Causal Inference: prediction, explanation, and intervention

Lecture 4: Graphical models

Samantha Kleinberg
samantha.kleinberg@stevens.edu

Main topics today

- Representation
- Inference
- Learning

Why graphical models?

Smoking	Chronic Bronchitis	Lung Cancer	Fatigue	Mass on X-ray	P(S,CB,LC,F,M)
F	F	F	F	F	0.2
F	F	F	F	T	0.1
	…				…
T	T	T	T	T	0.4

n binary variables = 2^n entries

Applications

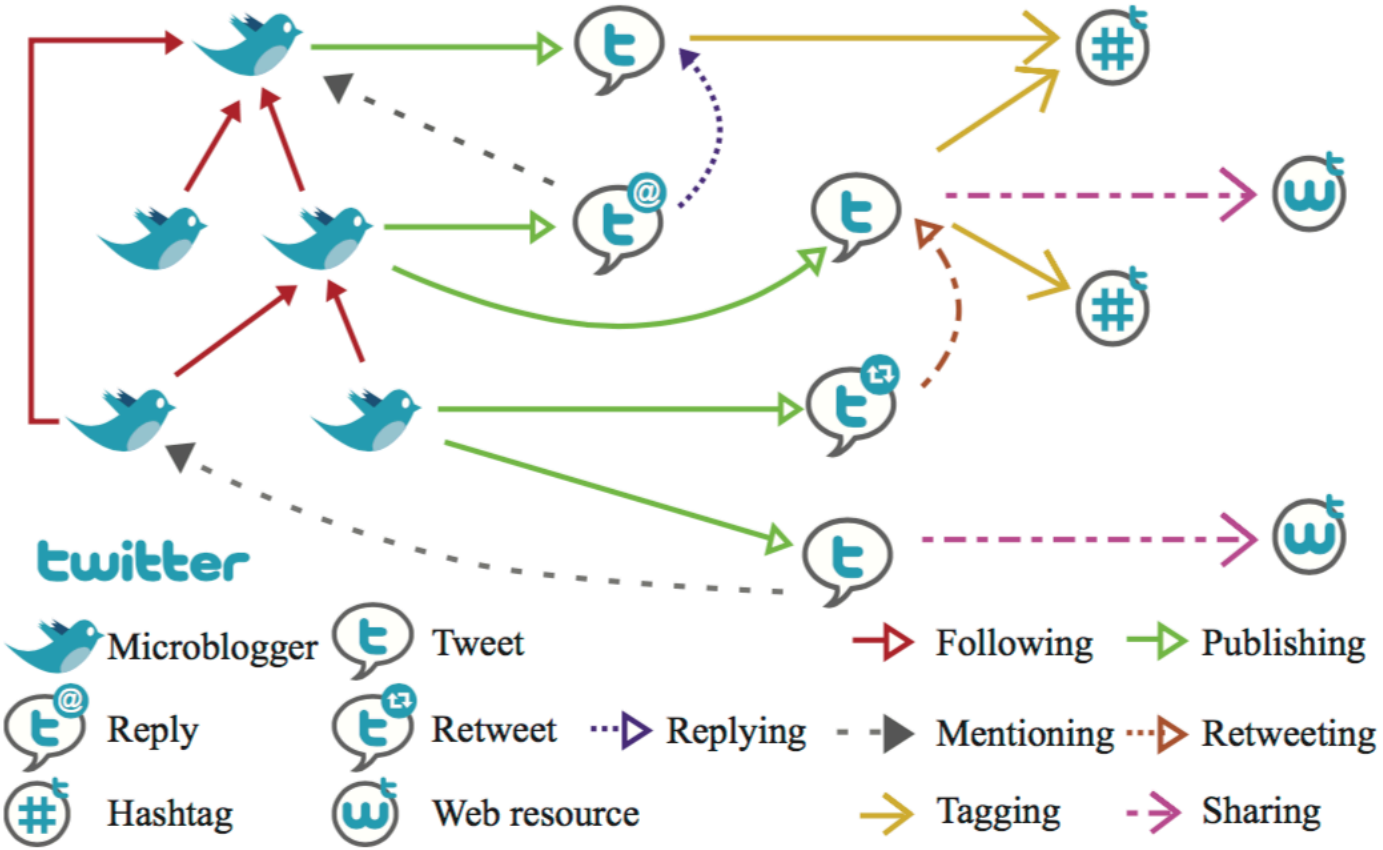


Figure 1: The social network of Twitter

Uprising microblogs: A Bayesian network retrieval model for tweet search

Lamjed Ben Jabeur
IRIT, Université Paul Sabatier
118 Route Narbonne
Toulouse, France
jabeur@irit.fr

Lynda Tamine
IRIT, Université Paul Sabatier
118 Route Narbonne
Toulouse, France
tamine@irit.fr

Mohand Boughanem
IRIT, Université Paul Sabatier
118 Route Narbonne
Toulouse, France
boughanem@irit.fr

Some notation

- X_1, X_2…X_n are a set of random variables
  - X_1 = variable, x=particular value of X_1
  - For coin, X={heads, tails}

$$\sum_{x} P(x) = \sum_{x \in \{heads, tails\}} P(x)$$

- $P(X_1, X_2 \ldots X_n)$ is set of joint probability distributions over all assignments of variables
  - If each variable binary, need 2^n parameters
  - Gives set of equations

Some more notation

Binary variables X,Y

$$P(x) = P(x \mid y)P(y) + P(x \mid \neg y)P(\neg y)$$

General case

$$P(x) = \sum_{Y} P(x \mid y)P(y)$$

Chain rule

Last week: $P(A \wedge B) = P(A|B)P(B)$

Multiple variables:

$$P(X,Y,Z) = P(X|Y,Z)P(Y,Z)$$
$$= P(X|Y,Z)P(Y|Z)P(Z)$$

Chain rule (general)

$$P(X_1 \ldots X_n) = P(X_1 \mid X_2 \ldots X_n) P(X_2 \mid X_3 \ldots X_n) \ldots P(X_n)$$

$$= \prod_{i=1}^{n} P(X_i \mid X_{i+1} .. X_n)$$

Key observation: independence

- If variables independent, fewer parameters needed
- When $X_1 \ldots X_n$ are all independent

$$P(X_1 \ldots X_n) = \prod_{i=1}^{n} P(X_i)$$

Conditional independence

Yellowed fingers and lung cancer are dependent:

$$P(Y,L)>P(Y)P(L)$$

But they are independent conditioned on smoking (S)

$$P(S,Y,L)=P(Y|L,S)P(L|S)P(S)$$
$$=P(Y|S)P(L|S)P(S)$$

Conditional independence

In general, for all x,y,z, X and Y are conditionally independent if

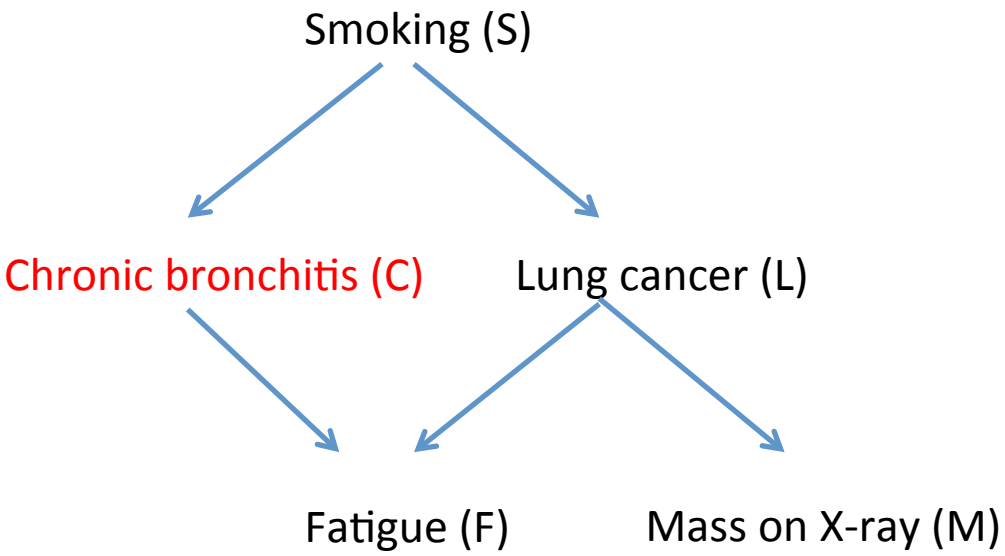
$$P(x \mid, y, z) = P(x \mid z)$$

Graphical model

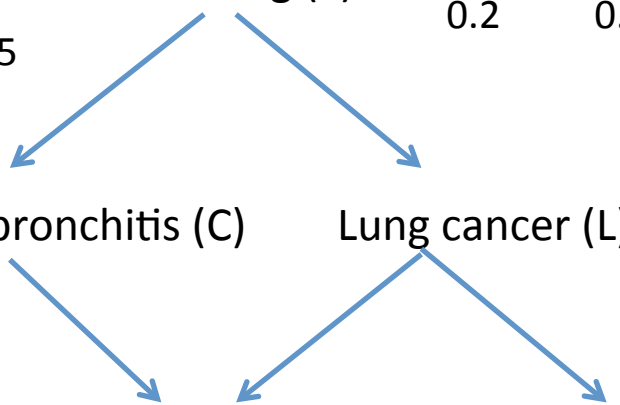
- Describes independencies in a set of variables

- Directed
  - Bayesian network
  - HMM (in a few weeks)
- Undirected

Graphical model and independencies



A node is independent of its non-descendants given its parents



	P(C=T)	P(C=F)
S=T	0.25	0.75
S=F	0.05	0.95

P(S=T)	P(S=F)
0.2	0.8

	P(L=T)	P(L=F)
S=T	0.003	0.997
S=F	0.000005	0.99995

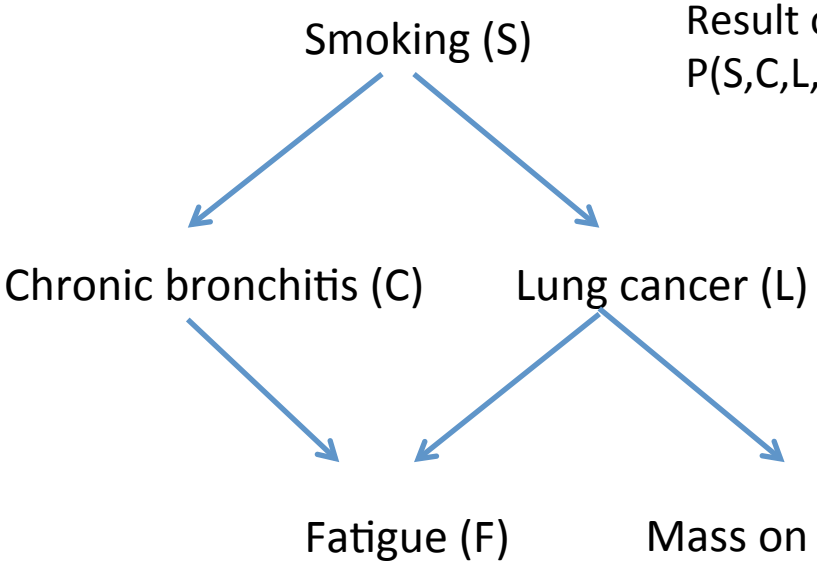
	P(F=T)	P(F=F)
C=T,L=T	0.75	0.25
C=T,L=F	0.10	0.90
C=F,L=T	0.50	0.50
C=F,L=F	0.05	0.95

	P(M=T)	P(M=F)
L=T	0.6	0.40
L=F	0.02	0.98

Instead of 2^5 parameters:

- P(S): 2
- P(C|S): 4
- P(L|S): 4
- P(F|C,L)=8
- P(M|L)=2

Graphical model and joint distribution



Result of applying chain rule:

$$P(S,C,L,F,M)=P(M|F,L,C,S)P(F|L,C,S)P(L|C,S)\ P(C|S)P(S)$$

$$P(S,C,L,F,M)=$$
$$P(M|L)P(F|C,L)P(L|S)\ P(C|S)P(S)$$

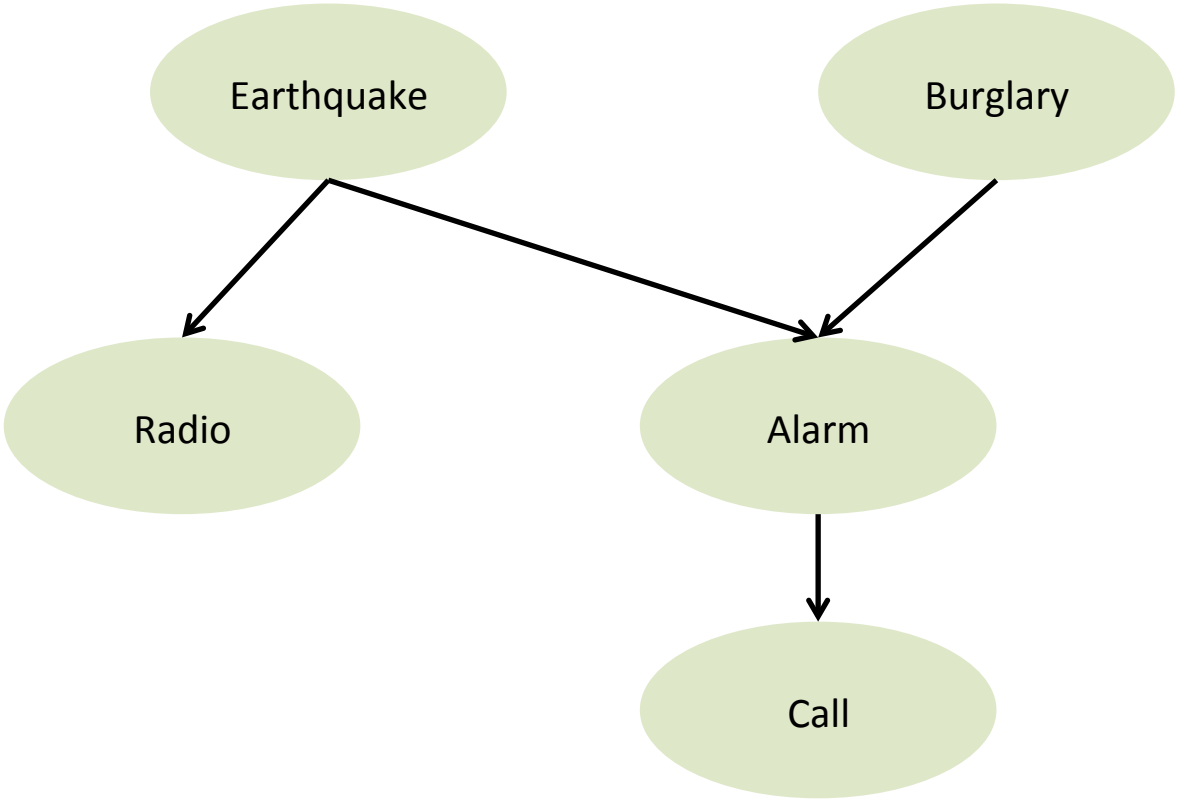
DAG and joint probability

$$P(X_1 \ldots X_n) = \prod_{i=1}^{n} P(X_i \mid X_{i+1} .. X_n)$$

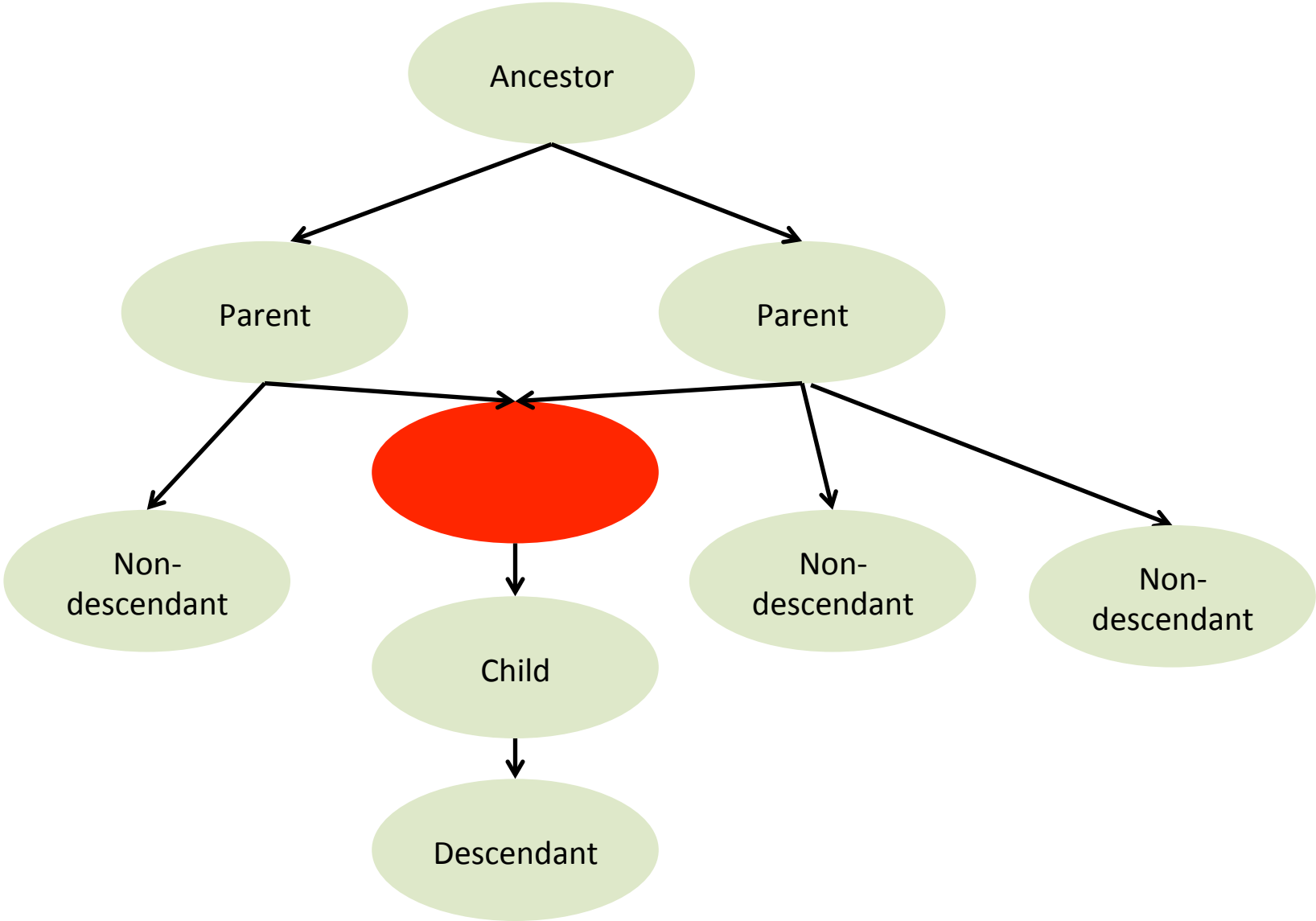
Factorize:

$$P(X_1 \ldots X_n) = \prod_{i}^{n} P(X_i \mid pa(X_i))$$

Example

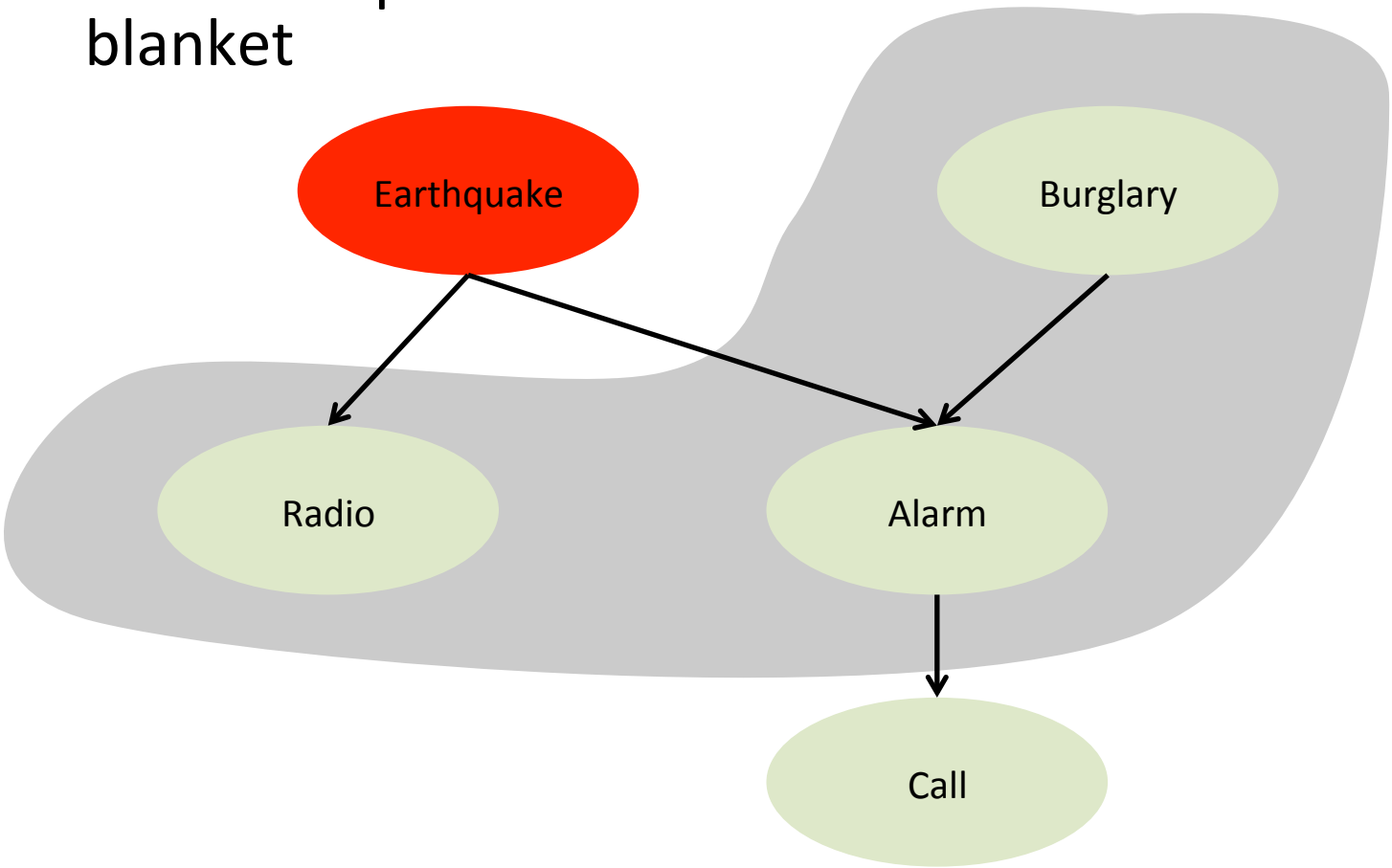


Terminology

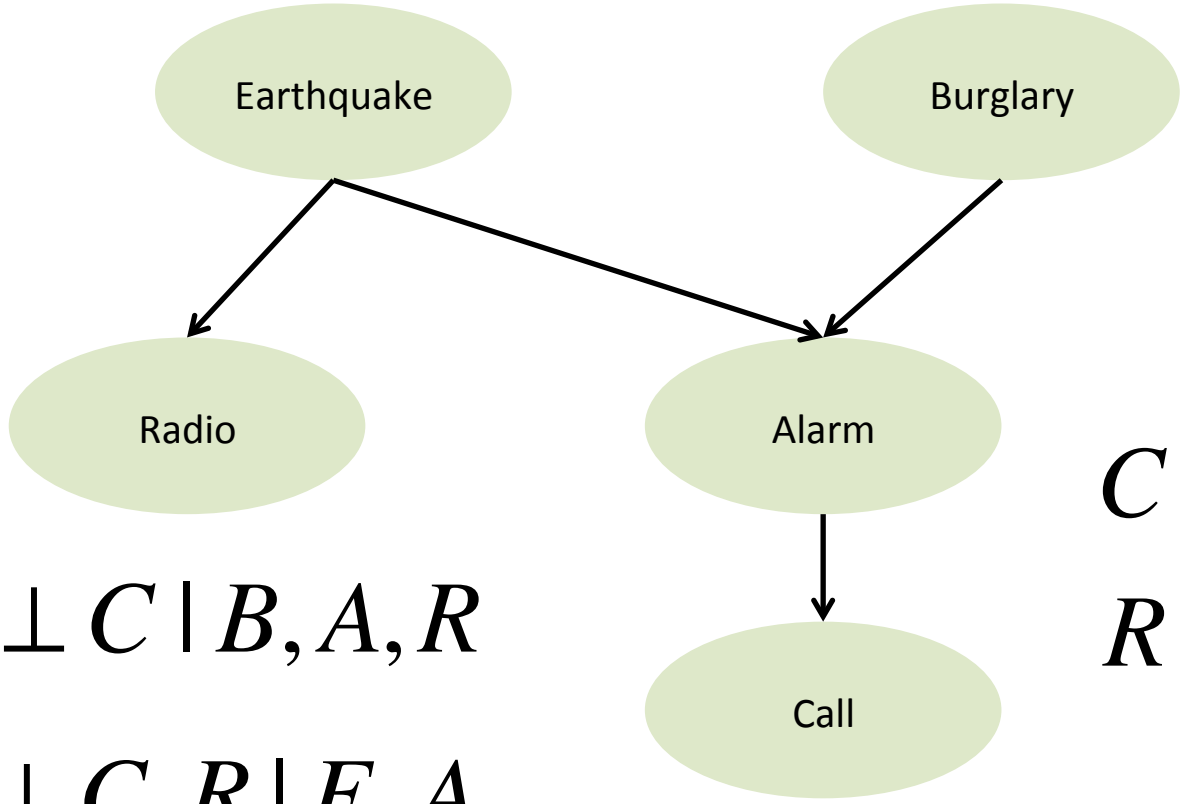


Markov blanket

- Markov blanket: parents, children, children's parents
- Node independent of all others conditioned on Markov blanket



Conditional independence from graph



$$E \perp B$$

$$A \perp R \mid E, B$$

$$B \perp E, R$$

$$C \perp B, E, R \mid A$$

$$R \perp A, B, C \mid E$$

$$E \perp C \mid B, A, R$$

$$B \perp C, R \mid E, A$$

Components of a Bayesian network

1. Directed acyclic graph
2. Conditional probability distributions

$$P(X_1 \ldots X_n) = \prod_i^n P(X_i \mid pa(X_i))$$

Conditional probability distribution

- In this lecture: discrete (usually binary valued) variables
- Conditional probability tables

X	Y	P(Z=T)	P(Z=F)
T	T	0.5	0.5
T	F	0.2	0.8
F	T	0.1	0.9
F	F	0.3	0.7

Simple method for building a BN

For $X_1, X_2, \ldots X_{n,}$:

-Add X_i to graph

-Add edges from $pa(X_i)$ where

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

-Add conditional probability $P(X_i | pa(X_i))$

Pearl, J (1988) Probabilistic Reasoning in Intelligent Systems

Example (1)

- Coin A, Coin B

$$A \perp B$$

- What's Bayesian network?

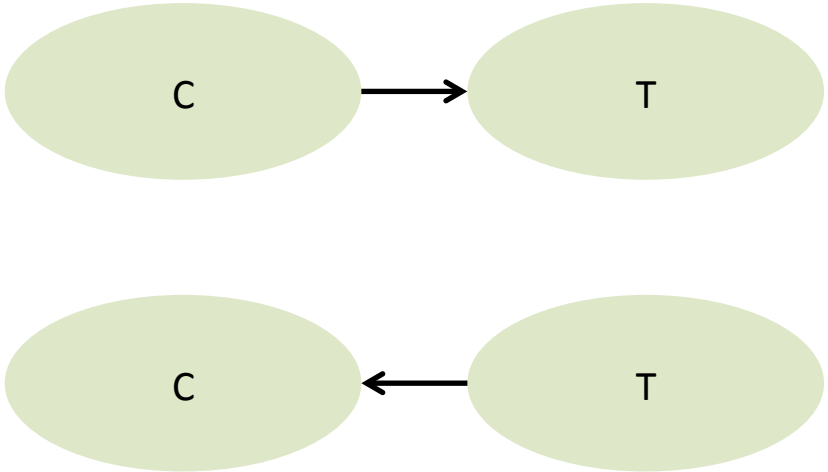


Example (2)

- Cavity, Toothache

$$C \not\perp T$$

- Possible BNs?

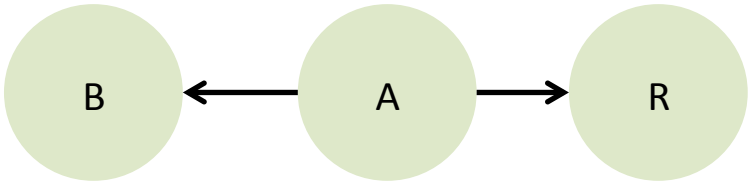


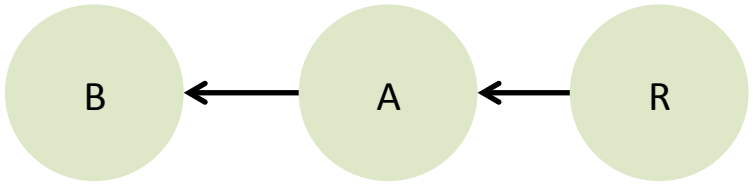
Example (3)

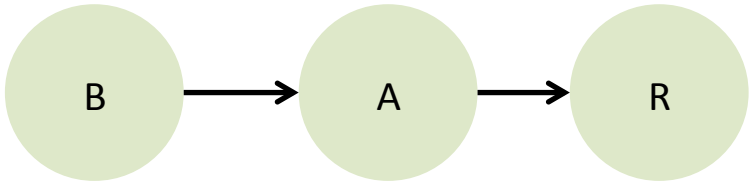
- Barometer, Rain, Air Pressure

$$B \perp R|A$$

- Possible BNs?



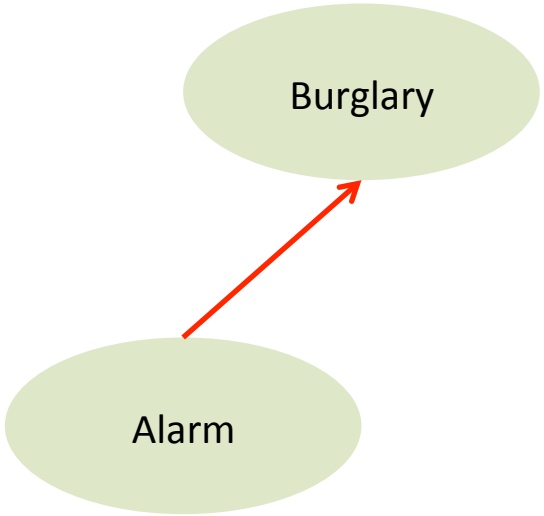


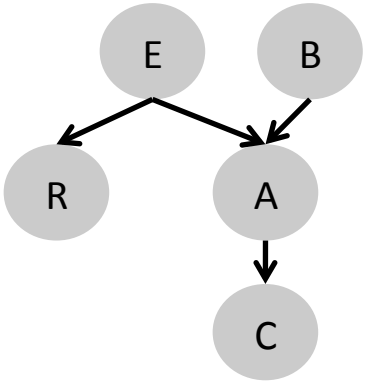


Example (4)

Order: Alarm, Burglary, Call, Earthquake, Radio

P(B|A)=P(B)?

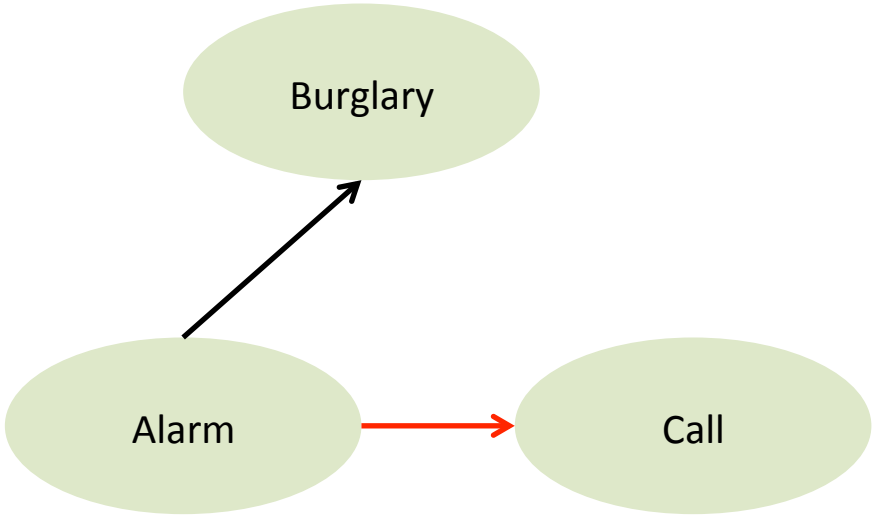


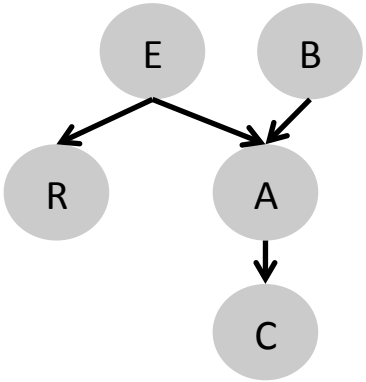


Example (4)

- Alarm, Burglary, Call, Earthquake, Radio

P(C|A,B)=P(C|A)?
=P(C|B)?
=P(C)?

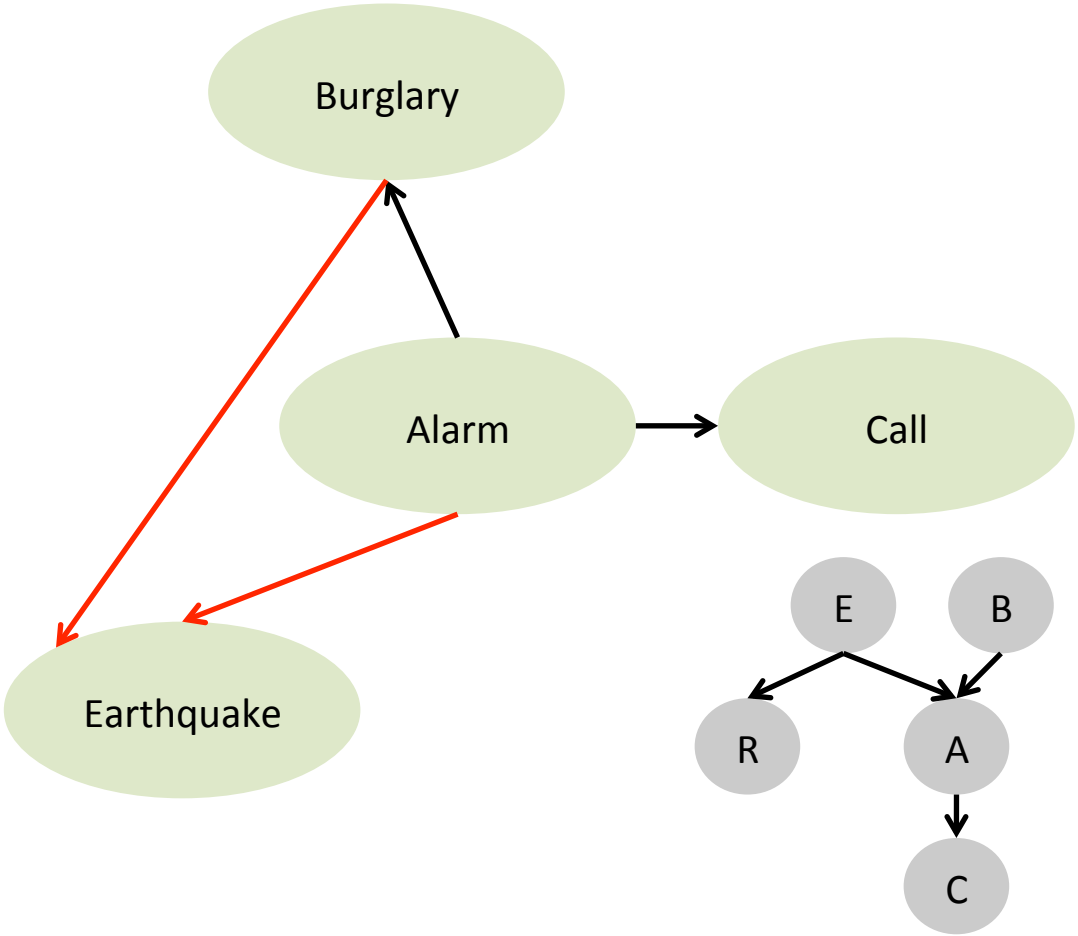




Example (4)

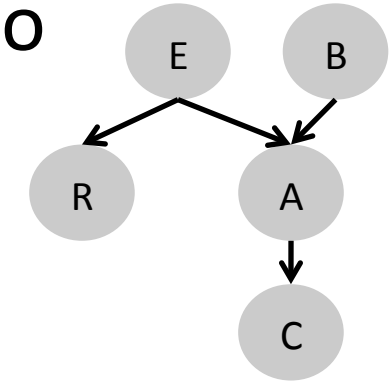
- Alarm, Burglary, Call, Earthquake, Radio

P(E|B,A,C)
=P(E|B,C)?
=P(E|A)?



Example (4)

- Alarm, Burglary, Call, Earthquake, Radio



P(R|B,A,E,C)
=P(R|E)?

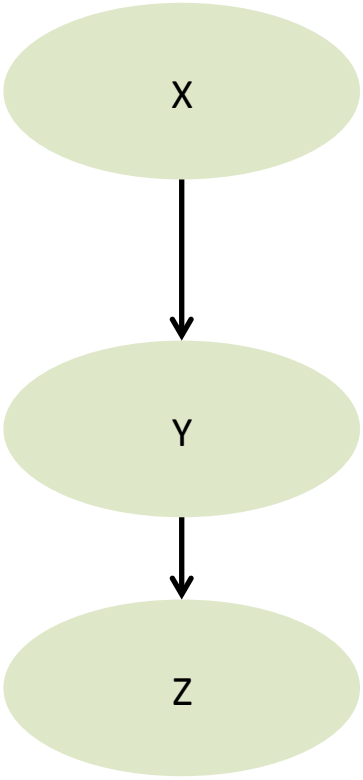
Notes on creating a BN

- Multiple BNs consistent with same relationships
- Order of operations can matter

Inference

- Unconditional probability
  - Probability of variable(s) taking particular values
  - Probability of smoking and asbestos exposure
- How to find probability of variable(s) of interest given what is known (posterior)
  - Probability of query given evidence
  - Probability of disease | symptoms, test result
- Finding most likely value of variables given evidence
  - Diagnosis given set of symptoms
  - If someone has a cough, what's the most probable dx

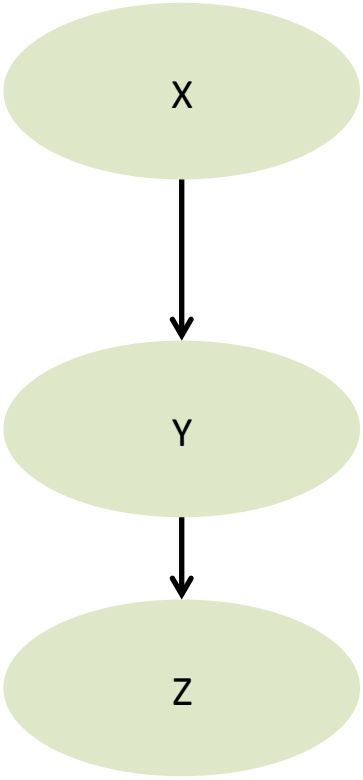
Unconditional probability



P(Y=T)?

$$P(Y=T) = \sum_X P(X=x, Y=T)$$
$$= \sum_X P(Y=T|X=x)P(X=x)$$

Unconditional probability



P(Z=T)?

From before: $P(Y = T) = \sum_X P(X = x, Y = T)$

$$= \sum_X P(Y = T|X = x)P(X = x)$$

$$P(Z = T) = \sum_{X,Y} P(X = x, Y - y, Z = T)$$

$$= \sum_{X,Y} P(Z = T|X = x, Y = y)P(X = x, Y = y)$$

$$= \sum_{X,Y} P(Z = T|X = x, Y = y)P(Y = y|X = x)P(x = x)$$

$$= \sum_{X,Y} P(Z = T|Y = y)P(Y = y|X = x)P(x = x)$$

$$= \sum_Y P(Z = T|Y = y)P(Y = y)$$

General case

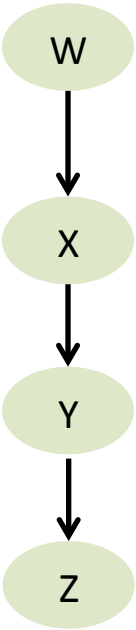
Goal: Calculate $P(X_1)$ with variables $X_1 \ldots X_n$

Naïve approach:

$$P(X_1 = x_1) = \sum_{X_2 \ldots X_n} P(x_1 \ldots X_n) = \sum_{X_2 \ldots X_n} P(x_1 \mid X_2 \ldots X_N) P(X_2 \ldots X_N)$$

Better approach: use conditional independence to rewrite joint probability

Key observation



$$P(Z=T)=\sum_{Y,X,W} P(y,x,w,Z=T)$$

$$=\sum_{y}\sum_{x}\sum_{w} P(Z=t\mid y)P(y\mid x)P(x\mid w)P(w)$$

$$=\sum_{y}\sum_{x} P(Z=t\mid y)P(y\mid x)\sum_{w} P(x\mid w)P(w)$$

But $$P(x)=\sum_{w} P(x\mid w)P(w)$$

$$=\sum_{y}\sum_{x} P(Z=t\mid y)P(y\mid x)P(x)$$

$$=\sum_{y} P(Z=t\mid y)P(y)$$

General approach

$$P(X_1 \ldots X_n) = \sum_{X_1} \sum_{X_2} \ldots \sum_{X_n} \prod_{i}^{n} P(x_i \mid pa(x_i))$$

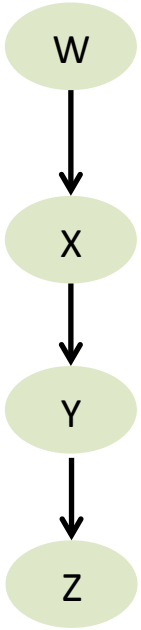
- Iteratively remove variables
- Why useful?
  - Save and reuse results! (can use dynamic programming)

Variable elimination

1. Order variables $X_1 \ldots X_n$
2. For each X_i in $X_1 \ldots X_n$
   - Eliminate X_i by taking all factors that include X_i and replacing with new factor
   - New factor is product of the m factors, summed over values of X_i

$$f_{Xi}(X_1 \ldots X_k) = \sum_{X_i} \prod_{j=1}^{m} f_{Xi}(X_i, X_1 \ldots X_k)$$

Variable elimination steps



$$P(Z=T) = \sum_{Y,X,W} P(y,x,w,Z=T)$$

$$= \sum_{y} \sum_{x} \sum_{w} P(Z=t \mid y)P(y \mid x)P(x \mid w)P(w)$$

1. Eliminate w

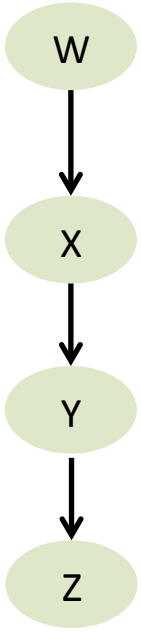
$$f_W(x) = \sum_{w} P(x \mid w)P(w)$$

$$f_W(x) = P(x)$$

Now:

$$\sum_{y} \sum_{x} P(Z=t \mid y)P(y \mid x)f_w(x)$$

Variable elimination steps



$$P(Z=T) = \sum_{Y,X,W} P(y,x,w,Z=T)$$

$$= \sum_{y} \sum_{x} \sum_{w} P(Z=t \mid y)P(y \mid x)P(x \mid w)P(w)$$

$$= \sum_{y} \sum_{x} P(Z=t \mid y)P(y \mid x)f_w(x)$$

2. Eliminate x

$$f_x(y) = \sum_{x} P(y \mid x)f_w(x)$$

Now:

$$\sum_{y} P(Z=t \mid y)f_x(y)$$

Variable elimination summary

- Exact inference
  - NP-hard
- Ordering matters
  - Algorithms for optimizing ordering
  - Finding best order NP-complete

Inference

- Unconditional probability
  - Probability of variable(s) taking particular values
  - Probability of smoking and asbestos exposure
- How to find probability of variable(s) of interest given what is known (posterior)
  - Probability of query given evidence
  - Probability of disease | symptoms, test result
- Finding most likely value of variables given evidence
  - Diagnosis given set of symptoms

Variable elimination with evidence

- Now want to calculate probability of variable given what we know about value of other variables
  - Evidence = value for some variables (T/F here)
- Say we know someone visited Asia and had Dyspnea, how to take that into account when calculating probability their x-ray will be positive?

- Last week's mystery!

Evidence and Bayes rule

$$P(X = T \mid V = T, D = T) = \frac{P(X = T, V = T, D = T)}{P(V = T, D = T)}$$

$$= \frac{P(X = T, V = T, D = T)}{\sum_{X} P(x, V = T, D = T)}$$

Normalization

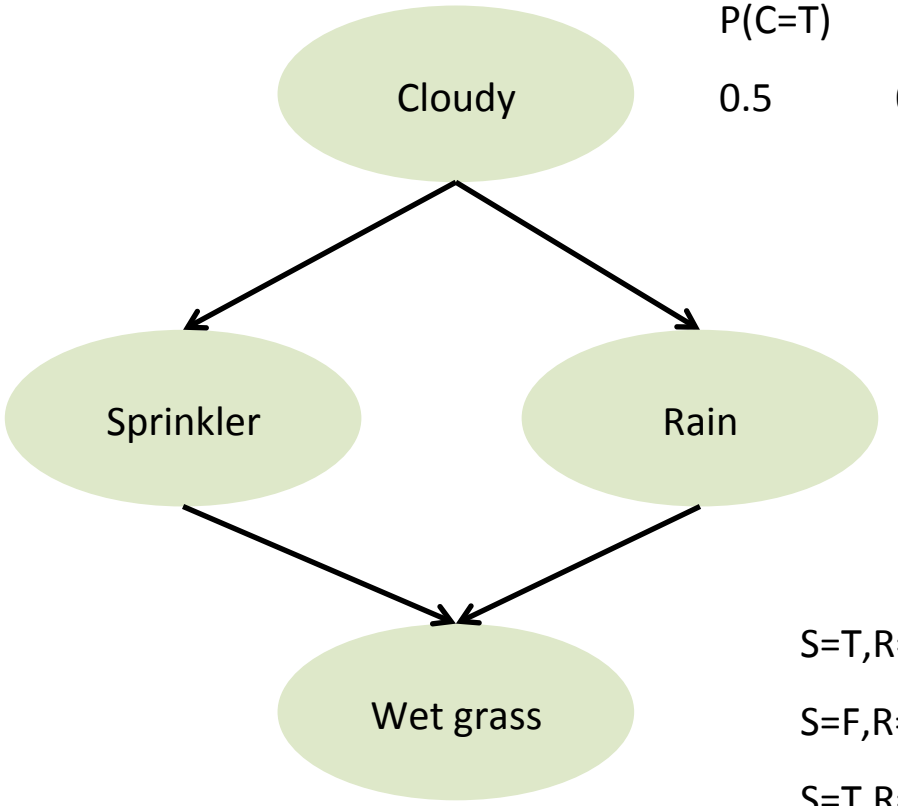
- P(X|e) is proportional to P(X,e)

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

- Calculate P(X,e), at very end divide by P(e) to get conditional probability

Example with evidence

P(R=T|W=T)



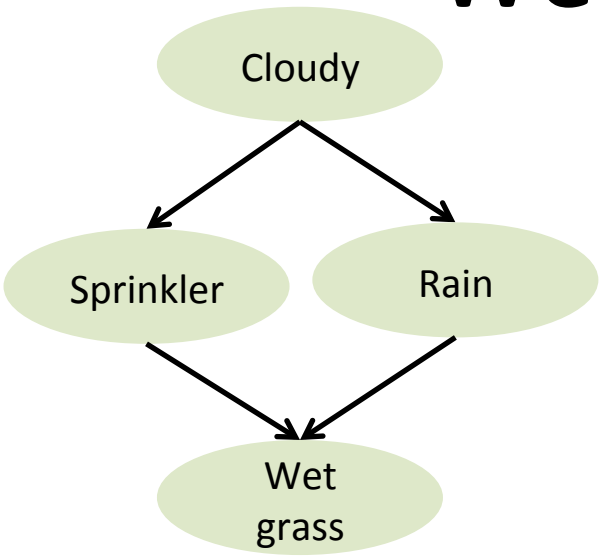
	P(C=T)	P(C=F)
	0.5	0.5

	P(S=T)	P(S=F)
C=T	0.1	0.9
C=F	0.5	0.5

	P(R=T)	P(R=F)
C=T	0.8	0.2
C=F	0.2	0.8

	P(W=T)	P(W=F)
S=T,R=T	0.99	0.01
S=F,R=T	0.9	0.1
S=T,R=F	0.9	0.1
S=F,R=F	0.01	0.99

Wet grass and rain



$$P(R=T \mid W=T) = \frac{P(R=T, W=T)}{P(W=T)}$$

$$P(R=T, W=T) = \sum_S \sum_C P(R=T, W=T, s, c)$$

$$= \sum_S \sum_C P(c)P(s \mid c)P(R=T \mid c)P(W \mid s, R=T)$$

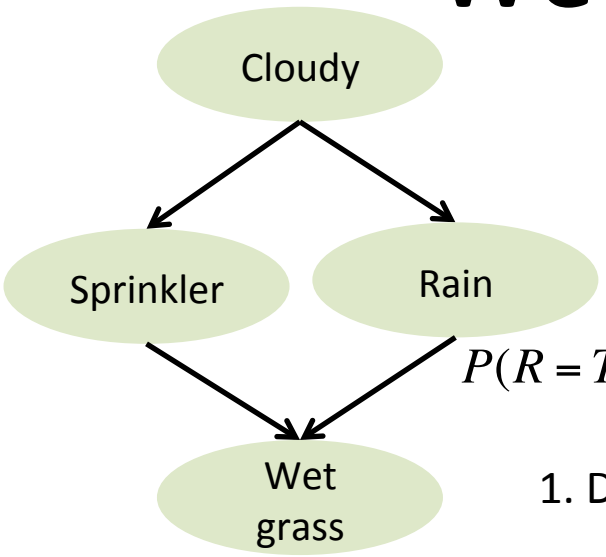
	P(S=T)	P(S=F)
C=T	0.1	0.9
C=F	0.5	0.5

	P(W=T)	P(W=F)
S=T,R=T	0.99	0.01
S=F,R=T	0.9	0.1
S=T,R=F	0.9	0.1
S=F,R=F	0.01	0.99

P(C=T)	P(C=F)
0.5	0.5

	P(R=T)	P(R=F)
C=T	0.8	0.2
C=F	0.2	0.8

Wet grass and rain



$$P(R=T \mid W=T) = \frac{P(R=T, W=T)}{P(W=T)}$$

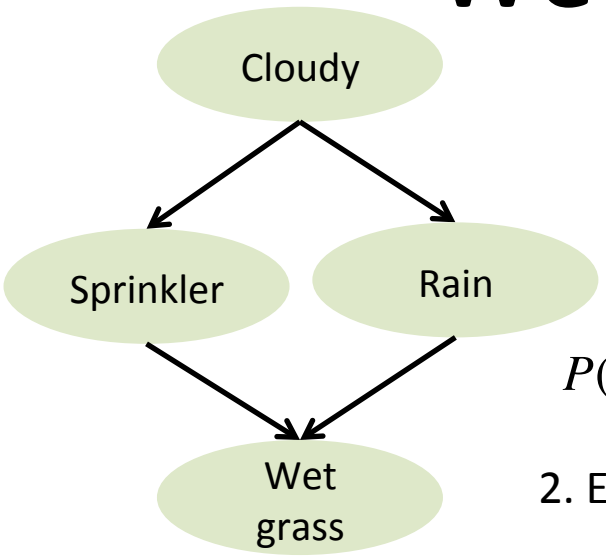
$$P(R=T, W=T) = \sum_C \sum_S P(c)P(s \mid c)P(R=T \mid c)P(W=T \mid s, R=T)$$

1. Deal with evidence variables R,W

$$f_{P(R=T|c)}(c) = P(R=T \mid c) \qquad f_{P(W=T|S,R=T)}(s) = P(W=T \mid s, R=T)$$

$$P(R=T, W=T) = \sum_C \sum_S P(c)P(s \mid c) f_{P(R=T|c)}(c) f_{P(W=T|s,R=T)}(s)$$

Wet grass and rain



$$P(R=T \mid W=T) = \frac{P(R=T, W=T)}{P(W=T)}$$

$$P(R=T, W=T) = \sum_S \sum_C P(c)P(s \mid c) f_{P(R=T|c)}(c) f_{P(W=T|s,R=T)}(s)$$

2. Eliminate c

Reminder:

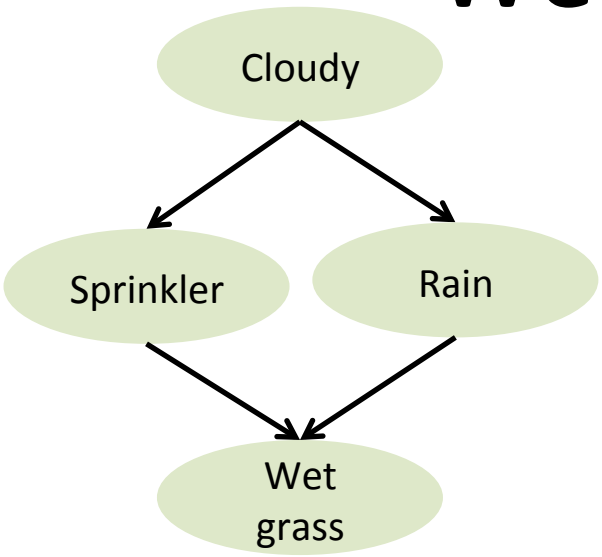
$$f_{P(R=T|c)}(c) = P(R=T \mid c)$$

$$f_{P(W=T|S,R=T)}(s) = P(W=T \mid s, R=T)$$

$$f_c(s) = \sum_C P(s \mid c)P(c) f_{P(R=T|c)}(c)$$

$$P(R=T, W=T) = \sum_S f_c(s) f_{P(W=T|s,R=T)}(s)$$

Wet grass and rain



$$P(R=T \mid W=T) = \frac{P(R=T, W=T)}{P(W=T)}$$

$$P(R=T, W=T) = \sum_S f_c(s) f_{P(W=T \mid S, R=T)}(s)$$

Now left with sum over values of S

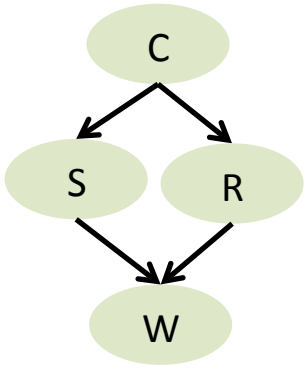
Reminder:

$$f_{P(R=T \mid c)}(c) = P(R=T \mid c)$$

$$f_{P(W=T \mid S, R=T)}(s) = P(W=T \mid S, R=T)$$

$$f_c(s) = \sum_C P(s \mid c) P(c) f_{P(R=T \mid c)}(c)$$

Let's calculate!



Wet grass and rain

Reminder:

$$f_{P(R=T|c)}(c) = P(R = T \mid c)$$

$$f_{P(W=T|s,R=T)}(s) = P(W = T \mid s, R = T)$$

$$f_c(s) = \sum_C P(s \mid c) P(c) f_{P(R=T|c)}(c)$$

$$P(R = T \mid W = T) = \frac{P(R = T, W = T)}{P(W = T)}$$

$$P(R = T, W = T) = \sum_S f_c(s) f_{P(W=T|s,R=T)}(s)$$

C	$f_{P(R=T\|c)}(c)$
T	0.8
F	0.02

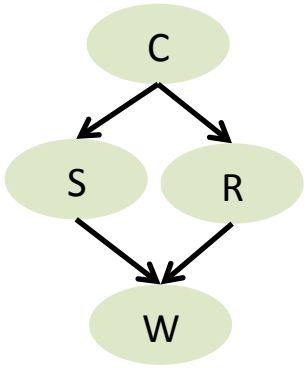
S	$f_{P(W=T\|s,R=T)}(s)$
T	0.99
F	0.9

	P(R=T)	P(R=F)
C=T	0.8	0.2
C=F	0.02	0.98

P(C=T)	P(C=F)
0.5	0.5

	P(S=T)	P(S=F)
C=T	0.1	0.9
C=F	0.5	0.5

	P(W=T)	P(W=F)
S=T,R=T	0.99	0.01
S=F,R=T	0.9	0.1
S=T,R=F	0.9	0.1
S=F,R=F	0.01	0.99



Wet grass and rain

$$P(R=T \mid W=T) = \frac{P(R=T, W=T)}{P(W=T)}$$

$$P(R=T, W=T) = \sum_S f_c(s) f_{P(W=T|s,R=T)}(s)$$

$$f_{P(R=T|c)}(c) = P(R=T \mid c)$$

$$f_{P(W=T|s,R=T)}(s) = P(W=T \mid s, R=T)$$

$$f_c(s) = \sum_C P(s \mid c) P(c) f_{P(R=T|c)}(c)$$

C	$f_{P(R=T\|c)}(c)$
T	0.8
F	0.02

S	$f_{P(W=T\|s,R=T)}(s)$
T	0.99
F	0.9

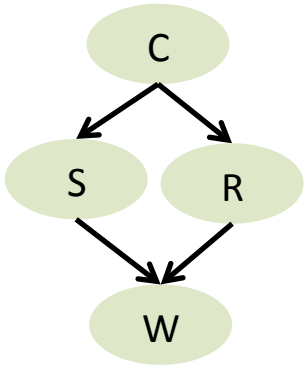
	P(R=T)	P(R=F)
C=T	0.8	0.2
C=F	0.02	0.98

P(C=T)	P(C=F)
0.5	0.5

	P(S=T)	P(S=F)
C=T	0.1	0.9
C=F	0.5	0.5

	P(W=T)	P(W=F)
S=T,R=T	0.99	0.01
S=F,R=T	0.9	0.1
S=T,R=F	0.9	0.1
S=F,R=F	0.01	0.99

C	S	$f_c(s)$
T	T	0.1*0.5*0.80=0.04
T	F	0.9*0.5*0.80=0.36
F	T	0.5*0.5*0.02=0.005
F	F	0.5*0.5*0.02=0.005



Wet grass and rain

$$P(R=T \mid W=T) = \frac{P(R=T, W=T)}{P(W=T)}$$

$$P(R=T, W=T) = \sum_S f_c(s) f_{P(W=T \mid S, R=T)}(s)$$

$$f_{P(R=T \mid c)}(c) = P(R=T \mid c)$$

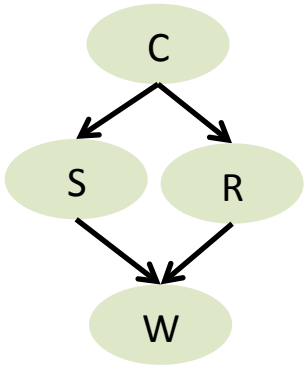
$$f_{P(W=T \mid S, R=T)}(s) = P(W=T \mid S, R=T)$$

$$f_c(s) = \sum_C P(s \mid c) P(c) f_{P(R=T \mid c)}(c)$$

S	$f_{P(W=T \mid S, R=T)}(s)$
T	0.99
F	0.9

C	S	$f_c(s)$
T	T	0.1*0.5*0.8=0.04
T	F	0.9*0.5*0.8=0.36
F	T	0.5*0.5*0.02=0.005
F	F	0.5*0.5*0.02=0.005

C	S	f_s
T	T	0.99*0.04=0.0396
T	F	0.9*0.36=0.324
F	T	0.99*0.005=0.00495
F	F	0.9*0.005=0.0045



Wet grass and rain

$$P(R=T \mid W=T) = \frac{P(R=T, W=T)}{P(W=T)}$$

$$P(R=T, W=T) = \sum_S f_c(s) f_{P(W=T \mid S, R=T)}(s)$$

$$f_{P(R=T \mid c)}(c) = P(R=T \mid c)$$

$$f_{P(W=T \mid S, R=T)}(s) = P(W=T \mid S, R=T)$$

$$f_c(s) = \sum_C P(s \mid c) P(c) f_{P(R=T \mid c)}(c)$$

$$= 0.0396 + 0.324 + 0.00495 + 0.225$$

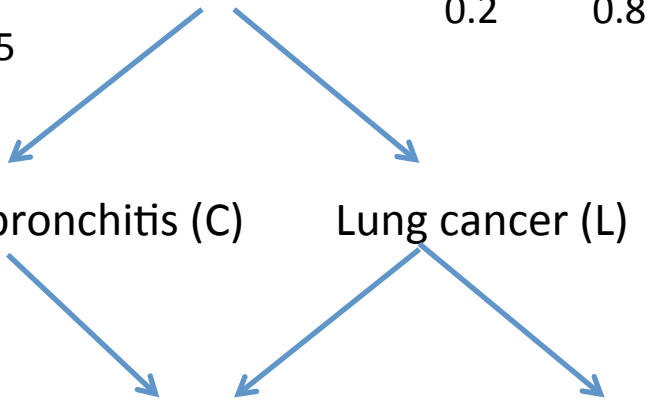
$$= 0.59355$$

C	S	f_s
T	T	0.99*0.04=0.0396
T	F	0.9*0.36=0.324
F	T	0.99*0.005=0.00495
F	F	0.9*0.005=0.0045

To get conditional probability, divide by P(W=T).

Another example

Probability of cancer for a smoker with fatigue?



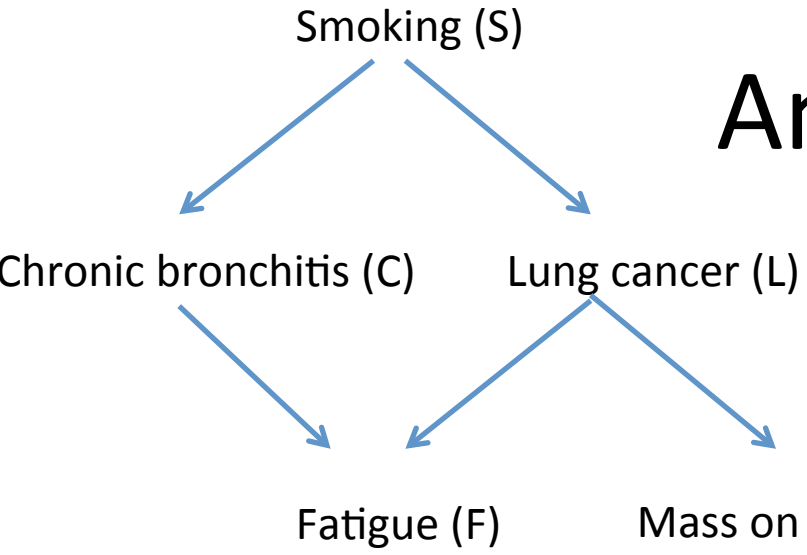
	P(C=T)	P(C=F)
S=T	0.25	0.75
S=F	0.05	0.95

P(S=T)	P(S=F)
0.2	0.8

	P(L=T)	P(L=F)
S=T	0.003	0.997
S=F	0.000005	0.99995

	P(F=T)	P(F=F)
C=T,L=T	0.75	0.25
C=T,L=F	0.10	0.90
C=F,L=T	0.50	0.50
C=F,L=F	0.05	0.95

	P(M=T)	P(M=F)
L=T	0.6	0.40
L=F	0.02	0.98



Answer(1)

$$P(L=T \mid S=T, F=T) = \frac{P(L=T, S=T, F=T)}{P(S=T, F=T)}$$

$$P(L=T, S=T, F=T) = \sum_M \sum_C P(S=T, L=T, F=T, m, c)$$

$$= \sum_M \sum_C P(S=T)P(c \mid S=T)P(L=T \mid S=T)P(F=T \mid L=T, c)P(m \mid L=T)$$

$$= f_{P(S)} f_{P(L=T \mid S=T)} \sum_C P(c \mid S=T)P(F=T \mid L=T, c)$$

$$f_{P(F=T \mid L=T, c)}(c) = P(F=T \mid L=T, c)P(c \mid S=T)$$

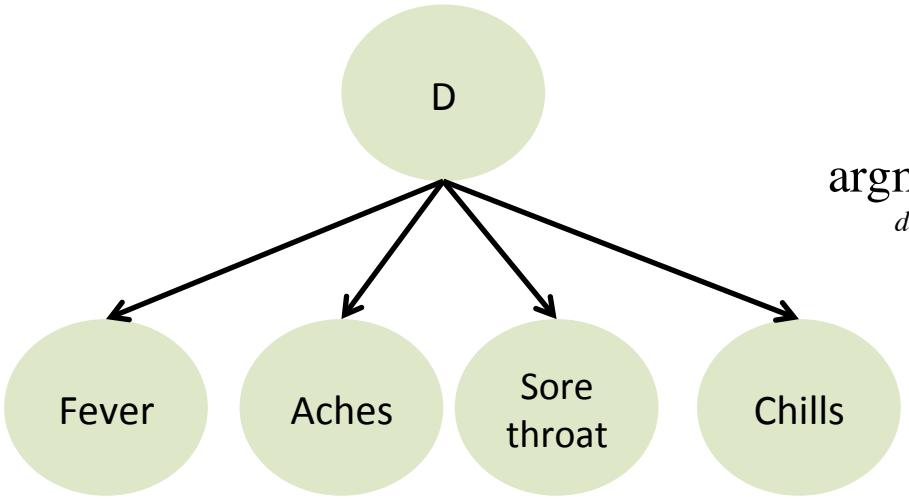
$$f_{P(S)} = P(S=T)$$

$$f_{P(L=T \mid S=T)} = P(L=T \mid S=T)$$

Inference

- Unconditional probability
  - Probability of variable(s) taking particular values
  - Probability of smoking and asbestos exposure
- How to find probability of variable(s) of interest given what is known (posterior)
  - Probability of query given evidence
  - Probability of disease | symptoms, test result
- Finding most likely value of variables given evidence
  - Diagnosis given set of symptoms

Naïve Bayes and diagnosis



What's most likely D if F=T,S=T,A=F,C=F?

$$\underset{d}{\operatorname{argmax}} \frac{P(D)P(F \mid D)P(S \mid D)P(\neg A \mid D)P(\neg C \mid D)}{P(F,S,\neg A,\neg C)}$$

Real life example: spam detection

- Class: Spam, Ham (T, type of message)
- Message comprised of words $W_1 \ldots W_n$

$$P(T \mid W_1 \ldots W_n) = \underset{T}{\operatorname{argmax}} \prod_{i=1}^{n} P(W_i \mid T) P(T)$$

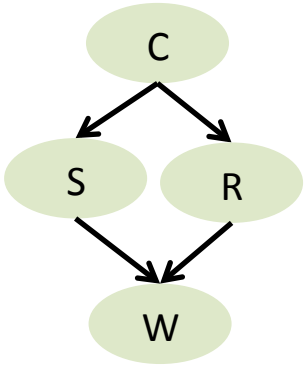
Maximum a posteriori (MAP)

- Find value assignments for variables that maximize probability of evidence

- For naïve Bayes:

$$y = \underset{y}{\operatorname{argmax}} P(Y = y \mid X_1 = x_1 \ldots X_n = x_n)$$

$$= \underset{y}{\operatorname{argmax}} P(Y = y) \prod_{i=1}^{n} (X_i = x_i \mid Y = y)$$



MAP example

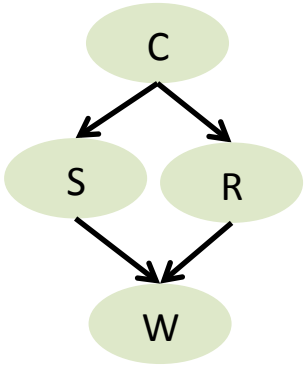
What's likeliest value for S given W=T,R=F?

$$\operatorname*{argmax}_S P(S \mid W = T, R = F)$$

$$= \operatorname*{argmax}_S P(S, W = T, R = F)$$

$$\operatorname*{argmax}_S \sum_C P(c, S, W = T, R = F)$$

$$= \sum_C P(c)P(S \mid c)P(R = F \mid c)P(W = T \mid R = F, S)$$



MAP example

What's likeliest value for S given W=T,R=F?

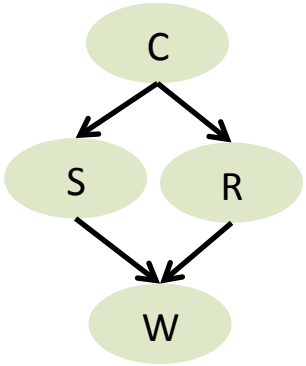
$$\operatorname*{argmax}_S \sum_C P(c,S,W=T,R=F)$$

$$= \sum_C P(c)P(S\mid c)P(R=F\mid c)P(W=T\mid R=F,S)$$

1. Evidence variables W,R

$$f_{P(R=F|c)}(c) = P(R=F\mid c)$$

$$f_{P(W=T|R=F,S)}(s) = P(W=T\mid R=F,S)$$



MAP example

What's likeliest value for S given W=T,R=F?

$$\operatorname*{argmax}_S \sum_C P(c,S,W=T,R=F)$$

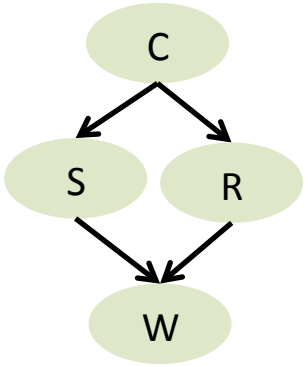
$$= \sum_C P(c)P(S \mid c)P(R=F \mid c)P(W=T \mid R=F,S)$$

$$= \sum_C P(c)P(S \mid c) f_{P(R=F|C)}(c) f_{P(W=T|R=F,S)}(s)$$

2. Sum over c

$$f_c = \sum_C P(c)P(S \mid c) f_{P(R=F|C)}(c)$$

$$\operatorname*{argmax}_S f_{P(W=T|R=F,S)}(s) f_c(s)$$



MAP example – with numbers

What's likeliest value for S given W=T,R=F?

$$\operatorname*{argmax}_S f_{P(W=T|R=F,S)}(s) f_c(s)$$

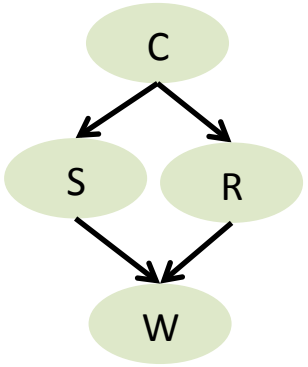
$$f_{P(R=F|c)}(c) = P(R=F \mid c)$$

$$f_{P(W=T|R=F,S)}(s) = P(W=T \mid R=F,S)$$

$$f_c = \sum_C P(c)P(S \mid c) f_{P(R=F|C)}(c)$$

	P(R=T)	P(R=F)
C=T	0.8	0.2
C=F	0.02	0.98

C	$f_{P(R=F\|c)}(c)$
T	0.2
F	0.98



MAP example – with numbers

What's likeliest value for S given W=T,R=F?

$$\operatorname*{argmax}_S f_{P(W=T|R=F,S)}(s) f_c(s)$$

$$f_{P(R=F|c)}(c) = P(R=F \mid c)$$

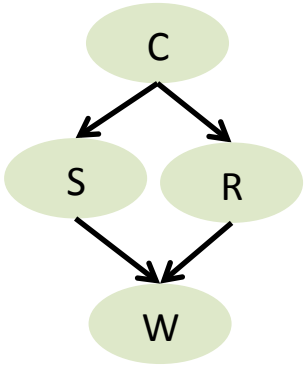
$$f_{P(W=T|R=F,s)}(s) = P(W=T \mid R=F, s)$$

$$f_c = \sum_C P(c)P(s \mid c) f_{P(R=F|C)}(c)$$

C	$f_{P(R=F\|c)}(c)$
T	0.2
F	0.98

	P(W=T)	P(W=F)
S=T,R=T	0.99	0.01
S=F,R=T	0.9	0.1
S=T,R=F	0.9	0.1
S=F,R=F	0.01	0.99

S	$f_{P(W=T\|R=F,S)}(s)$
T	0.9
F	0.01



MAP example – with numbers

What's likeliest value for S given W=T,R=F?

$$\operatorname*{argmax}_S f_{P(W=T|R=F,S)}(s) f_c(s)$$

$$f_{P(R=F|c)}(c) = P(R=F \mid c)$$

$$f_{P(W=T|R=F,S)}(s) = P(W=T \mid R=F, s)$$

$$f_c = \sum_C P(c)P(s \mid c) f_{P(R=F|C)}(c)$$

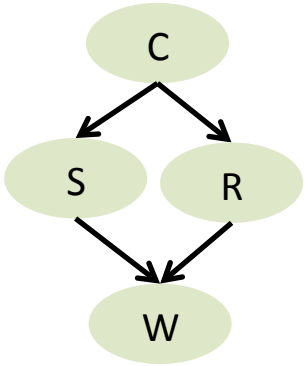
C	$f_{P(R=F\|c)}(c)$
T	0.2
F	0.98

S	$f_{P(W=T\|R=F,S)}(s)$
T	0.9
F	0.01

	P(S=T)	P(S=F)
C=T	0.1	0.9
C=F	0.5	0.5

P(C=T)	P(C=F)
0.5	0.5

C	S	fc
T	T	0.5*0.1*
T	F	0.5*0.9*
F	T	0.5*0.5*
F	F	0.5*0.5*



MAP example – with numbers

What's likeliest value for S given W=T,R=F?

$$\operatorname*{argmax}_S f_{P(W=T|R=F,S)}(s) f_c(s)$$

$$f_{P(R=F|c)}(c) = P(R=F \mid c)$$

$$f_{P(W=T|R=F,S)}(s) = P(W=T \mid R=F, s)$$

$$f_c = \sum_C P(c)P(s \mid c) f_{P(R=F|C)}(c)$$

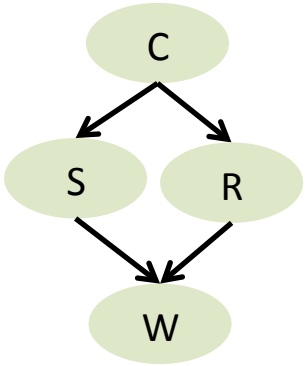
C	$f_{P(R=F\|c)}(c)$
T	0.2
F	0.98

S	$f_{P(W=T\|R=F,s)}(s)$
T	0.9
F	0.01

	P(S=T)	P(S=F)
C=T	0.1	0.9
C=F	0.5	0.5

P(C=T)	P(C=F)
0.5	0.5

C	S	fc
T	T	0.5*0.1*0.2=0.01
T	F	0.5*0.9*0.2=0.09
F	T	0.5*0.5*0.98=0.245
F	F	0.5*0.5*0.98=0.245



MAP example – with numbers

What's likeliest value for S given W=T,R=F?

$$\operatorname*{argmax}_S f_{P(W=T|R=F,S)}(s) f_c(s)$$

$$f_{P(R=F|c)}(c) = P(R=F \mid c)$$

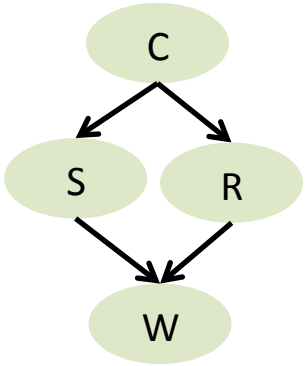
$$f_{P(W=T|R=F,S)}(s) = P(W=T \mid R=F,S)$$

$$f_c = \sum_C P(c)P(S \mid c) f_{P(R=F|C)}(c)$$

C	S	fc
T	T	0.5*0.1*0.2=0.01
T	F	0.5*0.9*0.2=0.09
F	T	0.5*0.5*0.98=0.245
F	F	0.5*0.5*0.98=0.245

C	S	fc*fpw
T	T	0.01*0.9=0.009
T	F	0.09*0.01=0.0009
F	T	0.245*0.9=0.2205
F	F	0.245*0.01=0.00245

S	$f_{P(W=T\|R=F,S)}(s)$
T	0.9
F	0.01



MAP example – with numbers

What's likeliest value for S given W=T,R=F?

$$\operatorname*{argmax}_S f_{P(W=T|R=F,S)}(s) f_c(s)$$

$$f_{P(R=F|c)}(c) = P(R=F \mid c)$$

$$f_{P(W=T|R=F,S)}(s) = P(W=T \mid R=F,S)$$

$$f_c = \sum_C P(c)P(S \mid c) f_{P(R=F|C)}(c)$$

$$S=T, P=0.2295$$

$$S=F, P=0.00335$$

C	S	fc*fpw
T	T	0.01*0.9=0.009
T	F	0.09*0.01=0.0009
F	T	0.245*0.9=0.2205
F	F	0.245*0.01=0.00245

MAP note

- What did we do?
  - Set value of evidence variables
  - Eliminated rest
  - Did all multiplication explicitly
- Complexity
  - Exponential
  - What happens with multiple variables?
- Better idea
  - Dynamic programming, similar to variable elimination

MAP/MPE

- MAP
  - Likeliest value for **a** variable(s) given some evidence
- MPE (most probable explanation)
  - Likeliest value for **all** unassigned variables given evidence

Complexity of inference

- Why can't we do exact inference in general?
  - Exponential in number of variables in each factor

- Recall:

$$f_{Xi}(X_1 \ldots X_k) = \sum_{X_i} \prod_{j=1}^{m} f_{Xi}(X_i, X_1 \ldots X_k)$$

m products (number of factors) for each possible assignment of variables

Approximate inference

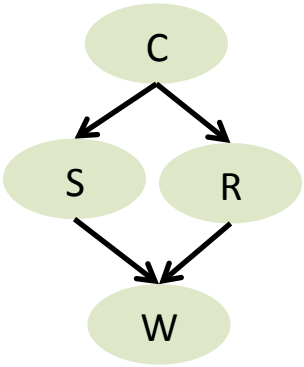
- Sampling
  - Basic idea: Generate set of samples (instantiations of graph), use to calculate probability. As sample size increases, approaches actual values.
  - Rejection sampling: Have values for some variables (evidence), ignore samples with different values
  - Importance sampling: Use evidence to weight sampling

Learning

- Parameters
- Structure

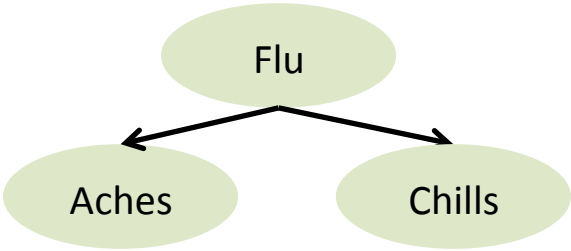
Parameter learning

- Have structure, what's CPT for each node?
- Two cases
  - Complete data (no missing variables or values)
  - Incomplete data



Parameter learning: complete data

- Three parameters
  - P(F), P(A|F) , P(C|F)

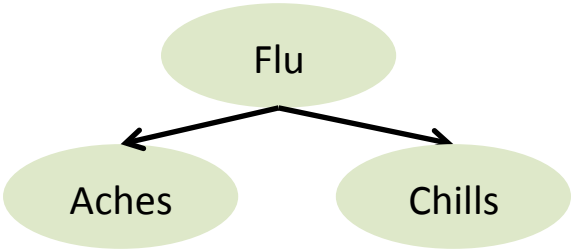


A	C	F
T	T	T
T	F	T
F	T	T
F	F	F
T	F	F
F	F	F
T	F	T

A	C	F	P
T	T	T	1/7
T	T	F	0/7
T	F	T	2/7
F	T	T	1/7
F	T	F	0/7
F	F	T	0/7
T	F	F	1/7
F	F	F	2/7

Parameter learning: complete data

- Three parameters
  - P(F), P(A|F) , P(C|F)

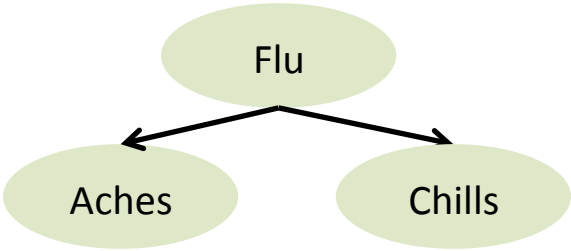


A	C	F	P
T	T	T	1/7
T	T	F	0/7
T	F	T	2/7
F	T	T	1/7
F	T	F	0/7
F	F	T	0/7
T	F	F	1/7
F	F	F	2/7

P(F=T)	P(F=F)
4/7	3/7

Parameter learning: complete data

- Three parameters
  - P(F), P(A|F) , P(C|F)



A	C	F	P
T	T	T	1/7
T	T	F	0/7
T	F	T	2/7
F	T	T	1/7
F	T	F	0/7
F	F	T	0/7
T	F	F	1/7
F	F	F	2/7

P(F=T)	P(F=F)
4/7	3/7

F	P(A=T)	P(A=F)
T	(3/7)/(4/7) =3/4	(1/7)/(4/7) =1/4
F	(1/7)/(3/7) =1/3	(2/7)/(3/7) =2/3

Structure learning approaches

- Search and score
  - Define scoring function for how well structure fits data
  - Search over set of graphs, maximizing scoring function
- Notes
  - Computational complexity

Structure learning approaches

- Constraint based
  - Use repeated conditional independence tests
  - E.g. Connect all nodes with undirected edges, repeatedly do conditional independence tests to remove/orient edges
- Notes
  - An incorrect edge removal/orientation affects later tests

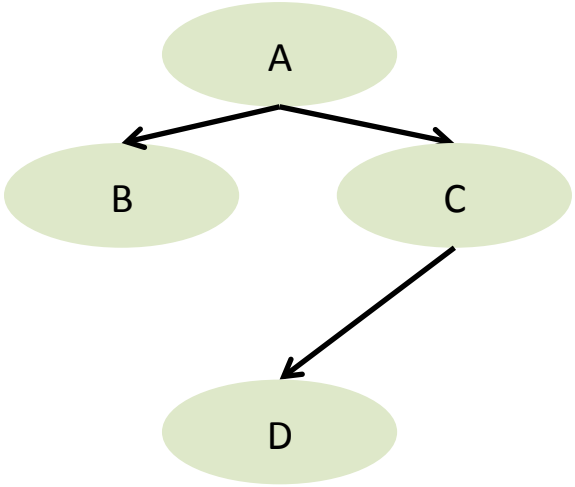
Search and score

- Define scoring function
  - Bayesian information criterion (BIC)
    - Aims to find most compact graph fitting data
    - Note minimality condition
    - $\log P(D|G) \approx \log P(D|G,\theta_G) - \log N/2 * \mathrm{Dim}(G)$

    Dim=#parameters, m=data size
- How to search over set of graphs?
- Overfitting

Heuristics

- Can't search over all graphs
- But
  - Can explore nearby graphs: add, remove, reorient edges



Greedy hillclimbing

- Apply all possible alterations, pick highest scoring change, iterate with new graph until no changes improve score
- Can get stuck in local maxima
  - But can randomly restart
  - Simulated annealing

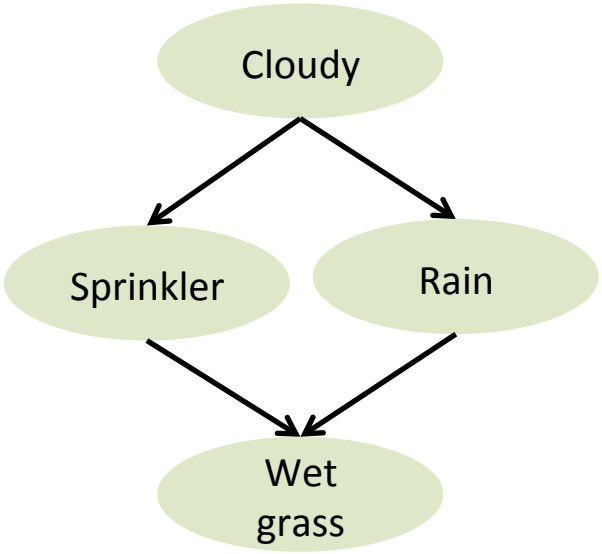
Constraint-based learning

- Find conditional independencies
  - Note: have to decide when to accept/reject Ind
  - Add edges (or remove) according to these
- Get CPT after
- More on this next week!

Software

Overview: http://www.cs.ubc.ca/~murphyk/Software/bnsoft.html

For next week



What's P(C) if I turn the sprinkler on?

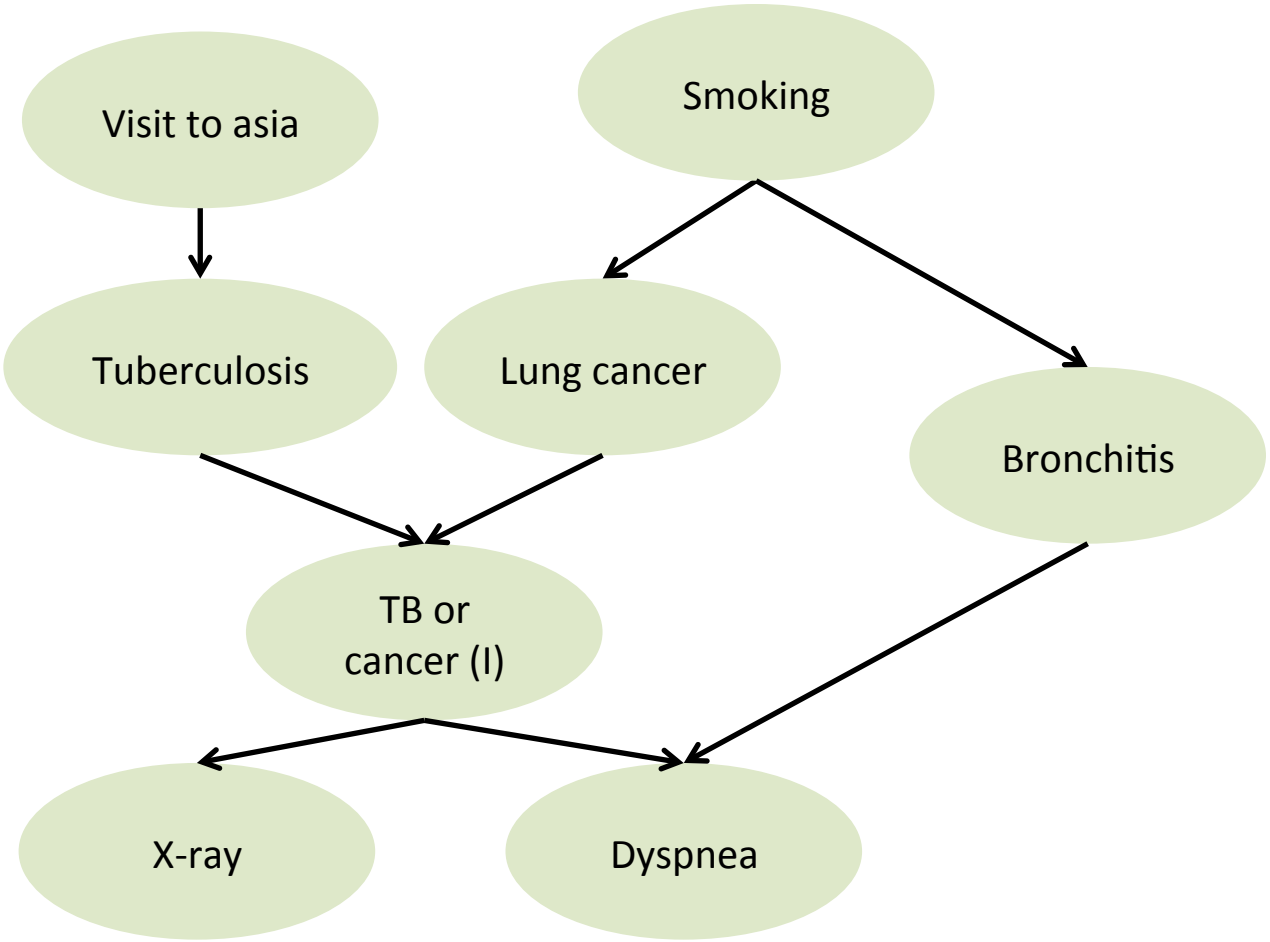
Is this the same as P(C|S=T)?

Read CPT Sections 2.4.1 and R. Scheines. An introduction to causal inference. [linked from syllabus]

Further reading on graphical models:
Koller and Friedman (2009) Probabilistic Graphical Models.
Darwiche (2009) Modeling and Reasoning with Bayesian Networks

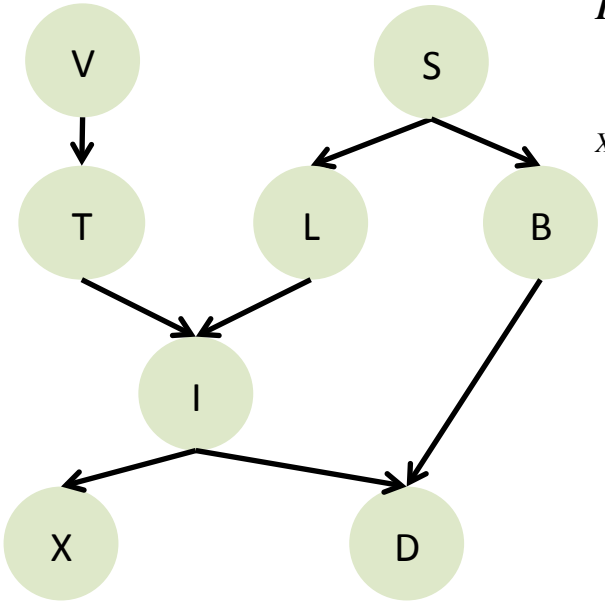
Extra slides

Another example



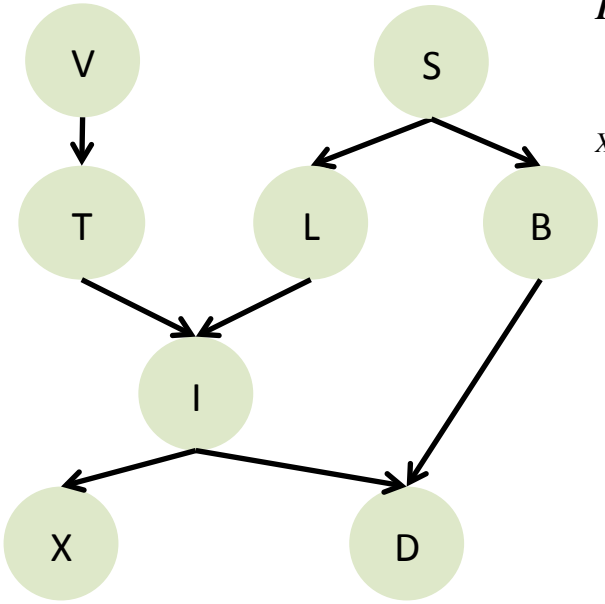
S. L. Lauritzen and D. J. Spiegelhalter (1988) Local Computations with Probabilities on Graphical Structures and Their Application to Expert Systems

Another example



$$P(d) = \sum_{X,I,L,B,T,S,V} P(s)P(v)P(t \mid v)P(l \mid s)P(b \mid s)P(i \mid t,l)P(x \mid i)P(d \mid i,b)$$

Another example



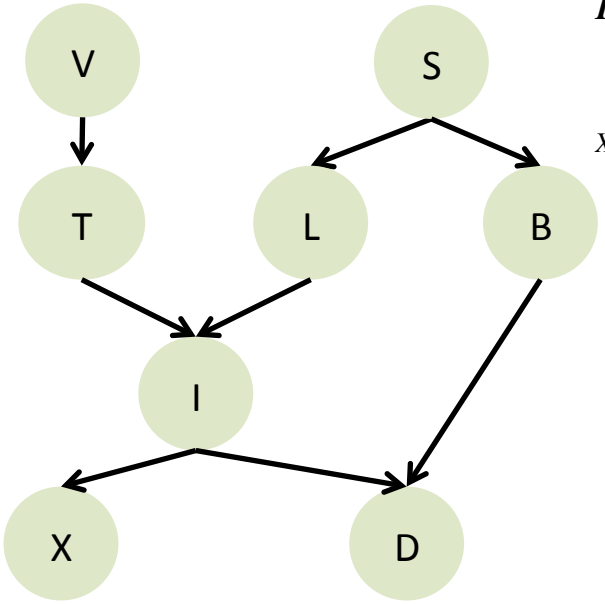
$$P(d) = \sum_{X,I,L,B,T,S,V} P(s)P(v)P(t \mid v)P(l \mid s)P(b \mid s)P(i \mid t,l)P(x \mid i)P(d \mid i,b)$$

1. Eliminate V

$$f_V(t) = \sum_V P(t \mid v)P(v)$$

$$P(d) = \sum_{X,I,L,B,T,S} P(s)f_V(t)P(l \mid s)P(b \mid s)P(i \mid t,l)P(x \mid i)P(d \mid i,b)$$

Another example



$$P(d) = \sum_{X,I,L,B,T,S,V} P(s)P(v)P(t \mid v)P(l \mid s)P(b \mid s)P(i \mid t,l)P(x \mid i)P(d \mid i,b)$$

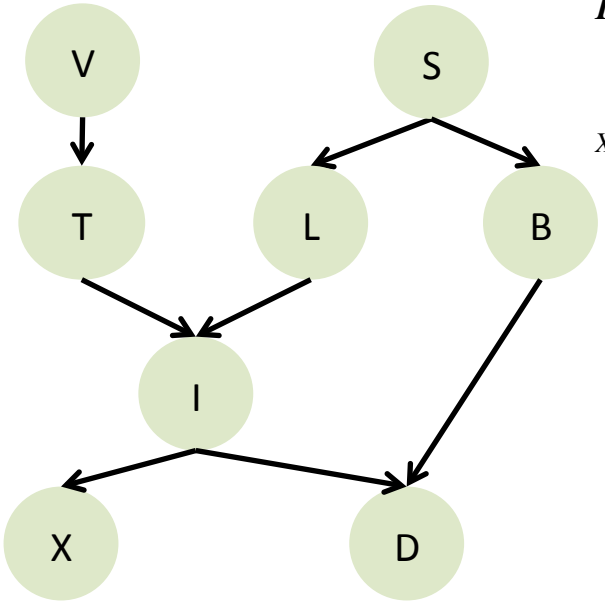
$$= \sum_{X,I,L,B,T,S} P(s)f_V(t)P(l \mid s)P(b \mid s)P(i \mid t,l)P(x \mid i)P(d \mid i,b)$$

2. Eliminate S

$$f_S(l,b) = \sum_S P(l \mid s)P(b \mid s)P(s)$$

$$= \sum_{X,I,L,B,T} f_s(l,b)f_V(t)P(i \mid t,l)P(x \mid i)P(d \mid i,b)$$

Another example



$$P(d) = \sum_{X,I,L,B,T,S,V} P(s)P(v)P(t \mid v)P(l \mid s)P(b \mid s)P(i \mid t,l)P(x \mid i)P(d \mid i,b)$$

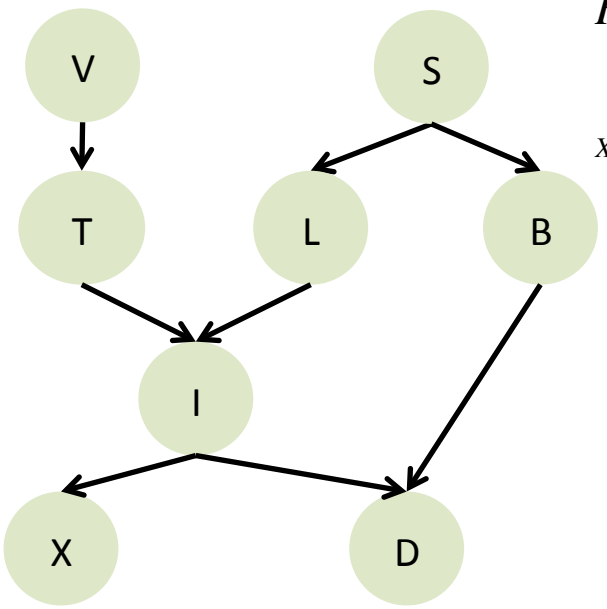
$$= \sum_{X,I,L,B,T,S} P(s)f_V(t)P(l \mid s)P(b \mid s)P(i \mid t,l)P(x \mid i)P(d \mid i,b)$$

$$= \sum_{X,I,L,B,T} f_s(l,b)f_V(t)P(i \mid t,l)P(x \mid i)P(d \mid i,b)$$

3. Eliminate T

$$f_T(i,l) = \sum_T f_v(t)P(i \mid t,l)$$

Another example



$$P(d) = \sum_{X,I,L,B,T,S,V} P(s)P(v)P(t \mid v)P(l \mid s)P(b \mid s)P(i \mid t,l)P(x \mid i)P(d \mid i,b)$$

$$= \sum_{X,I,L,B,T,S} P(s)f_V(t)P(l \mid s)P(b \mid s)P(i \mid t,l)P(x \mid i)P(d \mid i,b)$$

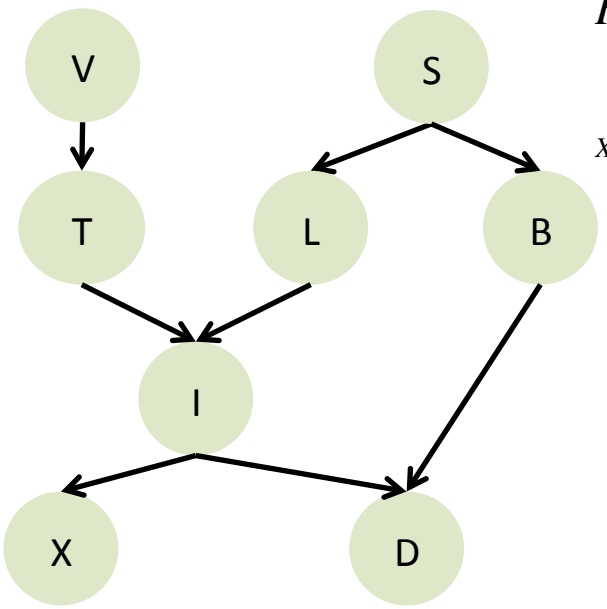
$$= \sum_{X,I,L,B,T} f_s(l,b)f_V(t)P(i \mid t,l)P(x \mid i)P(d \mid i,b)$$

$$= \sum_{X,I,L,B} f_s(l,b)f_T(i,l)P(x \mid i)P(d \mid i,b)$$

4. Eliminate B

$$f_B(d,i,l) = \sum_B f_s(l,b)P(d \mid i,b)$$

Another example



$$P(d) =$$

$$\sum_{X,I,L,B,T,S,V} P(s)P(v)P(t \mid v)P(l \mid s)P(b \mid s)P(i \mid t,l)P(x \mid i)P(d \mid i,b)$$

$$= \sum_{X,I,L,B,T,S} P(s)f_V(t)P(l \mid s)P(b \mid s)P(i \mid t,l)P(x \mid i)P(d \mid i,b)$$

$$= \sum_{X,I,L,B,T} f_s(l,b)f_V(t)P(i \mid t,l)P(x \mid i)P(d \mid i,b)$$

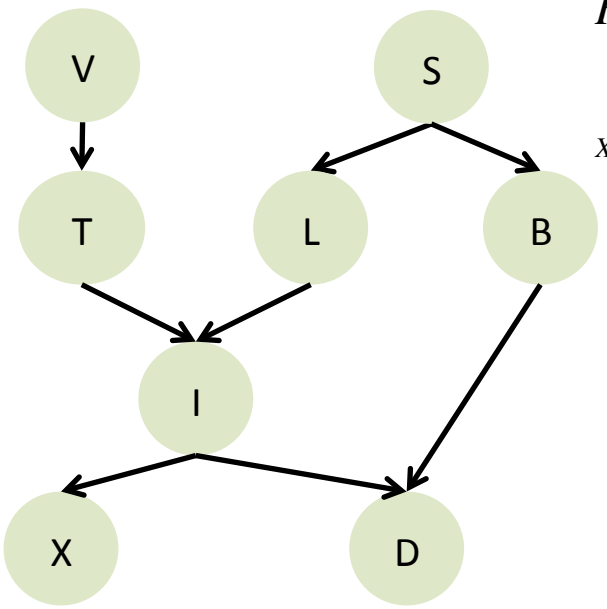
$$= \sum_{X,I,L,B} f_s(l,b)f_T(i,l)P(x \mid i)P(d \mid i,b)$$

$$= \sum_{X,I,L} f_B(l,d,i)f_T(i,l)P(x \mid i)$$

5. Eliminate L

$$f_L(i,d) = \prod_L f_B(l,d,i)f_T(i,l)$$

Another example



$$P(d) =$$

$$\sum_{X,I,L,B,T,S,V} P(s)P(v)P(t \mid v)P(l \mid s)P(b \mid s)P(i \mid t,l)P(x \mid i)P(d \mid i,b)$$

$$= \sum_{X,I,L,B,T,S} P(s)f_V(t)P(l \mid s)P(b \mid s)P(i \mid t,l)P(x \mid i)P(d \mid i,b)$$

$$= \sum_{X,I,L,B,T} f_s(l,b)f_V(t)P(i \mid t,l)P(x \mid i)P(d \mid i,b)$$

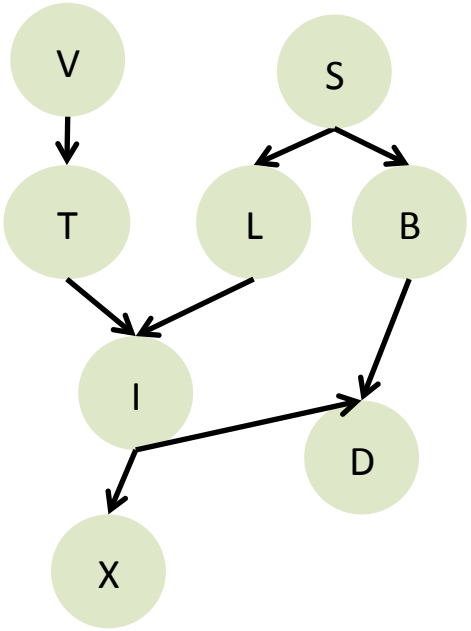
$$= \sum_{X,I,L,B} f_s(l,b)f_T(i,l)P(x \mid i)P(d \mid i,b)$$

$$= \sum_{X,I,L} f_B(l,d,i)f_T(i,l)P(x \mid i)$$

$$= \sum_{X,I} f_L(i,d)P(x \mid i)$$

6. Eliminate I

Evidence example



Evidence: asia network

$$P(x,v,d) = \sum_{I,T,L,B,S} P(s)P(v)P(t\,|\,v)P(i\,|\,t,l)P(l\,|\,s)P(b\,|\,s)P(x\,|\,i)P(d\,|\,i,b)$$

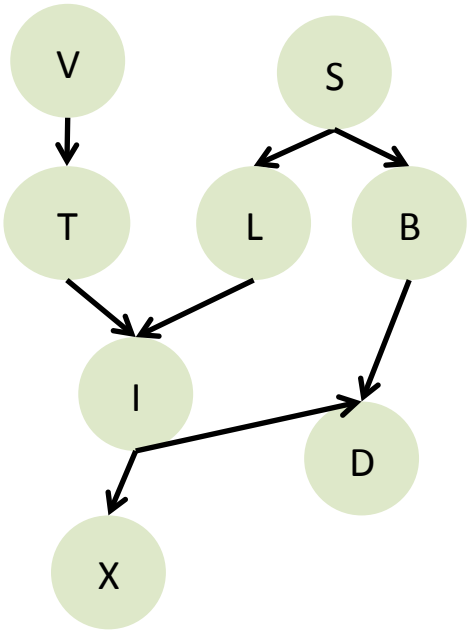
1. First deal with evidence variables: v, d

$$f_v(t) = P(t\,|\,v = T)$$

$$f_{P(V)} = P(v = T)$$

$$f_{P(d|i,b)}(i,b) = P(d = T\,|\,i,b)$$

$$P(x,v,d) = \sum_{I,T,L,B,S} P(s)f_{P(t|v)}(t)f_{P(v)}P(i\,|\,t,l)P(l\,|\,s)P(b\,|\,s)P(x\,|\,i)f_{P(d|i,b)}(i,b)$$



Evidence

$$P(x,v,d) =$$

$$\sum_{I,T,L,B,S} P(s)P(v)P(t\,|\,v)P(i\,|\,t,l)P(l\,|\,s)P(b\,|\,s)P(x\,|\,i)P(d\,|\,i,b)$$

$$P(x,v,d) =$$

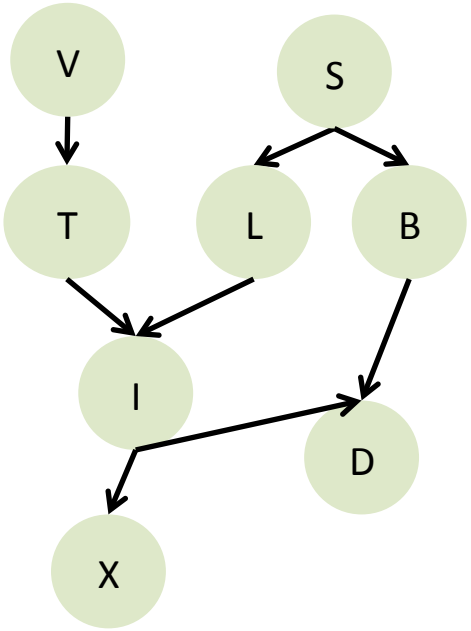
$$\sum_{I,T,L,B,S} P(s)f_{P(t|v)}(t)f_{P(v)}P(i\,|\,t,l)P(l\,|\,s)P(b\,|\,s)P(x\,|\,i)f_{P(d|i,b)}(i,b)$$

2. Eliminate S

$$f_s(b,l) = \sum_s P(l\,|\,s)P(b\,|\,s)P(s)$$

$$P(x,v,d) =$$

$$\sum_{I,T,L,B} f_s(b,l)f_{P(t|v)}(t)f_{P(v)}P(i\,|\,t,l)P(x\,|\,i)f_{P(d|i,b)}(i,b)$$



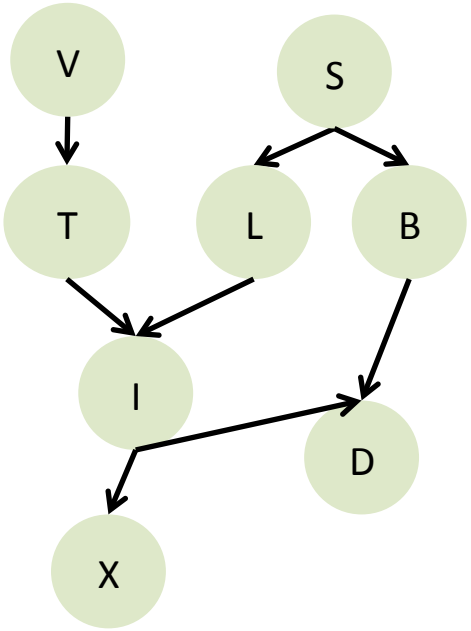
Evidence

$$P(x,v,d) = \sum_{I,T,L,B,S} f_s(b,l) f_{P(t|v)}(t) f_{P(v)} P(i \mid t,l) P(x \mid i) f_{P(d|i,b)}(i,b)$$

3. Eliminate B

$$f_B(i,l) = \sum_B f_{P(d|i,b)}(i,b) f_s(b,l)$$

$$P(x,v,d) = \sum_{I,T,L} f_B(i,l) f_{P(t|v)}(t) f_{P(v)} P(i \mid t,l) P(x \mid i)$$



Evidence

$$P(x,v,d) =$$

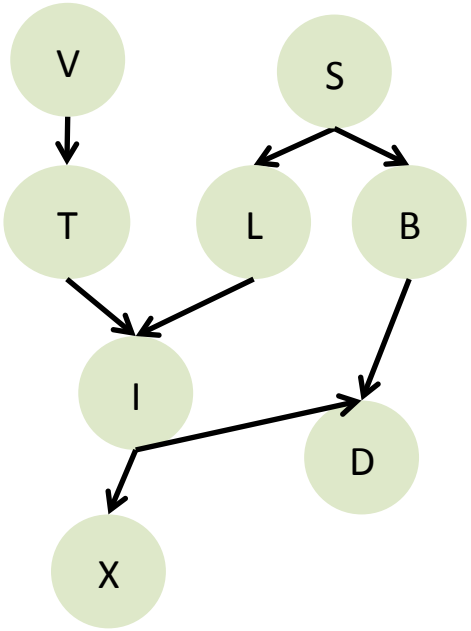
$$\sum_{I,T,L} f_B(i,l) f_{P(t|v)}(t) f_{P(v)} P(i \mid t,l) P(x \mid i)$$

4. Eliminate L

$$f_L(i,t) = \sum_{L} f_B(i,l) P(i \mid t,l)$$

$$P(x,v,d) =$$

$$\sum_{I,T} f_L(i,t) f_{P(t|v)}(t) f_{P(v)} P(x \mid i)$$

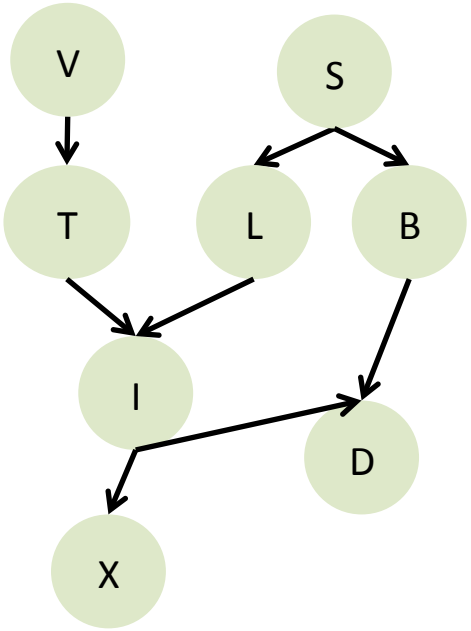


Evidence

$$P(x,v,d) = \sum_{I,T} f_L(i,t) f_{P(t|v)}(t) f_{P(v)} P(x \mid i)$$

5. Eliminate T

$$f_T(i) = \sum_{T} f_L(i,t) f_{P(t|V)}(t)$$



Evidence

$$P(x,v,d) = \sum_I f_T(i) f_{P(v)} P(x \mid i)$$

6. Finally

$$f_I(x) = \sum_I f_T(i) P(x \mid i)$$

$$P(x,v,d) = f_{P(v)} f_I(x)$$