Introduction to

Artificial Intelligence

Lecture 11 – Bayesian Networks

CS/CNS/EE 154

Andreas Krause

Announcements

- Homework 2 out; due Nov 10.
- Milestone due Nov 3

Probabilistic propositional logic

- Suppose we would like to express uncertainty about *logical propositions*
- Birds can typically fly $P(Bird \Rightarrow CanFly) = .95$
- Propositional symbols ➔ Bernoulli random variables
  - Specify $P(Bird = b, CanFly = f)$
    for all $b, f \in \{true, false\}$

- Probability of a proposition ϕ is the probability mass of all models of ϕ (i.e., all ω that make ϕ true)
- Allows us to avoid specifying large numbers of excepts ("Birds can fly unless X and …")

Random variables

- Bernoulli distribution: "(biased) coin flips"

  D = {H,T}

  Specify P(X = H) = p. Then P(X = T) = 1-p.

  Note: can identify atomic events ω with {X=H}, {X=T}

- Binomial distribution counts the number of heads S

  $p(S=k) = \binom{n}{k} p^k (1-p)^{n-k}$

- Categorical distribution: "(biased) m-sided dice"

  D = {1,…,m}

  Specify P(X = i) = p_i, s.t. $\sum_\iota p_i = 1$

- Multinomial distribution counts the number of outcomes for each side

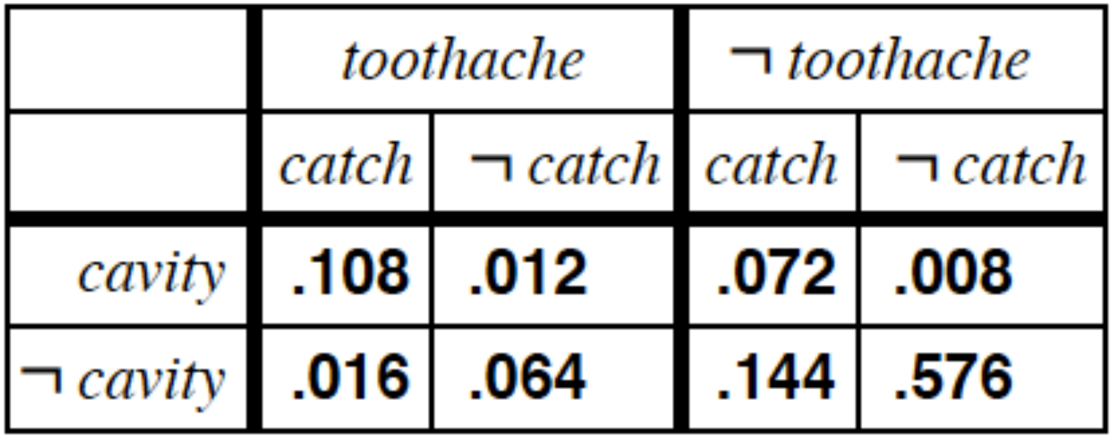
Joint distributions

- Instead of random variable, have random vector

$$\mathbf{X}(\omega) = [X_1(\omega), \ldots, X_n(\omega)] \in D^n$$

- Can specify $P(X_1=x_1,\ldots,X_n=x_n)$ directly

  (atomic events are assignments $x_1,\ldots,x_n$)
- Joint distribution describes relationship among all variables
- Example:



	toothache		*¬ toothache*	
	catch	*¬ catch*	*catch*	*¬ catch*
cavity	.108	.012	.072	.008
¬ cavity	.016	.064	.144	.576

Problems with high-dim. distributions

- Suppose we have *n* propositional symbols
- How many parameters do we need to specify $P(X_1=x_1,\dots,X_n=x_n)$?

X_1	X_2	…	X_{n-1}	X_n	P(X)
0	0	…	0	0	.01
0	0	…	1	0	.001
0	0	…	1	1	.213
…	…	…	…	…	
1	1	…	1	1	.0003

2^n-1 parameters! ☹

Marginal distributions

- Suppose we have joint distribution P(X_1,…,X_n)
- Then

$$P(X_i = x_i) = \sum_{x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_n} P(x_1,\ldots,x_n)$$

Need, because want to compute

$$P(X_1 = T \mid X_3 = F, X_5 = F) = \frac{P(X_1 = T, X_3 = F, X_5 = F)}{P(X_3 = F, X_5 = F)}$$

Marginal Dist.

- If all X_i binary: How many terms?

2^{n-1}

Independent RVs

- What if RVs are independent?

  $P(X_1=x_1,\ldots,X_n=x_n) = P(x_1)\, P(x_2) \ldots P(x_n)$

- How many parameters are needed in this case?

  n

- How about computing $P(x_i)$?

  Indep: $P(X \mid Y, Z) = P(X)$

- Independence too strong assumption… Is there something weaker?

Key concept: Conditional independence

- How many parameters? $P(Toothache, Cavity, Catch)$
- If I know there's a *cavity*, knowing *toothache* won't help predict whether the probe *catches*

- *P(Catch | Cavity, Toothache) = P(Catch | Cavity)*
  for all values of *Catch, Cavity* and *Toothache*

Key concept: Conditional independence

- Random variables X and Y cond. indep. given Z if for all x, y, z:

  P(X = x, Y = y | Z = z) = P(X =x | Z = z) P(Y = y| Z= z)

- If P(Y=y |Z=z)>0, that's equivalent to

  P(X = x | Z = z, Y = y) = P(X = x | Z = z)

Similarly for sets of random variables **X**, **Y**, **Z**

We write: $P \models \mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z}$

Properties of Conditional Independence

- **Symmetry**

$$\mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z} \Rightarrow \mathbf{Y} \perp \mathbf{X} \mid \mathbf{Z}$$

- **Decomposition**

$$\mathbf{X} \perp \mathbf{Y}, \mathbf{W} \mid \mathbf{Z} \Rightarrow \mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z}$$

- **Contraction**

$$(\mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z}) \wedge (\mathbf{X} \perp \mathbf{W} \mid \mathbf{Y}, \mathbf{Z}) \Rightarrow \mathbf{X} \perp \mathbf{Y}, \mathbf{W} \mid \mathbf{Z}$$

- **Weak union**

$$\mathbf{X} \perp \mathbf{Y}, \mathbf{W} \mid \mathbf{Z} \Rightarrow \mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z}, \mathbf{W}$$

- **Intersection**

$$(\mathbf{X} \perp \mathbf{Y} \mid \mathbf{W}, \mathbf{Z}) \wedge (\mathbf{X} \perp \mathbf{W} \mid \mathbf{Y}, \mathbf{Z}) \Rightarrow \mathbf{X} \perp \mathbf{Y}, \mathbf{W} \mid \mathbf{Z}$$

Holds only if distribution is positive, i.e., P>0

Example: Naïve Bayes Models

- Suppose we have multiple effects with a single cause
- E.g.: Flu causes fever, runny nose, cough, …
- Effects are *conditionally independent* given cause

Cause Y

Effects $X_1 \ldots X_n$

$X_A \perp X_B \mid Y$, where $A \subset \{1, \ldots n\}$

Eg: $A = \{i_1, \ldots i_\ell\}$

$X_A = [X_{i_1}, \ldots X_{i_\ell}]$

$$P(Y, X_1, \ldots X_n) = P(Y)\, P(X_1 \mid Y)\, \underbrace{P(X_2 \mid Y, X_1)}_{P(X_2 \mid Y)} \cdot \ldots \underbrace{P(X_n \mid Y, X_1, \ldots X_{n-1})}_{P(X_n \mid Y)}$$

$\Rightarrow$ $2n+1$ parameters

Inference in the Naïve Bayes model

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

$$P(Y | X_1 = T) = \frac{1}{Z} P(Y, X_1 = T) = \frac{1}{Z} \sum_{x_2} \sum_{x_3} \sum_{x_4} \dots \sum_{x_n} P(Y) P(X_1 = T | Y) \prod_{i=2}^{n} P(X_i | Y)$$

$$= \frac{1}{Z} P(Y) P(X_1 = T | Y) \sum_{x_2} \sum_{x_3} \dots \sum_{x_n} P(X_2 | Y) \dots P(X_n | Y)$$

$$= \frac{1}{Z} P(Y) P(X_1 = T | Y) \underbrace{\sum_{x_2} P(X_2 | Y) \underbrace{\sum_{x_3} \underbrace{P(X_3 | Y) \dots \underbrace{\sum_{x_n} P(X_n | Y)}_{=1}}_{=1}}_{=1}}_{>1}$$

$$= \frac{1}{Z} P(Y) P(X_1 = T | Y)$$

$\Rightarrow$ Summed only over $O(n)$ terms !!

Does this work in general?

- Conditional parameterization
  (instead of joint parameterization)
- For each RV, specify $P(X_i \mid \mathbf{X}_{\mathbf{A}_i})$ for set $\mathbf{X}_{\mathbf{A}_i}$ of RVs
- Then use chain rule to get joint parametrization

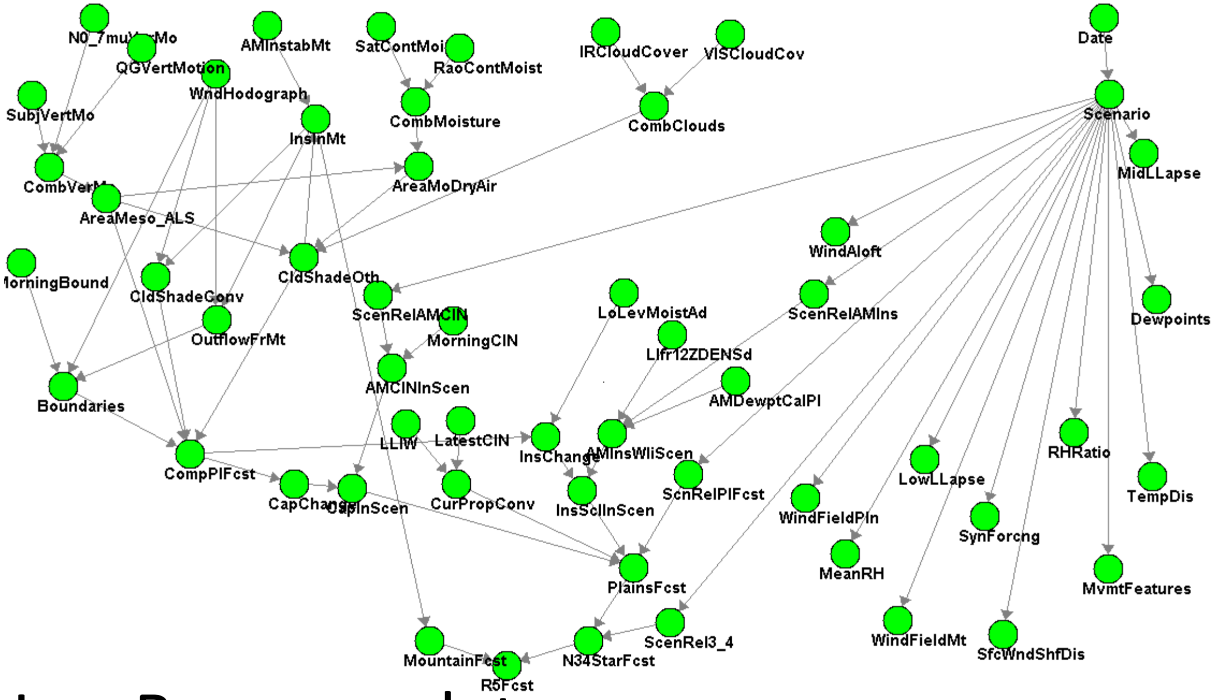
$$P(X_1 \dots X_n) = \prod P(X_i \mid X_{A_i})$$

- Number of parameters? $= \sum_i 2^{|A_i|}$
- Have to be careful to guarantee legal distribution…

If one chooses arbitrary $P(X|Y)$ and $P(Y|X)$ in general $\exists$ $P(X,Y)$ with those cond. distributions

Bayesian networks

- Compact representation of distributions over large number of variables
- (Often) allows efficient exact inference (computing marginals, etc.)



JavaBayes applet

HailFinder

56 vars

~ 3 states each

➔~10^{26} terms

> **10.000 years**

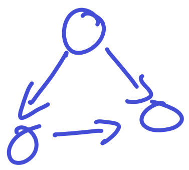
on Top

supercomputers

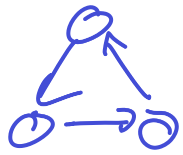
Causal parametrization

- Graph with directed edges from (immediate) causes to (immediate) effects

DAG



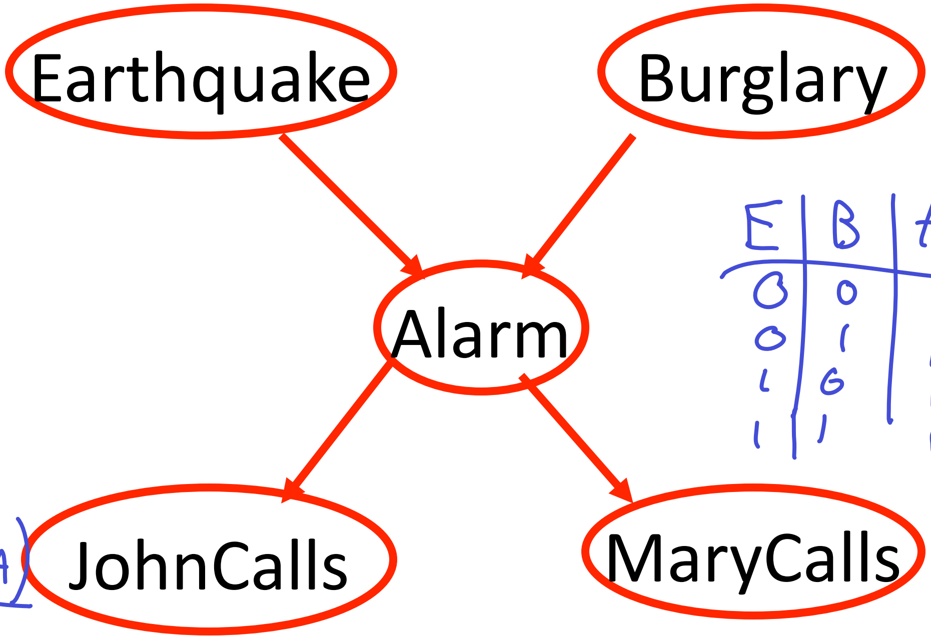
ok



not ok

E	P(E)
1	.01
0	.99

B	P(B)
1	.1
0	.9



E	B	A	P(A|E,B)
0	0	1	.0001
0	1	1	.8
1	0	1	.6
1	1	1	.9

A	J	P(J|A)
0	1	.3
1	1	.9

Bayesian networks

- A **Bayesian network structure** is a directed, acyclic graph G, where each vertex s of G is interpreted as a random variable X_s (with unspecified distribution)

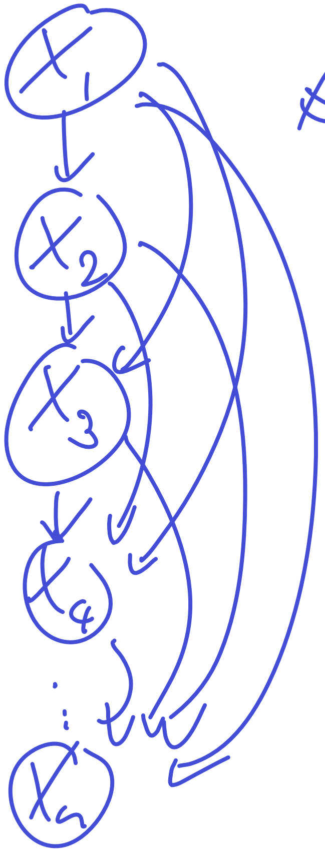
- A **Bayesian network** (G,P) consists of
  - A BN structure G and ..
  - ..a set of conditional probability distributions (CPTs) $P(X_s \mid \mathbf{Pa}_{X_s})$, where $\mathbf{Pa}_{X_s}$ are the parents of node X_s such that
  - (G,P) defines joint distribution

$$P(X_1, ..., X_n) = \prod_i P(X_i \mid \mathbf{Pa}_{X_i})$$

Bayesian networks

- Can every probability distribution be described by a BN?

$P(X_1 \ldots X_n) = P(X_1)\, P(X_2 \mid X_1) \cdots P(X_n \mid X_1 \ldots X_{n-1})$



# params:

$1 + 2 + 4^2 + 2^3 + \ldots 2^{n-1}$

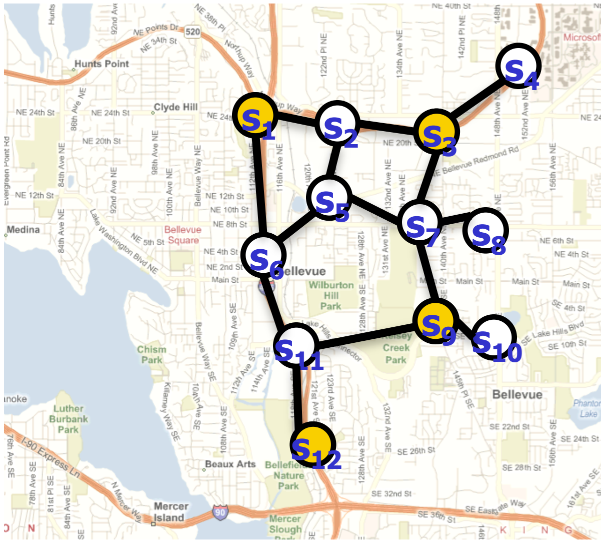
$= 2^n - 1$

Yes!

Representing the world using BNs

True distribution P’
with cond. ind. I(P’)

represent



Bayes net (G,P)
with I(P)

- Want to make sure that I(P) is a subset of I(P’)
- Need to understand conditional independence properties of BN (G,P)

Defining a Bayes Net

- Given random variables and known conditional independences
- Pick ordering $X_1,\ldots,X_n$ of the variables
- For each X_i
  - Find minimal subset A of $\{X_1,\ldots,X_{i-1}\}$ such that $X_i \perp \mathbf{X}_{\bar{A}} \mid \mathbf{X}_A$ where $\bar{A} = \{1, \ldots, \cancel{n}i\} \setminus (A \cup \{i\})$
  - Specify / learn $P(X_i \mid A)$

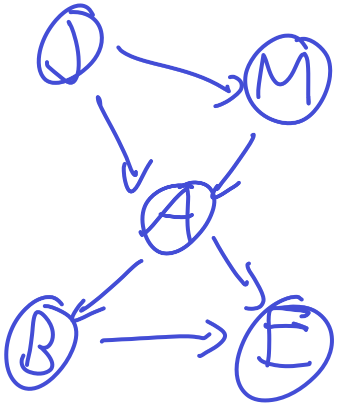
Theorem: Bayes' Nets defined this way are *sound*

- *Does only encode cond. indep. present in P*

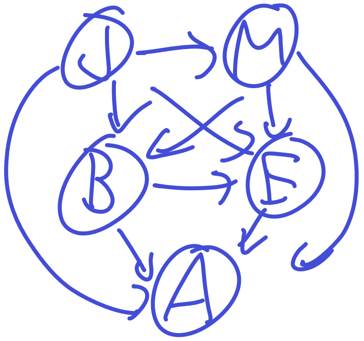
Ordering matters a lot for compactness of representation! More later this course.

Example

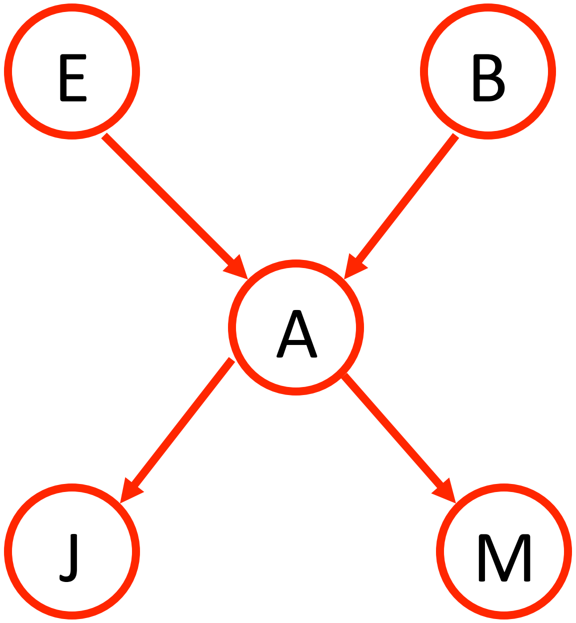
- Suppose we use the ordering JohnCalls, MaryCalls, Alarm, Burglary, Earthquake



- What if ordering is J, M, B, E, A?



Which kind of CI does a BN imply?



$E \perp B$?

$$P(E,B) = \sum_{a,j,m} P(E,B,a,j,m)$$

$$= \sum_{a,j,m} P(E)\cdot P(B)\cdot P(a|E,B)\cdot P(j|a)\,P(m|a)$$

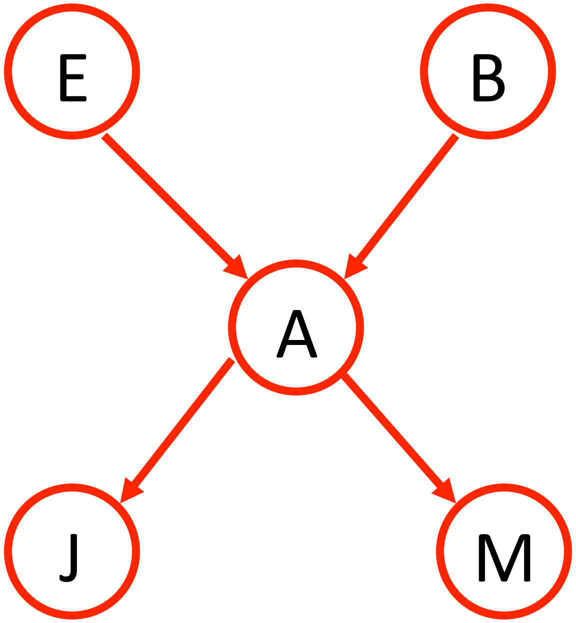
$$= P(E)\,P(B) \sum_{a,j,m} P(a|EB)\,P(j|a)\,P(m|a)$$

$$= P(E)P(B) \underbrace{\sum_{a} P(a|EB)}_{=1} \underbrace{\sum_{j} P(j|a)}_{=1} \underbrace{\sum_{m} P(m|a)}_{=1}$$

$$= P(E)P(B)$$

$\square$

Which kind of CI does a BN imply?



$J \perp M \mid A$?

$$\boxed{P(J \mid A\,M) = \frac{P(J,A,M)}{P(A,M)}}$$

$$P(J,A,M) = \sum_{eb} P(J,A,M,e,b)$$

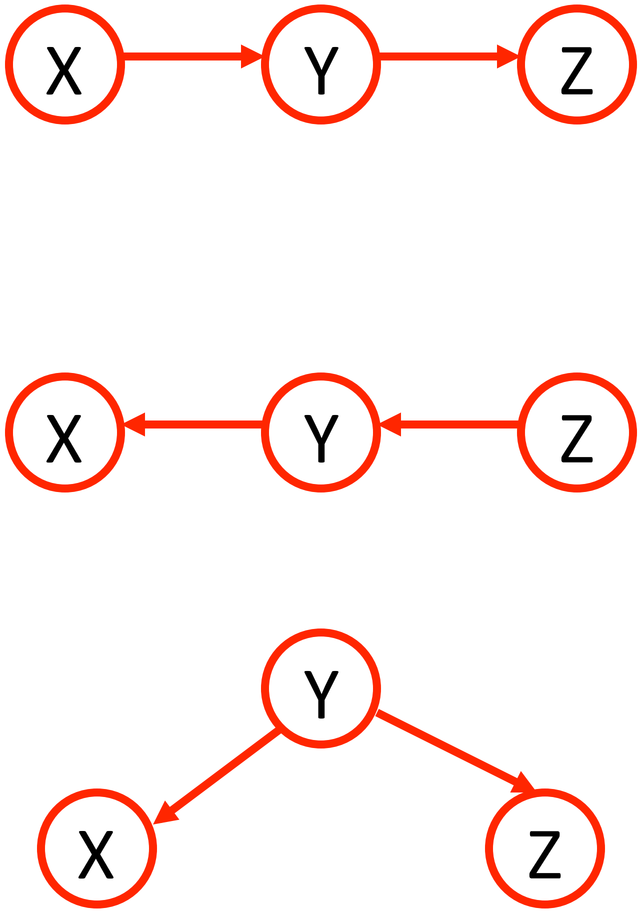
$$= \sum_{eb} P(e)\,P(b)\,P(A|e,b)\,P(J|A)\,P(M|A)$$

$$= P(J|A)\,\underbrace{P(M|A) \underbrace{\sum_{eb} \underbrace{P(A|e,b)\,P(e)\,P(b)}_{P(A,e,b)}}_{P(A)}}_{P(A,M)}$$

$$\Rightarrow P(J,A,M) = P(J|A) \cdot P(A,M)$$

$$\Rightarrow P(J|AM) = P(J|A) \qquad \square$$

BNs with 3 nodes

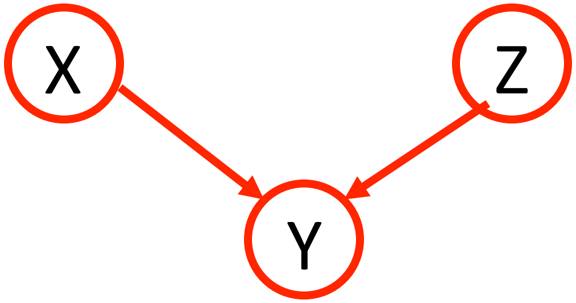


$X \perp Z \mid Y$

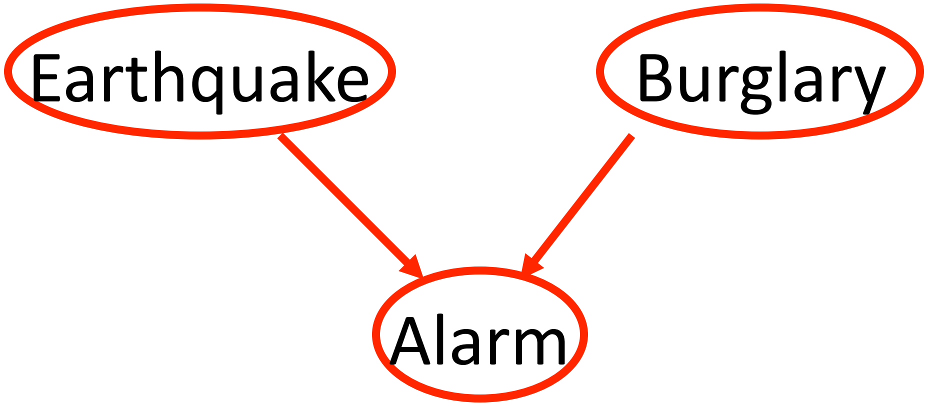
$\neg(X, Z)$

$X \perp Z$

$\neg(X \perp Z \mid Y)$



V-structures



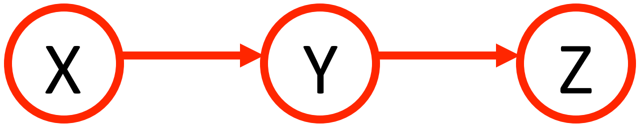
Can happen that

$P(B=T \mid A=T, E=T) < P(B=T \mid A=T)$

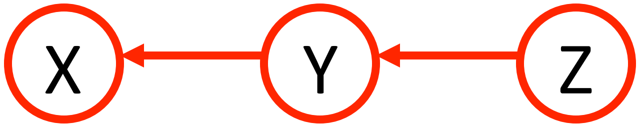
"Explaining away"

BNs with 3 nodes

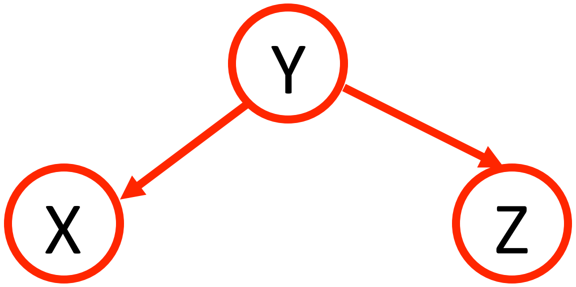
Indirect causal effect



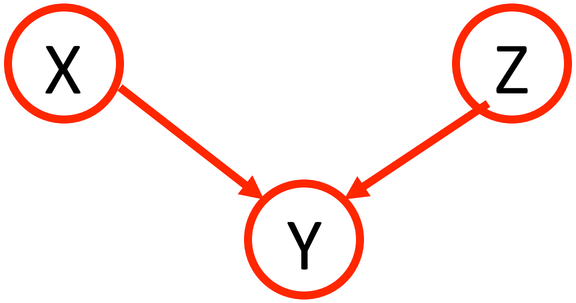
Indirect evidential effect



Common cause

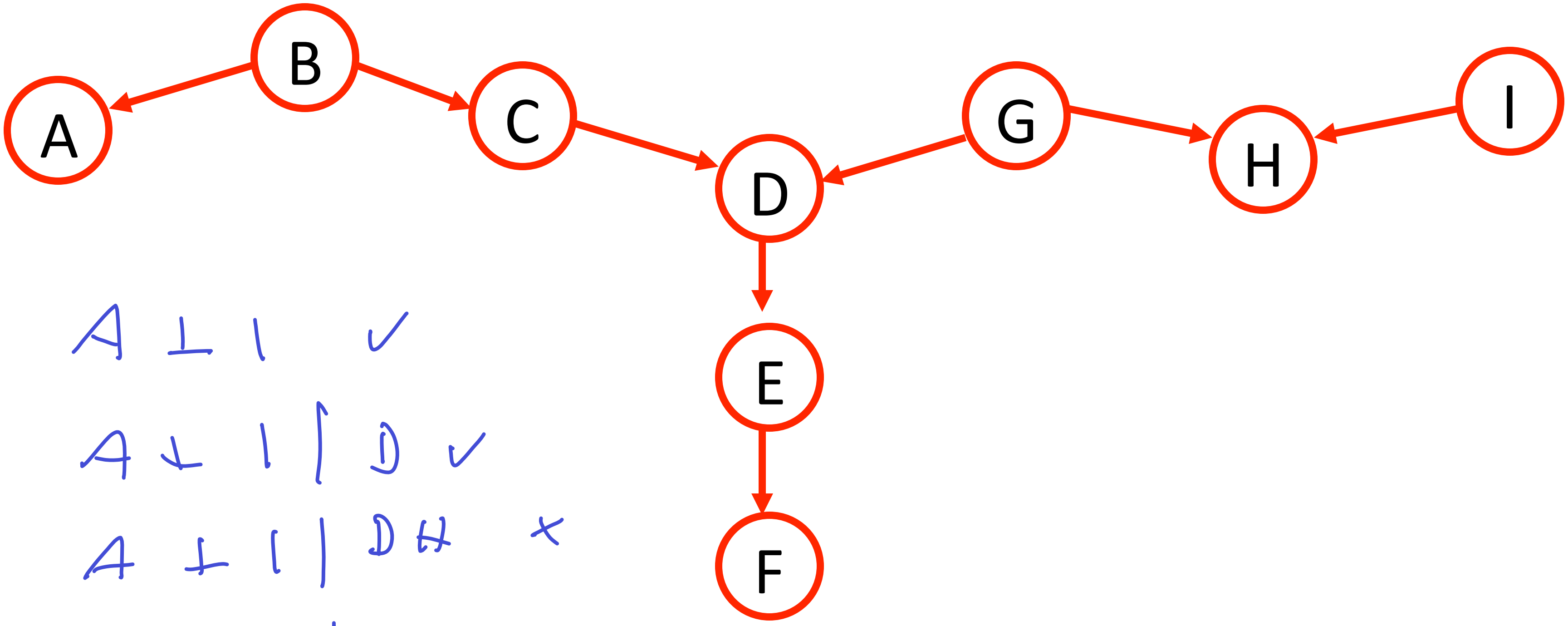


Common effect



Active trails

- When are A and I independent?



$A \perp I$ ✓

$A \perp I \mid D$ ✓

$A \perp I \mid D\,H$ ✗

$A \perp I \mid H, F$

Active trails

- An undirected path in BN structure G is called active trail for observed variables **O** µ $\{X_1,\ldots,X_n\}$, if for every consecutive triple of vars X,Y,Z on the path
    - X → Y → Z and Y is unobserved (Y ∉ **O**)
    - X ← Y ← Z and Y is unobserved (Y ∉ **O**)
    - X ← Y → Z and Y is unobserved (Y ∉ **O**)
    - X → Y ← Z and Y *or any of Y's descendants* is observed

- Any variables X_i and X_j for which there is no active trail for observations **O** are called d-separated by **O**
  We write d-sep(X_i;X_j | **O**)
- Sets **A** and **B** are d-separated given **O** if d-sep(X,Y |**O**) for all X in **A**, Y in **B**. Write d-sep(**A; B | O**)