Introduction to

Artificial Intelligence (AI)

Computer Science cpsc502, Lecture 8

Oct, 6, 2011

Slide credit Approx. Inference : S. Thrun, P, Norvig, D. Klein

Today Oct 6

- **R&R systems in Stochastic environments**
  - Bayesian Networks Representation
  - Bayesian Networks Exact Inference
  - Bayesian Networks Approx. Inference

R&Rsys we'll cover in this course

Environment

Problem		Deterministic	Stochastic
Static	Constraint Satisfaction	*Vars + Constraints* Arc Consistency SLS Search	
	Query	*Logics* → Propositional → First Order → . . . . Search	*Belief Nets* Var. Elimination Approx. Inference Temporal. Inference
Sequential	Planning	*STRIPS* actions precs effects Search	aka influence diagrams *Decision Nets* Var. Elimination *Markov Processes* Value Iteration

Representation

Reasoning Technique

Key points Recap

- We model the environment as a set of random vars

  $X_1 \ldots X_n$ JPD $P(X_1 \ldots X_n)$

- Why the joint is not an adequate representation ?

"Representation, reasoning and learning" are "exponential" in the number of variables

Solution: Exploit marginal&**conditional** independence

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

But how does independence allow us to simplify the joint?

CHAIN RULE!

Belief Nets: Burglary Example

There might be a **burglar** in my house

B

The **anti-burglar alarm** in my house may go off

A

I have an agreement with two of my neighbors, **John** and **Mary**, that they **call** me if they hear the alarm go off when I am at work

M J

Minor earthquakes may occur and sometimes they set off the alarm.

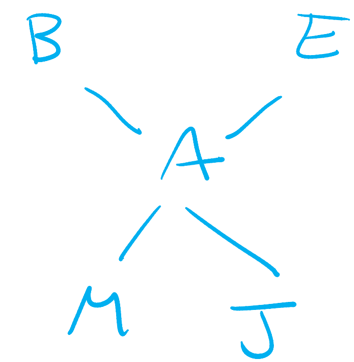
E

Variables: B A M J E $n = 5$

Joint has $2^5 - 1$ entries/probs $2^n - 1$

Belief Nets: Simplify the joint

- Typically order vars to reflect causal knowledge (i.e., causes *before effects)*
  - A burglar (B) can set the alarm (A) off
  - An earthquake (E) can set the alarm (A) off
  - The alarm can cause Mary to call (M)
  - The alarm can cause John to call (J)



$P(B, E, A, M, J)$

- Apply Chain Rule

marginal indep.

conditional indep.

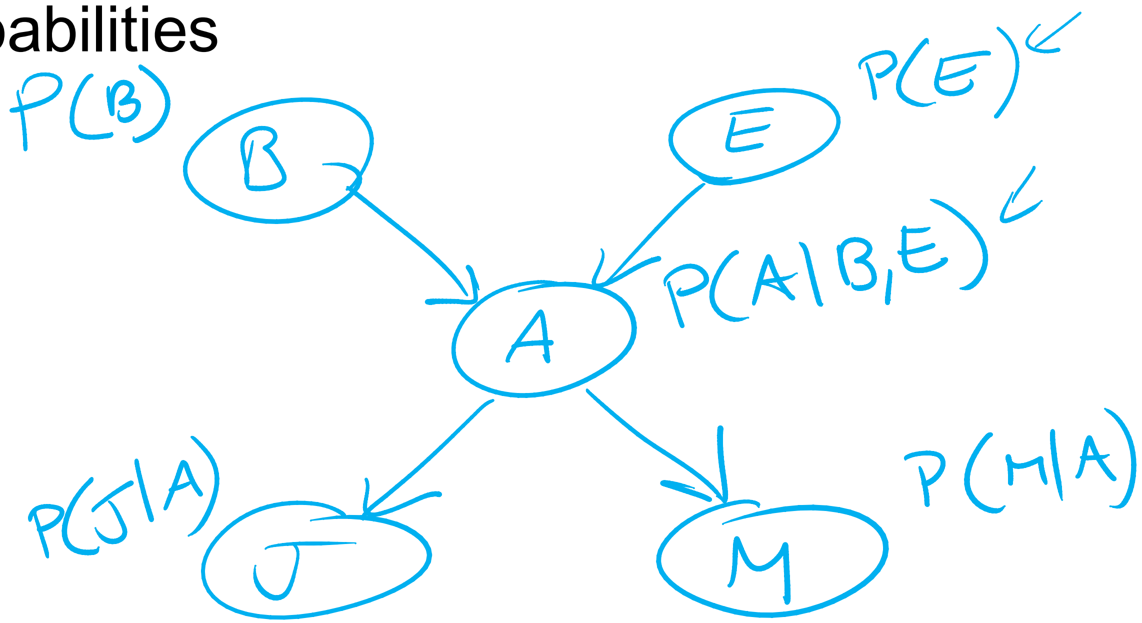
$P(B)\, P(E \mid \cancel{B})\, P(A \mid B, E)\, P(M \mid A \cancel{E} \cancel{B})\, P(J \mid \cancel{M} A \cancel{E} \cancel{B})$

- Simplify according to marginal&conditional independence

Belief Nets: Structure + Probs

$$P(B) * P(E) * P(A|B,E) * P(M|A) * P(J|A)$$

- Express remaining dependencies as a network
  - Each var is a node
  - For each var, the conditioning vars are its parents
  - Associate to each node corresponding conditional probabilities



- Directed Acyclic Graph (DAG)

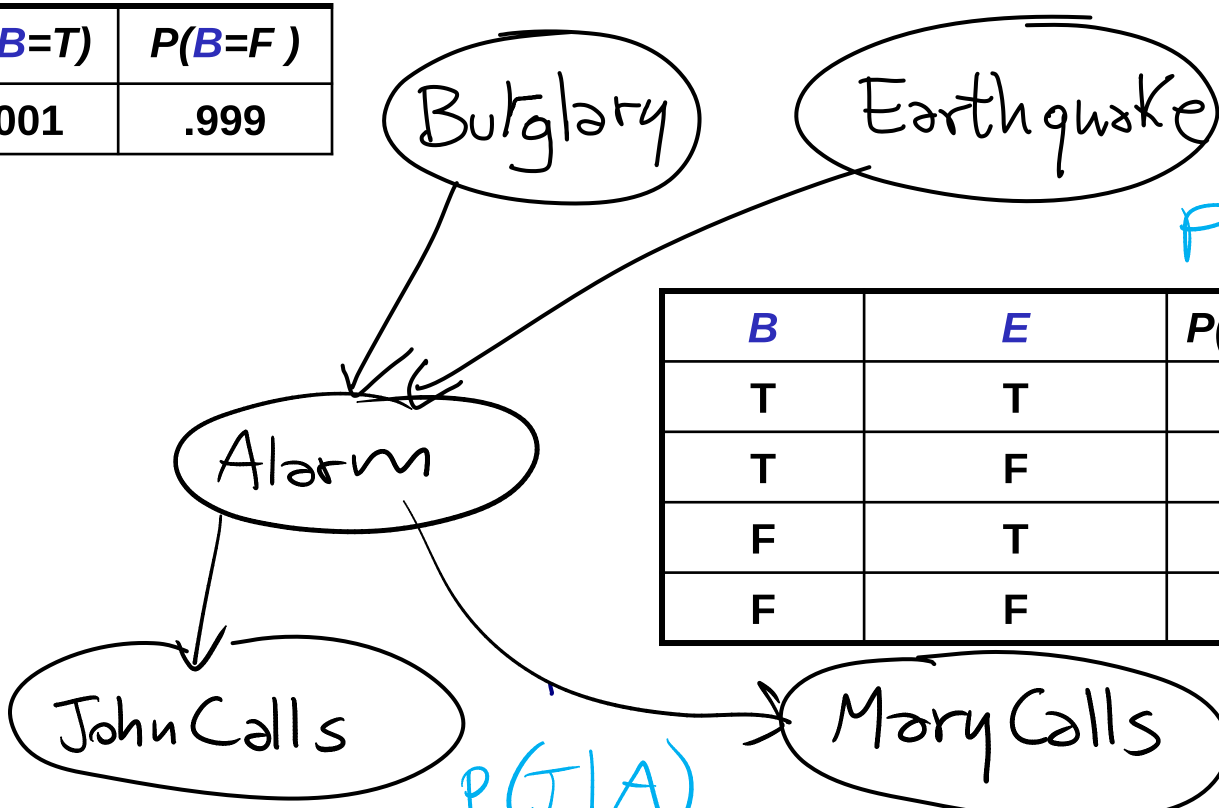
Burglary: complete BN

P(B)

P(B=T)	P(B=F)
.001	.999

P(E)

P(E=T)	P(E=F)
.002	.998



P(A | B, E)

B	E	P(A=T | B,E)	P(A=F | B,E)
T	T	.95	.05
T	F	.94	.06
F	T	.29	.71
F	F	.001	.999

P(J | A)

A	P(J=T | A)	P(J=F | A)
T	.90	.10
F	.05	.95

P(M | A)

A	P(M=T | A)	P(M=F | A)
T	.70	.30
F	.01	.99

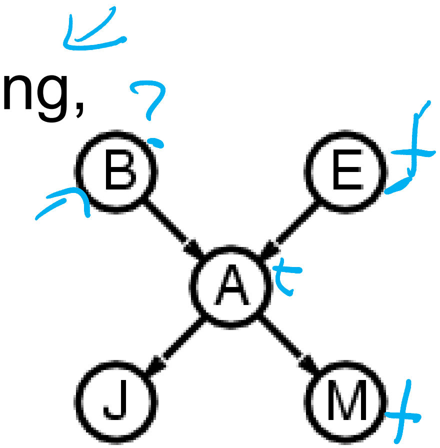
call for any other reasons

Burglary Example: Bnets inference

Our BN can answer any probabilistic query that can be answered by processing the joint!

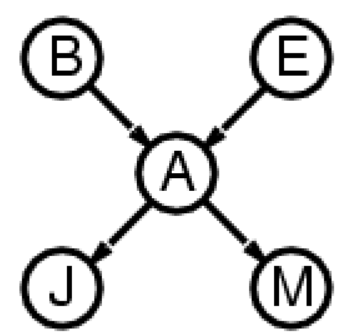
(Ex1) I'm at work,

- neighbor John calls to say my alarm is ringing,
- neighbor Mary doesn't call.
- No news of any earthquakes.
- Is there a burglar?

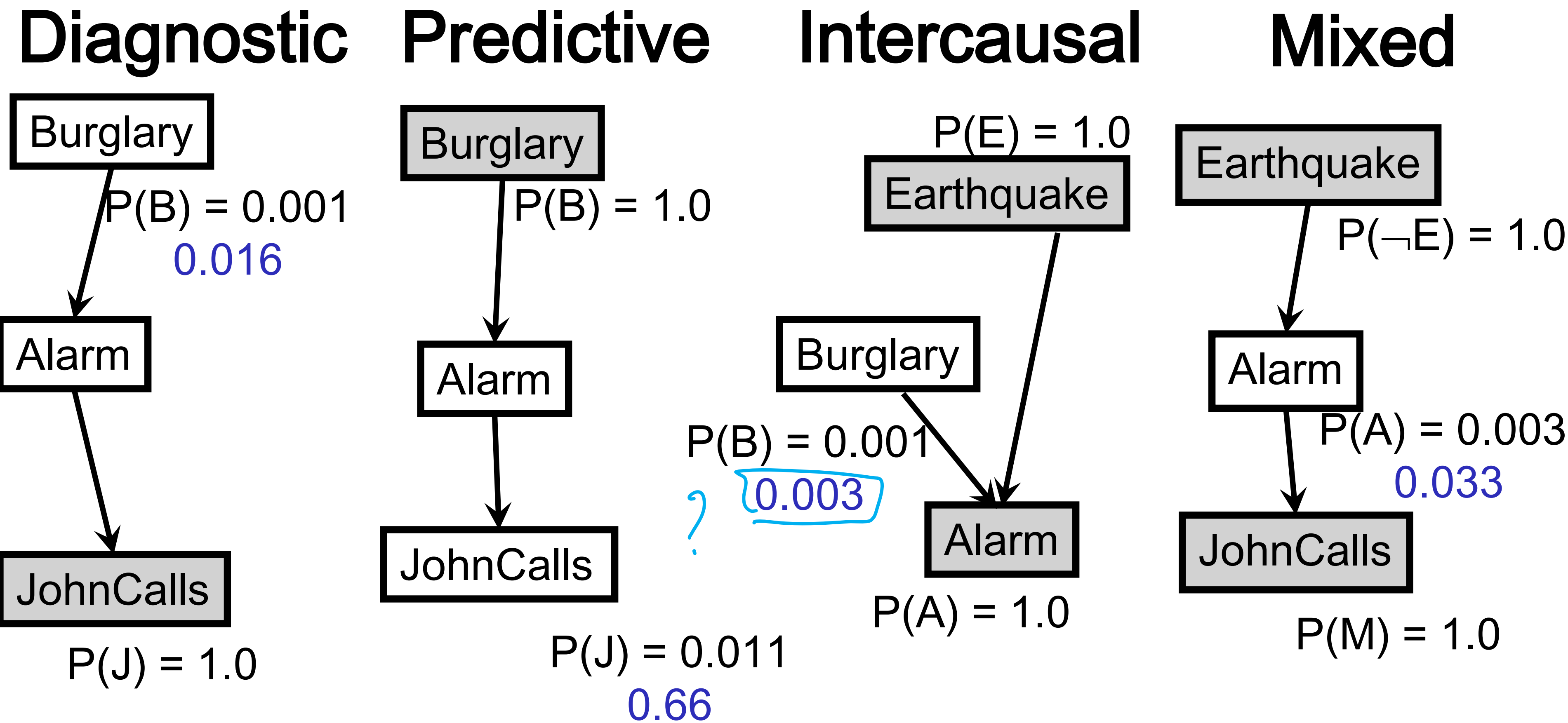


(Ex2) I'm at work,

- Receive message that neighbor John called ,
- News of minor earthquakes.
- Is there a burglar?



Bayesian Networks – Inference Types



Revised probability

BNnets: Compactness

P(B=T)	P(B=F)
.001	.999

1

Burglary

Earthquake

P(E=T)	P(E=F)
.002	.998

1

B	E	P(A=T | B,E)	P(A=F | B,E)
T	T	.95	.05
T	F	.94	.06
F	T	.29	.71
F	F	.001	.999

$2^2 = 4$

4

Alarm

JohnCalls

MaryCalls

A	P(J=T | A)	P(J=F | A)
T	.90	.10
F	.05	.95

2

2

A	P(M=T | A)	P(M=F | A)
T	.70	.30
F	.01	.99

BNet

$2+2+4+1+1=10$

$|JPD| = 2^5 - 1$

BNets: Compactness

Conditional Probability Table

In General:

A CPT for boolean X_i with k boolean parents has 2^k rows for the combinations of parent values

Each row requires one number p_i for $X_i = true$ (the number for $X_i = false$ is just $1-p_i$)

If each on the n variable has no more than k parents, the complete network requires $O(n\,2^k)$ numbers

For $k << n$, this is a substantial improvement,

- the numbers required grow linearly with n, vs. $O(2^n)$ for the full joint distribution

Realistic BNet: Liver Diagnosis

JPD

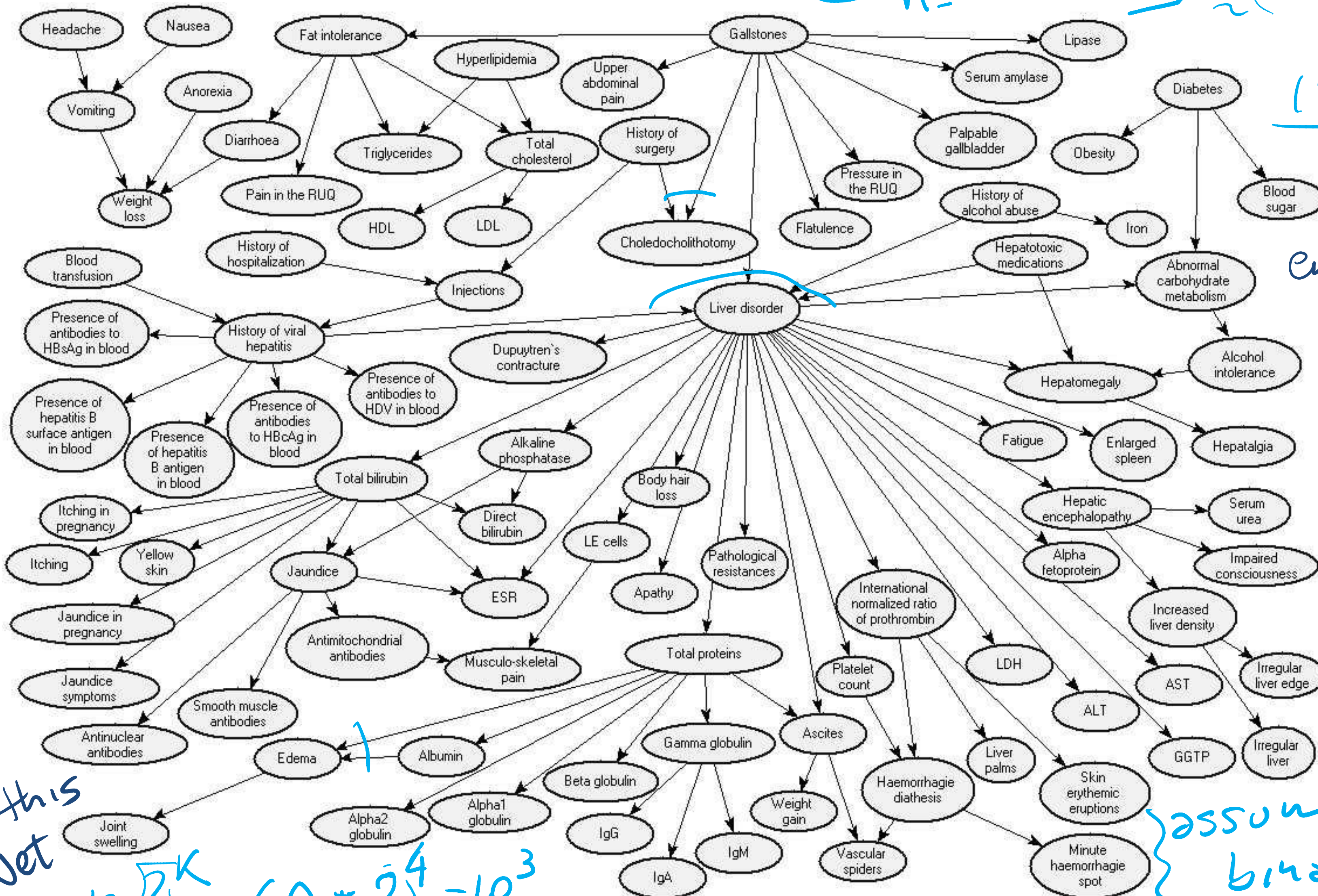
Source: Onisko et al., 1999

~60 nodes

$n = \sim 60$ $\sim 2^{60} = (2^{10})^6$

10^{18}

entries



for this BNet

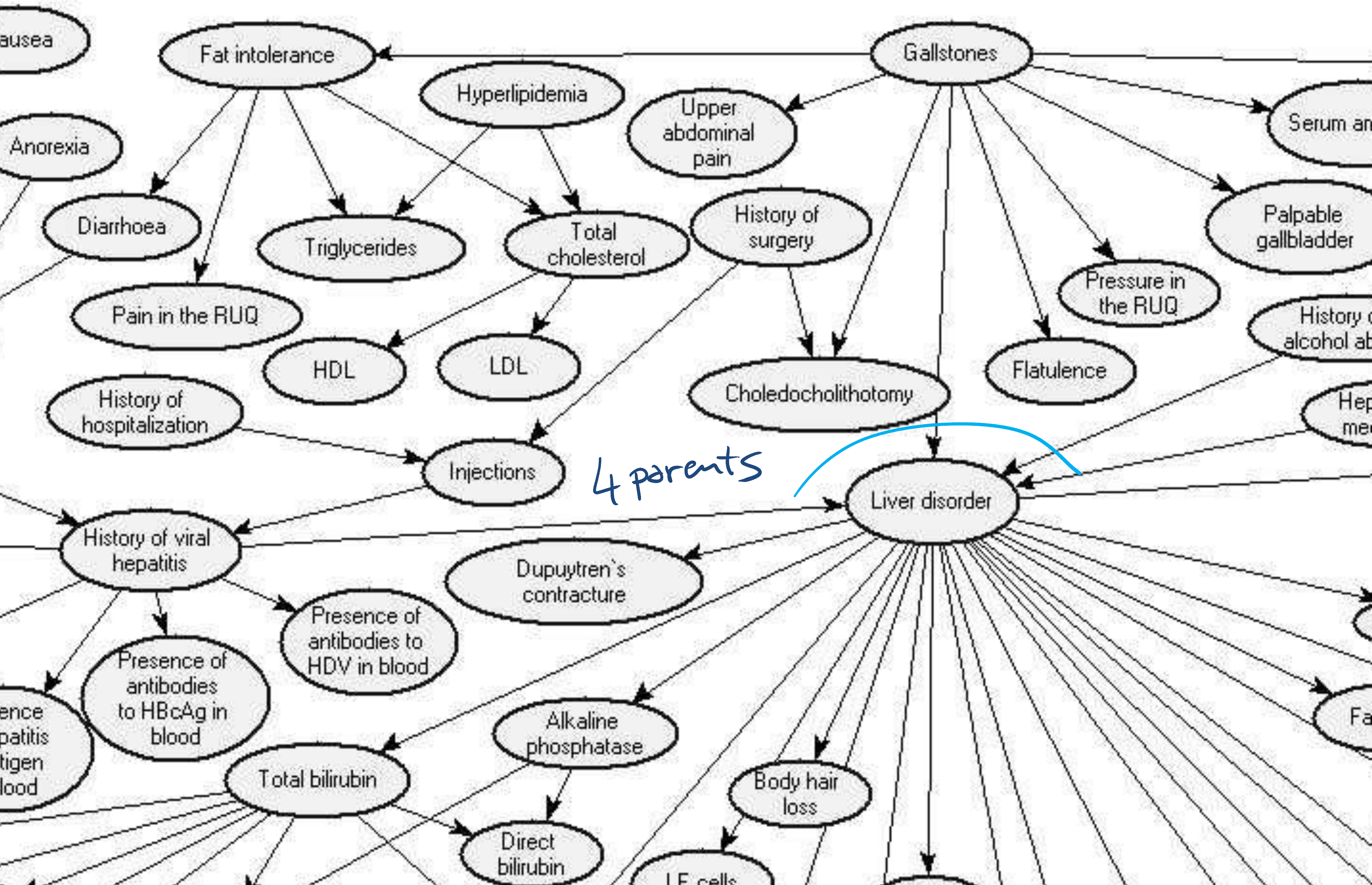
$n\,2^k$

$60 * 2^4 = 10^3$

assuming binary

Realistic BNet: Liver Diagnosis

Source: Onisko et al., 1999



BNets: Construction General Semantics

The full joint distribution can be defined as the product of conditional distributions:

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \mid X_1, \ldots, X_{i-1}) \text{ (chain rule)}$$

Simplify according to marginal&conditional independence

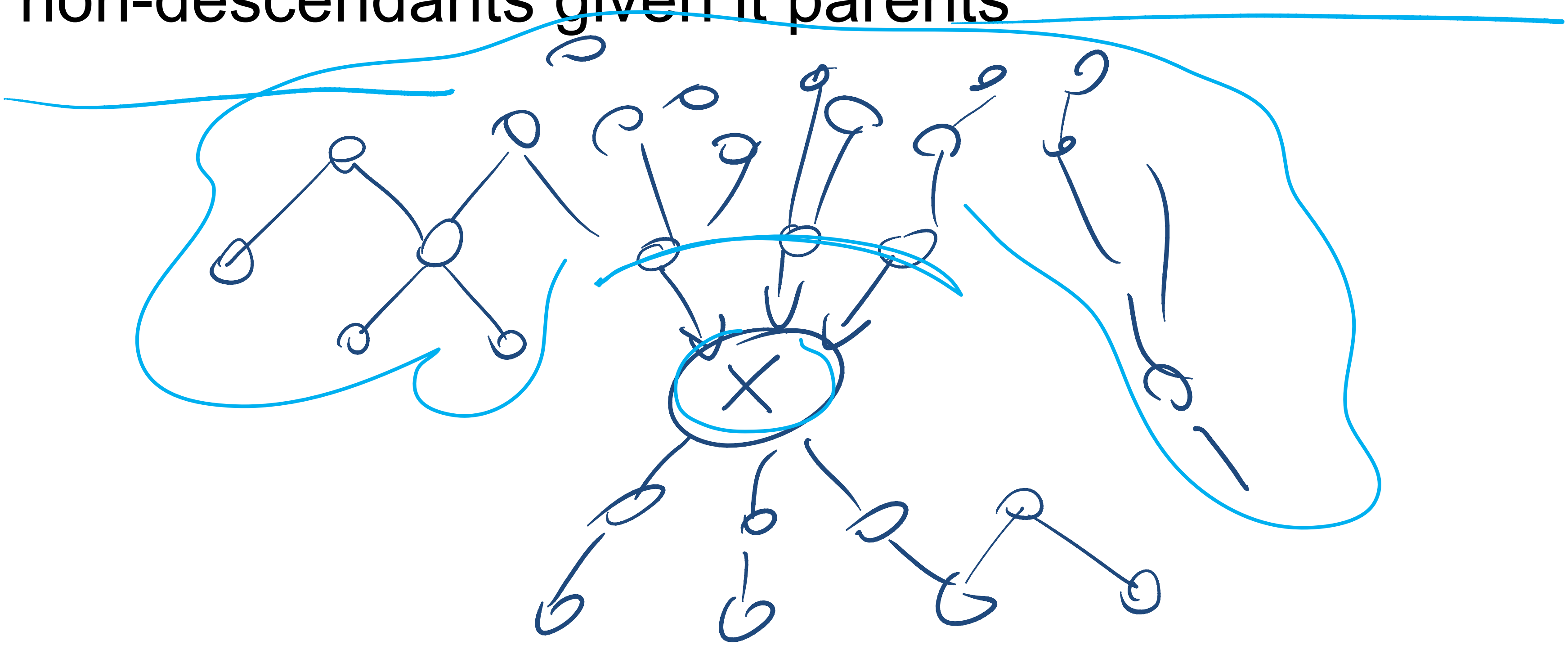
- Express remaining dependencies as a network
  - Each var is a node
  - For each var, the conditioning vars are its parents
  - Associate to each node corresponding conditional probabilities

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

BNets: Construction General Semantics (cont')

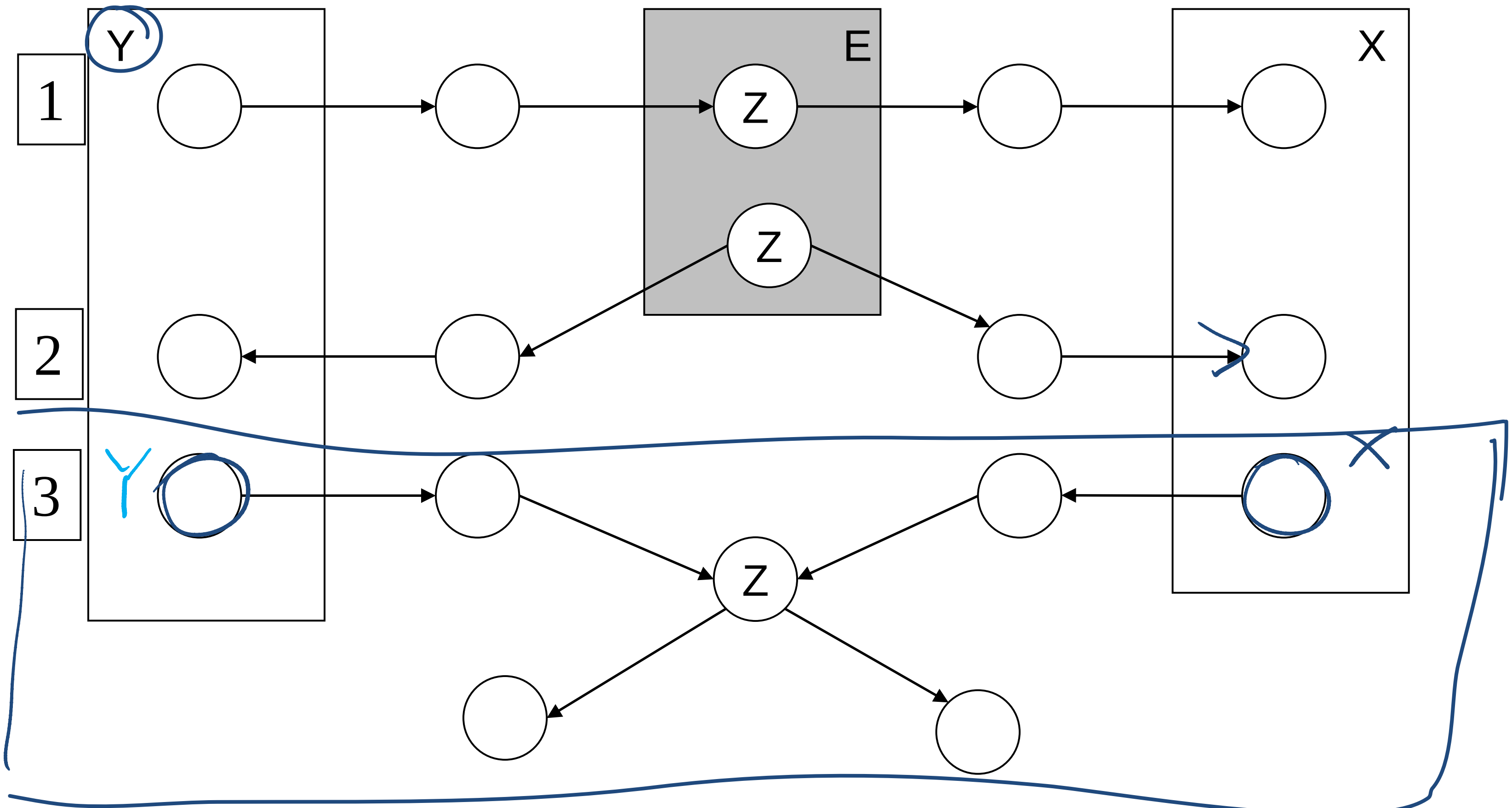
$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \mid Parents(X_i))$$

- **By construction**: Every node is independent from its non-descendants given it parents



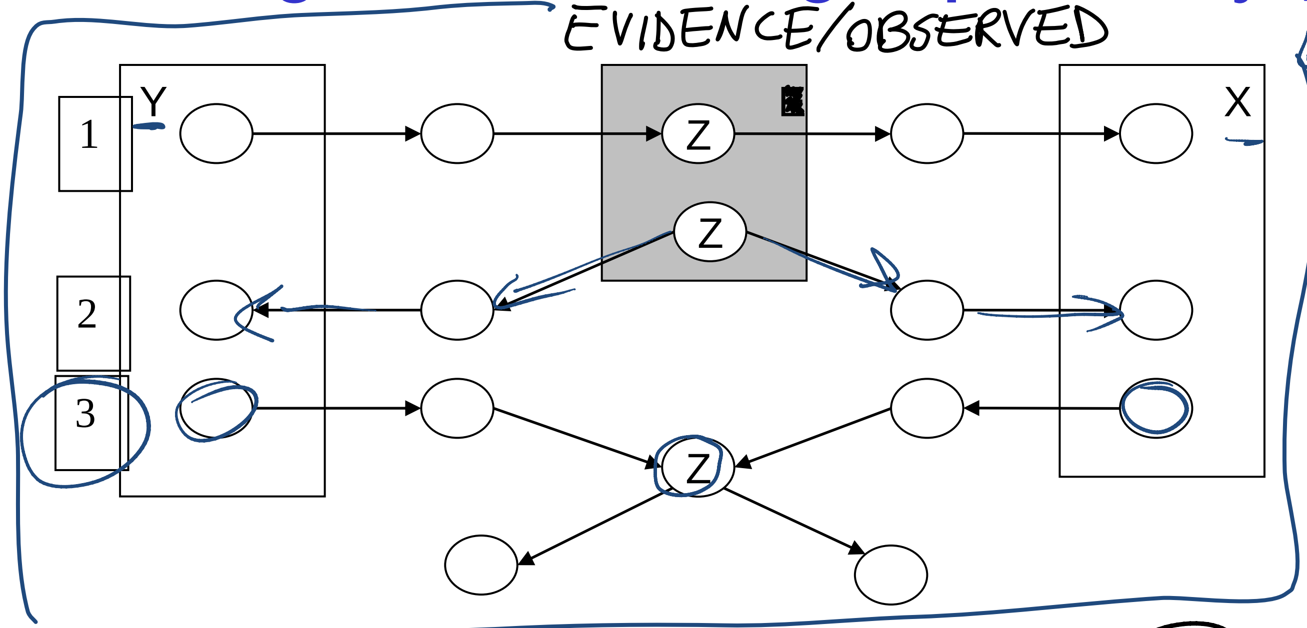
Additional Conditional Independencies

Or, blocking paths for probability propagation. Three ways in which a path between X to Y can be blocked, (1 and 2 given evidence E)



Note that, in 3, X and Y become dependent as soon as I get evidence on Z or on *any of its descendants*

3 Configuration blocking dependency (belief propagation)



true $\boxed{2}$

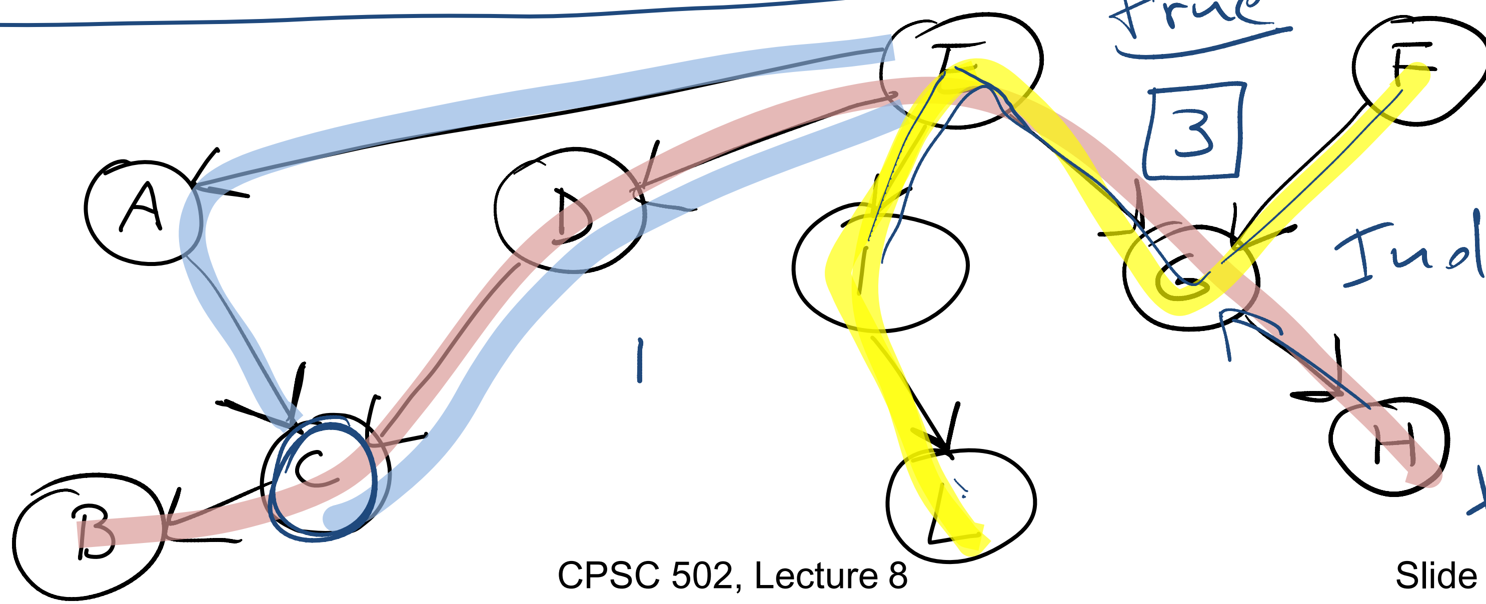
Indep(B, H {E})

false

Indep(E, C {D})

Indep(L, F {C})

true



$\boxed{3}$

G

false

Ind(L, F {H})

false

Today Oct 6

- **R&R systems in Stochastic environments**
  - Bayesian Networks Representation
  - Bayesian Networks Exact Inference
  - Bayesian Networks Approx. Inference

Bnet Inference: General

- Suppose the variables of the belief network are $\boxed{X_1, \ldots, X_n}$.
- Z is the query variable
- $Y_1 = v_1, \ldots, Y_j = v_j$ *are the observed variables (with their values)*
- $Z_1, \ldots, Z_k$ are the remaining variables
- What we want to compute: $P(Z \mid Y_1 = v_1, \ldots, Y_j = v_j)$
- We can actually compute: $\boxed{P(Z, Y_1 = v_1, \ldots, Y_j = v_j)}$

$$P(Z \mid Y_1 = v_1, \ldots, Y_j = v_j) = \frac{P(Z, Y_1 = v_1, \ldots, Y_j = v_j)}{\boxed{P(Y_1 = v_1, \ldots, Y_j = v_j)}} = \frac{P(Z, Y_1 = v_1, \ldots, Y_j = v_j)}{\sum_Z P(Z, Y_1 = v_1, \ldots, Y_j = v_j)}$$

What do we need to compute?

Remember conditioning and marginalization…

$$P(L \mid S = t, R = f) = \frac{P(L, S=t, R=f)}{P(S=t, R=f)}$$

1, 2, 3

L	*S*	*R*	*P(L, S=t, R=f)*
t	t	f	.3
f	t	f	.2

Do they have to sum up to one?

no

2 = .5

3

L	*S*	*R*	*P(L | S=t, R=f)*
t	t	f	.6
f	t	f	.4

Variable Elimination Intro

- Suppose the variables of the belief network are $X_1,\ldots,X_n$.
- Z is the query variable
- $Y_1=v_1, \ldots, Y_j=v_j$ *are the observed variables (with their values)*
- $Z_1, \ldots, Z_k$ are the remaining variables
- What we want to compute: $\boxed{P(Z \mid Y_1 = v_1, \ldots, Y_j = v_j)}$
- We just showed before that what we actually need to compute is

$$\boxed{P(Z, Y_1 = v_1, \ldots, Y_j = v_j)}$$

This can be computed in terms of **operations between factors** (that satisfy the semantics of probability)

Factors

- A **factor** is a representation of a function from a tuple of random variables into a number. $[0, 1]$
- We will write factor f on variables $X_1, \ldots, X_j$ as $f(X_1 \ldots X_j)$
- A factor denotes one or more (possibly partial) distributions over the given tuple of variables

$P(Z \mid X\,Y)$

- e.g., $P(X_1, X_2)$ is a factor $f(X_1, X_2)$ *Distribution*
- e.g., $P(X_1, X_2, X_3 = v_3)$ is a factor $f(X_1, X_2)_{X3 = v3}$ *Partial distribution*
- *e.g., $P(Z \mid X, Y)$ is a factor $f(Z,X,Y)$* *Set of Distributions*
- e.g., $P(X_1, X_3 = v_3 \mid X_2)$ is a factor $f(X_1, X_2)_{X3 = v3}$ *Set of partial Distributions*

f(X,Y,Z)

X	Y	Z	val
t	t	t	0.1
t	t	f	0.9
t	f	t	0.2
t	f	f	0.8
f	t	t	0.4
f	t	f	0.6
f	f	t	0.3
f	f	f	0.7

Manipulating Factors:

We can make new factors out of an existing factor

- Our first operation: we can *assign* some or all of the variables of a factor.

f(X,Y,Z):

X	Y	Z	val
t	t	t	0.1
t	t	f	0.9
t	f	t	0.2
t	f	f	0.8
~~f~~	~~t~~	~~t~~	~~0.4~~
~~f~~	~~t~~	~~f~~	~~0.6~~
~~f~~	~~f~~	~~t~~	~~0.3~~
~~f~~	~~f~~	~~f~~	~~0.7~~

What is the result of assigning X= t ?

$f(X=t,Y,Z)$

$f(X, Y, Z)_{X = t}$

Summing out a variable example

Our second operation: we can ***sum out*** a variable, say X_1 with domain $\{v_1, \ldots, v_k\}$, from factor $f(X_1, \ldots, X_j)$, resulting in a factor on $X_2, \ldots, X_j$ defined by:

$f_3(B,A,C)$:

B	A	C	val
t	t	t	0.03
t	t	f	0.07
f	t	t	0.54
f	t	f	0.36
t	f	t	0.06
t	f	f	0.14
f	f	t	0.48
f	f	f	0.32

$\sum_B f_3(A,C)$:

A	C	val
t	t	.57
t	f	.43
f	t	
f	f	

$$\left(\sum_{X_1} f\right)(X_2, \ldots, X_j) = f(X_1 = v_1, X_2, \ldots, X_j) + \ldots + f(X_1 = v_k, X_2, \ldots, X_j)$$

Multiplying factors

•Our third operation: factors can be *multiplied* together.

$f_1(A,B)$:

A	B	Val
t	t	0.1
t	f	0.9
f	t	0.2
f	f	0.8

$f_2(B,C)$:

B	C	Val
t	t	0.3
t	f	0.7
f	t	0.6
f	f	0.4

$f_1(A,B) \times f_2(B,C)$:

A	B	C	val
t	t	t	.03
t	t	f	.07
t	f	t	.054
t	f	f	
f	t	t	
f	t	f	
f	f	t	
f	f	f	

Factors Summary

- A factor is a representation of a function from a tuple of random variables into a number.
  - $f(X_1, \ldots, X_j)$.
- We have defined three operations on factors:
  1. Assigning one or more variables
     - $f(X_1=v_1, X_2, \ldots, X_j)$ is a factor on $X_2, \ldots, X_j$, also written as $f(X_1, \ldots, X_j)_{X_1=v_1}$
  2. Summing out variables
     - $(\sum_{X_1} f)(X_2, .., X_j) = f(X_1=v_1, X_2, ,X_j) + \ldots + f(X_1=v_k, X_2, ,X_j)$
  3. Multiplying factors
     - $f_1(A, B)\ f_2\ (B, C) = (f_1 \times f_2)(A, B, C)$

Variable Elimination Intro

- If we express the joint as a factor,

sum out

$$f(Z, \overbrace{Y_1,\dots,Y_j}^{\text{observed}}, Z_1,\dots,Z_j)$$

assign

- We can compute $P(Z, Y_1{=}v_1, \dots, Y_j{=}v_j)$ by ??
    - **assigning** $Y_1{=}v_1, \dots, Y_j{=}v_j$
    - and **summing out** the variables $Z_1, \dots, Z_k$

$$P(Z, Y_1 = v_1, \dots, Y_j = v_j) = \boxed{\sum_{Z_k} \cdots \sum_{Z_1}} f(Z, Y_1, .., Y_j, Z_1, .., Z_k)_{Y_1 = v_1, \dots, Y_j = v_j}$$

this is the joint TOO BIG!

Are we done? no

Variable Elimination Intro (1)

$$P(Z, Y_1 = v_1, \ldots, Y_j = v_j) = \sum_{Z_k} \cdots \sum_{Z_1} f(Z, Y_1, .., Y_j, Z_1, .., Z_k)_{Y_1 = v_1, \ldots, Y_j = v_j}$$

- Using the chain rule and the definition of a Bnet, we can write *P(X₁, …, Xₙ)* as $\prod_{i=1}^{n} P(X_i \mid pX_i)$

n

- We can express the joint factor as a product of factors

$f(Z, Y_1 \ldots, Y_j, Z_1 \ldots, Z_j)$ $\qquad \prod_{i=1}^{n} f(X_i, pX_i)$

$$P(Z, Y_1 = v_1, \ldots, Y_j = v_j) = \sum_{Z_k} \cdots \sum_{Z_1} \prod_{i=1}^{n} f(X_i, pX_i)_{Y_1 = v_1, \ldots, Y_j = v_j}$$

Variable Elimination Intro (2)

Inference in belief networks thus reduces to computing "the sums of products…."

$$\boxed{P(Z, Y_1 = v_1, \ldots, Y_j = v_j)} = \underbrace{\sum_{Z_k}^{4} \cdots \sum_{Z_1} \prod_{i=1}^{n}{}^{3} \underline{f(X_i, pX_i)}^{1} \;\; {}^{2}_{Y_1 = v_1, \ldots, Y_j = v_j}}$$

1. Construct a factor for each conditional probability.
2. In each factor assign the observed variables to their observed values.
3. Multiply the factors
4. For each of the other variables $Z_i \in \{Z_1, \ldots, Z_k\}$, sum out Z_i

Key Simplification Step

$$P(G,D=t) = \sum_{A,B,C,} f(A,G)\, f(B,A)\, f(C,G)\, f(B,C)$$

$$P(G,D=t) = \sum_A f(A,G) \sum_B f(B,A) \sum_C f(C,G)\, f(B,C)$$

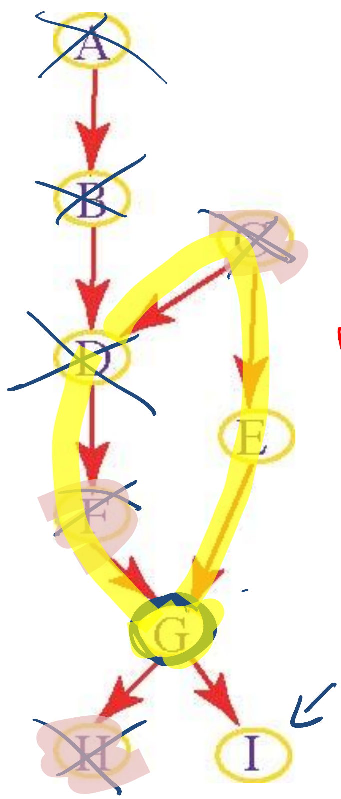
$f(\cancel{C}\ B\ G)$

I will add to the online slides a complete example of VE

ENDED HERE

Another Simplification before starting VE

- All the variables from which the query is conditional independent given the observations can be pruned from the Bnet



e.g., $P(G \mid H=v_1, F=v_2, C=v_3)$.

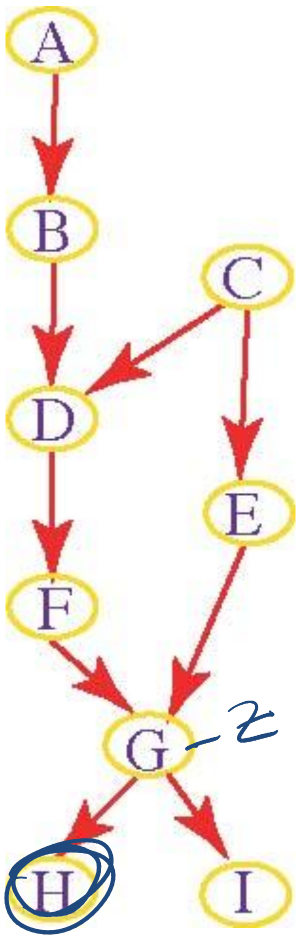
both paths from G to D are blocked

G is conditionally independent from A B D given the observed vars H, F, C

Variable elimination example

Compute $P(G \mid H{=}h_1)$.

- $P(G,H) = \sum_{A,B,C,D,E,F,I} P(A,B,C,D,E,F,G,H,I)$



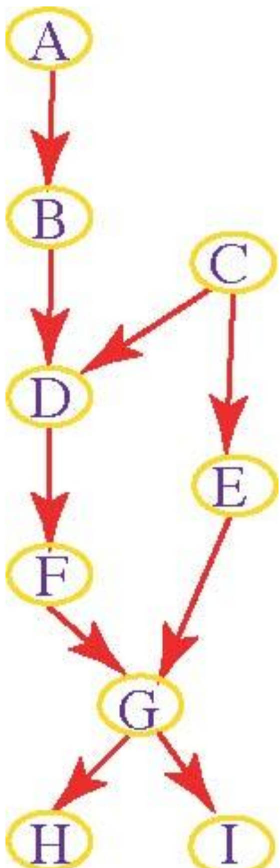
Variable elimination example

Compute $P(G \mid H=h_1)$.

- $P(G,H) = \sum_{A,B,C,D,E,F,I} P(A,B,C,D,E,F,G,H,I)$

Chain Rule + Conditional Independence:

$P(G,H) = \sum_{A,B,C,D,E,F,I} P(A)P(B|A)P(C)P(D|B,C)P(E|C)P(F|D)P(G|F,E)P(H|G)P(I|G)$



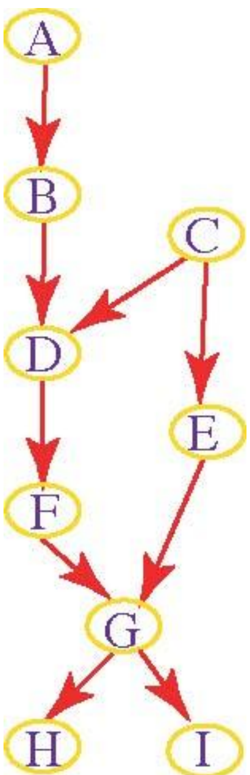
Variable elimination example (step1)

Compute $P(G \mid H=h_1)$.

- $P(G,H) = \sum_{A,B,C,D,E,F,I} P(A)P(B|A)P(C)P(D|B,C)P(E|C)P(F|D)P(G|F,E)P(H|G)P(I|G)$

Factorized Representation:

$P(G,H) = \sum_{A,B,C,D,E,F,I} f_0(A)\ f_1(B,A)\ f_2(C)\ f_3(D,B,C)\ f_4(E,C)\ f_5(F, D)\ f_6(G,F,E)\ f_7(H,G)\ f_8(I,G)$



- $f_0(A)$
- $f_1(B,A)$
- $f_2(C)$
- $f_3(D,B,C)$
- $f_4(E,C)$
- $f_5(F, D)$
- $f_6(G,F,E)$
- $f_7(H,G)$
- $f_8(I,G)$

Variable elimination example (step 2)

Compute $P(G \mid H{=}h_1)$.

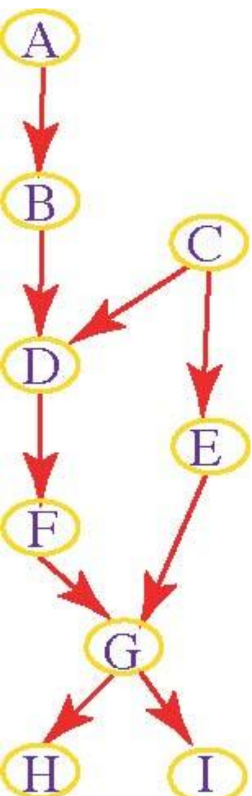
Previous state:

$$P(G,H) = \sum_{A,B,C,D,E,F,I} f_0(A)\, f_1(B,A)\, f_2(C)\, f_3(D,B,C)\, f_4(E,C)\, f_5(F,D)\, f_6(G,F,E)\, \underline{f_7(H,G)}\, f_8(I,G)$$

Observe H :

$$P(G,H{=}h_1) = \sum_{A,B,C,D,E,F,I} f_0(A)\, f_1(B,A)\, f_2(C)\, f_3(D,B,C)\, f_4(E,C)\, f_5(F,D)\, f_6(G,F,E)\, \underline{f_9(G)}\, f_8(I,G)$$

New factor



- $f_0(A)$
- $f_1(B,A)$
- $f_2(C)$
- $f_3(D,B,C)$
- $f_4(E,C)$
- $f_5(F, D)$
- $f_6(G,F,E)$
- $f_7(H,G)$
- $f_8(I,G)$

- $f_9(G)$

Variable elimination example (steps 3-4)

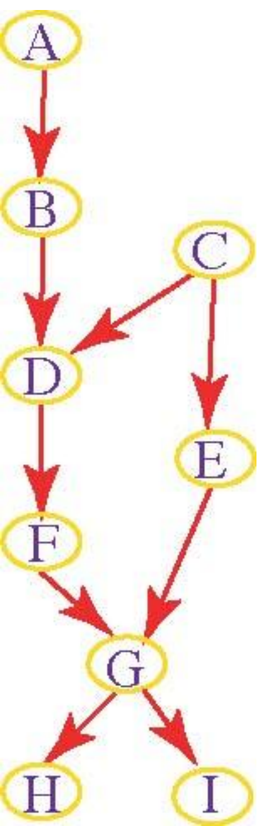
Compute $P(G \mid H=h_1)$.

Previous state:

$$P(G,H) = \sum_{A,B,C,D,E,F,I} f_0(A)\, f_1(B,A)\, f_2(C)\, f_3(D,B,C)\, f_4(E,C)\, f_5(F, D)\, f_6(G,F,E)\, f_9(G)\, f_8(I,G)$$

Elimination ordering A, C, E, I, B, D, F :

$$P(G,H=h_1) = f_9(G) \sum_F \sum_D f_5(F, D) \sum_B \sum_I f_8(I,G) \sum_E f_6(G,F,E) \sum_C f_2(C)\, f_3(D,B,C)\, f_4(E,C) \boxed{\sum_A f_0(A)\, f_1(B,A)}$$



- $f_0(A)$
- $f_1(B,A)$
- $f_2(C)$
- $f_3(D,B,C)$
- $f_4(E,C)$
- $f_5(F, D)$
- $f_6(G,F,E)$
- $f_7(H,G)$
- $f_8(I,G)$
- $f_9(G)$

Variable elimination example(steps 3-4)

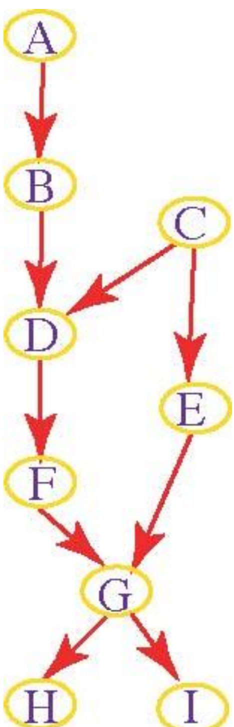
Compute $P(G \mid H=h_1)$. Elimination ordering A, C, E, I, B, D, F.

Previous state:

$$P(G,H=h_1) = f_9(G) \sum_F \sum_D f_5(F,D) \sum_B \sum_I f_8(I,G) \sum_E f_6(G,F,E) \sum_C f_2(C) f_3(D,B,C) f_4(E,C) \sum_A f_0(A) f_1(B,A)$$

Eliminate A:

$$P(G,H=h_1) = f_9(G) \sum_F \sum_D f_5(F,D) \sum_B f_{10}(B) \sum_I f_8(I,G) \sum_E f_6(G,F,E) \sum_C f_2(C) f_3(D,B,C) f_4(E,C)$$



- $f_0(A)$
- $f_1(B,A)$
- $f_2(C)$
- $f_3(D,B,C)$
- $f_4(E,C)$
- $f_5(F,D)$
- $f_6(G,F,E)$
- $f_7(H,G)$
- $f_8(I,G)$
- $f_9(G)$
- $f_{10}(B)$

Variable elimination example(steps 3-4)

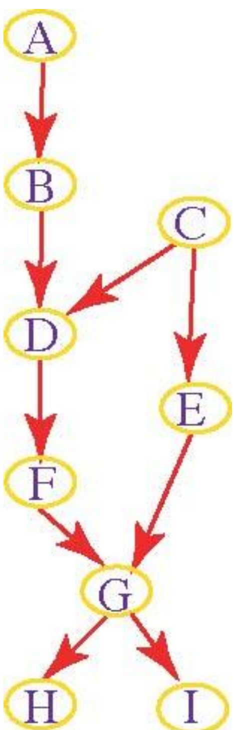
Compute $P(G \mid H{=}h_1)$. Elimination ordering A, C, E, I, B, D, F.

Previous state:

$$P(G,H{=}h_1) = f_9(G) \sum_F \sum_D f_5(F, D) \sum_B f_{10}(B) \sum_I f_8(I,G) \sum_E f_6(G,F,E) \sum_C f_2(C)\, f_3(D,B,C)\, f_4(E,C)$$

Eliminate C:

$$P(G,H{=}h_1) = f_9(G) \sum_F \sum_D f_5(F, D) \sum_B f_{10}(B) \sum_I f_8(I,G) \sum_E f_6(G,F,E)\, f_{12}(B,D,E)$$



- $f_0(A)$
- $f_1(B,A)$
- $f_2(C)$
- $f_3(D,B,C)$
- $f_4(E,C)$
- $f_5(F, D)$
- $f_6(G,F,E)$
- $f_7(H,G)$
- $f_8(I,G)$

- $f_9(G)$
- $f_{10}(B)$
- $f_{12}(B,D,E)$

Variable elimination example(steps 3-4)

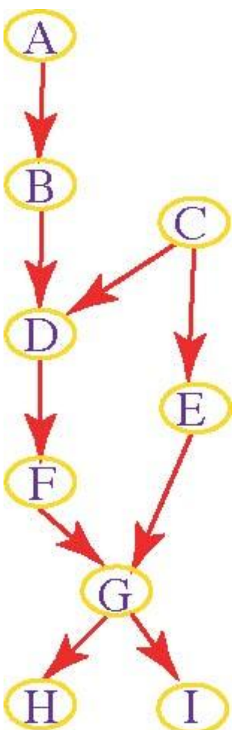
Compute $P(G \mid H=h_1)$. Elimination ordering A, C, E, I, B, D, F.

Previous state:

$P(G,H=h_1) = f_9(G) \sum_F \sum_D f_5(F, D) \sum_B f_{10}(B) \sum_I f_8(I,G) \sum_E f_6(G,F,E)\, f_{12}(B,D,E)$

Eliminate E:

$P(G,H=h_1) = f_9(G) \sum_F \sum_D f_5(F, D) \sum_B f_{10}(B)\, f_{13}(B,D,F,G)\, \sum_I f_8(I,G)$



- $f_0(A)$
- $f_1(B,A)$
- $f_2(C)$
- $f_3(D,B,C)$
- $f_4(E,C)$
- $f_5(F, D)$
- $f_6(G,F,E)$
- $f_7(H,G)$
- $f_8(I,G)$
- $f_9(G)$
- $f_{10}(B)$
- $f_{12}(B,D,E)$
- $f_{13}(B,D,F,G)$

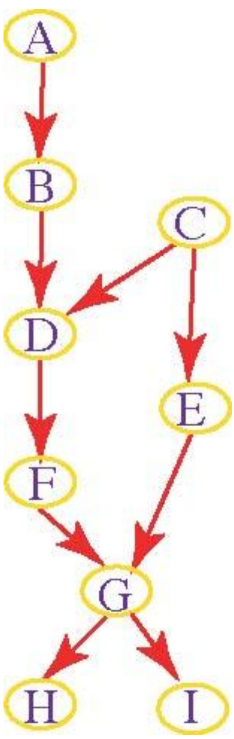
Variable elimination example(steps 3-4)

Compute $P(G \mid H=h_1)$. Elimination ordering A, C, E, I, B, D, F.

Previous state: $P(G,H=h_1) = f_9(G) \sum_F \sum_D f_5(F, D) \sum_B f_{10}(B)\, f_{13}(B,D,F,G) \sum_I f_8(I,G)$

Eliminate I:

$P(G,H=h_1) = f_9(G)\, f_{14}(G) \sum_F \sum_D f_5(F, D) \sum_B f_{10}(B)\, f_{13}(B,D,F,G)$



- $f_0(A)$
- $f_1(B,A)$
- $f_2(C)$
- $f_3(D,B,C)$
- $f_4(E,C)$
- $f_5(F, D)$
- $f_6(G,F,E)$
- $f_7(H,G)$
- $f_8(I,G)$
- $f_9(G)$
- $f_{10}(B)$
- $f_{12}(B,D,E)$
- $f_{13}(B,D,F,G)$
- $f_{14}(G)$

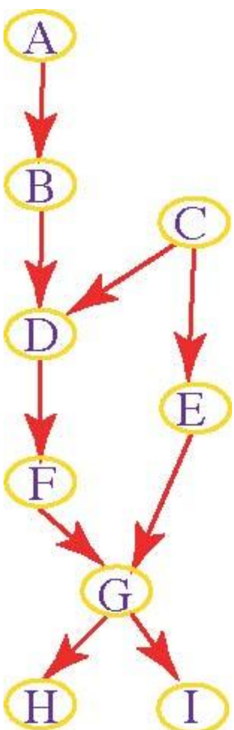
Variable elimination example(steps 3-4)

Compute $P(G \mid H=h_1)$. Elimination ordering A, C, E, I, B, D, F.

Previous state: $P(G,H=h_1) = f_9(G)\, f_{14}(G) \sum_F \sum_D f_5(F, D) \sum_B f_{10}(B)\, f_{13}(B,D,F,G)$

Eliminate B:

$$P(G,H=h_1) = f_9(G)\, f_{14}(G) \sum_F \sum_D f_5(F, D)\, f_{15}(D,F,G)$$



- $f_0(A)$
- $f_1(B,A)$
- $f_2(C)$
- $f_3(D,B,C)$
- $f_4(E,C)$
- $f_5(F, D)$
- $f_6(G,F,E)$
- $f_7(H,G)$
- $f_8(I,G)$
- $f_9(G)$
- $f_{10}(B)$
- $f_{12}(B,D,E)$
- $f_{13}(B,D,F,G)$
- $f_{14}(G)$
- $f_{15}(D,F,G)$

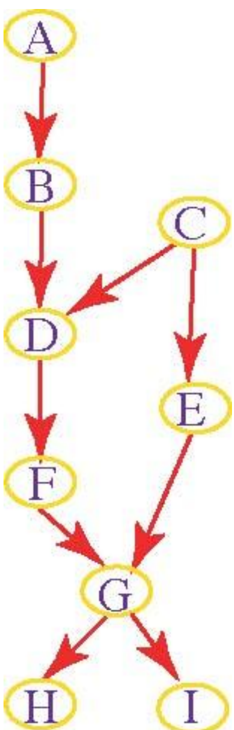
Variable elimination example(steps 3-4)

Compute $P(G \mid H=h_1)$. Elimination ordering A, C, E, I, B, D, F.

Previous state: $P(G,H=h_1) = f_9(G)\, f_{14}(G) \sum_F \sum_D f_5(F, D)\, f_{15}(D,F,G)$

Eliminate D:

$P(G,H=h_1) = f_9(G)\, f_{14}(G) \sum_F f_{16}(F, G)$



- $f_0(A)$
- $f_1(B,A)$
- $f_2(C)$
- $f_3(D,B,C)$
- $f_4(E,C)$
- $f_5(F, D)$
- $f_6(G,F,E)$
- $f_7(H,G)$
- $f_8(I,G)$
- $f_9(G)$
- $f_{10}(B)$
- $f_{12}(B,D,E)$
- $f_{13}(B,D,F,G)$
- $f_{14}(G)$
- $f_{15}(D,F,G)$
- $f_{16}(F, G)$

Variable elimination example(steps 3-4)

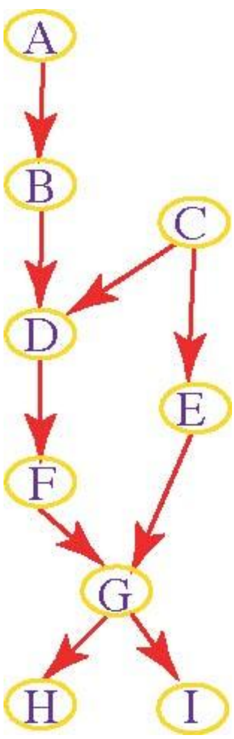
Compute $P(G \mid H=h_1)$. Elimination ordering A, C, E, I, B, D, F.

Previous state: $P(G,H=h_1) = f_9(G)\, f_{14}(G) \sum_F f_{16}(F, G)$

Eliminate F:

$$P(G,H=h_1) = \underbrace{f_9(G)\, f_{14}(G)\ \ f_{17}(G)}$$

- $f_0(A)$
- $f_1(B,A)$
- $f_2(C)$
- $f_3(D,B,C)$
- $f_4(E,C)$
- $f_5(F, D)$
- $f_6(G,F,E)$
- $f_7(H,G)$
- $f_8(I,G)$
- $f_9(G)$
- $f_{10}(B)$
- $f_{12}(B,D,E)$
- $f_{13}(B,D,F,G)$
- $f_{14}(G)$
- $f_{15}(D,F,G)$
- $f_{16}(F, G)$
- $f_{17}(G)$



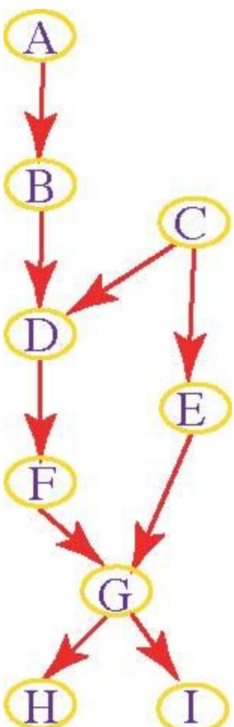
Variable elimination example (step 5)

Compute $P(G \mid H=h_1)$. Elimination ordering A, C, E, I, B, D, F.

Previous state: $P(G,H=h_1) = f_9(G)\ f_{14}(G)\ f_{17}(G)$

Multiply remaining factors:

$P(G,H=h_1) = f_{18}(G)$



- $f_0(A)$
- $f_1(B,A)$
- $f_2(C)$
- $f_3(D,B,C)$
- $f_4(E,C)$
- $f_5(F, D)$
- $f_6(G,F,E)$
- $f_7(H,G)$
- $f_8(I,G)$
- $f_9(G)$
- $f_{10}(B)$
- $f_{12}(B,D,E)$
- $f_{13}(B,D,F,G)$
- $f_{14}(G)$
- $f_{15}(D,F,G)$
- $f_{16}(F, G)$
- $f_{17}(G)$
- $f_{18}(G)$

Variable elimination example (step 6)

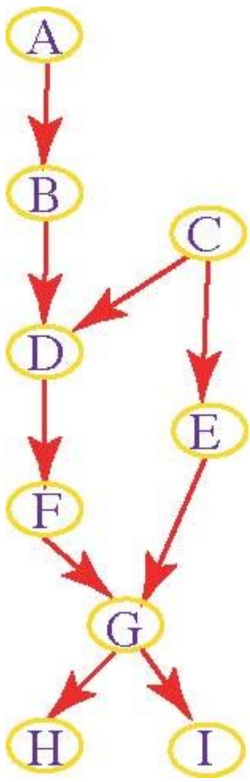
Compute $P(G \mid H=h_1)$. Elimination ordering A, C, E, I, B, D, F.

Previous state:

$P(G,H=h_1) = f_{18}(G)$

Normalize:

$$P(G \mid H=h_1) = f_{18}(G) / \sum_{g \in dom(G)} f_{18}(G)$$



- $f_0(A)$
- $f_1(B,A)$
- $f_2(C)$
- $f_3(D,B,C)$
- $f_4(E,C)$
- $f_5(F, D)$
- $f_6(G,F,E)$
- $f_7(H,G)$
- $f_8(I,G)$
- $f_9(G)$
- $f_{10}(B)$
- $f_{12}(B,D,E)$
- $f_{13}(B,D,F,G)$
- $f_{14}(G)$
- $f_{15}(D,F,G)$
- $f_{16}(F, G)$
- $f_{17}(G)$
- $f_{18}(G)$

Today Oct 6

- **R&R systems in Stochastic environments**
  - Bayesian Networks Representation
  - Bayesian Networks Exact Inference
  - **Bayesian Networks Approx. Inference**

Approximate Inference

- Basic idea:
  - Draw N samples from a sampling distribution S
  - Compute an approximate posterior probability
  - Show this converges to the true probability P

- Why sample?
  - Inference: getting a sample is faster than computing the right answer (e.g. with variable elimination)

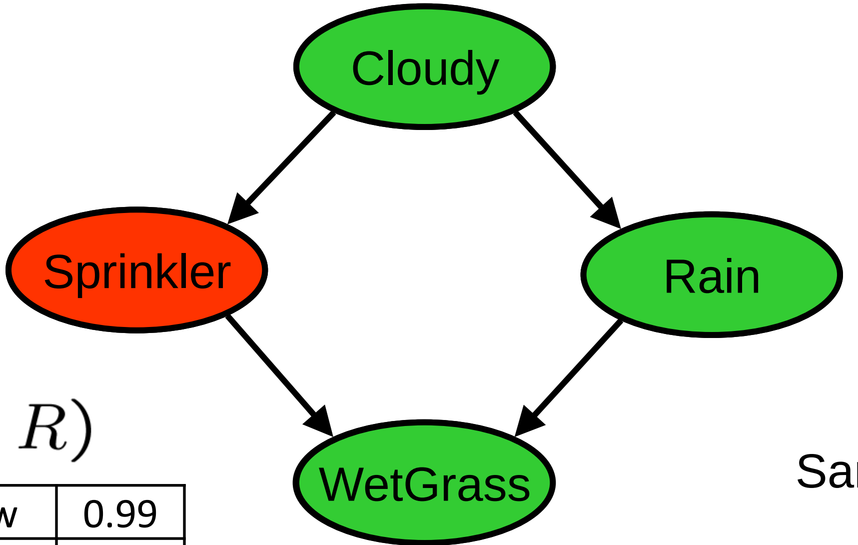
Prior Sampling

$P(C)$

+c	0.5
-c	0.5

$P(S|C)$

+c	+s	0.1
	-s	0.9
-c	+s	0.5
	-s	0.5



$P(R|C)$

+c	+r	0.8
	-r	0.2
-c	+r	0.2
	-r	0.8

$P(W|S,R)$

+s	+r	+w	0.99
		-w	0.01
	-r	+w	0.90
		-w	0.10
-s	+r	+w	0.90
		-w	0.10
	-r	+w	0.01
		-w	0.99

Samples:

+c, -s, +r, +w

-c, +s, -r, +w

…

Example

- We'll get a bunch of samples from the BN:

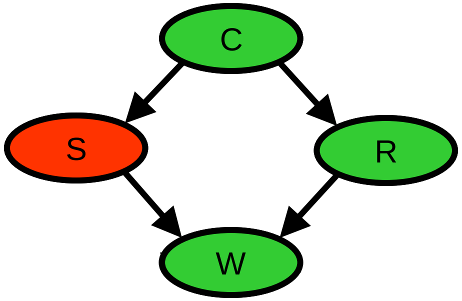
  +c, -s, +r, +w

  +c, +s, +r, +w

  -c, +s, +r, -w

  +c, -s, +r, +w

  -c, -s, -r, +w

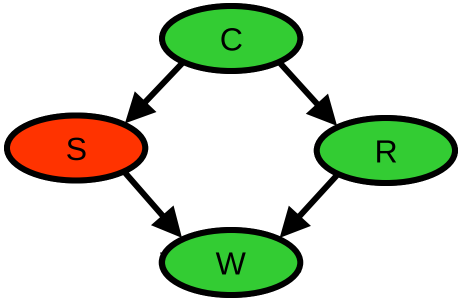


- If we want to know P(W)
  - We have counts <+w:4, -w:1>
  - Normalize to get P(W) = <+w:0.8, -w:0.2>
  - This will get closer to the true distribution with more samples
  - Can estimate anything else, too
  - What about P(C| +w)? P(C| +r, +w)? P(C| -r, -w)?

what's the drawback? Can use fewer samples ?

Rejection Sampling

- Let's say we want P(C)
  - No point keeping all samples around
  - Just tally counts of C as we go



- Let's say we want P(C| +s)
  - Same thing: tally C outcomes, but ignore (reject) samples which don't have S=+s
  - This is called rejection sampling
  - It is also consistent for conditional probabilities (i.e., correct in the limit)

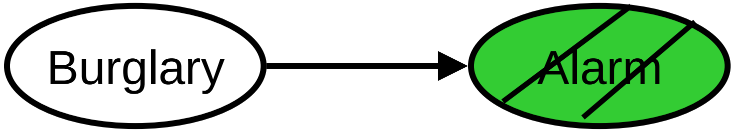
+c, -s, +r, +w
+c, +s, +r, +w
-c, +s, +r, -w
+c, -s, +r, +w
-c, -s, -r, +w

Likelihood Weighting

- Problem with rejection sampling:
  - If evidence is unlikely, you reject a lot of samples
  - You don't exploit your evidence as you sample
  - Consider P(B|+a)

-b, -a
-b, -a
-b, -a
-b, -a
+b, +a

- Idea: fix evidence variables and sample the rest



-b +a
-b, +a
-b, +a
-b, +a
+b, +a

- Problem: sample distribution not consistent!
- Solution: weight by probability of evidence given parents

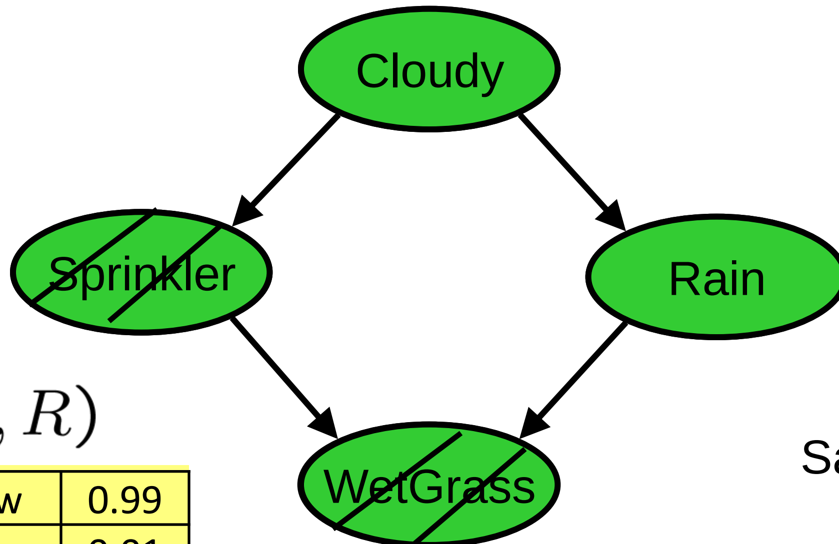
Likelihood Weighting

$P(C)$

+c	0.5
-c	0.5

$P(S|C)$

+c	+s	0.1
	-s	0.9
-c	+s	0.5
	-s	0.5



$P(R|C)$

+c	+r	0.8
	-r	0.2
-c	+r	0.2
	-r	0.8

$P(W|S,R)$

+s	+r	+w	0.99
		-w	0.01
	-r	+w	0.90
		-w	0.10
-s	+r	+w	0.90
		-w	0.10
	-r	+w	0.01
		-w	0.99

Samples:

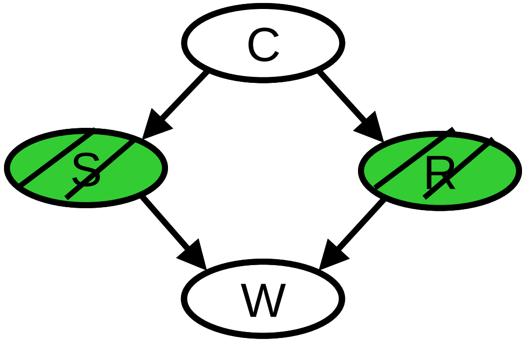
+c, +s, +r, +w

…

$$w = 1.0 \times 0.1 \times 0.99$$

Likelihood Weighting

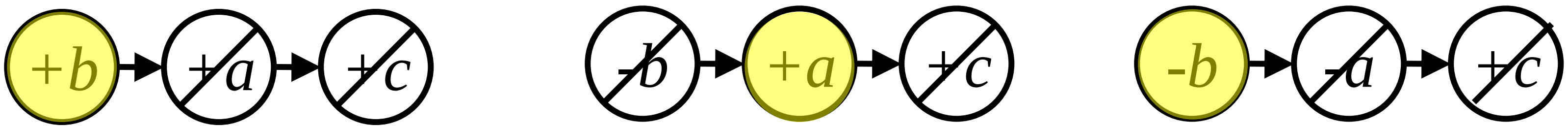
- Likelihood weighting is good
  - We have taken evidence into account as we generate the sample
  - E.g. here, W's value will get picked based on the evidence values of S, R
  - More of our samples will reflect the state of the world suggested by the evidence
- Likelihood weighting doesn't solve all our problems
  - Evidence influences the choice of downstream variables, but not upstream ones (C isn't more likely to get a value matching the evidence)
- We would like to consider evidence when we sample *every* variable



Markov Chain Monte Carlo

- *Idea*: instead of sampling from scratch, create samples that are each like the last one.

- *Procedure*: resample one variable at a time, conditioned on all the rest, but keep evidence fixed. E.g., for P(b|+c):



- *Properties*: Now samples are not independent (in fact they're nearly identical), but sample averages are still consistent estimators! And can be computed efficiently

- *What's the point*: both upstream and downstream variables condition on evidence.

TODO for this Tue

Finish Reading Chp 6 of textbook

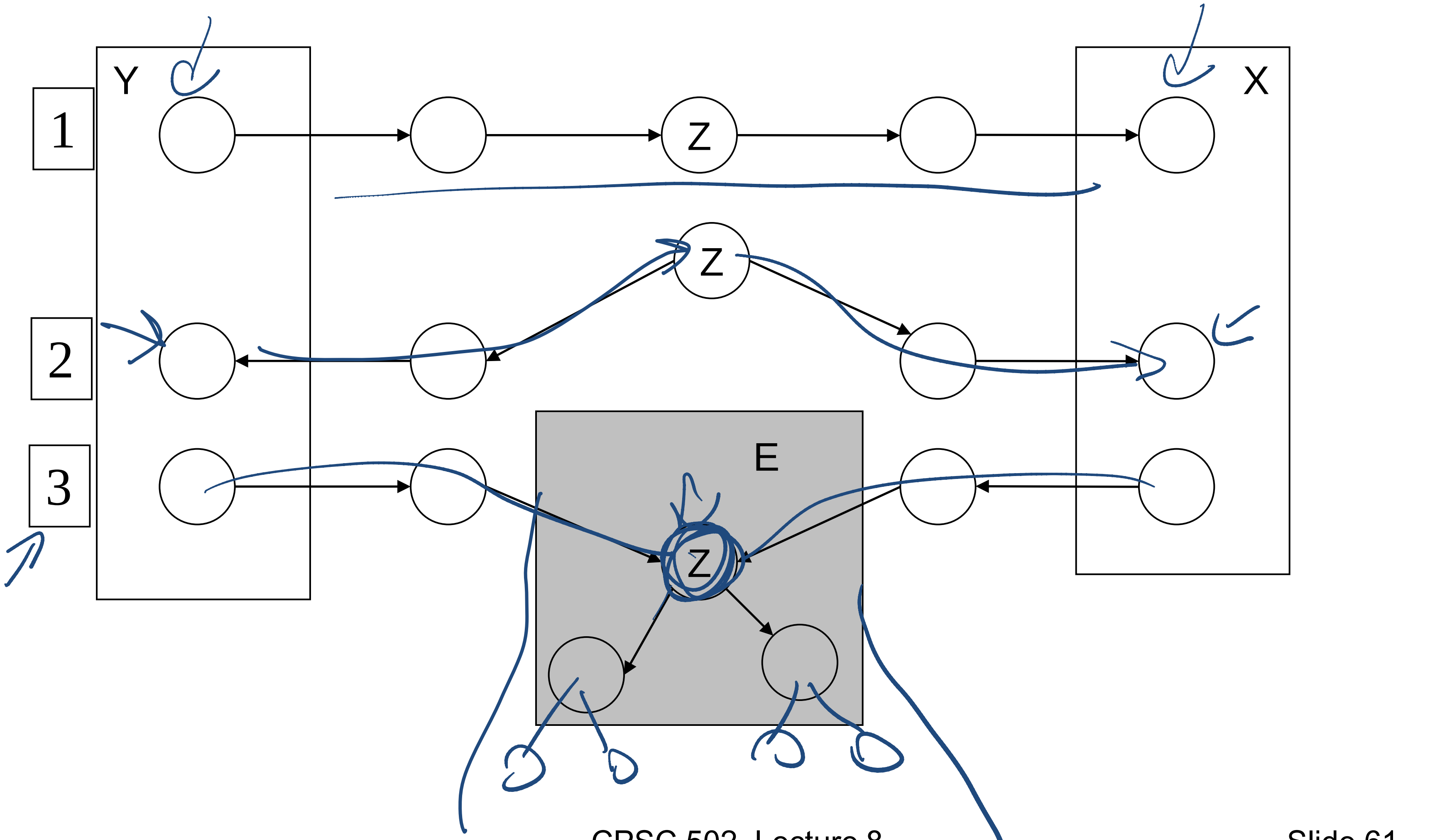
(Skip 6.4.2.5 Importance Sampling 6.4.2.6 Particle Filtering, we have covered instead likelihood weighting and MCMC methods)

Also Do exercises 6.E

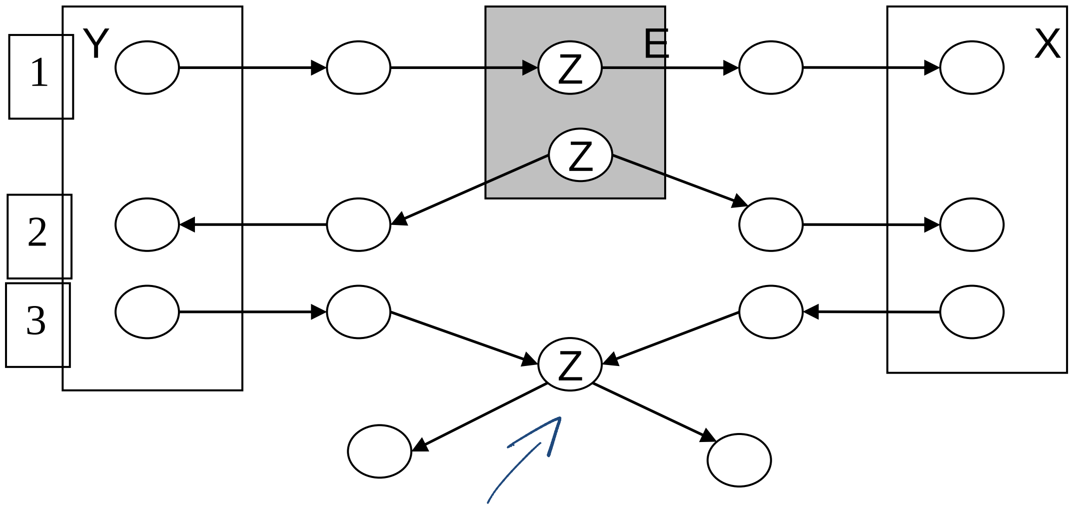
http://www.aispace.org/exercises.shtml

Or ….Conditional Dependencies

In 1,2,3 X Y are dependent

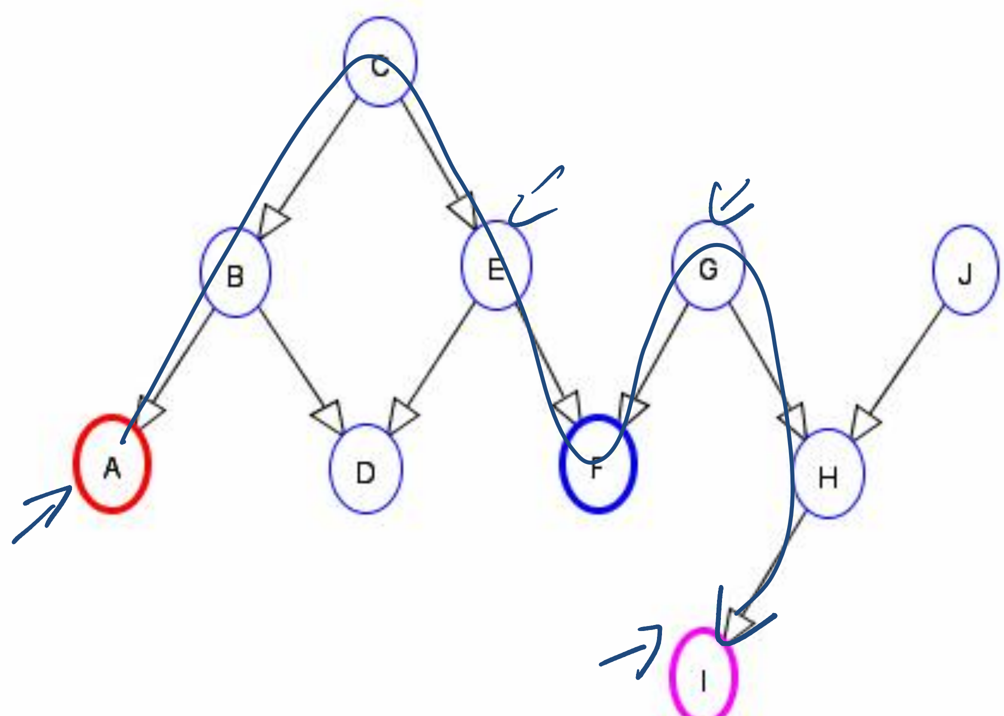


In/Dependencies in a Bnet : Example 1

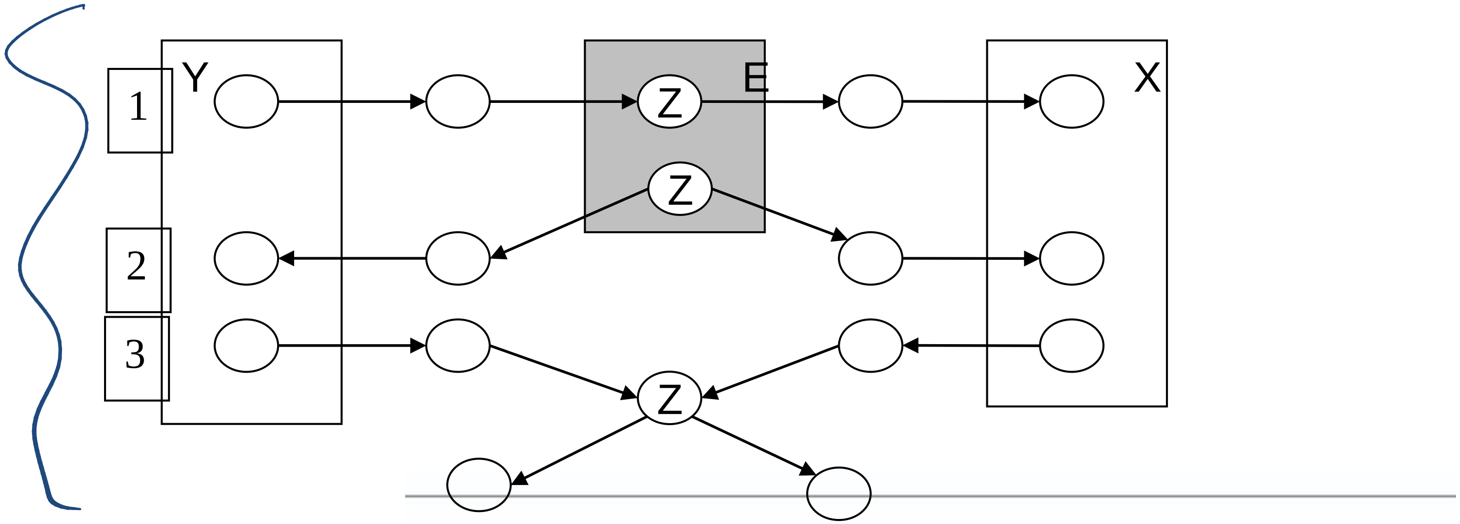


Is A conditionally independent of I given F?

false

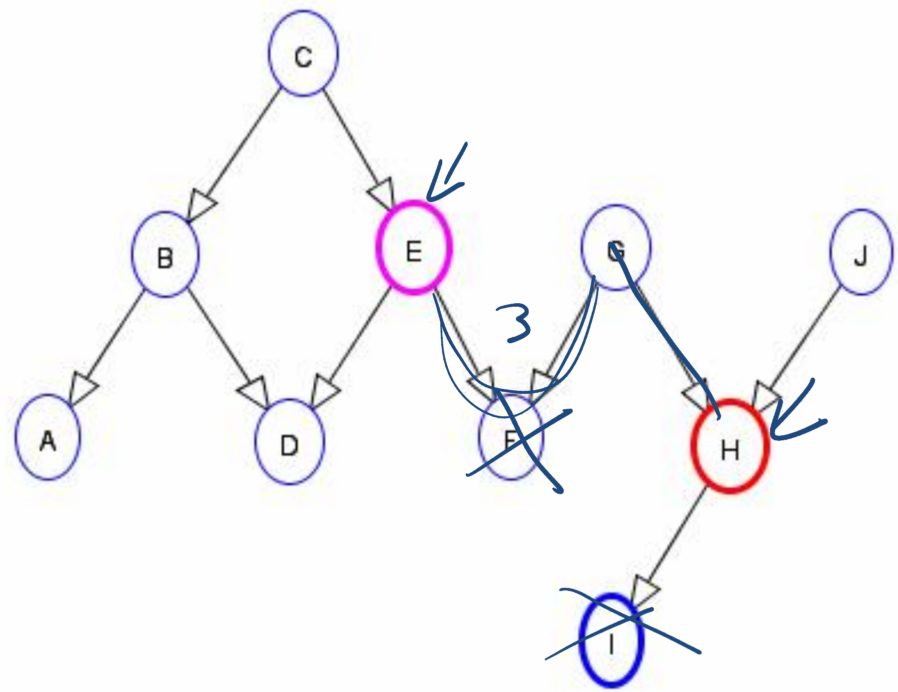


In/Dependencies in a Bnet : Example 2



Is H conditionally independent of E given ~~I~~?

true

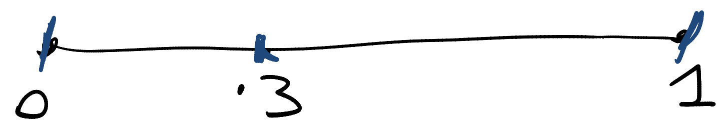


Sampling a discrete probability distribution

e.g. Sim. Annealing. Select n' with probability p

$p = .3$

generate random number in $[0,1]$



If $< .3$ accept n'

e.g. Beam Search: Select K individuals. Probability of selection proportional to their value

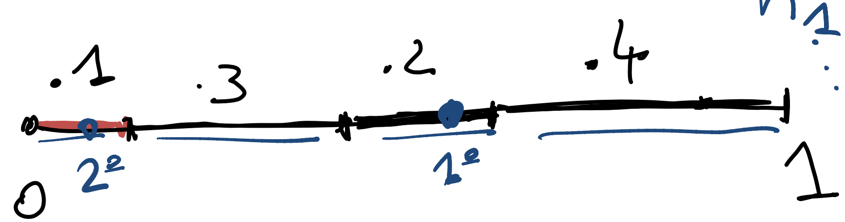
SAME HERE

→ n_1 $\boxed{p_1 = .1}$

→ n_2 $\boxed{p_2 = .3}$

→ n_3 $\boxed{p_3 = .2}$

→ n_4 $\boxed{p_4 = .4}$



n_3 first sample

n_1 second sample

Problem and Solution Plan

- We model the environment as a set of random vars

  $X_1 \ldots X_n$ JPD $P(X_1 \ldots X_n)$
- Why the joint is not an adequate representation ?

"Representation, reasoning and learning" are "exponential" in the number of variables

Solution: Exploit marginal&**conditional** independence

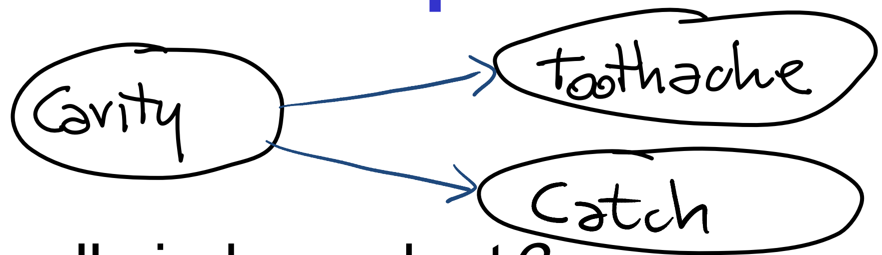
$P(X|Y) = P(X)$ $P(X|YZ) = P(X|Z)$

But how does independence allow us to simplify the joint?

CHAIN RULE!

Look for weaker form of independence

P(*Toothache, Cavity, Catch*)



Are *Toothache* and *Catch* marginally independent?

$P(\quad | \quad) = P(\text{Toothache})$?

BUT If I have a cavity, does the probability that the probe catches depend on whether I have a toothache?

(1) P(*catch* / *toothache, cavity*) = $P(\text{catch} \mid \text{cavity})$

What if I haven't got a cavity?

(2) P(*catch* / *toothache,* $\neg$*cavity*) = $P(\text{catch} \mid \neg\text{cavity})$

- *Each is directly caused by the cavity, but neither has a direct effect on the other*

Conditional independence

In general, *Catch* is conditionally independent of *Toothache* given *Cavity*:

1 **P**(*Catch* | *Toothache,Cavity*) = **P**(*Catch* | *Cavity*)

Equivalent statements:

2 **P**(*Toothache* | *Catch*, *Cavity*) = **P**(*Toothache* | *Cavity*)

3 **P**(*Toothache*, *Catch* | *Cavity*) =
P(*Toothache* | *Cavity*) **P**(*Catch* | *Cavity*)

$$P(X, Y \mid Z) = P(X \mid Z)\, P(Y \mid Z)$$

Proof of equivalent statements

1 If $\boxed{P(X|YZ) = P(X|Z)} \Rightarrow$

$$\Rightarrow \text{A} \quad \frac{P(X,Y,Z)}{P(Y,Z)} = \frac{P(X,Z)}{P(Z)} \Rightarrow \quad 2$$

$$\Rightarrow \frac{P(X,Y,Z)}{P(X,Z)} = \frac{P(Y,Z)}{P(Z)} \Rightarrow \boxed{P(Y|X,Z) = P(Y|Z)}$$

3 $$P(X,Y|Z) = \frac{P(X,Y,Z)}{P(Z)} \overset{\text{from A}}{\Rightarrow} \frac{P(Y,Z)\,P(X,Z)}{P(Z)} \cdot \frac{1}{P(Z)}$$

$$= \frac{P(Y,Z)}{P(Z)} \cdot \frac{P(X,Z)}{P(Z)} = \boxed{P(Y|Z) \cdot P(X|Z)}$$

Conditional Independence: Formal Def.

Sometimes, two variables might not be marginally independent. However, they *become* independent after we observe some third variable

DEF. Random variable **X** is conditionally independent of random variable **Y** given random variable **Z** if, for all $x_i \in dom(X)$, $y_k \in dom(Y)$, $z_m \in dom(Z)$

$$P(X = x_i \mid Y = y_k, Z = z_m) = P(X = x_i \mid Z = z_m)$$

That is, knowledge of **Y**'s value doesn't affect your belief in the value of **X,** given a value of **Z**

Conditional independence: Use

Write out full joint distribution using chain rule:

$$\mathbf{P}(Cavity, Catch, Toothache)$$

$$= \mathbf{P}(Toothache \mid Catch, Cavity)\ \mathbf{P}(Catch \mid Cavity)\ \mathbf{P}(Cavity)$$

$$= \underbrace{\mathbf{P}(Toothache \mid \text{Cavity})}_{2}\ \underbrace{\mathbf{P}(Catch \mid Cavity)}_{2}\ \underbrace{\mathbf{P}(\text{Cavity})}_{1}$$

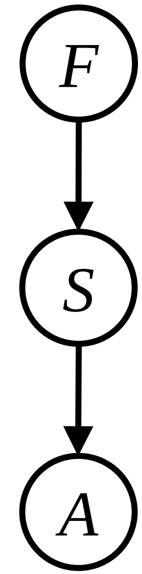
how many probabilities? $2^3 - 1 = 7$

$2 + 2 + 1 = 5$

The use of conditional independence often reduces the size of the representation of the joint distribution from exponential in *n* to linear in *n*. **n is the number of vars**

Conditional independence is our **most basic** and **robust** form of **knowledge** about **uncertain environments**.

Approximate Inference



Sampling / Simulating / Observing

Sampling is a hot topic in machine learning, and it's really simple

Basic idea:

- Draw N samples from a sampling distribution S
- Compute an approximate posterior probability
- Show this converges to the true probability P

Why sample?

- Learning: get samples from a distribution you don't know
- Inference: getting a sample is faster than computing the right answer (e.g. with variable elimination)