Machine Learning 10-601

Tom M. Mitchell
Machine Learning Department
Carnegie Mellon University

February 23, 2015

Today:

- Graphical models
- Bayes Nets:
  - Representing distributions
  - Conditional independencies
  - Simple inference
  - Simple learning

Readings:

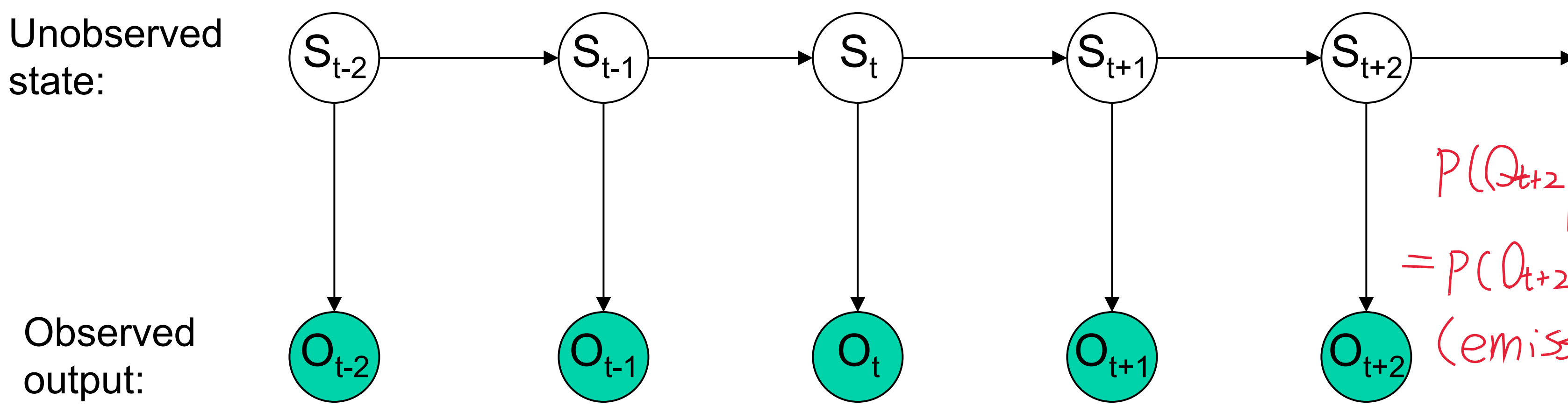
- Bishop chapter 8, through 8.2
- Mitchell chapter 6

Dynamic BN time-series data.

Bayes Network for a Hidden Markov Model (HMM)

Implies the future is conditionally independent of the past, given the present

$S_{t+1} \perp\!\!\!\perp \{S_{t-1}, S_{t-2}, \dots,$



$P(O_{t+2} \mid S_{t+2}, S_{t+1}, O_{t+1}, \dots)$

$= P(O_{t+2} \mid S_{t+2})$

(emission prob.)

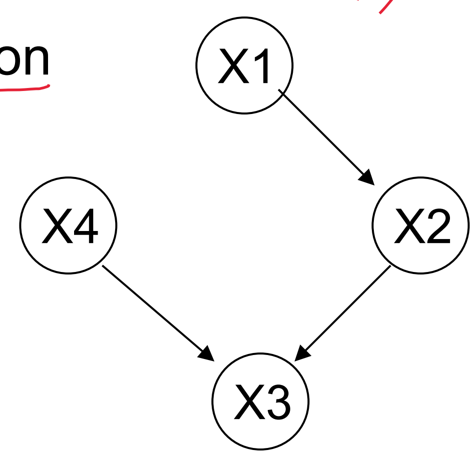
$$P(S_{t-2}, O_{t-2}, S_{t-1}, \ldots, O_{t+2}) =$$

$$P(S_1, O_1, S_2, O_2, \ldots, S_T, O_T) = P(S_1)\, P(O_1 \mid S_1) \prod_{t=2}^{T} P(S_t \mid S_{t-1})\, P(O_t \mid S_t)$$

Conditional Independence, Revisited

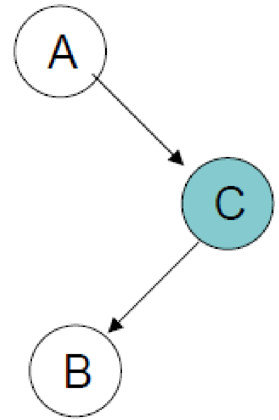
- We said:
  - Each node is conditionally independent of its non-descendents, given its immediate parents.

- Does this rule give us all of the conditional independence relations implied by the Bayes network?
  - No!
  - E.g., X1 and X4 are conditionally indep given {X2, X3} $X_1 \perp\!\!\!\perp X_4 \mid \{X_2, X_3\}$
  - But X1 and X4 not conditionally indep given X3 $X_1 \not\perp\!\!\!\perp X_4 \mid X_3$
  - For this, we need to understand D-separation



- Head-Tail (H-T): $\circ \rightarrow \circ \rightarrow \circ$
- Tail-Tail (T-T): $\circ \leftarrow \circ \rightarrow \circ$
- Head-Head (H-H): $\circ \rightarrow \circ \leftarrow \circ$

Easy Network 1: Head to Tail



prove A cond indep of B given C?

ie., p(a,b|c) = p(a|c) p(b|c)

$$p(a,b|c) = \frac{p(a,b,c)}{p(c)}$$

$$= \frac{p(a)\,p(c|a)\,p(b|c)}{p(c)}$$

$$= \frac{\cancel{p(c)}\,p(a|c)\,p(b|c)}{\cancel{p(c)}}$$

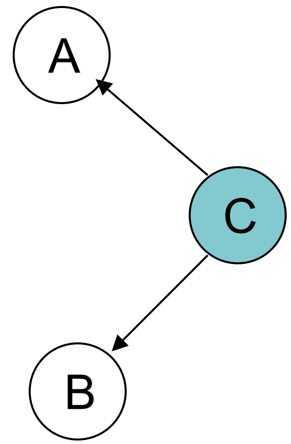
$$p(a,b) \neq p(a)\,p(b)$$

$$P(A,B,C) = p(A)\,p(C|A)\,p(B|C)$$

$$\text{H-T} \Rightarrow \begin{cases} A \perp\!\!\!\perp B \mid C \\ A \not\perp\!\!\!\perp B \end{cases}$$

let's use p(a,b) as shorthand for p(A=a, B=b)

Easy Network 2: Tail to Tail



prove A cond indep of B given C? ie., p(a,b|c) = p(a|c) p(b|c)

$$p(a,b|c) = \frac{p(a,b,c)}{p(c)}$$

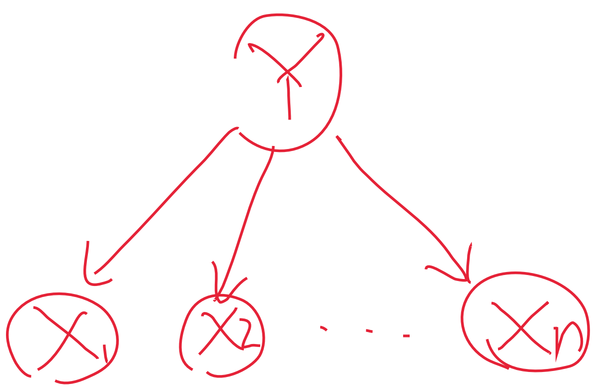
$$= \frac{\cancel{p(c)}\, p(a|c)\, p(b|c)}{\cancel{p(c)}}$$

$$P(A,B,C) = P(C)\,P(A|C)\,P(B|C)$$

$$\text{T-T} \Rightarrow \begin{cases} A \perp\!\!\!\perp B \mid C \\ A \not\perp\!\!\!\perp B \end{cases}$$

$$p(a,b) \neq p(a)\,p(b)$$

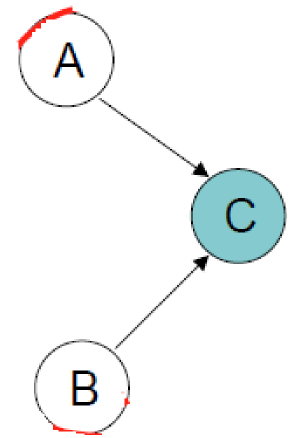
NB



let's use p(a,b) as shorthand for p(A=a, B=b)

Easy Network 3: Head to Head

prove A cond indep of B given C?   ie., p(a,b|c) = p(a|c) p(b|c)



$$p(a,b) = \sum_c p(a,b,c)$$

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

$$= p(a)\, p(b) \underbrace{\sum_c p(c|a)}_{=1}$$

$$= p(a)\, p(b)$$

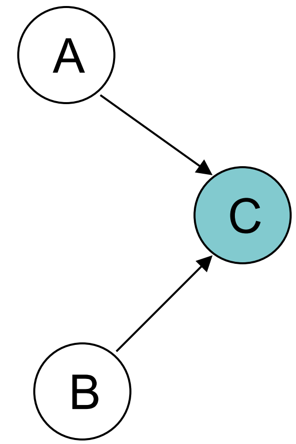
$$p(a,b|c) \neq p(a|c)\, p(b|c)$$

H-H:

$$\begin{cases} A \perp\!\!\!\perp B \\ A \not\perp\!\!\!\perp B \mid C \end{cases}$$

let's use p(a,b) as shorthand for p(A=a, B=b)

Easy Network 3: Head to Head



prove A cond indep of B given C?   NO!

Summary:

- p(a,b)=p(a)p(b)
- p(a,b|c) NotEqual p(a|c)p(b|c)

Explaining away.

e.g.,

[illegible]:

$$\begin{cases} A \perp\!\!\!\perp B \\ A \not\perp\!\!\!\perp B \mid C \end{cases}$$

- A=earthquake
- B=breakIn
- C=motionAlarm

X and Y are conditionally independent given Z, **<u>if and only if</u>** X and Y are D-separated by Z.

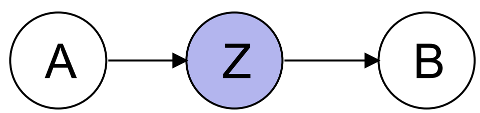
[Bishop, 8.2.2]

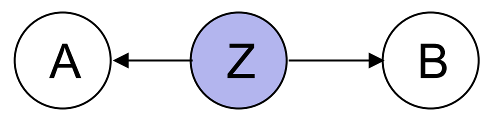
Suppose we have three sets of random variables: X, Y and Z

X and Y are **<u>D-separated</u>** by Z (and therefore conditionally indep, given Z) iff every path from every variable in X to every variable in Y is **<u>blocked</u>**

$X \perp\!\!\!\perp Y \mid Z$

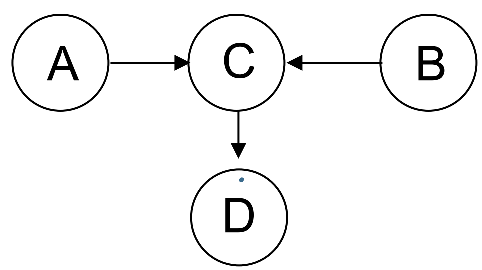
A path from variable X to variable Y is **blocked** if it includes a node in Z such that either





1. arrows on the path meet either head-to-tail or tail-to-tail at the node and this node is in Z

2. or, the arrows meet head-to-head at the node, and neither the node, nor any of its descendants, is in Z



X and Y are **<u>D-separated</u>** by Z (and therefore conditionally indep, given Z) iff every path from every variable in X to every variable in Y is **<u>blocked</u>**

A path from variable A to variable B is **blocked** if it includes a node such that either

1. arrows on the path meet either head-to-tail or tail-to-tail at the node and this node is in Z

2. or, the arrows meet head-to-head at the node, and neither the node, nor any of its descendants, is in Z

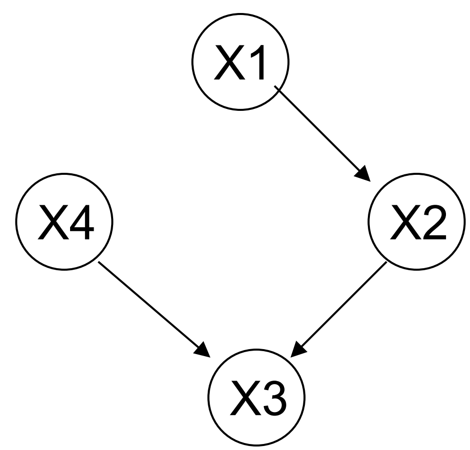
X Y Z

✓ X1 indep of X3 given X2? $X_1 \perp\!\!\!\perp X_3 \mid X_2$

✓ X3 indep of X1 given X2?

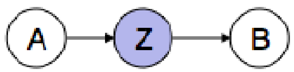
✓ X4 indep of X1 given X2? $X_1 \perp\!\!\!\perp X_4 \mid X_2$

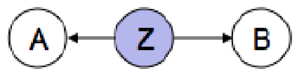
X Y Z

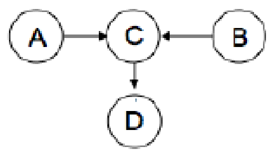


X and Y are <u>**D-separated**</u> by Z (and therefore conditionally indep, given Z) iff every path from any variable in X to any variable in Y is <u>**blocked**</u> by Z

A path from variable A to variable B is **blocked** by Z if it includes a node such that either

1.arrows on the path meet either head-to-tail or tail-to-tail at the node and this node is in Z 

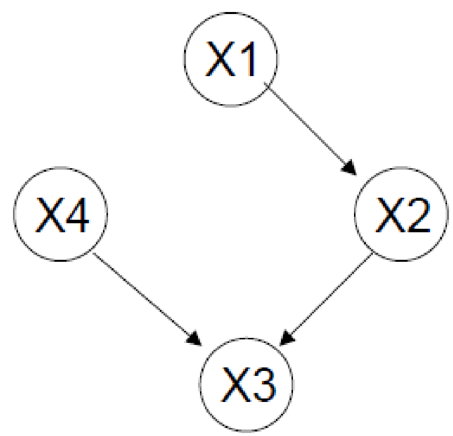


2.the arrows meet head-to-head at the node, and neither the node, nor any of its descendants, is in Z 

✗ X4 indep of X1 given X3?

✓ X4 indep of X1 given {X3, X2}?

✓ X4 indep of X1 given {}?



X and Y are **<u>D-separated</u>** by Z (and therefore conditionally indep, given Z) iff every path from any variable in X to any variable in Y is **<u>blocked</u>**

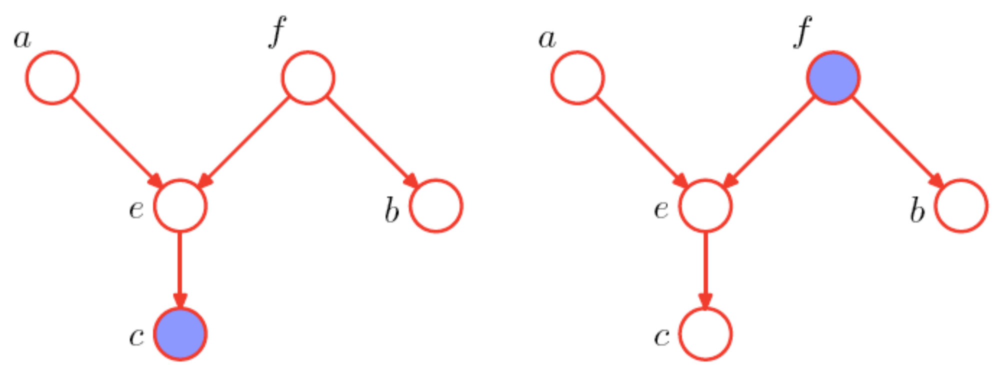
A path from variable A to variable B is **blocked** if it includes a node such that either

1. arrows on the path meet either head-to-tail or tail-to-tail at the node and this node is in Z

2. or, the arrows meet head-to-head at the node, and neither the node, nor any of its descendants, is in Z

a indep of b given c? ✗

a indep of b given f ? ✓



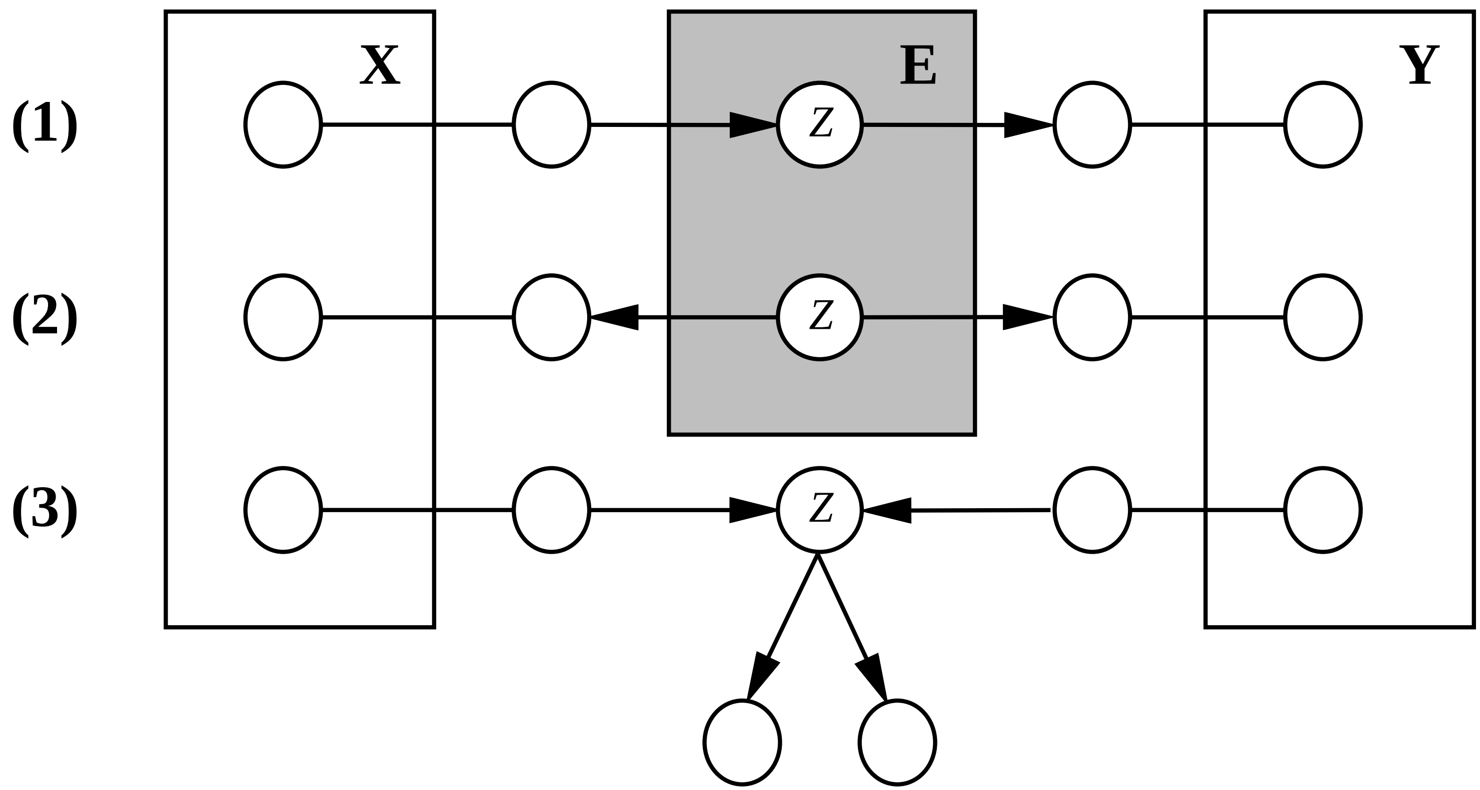
(a) (b)

D-separation

Q: When are nodes X independent of nodes Y given nodes E?

A: When every undirected path from a node in X to a node in Y is **d-separated** by E.

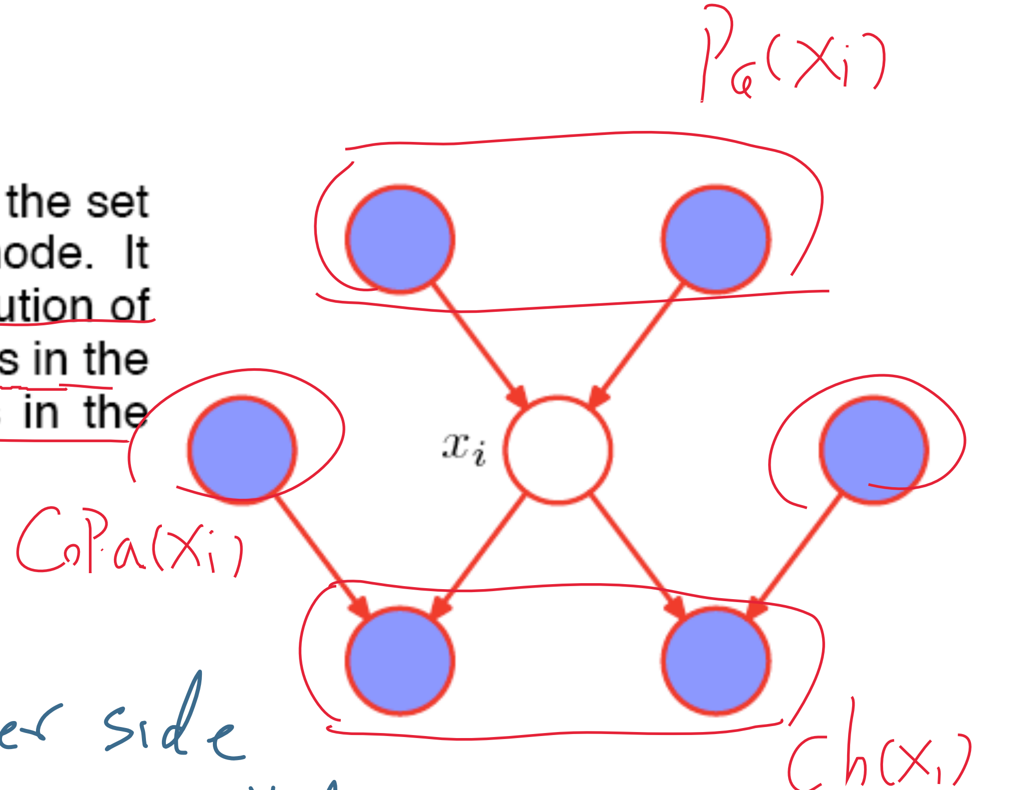
$X \perp\!\!\!\perp Y \mid E$



$MB(X_i) = Pa(X_i) \cup CoPa(X_i) \cup Ch(X_i)$

Markov Blanket

The Markov blanket of a node $\mathbf{x}_i$ comprises the set of parents, children and co-parents of the node. It has the property that the conditional distribution of $\mathbf{x}_i$, conditioned on all the remaining variables in the graph, is dependent only on the variables in the Markov blanket.



$P(X_i \mid X_{\{j \neq i\}})$

co-parent = other side of X_i's colliders

$= P(X_i \mid MB(X_i), \overline{MB}(X_i))$

$= P(X_i \mid MB(X_i)) \Rightarrow X_i \perp\!\!\!\perp \overline{MB}(X_i) \mid MB(X_i)$

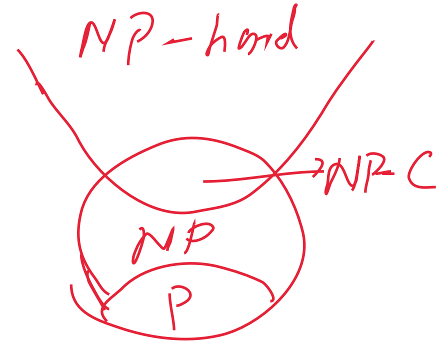
from [Bishop, 8.2]

What You Should Know

- Bayes nets are convenient representation for encoding dependencies / conditional independence
- BN = Graph plus parameters of CPD's
  - Defines joint distribution over variables
  - Can calculate everything else from that
  - Though inference may be intractable
- Reading conditional independence relations from the graph
  - Each node is cond indep of non-descendents, given only its parents
  - X and Y are conditionally independent given Z if Z D-separates every path connecting X to Y
  - Marginal independence : special case where Z={}

$O(\log n) < O(n) < O(n^2) < O(e^n) < O(n!)$, $n \to \infty$

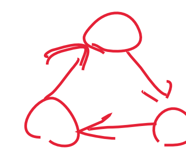
Inference in Bayes Nets



$(P \neq NP)$

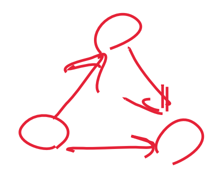
- In general, intractable (NP-~~complete~~ Hard)
- For certain cases, tractable
  - Assigning probability to fully observed set of variables
  - Or if just one variable unobserved
  - Or for singly connected graphs (ie., no undirected loops)
    - Belief propagation   max-sum

(DAG)



X

- Sometimes use Monte Carlo methods
  - Generate many samples according to the Bayes Net distribution, then count up the results



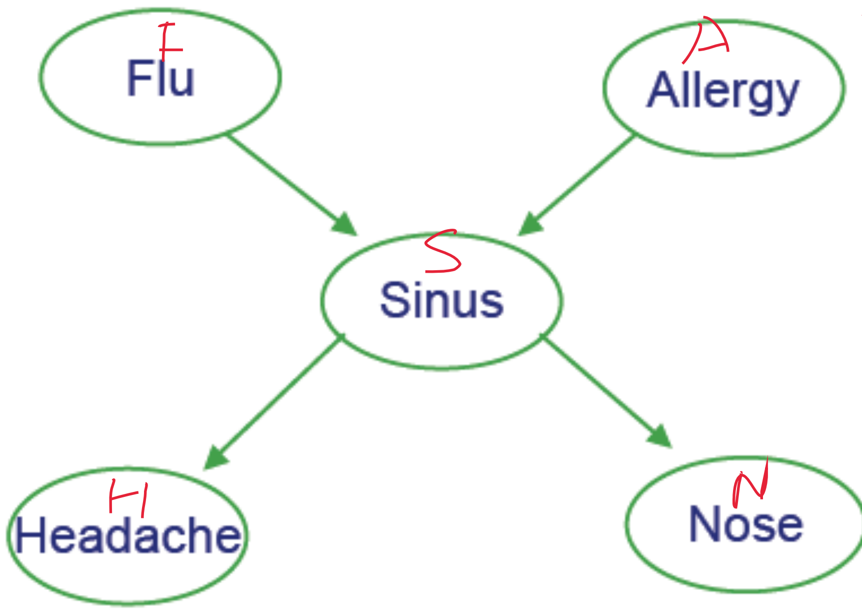
✓

- Variational methods for tractable approximate solutions

Example

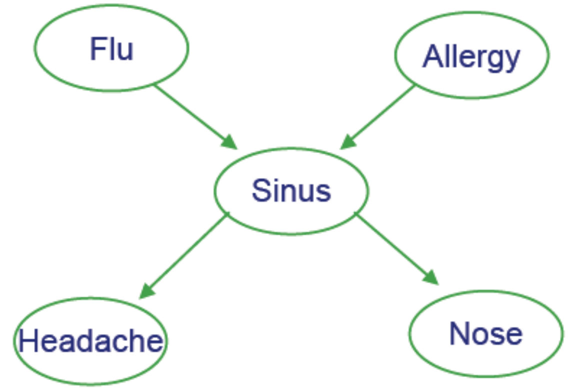
- Bird flu and Allegies both cause Sinus problems
- Sinus problems cause Headaches and runny Nose

BN: Graph, CPD



P(A)	A=1	A=0
	θ_a	$1-\theta_a$

Prob. of joint assignment: easy



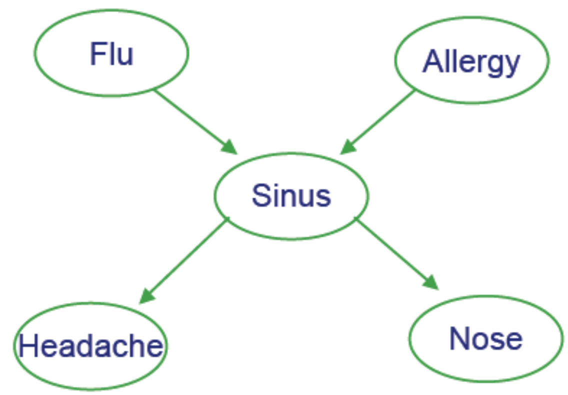
- Suppose we are interested in joint assignment <F=f,A=a,S=s,H=h,N=n>

What is P(f,a,s,h,n)? (n-1) multiplication $O(n-1)$

$$p(f,a,s,h,n) = p(f)\cdot p(a)\cdot p(s|f,a)\, p(h|s)\, p(n|s) = \cdots$$

let's use p(a,b) as shorthand for p(A=a, B=b)

Prob. of marginals: not so easy



- How do we calculate P(N=n) ?

$$p(n) = \sum_{f,a,s,h} p(f,a,s,h,n)$$

$O(2^n n)$.

$$2^4 \left\{ \begin{array}{l} p(f=0)\, p(a=0)\, p(s=0 \mid a=0, f=0) \cdot p(h=0 \mid s=0)\, p(n=0 \mid s=0) \\ \vdots \\ p(f=1)\, p(a=1)\, p(s=1 \mid a=1, f=1)\, p(h=1 \mid s=1)\, p(n=1 \mid s=1) \end{array} \right.$$

summation.

$2^4 \to 2^{n-1}$

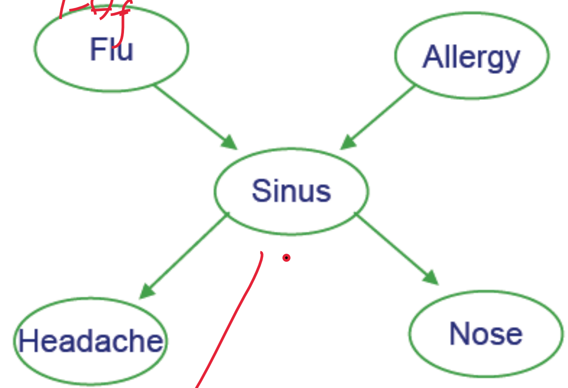
multiplication

$2^4 \cdot 4 \to 2^{n-1}(n-1)$

let's use p(a,b) as shorthand for p(A=a, B=b)

Generating a sample from joint distribution: easy

P(F)	F=1	F=0
	θ_f	$1-\theta_f$



How can we generate random samples drawn according to P(F,A,S,H,N)?

Hint: random sample of F according to P(F=1) = $\theta_{F=1}$: $= 0.7$

- draw a value of r uniformly from [0,1]
- if r<θ then output F=1, else F=0

0   0.7   1

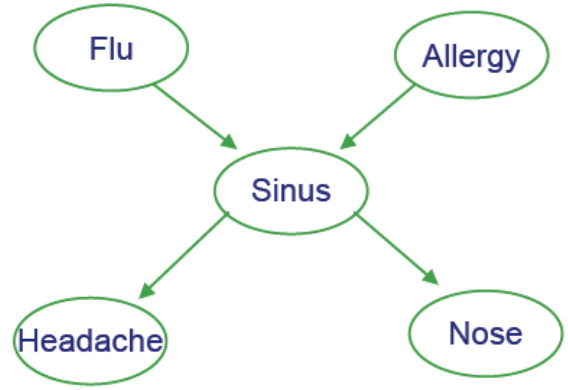
$D = \{x_j\}_{j=1}^{m}$

$x_j = (f_j, a_j, s_j, h_j, n_j)^T$

P(S|F,A)	S=1	S=0
F=0, A=0	θ_{00}	$1-\theta_{00}$
F=0, A=1	θ_{01}	$1-\theta_{01}$
F=1, A=0	θ_{10}	$1-\theta_{10}$
F=1, A=1	θ_{11}	$1-\theta_{11}$

let's use p(a,b) as shorthand for p(A=a, B=b)

Generating a sample from joint distribution: easy



How can we generate random samples drawn according to P(F,A,S,H,N)?

Hint: random sample of F according to $P(F=1) = \theta_{F=1}$:

- draw a value of r uniformly from [0,1]
- if $r<\theta$ then output F=1, else F=0

$\to P(N=n) = \theta^{N}(1-\theta)^{1-N}$

$\max_{\theta} L(\theta) = P(D|\theta)$

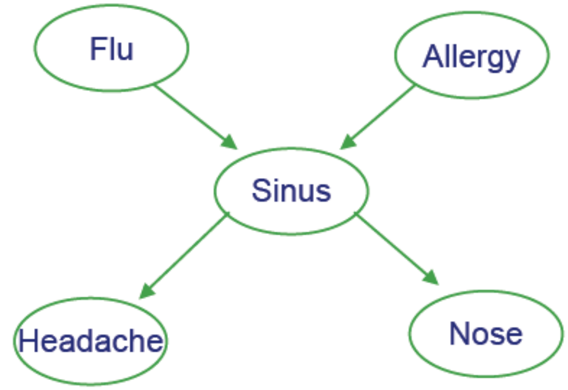
Solution: $D = \{(f_j, a_j, s_j, h_j, n_j)\}_{j=1}^{m}$

$\ell(\theta) = \sum_{j=1}^{m} \ln \theta^{\alpha_n}(1-\theta)^{m-\alpha_n}$

- draw a random value f for F, using its CPD
- then draw values for A, for S|A,F, for H|S, for N|S

$$\frac{\partial \ell(\theta)}{\partial \theta} = 0 \Rightarrow \theta_n = \frac{\alpha_n}{m} = \frac{|D_{N=1}|}{|D|}$$

Generating a sample from joint distribution: easy



Note we can estimate marginals

like P(N=n) by generating many samples

from joint distribution, then count the fraction of samples for which N=n

Similarly, for anything else we care about

MLE → P(F=1|H=1, N=0) $= \frac{|D_{F=1, H=1, N=0}|}{|D_{H=1, N=0}|}$

→ weak but general method for estimating <u>any</u> probability term…

Inference in Bayes Nets

- In general, intractable (NP-complete)
- For certain cases, tractable
  - Assigning probability to fully observed set of variables
  - Or if just one variable unobserved
  - Or for singly connected graphs (ie., no undirected loops)
    - Variable elimination
    - Belief propagation
- Often use Monte Carlo methods
  - e.g., Generate many samples according to the Bayes Net distribution, then count up the results
  - Gibbs sampling
- Variational methods for tractable approximate solutions

see Graphical Models course 10-708