Machine Learning 10-601

Tom M. Mitchell
Machine Learning Department
Carnegie Mellon University

February 23, 2015

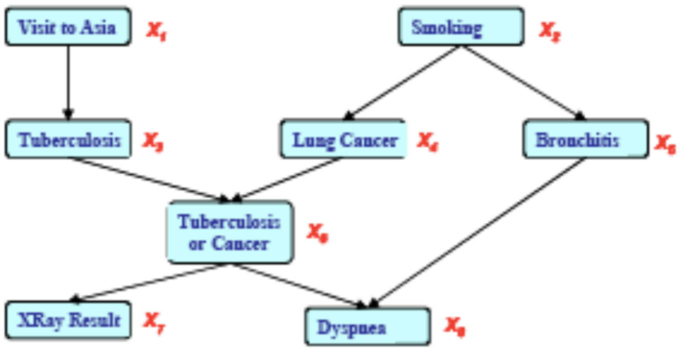
Today:

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

Readings:

- Bishop chapter 8, through 8.2
- Mitchell chapter 6

Bayes Nets define Joint Probability Distribution in terms of this graph, plus parameters



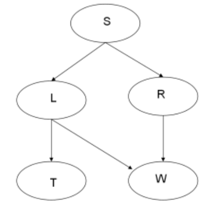
$$P(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8)$$

$$= P(X_1)\, P(X_2)\, P(X_3 | X_1)\, P(X_4 | X_2)\, P(X_5 | X_2)\, P(X_6 | X_3, X_4)\, P(X_7 | X_6)\, P(X_8 | X_5, X_6)$$

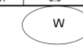
Benefits of Bayes Nets:

- Represent the full joint distribution in fewer parameters, using prior knowledge about dependencies
- Algorithms for inference and learning

Bayesian Networks Definition



Parents	P(W|Pa)	P(¬W|Pa)
L, R	0	1.0
L, ¬R	0	1.0
¬L, R	0.2	0.8
¬L, ¬R	0.9	0.1



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

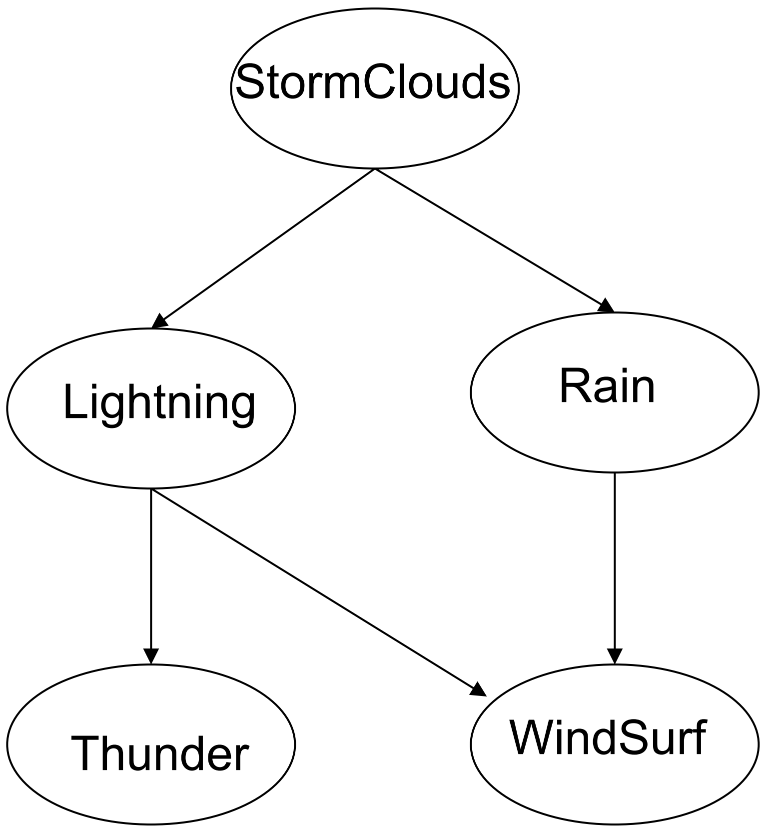
A Bayes network is a directed acyclic graph and a set of conditional probability distributions (CPD's)

- 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

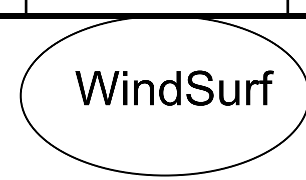
Bayesian Network



Nodes = random variables

A conditional probability distribution (CPD) is associated with each node N, defining P(N | Parents(N))

Parents	P(W|Pa)	P(¬W|Pa)
L, R	0	1.0
L, ¬R	0	1.0
¬L, R	0.2	0.8
¬L, ¬R	0.9	0.1

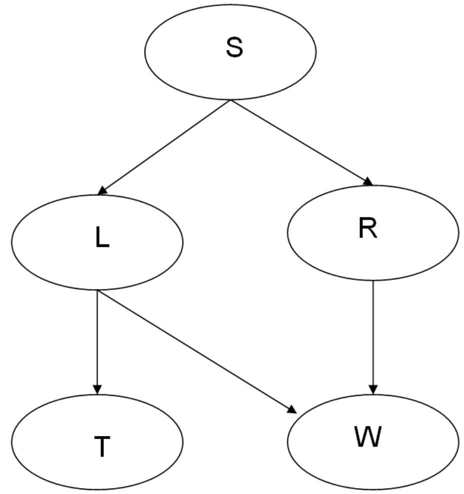


The joint distribution over all variables:

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

Bayesian Networks

- CPD for each node X_i describes $P(X_i \mid Pa(X_i))$



Parents	P(W|Pa)	P(¬W|Pa)
L, R	0	1.0
L, ¬R	0	1.0
¬L, R	0.2	0.8
¬L, ¬R	0.9	0.1

W

Chain rule of probability:

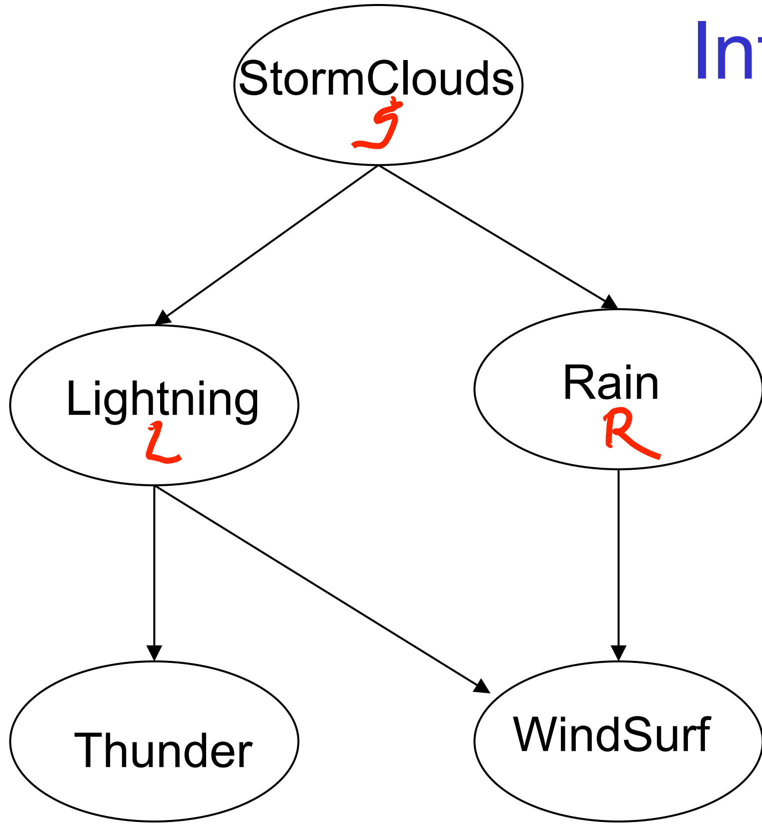
$$P(S,L,R,T,W) = P(S)P(L|S)P(R|S,L)P(T|S,L,R)P(W|S,L,R,T)$$

But in a Bayes net: $P(X_1 \ldots X_n) = \prod_i P(X_i|Pa(X_i))$

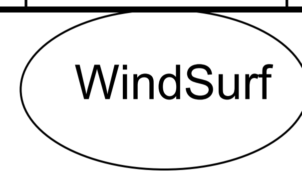
$$P(S\,L\,R\,T\,W) = P(S)\,P(L|S)\,P(R|S)\,P(T|L)\,P(W|L,R)$$

$$(\forall s\,l\,r\,t\,w)\; P(S{=}s, L{=}l \cdots) = P(S{=}s)\,P(L{=}l|S{=}s) \ldots \quad ''$$

Inference in Bayes Nets

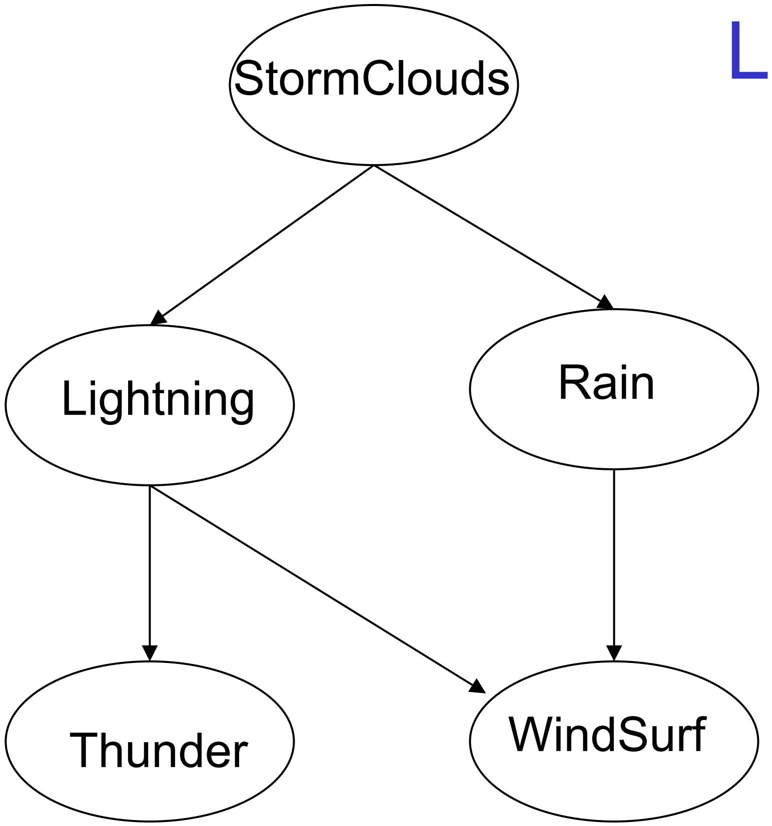


Parents	P(W|Pa)	P(¬W|Pa)
L, R	0	1.0
L, ¬R	0	1.0
¬L, R	0.2	0.8
¬L, ¬R	0.9	0.1

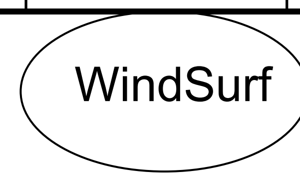


$$P(S=1, L=0, R=1, T=0, W=1) = P(S=1)\,P(L=0 \mid S=1)\,P(R=1 \mid S=1)\,P(T=0 \mid L=0)\,\underbrace{P(W=1 \mid L=0, R=1)}_{0.2}$$

Learning a Bayes Net



Parents	P(W|Pa)	P(¬W|Pa)
L, R	0	1.0
L, ¬R	0	1.0
¬L, R	0.2	0.8
¬L, ¬R	0.9	0.1



Consider learning when graph structure is given, and data = { <s,l,r,t,w> }

What is the MLE solution? MAP?

Algorithm for Constructing Bayes Network

- Choose an ordering over variables, e.g., $X_1, X_2, \ldots X_n$
- For i=1 to n
  - Add X_i to the network
  - Select parents $Pa(X_i)$ as minimal subset of $X_1 \ldots X_{i-1}$ such that

$$P(X_i|Pa(X_i)) = P(X_i|X_1, \ldots, X_{i-1})$$

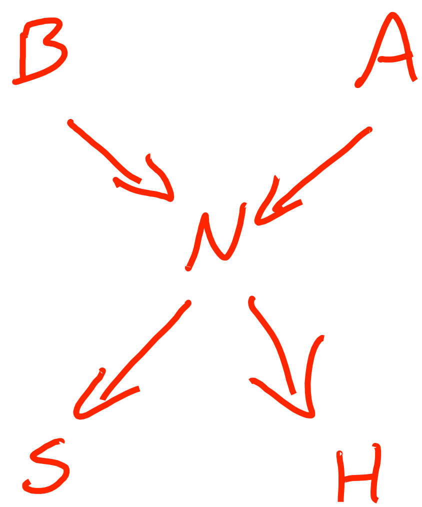
Notice this choice of parents assures

$$P(X_1 \ldots X_n) = \prod_i P(X_i|X_1 \ldots X_{i-1}) \quad \text{(by chain rule)}$$

$$= \prod_i P(X_i|Pa(X_i)) \quad \text{(by construction)}$$

Example

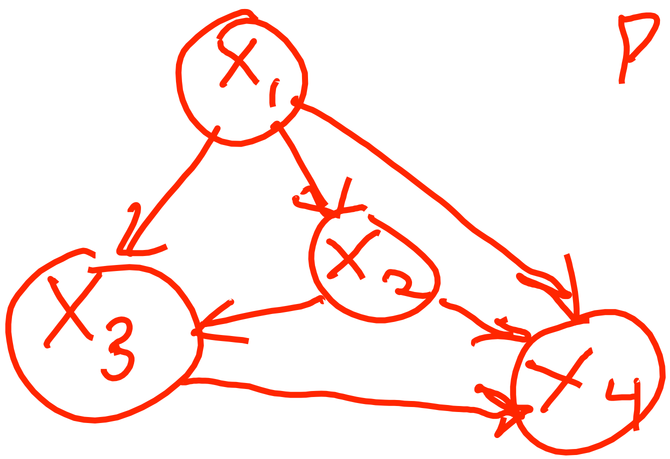
- Bird flu and Allegies both cause Nasal problems
- Nasal problems cause Sneezes and Headaches



What is the Bayes Network for X1,…X4 with NO assumed conditional independencies?

$$P(X_1, X_2, X_3, X_4) = P(X_1)\,P(X_2|X_1)\,P(X_3|X_1,X_2)\,P(X_4|X_1,X_2,X_3)$$

$$P(X_3)\,P(X_2|X_3)\,P(X_4|X_2,X_3)\,P(X_1|X_2,X_3,X_4)$$

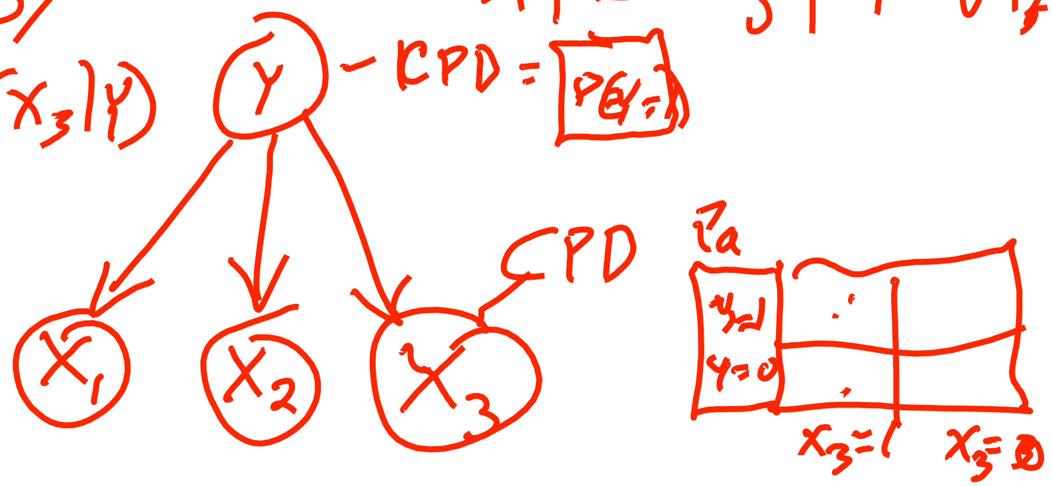


What is the Bayes Network for Naïve Bayes?

$$P(Y, X_1, X_2, X_3)$$

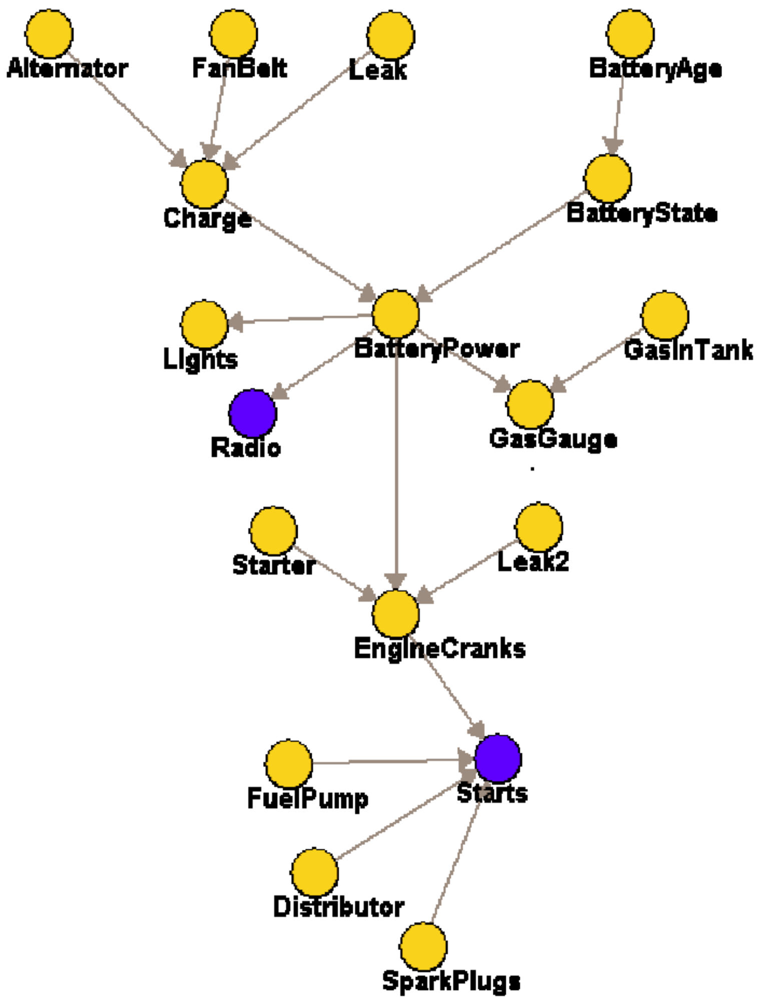
$$= P(Y)\,P(X_1|Y)\,P(X_2|Y)\,P(X_3|Y)$$

$$X_i \perp X_j \mid Y \quad \forall i \neq j$$



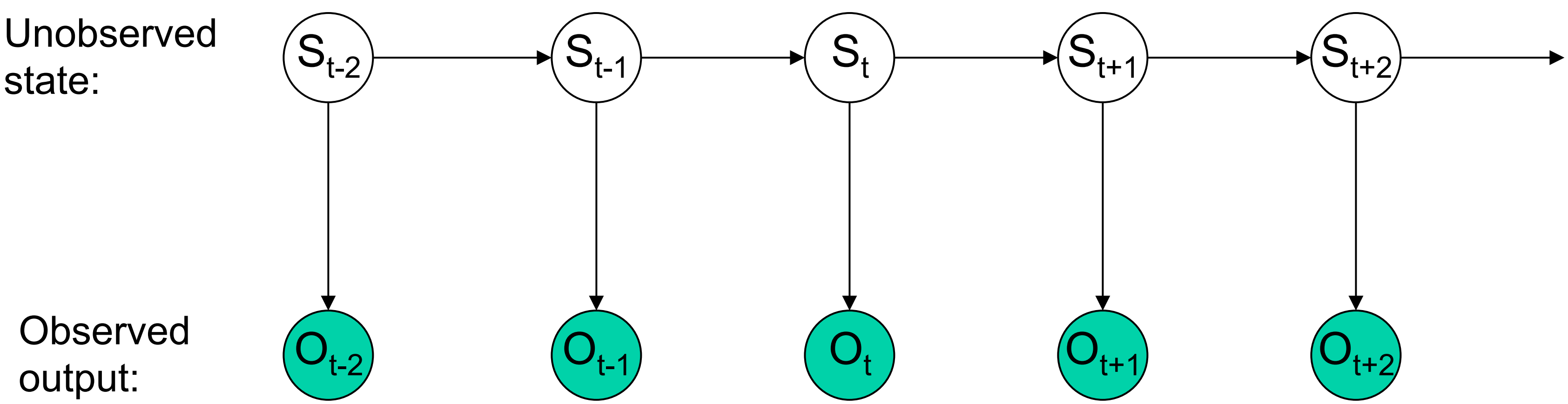
$$P(Y=1 \mid X_1=a, X_2=b, X_3=c) = \frac{P(Y=1, X_1=a, X_2=b, X=c)}{P(Y=1, X_1=a, X_2=b, X_3=c) + P(Y=0, X_1=a, X_2=b, X_3=c)}$$

What do we do if variables are mix of discrete and real valued?



Bayes Network for a Hidden Markov Model

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

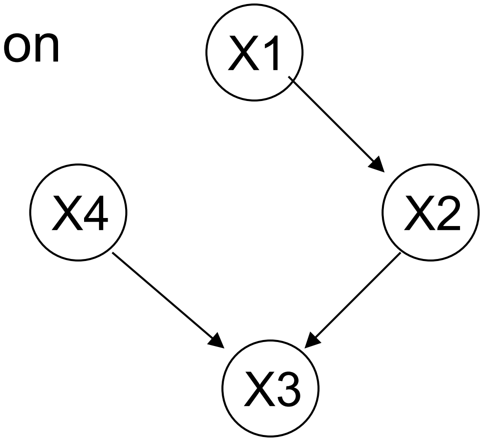


$$P(S_{t-2}, O_{t-2}, S_{t-1}, \ldots, O_{t+2}) =$$

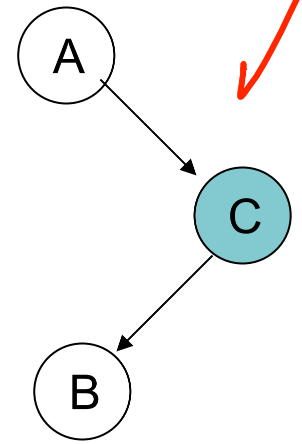
Conditional Independence, Revisited

- 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}
  - But X1 and X4 not conditionally indep given X3
  - For this, we need to understand D-separation



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)

$P(A=a)$ ← simply $P(a)$

$$p(ab|c) = P(a|c)\,P(b|c) \leftarrow A \perp B \mid C$$

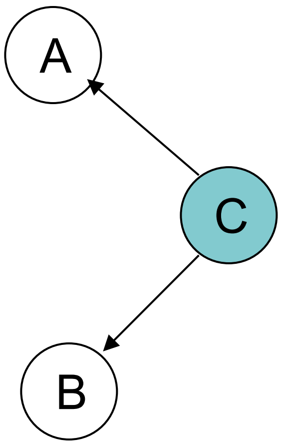
$$p(ab|c) = \frac{P(abc)}{P(c)} = \frac{P(a)\,P(c|a)}{P(c)}\,P(b|c)$$

$$\frac{P(ac)}{P(c)} = P(a|c)$$

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

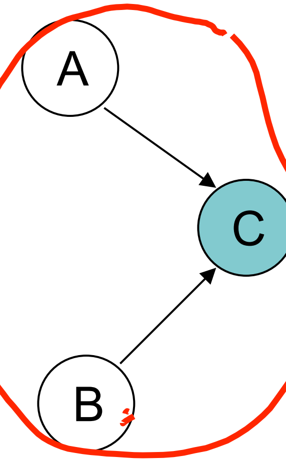
Easy Network 2: Tail to Tail

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



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)

No. False – $P(ab|c) \neq P(a|c)\,P(b|c)$

But $P(a,b) = P(a)\,P(b)$

$$P(a,b) = P(A=a,B=b,C=0) + P(A=a,B=b,C=1)$$

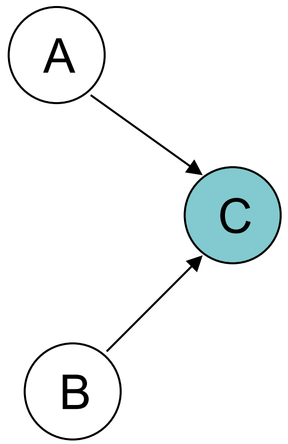
$$P(a)P(b)P(c=0|a,b) + P(a)P(b)P(c=1|ab)$$

$$P(ab) = P(a)P(b)\left[P(c=0|ab) + P(c=1|ab)\right]$$

1

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.,

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

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

[Bishop, 8.2.2]

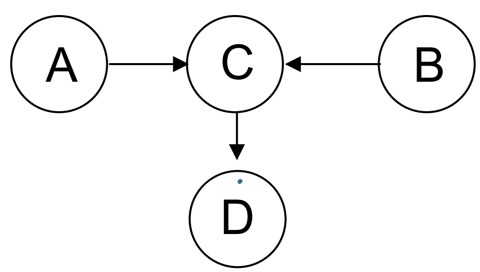
Suppose we have three sets of random variables: X, Y and Z

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

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

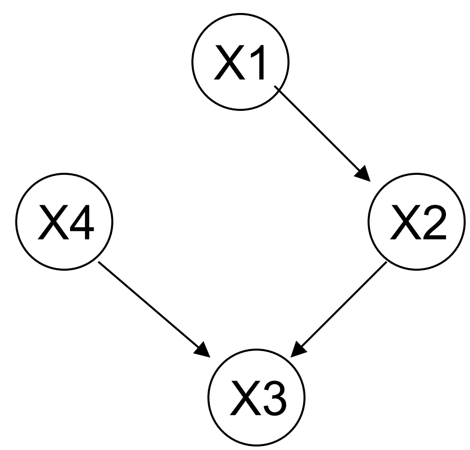
1.arrows on the path meet either head-to-tail or tail-to-tail at the node and this node is in Z

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

X1 indep of X3 given X2?

X3 indep of X1 given X2?

X4 indep of X1 given X2?

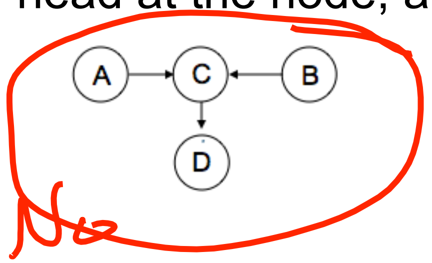


X and Y are **<u>D-separated</u>** by Z (and therefore conditionally indep, given Z) iff every path from any variable in X to any variable in Y is **<u>blocked</u>** by Z

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

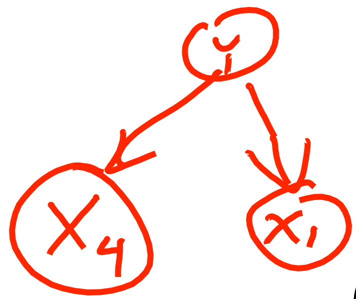
2. the arrows meet head-to-head at the node, and neither the node, nor any of its descendants, is in Z

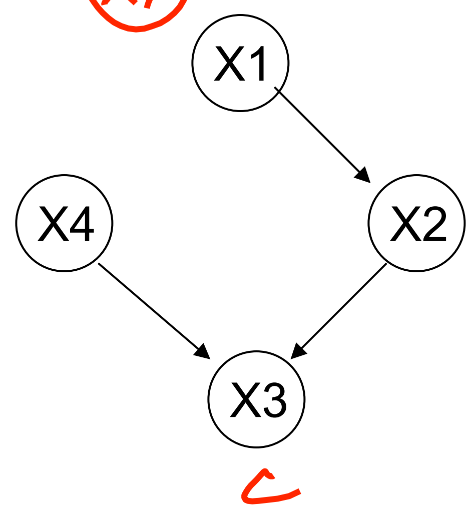


X4 indep of X1 given X3? No

X4 indep of X1 given {X3, X2}? Yes

X4 indep of X1 given {}? Yes





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>**

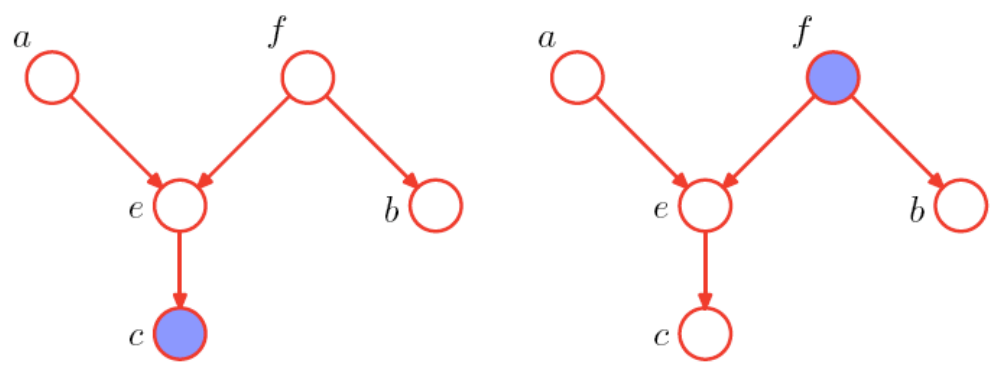
A path from variable A to variable B is **blocked** if it includes a node such that either

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?

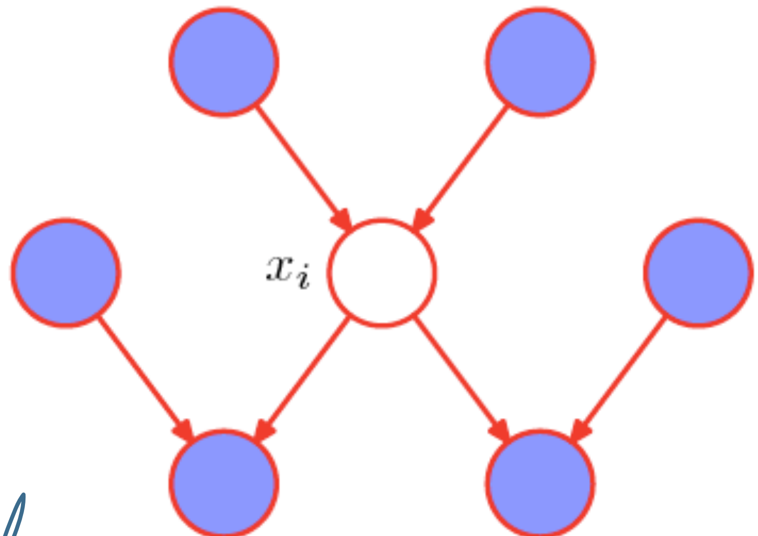
a indep of b given f ?



(a) (b)

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.



co-parent = other side of x_i's colliders

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
  - D-separation
  - 'Explaining away'

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)
    - Belief propagation
- For multiply connected graphs
    - Junction tree
- 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