}(X_i))$ for non-intervened variables (X_i not in X).



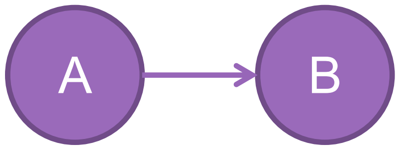


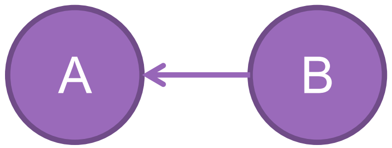
After intervention $do(X_3 = x_3)$

NYU

Causal Structure Learning from Observational Data

Key thing to keep in mind: we cannot distinguish between networks that have the same conditional independence relationships but different causation.

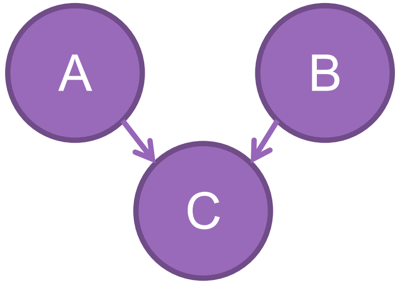




- We get an equivalence class of structures (some edges may be directed, some undirected).
- If we want to do better, need prior knowledge, additional assumptions, or different data (time series, intervention).
- How can we ever get a directed edge?

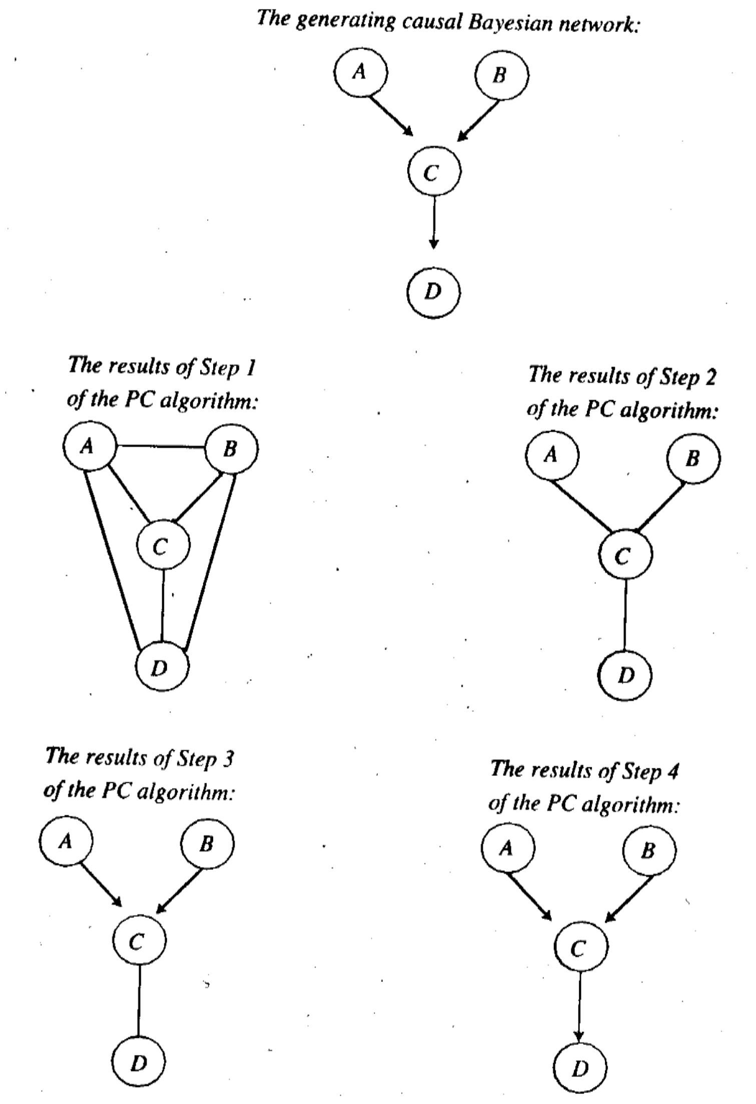
  Answer: V-structures!

  CI(A,B) but CD(A,B | C)



Constraint-Based Structure Learning

- Relies on results of statistical tests for conditional independence between variables; finds an equivalence class of structures satisfying these constraints.
- PC algorithm (Spirtes et al.):
  - Start with complete undirected graph.
  - For each pair of variables X and Y, delete the edge if they are conditionally independent given any subset of the other vars.
  - For any triplet X – Y – Z without X – Z, if X and Z are conditionally dependent given Y (and any subset of other vars), replace with V-structure X → Y ← Z.
  - Use this information to direct other edges (avoid creating directed cycles and additional V-structures).



Assumptions Made by PC (and most other causal learning methods)

- **Causal Markov:** a variable is probabilistically independent of its non-descendants (non-effects) and conditional on its direct causes.
  - Permits inference from dependence to causal connection.
- **Causal Faithfulness:** conditional independence between variables does not occur by accident (e.g., via "canceling out" settings of parameters), but only because of the lack of a (direct) causal relationship.
  - Permits inference from independence to causal separation.
- **Causal Sufficiency:** no unmeasured common causes
- **Acyclicity:** no variable is an (indirect) cause of itself.
  - It would be violated, for example, if $X \rightarrow Y$, $Y \rightarrow Z$, and $Z \rightarrow X$.

(Some text on this slide was borrowed from Fredrick Eberhardt's presentation, "All of Causal Discovery", and from Stephen Fancsali's doctoral dissertation.)

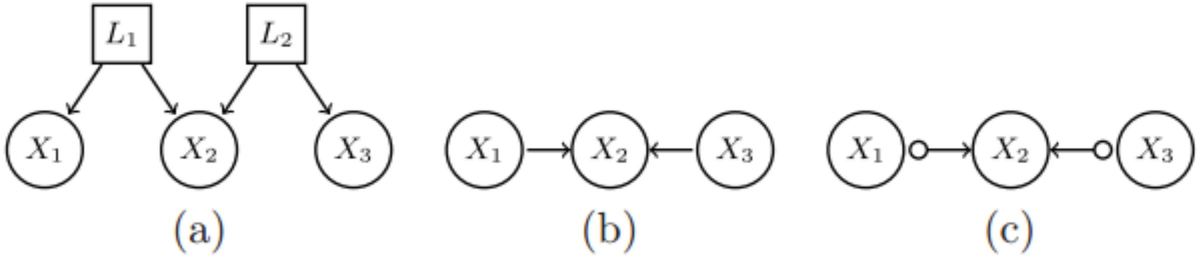
Assumptions Made by PC

Two ways to proceed from here:

Weaker assumptions, such as allowing unmeasured common causes, lead to a larger equivalence class of structures. (Fewer edges can be oriented)

Stronger assumptions, such as parametric model assumptions, lead to a smaller equivalence class of structures. (More edges can be oriented)

Fast Causal Inference (FCI)



Like PC but handles selection bias and unobserved confounders.

Fewer assumptions (does not require causal sufficiency).
But results in a larger equivalence class of structures.

Learns a **partial ancestral graph** (PAG): A → B means A is an ancestor of B.
Can sometimes distinguish this from unobserved common cause A ← X → B. Cannot ever rule out unobserved intervening causes A → X → B.

Not very fast, doesn't scale to many variables.

NYU

Some text and the figure on this slide were borrowed from Seth Flaxman's presentation, "Algorithmic Approaches to Causal Inference")

Causal Orientation Methods

With **additional parametric model assumptions**, can distinguish between X → Y and Y → X even without the presence of a third variable. These methods work by exploiting **asymmetries** in the shapes of the conditional probability densities.

Statnikov et al. (2012) does a big bake-off to compare many of these methods for a genomics application (X: transcription factor; Y: target gene).

<u>Example 1</u>: LiNGaM (assumes **linear** model with **non-Gaussian** errors)

Estimate models Y= bX + ε and X = b'Y + ε', where ε and ε' are independent. Choose direction with smaller slope b.

<u>Example 2</u>: ANM (assumes **non-linear** model with **additive noise**)

If x and y are dependent:

Estimate residuals from non-linear regression y = f(x) + ε
Check whether residuals and x are independent
Independent? Accept model x → y

Repeat with x and y switched.

Some text and the figure on this slide were borrowed from Seth Flaxman's presentation, "Algorithmic Approaches to Causal Inference")

The Many Uses of Bayes Nets

Bayes Nets provide a useful graphical representation of the probabilistic (+ causal) relationships between variables.

Automatic learning of Bayes net structure can be used for exploratory analysis of datasets with many attributes.

We can often improve the performance of model-based classification by moving from Naïve Bayes to Bayes Nets.

We can also use Bayes Nets to detect **anomalies**, by finding points with low probabilities given the Bayes Net.

Bayes Nets provide a compact structure that enables us to efficiently compute probability distributions for any unobserved variables given observations of others.

Medical diagnosis Failure troubleshooting Drug discovery

Environmental modeling Computational biology Transportation

Lab Time

NYU

For the Next Week (Week 12)

1. Assignment 3

   Due: April 14, 2024 (11:59pm)

References

Bayes Nets:

- J. Pearl. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann, 1988.
- S. Russell and P. Norvig. AI: A Modern Approach, Ch. 15.
- Several excellent tutorials on Bayes Nets available at https://www.cs.cmu.edu/~./awm/tutorials/

Causality:

- P. Spirtes, C. Glymour, and R. Scheines. Causation, Prediction, and Search. MIT Press, 2000.
- J. Pearl. Causality: Models, Reasoning, and Inference. Cambridge University Press, 2009.
- P. Spirtes. Introduction to causal inference. Journal of Machine Learning Research 11: 1643-1662, 2010.
- M. Kalish, P. Buhlmann. Causal structure learning and inference: a selective review. Qual Technol Quant Manag. 11:3–21, 2014.
- A. Statnikov et al. New methods for separating causes from effects in genomics data. BMC Genomics 2012, 13(Suppl 8):S22. http://www.biomedcentral.com/1471-2164/13/S8/S22