

Lecture 7-8: Probabilistic Reasoning

CS 486/686: Introduction to Artificial Intelligence

Introduction

- Plan for the next few lectures
 - Quick review of basic probability
 - Bayesian Networks
 - What they are?
 - How to use them?
 - Hidden Markov Models

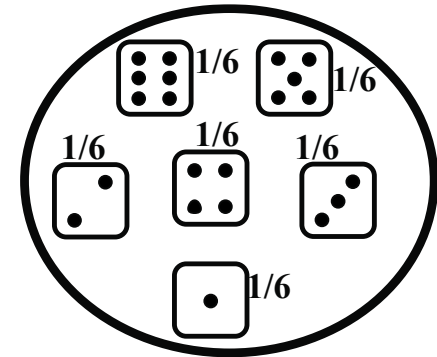
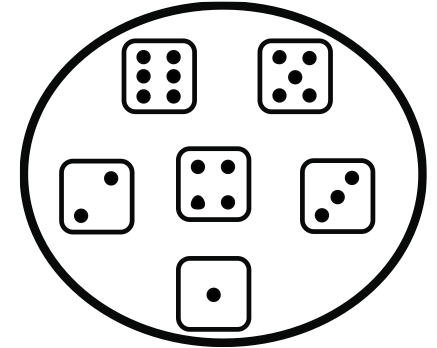
Warning: If you have forgotten everything you learned in STAT230/240/206 now is the time for some review!

Uncertainty

- The world is rife with uncertainty!
 - Example: If I leave for YYZ 60 minutes before my flight, will I miss my flight?
- Problems:
 - Partial observability (state of the 401, other drivers' plans, etc)
 - Noisy sensors (traffic reports, Google maps)
 - Complexity of modelling and predicting traffic, security line, etc
 - Lack of knowledge of world dynamics (will I get a flat tire?)
- We need some way to reason about this uncertainty

Probability Theory

- Begin with a set Ω of possible worlds
 - E.g. 6 possible rolls of a die $\{1,2,3,4,5,6\}$
- A **probability model** assigns a number $P(\omega)$ to each world ω
 - E.g. $P(1)=P(2)=P(3)=P(4)=P(5)=P(6) = 1/6$
- These numbers must satisfy
 - $0 \leq P(\omega) \leq 1$
 - $\sum_{\omega \in \Omega} P(\omega) = 1$



Probability Theory

- An event is any subset of Ω

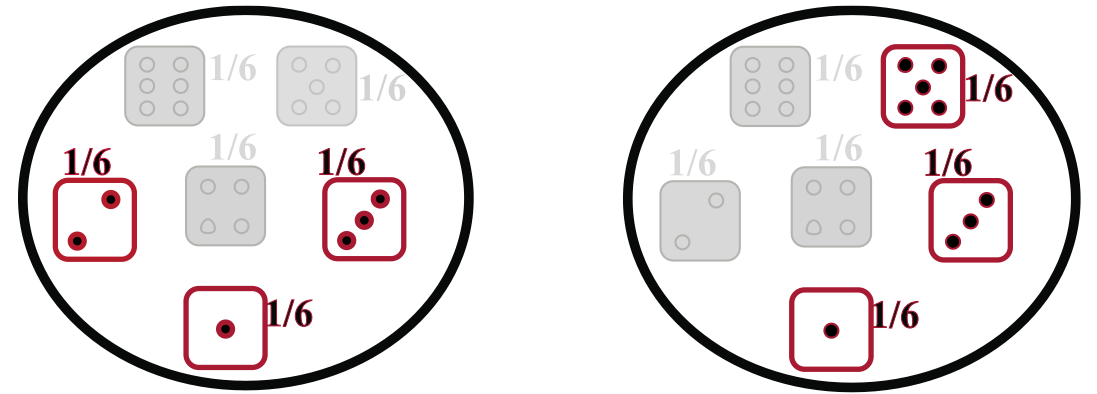
- E.g. “roll < 4” is the set {1,2,3}
- E.g. “roll is odd” is the set {1,3,5}

- The probability of an event is the sum of probabilities over its worlds

- $P(A) = \sum_{\omega \in A} P(\omega)$
- E.g. $P(\text{roll} < 4) = P(1) + P(2) + P(3) = \frac{1}{2}$

- De Finetti (1931) (paraphrased)

- If you don't believe in probability theory, don't gamble! You will lose.



Probability Theory

- A random variable (usually denoted by a capital letter) is some aspect of the world we may be uncertain about
- The range of a random variable is the set of possible values
 - Odd= “Is the dice roll an odd number?” -> {true, false}
 - D = “How long will it take me to get to the airport?” -> $[0, \infty)$
- A **probability distribution** of a random variable X specifies a probability for each value x in its range (probability that $X=x$)
 - $P(X = x) = \sum_{\{\omega: X(\omega)=x\}} P(\omega)$
 - $P(x)$ for short
 - **P(X)** refers to the entire distribution (think of this as a vector or table)

Probability Distributions

- Temperature:

$P(T)$

T	P
hot	0.5
cold	0.5



- Weather:

$P(W)$

W	P
sun	0.6
rain	0.1
fog	0.3
meteor	0.0



- Joint distribution*

$P(T,W)$

		Temperature	
		hot	cold
Weather	sun	0.45	0.15
	rain	0.02	0.08
	fog	0.03	0.27
	meteor	0.00	0.00

Marginalization

- Given a joint distribution over a set of variables, we can compute every event probability (marginalization)
- Given joint distribution $P(A,B)$
 - $P(A = a) = \sum_b P(A = a, B = b)$
 - $P(B = b) = \sum_a P(A = a, B = b)$
- Examples
 - $P(T=\text{hot}, W=\text{sun}) = 0.45$
 - $P(W=\text{fog}) = P(T=\text{hot}, W=\text{fog}) + P(T=\text{cold}, W=\text{fog}) = 0.03 + 0.2 = 0.23$
 - What about the probability it is not hot or it is foggy?
 - Remember that $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$ and $P(\sim A) = 1 - P(A)$

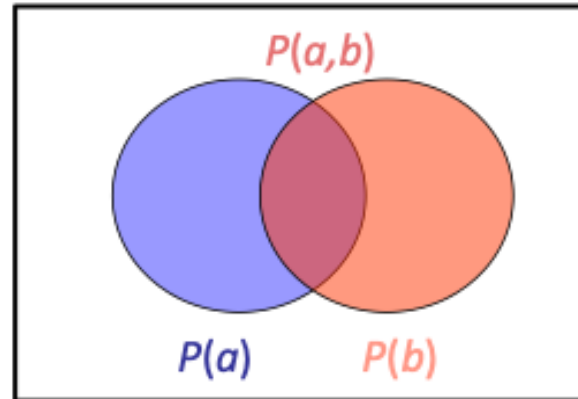
$P(T,W)$

		Temperature	
		hot	cold
Weather	sun	0.45	0.15
	rain	0.02	0.08
	fog	0.03	0.27
	meteor	0.00	0.00

Conditional Probabilities

- There is a relationship between joint and conditional probabilities

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



$P(T,W)$

		Temperature	
		hot	cold
Weather	sun	0.45	0.15
	rain	0.02	0.08
	fog	0.03	0.27
	meteor	0.00	0.00

$$P(W=s \mid T=c) = \frac{P(W=s, T=c)}{P(T=c)} = 0.15/0.50 = 0.3$$

$$\begin{aligned} &= P(W=s, T=c) + P(W=r, T=c) + P(W=f, T=c) + P(W=m, T=c) \\ &= 0.15 + 0.08 + 0.27 + 0.00 = 0.50 \end{aligned}$$

Conditional Distributions

- Distributions for one set of variables given another set of variables

		Temperature	
		hot	cold
Weather	sun	0.45	0.15
	rain	0.02	0.08
	fog	0.03	0.27
	meteor	0.00	0.00

$P(W \mid T=h)$

hot

0.90
0.04
0.06
0.00

$P(W \mid T=c)$

cold

0.30
0.16
0.54
0.00

$P(W \mid T)$

hot

cold

0.90	0.30
0.04	0.16
0.06	0.54
0.00	0.00

Notice how the values in the tables have been re-normalized!

The Chain Rule

- Note that $P(A|B) = \frac{P(A,B)}{P(B)}$
 - Thus, we also have $P(A, B) = P(A|B)P(B) = P(B|A)P(A)$
- This **Chain Rule** generalizes
 - $P(x_1, x_2, \dots, x_n) = \prod_i P(x_i | x_1, \dots, x_{i-1})$
 - Why? Consider $P(x_1, x_2, x_3) = P(x_3 | x_1, x_2)P(x_1, x_2)$
 - Note that $P(x_1, x_2) = P(x_2 | x_1)P(x_1)$

Bayes Rule

- Note that $P(a|b)P(b) = P(a, b) = P(b|a)P(a)$
 - Therefore

$$P(a|b) = \frac{P(b|a)P(a)}{P(b)}$$

- Bayes Rule is critically important for inference (reasoning)

A diagram showing the equation $P(H|e) = \frac{P(e|H)P(H)}{P(e)}$ with red arrows pointing to each part. The arrow from 'Likelihood' points to $P(e|H)$. The arrow from 'Prior probability' points to $P(H)$. The arrow from 'Posterior' points to $P(H|e)$. The arrow from 'Normalizing constant' points to $P(e)$.

$$P(H|e) = \frac{P(e|H)P(H)}{P(e)}$$

Likelihood

Prior probability

Posterior

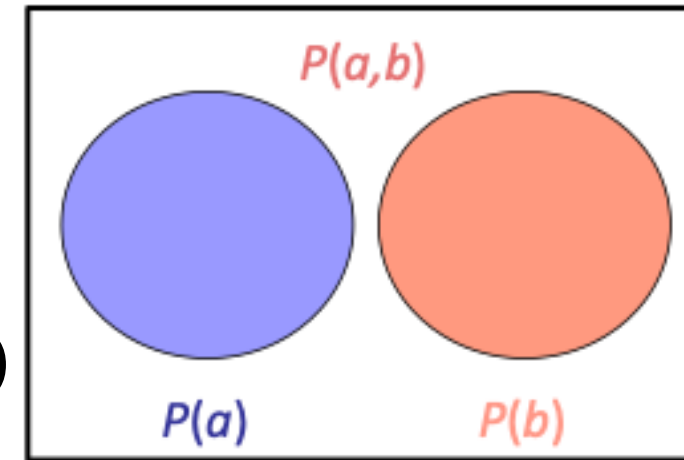
Normalizing constant

Independence

- Two variables, X and Y , if knowledge of X does not change uncertainty of Y (and vice versa)

$$\forall x, y P(x, y) = P(x)P(y)$$

$$P(x|y) = P(x) \text{ and } P(y|x) = P(y)$$



More Generally:

$$P(x_1, x_2, \dots, x_n) = \prod_i P(x_i)$$

Example: Two dice rolls Roll1 and Roll2

- $P(\text{Roll1}=5, \text{Roll2}=3) = P(\text{Roll1}=5)P(\text{Roll2}=3) = 1/6 \times 1/6 = 1/36$
- $P(\text{Roll2}=3 \mid \text{Roll1}=5) = P(\text{Roll2}=3) = 1/6$

Independence

- If variables $X_1, \dots, X_n$ are independent then we can easily represent the full joint distribution $P(X_1, \dots, X_n) = \prod_i P(X_i)$
 - If these are Boolean random variables then we need only n parameters to represent the joint distribution!
 - And if we know the joint distribution we can answer any probabilistic query!

4 independent Boolean random vars X_1, X_2, X_3, X_4

$$\Pr(x_1) = 0.4, \Pr(x_2) = 0.2, \Pr(x_3) = 0.5, \Pr(x_4) = 0.8$$

$$\begin{aligned}\Pr(x_1, \sim x_2, x_3, x_4) &= \Pr(x_1) (1 - \Pr(x_2)) \Pr(x_3) \Pr(x_4) \\ &= (0.4)(0.8)(0.5)(0.8) \\ &= 0.128\end{aligned}$$

$$\begin{aligned}\Pr(x_1, x_2, x_3 | x_4) &= \Pr(x_1) \Pr(x_2) \Pr(x_3) \mathbf{1} \\ &= (0.4)(0.2)(0.5)(1) \\ &= 0.04\end{aligned}$$

Conditional Independence

- Full independence is actually quite rare.
- But conditional independence is not!
 - X and Y are conditionally independent given Z if knowing the value of Y does not change the uncertainty of X (and vice versa) if the value of Z is known.

$$\forall x, y, z P(x|y, z) = P(x|z)$$

$$\forall x, y, z P(x, y|z) = P(x|z)P(y|z)$$

Probabilistic Inference and Leveraging Independence

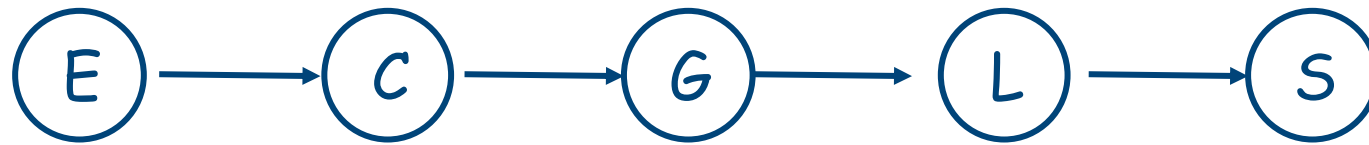
- Probabilistic inference simply means answering probabilistic questions
 - $P(a,b,c)?$
 - $P(a|\sim b)?$
- Often we have some evidence (**E**), and some query variables (**Q**) we are interested in. We ask $P(\mathbf{Q}|\mathbf{E})$?
- Note that if we have a full joint distribution over all possible variables of interest, we can compute the answer to $P(\mathbf{Q}|\mathbf{E})$
 - Apply marginalization, chain rule, etc
 - This is expensive unless we have some additional structure.... independence and conditional independence!

Notation Break

- $P(X)$ will be a distribution over X (X might be a variable or a set of variables)
 - Note the difference between $P(X)$ (distribution) and $P(x)$ (number)
 - Think of $P(X)$ as a function that accepts any x_i in $\text{Dom}(X)$ and returns a number
- $P(X|Y)$ is the family of conditional distributions over X (one for each y in $\text{Dom}(Y)$)
 - Think of $P(X|Y)$ as a function that accepts any x_i and y_j and returns $P(x_i|y_j)$

A Morning Story

If Kate work up too early (E), she probably needs coffee (C); if Kate needs coffee (C), she is likely to be grumpy (G). If she is grumpy, then it is possible that the lecture won't go smoothly (L). If the lecture does not go smoothly, then the students will likely be sad (S).



E - Kate woke too early G - Kate is grumpy S - Students are sad
C - Kate needs coffee L - The lecture did not go smoothly

Conditional Independence

- If you learned any of E, C, G, or L then your assessment of $P(S)$ would change
 - Therefore, S is not independent of E, C, G, or L



Conditional Independence

- But if you knew the value of L (true or false) then learning the values of E , C , or G would not influence $P(S)$.
 - Students are not sad because Kate did not have a coffee, they are sad because of the lecture
 - So S is independent of E , C , and G given L .
 - Similarly, L is independent of E and C given G , and G is independent of E given C

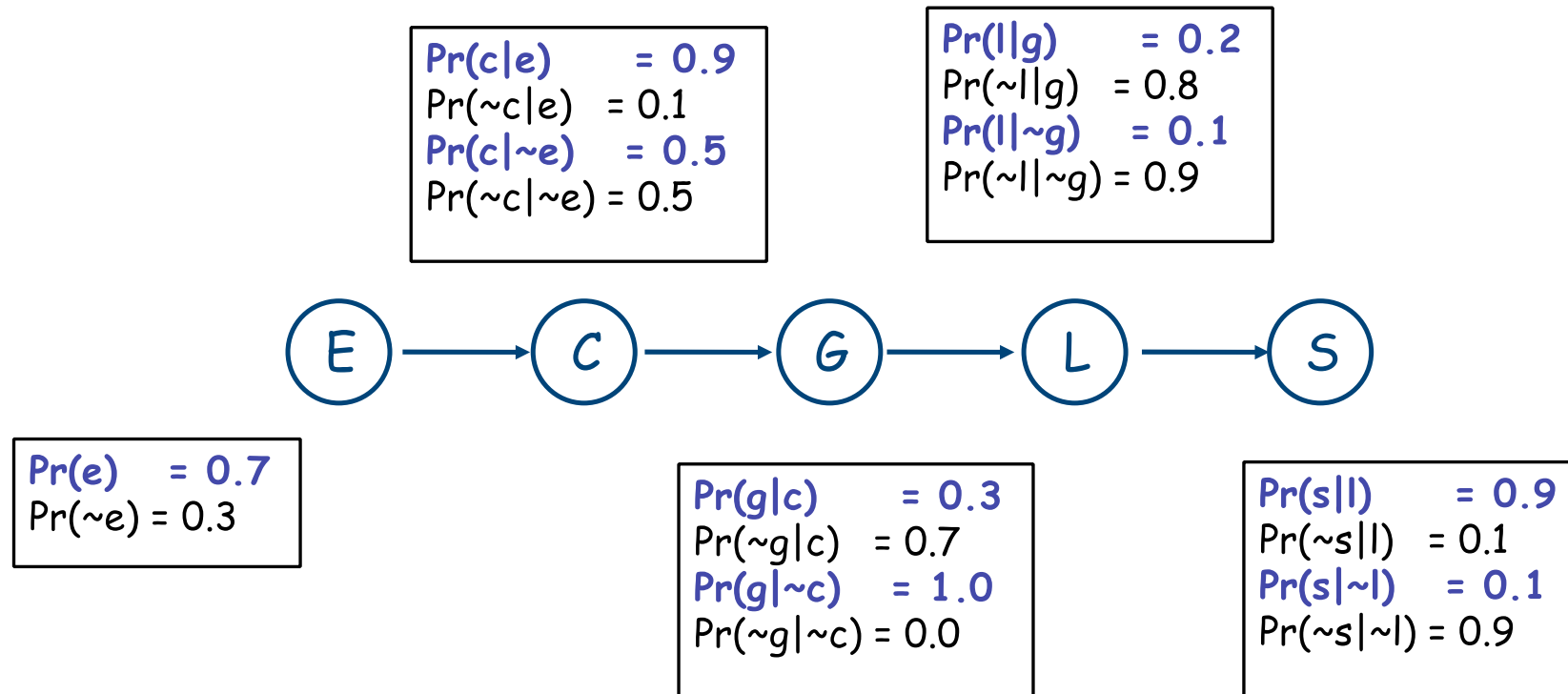


Conditional Independence



- $P(S, L, G, C, E) = P(S | L, G, C, E)P(L | G, C, E)P(G | C, E)P(C | E)P(E)$
 - Why?
- $P(S, L, G, C, E) = P(S | L)P(L | G)P(G | C)P(C | E)P(E)$
 - Why?
- We can specify the full joint distribution by specifying five conditional distributions!

Example



Specifying the joint requires only 9 parameters instead of 31 for explicit representation

Inference is Easy



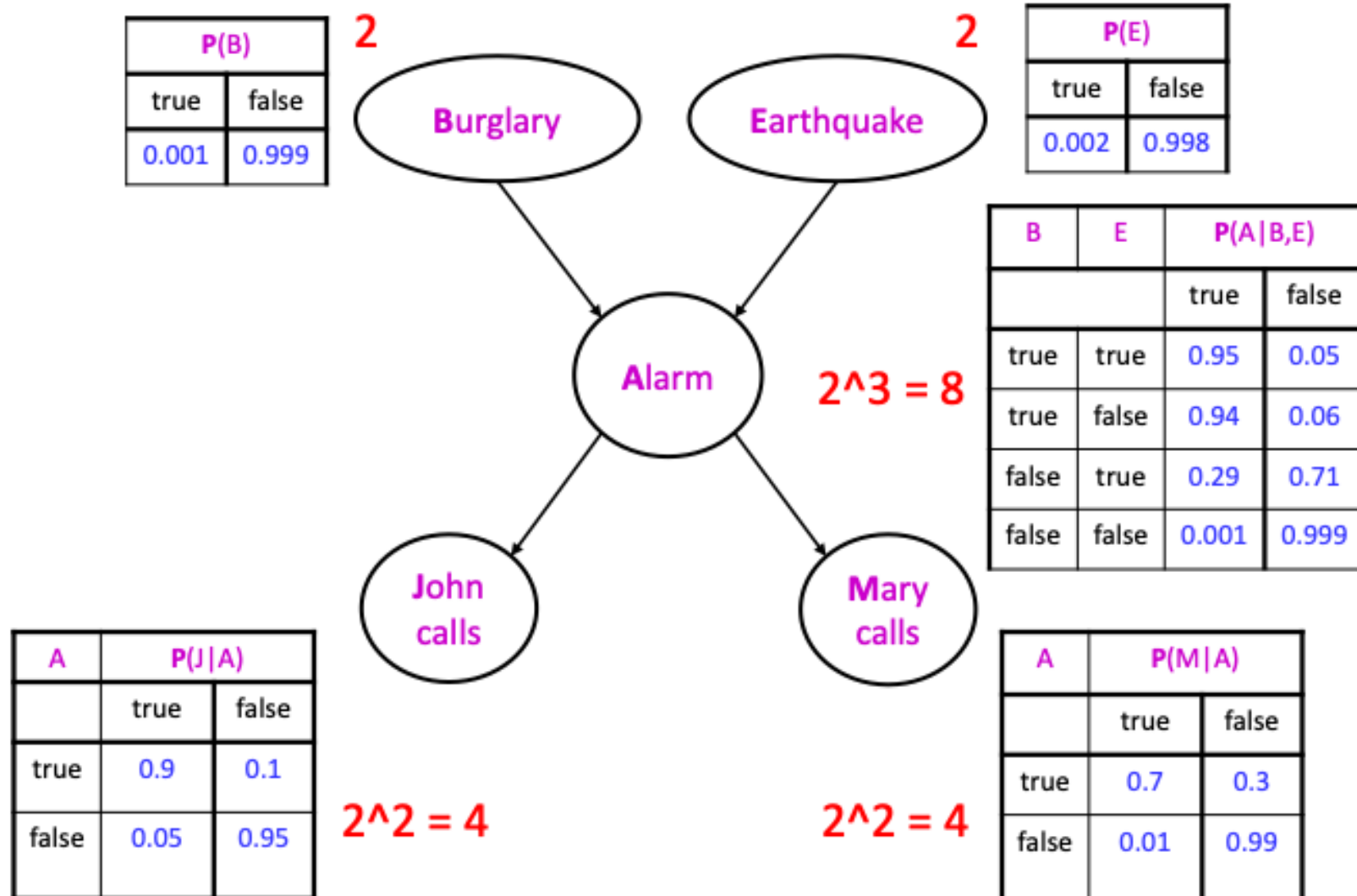
$$\begin{aligned} P(g) &= \sum_{c_i \in \text{Dom}(C)} \Pr(g \mid c_i) \Pr(c_i) \\ &= \sum_{c_i \in \text{Dom}(C)} \Pr(g \mid c_i) \sum_{e_i \in \text{Dom}(E)} \Pr(c_i \mid e_i) \Pr(e_i) \end{aligned}$$

These are all terms specified in our local distributions!

Bayesian Network (Bayes Net/Belief Nets)

- Bayes Nets are a technique for describing complex joint distributions (models) using simple, conditional distributions
 - A subset of the general class of graphical models
 - Use local causality/conditional independence
- A Bayes Net consists of
 - Nodes: variables (with domains)
 - Arcs: interactions between variables
 - Conditional Probability Tables (CPT) associated with each node $P(X | \text{Parents}(X))$

Example: Alarm Network



Factor size of each CPT:

$$d \prod d_i$$

Parent range sizes: $d_1, \dots, d_k$

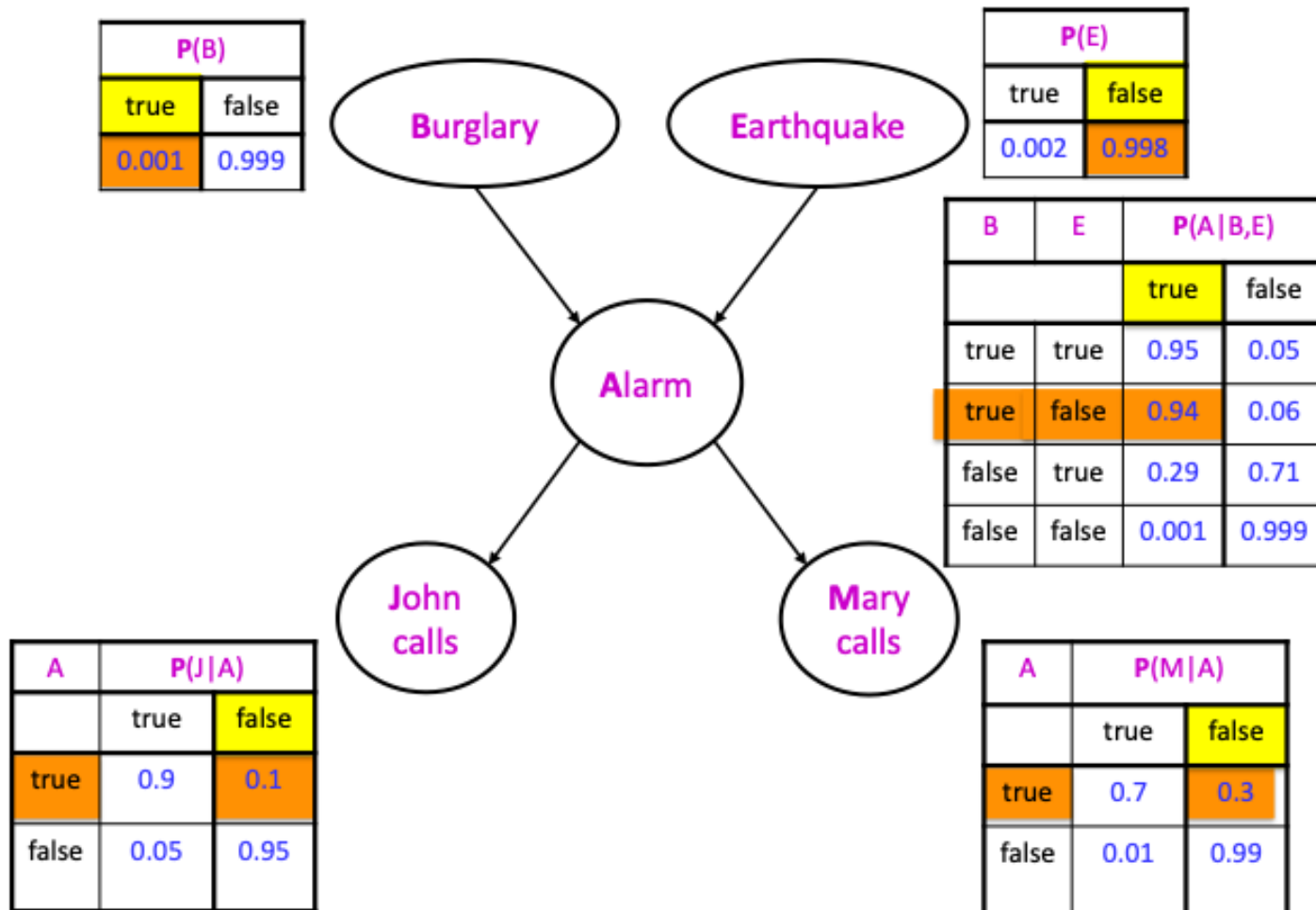
Child range size: d

Each table row must sum to 1

Bayes Nets Semantics

- Every X_i is conditionally independent of all its non-descendants given its parents
 - $P(X_i | S \cup \text{Par}(X_i)) = P(X_i | \text{Par}(X_i))$ for $S \subseteq \text{NonDes}(X_i)$
- Bayes Nets encode full joint distributions as a product of the conditional distributions on each variable
 - $P(X_1, \dots, X_n) = \prod_i P(X_i | \text{Par}(X_i))$

Example: Alarm Network



$$P(b, \neg e, a, \neg j, \neg m) =$$

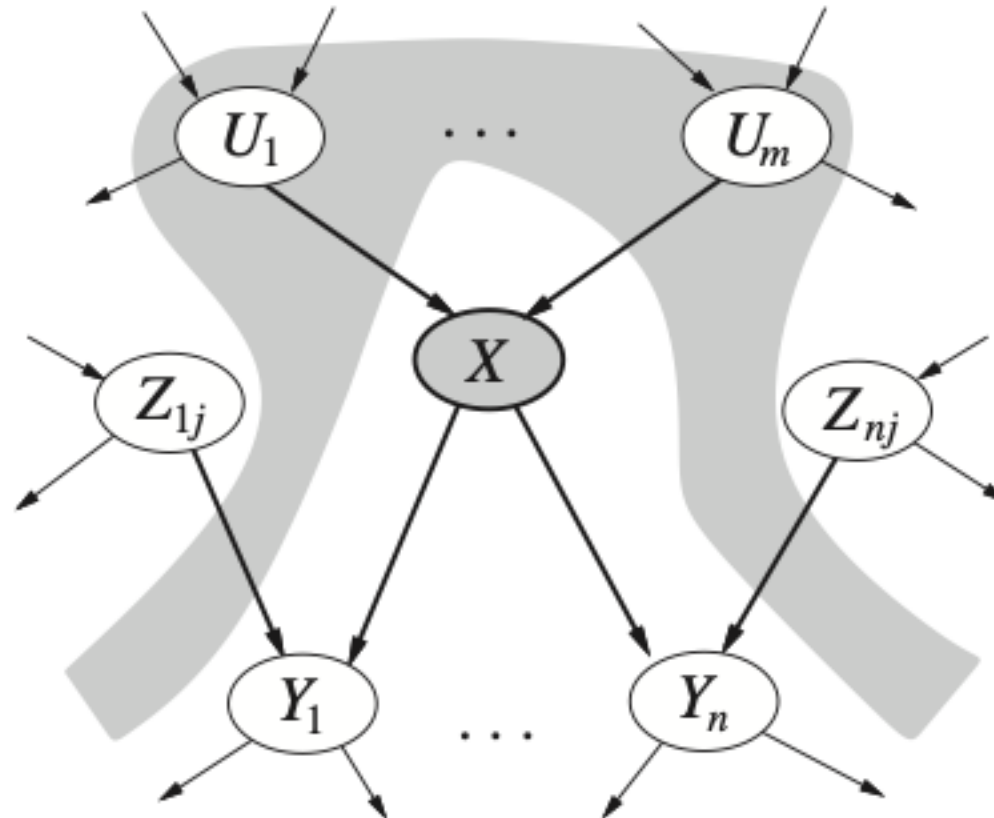
$$P(b) P(\neg e) P(a|b, \neg e) P(\neg j|a) P(\neg m|a)$$

$$=.001 \times .998 \times .94 \times .1 \times .3$$

$$=.000028$$

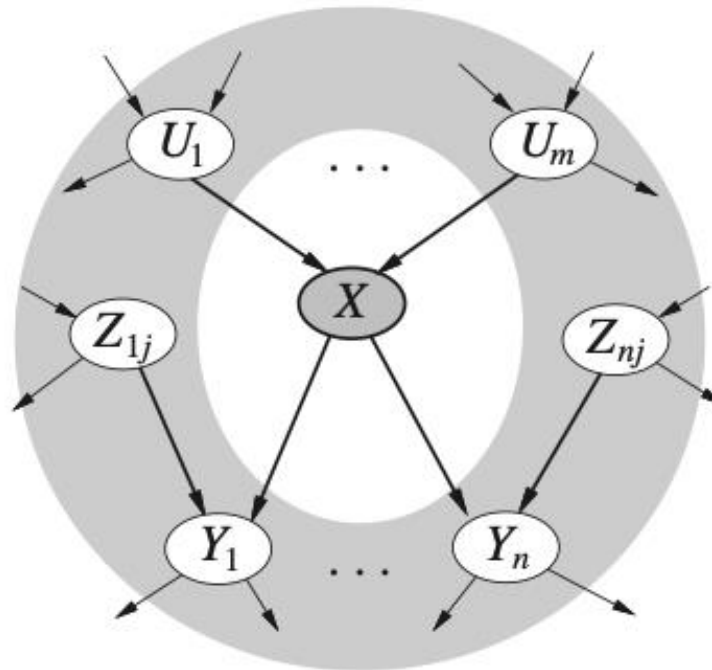
Independence

- We can look at the structure of a Bayes Net and understand conditional independence relationships between variables.



Markov Blanket

- A variable's Markov blanket consists of its parents, its children, and its children's parents
- Every variable is conditionally independent of all other variables given its Markov blanket



Other Independence Relationships

- Are X and Z necessarily independent?

- No: low pressure causes rain which causes traffic
- X can influence Z, Z can influence X (through Y)
- But they could be independent? How?

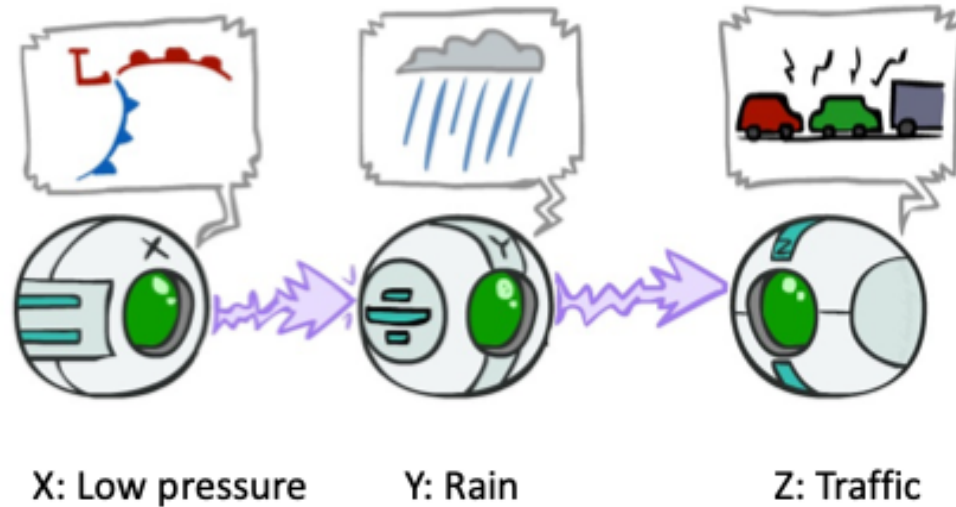


- We use an approach called D-separation to identify relationships between triplets of variables

- We look for patterns in paths in the underlying graph structure of the Bayes Net

Causal Chains

- This configuration is a “causal chain”



$$P(x, y, z) = P(y)P(x|y)P(z|y)$$

- Guaranteed X independent of Z given Y?

$$\begin{aligned} P(z|x, y) &= \frac{P(x, y, z)}{P(x, y)} \\ &= \frac{P(x)P(y|x)P(z|y)}{P(x)P(y|x)} \\ &= P(z|y) \end{aligned}$$

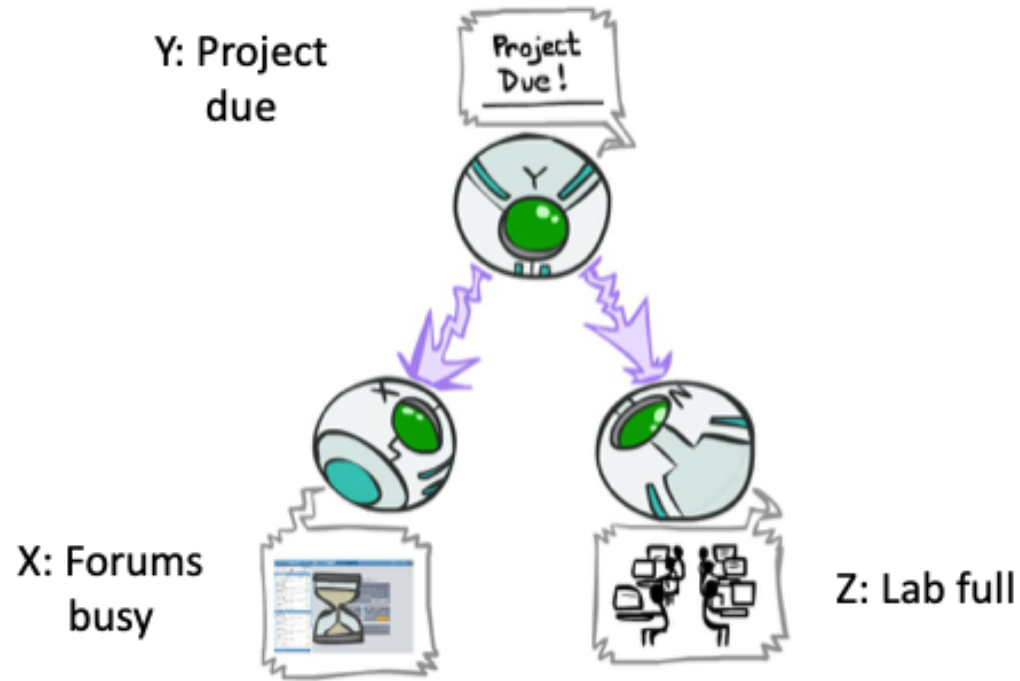
Yes!

- Evidence along the chain “blocks” the influence

Common Cause

- This configuration is a “common cause”

- Guaranteed X and Z independent given Y?



$$P(x, y, z) = P(y)P(x|y)P(z|y)$$

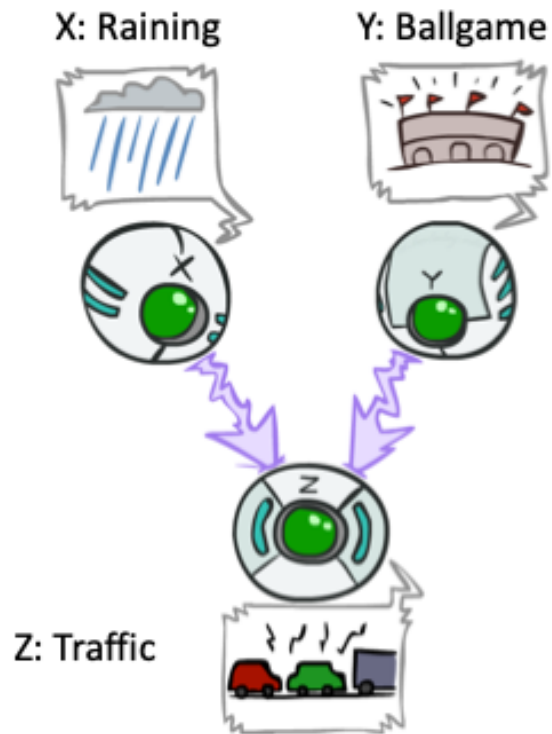
$$\begin{aligned} P(z|x, y) &= \frac{P(x, y, z)}{P(x, y)} \\ &= \frac{P(y)P(x|y)P(z|y)}{P(y)P(x|y)} \\ &= P(z|y) \end{aligned}$$

Yes!

- Observing the cause blocks influence between effects.

Common Effect

- Last configuration: two causes of one effect (v-structures)

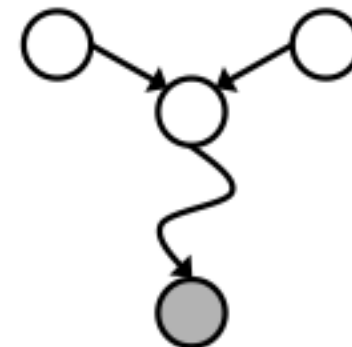
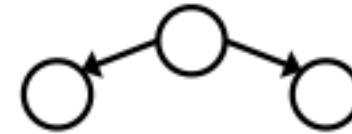


- Are X and Y independent?
 - **Yes**: the ballgame and the rain cause traffic, but they are not correlated
 - (Proved previously)
- Are X and Y independent given Z?
 - **No**: seeing traffic puts the rain and the ballgame in competition as explanation.
- **This is backwards from the other cases**
 - Observing an effect **activates** influence between possible causes.

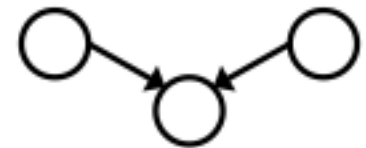
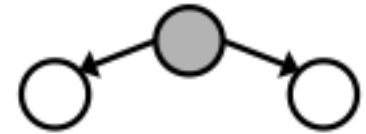
Active/Inactive Paths

- Are X and Y conditionally independent given evidence variables {Z}?
 - Yes, if X and Y are d-separated by Z
 - Consider all (undirect) paths from X to Y
 - No active paths = independence!
- A path is active if each triple is active:
 - Causal chain: $A \rightarrow B \rightarrow C$ where B is unobserved
 - Common cause: $A \leftarrow B \rightarrow C$ where B is unobserved
 - Common effect: $A \rightarrow B \leftarrow C$ where B is or one of its descendants is observed
- All it takes to block a path is a single inactive path

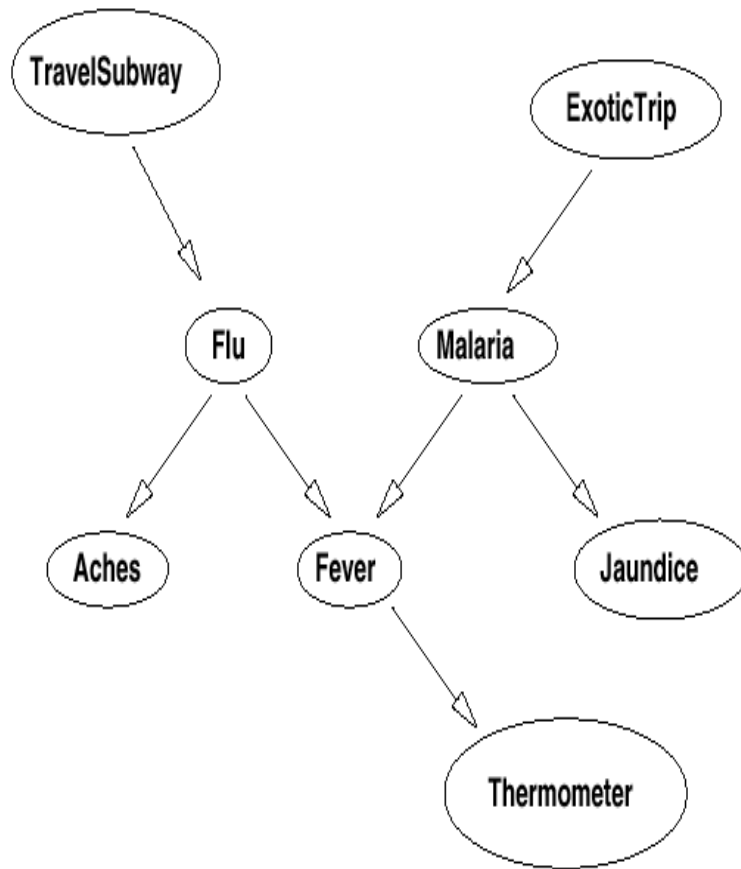
Active Triples



Inactive Triples

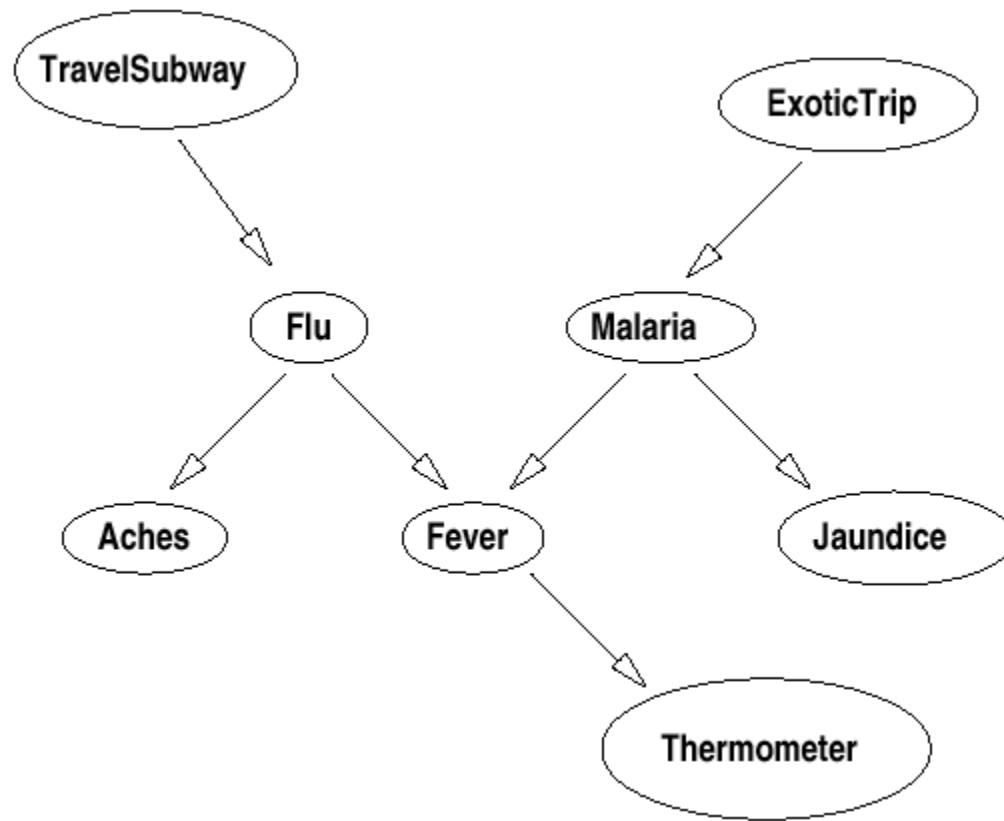


Example



1. Subway and Thermometer?
2. Aches and Fever?
3. Aches and Thermometer?
4. Flu and Malaria?
5. Subway and ExoticTrip?

Example



$P(J)=$

$P(J|ET) =$

$P(Fev)=$

$P(Fev| TS=ts, Mal=\sim m) =$

$P(ET|J=j)=$

Variable Elimination

- To do probabilistic inference using a Bayes Net we often use the Variable Elimination algorithm
 - Applies repeated marginalization steps while exploiting the structure of the Bayes Net (conditional independence)
- Factors: a function $f(X_1, \dots, X_k)$
 - Think of this as a table of numbers, one number for each instantiation of the variables
 - CPTs in a BN are factors
 - $P(C|A,B)$ is a function of 3 variables A, B, C (represented as $f(A,B,C)$)
 - We will operate on factors (multiply, sum-out, restrict (and normalize))

Products of Factors

- Let $f(\mathbf{X}, \mathbf{Y})$ and $g(\mathbf{Y}, \mathbf{Z})$ be two factors with shared variables $\mathbf{Y}$
- The product factor is $h(\mathbf{X}, \mathbf{Y}, \mathbf{Z}) = f(\mathbf{X}, \mathbf{Y})g(\mathbf{Y}, \mathbf{Z})$ is computed via pointwise multiplication

$f(A,B)$		$g(B,C)$		$h(A,B,C)$			
ab	0.9	bc	0.7	abc	0.63	$ab \sim c$	0.27
$a \sim b$	0.1	$b \sim c$	0.3	$a \sim bc$	0.08	$a \sim b \sim c$	0.02
$\sim ab$	0.4	$\sim bc$	0.8	$\sim abc$	0.28	$\sim ab \sim c$	0.12
$\sim a \sim b$	0.6	$\sim b \sim c$	0.2	$\sim a \sim bc$	0.48	$\sim a \sim b \sim c$	0.12

Summing Out

- Let $f(X, \mathbf{Y})$ be a factor with variable X and variable set $\mathbf{Y}$
- $h(X, \mathbf{Y}) = \sum_{x \in X} P(x, Y)$

f(A,B)		h(B)	
ab	0.9	b	1.3
a~b	0.1	~b	0.7
~ab	0.4		
~a~b	0.6		

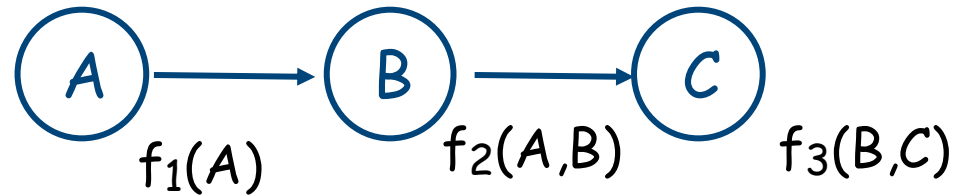
Restricting a Factor

- Let $f(X,Y)$ be a factor with variable X
- Restrict f to $X=x$ by setting $X=x$ and “deleting”, $h(Y) = f(X=x, Y)$

$f(A,B)$		$h(B) = f_{A=a}$	
ab	0.9	b	0.9
a~b	0.1	~b	0.1
~ab	0.4		
~a~b	0.6		

Variable Elimination: No Evidence

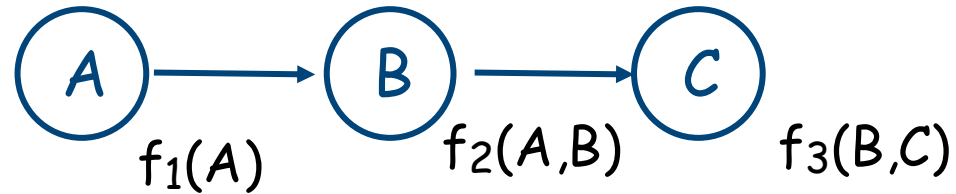
- Computing a prior probability on variable X can be seen as applying factor operations on the CPTs



- $$\begin{aligned} P(C) &= \sum_{A,B} P(C|B) P(B|A) P(A) \\ &= \sum_B P(C|B) \sum_A P(B|A) P(A) \\ &= \sum_B f_3(B,C) \sum_A f_2(A,B) f_1(A) \\ &= \sum_B f_3(B,C) f_4(B) \\ &= f_5(C) \end{aligned}$$

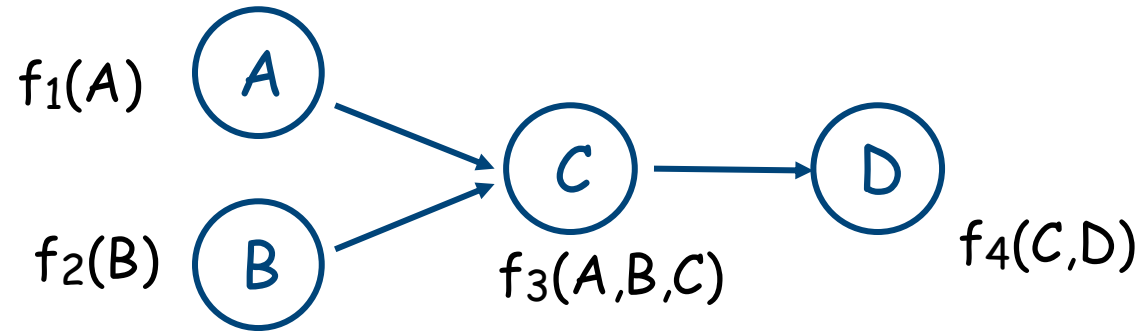
Define new factors: $f_4(B) = \sum_A f_2(A,B) f_1(A)$ and $f_5(C) = \sum_B f_3(B,C) f_4(B)$

Variable Elimination: No Evidence



$f_1(A)$		$f_2(A,B)$		$f_3(B,C)$		$f_4(B)$		$f_5(C)$	
a	0.9	ab	0.9	bc	0.7	b	0.85	c	0.625
$\sim a$	0.1	$a \sim b$	0.1	$b \sim c$	0.3	$\sim b$	0.15	$\sim c$	0.375
		$\sim ab$	0.4	$\sim bc$	0.2				
		$\sim a \sim b$	0.6	$\sim b \sim c$	0.8				

Variable Elimination: No Evidence



$$\begin{aligned} P(D) &= \sum_{A,B,C} P(D|C) P(C|B,A) P(B) P(A) \\ &= \sum_C P(D|C) \sum_B P(B) \sum_A P(C|B,A) P(A) \\ &= \sum_C f_4(C,D) \sum_B f_2(B) \sum_A f_3(A,B,C) f_1(A) \\ &= \sum_C f_4(C,D) \sum_B f_2(B) f_5(B,C) \\ &= \sum_C f_4(C,D) f_6(C) \\ &= f_7(D) \end{aligned}$$

Define new factors: $f_5(B,C)$, $f_6(C)$, $f_7(D)$, in the obvious way

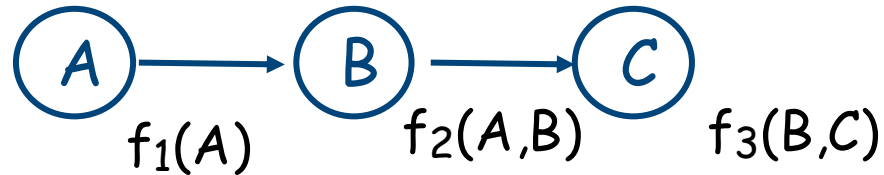
Variable Elimination: No Evidence

Given Query variable Q , remaining variables Z . Let F be the set of factors corresponding to the CPTs for Q and Z .

1. Choose an elimination ordering $Z_1, \dots, Z_n$ of variables in Z .
2. For each Z_j -- in the order given -- eliminate $Z_j \in Z$ as follows:
 - (a) Compute new factor $g_j = \sum_{Z_j} f_1 \times f_2 \times \dots \times f_k$,
where the f_i are the factors in F that include Z_j
 - (b) Remove the factors f_i (that mention Z_j) from F
and add new factor g_j to F
3. The remaining factors refer only to the query variable Q .
Take their product and normalize to produce $P(Q)$

Variable Elimination: Evidence

- Computing the posterior of a query variable given some evidence is similar.
Suppose we observe $C=c$



$$\begin{aligned} P(A|c) &= \alpha P(A) P(c|A) \\ &= \alpha P(A) \sum_B P(c|B) P(B|A) \\ &= \alpha f_1(A) \sum_B f_3(B,c) f_2(A,B) \\ &= \alpha f_1(A) \sum_B f_4(B) f_2(A,B) \\ &= \alpha f_1(A) f_5(A) \\ &= \alpha f_6(A) \end{aligned}$$

New factors: $f_4(B) = f_3(B,c)$; $f_5(A) = \sum_B f_2(A,B) f_4(B)$;
 $f_6(A) = f_1(A) f_5(A)$

Variable Elimination: Evidence

- Given Query variable Q , Evidence $E=e$, remaining variables Z . Let F be the set of factors corresponding to the CPTs for Q and Z .

1. Replace each factor $f \in F$ that mentions a variable(s) in E with its restriction $f_{E=e}$ (somewhat abusing notation)
2. Choose an elimination ordering $Z_1, \dots, Z_n$ of variables in Z .
3. Run VE as before.
4. Until remaining factors refer only to Q . Take their product and normalize to produce $P(Q)$.

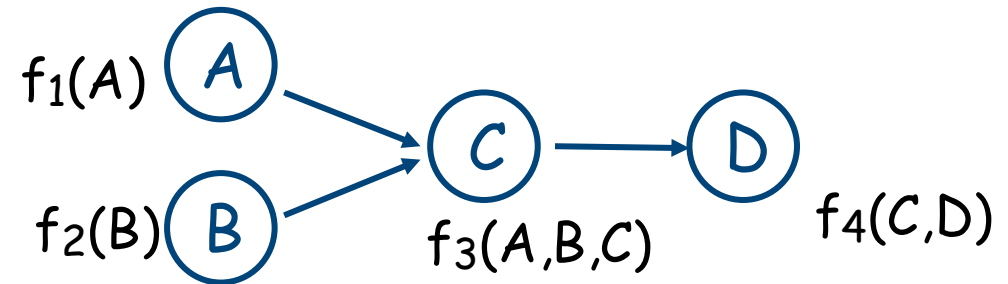
Example

Factors: $f_1(A)$ $f_2(B)$
 $f_3(A,B,C)$ $f_4(C,D)$

Query: $P(A)?$

Evidence: $D = d$

Elim. Order: C, B



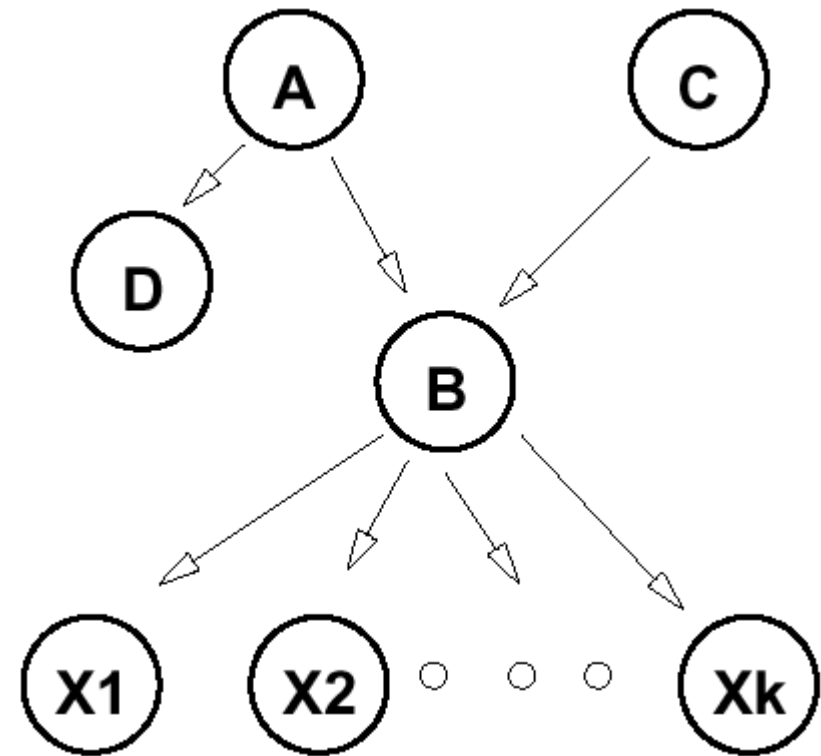
VE Notes

- Each iteration eliminates a single variable
 - Therefore the number of iterations is linear in the number of variables
 - But each iteration can be expensive (exponential in the size of the largest factors)
 - Can't do better than the size of the Bayes Net (the original factors are input to the VE algorithm) BUT when we create new factors they may grow in size significantly
 - We try to be careful in the order we eliminate factors so as to control the size of the factors

Example

What happens if we eliminate B first?

What happens if we eliminate some X_i first?



Where do Bayes Nets Come From?

- Handcrafted from expert knowledge
- Empirical data coupled with human guidance
- Learn them directly from data (both the CPTs and the topology)

From: [Applications of Bayesian network models in predicting types of hematological malignancies](#)

