Readings:

K&F: 3.4, 14.1, 14.2

BN Semantics 3 – Now it's personal!

Parameter Learning 1

Graphical Models – 10708

Carlos Guestrin

Carnegie Mellon University

September 22nd, 2006

Building BNs from independence properties

- From d-separation we learned:
  - Start from local Markov assumptions, obtain all independence assumptions encoded by graph
  - For most *P*'s that factorize over *G*, I(*G*) = I(*P*)
  - All of this discussion was for a given *G* that is an I-map for *P*

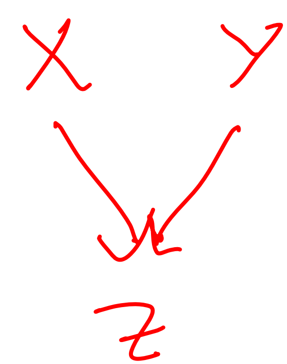
- Now, give me a *P*, how can I get a *G*?
  - i.e., give me the independence ~~assumptions~~ assertions entailed by *P*
  - Many *G* are "equivalent", how do I represent this?
  - Most of this discussion is not about practical algorithms, but useful concepts that will be used by practical algorithms
    - Practical algs next week

Minimal I-maps

- One option:
  - G is an I-map for P
  - G is as simple as possible

- G is a **minimal I-map** for P if deleting any edges from G makes it no longer an I-map

true P $X \perp Y$
and nothing else
vars X, Y, Z

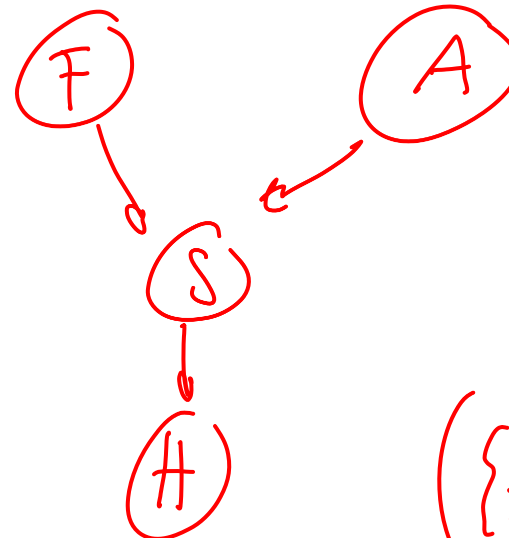


Obtaining a minimal I-map

- Given a set of variables and conditional independence ~~assumptions~~ assertions that P entails
- Choose an ordering on variables, e.g., $X_1, \ldots, X_n$
- For i = 1 to n
  - Add X_i to the network
  - Define parents of X_i, $\mathbf{Pa}_{X_i}$, in graph as the minimal subset of $\{X_1,\ldots,X_{i-1}\}$ such that local Markov assumption holds – X_i independent of rest of $\{X_1,\ldots,X_{i-1}\}$, given parents $\mathbf{Pa}_{Xi}$
  - Define/learn CPT – $P(X_i | \mathbf{Pa}_{Xi})$

1 2 3 4

Flu, Allergy, SinusInfection, Headache



$F \perp A$

$\neg(S \perp F)$

$\neg(S \perp A)$

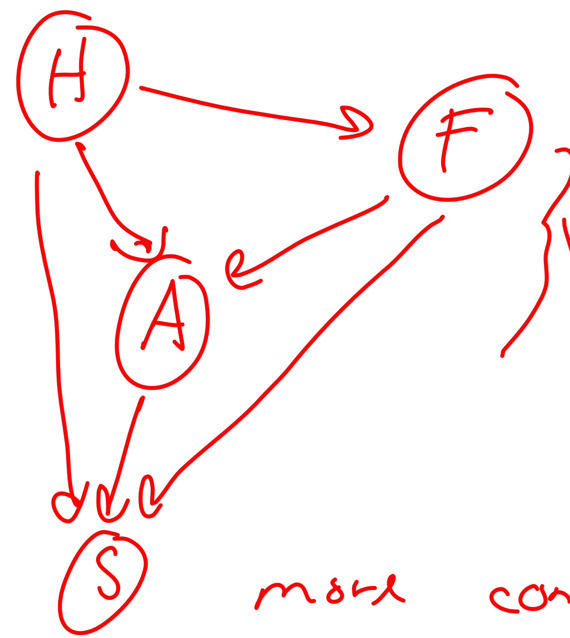
$(\{F, A\} \perp H \mid S)$

Minimal I-map not unique (or minimal)

- Given a set of variables and conditional independence assumptions
- Choose an ordering on variables, e.g., $X_1, \ldots, X_n$
- For i = 1 to n
  - Add X_i to the network
  - Define parents of X_i, $\mathbf{Pa}_{X_i}$, in graph as the minimal subset of $\{X_1,\ldots,X_{i-1}\}$ such that local Markov assumption holds – X_i independent of rest of $\{X_1,\ldots,X_{i-1}\}$, given parents $\mathbf{Pa}_{X_i}$
  - Define/learn CPT – $P(X_i \mid \mathbf{Pa}_{X_i})$

~~3~~ 2 ~~2~~ 3 4 1

Flu, Allergy, SinusInfection, Headache



truth

$(A \perp F)$

$\neg (A \perp F)$

$(A \perp F \mid H)$

more complicated

still minimal I-map

Perfect maps (P-maps)

minimal

- I-maps are not unique and often not simple enough
- Define "simplest" G that is I-map for P
  - A BN structure G is a **perfect map** for a distribution P if $I(P) = I(G)$
- Our goal:
  - Find a perfect map!
  - Must address equivalent BNs

Inexistence of P-maps 1

- XOR (this is a hint for the homework)

$Z = X \oplus Y$

Z true if exactly one of X or Y true

$(X \perp Y)$

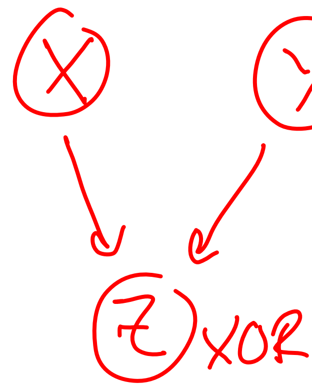
$(Z \perp Y)$

$(Z \perp X)$

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

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

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

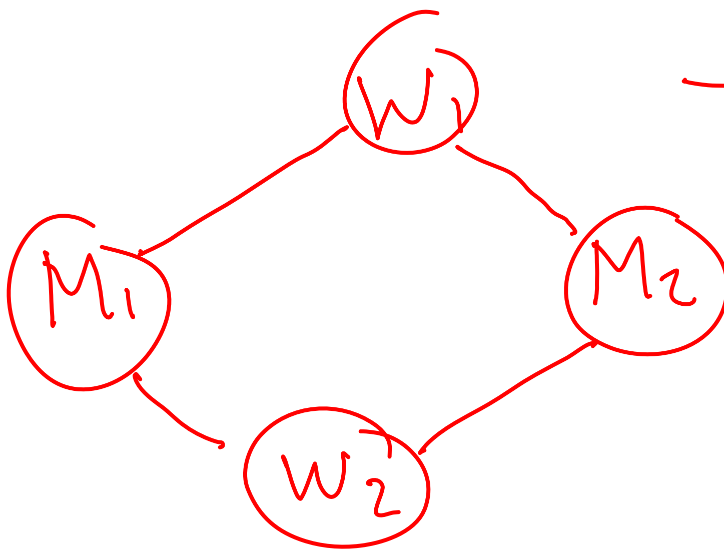


$\neg X \perp Z$

$\neg Y \perp Z$

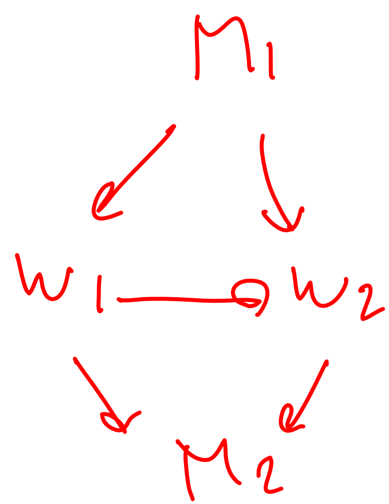
Inexistence of P-maps 2

- (Slightly un-PC) swinging couples example



$\neg(M_1 \perp M_2)$

$(M_1 \perp M_2 \mid W_1, W_2)$



Obtaining a P-map

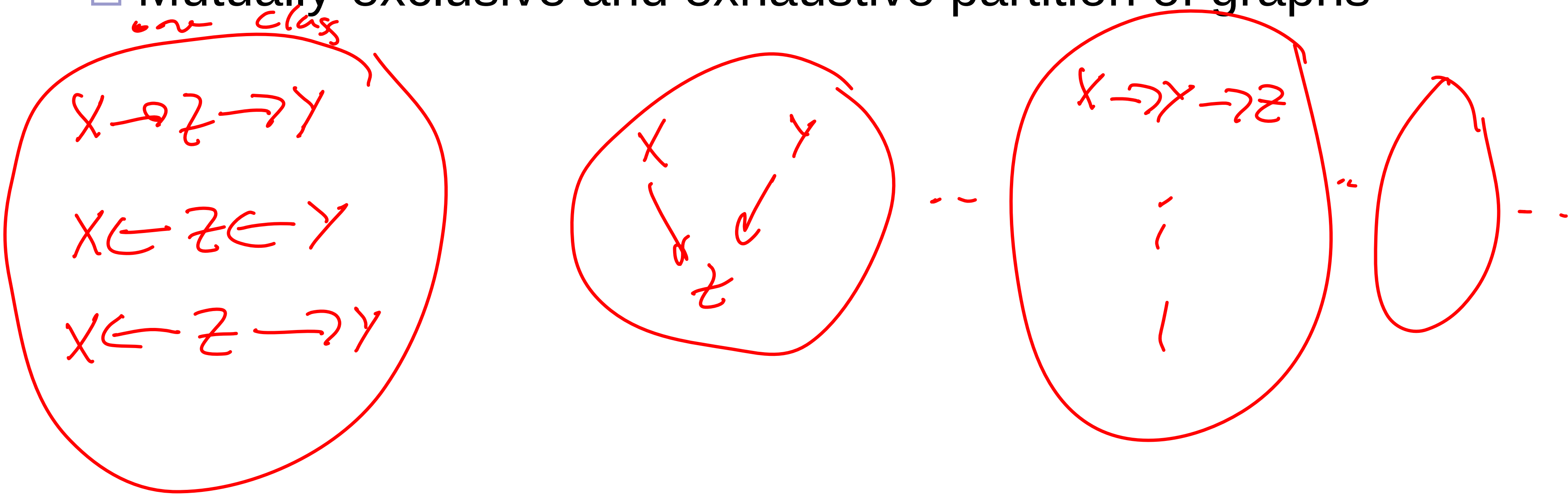
- Given the independence assertions that are true for *P*

- Assume that there exists a perfect map G^*
  - Want to find G^*

- Many structures may encode same independencies as G^*, when are we done?
  - Find all equivalent structures simultaneously!

I-Equivalence

- Two graphs G_1 and G_2 are **I-equivalent** if $I(G_1) = I(G_2)$
- **Equivalence class** of BN structures
  - Mutually-exclusive and exhaustive partition of graphs



- How do we characterize these equivalence classes?

Skeleton of a BN

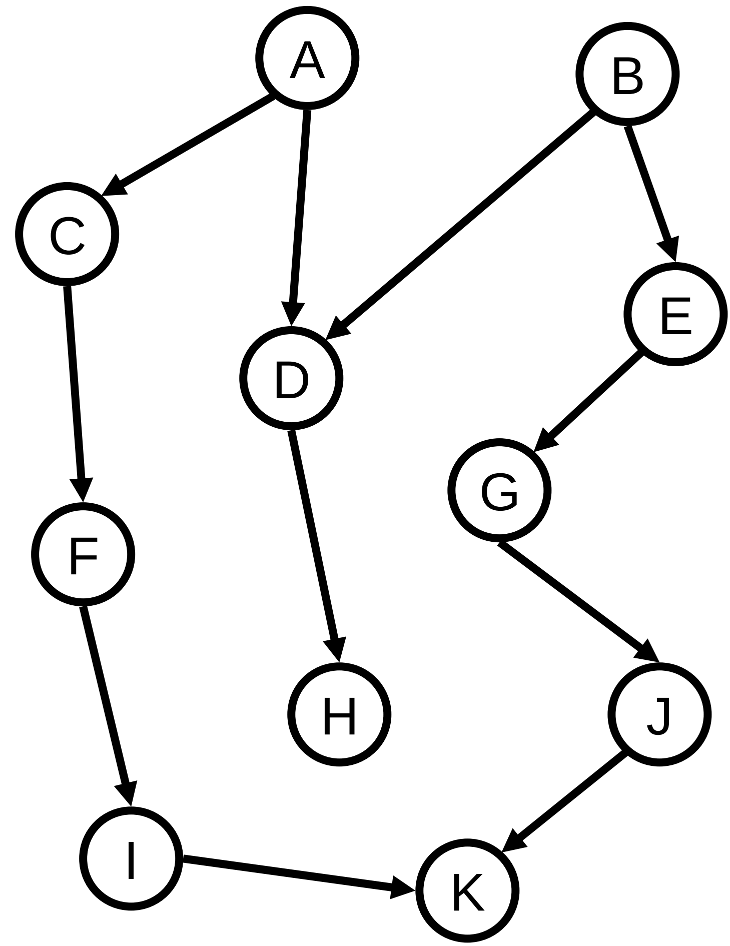
- **Skeleton** of a BN structure *G* is an **undirected graph** over the same variables that has an edge X–Y for every X→Y or Y→X in *G*

can make indep

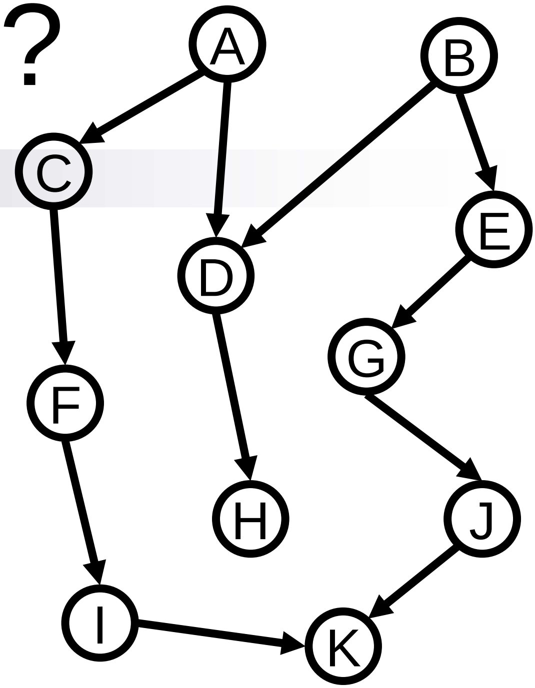
never indep.

- (Little) **Lemma:** Two I-equivalent BN structures must have the same skeleton

counter example



What about V-structures?



- **V-structures are key property of BN structure**

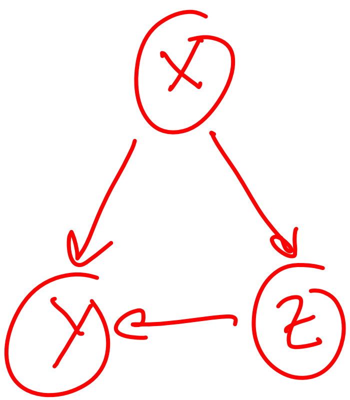
- **Theorem:** If G_1 and G_2 have the same skeleton and V-structures, then G_1 and G_2 are I-equivalent

not if and only if

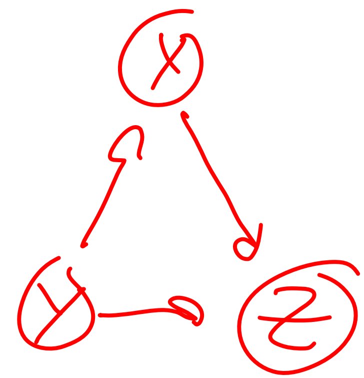
Same V-structures not necessary

- **Theorem:** If G_1 and G_2 have the same skeleton and V-structures, then G_1 and G_2 are I-equivalent

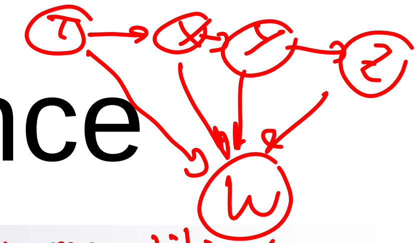
- Though sufficient, same V-structures not necessary



diff. V-structures
but
I-equiv.



Immoralities & I-Equivalence



Immoralities:
(T, Y) (T, Z) ...

- Key concept not V-structures, but "immoralities" (unmarried parents ☺)
  - edge
  - $X \rightarrow Z \leftarrow Y$, with no arrow between X and Y
  - Important pattern: X and Y independent given their parents, but not given Z
  - (If edge exists between X and Y, we have *covered* the V-structure)
- **Theorem:** G_1 and G_2 have the same skeleton and immoralities if and only if G_1 and G_2 are I-equivalent

Obtaining a P-map

- Given the independence assertions that are true for *P*
  - Obtain skeleton
  - Obtain immoralities
- From skeleton and immoralities, obtain every (and any) BN structure from the equivalence class

Identifying the skeleton 1

- When is there an edge between X and Y?

Edge $X - Y$ if $\nexists \mathbf{Z} \subseteq \{X_1 \ldots X_n\} / \{X, Y\}$ $(X \perp Y \mid \mathbf{Z})$

- When is there no edge between X and Y?

$\exists \mathbf{Z} \; : \; (X \perp Y \mid \mathbf{Z})$

Identifying the skeleton 2

- Assume d is max number of parents (d could be n)

- For each X_i and X_j
  - $E_{ij} \leftarrow$ true
  - For each $\mathbf{U} \subseteq \mathbf{X} - \{X_i, X_j\}$, $|\mathbf{U}| \le \cancel{2}d$
    - Is $(X_i \perp X_j \mid \mathbf{U})$?
      - $E_{ij} \leftarrow$ ~~true~~ false
  - If E_{ij} is true
    - Add edge X – Y to skeleton

Identifying immoralities

- Consider X – Z – Y in skeleton, when should it be an immorality?

other active path

$X \to Z \to Y$

$\Downarrow$

must check larger subsets

- Must be $X \to Z \leftarrow Y$ (immorality):
  - When X and Y are **never independent** given **U,** if Z∈**U**

$\nexists\, U \subseteq \{X_1, \ldots X_n\} - \{X, Y\}$ s.t. $Z \in U$ and $(X \perp Y \mid U)$

- Must **not** be $X \to Z \leftarrow Y$ (not immorality):
  - When there exists **U** with Z∈**U**, such that X and Y are **independent** given **U**

possible dirs:

$X \to Z \to Y$

$X \leftarrow Z \leftarrow Y$

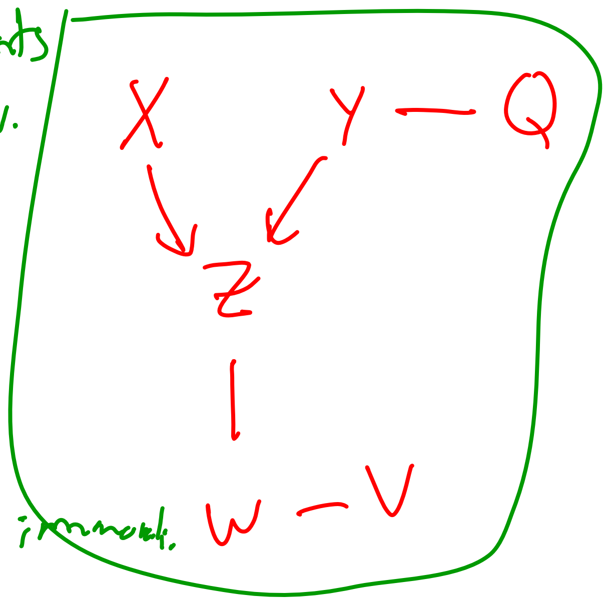
$X \leftarrow Z \to Y$

From immoralities and skeleton to BN structures

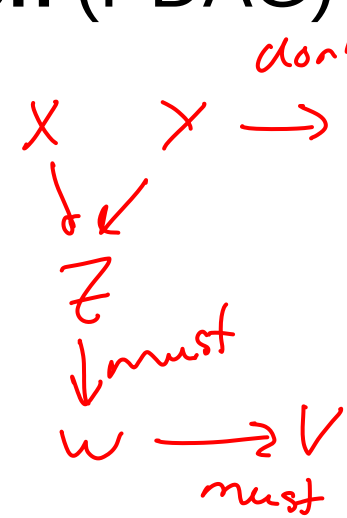
- Representing BN equivalence class as a **partially-directed acyclic graph** (PDAG)

represents equiv. class

all edge assignments
- acyclic
- no new immoral.
in class



⇒



- Immoralities force direction on some other BN edges
- Full (polynomial-time) procedure described in reading

What you need to know

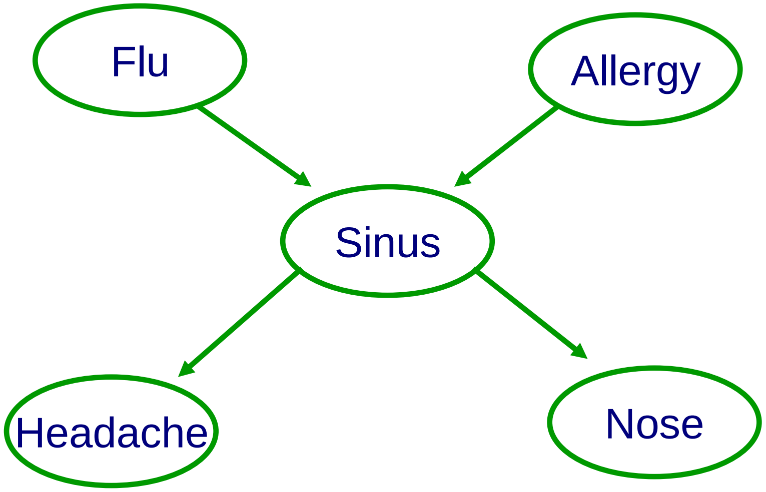
- Minimal I-map
  - every *P* has one, but usually many
- Perfect map
  - better choice for BN structure
  - not every *P* has one
  - can find one (if it exists) by considering I-equivalence
  - Two structures are I-equivalent if they have same skeleton and immoralities

Announcements

- I'll lead a special discussion session:
  - Today 2-3pm in NSH 1507
    - talk about homework, especially programming question

Review

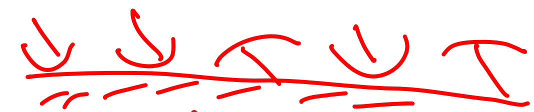
- Bayesian Networks
  - Compact representation for probability distributions
  - Exponential reduction in number of parameters
  - Exploits independencies



- Next – Learn BNs
  - parameters
  - structure

Thumbtack – Binomial Distribution

- P(Heads) = θ, P(Tails) = 1-θ



$\theta = \frac{3}{5} = .6$

- Flips are i.i.d.:
  - Independent events
  - Identically distributed according to Binomial distribution
- Sequence D of α_H Heads and α_T Tails

$$P(\mathcal{D} \mid \theta) = \theta^{\alpha_H}(1-\theta)^{\alpha_T}$$

Maximum Likelihood Estimation

- **Data:** Observed set *D* of α_H Heads and α_T Tails
- **Hypothesis:** space Binomial distributions
- Learning θ is an optimization problem
  - What's the objective function?

- MLE: Choose θ that maximizes the probability of observed data:

$$\begin{aligned} \hat{\theta} &= \arg\max_{\theta} \; P(\mathcal{D} \mid \theta) \\ &= \arg\max_{\theta} \; \ln P(\mathcal{D} \mid \theta) \end{aligned}$$

Your first learning algorithm

$$\hat{\theta} = \arg\max_{\theta} \ln P(\mathcal{D} \mid \theta)$$
$$= \arg\max_{\theta} \ln \theta^{\alpha_H}(1-\theta)^{\alpha_T}$$

$\ln a^b = b \ln a$

$\ln a \cdot b = \ln a + \ln b$

$\frac{\partial}{\partial\theta} \ln\theta = \frac{1}{\theta}$

$\frac{\partial}{\partial\theta} \ln(1-\theta) = \frac{-1}{1-\theta}$

- Set derivative to zero: $\boxed{\frac{d}{d\theta} \ln P(\mathcal{D} \mid \theta) = 0}$

$$\frac{\partial}{\partial\theta} \ln\left[\theta^{\alpha_H}(1-\theta)^{\alpha_T}\right]$$

$$= \frac{\partial}{\partial\theta}\left[\alpha_H \ln\theta + \alpha_T \ln(1-\theta)\right] = \frac{\partial}{\partial\theta}\alpha_H \ln\theta + \frac{\partial}{\partial\theta}\alpha_T \ln(1-\theta)$$

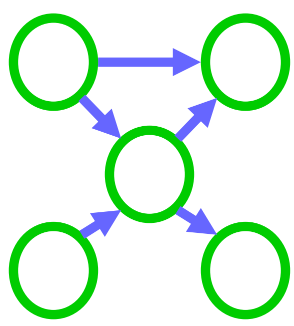
$$= \frac{\alpha_H}{\theta} - \frac{\alpha_T}{1-\theta} = 0 \Rightarrow \underline{\theta = \frac{\alpha_H}{\alpha_H + \alpha_T}}$$

Learning Bayes nets

	Known structure	Unknown structure
Fully observable data	easy	NP-hard but not so bad
Missing data	hard (things like EM)	very hard but → in a few weeks

↳ $\langle x_1 = t, x_2 = ?, x_3 = f \rangle$
⋮

Data
$\mathbf{x}^{(1)}$
…
$\mathbf{x}^{(m)}$



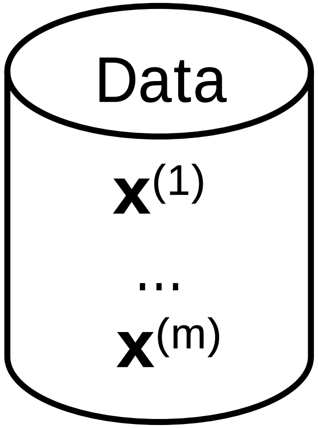
structure

+

CPTs – $P(X_i \mid \mathbf{Pa}_{X_i})$

parameters

Learning the CPTs

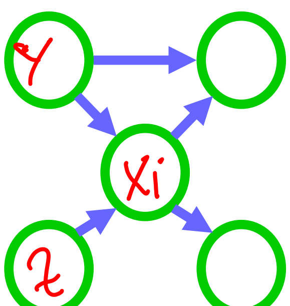


For each discrete variable X_i

$$P(X_i \mid Pa_{X_i}) = P(X_i \mid Y, Z)$$

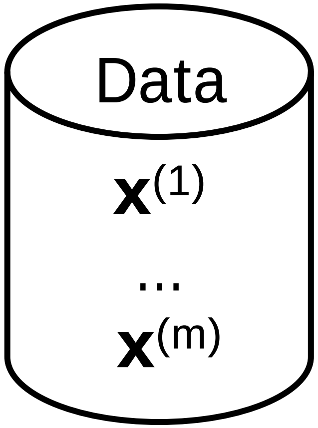
$$\overset{MLE}{\approx}$$

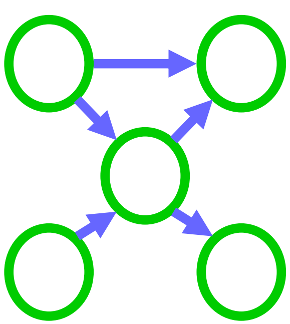
$$P(X_i = x_i \mid Y = y, Z = z) \overset{MLE}{\approx} \frac{\text{Count}(X_i = x_i, Y = y, Z = z)}{\text{Count}(Y = y, Z = z)}$$



MLE: $$P(X_i = x_i \mid X_j = x_j) = \frac{\text{Count}(X_i = x_i, X_j = x_j)}{\text{Count}(X_j = x_j)}$$

Learning the CPTs



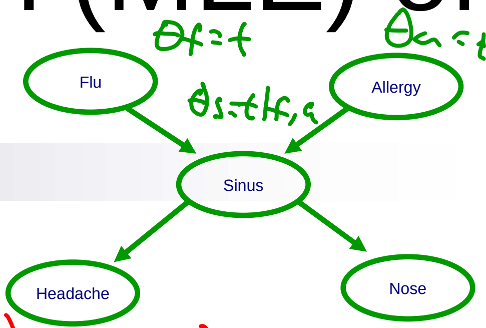


For each discrete variable X_i

MLE: $$P(X_i = x_i \mid X_j = x_j) = \frac{\text{Count}(X_i = x_i, X_j = x_j)}{\text{Count}(X_j = x_j)}$$

$$\boxed{\textbf{WHY??????????}}$$

Maximum likelihood estimation (MLE) of BN parameters – example



- Given structure, log likelihood of data:

$$\arg\max_{\theta} \log P(\mathcal{D} \mid \theta_{\mathcal{G}}, \mathcal{G}) \overset{iid}{=} \log \prod_j P(F=f^{(j)}, A=a^{(j)}, S=s^{(j)}, H=h^{(j)}, N=n^{(j)})$$

$$\overset{\text{structure}}{=} \log \prod_j P(F=f^{(j)})\, P(a^{(j)}) \cdot P(s^{(j)} \mid a^{(j)}, f^{(j)}) \cdot P(h^{(j)} \mid s^{(j)}) \cdot P(n^{(j)} \mid s^{(j)})$$

$$= \sum_j \Big[\log P(f^{(j)}) + \log P(a^{(j)}) + \log P(s^{(j)} \mid a^{(j)}, f^{(j)}) + \log P(h^{(j)} \mid s^{(j)}) + \log P(h^{(j)} \mid s^{(j)})\Big]$$

$$\arg\max_{\theta} = \Big[\sum_j \log P(f^{(j)}) + \sum_j \log P(a^{(j)}) + \sum_j \log P(s^{(j)} \mid a^{(j)}, f^{(j)}) \ldots\Big]$$

$$= \Big[\arg\max_{\theta_{f=t}} \sum_j \log P(f^{(j)} \mid \theta_{f=t})\Big] + \ldots \arg\max_{\theta_{a=t}} \ldots + \arg\max_{\theta_{s=t|f,a}} \ldots$$

Maximum likelihood estimation (MLE) of BN parameters – General case

- Data: $\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(m)}$
- Restriction: $\mathbf{x}^{(j)}[\mathbf{Pa}_{X_i}] \to$ assignment to $\mathbf{Pa}_{X_i}$ in $\mathbf{x}^{(j)}$
- Given structure, log likelihood of data:

$$\log P(\mathcal{D} \mid \theta_{\mathcal{G}}, \mathcal{G}) = \log \prod_j \prod_i P\left(X_i = x_i^{(j)} \mid Pa_{X_i} = x^{(j)}[Pa_{X_i}]\right)$$

Taking derivatives of MLE of BN parameters – General case

$$\log P(\mathcal{D} \mid \theta_{\mathcal{G}}, \mathcal{G}) = \sum_{j=1}^{m} \sum_{i=1}^{n} \log P\left(X_i = x_i^{(j)} \mid \mathbf{Pa}_{X_i} = \mathbf{x}^{(j)}\left[\mathbf{Pa}_{X_i}\right]\right)$$

General MLE for a CPT

- Take a CPT: P(X|**U**)
- Log likelihood term for this CPT

- Parameter $\theta_{X=x|\mathbf{U}=\mathbf{u}}$:

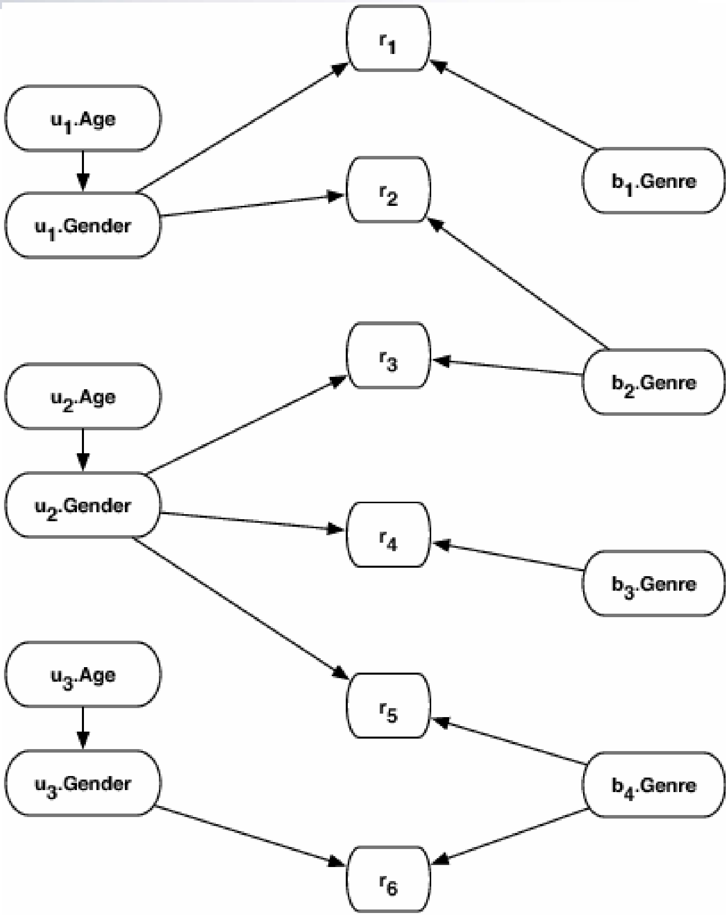
MLE: $$P(X = x \mid \mathbf{U} = \mathbf{u}) = \theta_{X=x|\mathbf{U}=\mathbf{u}} = \frac{\text{Count}(X = x, \mathbf{U} = \mathbf{u})}{\text{Count}(\mathbf{U} = \mathbf{u})}$$

Parameter sharing
(basics now, more later in the semester)

- Suppose we want to model customers' rating for books
- You know:
  - features of customers, e.g., age, gender, income,…
  - features of books, e.g., genre, awards, # of pages, has pictures,…
  - ratings: each user rates a few books
- A simple BN:

Using recommender system

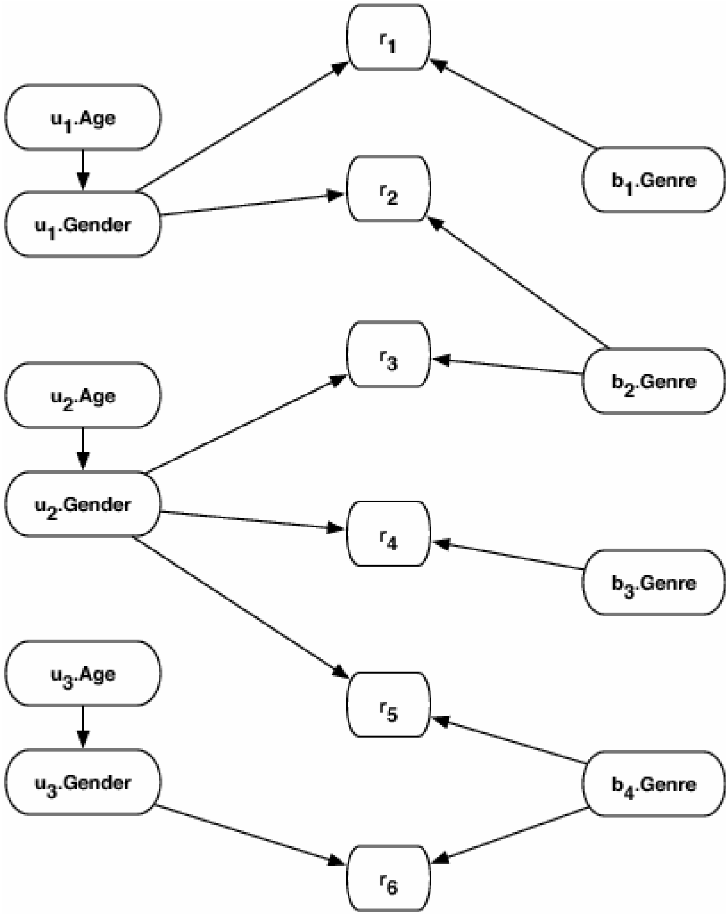
- Answer probabilistic question:



Learning parameters of recommender system BN

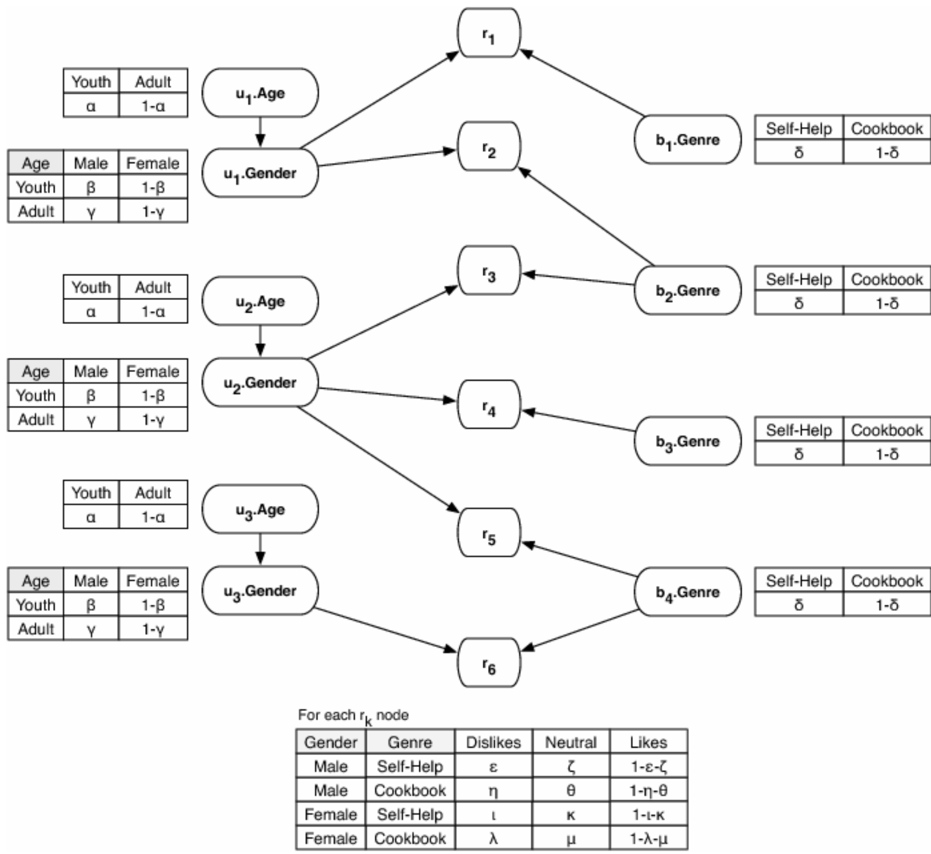
- How many parameters do I have to learn?

- How many samples do I have?



Parameter sharing for recommender system BN

- Use same parameters in many CPTs
- How many parameters do I have to learn?
- How many samples do I have?

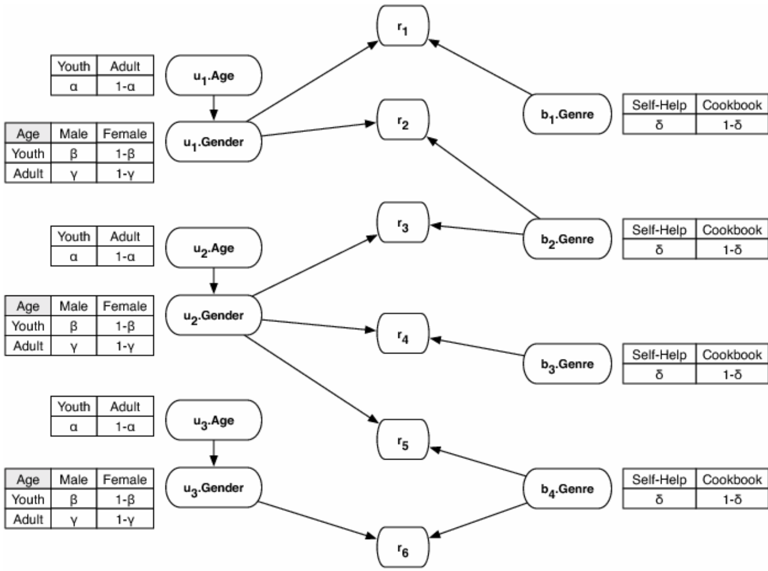


MLE with simple parameter sharing

- Estimating α:

- Estimating β:

- Estimating ε:



For each r_k node

Gender	Genre	Dislikes	Neutral	Likes
Male	Self-Help	ε	ζ	$1-\varepsilon-\zeta$
Male	Cookbook	η	θ	$1-\eta-\theta$
Female	Self-Help	ι	κ	$1-\iota-\kappa$
Female	Cookbook	λ	μ	$1-\lambda-\mu$

What you need to know about learning BNs thus far

- Maximum likelihood estimation
  - decomposition of score
  - computing CPTs
- Simple parameter sharing
  - why share parameters?
  - computing MLE for shared parameters