BAYES' NETS INFERENCE

Lara J. Martin (she/they)
TA: Aydin Ayanzadeh (he)

11/14/2023
CMSC 671

By the end of class today, you will be able to:

- Draw connections between inference by enumeration with probability (MLE) and Bayes' nets
- Eliminate variables for Bayes' net inference

Modified from slides by Dr. Cynthia Matuszek

REVIEW: INDEPENDENCE

What does it mean for A and B to be **independent** (P(A) ⫫ P(B))?

- A and B do not affect each other’s probability
  - $P(A,B) = P(A)\,P(B)$
  - $P(x|y) = P(x)$

CONDITIONING

- What does it mean for A and B to be **conditionally independent given C?**
- A and B don't affect each other **if C is known**
- $P(\mathrm{A}, \mathrm{B} \mid \mathrm{C}) = P(\mathrm{A} \mid \mathrm{C})\, P(\mathrm{B} \mid \mathrm{C})$

REVIEW: BAYES' RULE

- What is **Bayes' Rule**?

$$P(H_i \mid E_j) = \frac{P(E_j \mid H_i)P(H_i)}{P(E_j)}$$

- What's it useful for?
  - Diagnosis
  - Effect is perceived, want to know (probability of) cause

$$P(cause \mid effect) = \frac{P(effect \mid cause)P(cause)}{P(effect)}$$

R&N, 495–496

REVIEW: BAYES' RULE

- What is **Bayes' Rule**?

$$P(H_i \mid E_j) = \frac{P(E_j \mid H_i) P(H_i)}{P(E_j)}$$

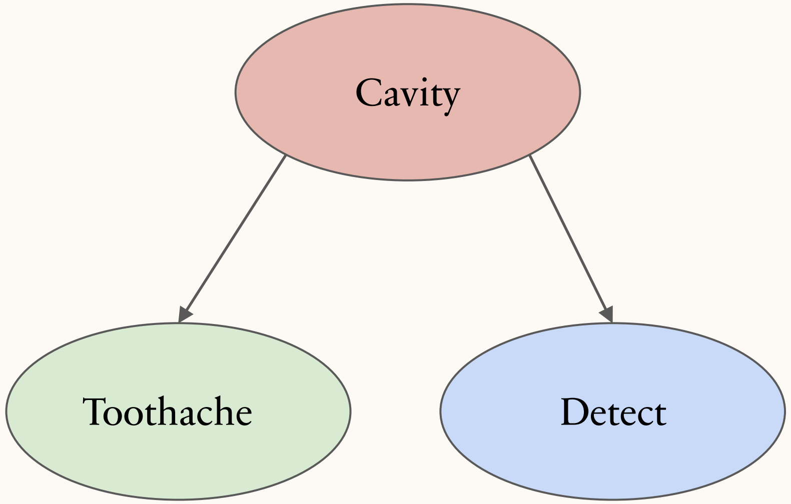
- What's it useful for?
  - Diagnosis
  - Effect is perceived, want to know (probability of) cause

$$P(hidden \mid observed) = \frac{P(observed \mid hidden) P(hidden)}{P(observed)}$$

R&N, 495–496

REVIEW: BAYES' NETS

- Bayesian Network (BN): **BN = (DAG, CPD)**
  - **DAG**: directed acyclic graph (BN's structure)
  - **CPT**: conditional probability table (BN's parameters)



p(Cav)	p(¬Cav)
0.2	0.8

	p(Det|Cav)	p(¬Det|Cav)
Cav=T	0.9	0.1
Cav=F	0.6	0.4

	p(Tth|Cav)	p(¬Tth|Cav)
Cav=T	0.6	0.4
Cav=F	0.1	0.9

BAYES’ NETS BIG PICTURE

- Two problems with using **full joint distribution tables** as our probabilistic models:
  - Unless there are only a few variables, the joint is *way* too big to represent explicitly
  - Hard to learn (estimate) anything empirically about more than a few variables at a time

- **Bayes’ nets**: a technique for describing complex joint distributions (models) using simple, local distributions (conditional probabilities)
  - More properly called graphical models
  - We describe how variables locally interact
  - Local interactions chain together to give global, indirect interactions

Slide from Dr. Chris Callison-Burch

BAYESIAN DIAGNOSTIC REASONING

- Bayes' rule (extended) says that
  - $P(H_i \mid E_1, \ldots, E_m) = P(E_1, \ldots, E_m \mid H_i)\, P(H_i) / P(E_1, \ldots, E_m)$
- Assume each piece of evidence E_i is conditionally independent of the others, **given** a hypothesis H_i, then:
  - $P(E_1, \ldots, E_m \mid H_i) = \prod^{l}_{j=1} P(E_j \mid H_i)$
- If we only care about relative probabilities for the H_i, then we have:
  - $P(H_i \mid E_1, \ldots, E_m) = \alpha\, P(H_i) \prod^{l}_{j=1} P(E_j \mid H_i)$

REVIEW: THE CHAIN RULE

- $P(\alpha_{1,}\alpha_{2,}\ldots,\alpha_n) =$

$$P(\alpha_1) \times$$
$$P(\alpha_2 \mid \alpha_1) \times$$
$$P(\alpha_3 \mid \alpha_{1,}\alpha_2) \times \ldots \times$$
$$P(\alpha_n \mid \alpha_{1,}\cdots,\alpha_{n\text{-}1})$$

$$= \quad \prod\nolimits_{i=1..n} P(\alpha_i \mid \alpha_{1,}\cdots,\alpha_{i\text{-}1})$$

$$= \quad P(x_1,\ldots,x_n) = \mathrm{P}_{i=1}^{n} P(x_i \mid p_i)$$

artint.info/html/ArtInt_143.html

REVIEW: THE CHAIN RULE

$$P(x_1,\ldots,x_n) = \Pi_{i=1}^{n} P(x_i \mid \pi_i)$$

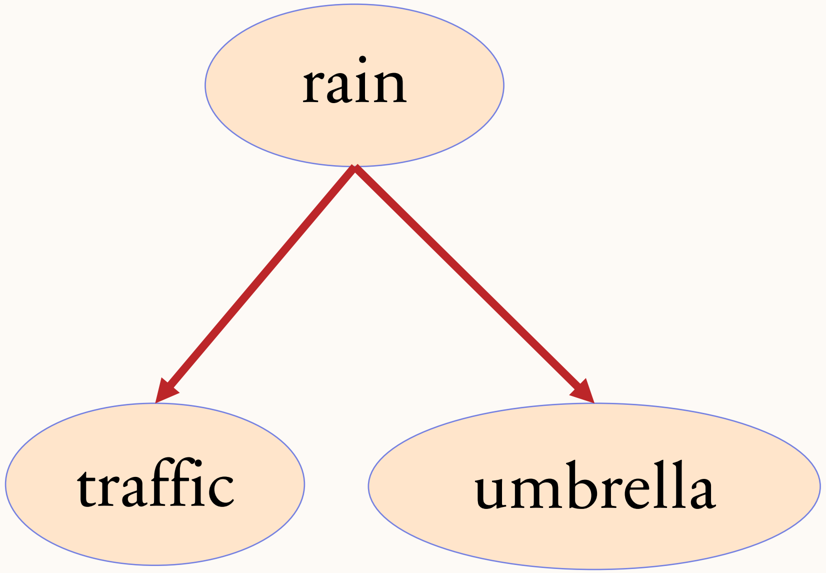
- Decomposition: $P(x_1,\ldots,x_n) = P(x_1)P(x_2 \mid x_1)P(x_3 \mid x_1, x2)\ldots$

  $P(\text{Traffic, Rain, Umbrella}) =$
  $P(\text{Rain})\ P(\text{Traffic} \mid \text{Rain})\ P(\text{Umbrella} \mid \text{Rain, Traffic})$

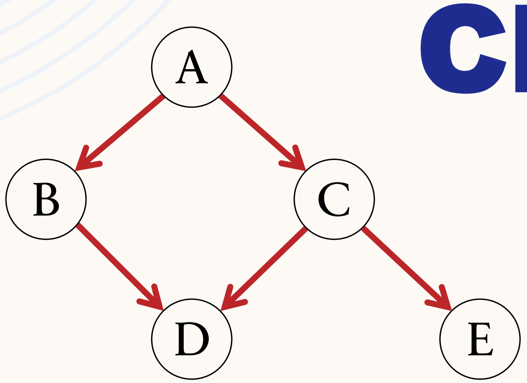
- With assumption of conditional independence:

  $P(\text{Traffic, Rain, Umbrella}) =$
  $P(\text{Rain})\ P(\text{Traffic} \mid \text{Rain})\ P(\text{Umbrella} \mid \text{Rain})$

- Bayes’ nets express conditional independences
  - (Assumptions)



Slides derived from Matt E. Taylor, U Alberta



CHAINING: EXAMPLE

Computing the joint probability for all variables is easy:

$$
\begin{aligned}
P(a, b, c, d, e) &= P(e \mid a, b, \boldsymbol{\mathit{c}}, d)\, P(a, b, c, d) \\
&= P(e \mid c)\, P(a, b, c, d) \\
&= P(e \mid c)\, P(d \mid a, \boldsymbol{\mathit{b}}, \boldsymbol{\mathit{c}})\, P(a, b, c) \\
&= P(e \mid c)\, P(d \mid b, c)\, P(c \mid \boldsymbol{\mathit{a}}, b)\, P(a, b) \\
&= P(e \mid c)\, P(d \mid b, c)\, P(c \mid a)\, P(b \mid a)\, P(a)
\end{aligned}
$$

By product rule

By conditional independence assumption

We're reducing distributions–P(x,y)–to single values.

TOPOLOGICAL SEMANTICS

- Remember: A node is **conditionally independent** of its non-descendants given its parents
- A node is **conditionally independent** of all other nodes in the network given its parents, children, and children's parents (also known as its **Markov blanket**)

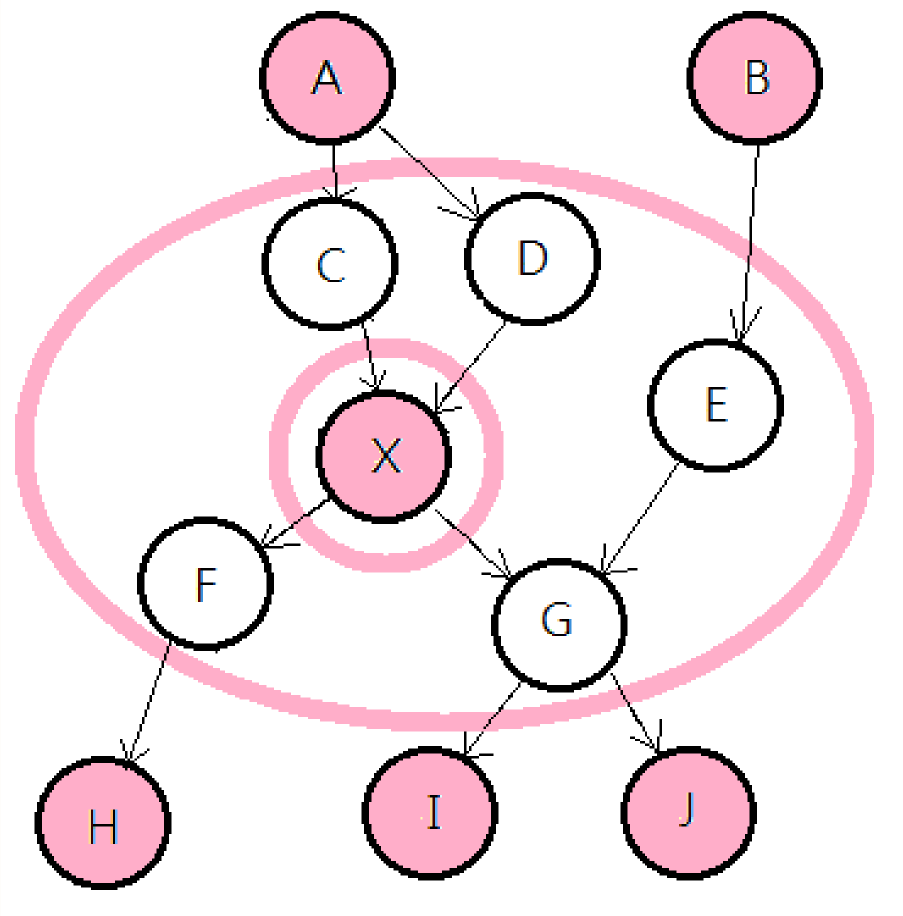
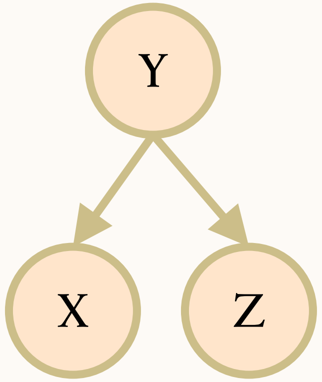
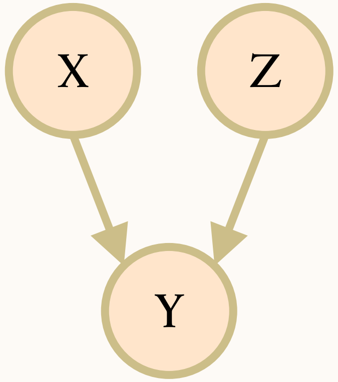


Image: mjtsai1974.github.io/DevBlog/2018/07/11/bayesian-ml-net-profound

INDEPENDENCE WITH CHILDREN

- Common Cause:
  - Y causes X and Y causes Z
  - Are X and Z independent? **No**
  - Are X and Z independent given Y? **Yes**
- Common Effect:
  - Two causes of one effect
  - Are X and Z independent? **Yes**
  - Are X and Z independent given Y? **No**





Observing an effect "activates" influence between possible causes.

Slides derived from Matt E. Taylor, WSU

REVIEW: BAYES' NETS EXAMPLE

What's the probability that

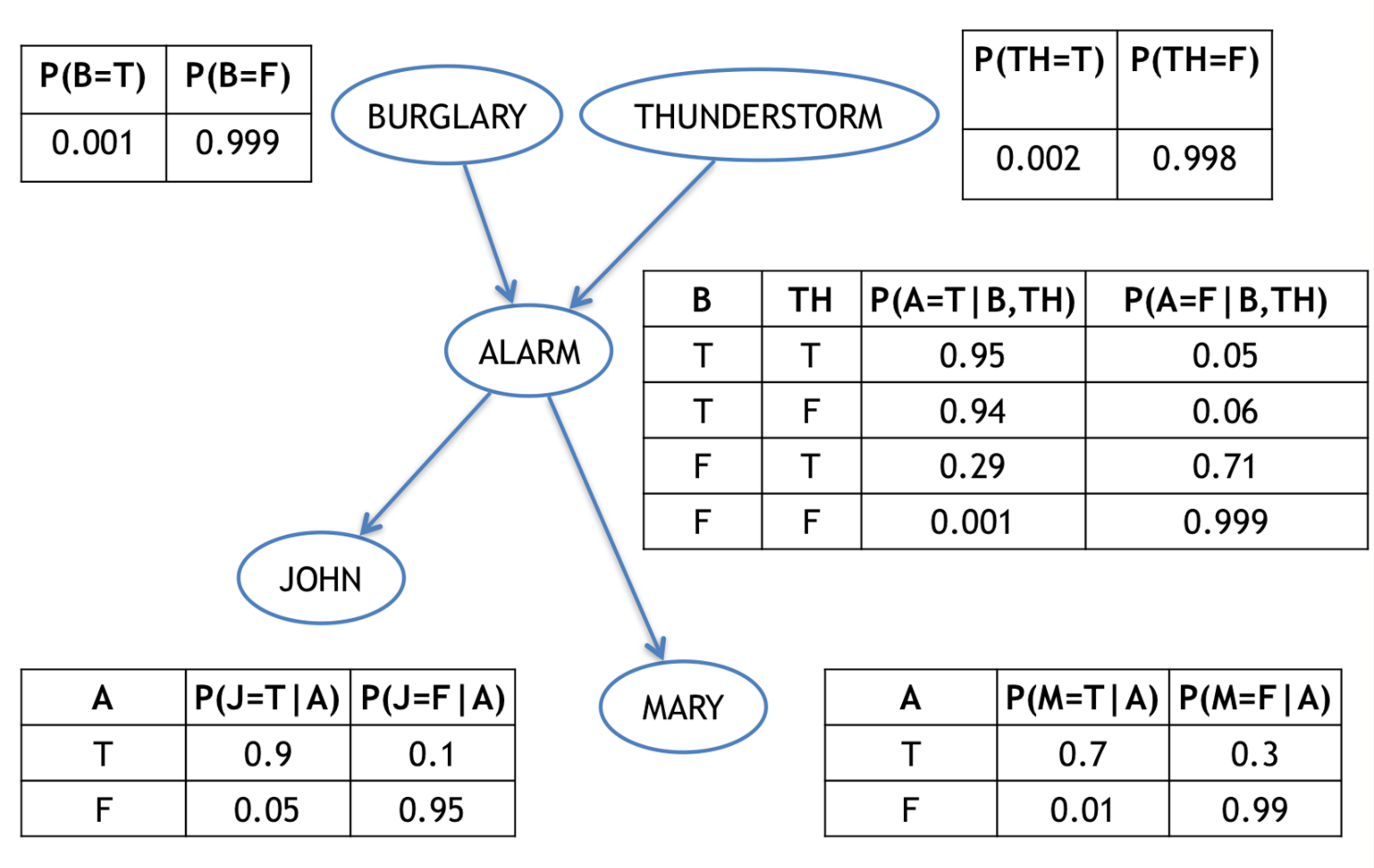
- Both neighbors call
- The alarm goes off
- There is no burglar
- There is no storm

p(j,m,a,¬b,¬t) =
p(j|a) p(m|a) p(a|¬b,¬t) p(¬b) p(¬t) =
(.9) (.7) (.001) (.999) (.998) = 0.00062

Joint probability table: 2^5=32 cells

CPT factorization: 20 cells

$$p(X_1, X_2, \ldots, X_D) = \prod_{i=1}^{D} p(X_i \mid \textsc{Parents}(X_i))$$



P(B=T)	P(B=F)
0.001	0.999

P(TH=T)	P(TH=F)
0.002	0.998

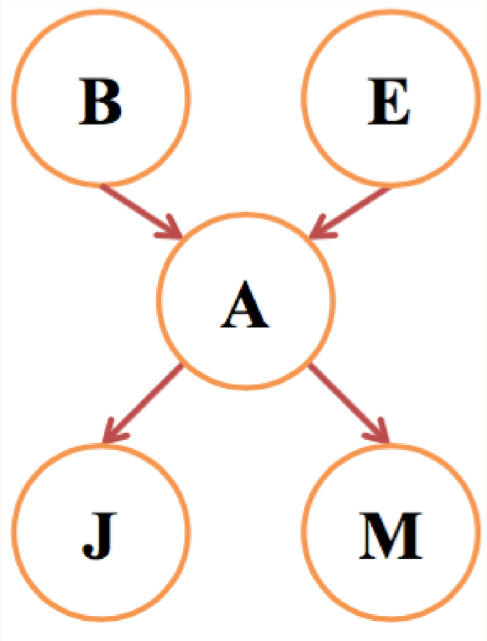
B	TH	P(A=T|B,TH)	P(A=F|B,TH)
T	T	0.95	0.05
T	F	0.94	0.06
F	T	0.29	0.71
F	F	0.001	0.999

A	P(J=T|A)	P(J=F|A)
T	0.9	0.1
F	0.05	0.95

A	P(M=T|A)	P(M=F|A)
T	0.7	0.3
F	0.01	0.99

CONDITIONALITY EXAMPLE

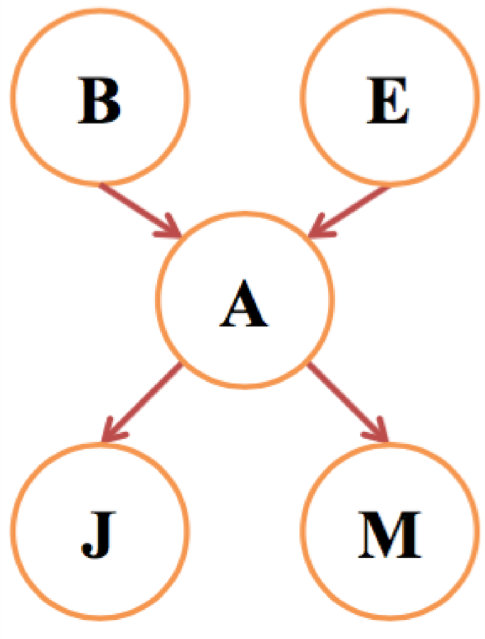
- Hidden: *A*, *B*, *E*. You don't know:
  - If there's a burglar.
  - If there was an earthquake.
  - If the alarm is going off.
- Observed: *J* and *M*.
  - John and/or Mary have some chance of calling if the alarm rings.
  - You know who called you.



Slides derived from Matt E. Taylor, WSU

CONDITIONALITY EXAMPLE 2

- At first:
  - Is the probability of John calling affected by whether there's an earthquake?
  - Is the probability of Mary calling affected by John calling?
- Your alarm is going off!
  - Is the probability of Mary calling affected by John calling?



INFERENCE TASKS

- **Simple queries:** Compute posterior marginal $P(X_i \mid E{=}\textit{value})$
  - E.g., P(NoGas | Gauge=*empty*, Lights=*on*, Starts=*false*)
- **Conjunctive queries:**
  - $P(X_i, X_j \mid E{=}\textit{value}) = P(X_i \mid E{=}\textit{value})\ P(X_j \mid X_i, E{=}\textit{value})$
- **Optimal decisions:**
  - *Decision networks* include utility information
  - Probabilistic inference gives P(outcome | action, evidence)
- **Value of information:** Which evidence should we seek next?
- **Sensitivity analysis:** Which probability values are most critical?
- **Explanation:** Why do I need a new starter motor?

DIRECT INFERENCE WITH BNS

- Instead of computing the joint, suppose we just want the probability for one variable.
- Exact methods of computation:
  - **Enumeration**
  - **Variable elimination**
  - **Join trees:** get the probabilities associated with every query variable

REVIEW: INFERENCE BY ENUMERATION

1. Find the relevant datapoints consistent with the evidence

   E.g., when it was raining and I was on time

2. Sum across all the *h*'s to get the joint probability of the query and the evidence

   E.g., total of *all* the times I was on time when it was raining

3. Normalize i.e., divide each instance by the sum of them all

   E.g., divide by the total across all queries (on time, not on time) with the same evidence (raining, etc.)

$$P(Q, e_1 \dots e_k) = \sum_{h_1 \dots h_r} P(Q, h_1 \dots h_r, e_1 \dots e_k)$$

$$P(Q|e_1 \dots e_k) = \frac{P(Q, e_1 \dots e_k)}{\sum_q P(Q, e_1 \dots e_k)}$$

With:

- "evidence" variables $E_1 \dots E_k = e_1 \dots e_k$
- "query" variable Q
- "hidden" variables $H_1 \dots H_r$

We want $P(Q| e_1 \dots e_k)$

Q in this example is "will I be on time?"

INFERENCE BY ENUMERATION

- Add all of the terms (atomic event probabilities) from the full joint distribution
- If **E** are the evidence (observed) variables and **Y** are the other (unobserved) variables, then:

$$P(X \mid \mathbf{E}) = \alpha\, P(X, \mathbf{E}) = \alpha \sum P(X, \mathbf{E}, \mathbf{Y})$$

- Each $P(X, \mathbf{E}, \mathbf{Y})$ term can be computed using the chain rule
- Computationally expensive!

P(E) is known (observed), so 1/P(E) is a constant that makes everything sum to 1: the *normalizing constant*

YOUR TURN: USING BAYES' NETS

In general, there's a 4 step process to solve **any** query about a Bayes' Net:

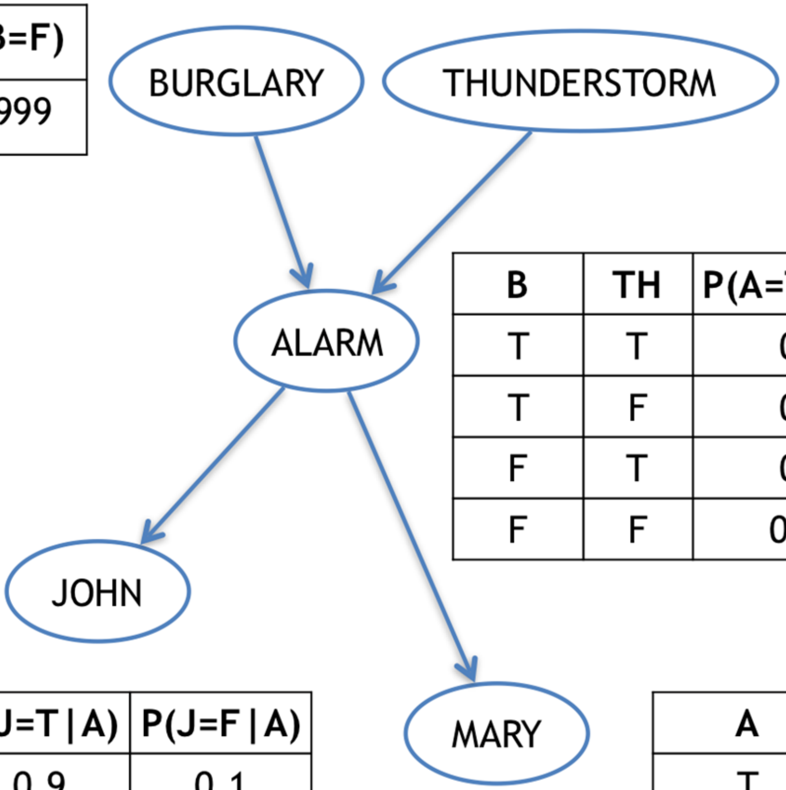
1. Write the query as a statement about probabilities
2. Rewrite statement in terms of the joint probability distribution
3. Figure out ALL the atomic probabilities you need
4. Simplify, and plug in numbers from CPTs

$$P(X \mid \mathbf{E}) = \alpha\, P(X, \mathbf{E}) = \alpha \sum P(X, \mathbf{E}, \mathbf{Y})$$

Calculate the probability that there is a burglar if both John and Mary call.

P(B=T)	P(B=F)
0.001	0.999

P(TH=T)	P(TH=F)
0.002	0.998



B	TH	P(A=T|B,TH)	P(A=F|B,TH)
T	T	0.95	0.05
T	F	0.94	0.06
F	T	0.29	0.71
F	F	0.001	0.999

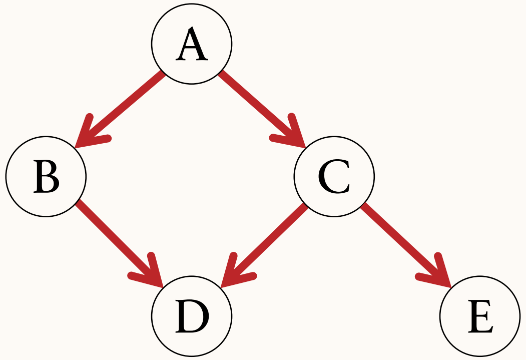
A	P(J=T|A)	P(J=F|A)
T	0.9	0.1
F	0.05	0.95

A	P(M=T|A)	P(M=F|A)
T	0.7	0.3
F	0.01	0.99

Slide from Dr. Brian Hrolenok, AIMA Book

ENUMERATION EXAMPLE

- $P(x_i) = \Sigma_{\pi\ i} P(x_i \mid \pi_i)\ P(\pi_i)$
- Say we want to know $P(\mathrm{D}=t)$
- Only E is *given* as true
- $\mathrm{P}\ (\mathrm{d} \mid \mathrm{e}) = \alpha\ \Sigma_{\mathrm{ABC}}\mathrm{P}(\mathrm{a, b, c, d, e})$
- $= \alpha\ \Sigma_{\mathrm{ABC}}\mathrm{P}(\mathrm{a})\ \mathrm{P}(\mathrm{b} \mid \mathrm{a})\ \mathrm{P}(\mathrm{c} \mid \mathrm{a})\ \mathrm{P}(\mathrm{d} \mid \mathrm{b, c})\ \mathrm{P}(\mathrm{e} \mid \mathrm{c})$
- With simple iteration, that's a lot of repetition!
  - P(e|c) has to be recomputed every time we iterate over C=true



VARIABLE ELIMINATION

- Basically just enumeration with caching of local calculations
- Linear for polytrees (singly connected BNs)
- Potentially exponential for multiply connected BNs
  - **Exact inference in Bayesian networks is NP-hard!**
- Join tree algorithms are an extension of variable elimination methods that compute posterior probabilities for all nodes in a BN simultaneously

VARIABLE ELIMINATION APPROACH

- Write query in the form

$$p(x_n) = \sum_{x_1} \dots \sum_{x_{n-1}} p(x_1) \prod_{i=2}^{n} p(x_i | x_{i-1})$$

  - Note that there is no α term here
  - It's a conjunctive probability, not a conditional probability…
- Iteratively,
  - Move all irrelevant terms outside of innermost sum
  - Perform innermost sum, getting a new term
  - Insert the new term into the product

A helpful description of variable elimination: https://ermongroup.github.io/cs228-notes/inference/ve/

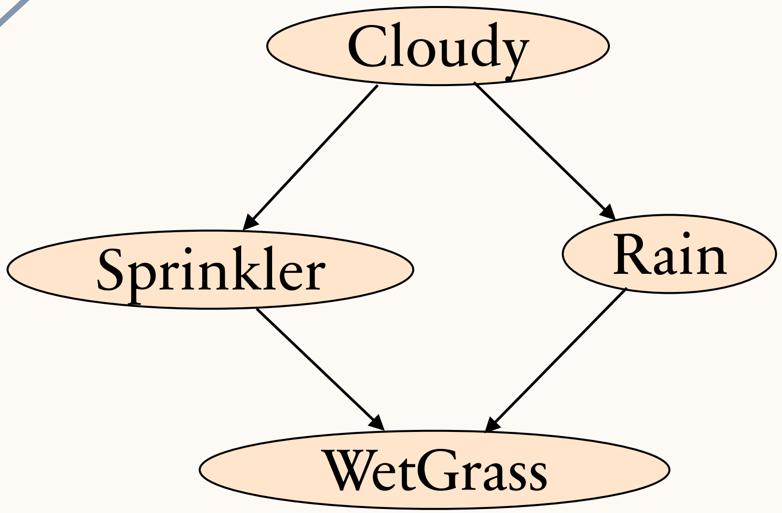
VARIABLE ELIMINATION: EXAMPLE

$$P(w) = \sum_{r,s,c} P(w|r,s)P(r|c)P(s|c)P(c)$$

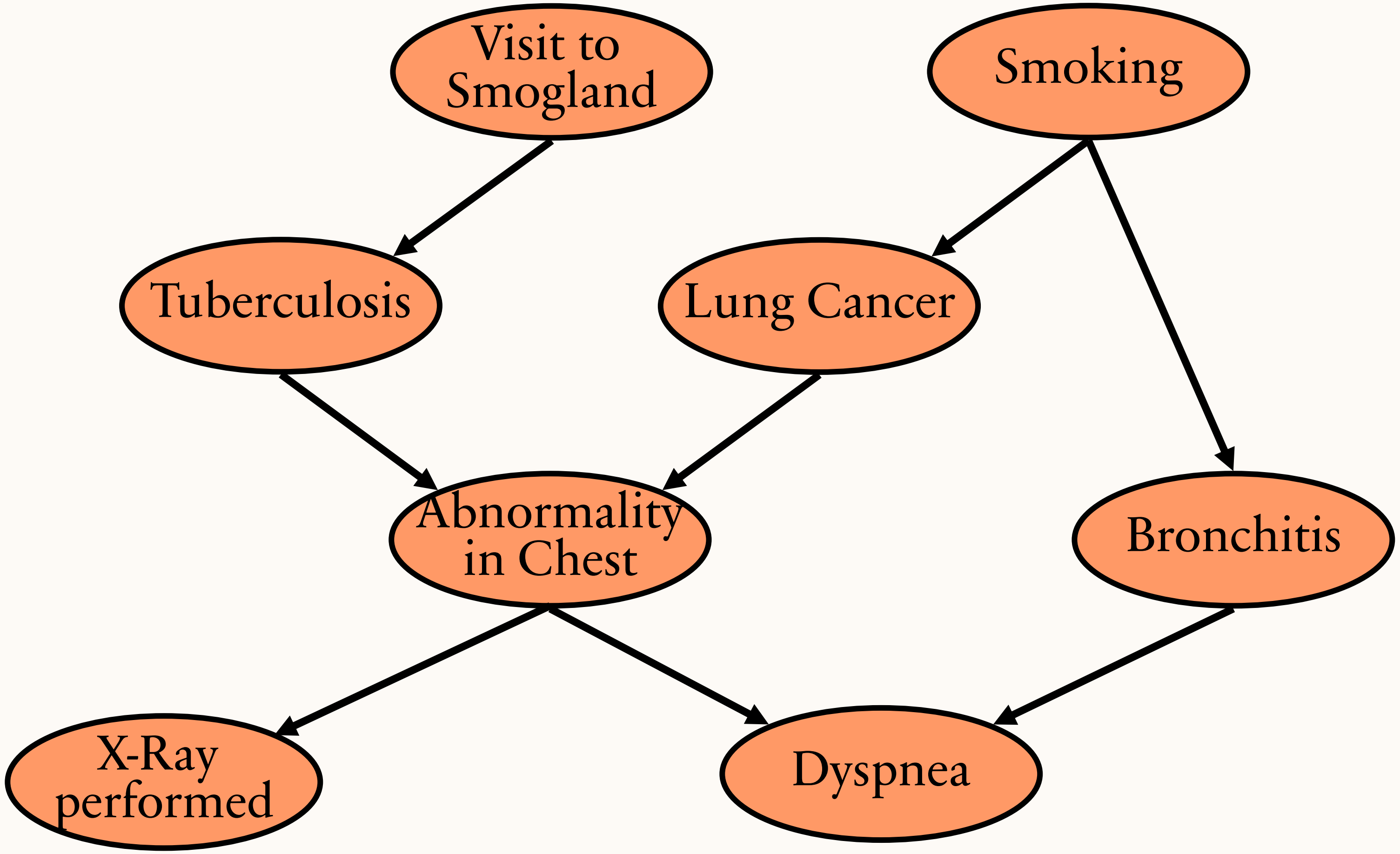
$$= \sum_{r,s} P(w|r,s)\underbrace{\sum_{c} P(r|c)P(s|c)P(c)}_{f_1(r,s)}$$

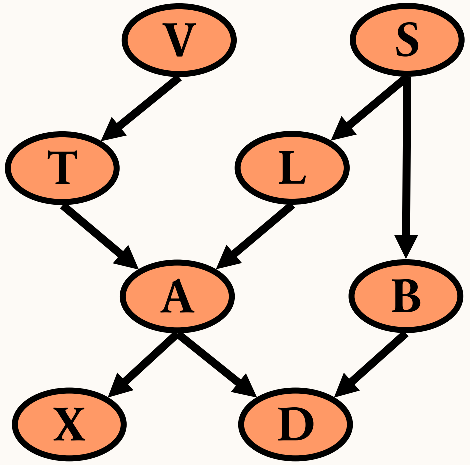
$$= \sum_{r,s} P(w|r,s)f_1(r,s)$$

"factors"



A MORE COMPLEX EXAMPLE

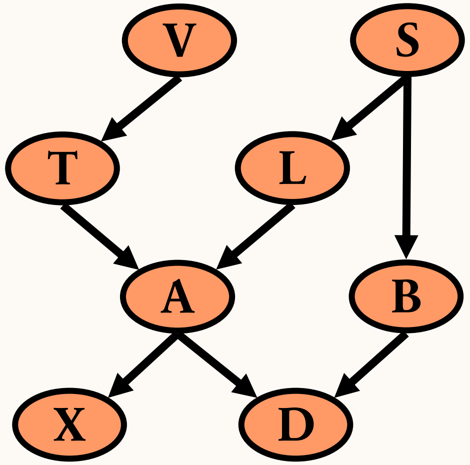




LUNGS 1

- We want to compute $P(d)$
- Need to eliminate: v,s,x,t,l,a,b
- Initial factors:

$$P(v)P(s)P(t\,|\,v)P(l\,|\,s)P(b\,|\,s)P(a\,|\,t\overset{\swarrow}{,}l)P(x\,|\,a)P(d\,|\,a\overset{\swarrow}{,}b)$$



LUNGS 2

- We want to compute $P(d)$
- Need to eliminate: v,s,x,t,l,a,b
- Initial factors:

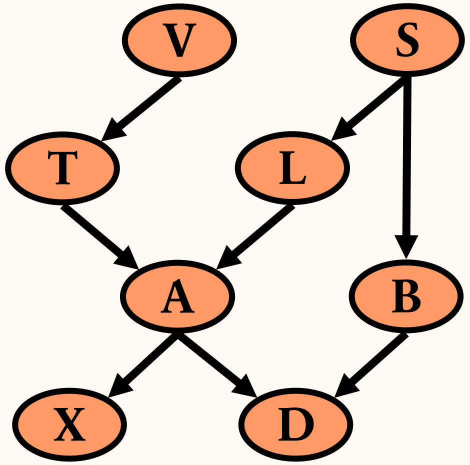
$$\underline{P(v)}P(s)\underline{P(t\,|\,v)}P(l\,|\,s)P(b\,|\,s)P(a\,|\,t,l)P(x\,|\,a)P(d\,|\,a,b)$$

- Eliminate: v
- Compute: $f_v(t) = \sum_v P(v)P(t\,|\,v)$

$$\Rightarrow\ \underline{f_v(t)}P(s)P(l\,|\,s)P(b\,|\,s)P(a\,|\,t,l)P(x\,|\,a)P(d\,|\,a,b)$$

- Note: $f_v(t) = P(t)$
- Result of elimination is not **necessarily** a probability term
  - For example, $f_v(t)$ might capture P(t) and P(¬t)

LUNGS 3



- We want to compute *P(d)*
- Need to eliminate: *s,x,t,l,a,b*
- Initial factors: $P(v)P(s)P(t\,|\,v)P(l\,|\,s)P(b\,|\,s)P(a\,|\,t,l)P(x\,|\,a)P(d\,|\,a,b)$

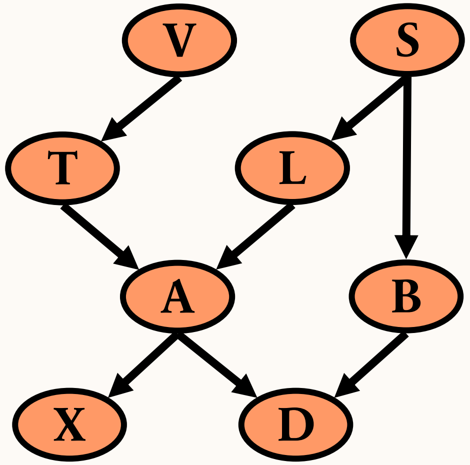
$$\text{Þ}\;\; f_v(t)\underline{P(s)}\,\underline{P(l\,|\,s)}\,\underline{P(b\,|\,s)}P(a\,|\,t,l)P(x\,|\,a)P(d\,|\,a,b)$$

- Eliminate: *s*
- Compute: $f_s(b,l) = \text{å}\; P(s)P(b\,|\,s)P(l\,|\,s)$

$$\text{Þ}\;\; f_v(t)\underline{f_s^{\,s}(b,l)}P(a\,|\,t,l)P(x\,|\,a)P(d\,|\,a,b)$$

- Summing on *s* results in a factor with two arguments *f*$f_s(b,l)$
- In general, result of elimination may be a function of several variables

LUNGS 4



- We want to compute *P(d)*
- Need to eliminate: *x,t,l,a,b*
- Initial factors

$$P(v)P(s)P(t \mid v)P(l \mid s)P(b \mid s)P(a \mid t,l)P(x \mid a)P(d \mid a,b)$$

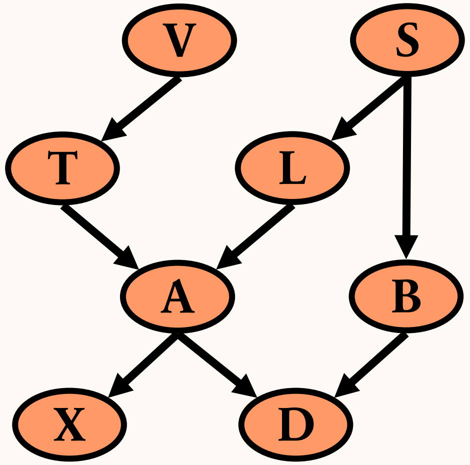
Þ $f_v(t)P(s)P(l \mid s)P(b \mid s)P(a \mid t,l)P(x \mid a)P(d \mid a,b)$

Eliminate: *x* Þ $f_v(t)f_s(b,l)P(a \mid t,l)P(x \mid a)P\underline{(d \mid a,b})$

Compute: $f_x(a) = \underset{x}{\text{å}}\, P(x \mid a)$

Þ $f_v(t)f_s(b,l)\underline{f_x(a)}P(a \mid t,l)P(d \mid a,b)$

LUNGS 5



- We want to compute *P(d)*
- Need to eliminate: *t,l,a,b*
- Initial factors $P(v)P(s)P(t\,|\,v)P(l\,|\,s)P(b\,|\,s)P(a\,|\,t,l)P(x\,|\,a)P(d\,|\,a,b)$

$$\text{Þ}\ \ f_v(t)P(s)P(l\,|\,s)P(b\,|\,s)P(a\,|\,t,l)P(x\,|\,a)P(d\,|\,a,b)$$

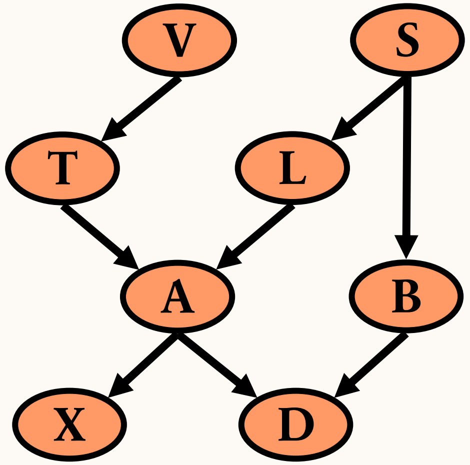
$$\text{Þ}\ \ f_v(t)f_s(b,l)P(a\,|\,t,l)P(x\,|\,a)P(d\,|\,a,b)$$

$$\text{Þ}\ \ \underline{f_v(t)}f_s(b,l)f_x(a)\underline{P(a\,|\,t,l)}P(d\,|\,a,b)$$

Eliminate: *t*

Compute: $f_t(a,l) = \underset{t}{\text{å}}\, f_v(t)P(a\,|\,t,l)$

$$\text{Þ}\ \ f_s(b,l)f_x(a)\underline{f_t(a,l)}P(d\,|\,a,b)$$



LUNGS 6

- We want to compute *P(d)*
- Need to eliminate: *l,a,b*
- Initial factors

$$P(v)P(s)P(t\,|\,v)P(l\,|\,s)P(b\,|\,s)P(a\,|\,t,l)P(x\,|\,a)P(d\,|\,a,b)$$

$$\text{Þ}\;\; f_v(t)P(s)P(l\,|\,s)P(b\,|\,s)P(a\,|\,t,l)P(x\,|\,a)P(d\,|\,a,b)$$

$$\text{Þ}\;\; f_v(t)f_s(b,l)P(a\,|\,t,l)P(x\,|\,a)P(d\,|\,a,b)$$

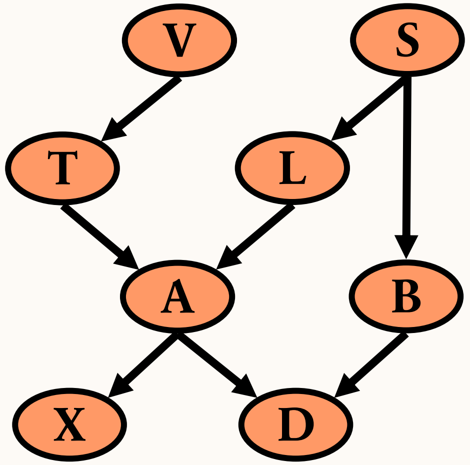
$$\text{Þ}\;\; f_v(t)f_s(b,l)f_x(a)P(a\,|\,t,l)P(d\,|\,a,b)$$

$$\text{Þ}\;\; \underline{f_s(b,l)}f_x(a)\underline{f_t(a,l)}P(d\,|\,a,b)$$

Eliminate: *l*

Compute: $f_l(a,b) = \text{å}_{l}\, f_s(b,l)f_t(a,l)$

$$\text{Þ}\;\; \underline{f_l(a,b)}f_x(a)P(d\,|\,a,b)$$



LUNGS FINALE

- We want to compute $P(d)$
- Need to eliminate: b
- Initial factors

$$P(v)P(s)P(t|v)P(l|s)P(b|s)P(a|t,l)P(x|a)P(d|a,b)$$
$$\Rightarrow f_v(t)P(s)P(l|s)P(b|s)P(a|t,l)P(x|a)P(d|a,b)$$
$$\Rightarrow f_v(t)f_s(b,l)P(a|t,l)P(x|a)P(d|a,b)$$
$$\Rightarrow f_v(t)f_s(b,l)f_x(a)P(a|t,l)P(d|a,b)$$
$$\Rightarrow f_s(b,l)f_x(a)f_t(a,l)P(d|a,b)$$
$$\Rightarrow \underline{f_l(a,b)}\,\underline{f_x(a)}\,\underline{P(d|a,b)} \Rightarrow \underline{f_a(b,d)} \Rightarrow \underline{f_b(d)}$$

Eliminate: a,b

Compute: $f_a(b,d)=\sum_a f_l(a,b)f_x(a)p(d|a,b) \quad f_b(d)=\sum_b f_a(b,d)$

VARIABLE ELIMINATION ALGORITHM

- Let $X_1,\ldots,X_m$ be an ordering on the non-query variables
- For i = m, …, 1 $\sum_{X_1}\sum_{X_2}\ldots\sum_{X_m}\prod_{j} P(X_j \mid Parents(X_j))$
  - In the summation for X_i, leave only factors mentioning X_i
  - Multiply the factors, getting a factor that contains a number for each value of the variables mentioned, including X_i
  - Sum out X_i, getting a factor f that contains a number for each value of the variables mentioned, not including X_i
  - Replace the multiplied factor in the summation