Bayesian Networks: Representation

Forest Agostinelli
University of South Carolina

Topics Covered in This Class

- **Part 1: Search**
  - Pathfinding
    - Uninformed search
    - Informed search
  - Adversarial search
  - Optimization
    - Local search
    - Constraint satisfaction
- **Part 2: Knowledge Representation and Reasoning**
  - Propositional logic
  - First-order logic
  - Prolog
- **Part 3: Knowledge Representation and Reasoning Under Uncertainty**
  - Probability
  - Bayesian networks
- **Part 4: Machine Learning**
  - Supervised learning
    - Inductive logic programming
    - Linear models
    - Deep neural networks
    - PyTorch
  - Reinforcement learning
    - Markov decision processes
    - Dynamic programming
    - Model-free RL
  - Unsupervised learning
    - Clustering
    - Autoencoders

Outline

- Bayes’ Rule
- Chain Rule and Conditional Independence
- Bayesian Networks

Bayes' Rule

- Product Rule
  - $P(a, b) = P(a|b)P(b)$
  - $P(a, b) = P(b|a)P(a)$
- Bayes' Rule
  - $P(b|a) = \frac{P(a|b)P(b)}{P(a)}$
- Often, we perceive evidence as the effect of some unknown cause
  - We perceive toothache, which may be due to a cavity
- It may be a lot easier to model the probability of the effect given the cause
  - I.e. P(symptoms|disease) may be known but P(disease|symptoms) may be unknown
- $P(cause|effect) = \frac{P(effect|cause)P(cause)}{P(effect)}$

Meningitis Example and Concept Check

- Suppose the probability of having a stiff next if you have meningitis is P(s|m)=0.7, the probability of having meningitis is P(m)=1/50000, and the probability of having a stiff neck is P(s)=0.01. What is the probability of having meningitis given a stiff neck, P(m|s)?
- $P(m|s) = \frac{P(s|m)P(m)}{P(s)} = \frac{0.7 \times \frac{1}{50000}}{0.01} = 0.0014$

Meningitis Example and Concept Check

- $P(m|s) = \frac{P(s|m)P(m)}{P(s)} = \frac{0.7 \times \frac{1}{50000}}{0.01} = 0.0014$
- What if P(s)=0.00001? Then P(m|s)=1.4?
  - But that's not possible!
  - What is wrong here?
  - $P(a) = \sum_b P(a,b)$ //**marginalization**
  - $\sum_b P(a|b)P(b)$ //**product rule**
- $P(s) = P(s|m)P(m) + P(s|\neg m)P(\neg m)$
- $P(s|m)$ and $P(m)$ affect the value of $P(s)$

Joint Distribution

- Assume we have n random variables that can take on d values
- If we are to store all of this in a table, we would need $O(d^n)$ entries
  - $P(X_1, X_2, \dots, X_n)$

Chain Rule

- Product rule: $P(X_1, X_2) = P(X_1|X_2)P(X_2) = P(X_2|X_1)P(X_1)$
- $P(X_1, X_2, \dots, X_n) = P(X_n|X_{n-1}, \dots, X_1)P(X_{n-1}, \dots, X_1)$
- $= P(X_n|X_{n-1}, \dots, X_1)P(X_{n-1}|X_{n-2}, \dots, X_1)P(X_{n-2}, \dots, X_1)$
- $= P(X_n|X_{n-1}, \dots, X_1)P(X_{n-1}|X_{n-2}, \dots, X_1) \dots P(X_2|X_1)P(X_1)$
- $= \prod_{i=1}^{n} P(X_i|X_{i-1}, \dots, X_1)$
- This is just one way to write the joint distribution as a product of conditional distributions
  - As long as we follow the product rule, we can write this as a product of different conditional distributions

Joint Distribution w/ Chain Rule

- $P(X_1, X_2, \dots, X_n) = \prod_{i=1}^{n} P(X_i | X_{i-1}, \dots, X_1)$
- For each table we need $\prod_{i=1}^{n} d^i$ probabilities
  - Still not less than $O(d^n)$

Conditional Independence

- $P(X_1, X_2|X_3) = P(X_1|X_3)P(X_2|X_3)$
- If X_1 and X_2 and conditionally independent given X_3, what about $P(X_1|X_2, X_3)$?
- $P(X_1, X_2|X_3) = P(X_1|X_2, X_3)P(X_2|X_3)$ //**product rule**
- $P(X_1|X_3)P(X_2|X_3) = P(X_1|X_2, X_3)P(X_2|X_3)$ //**conditional independence**
- $P(X_1|X_3) = P(X_1|X_2, X_3)$

Joint Distribution w/ Chain Rule and Conditional Independence

- $P(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P(X_i | X_{i-1}, \ldots, X_1)$
- We can reduce the size of the tables using conditional independence
- Suppose each variable is conditionally independent of all other variables it is conditioned on given, at most, r variables
- $P(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P(X_i | Vars(X_i))$
  - Where $Vars$ returns the (at most) r variables needed for X_i to be conditionally independent of all other variables
- For each table, we would need $\prod_{i=1}^{n} d^{(r+1)} \; = nd^{(r+1)}$
- Number of entries we need grows **linearly** with the number of variables instead of **exponentially**
- If I have 100 variables that can take on 2 different values
  - Need $2^{100} = 1.3 \times 10^{30}$ probabilities
- However, if each variable can be conditionally independent from others given 3 variables
  - $100 \times 2^{3+1} = 1600$

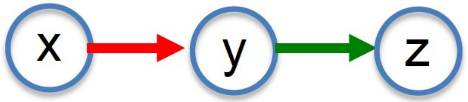
Bayesian Networks

- Bayesian networks give us a way of efficiently representing the full joint distribution using independence and conditional independence in the form of a graphical model
- Using Bayesian networks, we can also perform probabilistic inference in a manner that is efficient in many practical scenarios
- Because probabilistic inference can be computationally intractable in the worst case, we can use approximate inference algorithms when exact inference is infeasible

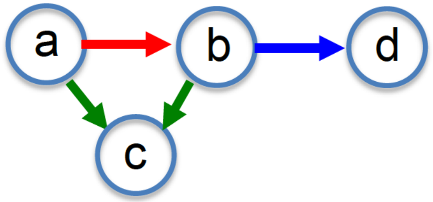
Bayesian Networks

- Directed acyclic graphical model
  - Directed acyclic graph (DAG)
- Directed edges connect pairs of nodes
  - If there is an arrow from X to Y, then X is said to be a parent of Y
- Nodes have a random variable X and a probability table specifying $P(X|Parents(X))$
- $P(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P(X_i|Parents(X_i))$
  - **Key assumption: a random variable is conditionally independent of all of its non-descendents given its parents**

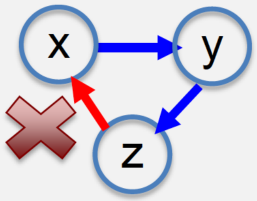
$$p(x,y,z) = p(x)\, p(y\,|\,x)\, p(z\,|\,y)$$



$$p(a,b,c,d) = p(a)\, p(b\,|\,a)\, p(c\,|\,a,b)\, p(d\,|\,b)$$



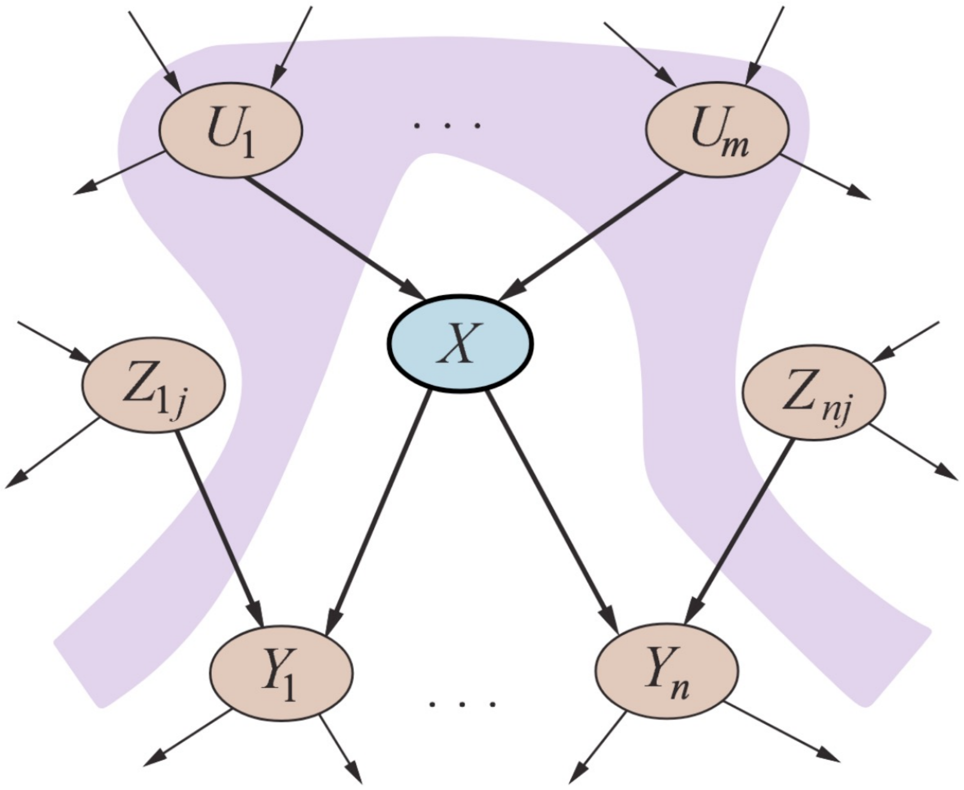
*Graph must be **acyclic***



Corresponds to an order over the variables (chain rule)

Bayesian Networks

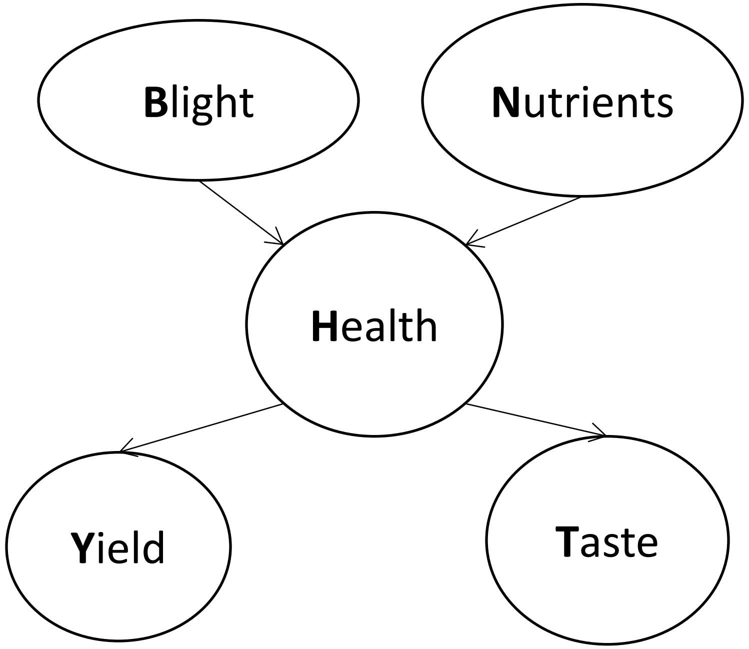
- A variable X is conditionally independent of its non-descendents (Zs) given its parents (Us)



Example: Plant Health Network

*modified from AIMA

P(B=+b)
0.1



P(N=+n)
0.95

B	N	P(H=+h|B,N)
+b	+n	0.5
+b	-n	0.05
-b	+n	0.99
-b	-n	0.6

H	P(Y=+y|H)
+h	0.7
-h	0.1

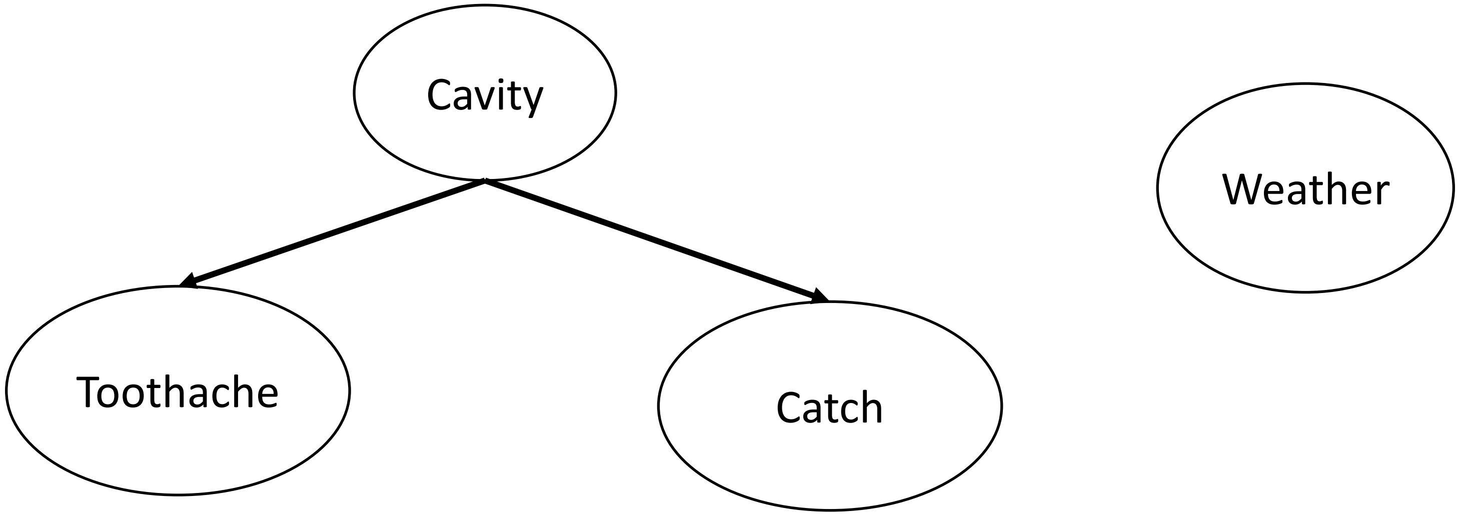
H	P(T=+t|H)
+h	0.9
-h	0.05

Joint Distribution w/ Bayesian Networks

- Chain rule shows: $P(X_1, X_2, \dots, X_n) = \prod_{i=1}^{n} P(X_i | X_{i-1}, \dots, X_1)$
- Conditional independence shows: $P(X_1, X_2, \dots, X_n) = \prod_{i=1}^{n} P(X_i | Vars(X_i))$
- Since Bayesian Networks encode conditional independence of all non-descendents given their parents, we put nodes in topological order
  - That is, any order consistent with the directed graph structure
- Since nodes are in topological order, they are only conditioned on their non-descendents
- Therefore, $P(X_1, X_2, \dots, X_n) = \prod_{i=1}^{n} P(X_i | Parents(X_i))$

Dentist Example

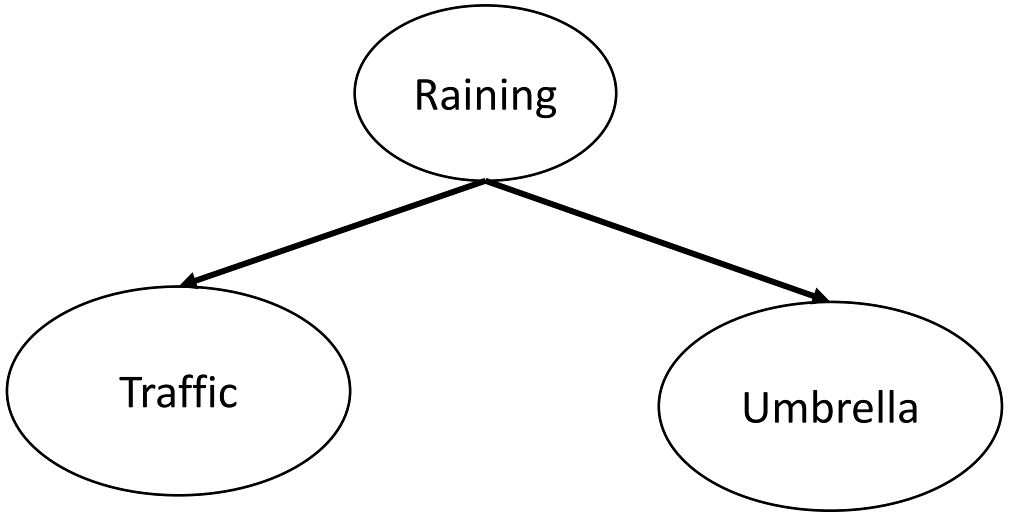
- Toothache
- Cavity
- Catch (dentist tool that catches in a hole in the teeth)
- Weather



- $P(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P(X_i | Parents(X_i))$
- $P(Cavity, Toothache, Catch, Weather)$
- A topological order: Weather, Cavity, Catch, Toothache
- $P(Weather)P(Cavity|Weather)P(Catch|Weather, Cavity)P(Toothache|Catch, Weather, Cavity)$
- $P(Weather)P(Cavity)P(Catch|Cavity)P(Toothache|Cavity)$
- Other node orders are possible as long as they are in topological order
  - E.g. Cavity, weather, toothache, catch

Traffic Example

- Traffic
- Umbrella
- Raining



Traffic Example 2

- T: Traffic
- R: It rains
- L: Low pressure (in atmosphere not in sensor)
- D: Roof leaks
- B: Ballgame
- C: Cavity

Coin Flip Example

- Suppose you flip n coins

X_1 $\quad$ X_2 $\quad \cdots \quad$ X_n

$P(X_1)$

h	0.5
t	0.5

$P(X_2)$

h	0.5
t	0.5

$\cdots$

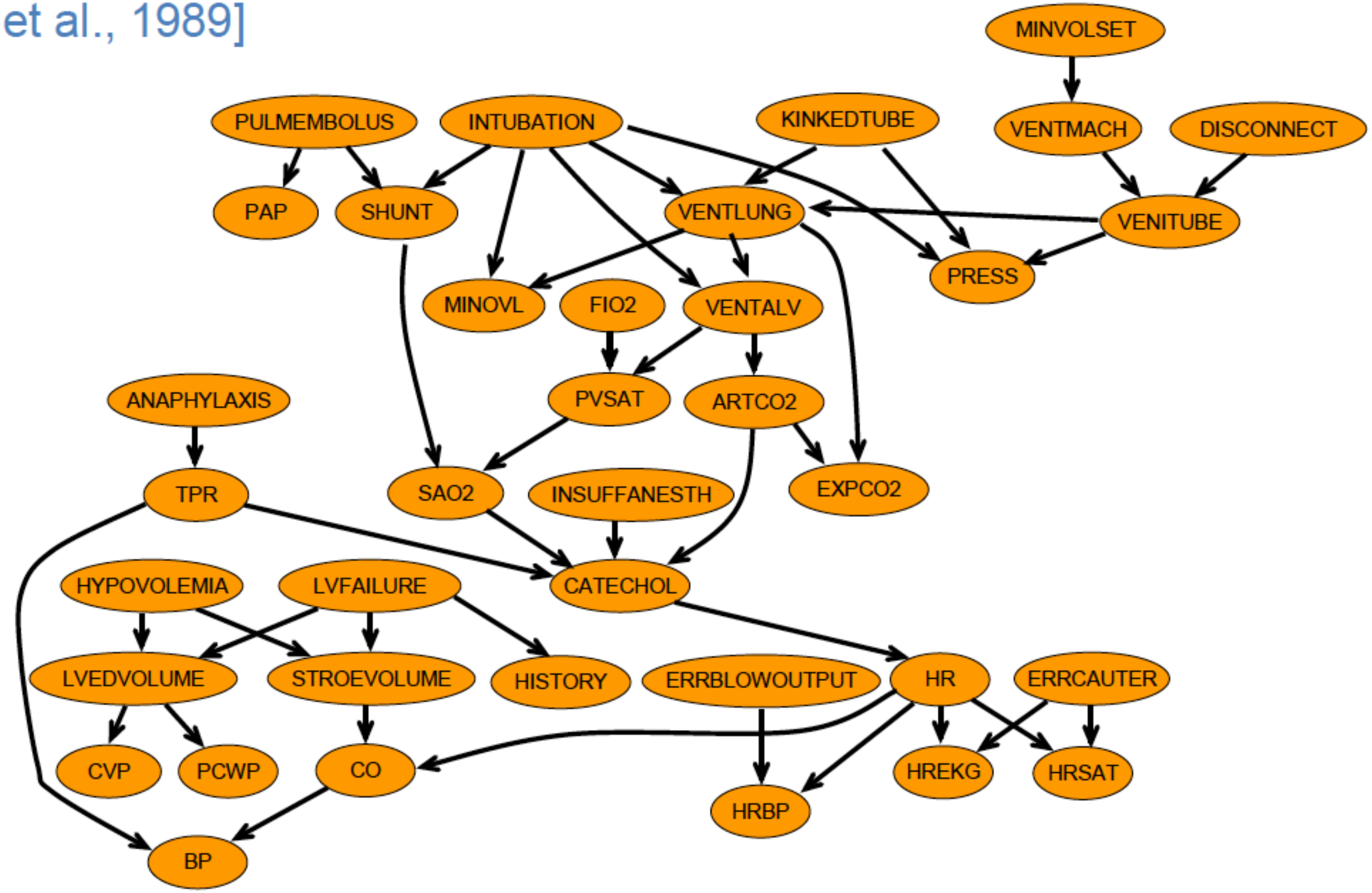
$P(X_n)$

h	0.5
t	0.5

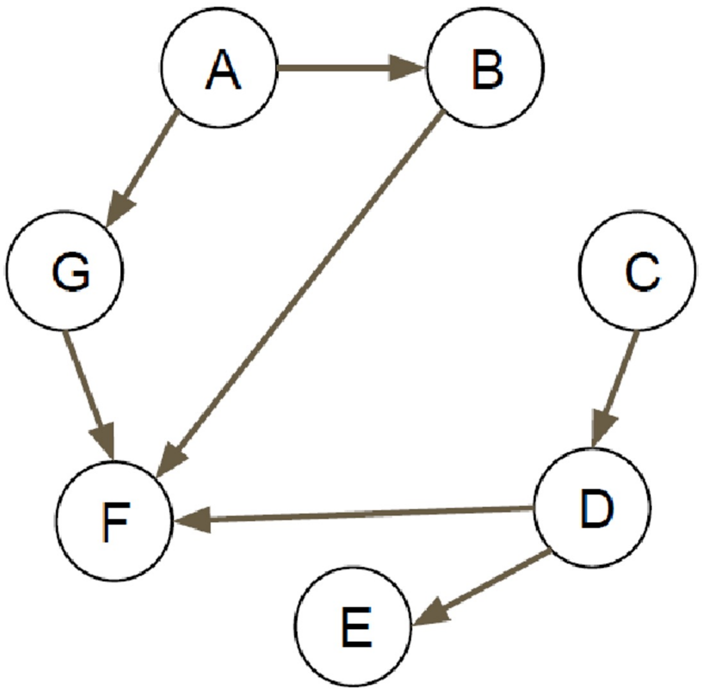
Hospital Alarm Example

- 37 variables with 509 probabilities (instead of $2^{37} \approx 10^{11}$)

[Beinlich et al., 1989]



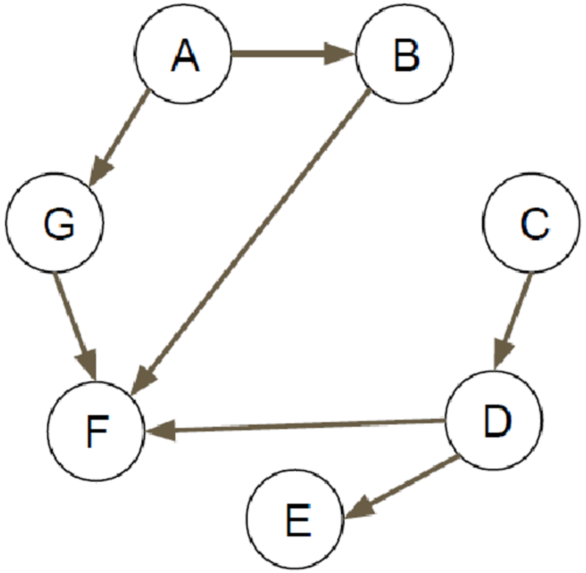
Quick Quiz



Factor the joint probability

$P(A, B, C, D, E, F, G)$

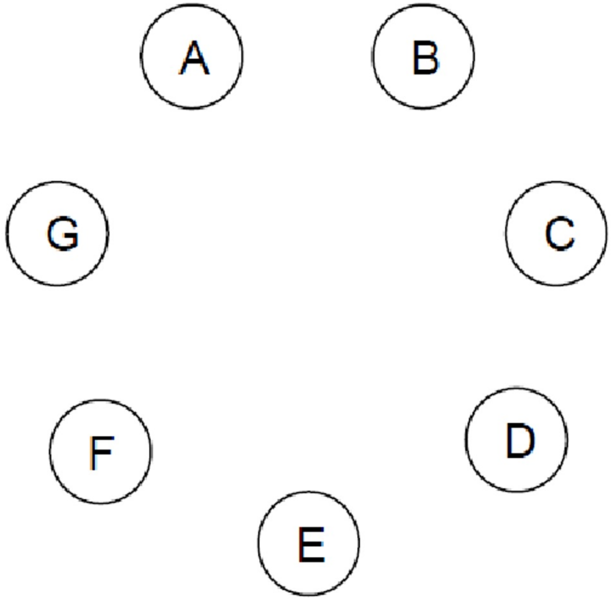
Quick Quiz



Factor the joint probability

$$P(A, B, C, D, E, F, G) \\ = P(A)\, P(B|A)\, P(G|A)\, P(C)\, P(D|C)\, P(E|D)\, P(F|G, B, D)$$

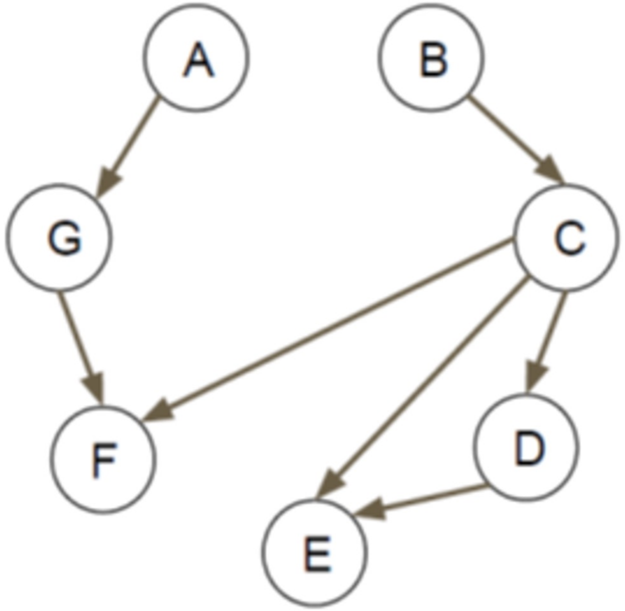
Quick Quiz



Draw the Bayesian network corresponding to the factored conditional probability

$$P(A, B, C, D, E, F, G) = P(A)\, P(B)\, P(G|A)\, P(C|B)\, P(D|C)\, P(E|C, D)\, P(F|G, C)$$

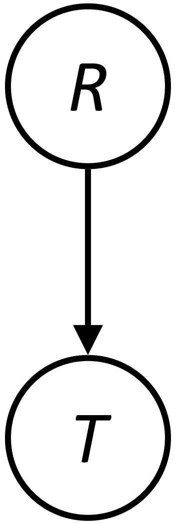
Quick Quiz



Draw the Bayesian network corresponding to the factored conditional probability

$$P(A, B, C, D, E, F, G) = P(A)\, P(B)\, P(G|A)\, P(C|B)\, P(D|C)\, P(E|C, D)\, P(F|G, C)$$

Simple Traffic Example: Causal Direction



$P(R)$

+r	1/4
-r	3/4

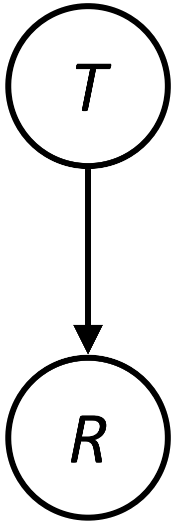
$P(T|R)$

+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

Simple Traffic Example: Reverse Direction



$P(T)$

+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)
  - 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**

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

Summary

- $P(X_1, X_2, \dots, X_n) = \prod_{i=1}^{n} P(X_i | Parents(X_i))$
  - If, for each conditional probability table, we need $\prod_{i=1}^{n} d^{(r+1)} = nd^{(r+1)}$
  - Number of entries we need grows **linearly** with the number of variables instead of **exponentially**
- **Key assumption: a random variable is conditionally independent of all of its non-descendents given its parents**