Bayesian Networks

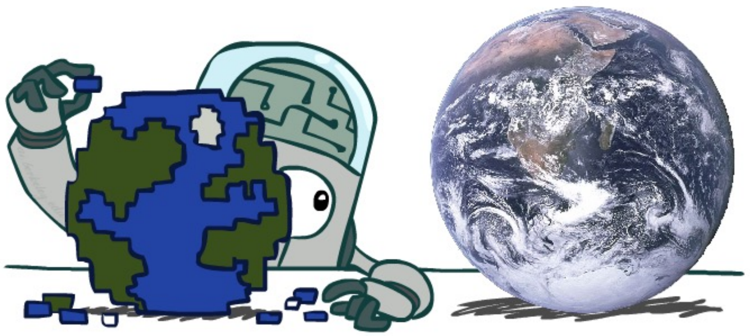
CE417: Introduction to Artificial Intelligence
Sharif University of Technology
Fall 2023

Soleymani

Slides have been adopted from Klein and Abdeel, CS188, UC Berkeley.

Probabilistic Models

- Models describe how (a portion of) the world works
- Models are always simplifications
  - May not account for every variable
  - May not account for all interactions between variables
  - " All models are wrong; but some are useful. " – George E. P. Box
- What do we do with probabilistic models?
  - We (or our agents) need to reason about unknown variables, given evidence
  - Example: explanation (diagnostic reasoning)
  - Example: prediction (causal reasoning)
  - Example: value of information



Independence

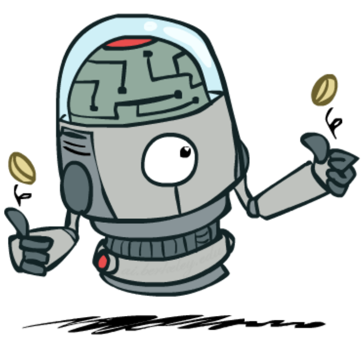
- Two variables are *independent* if:

$$\forall x, y : P(x, y) = P(x)P(y)$$

  - This says that their joint distribution *factors* into a product two simpler distributions
  - Another form:

$$\forall x, y : P(x|y) = P(x)$$

  - We write: $X \perp\!\!\!\perp Y$



- Independence is a simplifying *modeling assumption*
  - *Empirical* joint distributions: at best "close" to independent
  - What could we assume for {Weather, Traffic, Cavity, Toothache}?

Conditional Independence

- Unconditional (absolute) independence very rare (why?)
- *Conditional independence* is our most basic and robust form of knowledge about uncertain environments.
- X is conditionally independent of Y given Z $X \perp\!\!\!\perp Y | Z$

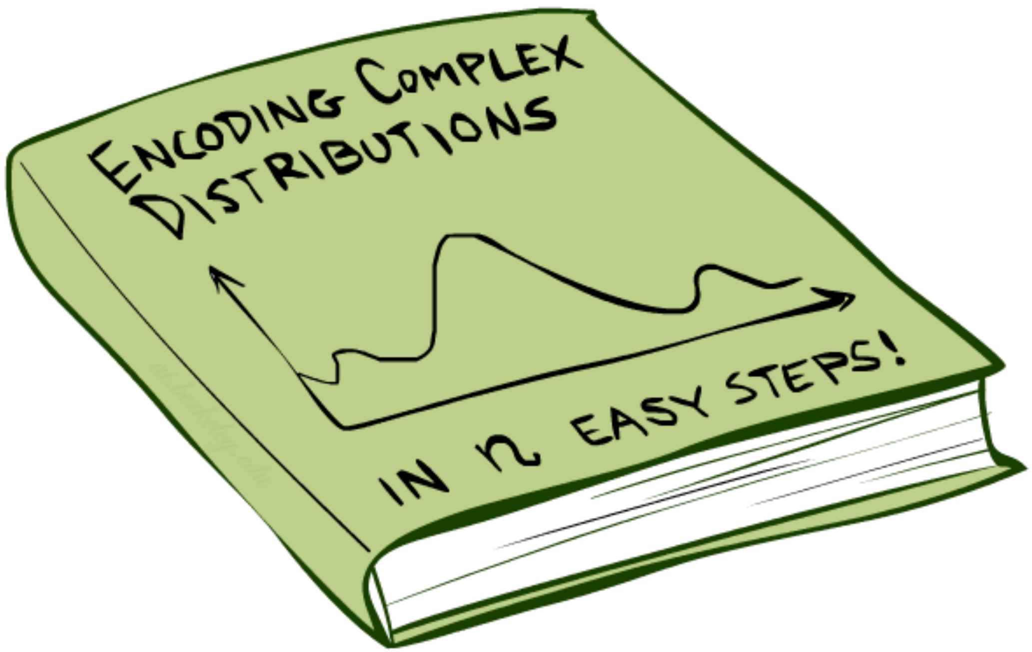
if and only if:

$$\forall x, y, z : P(x, y|z) = P(x|z)P(y|z)$$

or, equivalently, if and only if

$$\forall x, y, z : P(x|z, y) = P(x|z)$$

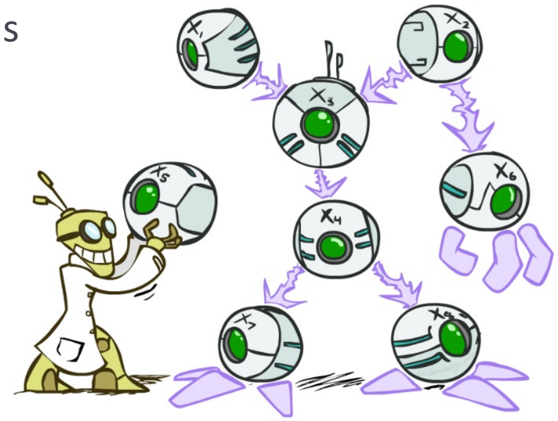
Bayesian Nets: Big Picture



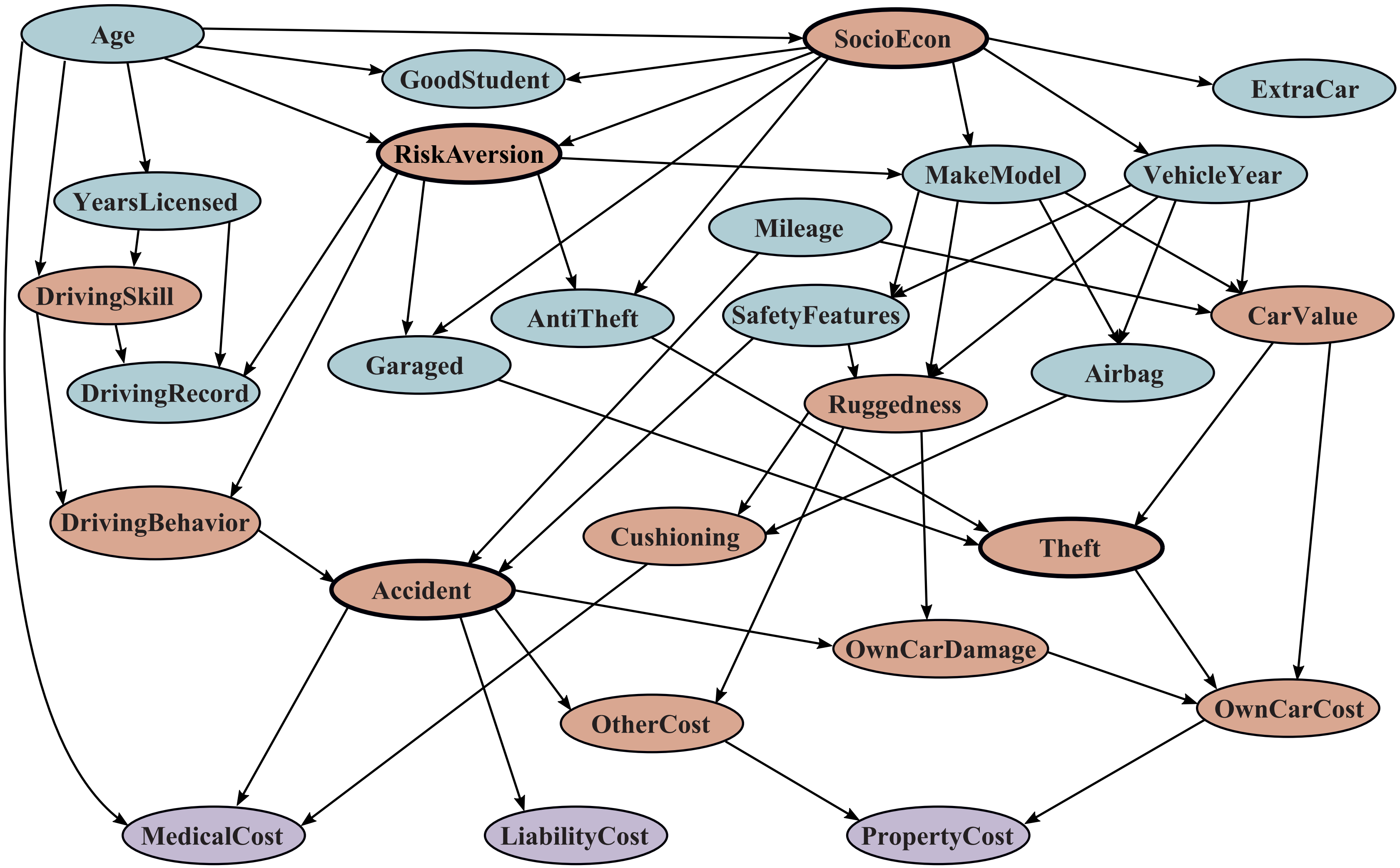
Bayesian Nets: Big Picture

- Two problems with using full joint distribution tables as our probabilistic models:
  - Unless there are only a few variables, the joint is WAY too big to represent explicitly
  - Hard to learn (estimate) anything empirically about more than a few variables at a time

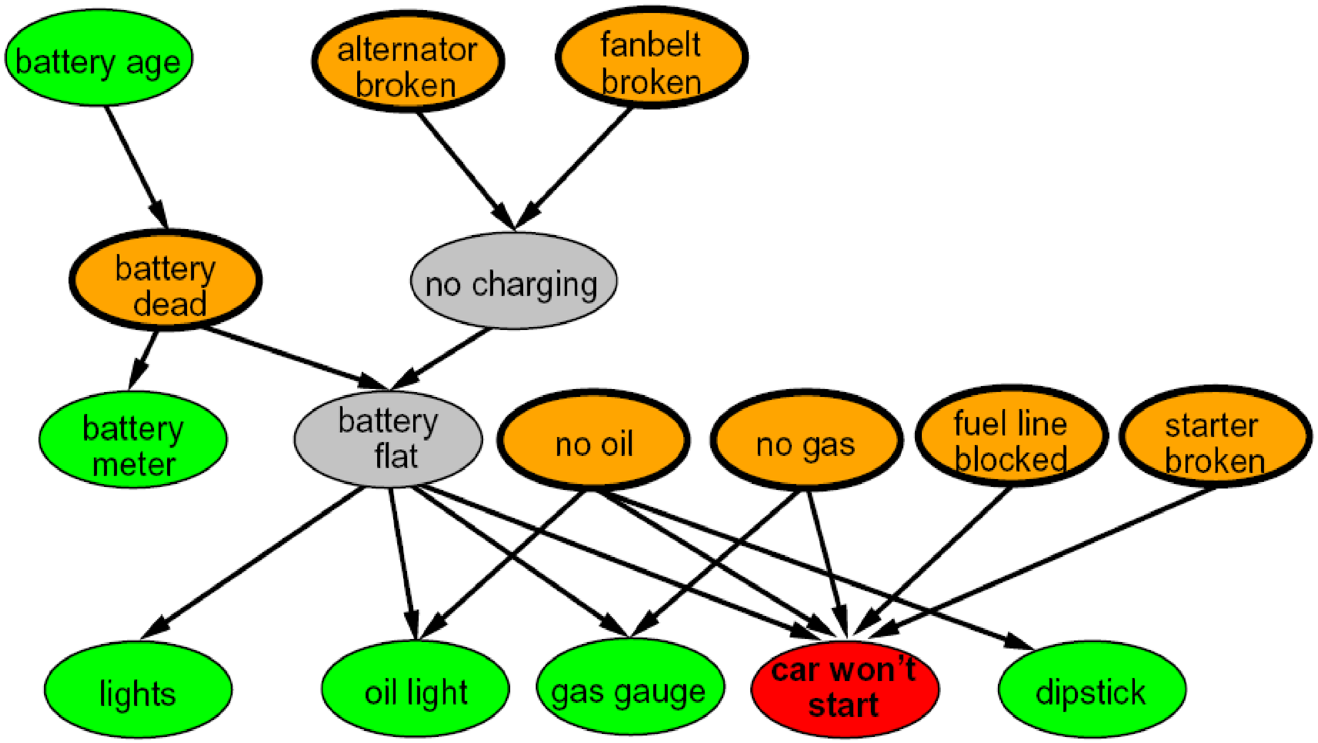
- Bayes' nets: a technique for describing complex joint distributions (models) using simple, local distributions (conditional probabilities)
  - More properly called graphical models
  - We describe how variables locally interact
  - Local interactions chain together to give global, indirect interactions



Example Bayesian Net: Car Insurance

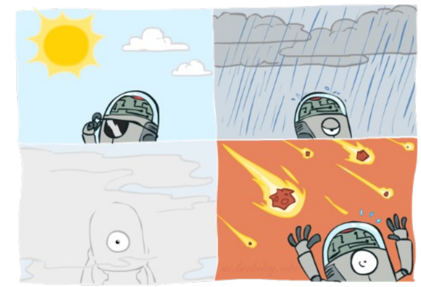


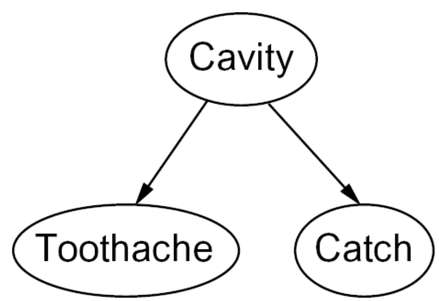
Example Bayesian Net: Car Won't Start

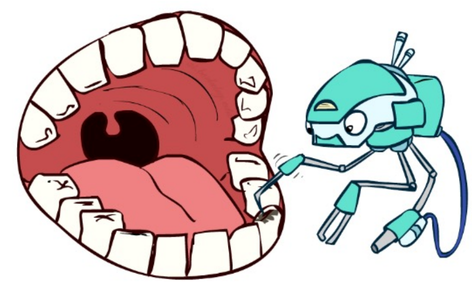


Graphical Model Notation

- Nodes: variables (with domains)
    - Can be assigned (observed) or unassigned (unobserved)
- Arcs: interactions
    - Similar to CSP constraints
    - Indicate "direct influence" between variables
    - Formally: encode conditional independence (more later)
- For now: imagine that arrows mean direct causation (in general, they don't!)

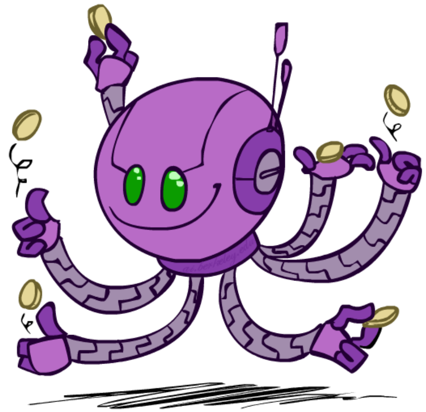






Example: Coin Flips

- N independent coin flips



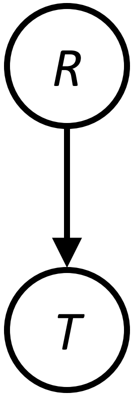
- No interactions between variables: absolute independence

Example: Traffic

- Variables:
  - R: It rains
  - T: There is traffic

- Model 1: independence

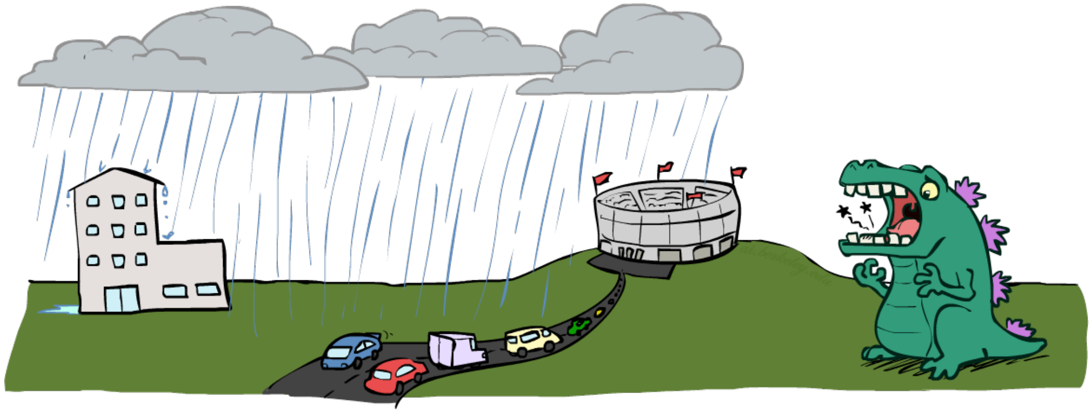
- Model 2: rain causes traffic



- Why is an agent using model 2 better?

Example: Traffic II

- Let's build a causal graphical model!
- Variables
  - T: Traffic
  - R: It rains
  - L: Low pressure
  - D: Roof drips
  - B: Ballgame
  - C: Cavity



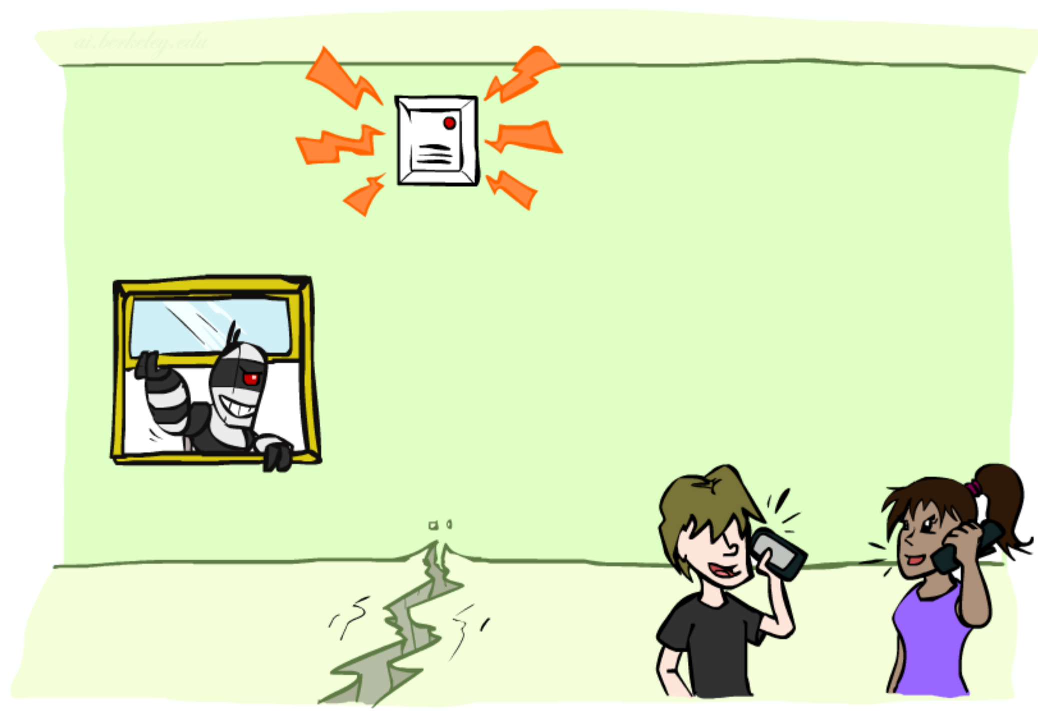
Example: Alarm Network

"A **burglar alarm**, respond occasionally to minor **earthquakes**.

Neighbors **John** and **Mary** call you when hearing the alarm.

John nearly always calls when hearing the alarm.

Mary often misses the alarm."



- Variables:
  - B: Burglary
  - A: Alarm goes off
  - M: Mary calls
  - J: John calls
  - E: Earthquake!

Bayes' Net Semantics

Bayesian Networks

- Importance of independence and conditional independence relationships (to simplify representation)

- Bayesian network: a graphical model to represent dependencies among variables
  - compact specification of full joint distributions
  - easier for human to understand

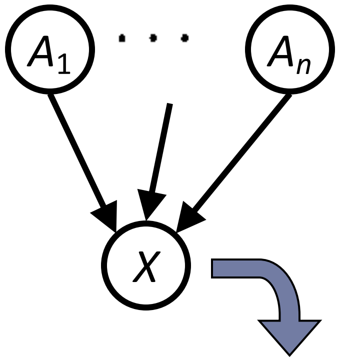
- **Bayesian network** is a directed acyclic graph
  - Each node shows a random variable
  - Each link from X to Y shows a "direct influence“ of X on Y (X is a parent of Y)
  - For each node, a conditional probability distribution $\mathbf{P}(X_i|Parents(X_i))$ shows the effects of parents on the node

Bayesian Net Semantics

- A set of nodes, one per variable X
- A directed, acyclic graph
- A conditional distribution for each node
  - A collection of distributions over X, one for each combination of parents' values

$$P(X|a_1 \dots a_n)$$

  - CPT(conditional probability table): each row is a distribution for child given values of its parents
  - Description of a noisy "causal" process



$$P(X|A_1 \dots A_n)$$

A Bayes net = Topology (graph) + Local Conditional Probabilities

Semantics of Bayesian Networks

- The full joint distribution can be defined as the product of the local conditional distributions (using chain rule):

$$P(X_1, \dots, X_n) = \prod_{i=1}^{n} P\,(X_i |\, Parents(X_i))$$

- Chain rule is derived by successive application of product rule:

$$
\begin{aligned}
& P(X_1, \dots, X_n) \\
& = P(X_1, \dots, X_{n-1}\,)\, P(X_n | X_1, \dots, X_{n-1}) \\
& = P(X_1, \dots, X_{n-2})\, P(X_{n-1} | X_1, \dots, X_{n-2})\, P(X_n | X_1, \dots, X_{n-1}) \\
& = \dots
\end{aligned}
$$

$$= P(X_1) \prod_{i=2}^{n} P(X_i | X_1, \dots, X_{i-1})$$

Probabilities in BNs

- Why are we guaranteed that setting

$$P(x_1, x_2, \ldots x_n) = \prod_{i=1}^{n} P(x_i|parents(X_i))$$

results in a proper joint distribution?

- Chain rule (valid for all distributions): $P(x_1, x_2, \ldots x_n) = \prod_{i=1}^{n} P(x_i|x_1 \ldots x_{i-1})$

- <u>Assume</u> conditional independences: $P(x_i|x_1, \ldots x_{i-1}) = P(x_i|parents(X_i))$

→ Consequence: $P(x_1, x_2, \ldots x_n) = \prod_{i=1}^{n} P(x_i|parents(X_i))$

- Not every BN can represent every joint distribution
  - The topology enforces certain conditional independencies

Example: Coin Flips

X_1 X_2 . . . X_n

$P(X_1)$

h	0.5
t	0.5

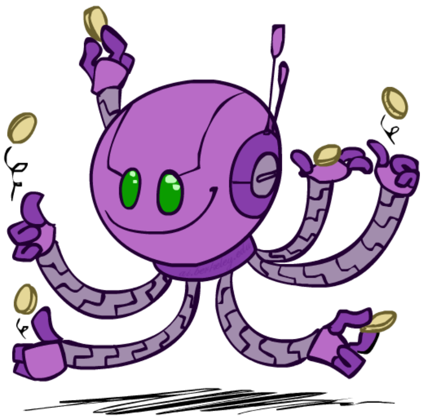
$P(X_2)$

h	0.5
t	0.5

. . .

$P(X_n)$

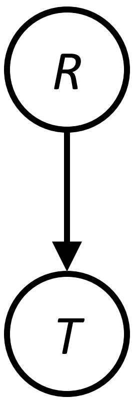
h	0.5
t	0.5



$$P(h, h, t, h) =$$

Only distributions whose variables are absolutely independent can be represented by a Bayes' net with no arcs.

Example: Traffic



$$P(R)$$

+r	1/4
-r	3/4

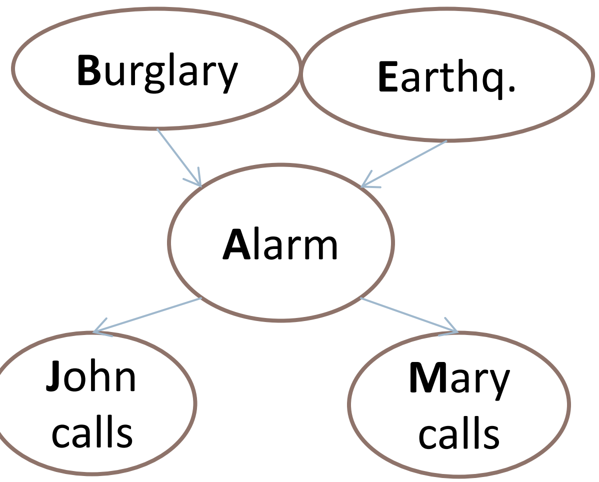
$$P(T|R)$$

+r	+t	3/4
	-t	1/4
-r	+t	1/2
	-t	1/2

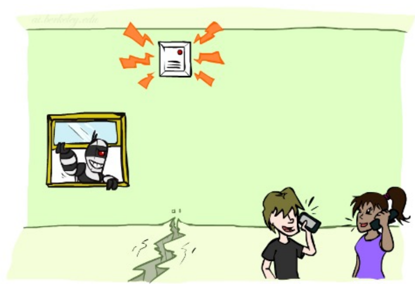
$$P(+r, -t) =$$

Example: Alarm Network

B	P(B)
+b	0.001
-b	0.999



E	P(E)
+e	0.002
-e	0.998



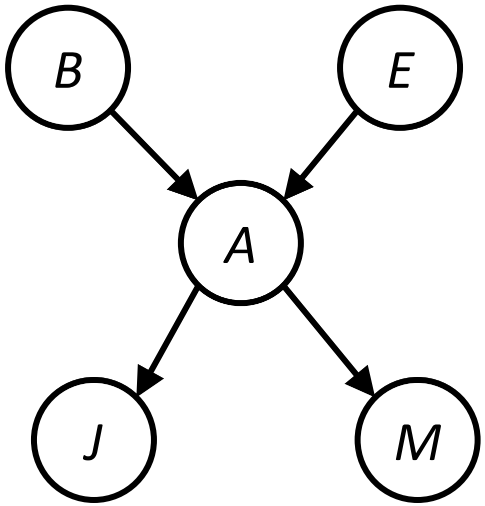
B	E	A	P(A|B,E)
+b	+e	+a	0.95
+b	+e	-a	0.05
+b	-e	+a	0.94
+b	-e	-a	0.06
-b	+e	+a	0.29
-b	+e	-a	0.71
-b	-e	+a	0.001
-b	-e	-a	0.999

A	J	P(J|A)
+a	+j	0.9
+a	-j	0.1
-a	+j	0.05
-a	-j	0.95

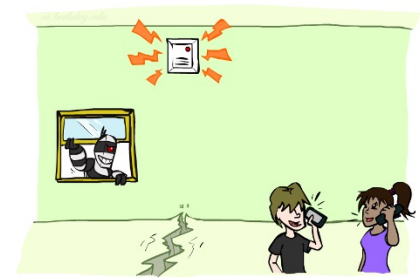
A	M	P(M|A)
+a	+m	0.7
+a	-m	0.3
-a	+m	0.01
-a	-m	0.99

Example: Alarm Network

B	P(B)
+b	0.001
-b	0.999



E	P(E)
+e	0.002
-e	0.998



A	J	P(J|A)
+a	+j	0.9
+a	-j	0.1
-a	+j	0.05
-a	-j	0.95

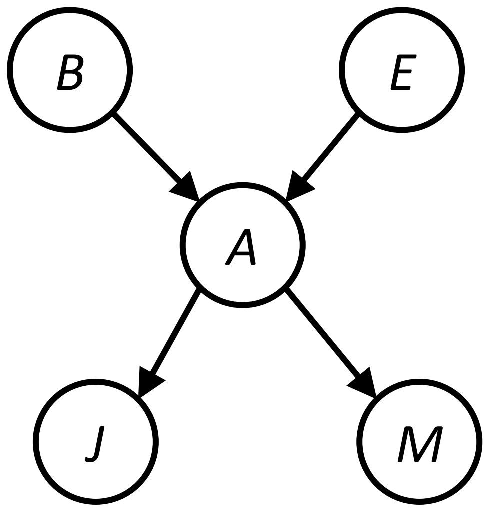
A	M	P(M|A)
+a	+m	0.7
+a	-m	0.3
-a	+m	0.01
-a	-m	0.99

B	E	A	P(A|B,E)
+b	+e	+a	0.95
+b	+e	-a	0.05
+b	-e	+a	0.94
+b	-e	-a	0.06
-b	+e	+a	0.29
-b	+e	-a	0.71
-b	-e	+a	0.001
-b	-e	-a	0.999

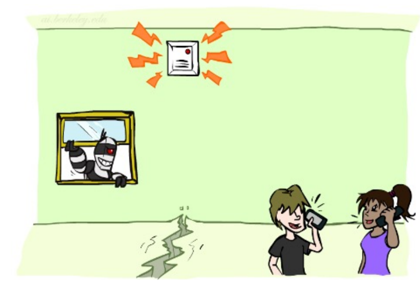
$$P(+b, -e, +a, -j, +m) =$$

Example: Alarm Network

B	P(B)
+b	0.001
-b	0.999



E	P(E)
+e	0.002
-e	0.998



A	J	P(J|A)
+a	+j	0.9
+a	-j	0.1
-a	+j	0.05
-a	-j	0.95

A	M	P(M|A)
+a	+m	0.7
+a	-m	0.3
-a	+m	0.01
-a	-m	0.99

B	E	A	P(A|B,E)
+b	+e	+a	0.95
+b	+e	-a	0.05
+b	-e	+a	0.94
+b	-e	-a	0.06
-b	+e	+a	0.29
-b	+e	-a	0.71
-b	-e	+a	0.001
-b	-e	-a	0.999

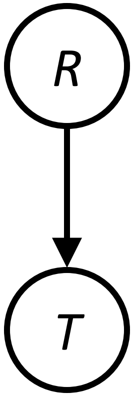
$$P(+b, -e, +a, -j, +m) =$$

$$P(+b)P(-e)P(+a|+b,-e)P(-j|+a)P(+m|+a) =$$

$$0.001 \times 0.998 \times 0.94 \times 0.1 \times 0.7$$

Example: Traffic

- Causal direction

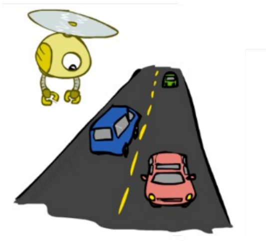


$P(R)$

+r	1/4
-r	3/4

$P(T|R)$

+r	+t	3/4
	-t	1/4
-r	+t	1/2
	-t	1/2

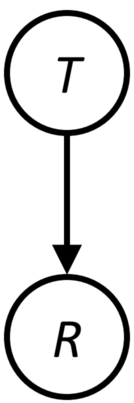


$P(T, R)$

+r	+t	3/16
+r	-t	1/16
-r	+t	6/16
-r	-t	6/16

Example: Reverse Traffic

- Reverse causality?



$P(T)$

+t	9/16
-t	7/16

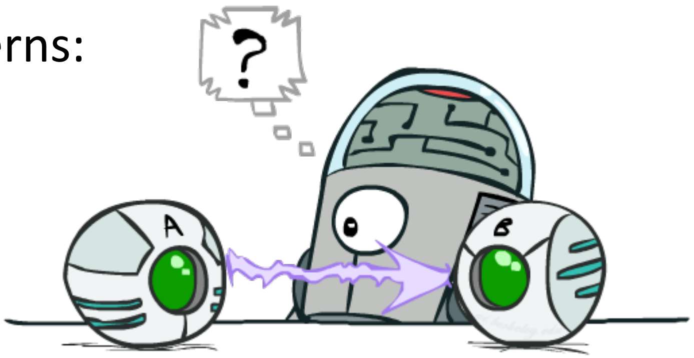
$P(R|T)$

+t	+r	1/3
	-r	2/3
-t	+r	1/7
	-r	6/7

$P(T, R)$

+r	+t	3/16
+r	-t	1/16
-r	+t	6/16
-r	-t	6/16

Causality?



- When Bayes nets reflect the true causal patterns:
  - Often simpler (nodes have fewer parents)
  - Often easier to think about
  - Often easier to elicit from experts
- BNs need not actually be causal
  - Sometimes no causal net exists over the domain (especially if variables are missing)
  - E.g. consider the variables *Traffic* and *Drips*
  - End up with arrows that reflect correlation, not causation
- What do the arrows really mean?
  - Topology may happen to encode causal structure
  - Topology really encodes conditional independence

$$P(x_i|x_1, \ldots x_{i-1}) = P(x_i|parents(X_i))$$

Constructing Bayesian Networks

I. **Nodes**:

determine the set of variables and order them as $X_1, \ldots, X_n$

(More compact network if causes precede effects)

II. **Links**:

for $i = 1$ to n

1) select a minimal set of parents for X_i from $X_1, \ldots, X_{i-1}$ such that

$$\mathbf{P}(X_i \mid Parents(X_i)) = \mathbf{P}(X_i \mid X_1, \ldots X_{i-1})$$

2) For each parent insert a link from the parent to X_i

3) CPT creation based on $\mathbf{P}(X_i \mid X_1, \ldots X_{i-1})$

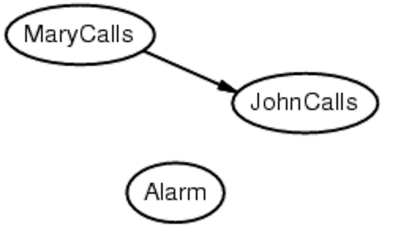
Node Ordering: Burglary example

- Suppose we choose the ordering M, J, A, B, E

- $\mathbf{P}(J \mid M) = \mathbf{P}(J)$?

Node Ordering: Burglary example

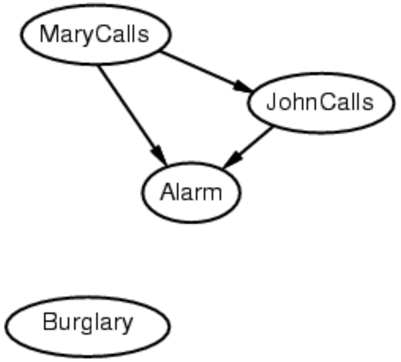
- Suppose we choose the ordering M, J, A, B, E



- $\mathbf{P}(J \mid M) = \mathbf{P}(J)$? **No**
- $\mathbf{P}(A \mid J, M) = \mathbf{P}(A \mid J)$?
- $\mathbf{P}(A \mid J, M) = \mathbf{P}(A)$?

Node Ordering: Burglary example

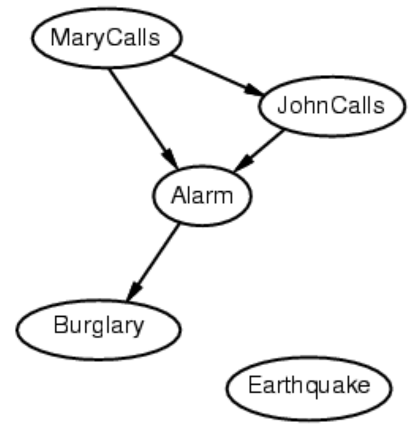
- Suppose we choose the ordering M, J, A, B, E



- $\mathbf{P}(J \mid M) = \mathbf{P}(J)$? **No**
- $\mathbf{P}(A \mid J, M) = \mathbf{P}(A \mid J)$? **No**
- $\mathbf{P}(A \mid J, M) = \mathbf{P}(A)$? **No**
- $\mathbf{P}(B \mid A, J, M) = \mathbf{P}(B \mid A)$?
- $\mathbf{P}(B \mid A, J, M) = \mathbf{P}(B)$?

Node Ordering: Burglary example

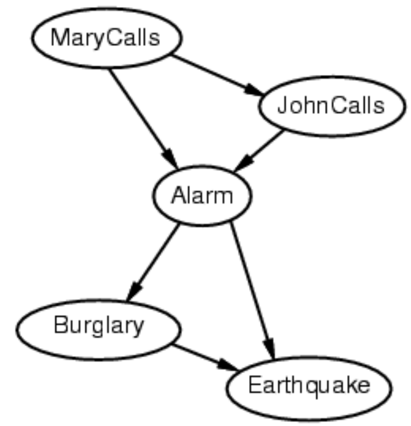
- Suppose we choose the ordering M, J, A, B, E



- $\mathbf{P}(J \mid M) = \mathbf{P}(J)$? **No**
- $\mathbf{P}(A \mid J, M) = \mathbf{P}(A \mid J)$? **No**
- $\mathbf{P}(A \mid J, M) = \mathbf{P}(A)$? **No**
- $\mathbf{P}(B \mid A, J, M) = \mathbf{P}(B \mid A)$? **Yes**
- $\mathbf{P}(B \mid A, J, M) = \mathbf{P}(B)$? **No**
- $\mathbf{P}(E \mid B, A, J, M) = \boldsymbol{P}(E \mid A)$?
- $\mathbf{P}(E \mid B, A, J, M) = \boldsymbol{P}(E \mid A, B)$?

Node Ordering: Burglary example

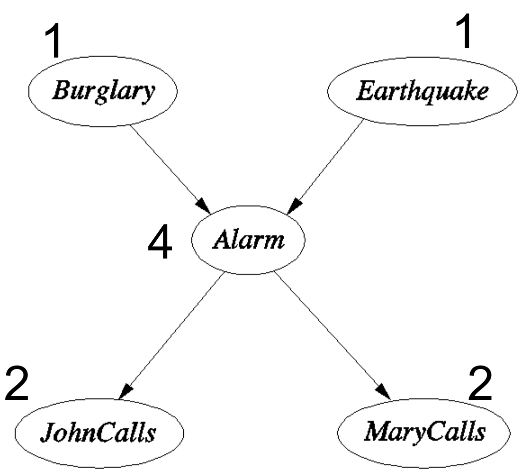
- Suppose we choose the ordering M, J, A, B, E



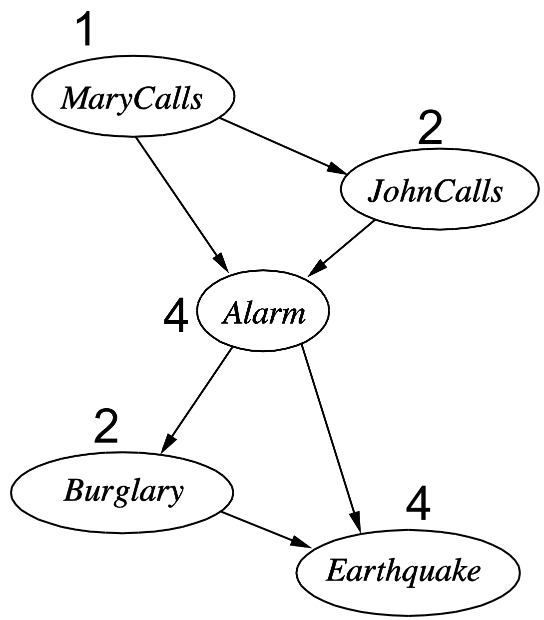
- $\mathbf{P}(J \mid M) = \mathbf{P}(J)$? **No**
- $\mathbf{P}(A \mid J, M) = \mathbf{P}(A \mid J)$? **No**
- $\mathbf{P}(A \mid J, M) = \mathbf{P}(A)$? **No**
- $\mathbf{P}(B \mid A, J, M) = \mathbf{P}(B \mid A)$? **Yes**
- $\mathbf{P}(B \mid A, J, M) = \mathbf{P}(B)$? **No**
- $\mathbf{P}(E \mid B, A, J, M) = \boldsymbol{P}(E \mid A)$? **No**
- $\mathbf{P}(E \mid B, A, J, M) = \boldsymbol{P}(E \mid A, B)$? **Yes**

Node Ordering: Burglary example

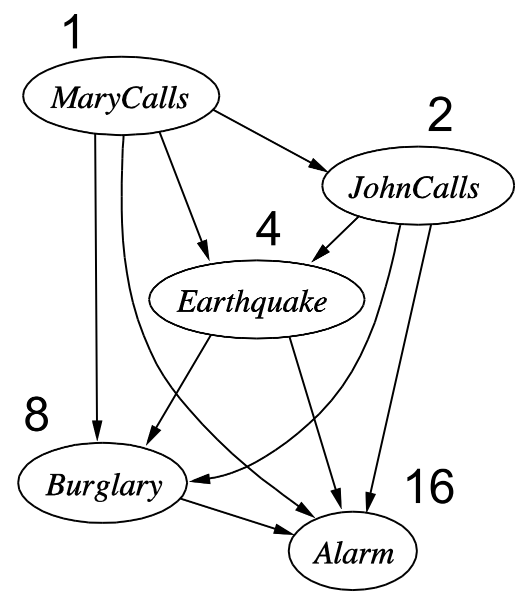
- The structure of the network and so the number of required probabilities for different node orderings can be different



1 + 1 + 4 + 2 + 2 = 10



1 + 2 + 4 + 2 + 4 = 13



1 + 2 + 4 + 8 + 16 = 31

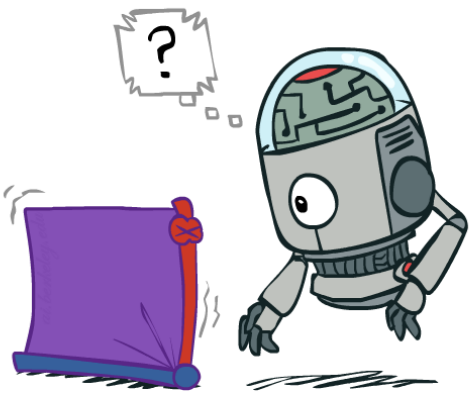
Size of a Bayes' Net

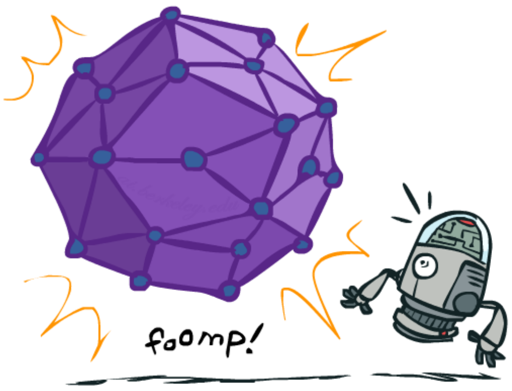
- How big is a joint distribution over N Boolean variables?
  - 2^N
- How big is an N-node net if nodes have up to k parents?
  - $O(N * 2^{k+1})$

- Both give you the power to calculate

$$P(X_1, X_2, \ldots X_n)$$

- BNs: Huge space savings!
- Also easier to elicit local CPTs
- Also faster to answer queries (coming)



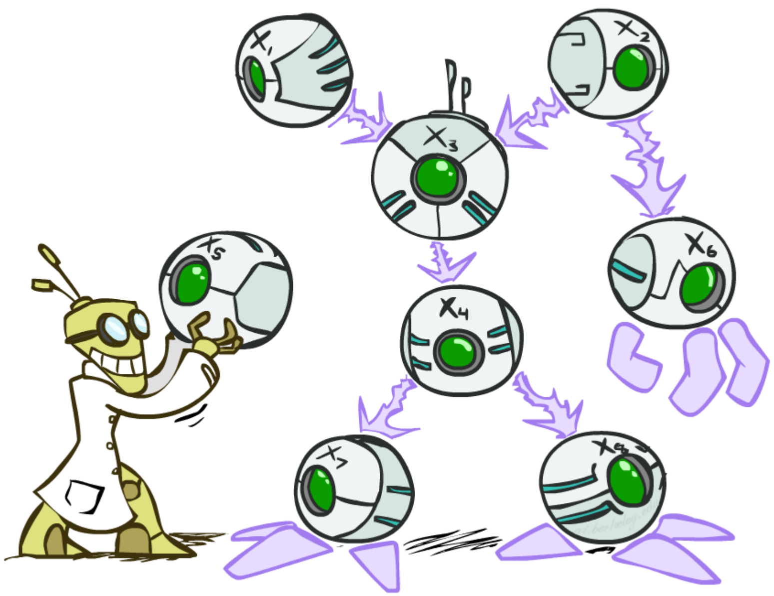


Bayes' Nets

- So far: how a Bayes' net encodes a joint distribution

- Next: how to answer queries about that distribution
  - Today:
    - First assembled BNs using an intuitive notion of conditional independence as causality
    - Then saw that key property is conditional independence
  - Main goal: answer queries about conditional independence and influence

- After that: how to answer numerical queries (inference)

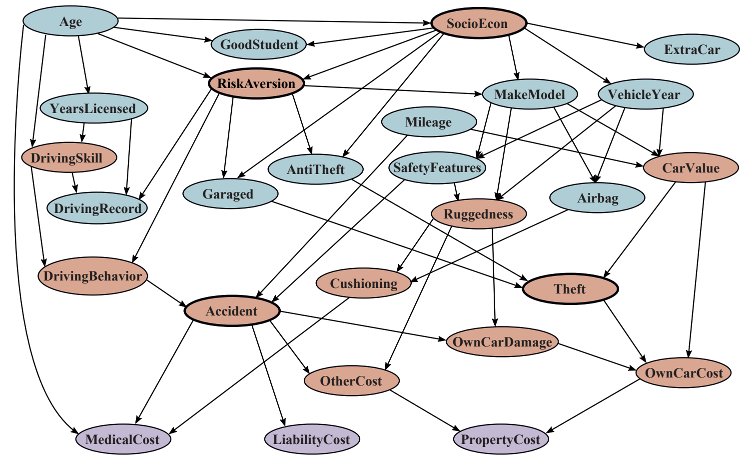


Probabilistic Models & Bayesian Networks: Summary

- Probability is the most common way to represent uncertain knowledge
- Whole knowledge: joint probability distribution in probability theory (instead of truth table for KB in two-valued logic)
- Independence and conditional independence can be used to provide a compact representation of joint probabilities
- Bayesian networks: representation for conditional independence
  - Compact representation of joint distribution by network topology and CPTs

Bayesian Nets

- A Bayes' net is an efficient encoding of a probabilistic model of a domain



- Questions we can ask:
  - **Representation**: given a BN graph, what kinds of distributions can it encode?
  - **Inference**: given a fixed BN, what is P(X | e)?
  - **Modeling**: what BN is most appropriate for a given domain?

Bayesian Nets

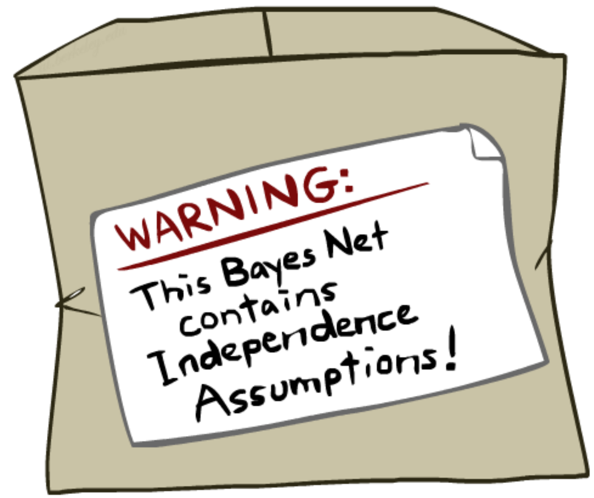
- ✓ Representation
- Conditional Independences
- Probabilistic Inference
- Learning Bayes' Nets from Data

Bayesian Nets: Assumptions

- Assumptions we are required to make to define the Bayes net when given the graph:

$$P(x_i|x_1 \cdots x_{i-1}) = P(x_i|parents(X_i))$$

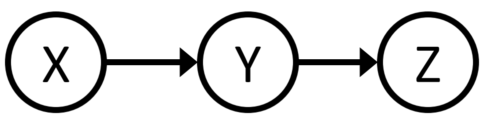
- Beyond above "chain rule → Bayes net" conditional independence assumptions
  - Often additional conditional independences
  - They can be read off the graph
- Important for modeling: understand assumptions made when choosing a Bayes net graph



Independence in a BN

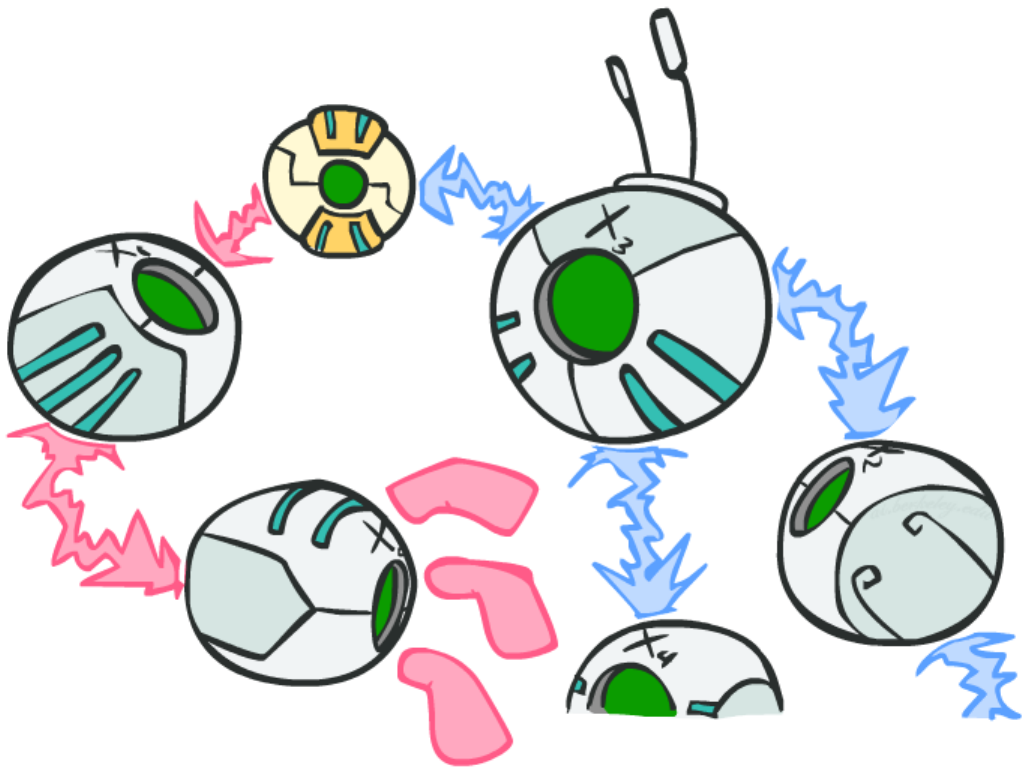
- Additional implied conditional independence assumptions?

- Important question about a BN:
  - Are two nodes independent given certain evidence?
  - If yes, can prove using algebra (tedious in general)
  - If no, can prove with a counter example
  - Example:



  - Question: are X and Z necessarily independent?
    - Answer: no. Example: low pressure causes rain, which causes traffic.
    - X can influence Z, Z can influence X (via Y)
    - Addendum: they *could* be independent: how?

D-separation: Outline



D-separation: Outline

- Study independence properties for triples
- Analyze complex cases in terms of member triples
- D-separation: a condition / algorithm for answering such queries

Causal Chains

- This configuration is a "causal chain"

X: Low pressure Y: Rain Z: Traffic

$$P(x, y, z) = P(x)P(y|x)P(z|y)$$

- Guaranteed X independent of Z ? *No!*
  - One example set of CPTs for which X is not independent of Z is sufficient to show this independence is not guaranteed.
  - Example:
    - Low pressure causes rain causes traffic, high pressure causes no rain causes no traffic
    - In numbers:

P(+y | +x) = 1, P(-y | - x) = 1,
P(+z | +y) = 1, P(-z | -y) = 1,
P(+x)= P(-x) = 0.5

Causal Chains

- This configuration is a "causal chain"

X: Low pressure Y: Rain Z: Traffic

$$P(x, y, z) = P(x)P(y|x)P(z|y)$$

- Guaranteed X independent of Z given Y?

$$P(z|x, y) = \frac{P(x, y, z)}{P(x, y)}$$

$$= \frac{P(x)P(y|x)P(z|y)}{P(x)P(y|x)}$$

$$= P(z|y)$$

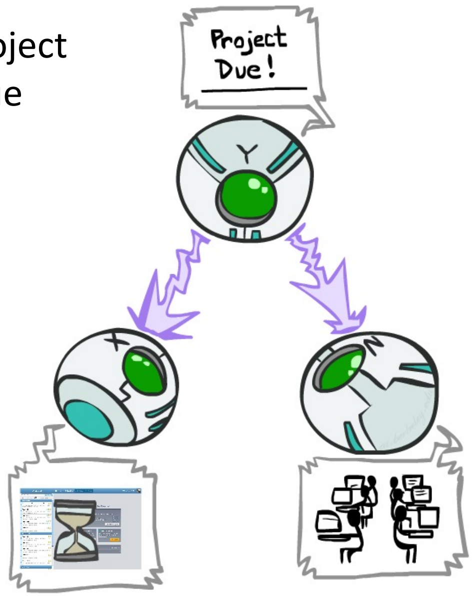
Yes!

Evidence along the chain "blocks" the influence

Common Cause

- This configuration is a "common cause"

Y: Project due



X: Forums busy

Z: Lab full

$$P(x, y, z) = P(y)P(x|y)P(z|y)$$

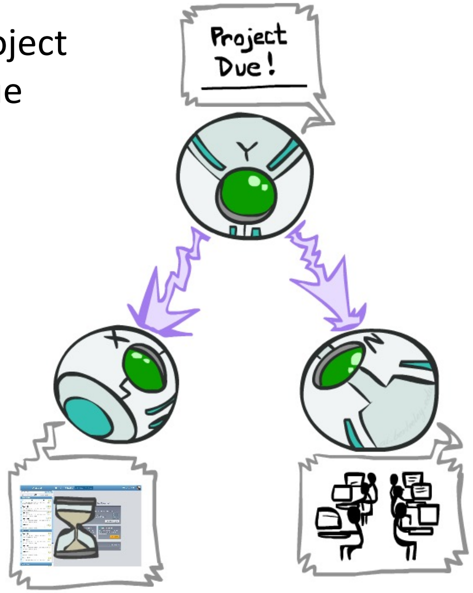
- Guaranteed X independent of Z ? *No!*
  - One example set of CPTs for which X is not independent of Z is sufficient to show this independence is not guaranteed.
  - Example:
    - Project due causes both forums busy and lab full
    - In numbers:

      P(+x | +y) = 1, P(-x | -y) = 1,
      P(+z | +y) = 1, P(-z | -y) = 1
      P(+y)= P(-y) = 0.5

Common Cause

- This configuration is a "common cause"

Y: Project due



X: Forums busy

Z: Lab full

$$P(x, y, z) = P(y)P(x|y)P(z|y)$$

- Guaranteed X and Z independent given Y?

$$P(z|x, y) = \frac{P(x, y, z)}{P(x, y)}$$

$$= \frac{P(y)P(x|y)P(z|y)}{P(y)P(x|y)}$$

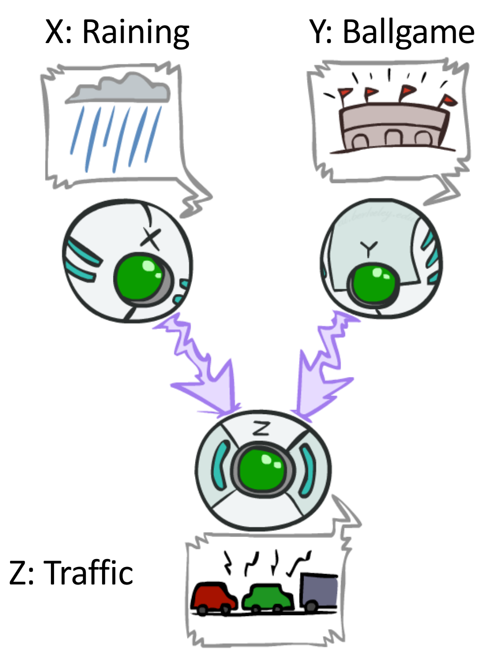
$$= P(z|y)$$

Yes!

Observing the cause blocks influence between effects.

Common Effect

- Last configuration: two causes of one effect (v-structures)



- Are X and Y independent?
  - *Yes*: the ballgame and the rain cause traffic, but they are not correlated
  - Still need to prove they must be (try it!)
- Are X and Y independent given Z?
  - *No*: seeing traffic puts the rain and the ballgame in competition as explanation.
- This is backwards from the other cases
  - Observing an effect activates influence between possible causes.

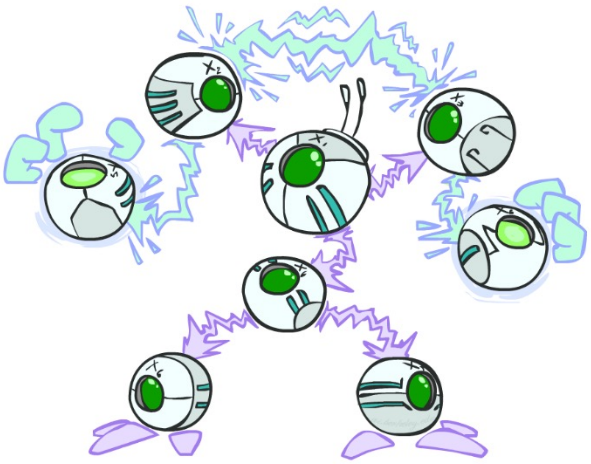
The General Case

The General Case

- General question: in a given BN, are two variables independent (given evidence)?

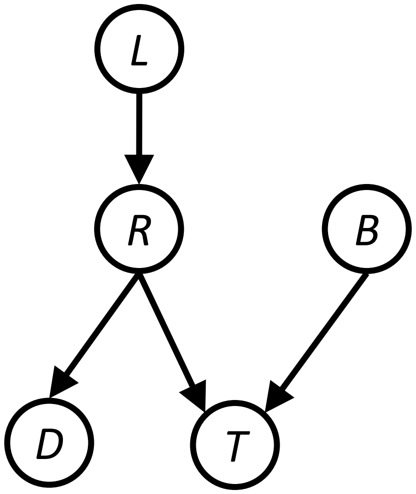
- Solution: analyze the graph

- Any complex example can be broken
-     into repetitions of the three canonical cases



Reachability

- Recipe: shade evidence nodes, look for paths in the resulting graph
- Attempt 1: if two nodes are connected by an undirected path not blocked by a shaded node, they are conditionally independent
- Almost works, but not quite
  - Where does it break?
  - Answer: the v-structure at T doesn't count as a link in a path unless "active"

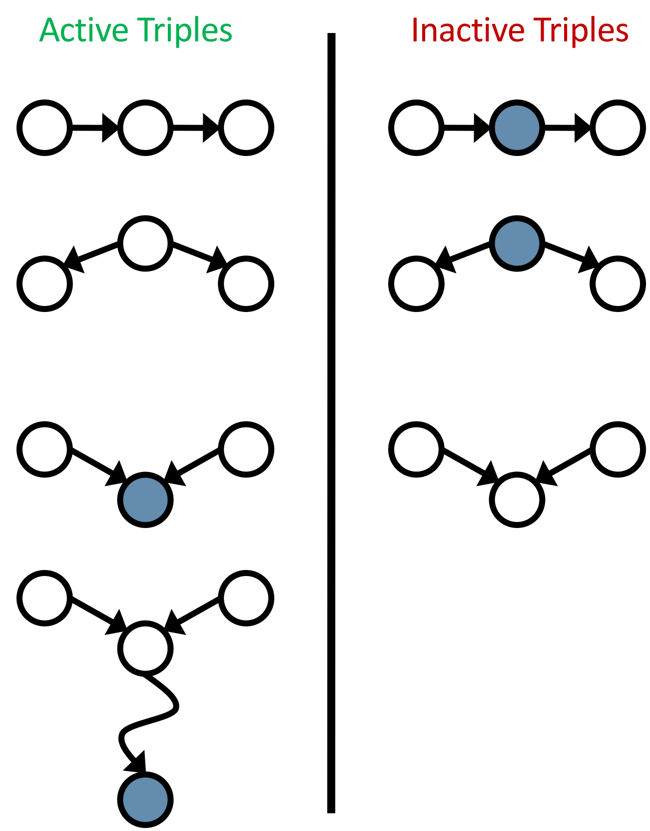


Active / Inactive Paths

- Question: Are X and Y conditionally independent given evidence variables {Z}?
  - Yes, if X and Y "d-separated" by Z
  - Consider all (undirected) paths from X to Y
  - No active paths = independence!

- A path is active if each triple is active:
  - Causal chain
    $A \to B \to C$ where B is unobserved (either direction)
  - Common cause
    $A \leftarrow B \to C$ where B is unobserved
  - Common effect (aka v-structure)
    $A \to B \leftarrow C$ where B *or one of its descendents* is observed

- All it takes to block a path is a single inactive segment



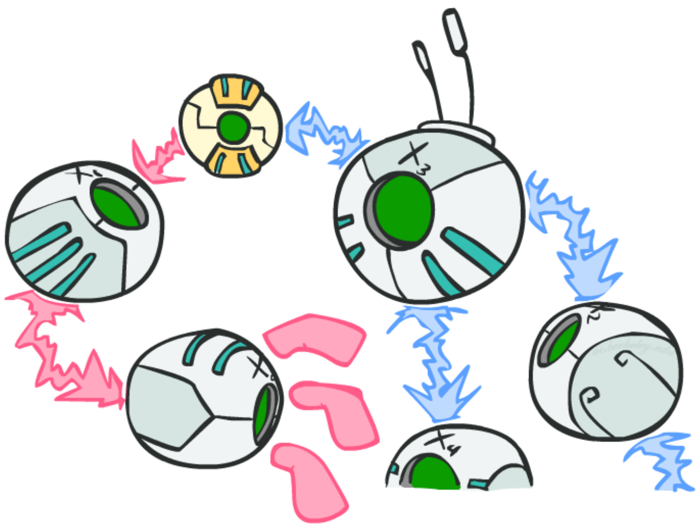
D-Separation

- Query: $X_i \perp\!\!\!\perp X_j | \{X_{k_1}, ..., X_{k_n}\}$?
- Check all (undirected!) paths between X_i and X_j
  - If one or more active, then independence not guaranteed

    $$X_i \not\!\perp\!\!\!\perp X_j | \{X_{k_1}, ..., X_{k_n}\}$$

  - Otherwise (i.e. if all paths are inactive), then independence is guaranteed

    $$X_i \perp\!\!\!\perp X_j | \{X_{k_1}, ..., X_{k_n}\}$$

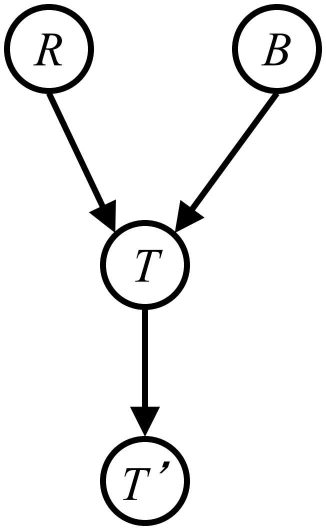


Example

$R \perp\!\!\!\perp B$ *Yes*

$R \perp\!\!\!\perp B | T$

$R \perp\!\!\!\perp B | T'$



Example

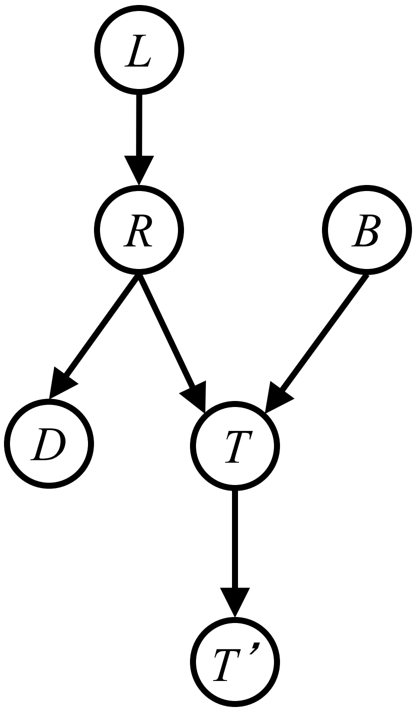
$L \perp\!\!\!\perp T' | T$ Yes

$L \perp\!\!\!\perp B$ Yes

$L \perp\!\!\!\perp B | T$

$L \perp\!\!\!\perp B | T'$

$L \perp\!\!\!\perp B | T, R$ Yes



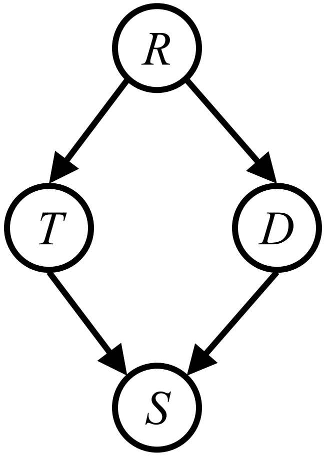
Example

- Variables:
  - R: Raining
  - T: Traffic
  - D: Roof drips
  - S: I'm sad
- Questions:

$T \perp\!\!\!\perp D$

$T \perp\!\!\!\perp D | R$ *Yes*

$T \perp\!\!\!\perp D | R, S$



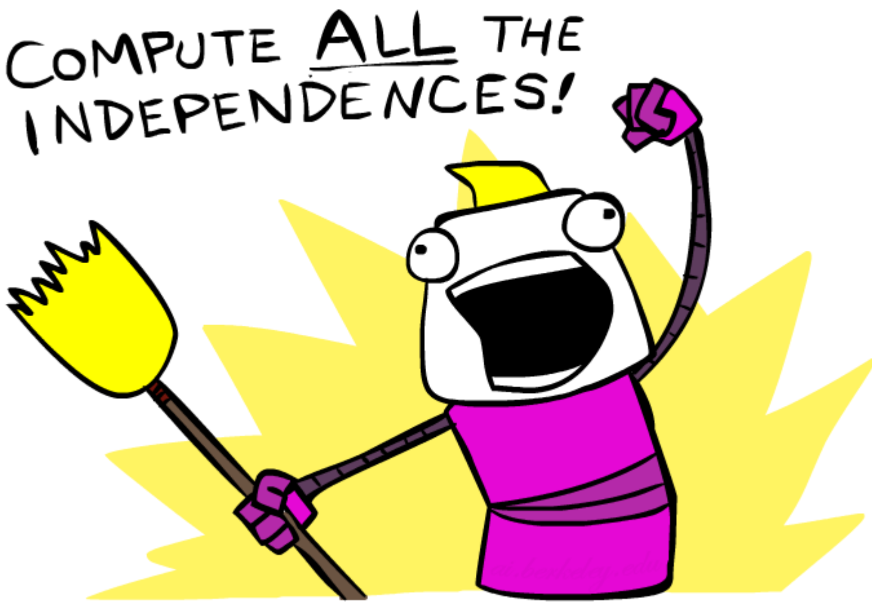
Structure Implications

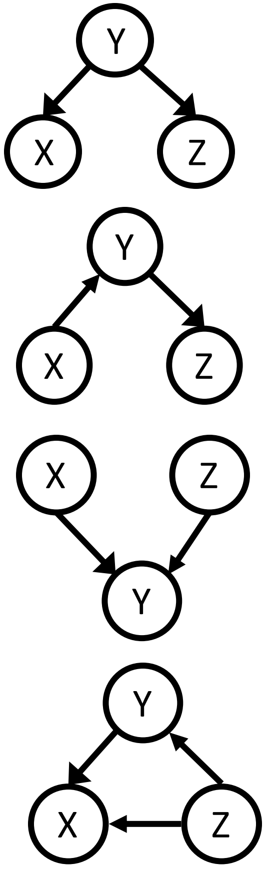
- Given a Bayes net structure, can run d-separation algorithm to build a complete list of conditional independences that are necessarily true of the form

$$X_i \perp\!\!\!\perp X_j | \{X_{k_1}, ..., X_{k_n}\}$$

- This list determines the set of probability distributions that can be represented

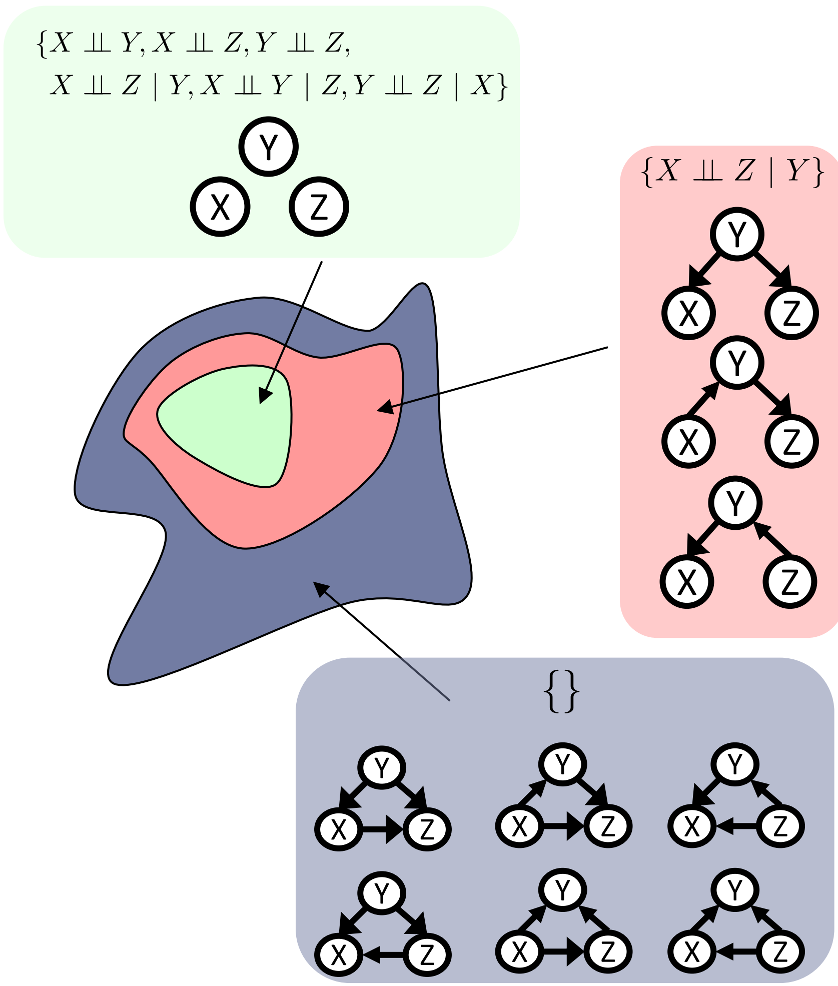
Computing All Independences





Topology Limits Distributions

- Given some graph topology G, only certain joint distributions can be encoded
- The graph structure guarantees certain (conditional) independences
- (There might be more independence)
- Adding arcs increases the set of distributions, but has several costs
- Full conditioning can encode any distribution



Bayes Nets Representation Summary

- Bayes nets compactly encode joint distributions
- Guaranteed independencies of distributions can be deduced from BN graph structure
- D-separation gives precise conditional independence guarantees from graph alone
- A Bayes' net's joint distribution may have further (conditional) independence that is not detectable until you inspect its specific distribution