CS 221: Artificial Intelligence

Lecture 3: Probability and Bayes Nets

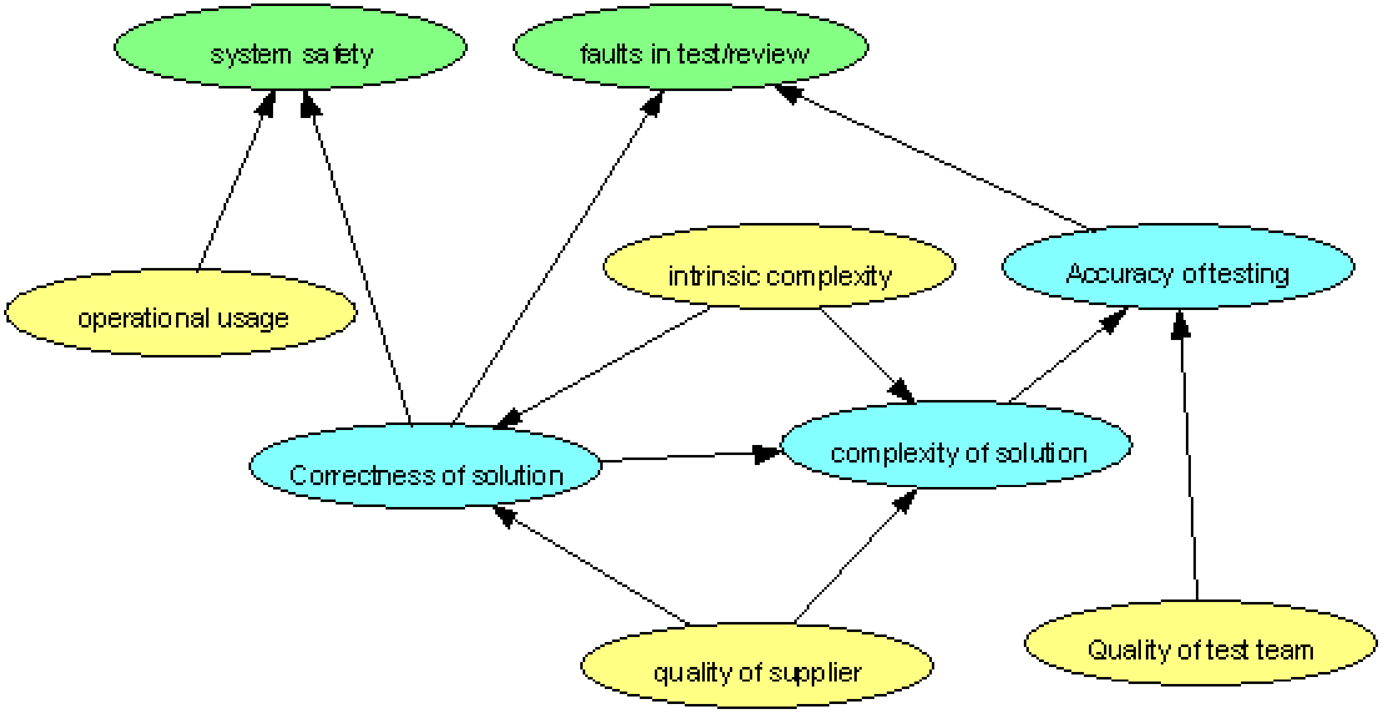
Sebastian Thrun and Peter Norvig

Naranbaatar Bayanbat, Juthika Dabholkar, Carlos Fernandez-Granda, Yan Largman, Cameron Schaeffer, Matthew Seal

Slide Credit: Dan Klein (UC Berkeley)

Goal of Today

- Structured representation of probability distribution



Probability

- Expresses uncertainty
- Pervasive in all of AI
  - Machine learning
  - Information Retrieval (e.g., Web)
  - Computer Vision
  - Robotics
- Based on mathematical calculus

- Disclaimer: We only discuss finite distributions

Probability

- Probability of a fair coin

$$P(\text{COIN} = \text{tail}) = \frac{1}{2}$$

$$P(\text{tail}) = \frac{1}{2}$$

Probability

- Probability of cancer

$$P(\text{has cancer}) = 0.02$$

$$\text{Þ} \quad P(\text{Ø}\,\text{has cancer}) = 0.98$$

Joint Probability

- Multiple events: cancer, test result

$$P(\text{has cancer, test positive})$$

Has cancer?	Test positive?	P(C,TP)
yes	yes	0.018
yes	no	0.002
no	yes	0.196
no	no	0.784

Joint Probability

- The problem with joint distributions

It takes 2^D-1 numbers to specify them!

Conditional Probability

- Describes the cancer test:

$$P(\text{test positive} \mid \text{has cancer}) = 0.9$$
$$P(\text{test positive} \mid \text{Øhas cancer}) = 0.2$$

- Put this together with: Prior probability

$$P(\text{has cancer}) = 0.02$$

$$P(\text{test negative} \mid \text{has cancer}) = 0.1$$

Conditional Probability

- We have:

$$P(\mathrm{C}) = 0.02 \qquad P(\varnothing \mathrm{C}) = 0.98$$

$$P(\mathrm{TP} \mid \mathrm{C}) = 0.9 \qquad P(\varnothing \mathrm{TP} \mid \mathrm{C}) = 0.1$$

$$P(\mathrm{TP} \mid \varnothing \mathrm{C}) = 0.2 \qquad P(\varnothing \mathrm{TP} \mid \varnothing \mathrm{C}) = 0.8$$

- We can now calculate joint probabilities

Has cancer?	Test positive?	P(TP, C)
yes	yes	0.018
yes	no	0.002
no	yes	0.196
no	no	0.784
no	no	

Conditional Probability

- "Diagnostic" question: How likely do is cancer given a positive test?

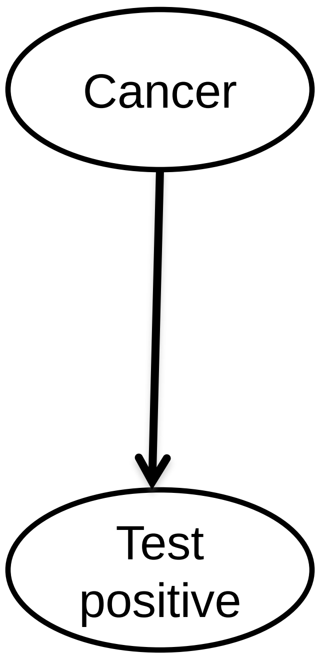
$$P(\text{has cancer} \mid \text{test positive}) = ?$$

Has cancer?	Test positive?	P(TP, C)
yes	yes	0.018
yes	no	0.002
no	yes	0.196
no	no	0.784

$$P(\mathrm{C} \mid \mathrm{TP}) = P(\mathrm{C}, \mathrm{TP}) / P(\mathrm{TP}) = 0.018 / 0.214 = 0.084$$

Bayes Network

- We just encountered our first Bayes network:



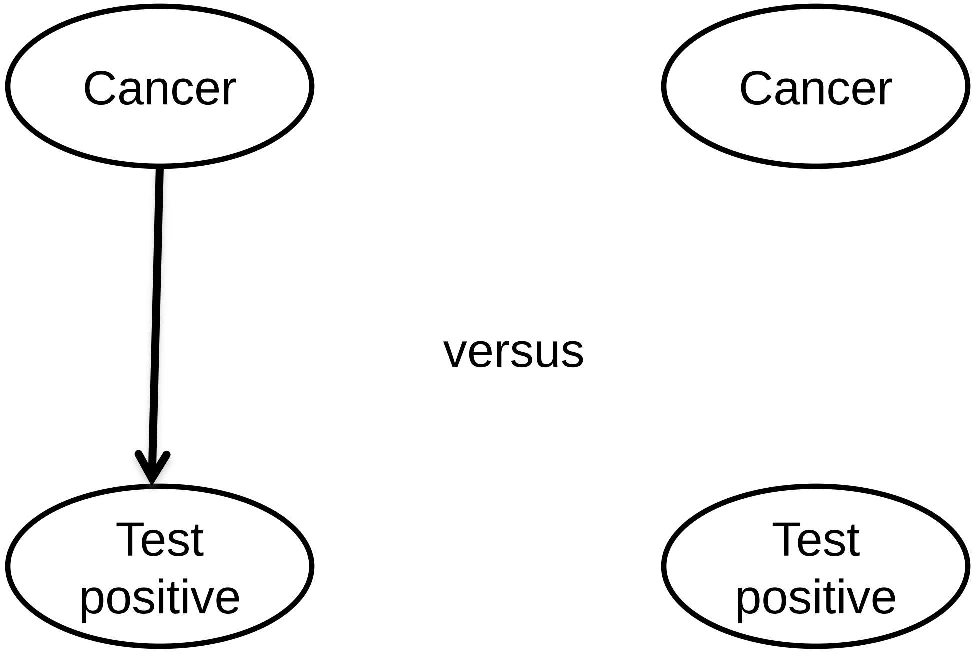
P(cancer) and P(Test positive | cancer) is called the "model"

Calculating P(Test positive) is called "prediction"

Calculating P(Cancer | test positive) is called "diagnostic reasoning"

Bayes Network

- We just encountered our first Bayes network:

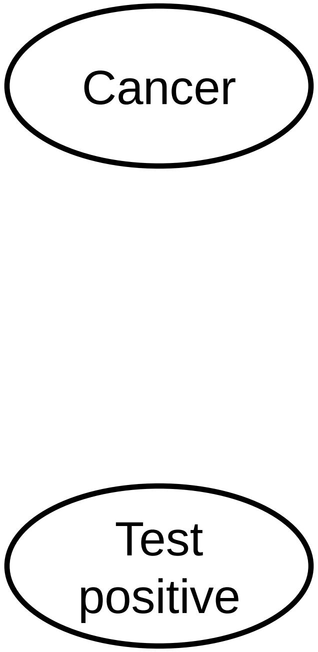


Independence

- Independence

$$P(\mathrm{C},\mathrm{TP}) = P(\mathrm{C}) \times P(\mathrm{TP})$$

- What does this mean for our test?
  - Don't take it!



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
  - This implies:

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

  - We write: $X \perp\!\!\!\perp Y$
- Independence is a simplifying *modeling assumption*
  - *Empirical* joint distributions: at best "close" to independent

Example: Independence

- N fair, independent coin flips:

$P(X_1)$

h	0.5
t	0.5

$P(X_2)$

h	0.5
t	0.5

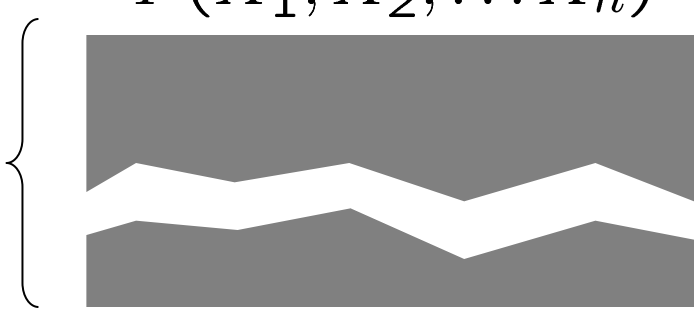
$\ldots$

$P(X_n)$

h	0.5
t	0.5

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

2^n



Example: Independence?

$P(T)$

T	P
warm	0.5
cold	0.5

$P_1(T, W)$

T	W	P
warm	sun	0.4
warm	rain	0.1
cold	sun	0.2
cold	rain	0.3

$P_2(T, W)$

T	W	P
warm	sun	0.3
warm	rain	0.2
cold	sun	0.3
cold	rain	0.2

$P(W)$

W	P
sun	0.6
rain	0.4

Conditional Independence

- P(Toothache, Cavity, Catch)
- If I have a Toothache, a dental probe might be more likely to catch
- But: if I have a cavity, the probability that the probe catches doesn't depend on whether I have a toothache:
  - P(+catch | +toothache, +cavity) = P(+catch | +cavity)
- The same independence holds if I don't have a cavity:
  - P(+catch | +toothache, $\neg$cavity) = P(+catch| $\neg$cavity)
- Catch is *conditionally independent* of Toothache given Cavity:
  - P(Catch | Toothache, Cavity) = P(Catch | Cavity)
- Equivalent *conditional independence* statements:
  - P(Toothache | Catch , Cavity) = P(Toothache | Cavity)
  - P(Toothache, Catch | Cavity) = P(Toothache | Cavity) P(Catch | Cavity)
  - One can be derived from the other easily
- We write: $X \perp\!\!\!\perp Y | Z$

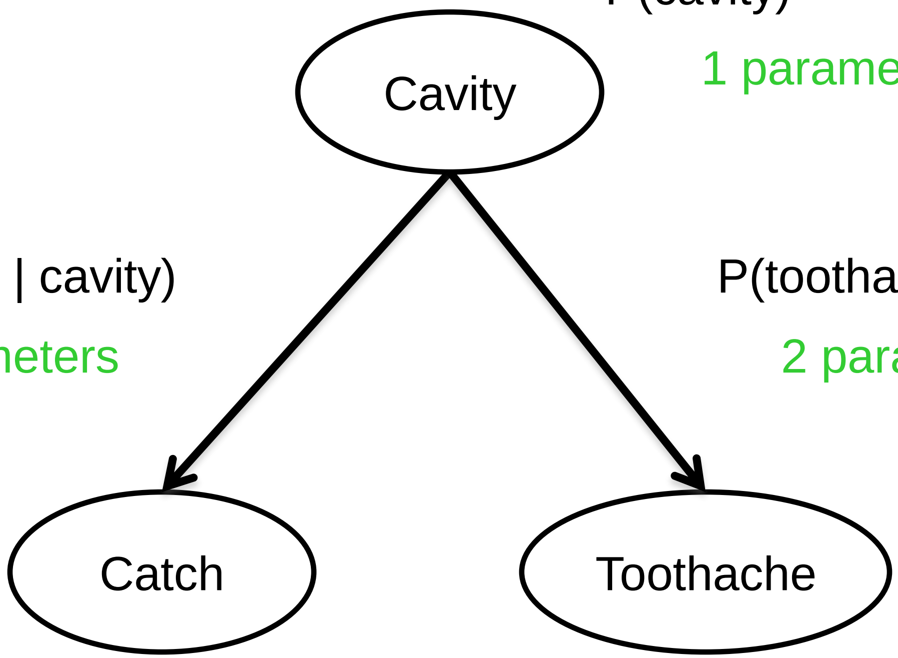
Bayes Network Representation

P(cavity)

1 parameter

P(catch | cavity)

2 parameters

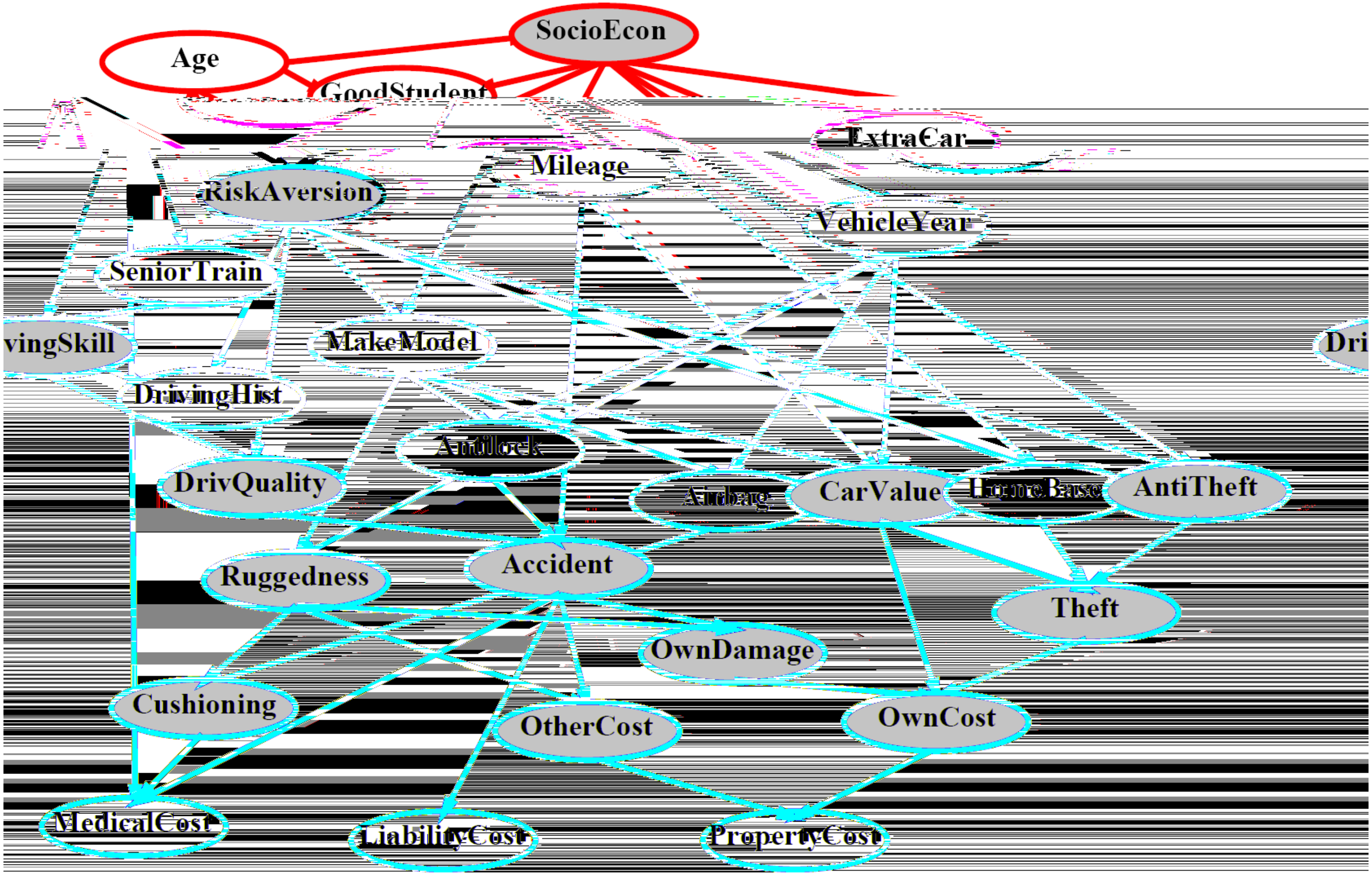


P(toothache | cavity)

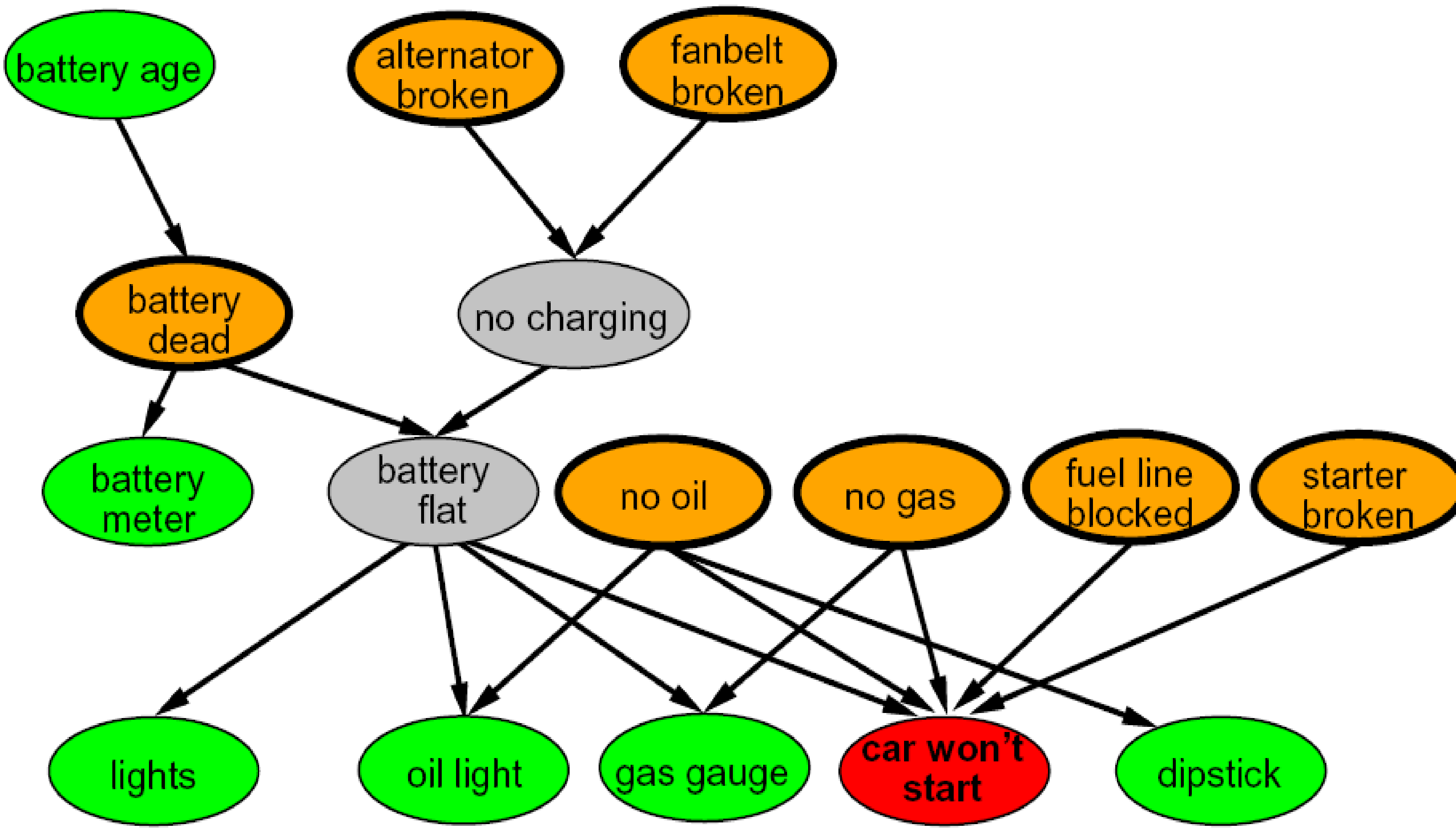
2 parameters

Versus: 2^3-1 = 7 parameters

A More Realistic Bayes Network

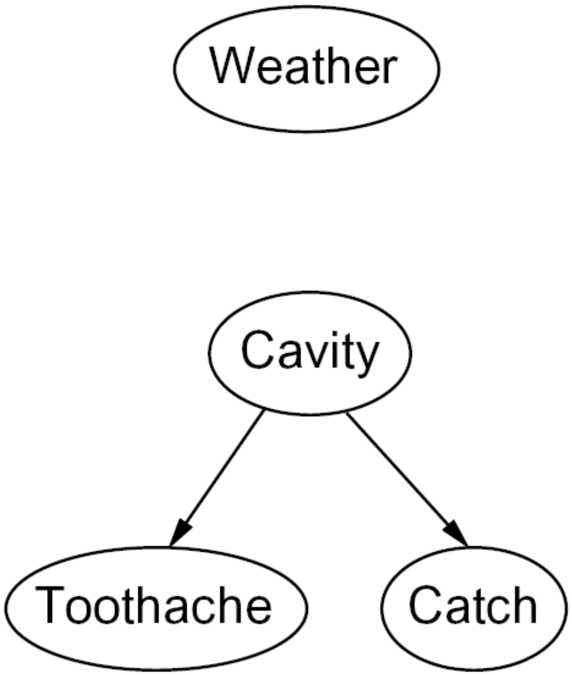


Example Bayes Network: Car



Graphical Model Notation

- Nodes: variables (with domains)
  - Can be assigned (observed) or unassigned (unobserved)
- Arcs: interactions
  - Indicate "direct influence" between variables
  - Formally: encode conditional independence (more later)
- For now: imagine that arrows mean direct causation (they may not!)



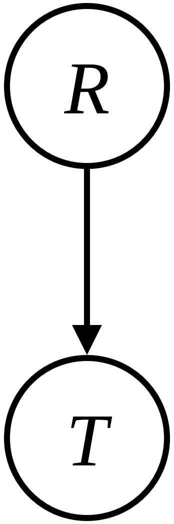
Example: Coin Flips

- N independent coin flips

- No interactions between variables: absolute independence

Example: Traffic

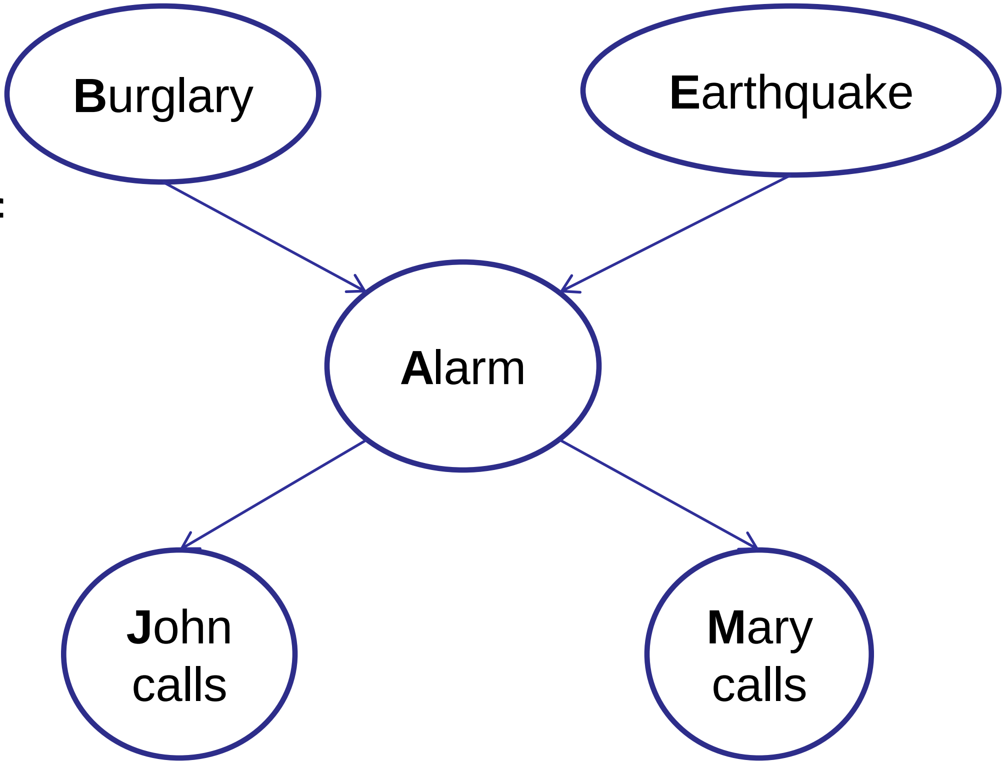
- Variables:
  - R: It rains
  - T: There is traffic



- Model 1: independence
- Model 2: rain causes traffic
- Why is an agent using model 2 better?

Example: Alarm Network

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

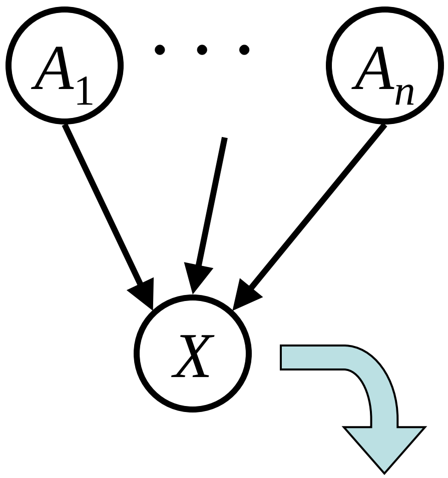


Bayes Net Semantics

- 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

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

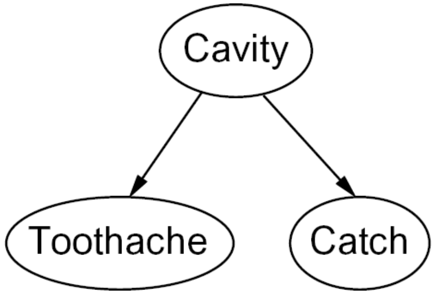
  - CPT: conditional probability table
  - Description of a noisy "causal" process



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

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

Probabilities in BNs



- Bayes nets implicitly encode joint distributions
  - As a product of local conditional distributions
  - To see what probability a BN gives to a full assignment, multiply all the relevant conditionals together:

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

  - Example:

$$P(+cavity, +catch, \neg toothache)$$

- This lets us reconstruct any entry of the full joint
- Not every BN can represent every joint distribution
  - The topology enforces certain conditional independencies

Example: Coin Flips

$P(X_1)$

h	0.5
t	0.5

$P(X_2)$

h	0.5
t	0.5

$\cdots$

$P(X_n)$

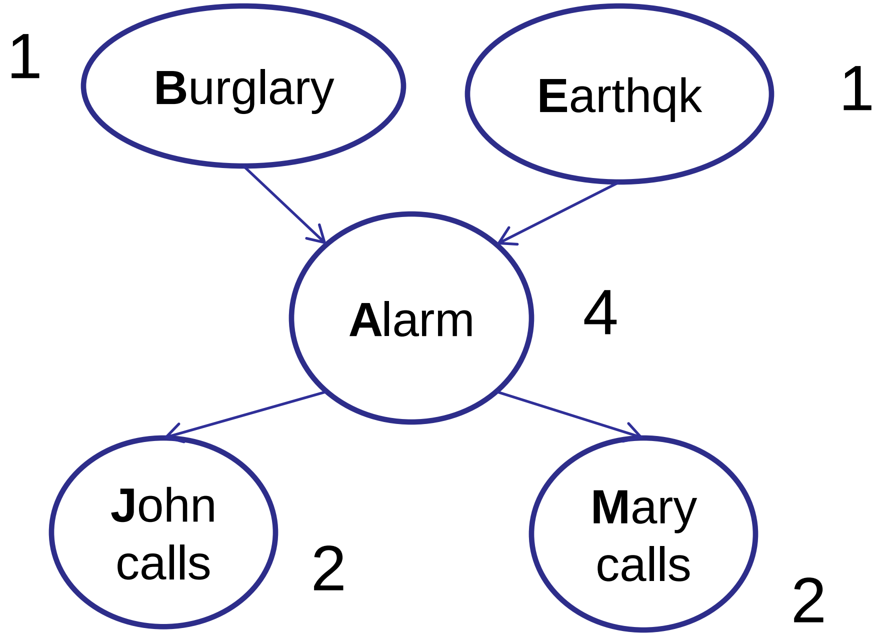
h	0.5
t	0.5

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

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

Example: Traffic

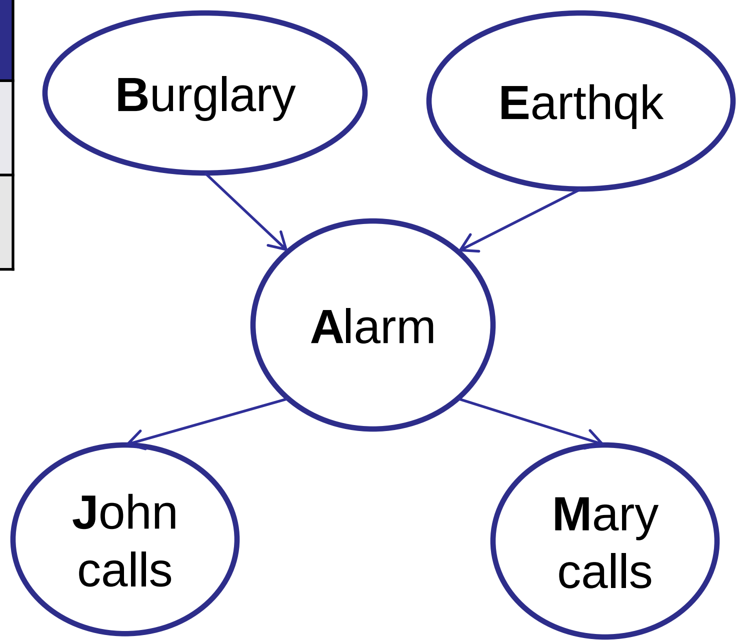
Example: Alarm Network



How many parameters? 10

Example: Alarm Network

B	P(B)
+b	0.001
¬b	0.999



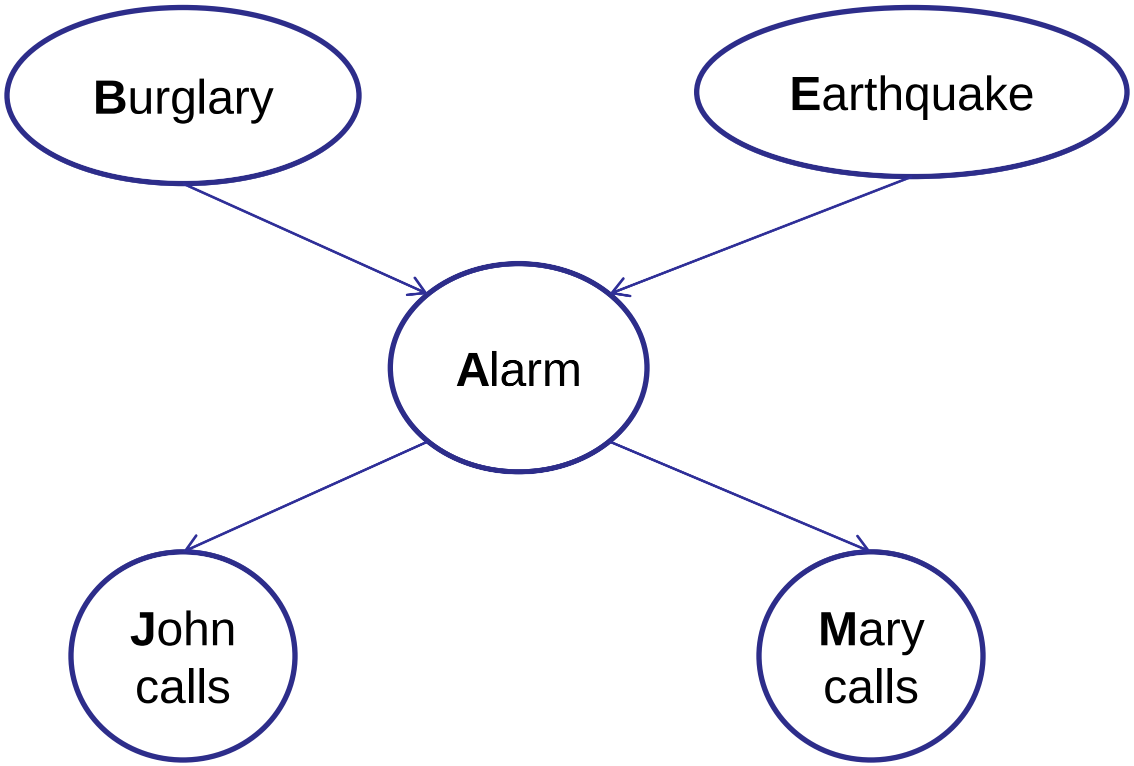
E	P(E)
+e	0.002
¬e	0.998

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

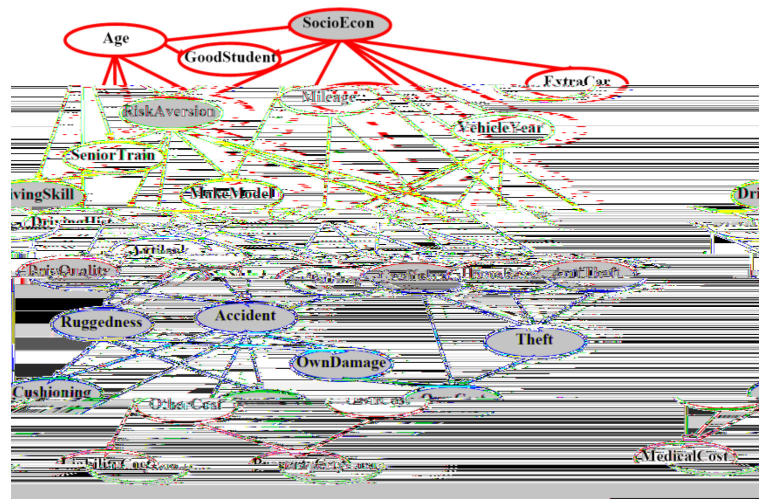
Example: Alarm Network



$$\prod_i P(X_i|\text{Parents}(X_i)) = P(B) \cdot P(E) \cdot P(A|B,E) \cdot P(J|A) \cdot P(M|A)$$

Bayes' Nets

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



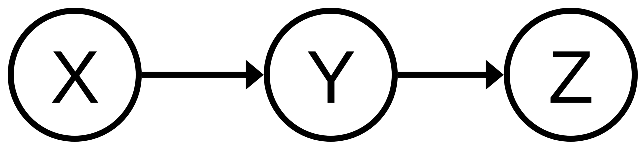
- Questions we can ask:
  - Inference: given a fixed BN, what is P(X | e)?
  - Representation: given a BN graph, what kinds of distributions can it encode?
  - Modeling: what BN is most appropriate for a given domain?

Remainder of this Class

- Find Conditional (In)Dependencies
  - Concept of “d-separation”

Causal Chains

- This configuration is a "causal chain"



X: Low pressure

Y: Rain

Z: Traffic

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

- Is X independent of Z given Y?

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

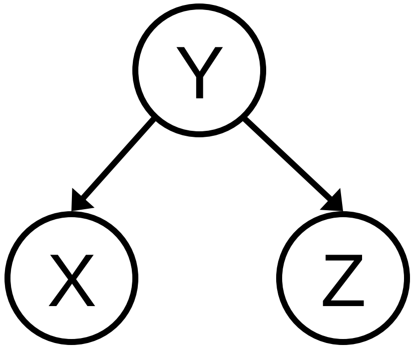
$$= P(z|y)$$

Yes!

- Evidence along the chain "blocks" the influence

Common Cause

- Another basic configuration: two effects of the same cause
  - Are X and Z independent?
  - Are 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!

Y: Alarm

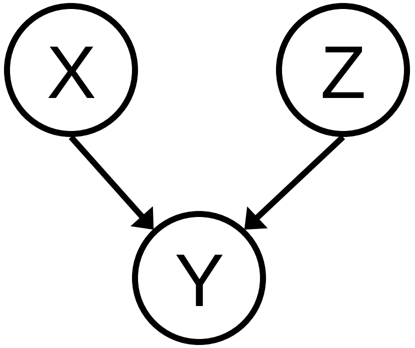
X: John calls

Z: Mary calls

  - Observing the cause blocks influence between effects.

Common Effect

- Last configuration: two causes of one effect (v-structures)
  - Are X and Z 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 Z independent given Y?
    - 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.



X: Raining

Z: Ballgame

Y: Traffic

The General Case

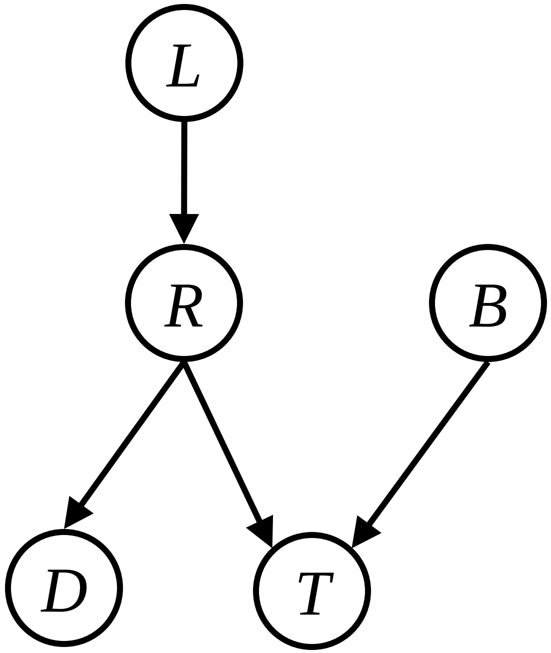
- Any complex example can be analyzed using these three canonical cases
- General question: in a given BN, are two variables independent (given evidence)?
- Solution: analyze the graph

Reachability

- Recipe: shade evidence nodes

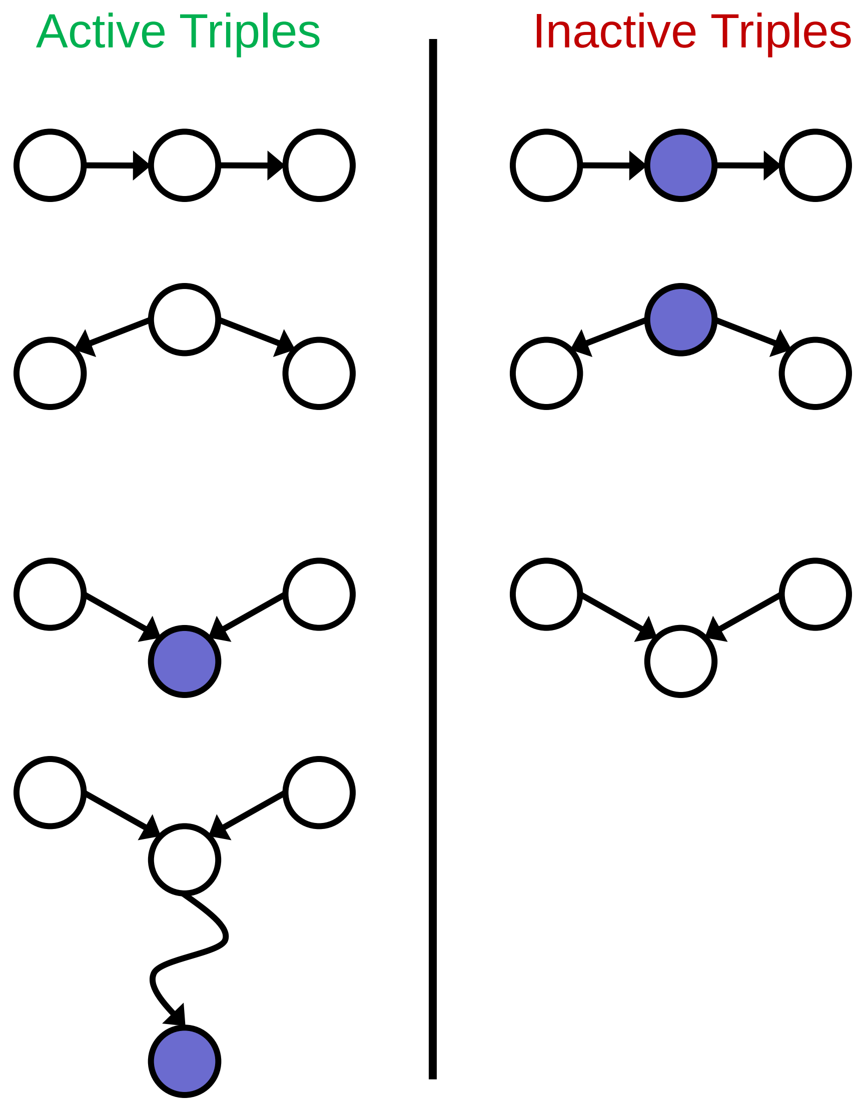
- Attempt 1: Remove shaded nodes. If two nodes are still connected by an undirected path, they are not conditionally independent

- 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"



Reachability (D-Separation)

- Question: Are X and Y conditionally independent given evidence vars {Z}?
  - Yes, if X and Y "separated" by Z
  - Look for active paths from X to Y
  - No active paths = independence!
- A path is active if each triple is active:
  - Causal chain $A \rightarrow B \rightarrow C$ where B is unobserved (either direction)
  - Common cause $A \leftarrow B \rightarrow C$ where B is unobserved
  - Common effect (aka v-structure) $A \rightarrow B \leftarrow C$ where B *or one of its descendents* is observed
- All it takes to block a path is a single inactive segment

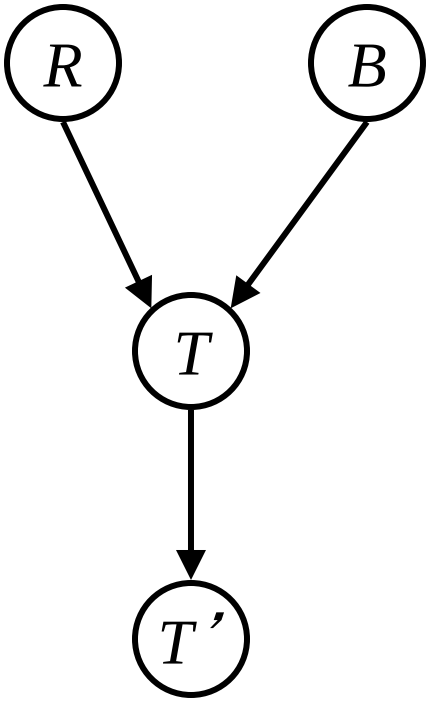


Example

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

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

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



Example

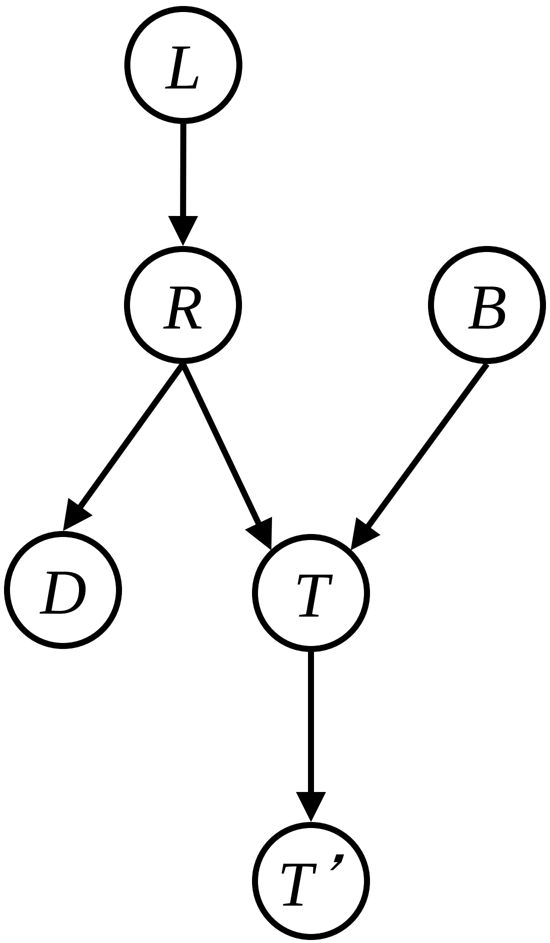
$L \perp\!\!\!\perp T' | T$ *Yes*

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

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

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

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



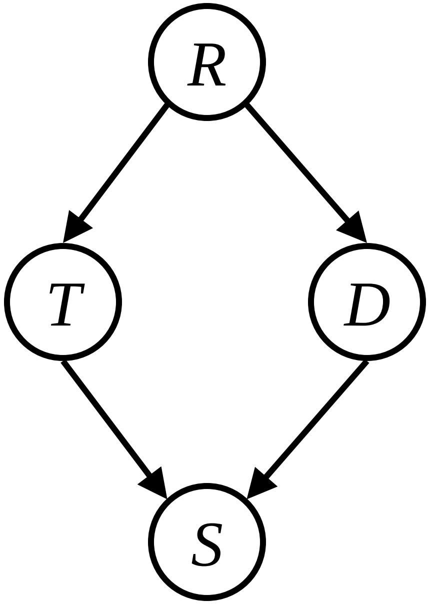
Example

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

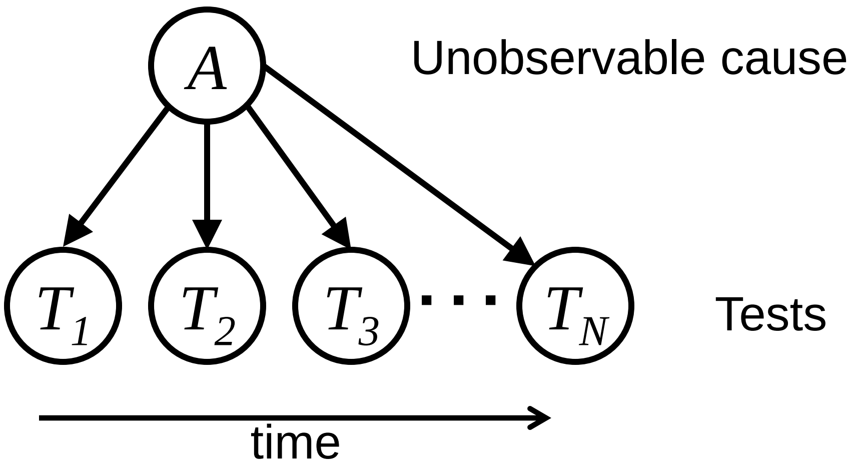
$$T \perp\!\!\!\perp D$$

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

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



A Common BN



Diagnostic Reasoning:

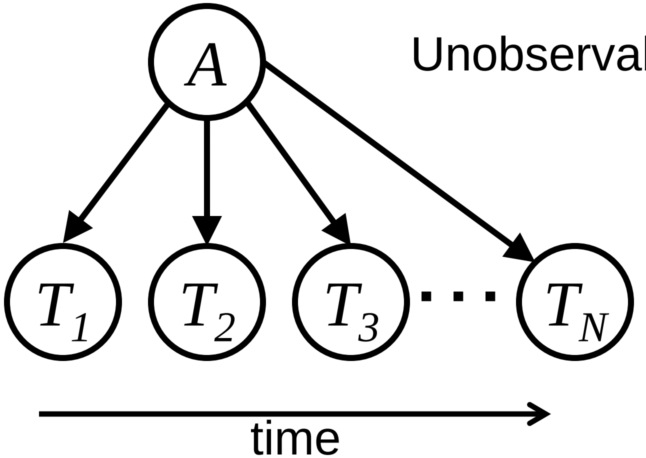
$$P(A \mid T_1, T_2, T_3, \ldots, T_N)$$

$$P(A \mid T_1 \ldots. T_N) = \frac{P(T_N \mid A, T_1 \ldots. T_{N-1})\, P(A \mid T_1 \ldots. T_{N-1})}{P(T_N \mid T_1 \ldots. T_{N-1})}$$

$$= \frac{1}{P(T_N \mid T_1 \ldots. T_{N-1})}\, P(T_N \mid A)\, P(A \mid$$

T_1

A Common BN



Unobservable cause

Tests

Diagnostic Reasoning:

$$P(A \mid T_1, T_2, T_3, \ldots, T_N)$$

$$\alpha_+ \leftarrow P(A) \prod_{n=1}^{N} P(T_n \mid A)$$

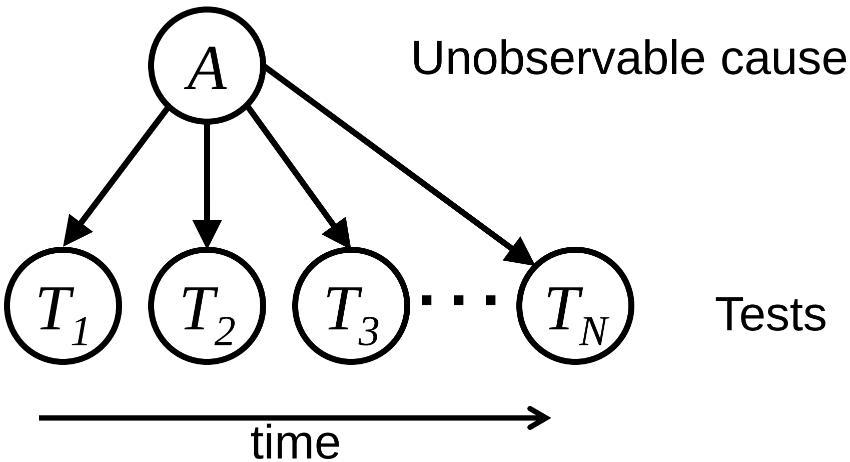
$$\alpha_- \leftarrow P(\neg A) \prod_{n=1}^{N} P(T_n \mid \neg A)$$

$$\eta \leftarrow \frac{1}{\alpha_+ + \alpha_-}$$

$$P(A \mid T_1 \ldots T_N) = \eta \alpha_+$$

$$P(\neg A \mid T_1 \ldots T_N) = \eta \alpha_-$$

A Common BN



Diagnostic Reasoning:

$$P(A \mid T_1, T_2, T_3, \ldots, T_N)$$

$$\beta_+ \leftarrow \log P(A) + \sum_{n=1}^{N} \log P(T_n \mid A)$$

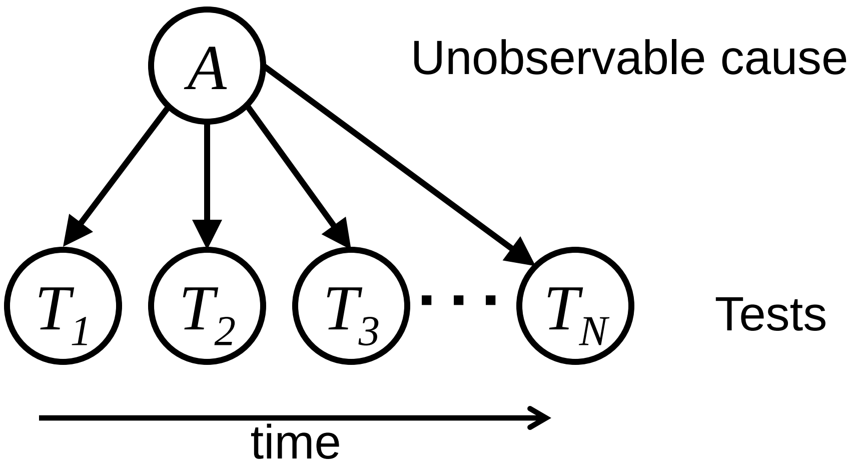
$$\beta_- \leftarrow \log P(\neg A) + \sum_{n=1}^{N} \log P(T_n \mid \neg A)$$

$$\eta \leftarrow \frac{1}{\exp \beta_+ + \exp \beta_-}$$

$$P(A \mid T_1 \ldots T_N) = \eta \exp \beta_+$$

$$P(\neg A \mid T_1 \ldots T_N) = \eta \exp \beta_-$$

A Common BN



Diagnostic Reasoning:

$$P(A \mid T_1, T_2, T_3, \ldots, T_N)$$

$$b = \log \frac{P(A \mid T_1 \ldots T_N)}{P(\neg A \mid T_1 \ldots T_N)} = \log \frac{P(A \mid T_1 \ldots T_N)}{1 - P(A \mid T_1 \ldots T_N)}$$

$$b \leftarrow \log P(A) - \log P(\neg A) + \sum_{n=1}^{N} \log P(T_n \mid A) - \log P(T_n \mid \neg A)$$

$$P(A \mid T_1 \ldots T_N) = 1 - \frac{1}{1 + \exp b}$$

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
  - End up with arrows that reflect correlation, not causation
- What do the arrows really mean?
  - Topology may happen to encode causal structure
  - Topology only guaranteed to encode conditional independence

Summary

- Bayes network:
  - Graphical representation of joint distributions
  - Efficiently encode conditional independencies
  - Reduce number of parameters from exponential to linear (in many cases)

  - Thursday: Inference in (general) Bayes networks