COMS W4701: Artificial Intelligence

Lecture 17: Inference in Bayes Nets

Tony Dear, Ph.D.
Department of Computer Science
School of Engineering and Applied Sciences

Today

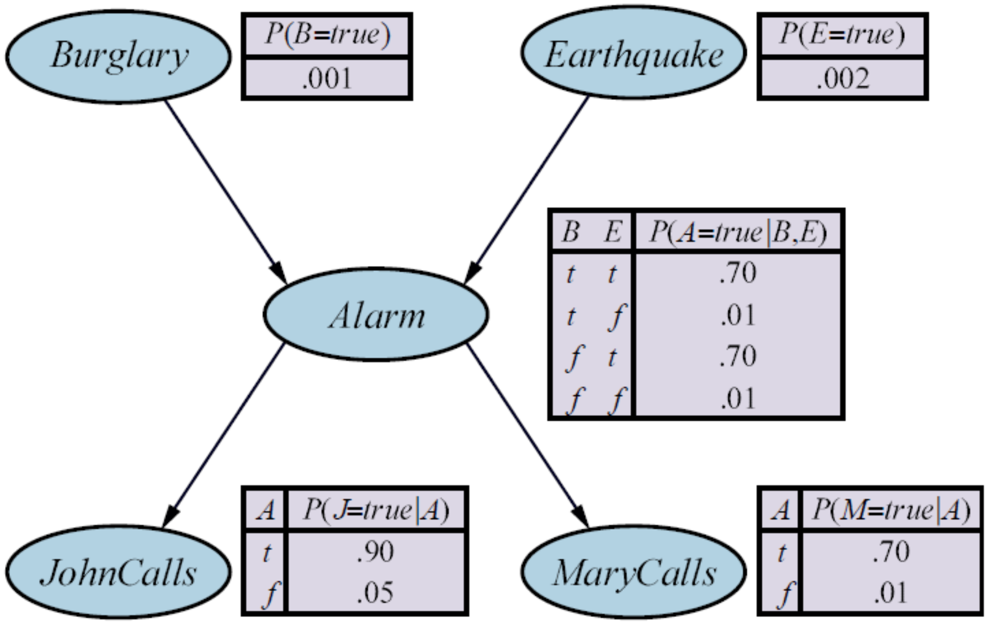
- Inference in Bayes nets
- Inference by enumeration
- Variable elimination

Inference in Bayes Nets

- We are interested in computing probabilities of **query** variables X

- General strategy: Apply the *chain rule* to the CPT parameters

$$P(+b,-e,+a) = P(+b)P(-e)P(+a|+b,-e)$$
$$= (.001)(.998)(.01) = 9.98 \times 10^{-6}$$



- May have to first include, then marginalize out, **hidden** variables Y (ancestors of X)

$$P(+a) = \sum_{b,e} P(b,e,+a) = \sum_{b,e} P(b)P(e)P(+a|b,e)$$
$$= (.001)(.002)(.7) + (.001)(.998)(.01) + (.999)(.002)(.7) + (.999)(.998)(.01) = .01138$$

Evidence Variables

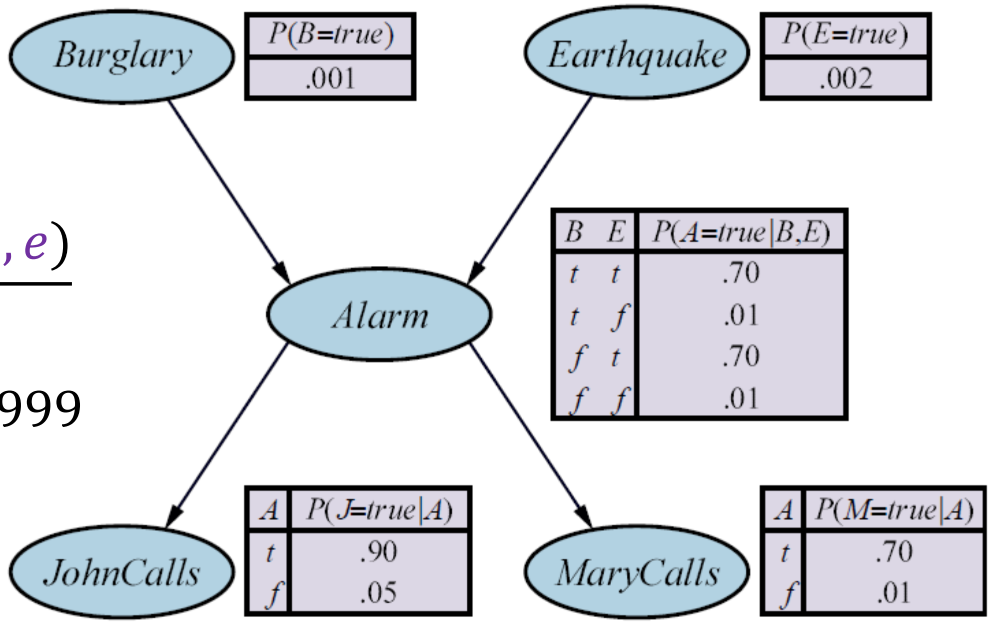
- *Posterior probabilities* condition on observed **evidence** variables E

e

- Hidden variables must now include ancestors of both X and E

$$P(-b|+a) = \frac{\sum_e P(-b,e,+a)}{P(+a)} = \frac{\sum_e P(-b)P(e)P(+a|-b,e)}{P(+a)}$$

$$= \big((.999)(.002)(.7) + (.999)(.998)(.01)\big)/.01138 = .999$$



- Can exploit conditional independences to simplify calculations

$$P(+j,+m|-a,+e,+b) = P(+j,+m|-a) = P(+j|-a)P(+m|-a) = (0.05)(0.01) = 0.0005$$

Inference by Enumeration

- General task: Find the *posterior* distribution of a set of **query** variables X given a set of observed **evidence** e

- *Enumeration* strategy: Compute joint probabilities using chain rule
- We may have to include **hidden** variables Y from ancestors of X and E

$$\boxed{P(\boldsymbol{X} \mid \boldsymbol{e}) \propto P(\boldsymbol{X}, \boldsymbol{e}) = \sum_{\boldsymbol{y}} P(\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{e})}$$

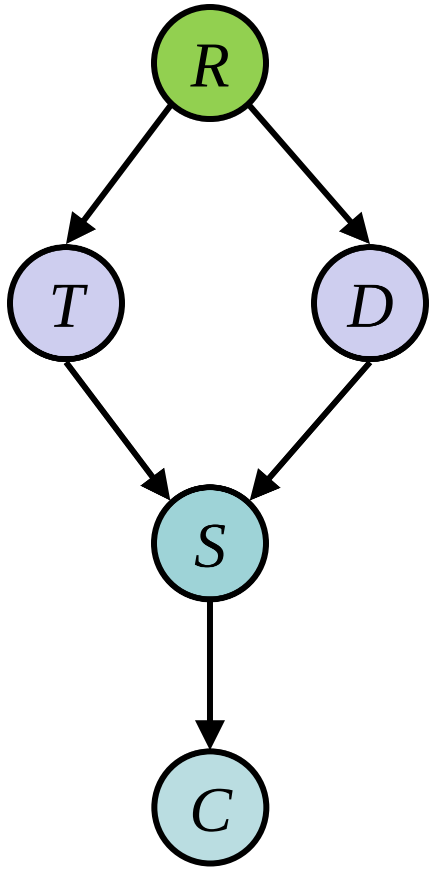
- After computing joint probs, we can simply *normalize* to obtain posterior

Inference by Enumeration

- Computational complexity of enumeration is generally exponential in number of query and hidden variables

$$P(R|+s) \propto P(R,+s) = \sum_{t,d} P(R,t,d,+s)$$
$$= \sum_{t,d} P(R)P(t|R)P(d|R)P(+s|t,d)$$

- Each term in the product is a *subset* of a Bayes net CPT
- Size of each term depends on how many non-evidence variables it depends on



Factor Representation

- Think of each term as a **factor** f_i indexed by the input variables:

$$P(R|+s) \propto \sum_{t,d} P(R)P(t|R)P(d|R)P(+s|t,d)$$

s不在input里 因为s是fixed的

$$= \sum_{t,d} f_R(R) f_T(R,t) f_D(R,d) f_S(t,d)$$

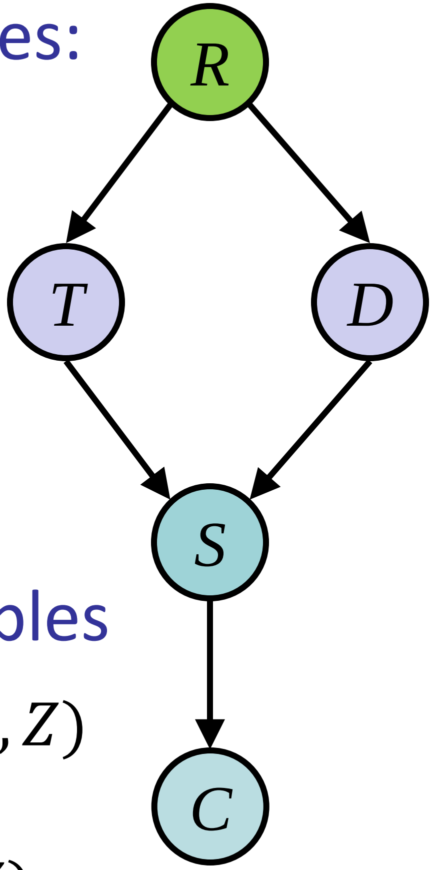
这些是factor representation

- *Multiplying* factors means *elementwise* multiplication over *common* variables, producing new factor in union of all variables

$$f_1(X,Y) \times f_2(Y,Z) = f_3(X,Y,Z)$$

- *Summing* over a factor is the same as marginalization of a joint distribution

$$\sum_y f_3(X,y,Z) = f_4(X,Z)$$



Example

$$P(R|+s) \propto P(R,+s) = \sum_{t,d} P(R)P(t|R)P(d|R)P(+s|t,d)$$

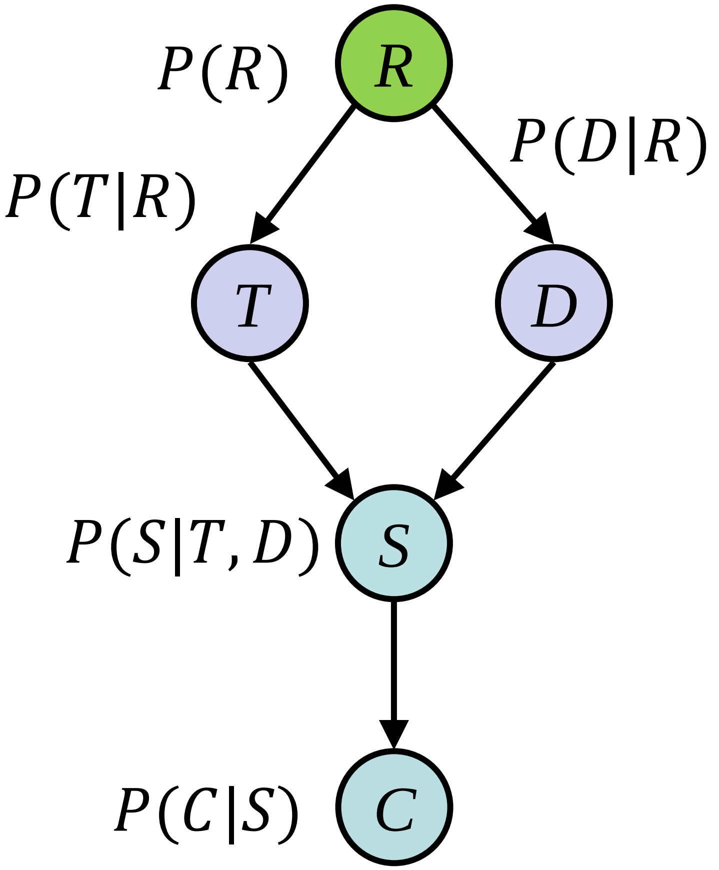
- Bayes net CPT parameters:

R	$P(R)$
+r	0.5
-r	0.5

T	R	$P(T\|R)$
+t	+r	0.7
+t	-r	0.6
-t	+r	0.3
-t	-r	0.4

D	R	$P(D\|R)$
+d	+r	0.7
+d	-r	0.6
-d	+r	0.3
-d	-r	0.4

T	D	S	$P(S\|T,D)$
+t	+d	+s	0.1
+t	-d	+s	0.4
-t	+d	+s	0.2
-t	-d	+s	0.9
+t	+d	-s	0.9
+t	-d	-s	0.6
-t	+d	-s	0.8
-t	-d	-s	0.1



Example

$$P(R|+s) \propto P(R,+s) = \sum_{t,d} P(R)P(t|R)P(d|R)P(+s|t,d)$$

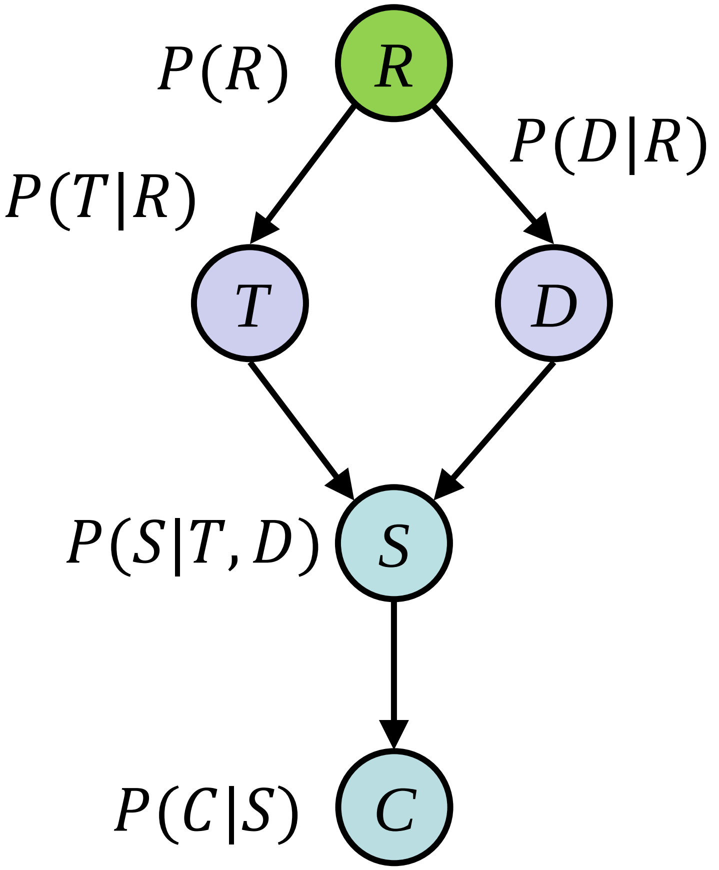
- Joint probabilities via factor multiplication

R	T	D	$P(R,T,D,+s)$
+r	+t	+d	(0.5)(0.7)(0.7)(0.1)
+r	+t	-d	(0.5)(0.7)(0.3)(0.4)
+r	-t	+d	(0.5)(0.3)(0.7)(0.2)
+r	-t	-d	(0.5)(0.3)(0.3)(0.9)
-r	+t	+d	(0.5)(0.6)(0.6)(0.1)
-r	+t	-d	(0.5)(0.6)(0.4)(0.4)
-r	-t	+d	(0.5)(0.4)(0.6)(0.2)
-r	-t	-d	(0.5)(0.4)(0.4)(0.9)

Joint CPT size: $2^3 = 8$ rows

- Marginalize over T and D and normalize to obtain $P(R|+s)$:

R	$P(R\|+s)$
+r	0.441
-r	0.559



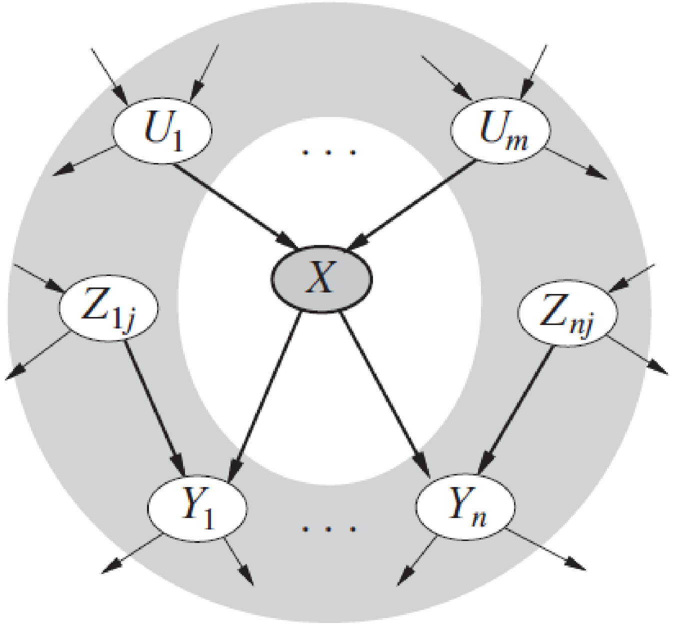
Example: Markov Blanket

- What if we want the posterior of X conditioned on *all* other variables?
- Observing *Markov blanket* of X makes it independent of other variables
- We simply compute a product of factors over X and its children nodes

$$P(X|mb(X)) \propto P(X|parents(X)) \times \prod_{Y_j} P(y_j|parents(Y_j))$$

P(parents(X) I …)

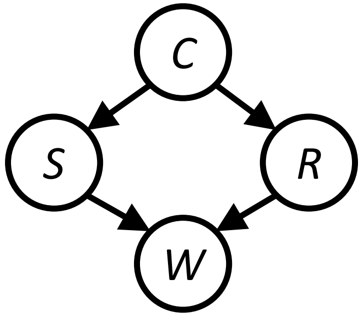
- Can exclude factors over $parents(X)$, since they are completely observed and just constants
- Factors over X and $children(X)$ all just have a single input variable in X



Example: Markov Blanket

C	P(+s|C)
+c	0.1
-c	0.5

C	P(C)
+c	0.5



C	P(+r|C)
+c	0.8
-c	0.2

$$P(C \mid mb(C)) = P(C \mid s, r) \propto P(C)P(s|C)P(r|C)$$

$$P(S \mid mb(S)) = P(S \mid c, r, w) \propto P(S|c)P(w|S, r)$$

$$P(R \mid mb(R)) = P(R|c, s, w) \propto P(R|c)P(w|s, R)$$

$$P(W \mid mb(W)) = P(W \mid s, r)$$

C	P(C,+s,+r)
+c	0.04
-c	0.05

=

C	P(C)
+c	0.5
-c	0.5

×

C	P(+s|C)
+c	0.1
-c	0.5

×

C	P(+r|C)
+c	0.8
-c	0.2

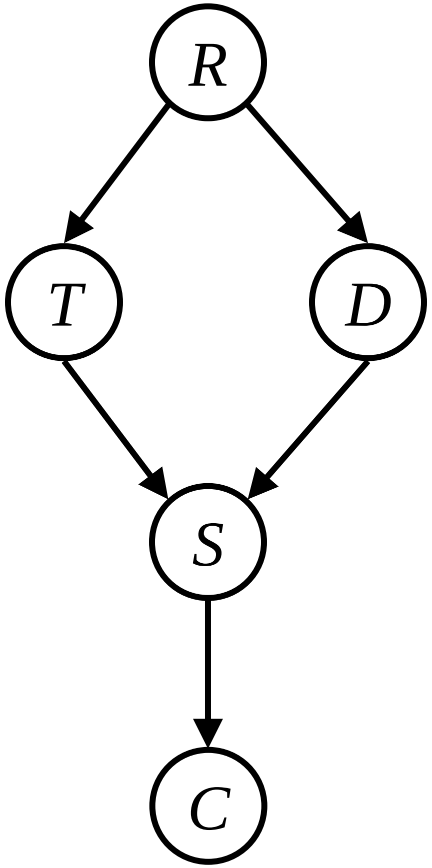
$$P(C|mb(C)) = \left(\frac{4}{9}, \frac{5}{9}\right)$$

Inference Complexity

- Inference complexity depends on the size of the joint distribution
- But we do not have to wait to sum over all variables at the end!

- Better idea: Perform summation over each variable independently
- Factors not dependent on X can be *taken out* of a summation over X

- Ex: $uwy + uwz + uxy + uxz + vwy + vwz + vxy + vxz$ has 16 multiplies and 7 adds
- $(u + v)(wy + wz + xy + xz)$ has 5 multiplies and 4 adds
- $(u + v)(w + x)(y + z)$ has 2 multiplies and 3 adds



Variable Elimination

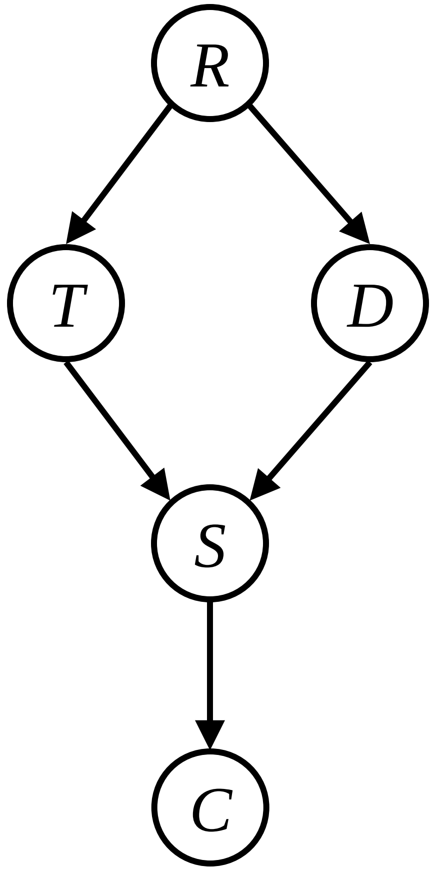
- Idea: Move summations as far *inwards* as possible
- Marginalization is done starting inside and moving outward

$$P(S|r) \propto P(S,r) = \sum_{t,d} P(r)P(t|r)P(d|r)P(S|t,d)$$

$$= P(r)\sum_{t} P(t|r)\sum_{d} P(d|r)P(S|t,d)$$

$$P(S|c) \propto P(S,c) = \sum_{r,t,d} P(r)P(t|r)P(d|r)P(S|t,d)P(c|S)$$

$$= P(c|S)\sum_{t,d} P(S|t,d)\sum_{r} P(r)P(t|r)P(d|r)$$



Example: Variable Elimination

$$P(S,c) = P(c|S)\sum_{t,d} P(S|t,d)\sum_{r} P(r)P(t|r)P(d|r) = f_C(S)\sum_{t,d} f_S(S,t,d)\ \sum_{r} f_R(r)f_T(t,r)f_D(d,r)$$

c is observed 是constant fixed的

R	T	D	$f_1(R,T,D)$
+r	+t	+d	0.245
+r	+t	-d	0.105
+r	-t	+d	0.105
+r	-t	-d	0.045
-r	+t	+d	0.180
-r	+t	-d	0.120
-r	-t	+d	0.120
-r	-t	-d	0.080

$\sum_r$

T	D	$f_2(T,D)$
+t	+d	0.425
+t	-d	0.225
-t	+d	0.225
-t	-d	0.125

$\times f_S$

S	T	D	$f_3(S,T,D)$
+s	+t	+d	0.0425
+s	+t	-d	0.3825
+s	-t	+d	0.09
+s	-t	-d	0.135
-s	+t	+d	0.045
-s	+t	-d	0.18
-s	-t	+d	0.1125
-s	-t	-d	0.0125

$\sum_{t,d}$

S	$f_4(S)$
+s	0.29
-s	0.71

$\times f_C$

S	$f_5(S)$
+s	0.116
-s	0.497

normalize

S	$P(S\|c)$
+s	0.189
-s	0.811

作业可以用 bayes.net ipynb

Variable Ordering

- Elimination ordering does not affect correctness of inference, but *does* greatly affect computational efficiency!

$$P(S,c) = \sum_{r,t,d} f_1(R) f_2(T,R) f_3(D,R) f_4(S,T,D) f_5(S)$$

orders 影响计算的复杂度

- R then T then D: $f_5(S) \sum_d \sum_t f_4(S,T,D) \sum_r \underbrace{f_1(R) f_2(T,R) f_3(D,R)}_{\text{8 rows}}$
- 26 multiplies, 10 adds

- T then D then R: $f_5(S) \sum_r f_1(R) \sum_d f_3(D,R) \sum_t \underbrace{f_2(T,R)\, f_4(S,T,D)}_{\text{16 rows}}$
- 30 multiplies, 14 adds

Improving Complexity

- Elimination complexity depends on size of the largest constructed CPT
- NP-hard in the worst case, as this can reduce to a satisfiability problem

- *Greedy* variable ordering can be a good heuristic: Select the next variable that minimizes the size of the constructed CPT
- Still no guarantee of optimal variable ordering

- If Bayes net is a **polytree** (replace all directed edges with undirected edges), elimination can be *linear* if we eliminate *leaves first, then root*

Summary

- Inference in Bayesian networks: Computing distributions over query variables given evidence variables (and marginalizing hidden variables)
- Inference by enumeration: Compute full joint distribution of all relevant variables using chain rule, then marginalize hidden variables
- Variable elimination: Alternate between building up and summing out CPTs to reduce computational complexity
- Overall still a NP-hard problem