Machine Learning 10-601 10-301

Tom M. Mitchell
Machine Learning Department
Carnegie Mellon University

April 14, 2021

Today:

- Graphical models 2
  - Probabilistic inference in Bayesian Networks

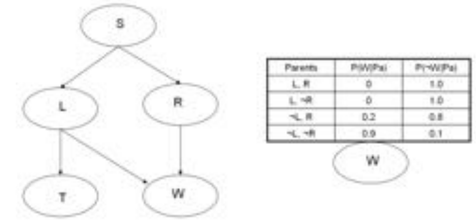
Readings:

- Bishop chapter 8.1 and 8.2

https://www.microsoft.com/en-us/research/wp-content/uploads/2016/05/Bishop-PRML-sample.pdf

Poll: Answer Question 1

Bayesian Networks Definition



A Bayes network represents the joint probability distribution over a collection of random variables

A Bayes network is a directed acyclic graph and a set of Conditional Probability Distributions (CPD's)

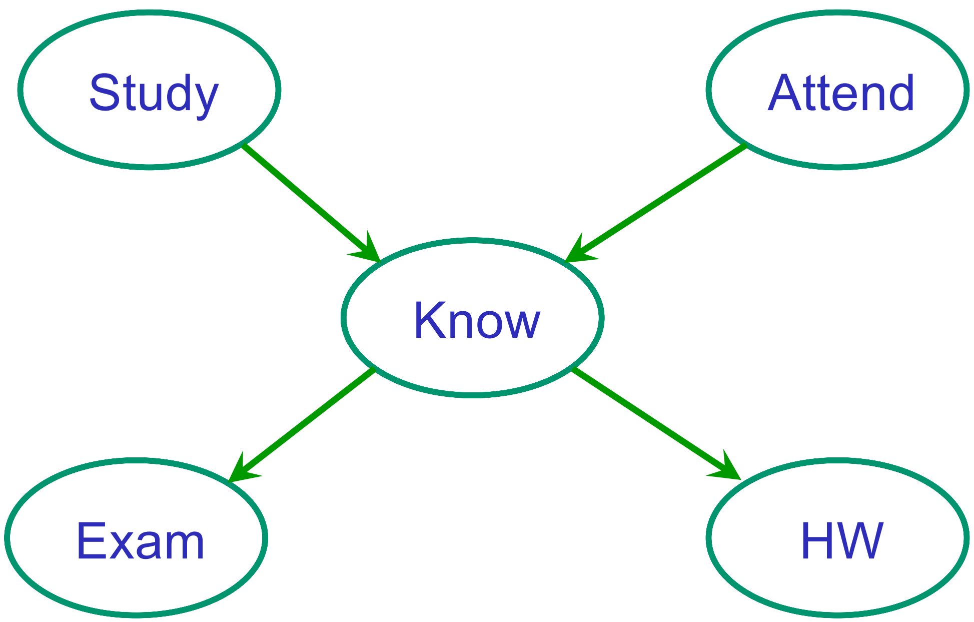
- Each node denotes a random variable
- Edges denote dependencies
- For each node X_i its CPD defines $P(X_i \mid Pa(X_i))$
- The joint distribution over all variables is defined to be

$$P(X_1 \ldots X_n) = \prod_i P(X_i | Pa(X_i))$$

Pa(X) = immediate parents of X in the graph

Example

- Attending class and Studying both cause you to Know the course material
- Knowing the course material determines whether you pass the Exam, and ace the HW



Parents <S, A>	P(K=1|S,A)	P(K=0|S,A)
<1, 1>	0.9	0.1
<1, 0>	0.8	0.2
<0, 1>	0.7	0.3
<0, 0>	0.1	0.9

Know

Inference in Bayes Nets

Given a joint distribution represented by a Bayes Net, how can we calculate arbitrary probabilities over subsets of variables?

- $P(X_3=1 \mid X_1=a, X_{17}=0)$
- $P(X_7=1)$
- …

Unfortunately, in general, intractable (NP-complete)

Fortunately, for certain types of graphs, tractable

Fortunately, we can sometimes *estimate* them tractably

Prob. of joint assignment: easy

Parents <S, A>	P(K=1|S,A)	P(K=0|S,A)
<1, 1>	0.9	0.1
<1, 0>	0.8	0.2
<0, 1>	0.7	0.3
<0, 0>	0.1	0.9

Know

Suppose we are interested in joint assignment <S=s,A=a,K=k,E=e,H=h>

What is P(s,a,k,e,h)?

we'll use p(a,b) as shorthand for p(A=a, B=b)

Prob. of joint assignment: easy

Parents <S, A>	P(K=1|S,A)	P(K=0|S,A)
<1, 1>	0.9	0.1
<1, 0>	0.8	0.2
<0, 1>	0.7	0.3
<0, 0>	0.1	0.9

Know

Suppose we are interested in joint assignment <S=s,A=a,K=k,E=e,H=h>

What is P(s,a,k,e,h)?

$$P(s,a,k,e,h) = P(s)P(a)P(k|s,a)P(e|k)P(h|k)$$

Efficient: O(n) for n variables.
i.e., look up n values, multiply them

we'll use p(a,b) as shorthand for p(A=a, B=b)

Marginal probabilities $P(X_i)$:

- How do we calculate P(H=1) ?

Parents <S, A>	P(K=1|S,A)	P(K=0|S,A)
<1, 1>	0.9	0.1
<1, 0>	0.8	0.2
<0, 1>	0.7	0.3
<0, 0>	0.1	0.9

Know

$$P(H=1) = \sum_{s=0}^{1}\sum_{a=0}^{1}\sum_{k=0}^{1}\sum_{e=0}^{1} P(S=s, A=a, K=k, E=e, H=1)$$

Marginal probabilities $P(X_i)$:

Parents <S, A>	P(K=1|S,A)	P(K=0|S,A)
<1, 1>	0.9	0.1
<1, 0>	0.8	0.2
<0, 1>	0.7	0.3
<0, 0>	0.1	0.9

Know

- How do we calculate P(H=1) ?

$$P(H=1) = \sum_{s=0}^{1}\sum_{a=0}^{1}\sum_{k=0}^{1}\sum_{e=0}^{1} P(S=s, A=a, K=k, E=e, H=1))$$

$$P(H=1) = \sum_{s=0}^{1}\sum_{a=0}^{1}\sum_{k=0}^{1}\sum_{e=0}^{1} P(S=s)P(A=a)P(K=k|S=s,A=a)P(E=e|K=k)P(H=1|K=k)$$

Inefficient: $O(n\ 2^{(n-1)})$ for n Boolean variables.

Only One Unobserved Variable:

Parents <S, A>	P(K=1|S,A)	P(K=0|S,A)
<1, 1>	0.9	0.1
<1, 0>	0.8	0.2
<0, 1>	0.7	0.3
<0, 0>	0.1	0.9

Know

How do we calculate P(K=1 | S=s, A=a, E=e, H=h) ?

$$P(K=1|S=s, A=a, E=e, H=h) = \frac{P(S=s, A=a, K=1, E=e, H=h)}{P(S=s, A=a, E=e, H=h)}$$

Only One Unobserved Variable:

Parents <S, A>	P(K=1|S,A)	P(K=0|S,A)
<1, 1>	0.9	0.1
<1, 0>	0.8	0.2
<0, 1>	0.7	0.3
<0, 0>	0.1	0.9

Know

How do we calculate P(K=1 | S=s, A=a, E=e, H=h) ?

$$P(K=1|S=s,A=a,E=e,H=h) = \frac{P(S=s,A=a,K=1,E=e,H=h)}{P(S=s,A=a,E=e,H=h)}$$

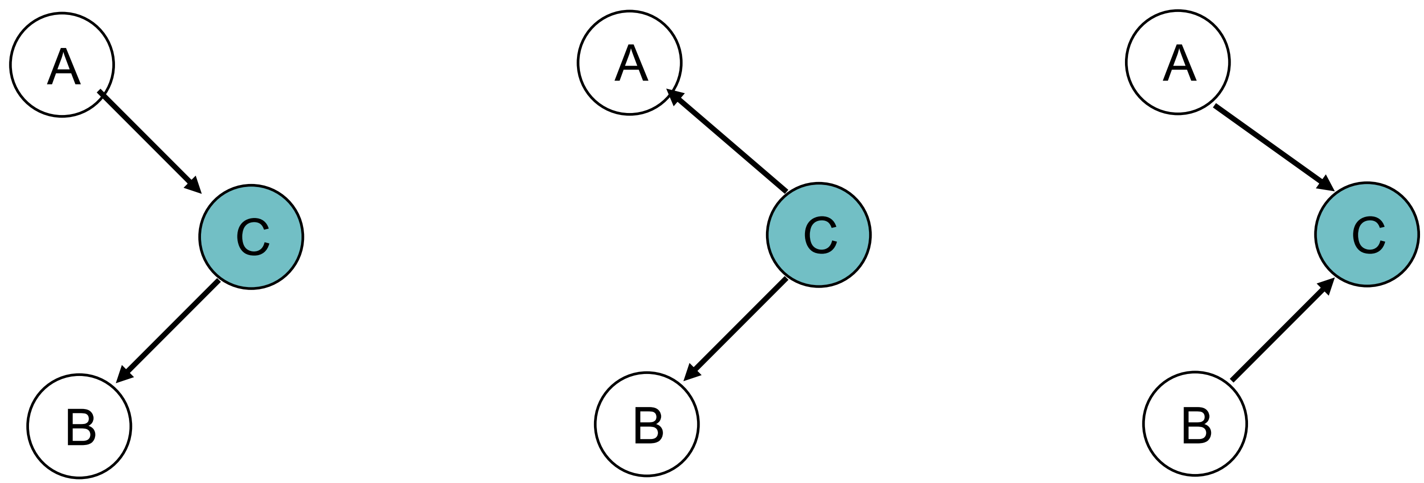
$$= \frac{P(S=s,A=a,K=1,E=e,H=h)}{P(S=s,A=a,K=1,E=e,H=h)+P(S=s,A=a,K=0,E=e,H=h)}$$

Efficient: O(2n) for n Boolean variables.

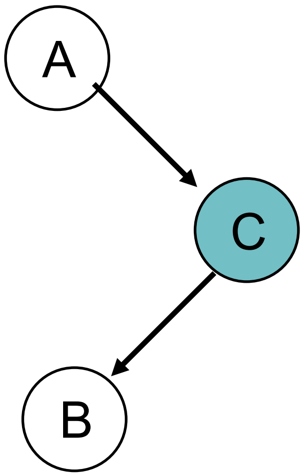
D-Separation

See recommended reading: Bishop Chapter 8.1-8.2

D-Separation Rule to determine Cond. Indep. is based on three simple subgraphs:



Simple Network 1: Head to Tail



prove A cond indep of B given C?

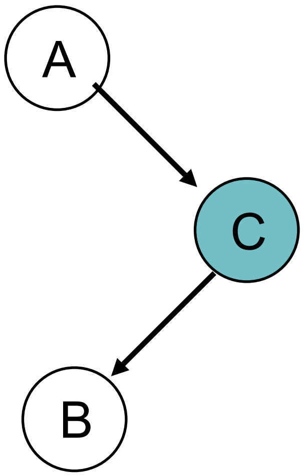
Ie., P(A=a,B=b|C=c) = P(A=a|C=c) P(B=b|C=c)

Which we'll write p(a,b|c) = p(a|c) p(b|c)

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

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

Simple Network 1: Head to Tail



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

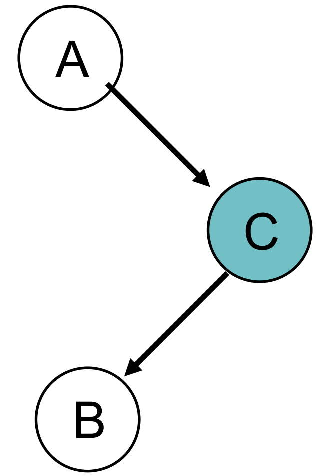
Which we'll write p(a,b|c) = p(a|c) p(b|c)

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

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

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

Simple Network 1: Head to Tail



prove A cond indep of B given C?

Ie., P(A=a,B=b|C=c) = P(A=a|C=c) P(B=b|C=c)

Which we'll write p(a,b|c) = p(a|c) p(b|c)

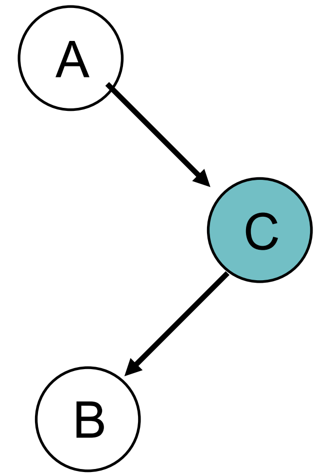
$$p(a,b|c) = \frac{p(a,b,c)}{p(c)} = \frac{p(a)p(c|a)p(b|c)}{p(c)}$$

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

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

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

Simple Network 1: Head to Tail



prove A cond indep of B given C?

Ie., P(A=a,B=b|C=c) = P(A=a|C=c) P(B=b|C=c)

Which we'll write p(a,b|c) = p(a|c) p(b|c)

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

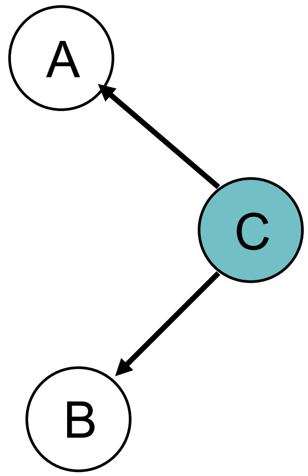
$$= \frac{p(a)p(c|a)}{p(c)}p(b|c)$$

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

$$= p(a|c)p(b|c)$$

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

Simple Network 2: Tail to Tail

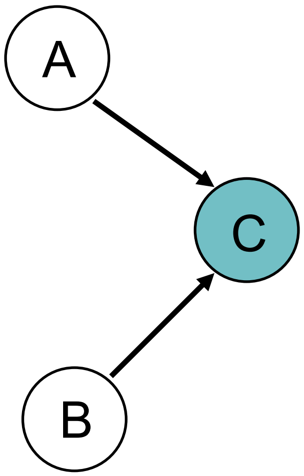


prove A cond indep of B given C? ie., $p(a,b|c) = p(a|c)\ p(b|c)$

This is also provable. [try it!!!]

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

Simple Network 3: Head to Head



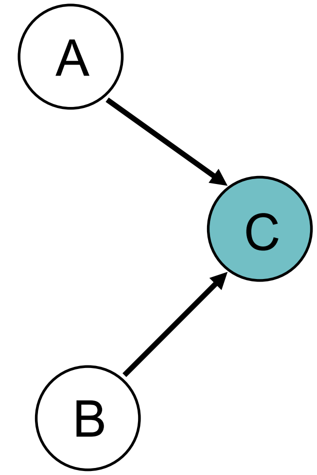
prove A cond indep of B given C? ie., $p(a,b|c) = p(a|c)\ p(b|c)$

This is NOT true!

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

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

Simple Network 3: Head to Head



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

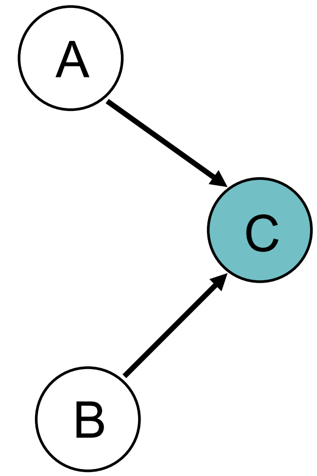
This is NOT true!

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

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

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

Simple Network 3: Head to Head



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

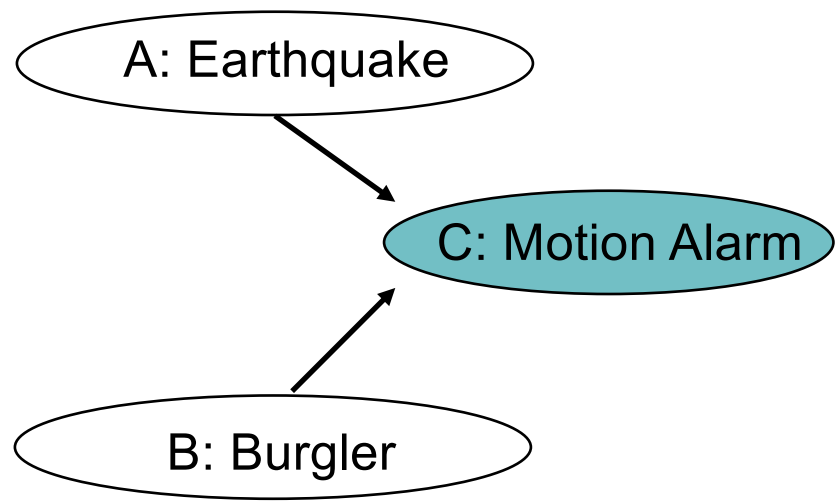
This is NOT true!

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

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

Intuition:

"Explaining away"



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

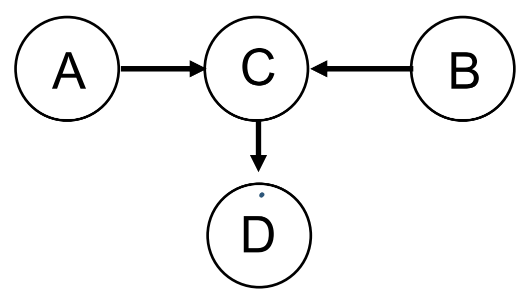
Suppose we have three <u>sets</u> of random variables: X, Y and Z

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>

A path from variable X to variable Y is <u>**blocked**</u> if it includes a node such that *either* (1) or (2) holds:

(1). arrows on the path meet either head-to-tail or tail-to-tail at a node in Z

(2). arrows on the path meet head-to-head at a node, and neither that node, nor any of its descendants, is in Z

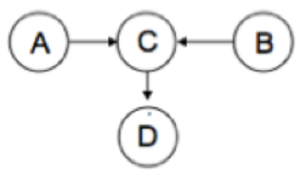


X and Y are **<u>D-separated</u>** by Z (and therefore conditionally indep, given Z) iff every path from any variable in X to any variable in Y is **<u>blocked</u>** by Z

A path from variable A to variable B is **blocked** by Z if it includes a node such that either (1) or (2)

(1) Arrows on the path meet either head-to-tail or tail-to-tail at a node from Z

(2) Arrows on the path meet head-to-head at a node, and neither that node, nor any of its descendants, is in Z

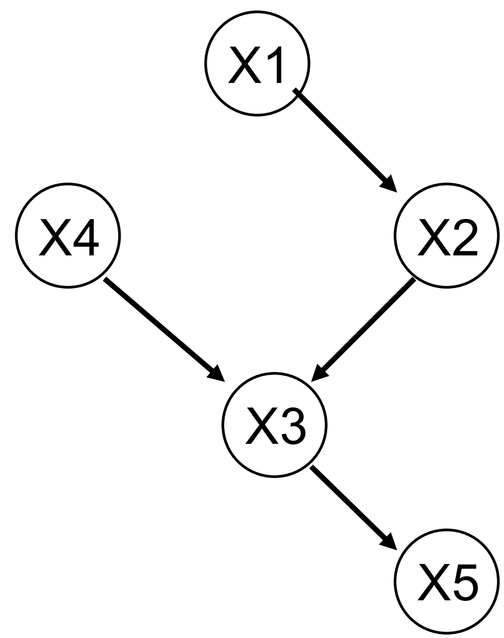


X4 indep of X1 given X2?

X4 indep of X1 given X3?

X4 indep of X1 given {}?

X4 indep of X1 given X5?

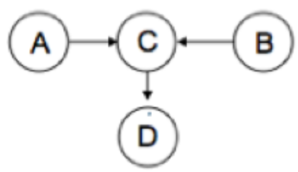


X and Y are **D-separated** by Z (and therefore conditionally indep, given Z) iff every path from any variable in X to any variable in Y is **blocked** by Z

A path from variable A to variable B is **blocked** by Z if it includes a node such that either (1) or (2)

(1) Arrows on the path meet either head-to-tail or tail-to-tail at a node from Z

(2) Arrows on the path meet head-to-head at a node, and neither the node, nor any of its descendants, is in Z

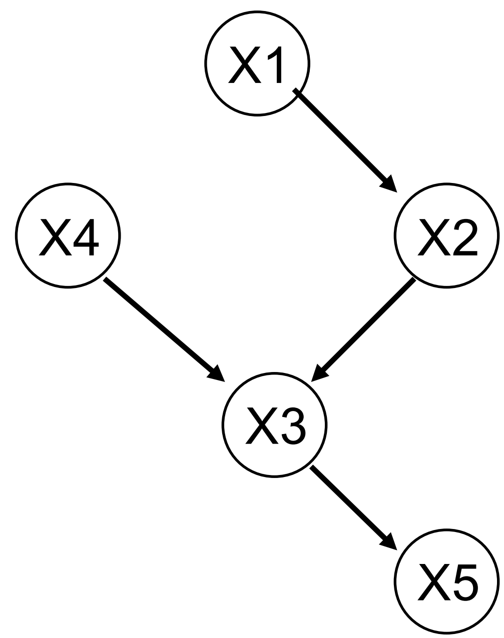


X4 indep of X1 given X2? YES (Condition 1)

X4 indep of X1 given X3? NO

X4 indep of X1 given {}? YES (Condition 2)

X4 indep of X1 given X5? NO

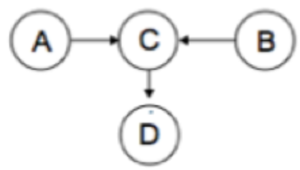


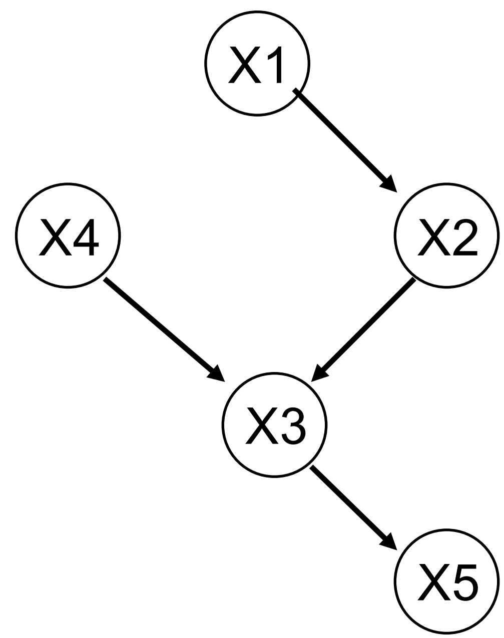
X and Y are **D-separated** by Z (and therefore conditionally indep, given Z) iff every path from any variable in X to any variable in Y is **blocked** by Z

A path from variable A to variable B is **blocked** by Z if it includes a node such that either (1) or (2)

(1) Arrows on the path meet either head-to-tail or tail-to-tail at a node from Z

(2) Arrows on the path meet head-to-head at a node, and neither the node, nor any of its descendants, is in Z





X1 indep of X3 given X2?

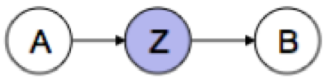
X3 indep of X1 given X2?

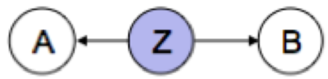
X4 indep of X2 given {}?

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>** by Z

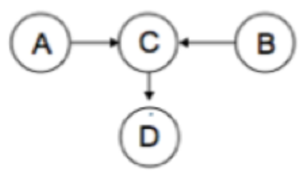
A path from variable A to variable B is **blocked** by Z if it includes a node such that either (1) or (2)

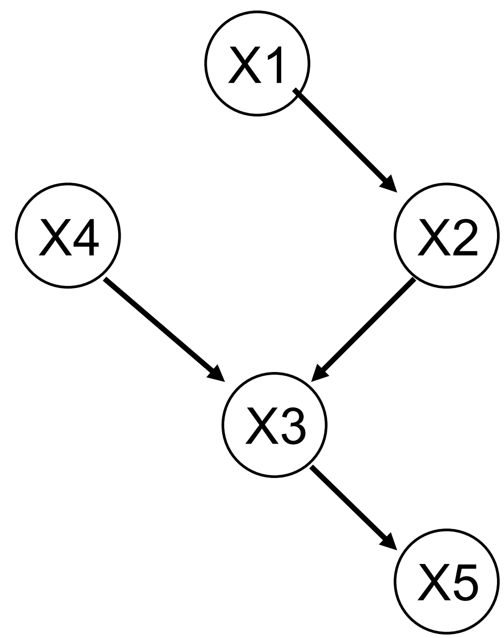
(1) Arrows on the path meet either head-to-tail or tail-to-tail at a node from Z





(2) Arrows on the path meet head-to-head at a node, and neither the node, nor any of its descendants, is in Z





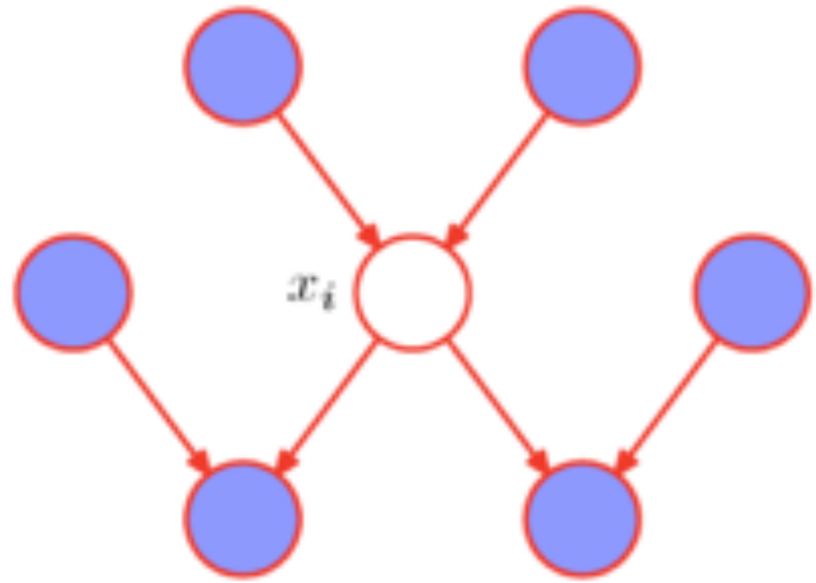
X1 indep of X3 given X2?   YES (1)

X3 indep of X1 given X2?   YES (1)

X4 indep of X2 given {}?   YES (2)

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.

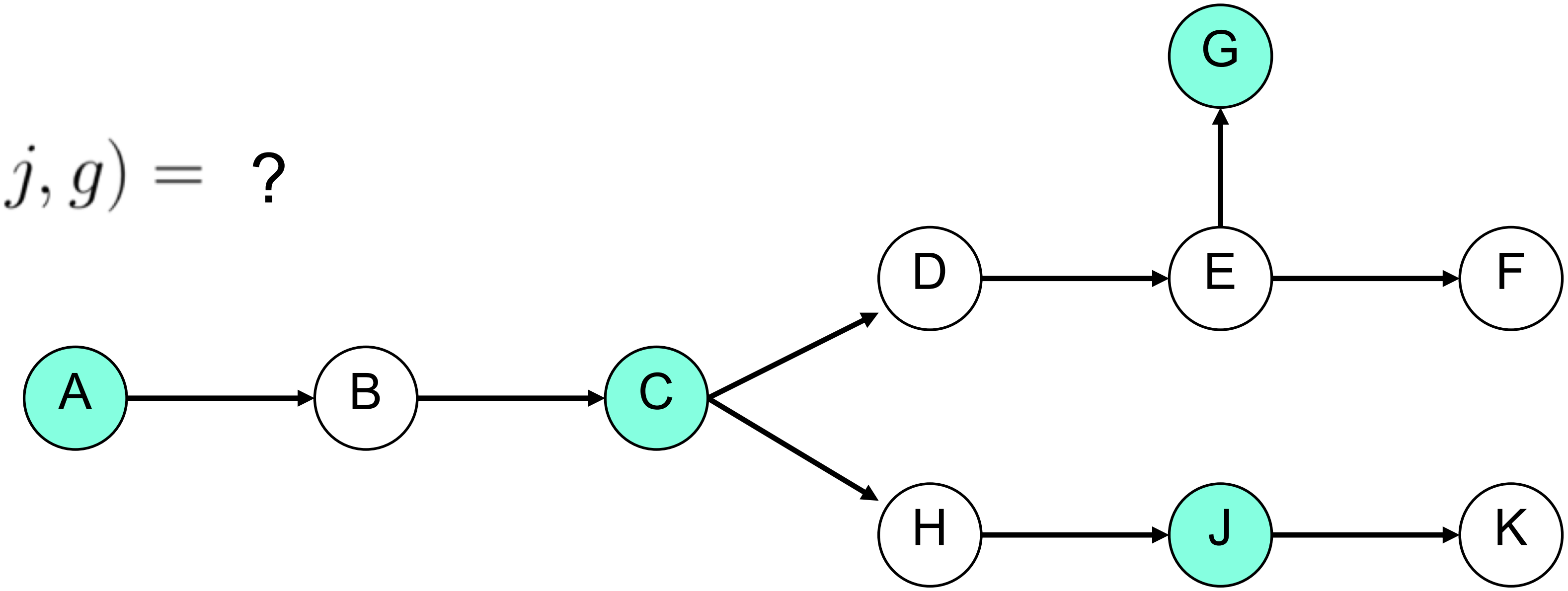


from [Bishop, 8.2]

→ Computational efficiency in many cases!

Why Markov Blanket is Useful for Inference

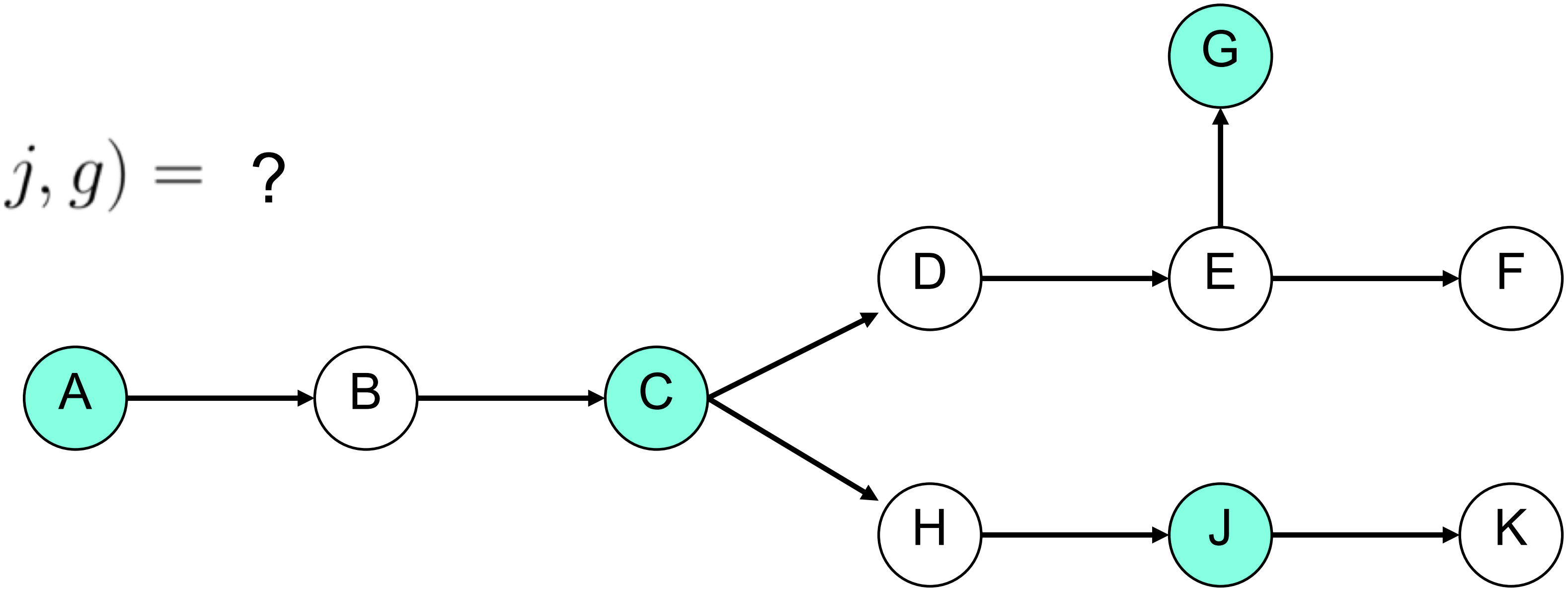
$$P(h|a, c, j, g) = \text{ ?}$$

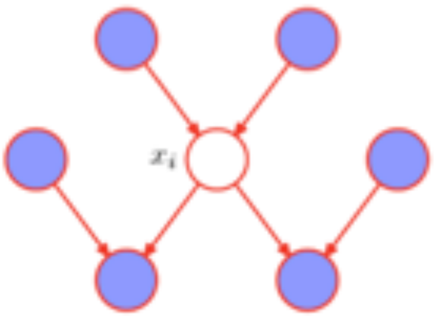


let's use shorthand P(a) to represent P(A=a)

Why Markov Blanket is Useful for Inference

$$P(h|a, c, j, g) = \ ?$$





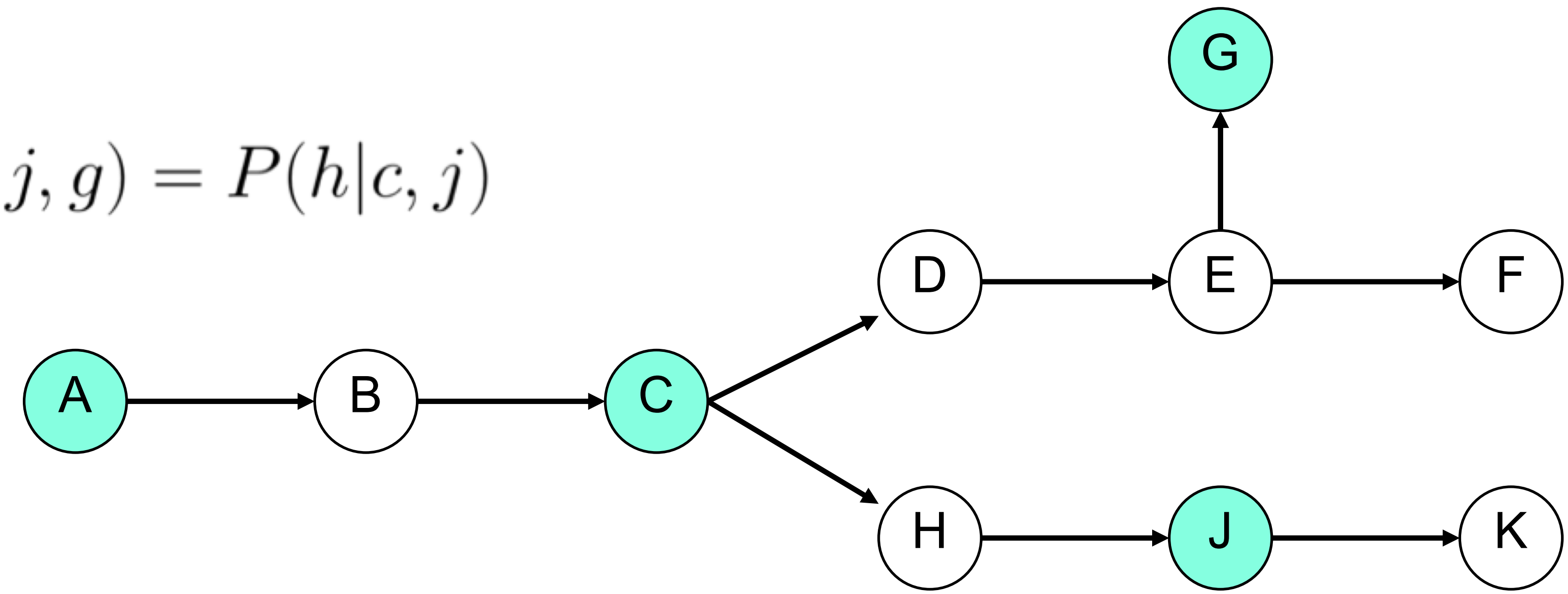
What is the Markov Blanket of H?

Poll: Answer Question 2

let's use shorthand P(a) to represent P(A=a)

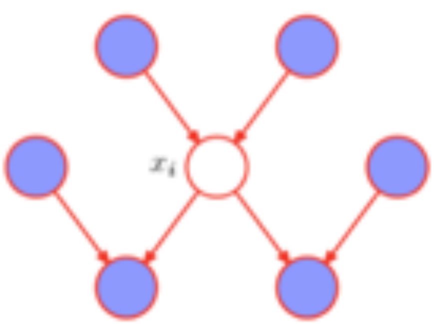
Why Markov Blanket is Useful for Inference

$$P(h|a,c,j,g) = P(h|c,j)$$



$$P(h|c,j) = \frac{P(h,c,j)}{P(c,j)} = \frac{P(c)P(h|c)P(j|h)}{P(c)P(h|c)P(j|h) + P(c)P(\neg h|c)P(j|\neg h)}$$

$$= \frac{P(h|c)P(j|h)}{P(h|c)P(j|h) + P(\neg h|c)P(j|\neg h)}$$



let's use shorthand P(a) to represent P(A=a)

Bayes Net Inference by generating data samples

So far: exact inference methods, sometimes expensive

Next: generate data by sampling joint distribution,
then estimate probabilities (MLE) from counts over
this data!

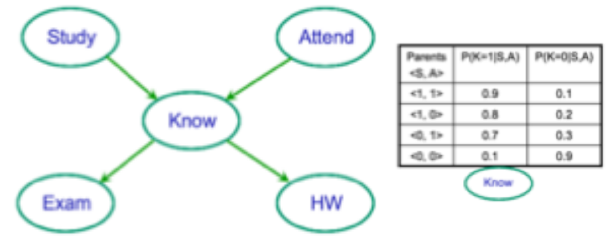
Generating a sample from joint distribution: easy

Parents <S, A>	P(K=1|S,A)	P(K=0|S,A)
<1, 1>	0.9	0.1
<1, 0>	0.8	0.2
<0, 1>	0.7	0.3
<0, 0>	0.1	0.9

Know

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

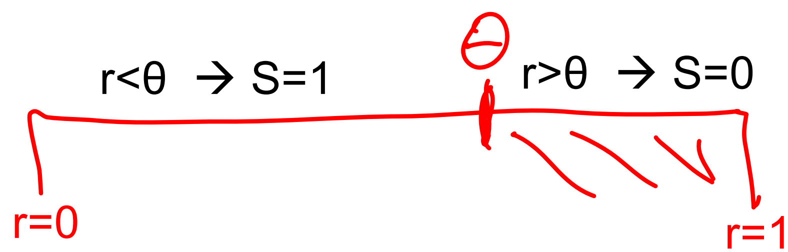
Generating a sample from joint distribution: easy



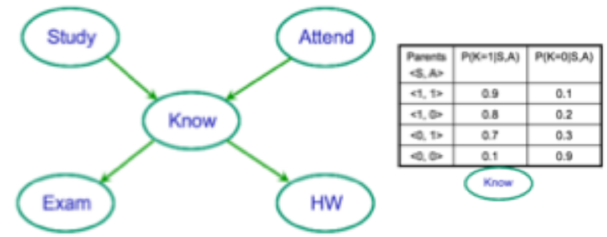
How can we generate random samples drawn according to P(S,A,K,E,H)?

To generate a random sample for roots of network (S or A):

1. let θ = P(S=1) # look up from CPD
2. r = random number drawn uniformly between 0 and 1
3. if $r<\theta$ then output S=1, else S=0



Generating a sample from joint distribution: easy

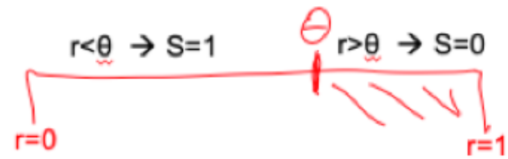


Parents <S,A>	P(K=1|S,A)	P(K=0|S,A)
<1, 1>	0.9	0.1
<1, 0>	0.8	0.2
<0, 1>	0.7	0.3
<0, 0>	0.1	0.9

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

To generate a random sample for roots of network (S or A):

1. let θ = P(S=1) # look up from CPD
2. r = random number drawn uniformly between 0 and 1
3. if r<θ then output S=1, else S=0



To generate a random sample for K, given S=s, A=a:

1. let θ = P(K=1|S=s,A=a) # look up from CPD
2. r = random number drawn uniformly between 0 and 1
3. if r<θ then output K=1, else K=0

Generating a sample from joint distribution: easy

We can estimate probabilities like P(E=e) by generating many samples from joint distribution, then counting the fraction of samples (MLE) for which E=e

Similarly, for <u>anything</u> else we care about, calculate its maximum likelihood estimate from many generated examples

e.g., P(A=1|E=1, H=0)

→ General method for <u>estimating</u> <u>any</u> probability term

→ Alternative to <u>exact</u> closed form solutions

→ Can be computationally expensive, depending on …

Generating a sample from joint distribution: easy

We can easily sample P(S,A,K,E,H)

We can use multiple samples to estimate P(S,A,K,E | H=1)

But if P(H=1) very small, most samples will have H=0 😫

Can we directly sample P(S,A,K,E | H=1), forcing H to be 1? 🤔

Gibbs Sampling:

Goal: Directly sample conditional distributions

$P(X_1,\ldots,X_n \mid X_{n+1}, \ldots, X_m)$

Approach:

- start with the fixed observed $X_{n+1}, \ldots, X_m$
  plus arbitrary initial values for the unobserved $X_1^{(0)},\ldots,X_n^{(0)}$
- Iterate: for sample s=0 to a big number:

$$X_1^{s+1} \sim P(X_1|X_2^s, X_3^s \ldots X_n^s, X_{n+1}, \ldots X_m)$$
$$X_2^{s+1} \sim P(X_2|X_1^{s+1}, X_3^s \ldots X_n^s, X_{n+1}, \ldots X_m)$$
$$\ldots$$
$$X_n^{s+1} \sim P(X_n|X_1^{s+1}, X_2^{s+1}, \ldots X_{n-1}^{s+1}, X_{n+1}, \ldots X_m)$$

Eventually (after burn-in), each sample will constitute a sample of the true $P(X_1,\ldots,X_n \mid X_{n+1}, \ldots, X_m)$

* but often use every 100th sample, since iters not independent

X ~ P(X) means X is drawn randomly from P(X)

Gibbs Sampling:

Goal: Directly sample conditional distributions

$P(X_1,\ldots,X_n \mid X_{n+1}, \ldots, X_m)$

Approach:

- start with the fixed observed $X_{n+1}, \ldots, X_m$
  plus arbitrary initial values for the unobserved $X_1^{(0)},\ldots,X_n^{(0)}$
- Iterate: for sample s=0 to a big number:

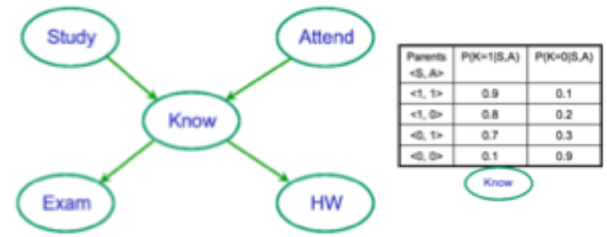
$$X_1^{s+1} \sim P(X_1|X_2^s, X_3^s \ldots X_n^s, X_{n+1}, \ldots X_m)$$
$$X_2^{s+1} \sim P(X_2|X_1^{s+1}, X_3^s \ldots X_n^s, X_{n+1}, \ldots X_m)$$
$$\ldots$$
$$X_n^{s+1} \sim P(X_n|X_1^{s+1}, X$$

By the way…
How would we generate this sample?

Eventually (after burn-in),
$P(X_1,\ldots,X_n \mid X_{n+1}, \ldots, X_m)$

* but often use every 100th sample, since iters not independent

Only One Unobserved Variable:



How do we calculate P(K=1 | S=s, A=a, E=e, H=h) ?

$$P(K=1|S=s,A=a,E=e,H=h) = \frac{P(S=s,A=a,K=1,E=e,H=h)}{P(S=s,A=a,E=e,H=h)}$$

$$= \frac{P(S=s,A=a,K=1,E=e,H=h)}{P(S=s,A=a,K=1,E=e,H=h)+P(S=s,A=a,K=0,E=e,H=h)}$$

Efficient: O(2n) for n Boolean variables.

Gibbs Sampling:

Goal: Directly sample conditional distributions

$P(X_1,\ldots,X_n \mid X_{n+1}, \ldots, X_m)$

Approach:

- start with the fixed observed $X_{n+1}, \ldots, X_m$
  plus arbitrary initial values for the unobserved $X_1^{(0)},\ldots,X_n^{(0)}$
- iterate for s=0 to a big number:

$$
\begin{aligned}
X_1^{s+1} &\sim P(X_1|X_2^s, X_3^s \ldots X_n^s, X_{n+1}, \ldots X_m) \\
X_2^{s+1} &\sim P(X_2|X_1^{s+1}, X_3^s \ldots X_n^s, X_{n+1}, \ldots X_m) \\
&\cdots \\
X_n^{s+1} &\sim P(X_n|X_1^{s+1}, X_2^{s+1}, \ldots X_{n-1}^{s+1}, X_{n+1}, \ldots X_m)
\end{aligned}
$$

Eventually (after burn-in), each sample will constitute a sample of the true $P(X_1,\ldots,X_n \mid X_{n+1}, \ldots, X_m)$

* but often use every 100th sample, since iterations not independent

Gibbs Sampling:

Goal: Directly sample conditional distributions

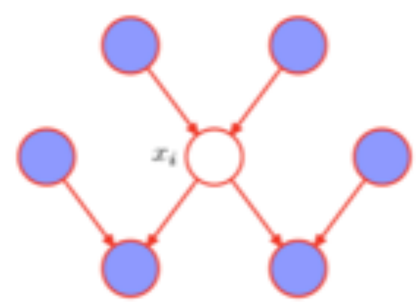
$P(X_1,\ldots,X_n \mid X_{n+1}, \ldots, X_m)$

Approach:

- start with the fixed observed $X_{n+1}, \ldots, X_m$
  plus arbitrary initial values for the unobserved $X_1^{(0)},\ldots,X_n^{(0)}$
- iterate for s=0 to a big number:

$$X_1^{s+1} \sim P(X_1|X_2^s, X_3^s \ldots X_n^s, X_{n+1}, \ldots X_m)$$
$$X_2^{s+1} \sim P(X_2|X_1^{s+1}, X_3^s \ldots X_n^s, X_{n+1}, \ldots X_m)$$
$$\ldots$$
$$X_n^{s+1} \sim P(X_n|X_1^{s+1}, X_2^{s+1}, \ldots X_{n-1}^{s+1}, X_{n+1}, \ldots X_m)$$

and need only the Markov Blanket at each step!



Gibbs is a special case of Markov Chain Monte Carlo methods

Inference in Bayes Nets

- Worst case is intractable (NP-complete)

- For certain cases, tractable <u>exact</u> solutions
  - Assigning probability to full joint assignment of variable values
  - Or if just one variable unobserved: $P(X_k | X_1, X_2, \ldots)$
  - Other special cases

- Can often <u>estimate</u> probabilities by sampling the probability distribution: Monte Carlo methods
  - Generate many samples, then use MLE estimates from samples
  - Gibbs sampling (example of Markov Chain Monte Carlo)

- Many other approaches beyond this class
  - Variational methods for tractable approximate solutions
  - Junction tree, Belief propagation, …
  - see Probabilistic Graphical Models course 10-708 (which Matt is teaching!)