Probabilistic Graphical Models

Lecture 2 – Bayesian Networks Representation

CS/CNS/EE 155
Andreas Krause

Announcements

- Will meet in Steele 102 for now
- Still looking for another 1-2 TAs..
- Homework 1 will be out soon. Start early!! ☺

Multivariate distributions

- Instead of random variable, have random vector

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

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

- Suppose all X_i are Bernoulli variables.
- How many parameters do we need to specify?

Marginal distributions

- Suppose we have joint distribution $P(X_1,\ldots,X_n)$
- Then

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

- If all X_i binary: How many terms?

Rules for random variables

- Chain rule

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

- Bayes' rule

$$P(X \mid Y) = \frac{P(Y \mid X)\, P(X)}{P(Y)}$$

How do we get P(Y)?

Key concept: Conditional independence

- Events α, β conditionally independent given γ if

$$P(\alpha \cap \beta \mid \gamma) = P(\alpha \mid \gamma)\, P(\beta \mid \gamma)$$

- Random variables X and Y cond. indep. given Z if for all x$\in$ Val(X), y$\in$ Val(Y), Z$\in$ Val(Z)

$$P(X = x, Y = y \mid Z = z) = P(X = x \mid Z = z)\, P(Y = y \mid Z = z)$$

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

$$P(X = x \mid Z = z, Y = y) = P(X = x \mid Z = z)$$

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

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

Why is conditional independence useful?

- $P(X_1,\ldots,X_n) = P(X_1)\, P(X_2 \mid X_1) \ldots \underbrace{P(X_n \mid X_1,\ldots,X_{n-1})}_{2^{n-1}}$

  How many parameters?

  $2^0 + 2^1 + 2^2 + \ldots + 2^{n-1} = 2^n - 1$

- Now suppose $X_1 \ldots X_{i-1} \perp X_{i+1} \ldots X_n \mid X_i$ for all i

  Then

  $P(X_1,\ldots,X_n) = \underset{1}{P(X_1)} \cdot \underset{2}{P(X_2 \mid X_1)} \cdot \underset{2}{P(X_3 \mid X_2)} \cdot \ldots \cdot \underset{2}{P(X_n \mid X_{n-1})}$

  $2n-1 \ll 2^n$

  How many parameters? Exponential reduction in # params

- Can we compute $P(X_n)$ more efficiently? Yes (often)

Properties of Conditional Independence

- **Symmetry**
  - $X \perp Y \mid Z \Rightarrow Y \perp X \mid Z$
- **Decomposition**
  - $X \perp Y,W \mid Z \Rightarrow X \perp Y \mid Z$
- **Contraction** "Inverse Decomposition"
  - $(X \perp Y \mid Z) \wedge (X \perp W \mid Y,Z) \Rightarrow X \perp Y,W \mid Z$
- **Weak union**
  - $X \perp Y,W \mid Z \Rightarrow X \perp Y \mid Z,W$
- **Intersection**
  - $(X \perp Y \mid Z,W) \wedge (X \perp W \mid Y,Z) \Rightarrow X \perp Y,W \mid Z$
  - Holds only if distribution is positive, i.e., $P>0$

Key questions

- How do we specify distributions that satisfy particular independence properties?

    → **Representation**

- How can we exploit independence properties for efficient computation?

    → **Inference**

- How can we identify independence properties present in data?

    → **Learning**

Will now see example: Bayesian Networks

Key idea

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

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

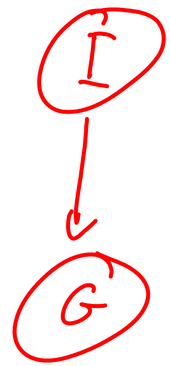
- Have to be careful to guarantee legal distribution…

$P(X \mid Y)$, $P(Y \mid X)$

Does there exist $P(X, Y)$ with above ~~dist~~ cond. distributions

Not in general

Example: 2 variables

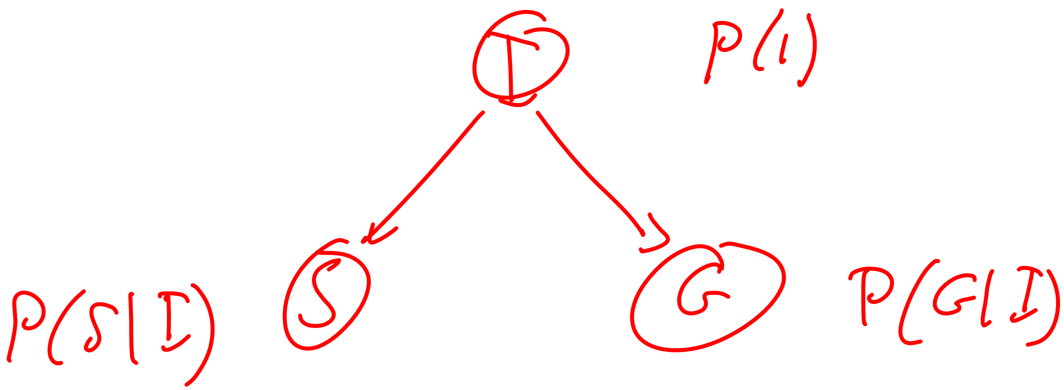


$P(I = VH) = 0.8$

$P(G \mid I)$

I \ G	A	B	
VH	0.8	0.2	$\Sigma = 1$
H	0.6	0.4	$\Sigma = 1$

Example: 3 variables



$$P(I, S, G) = P(I)\, P(G \mid I)\, P(S \mid I)$$

Example: Naïve Bayes models

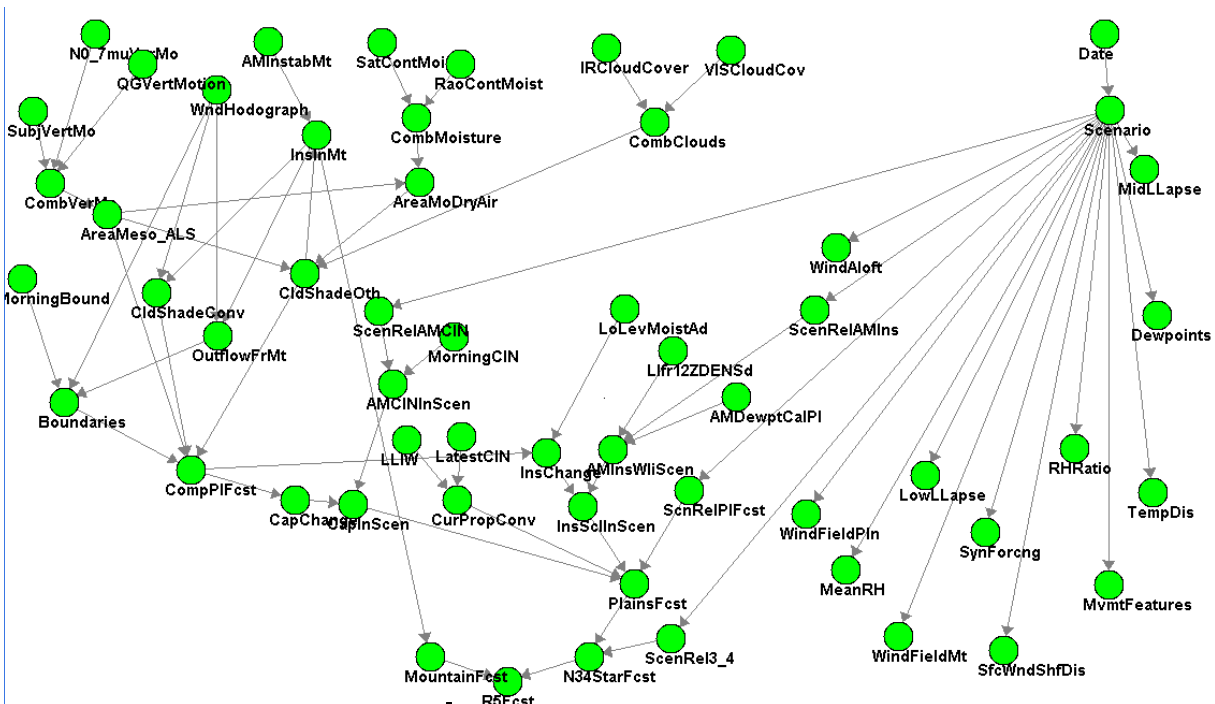
- Class variable Y
- Evidence variables $X_1,\ldots,X_n$
- Assume that $X_A \perp X_B \mid Y$
  for all subsets X_A, X_B of $\{X_1,\ldots,X_n\}$

- Conditional parametrization:
  - Specify $P(Y)$
  - Specify $P(X_i \mid Y)$
- Joint distribution

$$P(X_1 \ldots X_n, Y) = P(Y) \prod_i P(X_i \mid Y)$$

Today: 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

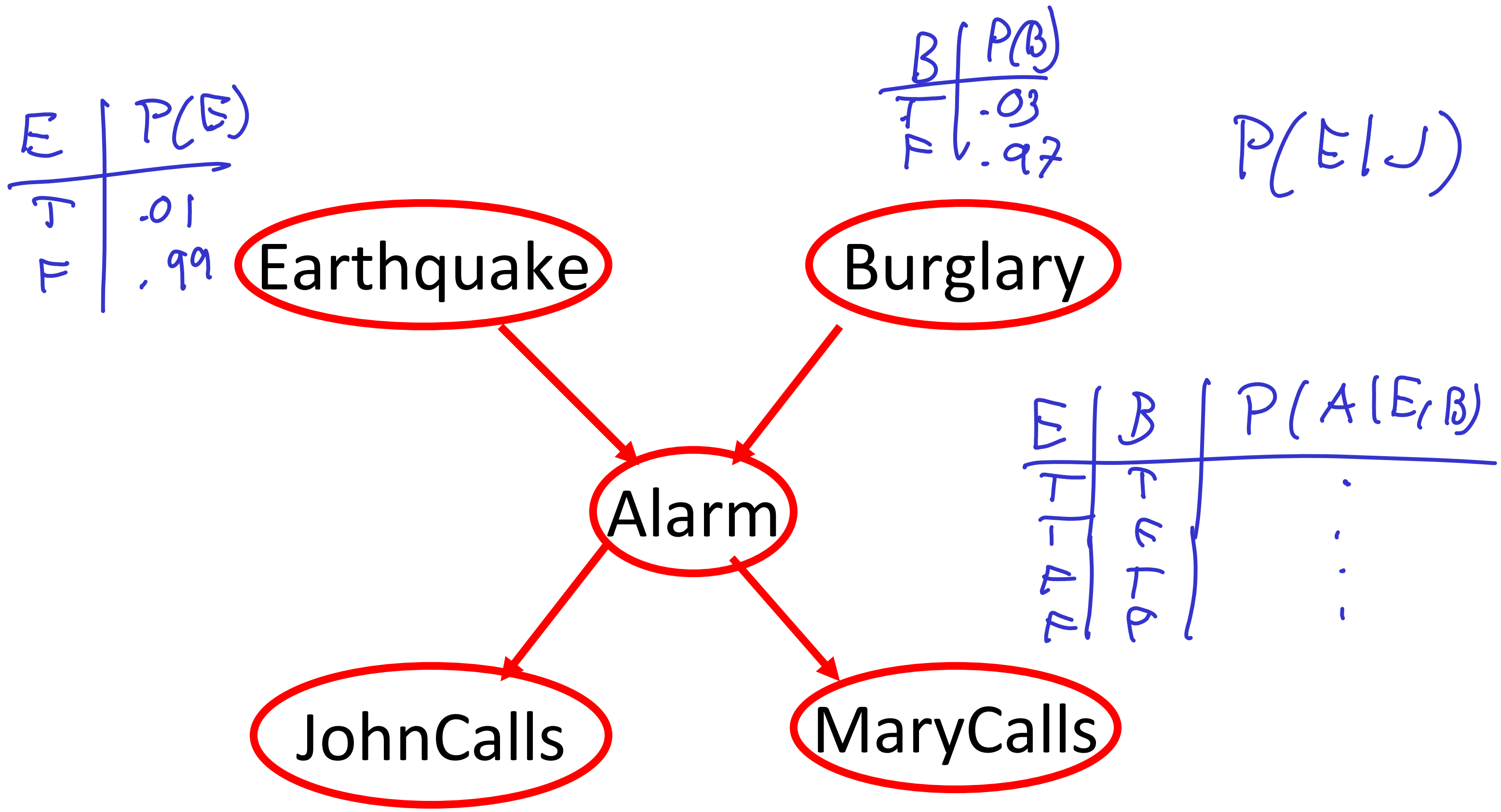
➔ ~10^{26} terms

> **10.000 years**

on Top supercomputers

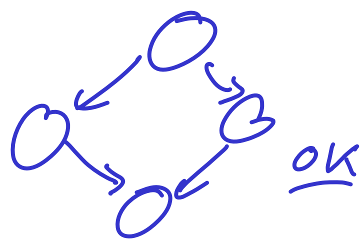
Causal parametrization

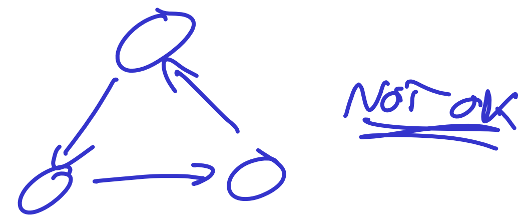
- Graph with directed edges from (immediate) causes to (immediate) effects



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)





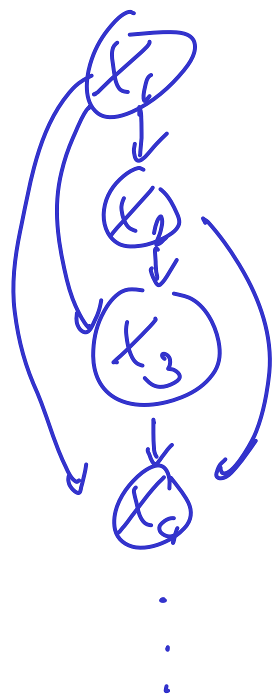
- A **Bayesian network** (G,P) consists of
  - A BN structure G and ..
  - ..a set of conditional probability distributions (CPTs) $P(X_s \mid \mathbf{Pa}_{X_s})$, where $\mathbf{Pa}_{X_s}$ are the parents of node X_s such that
  - (G,P) defines joint distribution

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

Bayesian networks

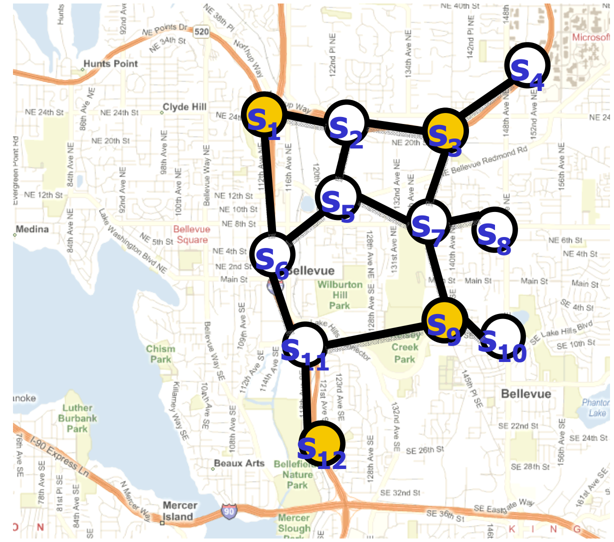
- Can every probability distribution be described by a BN?

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



Representing the world using BNs

represent

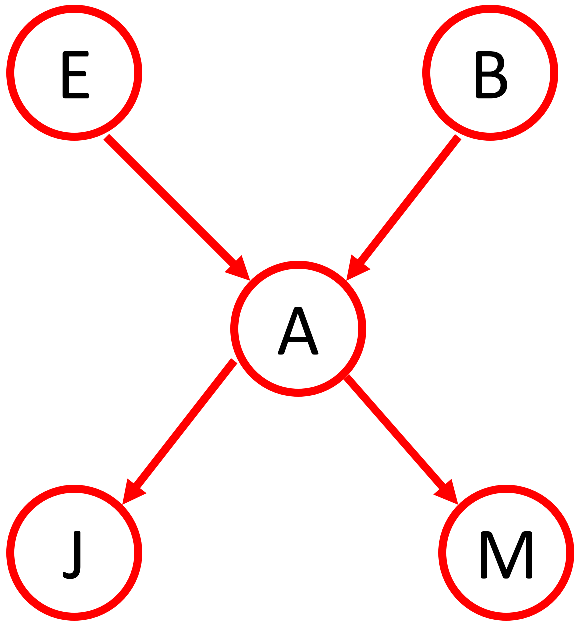


True distribution P'
with cond. ind. I(P')

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

- Want to make sure that I(P) $\subseteq$ I(P')
- Need to understand CI properties of BN (G,P)

Which kind of CI does a BN imply?



$$E \perp B$$

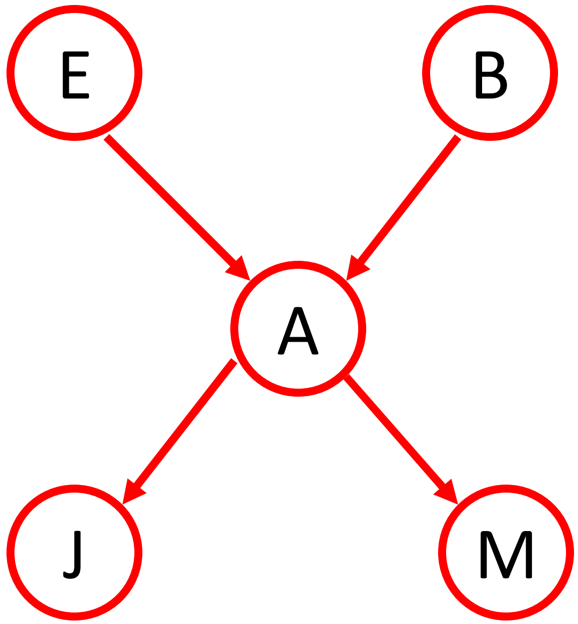
$$P(E,B) = \sum_{AJM} P(E,B,A,J,M)$$

$$= \sum_{AJM} \underline{P(E)P(B)}\, P(A|E,B)\, P(J|A)\, P(M|A)$$

$$= P(E)P(B) \underbrace{\sum_{AJM} P(A|E,B)\, P(J|A)\, P(M|A)}_{1}$$

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

Which kind of CI does a BN imply?



$$J \perp M \mid A$$

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

$$P(J,A,M) = \sum_{E,B} P(J,A,M,E,B)$$

$$= \sum_{E,B} P(E)\,P(B)\,P(A \mid E,B)\,P(J \mid A)\,P(M \mid A)$$

$$= P(J \mid A)\,\underbrace{P(M \mid A)\,\underbrace{\sum_{EB} P(E)\,P(B)\,P(A \mid E,B)}_{=P(A)}}_{=P(M,A)}$$

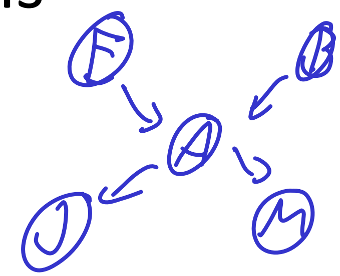
$$\Rightarrow P(J \mid AM) = \frac{P(J,A,M)}{P(AM)} = \frac{P(J \mid A)\,\cancel{P(A,M)}}{\cancel{P(A,M)}} = P(J \mid A) \quad \square$$

Local Markov Assumption

- Each BN Structure G is associated with the following conditional independence assumptions

$J \perp B \mid A$

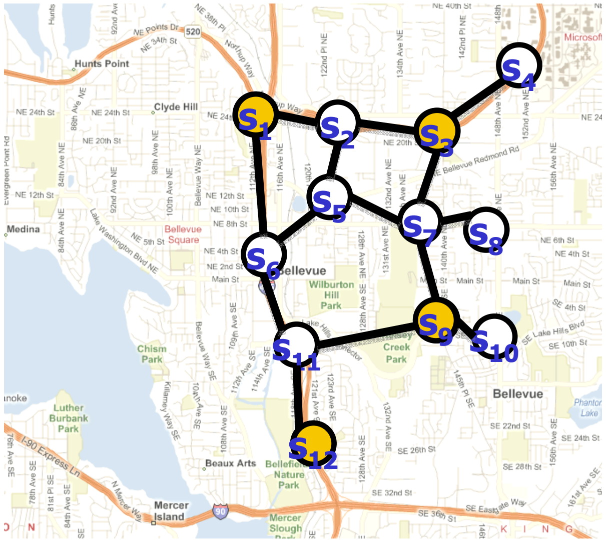
$$X \perp \mathbf{NonDescendents}_X \mid \mathbf{Pa}_X$$



- We write $I_{loc}(G)$ for these conditional independences

- Suppose (G,P) is a Bayesian network representing P
  Does it hold that $I_{loc}(G) \subseteq I(P)$?
  If this holds, we say G is an I-map for P.

Factorization Theorem




$I_{loc}(G) \subseteq I(P)$

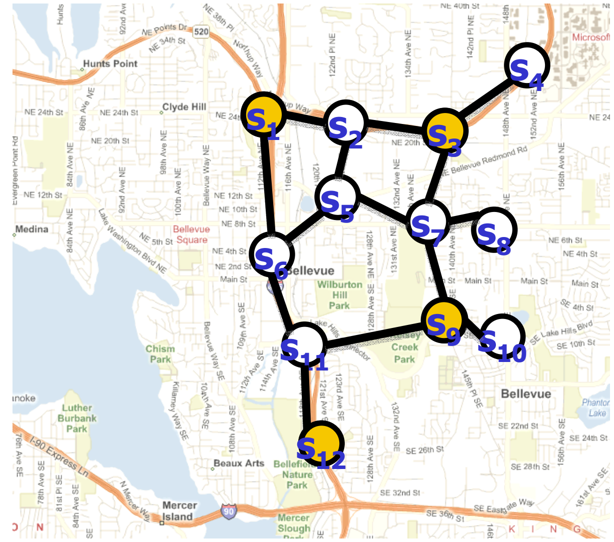
G is an **I-map** of P
(independence map)

True distribution P
can be represented exactly as

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

i.e., P can be represented as
a Bayes net (G,P)

Factorization Theorem



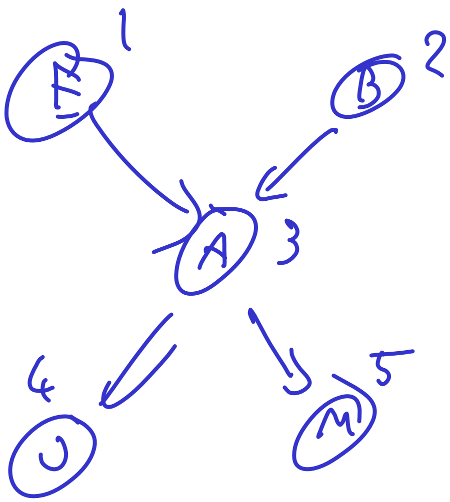
True distribution P
can be represented exactly as
a Bayes net (G,P)

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

$I_{loc}(G) \subseteq I(P)$

G is an **I-map** of P
(independence map)

Proof: I-Map to factorization



Ordering $\pi : \{1 \ldots n\} \to \{1 \ldots n\}$ topological

If: $\forall X_j$ descendant of X_i

$\pi(j) > \pi(i)$

Can find topological ordering in linear time

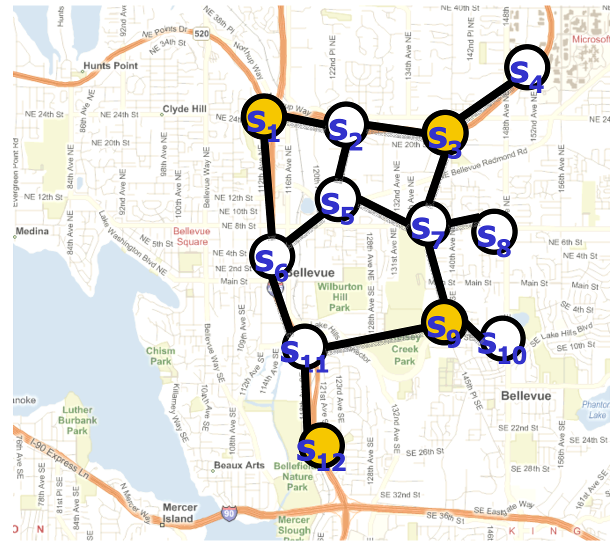
$$P(X_1 \ldots X_n) = \prod \underbrace{P(X_{\pi(i)} \mid X_{\pi(1)} \ldots X_{\pi(i-1)})}_{P(X_{\pi(i)} \mid Pa_{X_{\pi(i)}})}$$ Chain rule

□

Factorization Theorem

True distribution P
can be represented exactly as
a Bayes net (G,P)

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



$I_{loc}(G) \subseteq I(P)$

G is an **I-map** of P
(independence map)

The general case

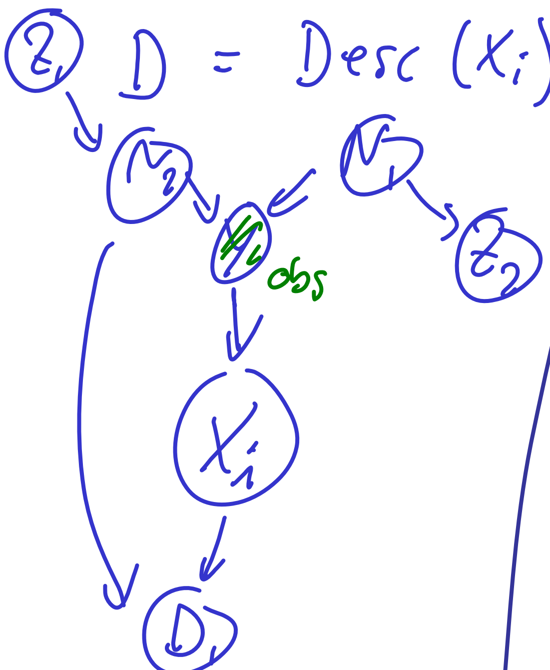
$$P(X_1 \ldots X_n) = \prod P(X_i \mid Pa_{X_i}) \stackrel{!}{\Rightarrow} \forall X_i : \forall N \subseteq \text{Nondesc}_{X_i}$$

$$X_i \perp N \mid Pa_{X_i}$$

$Y = Pa_{X_i}$

$Z = \text{Nondesc.}_{X_i} \setminus (Y \cup N)$

$D = \text{Desc}(X_i)$



$$P(X_i \mid Y,N) = \frac{P(X_i,Y,N)}{P(Y,N)}$$

$$P(X_i,Y,N) = \sum_{Z,D} P(X_i,Y,N,Z,D)$$

$$= \sum_{Z,D} P(X_i \mid Y) \prod_{X \in D} P(X \mid Pa_X) \prod_{X' \in (N \cup Z \cup Y)} P(X' \mid Pa_{X'})$$

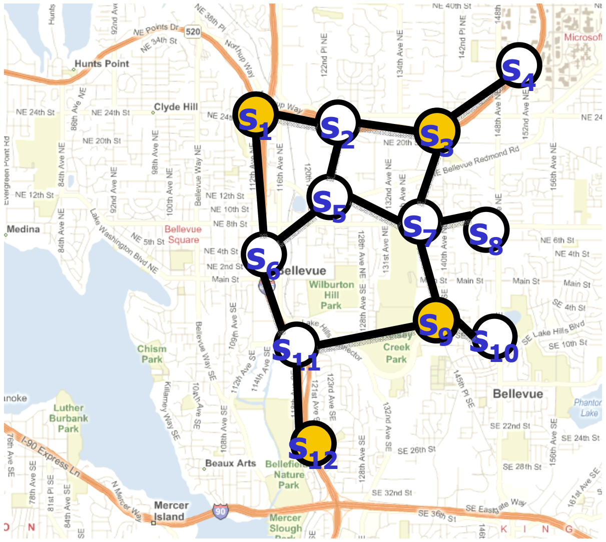
$$= P(X_i \mid Y) \underbrace{\sum_{Z} \prod_{X' \in (N \cup Z \cup Y)} P(X' \mid Pa_{X'})}_{(*)} \underbrace{\sum_{D} \prod_{X \in D} P(X \mid Pa_X)}_{=1}$$

$$P(Y,N) = \sum_{X_i} P(X_i,Y,N) = \underbrace{\sum_{Z} \prod_{X' \in (N \cup Z \cup Y)} P(X' \mid Pa_{X'})}_{=(*)} \underbrace{\sum_{X_i} P(X_i \mid Y)}_{=1}$$

$$\Rightarrow P(X_i \mid Y,N) = P(X_i \mid Y)$$

$\square$

Factorization Theorem



$I_{loc}(G) \subseteq I(P)$

G is an **I-map** of P (independence map)

True distribution P can be represented exactly as Bayesian network (G,P)

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

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 $A \subseteq \{X_1,\ldots,X_{i-1}\}$ such that $X_i \perp X_{\neg\mathcal{A}} \mid A$, where $\neg A = \{X_1,\ldots,X_n\} \setminus A$

    Ensure local Markov property holds!
  - Specify / learn CPD(X_i | A)

    A parents of X_i
- Ordering matters a lot for compactness of representation! More later this course.

Adding edges doesn't hurt

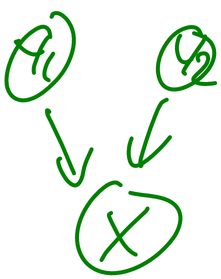
Theorem:

Let G be an I-Map for P, and G' be derived from G by adding an edge. Then G' is an I-Map of P

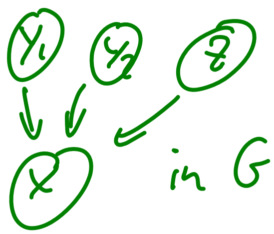
(G' is strictly more expressive than G)

- **Proof** : wtp: $I_{loc}(G') \subseteq I_{loc}(G)$

  Then $I_{loc}(G') \subset I(P)$ since $I_{loc}(G) \subset I(P)$



in G



in G'

wtp: $X \perp \text{Nondesc}(X;G) \mid Pa(X;G)$

$\Rightarrow X \perp \text{Nondesc}(X;G') \mid Pa(X;G')$

$X \perp N, Z \mid Y \Rightarrow X \perp N \mid Y, Z$

holds b/c: Weak Union property of C.I. $\square$

Additional conditional independencies

- BN specifies joint distribution through conditional parameterization that satisfies Local Markov Property
- But we also talked about additional properties of CI
  - Weak Union, Intersection, Contraction, …
- Which additional CI does a particular BN specify?
  - All CI that can be derived through algebraic operations

Local Markov prop. $I_{loc}(G) \subseteq I(G)$

What you need to know

- Bayesian networks
- Local Markov property
- I-Maps
- Factorization Theorem

Tasks

- Subscribe to Mailing list
  https://utils.its.caltech.edu/mailman/listinfo/cs155

- Read Koller & Friedman Chapter 3.1-3.3

- Form groups and think about class projects. If you have difficulty finding a group, email Pete Trautman