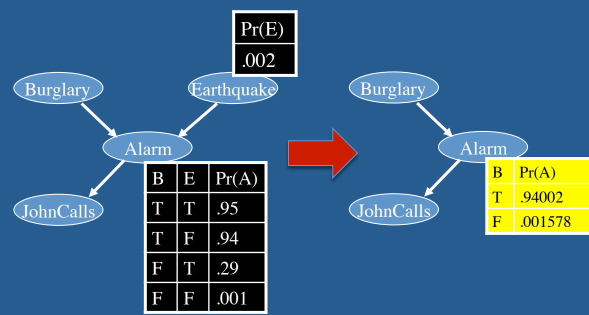
1



Pr(E)
.002

B	E	Pr(A)
T	T	.95
T	F	.94
F	T	.29
F	F	.001

B	Pr(A)
T	.94002
F	.001578

BAYESIAN INFERENCE

EECS 492
March 15, 2011

Bayesian Inference Methods

2

- Given some evidence, what is the probability of something happening?
  - Probability of a burglary given Mary calls.
  - Probability of an earthquake given there was no burglary.
  - What's the (marginal) probability of John calling?
- In General:
  - P(x | e)
- Many different algorithms…

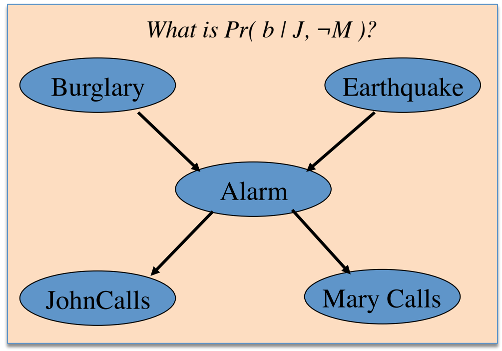
Method 1: Enumeration

3

- Express desired conditional in terms of marginals

$$P(b|J,\neg M) = \frac{P(b,J,\neg M)}{P(J,\neg M)}$$

- Obtain marginals from joint distribution
  - Encoded in Bayes net!



$$P(B,E,A,J,M) = P(B)P(E)P(A|BE)P(J|A)P(M|A)$$

$$P(b|J,\neg M) = \frac{\sum_{E,A} P(b)P(E)P(A|bE)P(J|A)P(\neg M|A)}{\sum_{E,A,B} P(b)P(E)P(A|bE)P(J|A)P(\neg M|A)}$$

Method 1: Enumeration

4

$$P(b|J,\neg M) = \frac{\sum_{E,A} P(b)P(E)P(A|bE)P(J|A)P(\neg M|A)}{\sum_{E,A,B} P(b)P(E)P(A|bE)P(J|A)P(\neg M|A)}$$

- Given N nodes, how expensive is this?
  - Space?
    - O(N)
  - Complexity?
    - $O(N2^N)$
    - Possibly as many as 2^N terms, each involving the product of N conditional probabilities.

Improving Enumeration

5

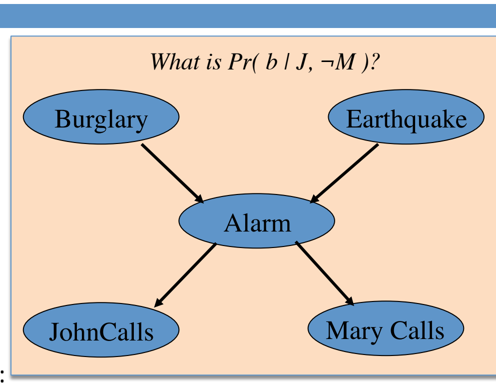
- Moving summations inwards reduces complexity:

$$\Pr(J \mid b) = \alpha \sum_{e,a,m} \Pr(b)\Pr(e)\Pr(a \mid b,e)\Pr(J \mid a)\Pr(m \mid a)$$

$$= \alpha \Pr(b) \sum_{e} \Pr(e) \sum_{a} \Pr(a \mid b,e)\Pr(J \mid a) \sum_{m} \Pr(m \mid a)$$

"Barren" Nodes

- Observation:
  - Only ancestors of X or E are relevant to query.
- Example:
  - Question: Pr(J|b)
  - Answer: M not relevant.
- Can also be seen from the joint:



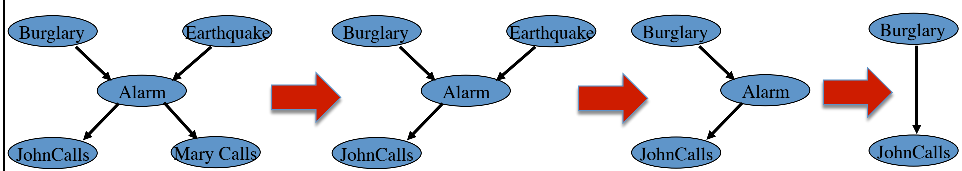
$$\Pr(J \mid b) = \alpha \sum_{e,a,m} \Pr(b)\Pr(e)\Pr(a \mid b,e)\Pr(J \mid a)\Pr(m \mid a)$$

$$= \alpha \Pr(b) \sum_{e} \Pr(e) \sum_{a} \Pr(a \mid b,e)\Pr(J \mid a) \boxed{\sum_{m} \Pr(m \mid a)}$$

Method 2: Variable Elimination

7

- Systematically remove all nodes in the graph that aren't part of our desired probability.
  - What is Pr(J | B)? Idea:

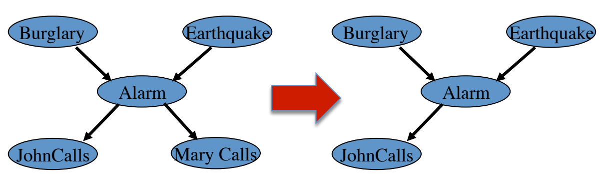


  - What's the term for summing over a variable in order to make its value irrelevant?
    - Marginalization

Variable Elimination: M

8

What is Pr(J | B)?



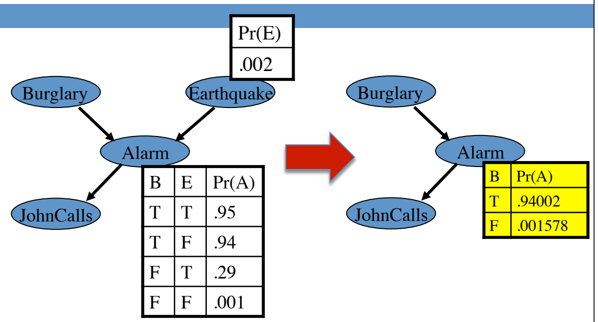
- Step 1: Eliminate M

- M is irrelevant (as we described before). We can just delete it.

Variable Elimination: E

9

What is Pr(J | B)?



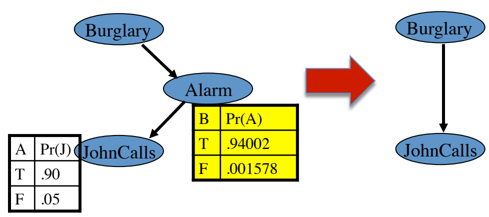
$$P(A|B) = \sum_E P(A,e|B) = \sum_E P(A|B,e)P(e)$$

$$P(A|B) = 0.95 \times 0.002 + 0.94 \times (1 - 0.002) = 0.94002$$
$$P(A|\neg B) = 0.29 \times 0.002 + 0.001 \times (1 - 0.002) = 0.001578$$

Your turn: Variable Elimination (A)

10

What is Pr(J | B)?



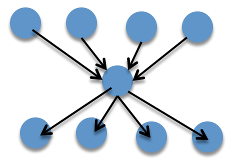
$$\begin{aligned} P(J|B) &= \sum_A P(J,a|B) \\ &= \sum_A P(J|a,B)P(a|B) \\ &= \sum_A (J|a)P(a|B) \end{aligned}$$

$$\begin{aligned} P(J|B) &= 0.9 \times 0.94002 + 0.05 \times (1 - 0.94002) = 0.849017 \\ P(J|\neg B) &= 0.9 \times 0.001578 + 0.05 \times (1 - 0.001578) = 0.051341 \end{aligned}$$

Variable Elimination

11

- Does the order in which we eliminate variables matter?



- Complexity
  - Space?
    - $O(2^N)$: we might have to store an enormous CPT if everything becomes dependent.
  - Complexity?
    - $O(2^N)$: Could have to build a mega CPT with 2^N entries.

Exploiting Problem Structure

12

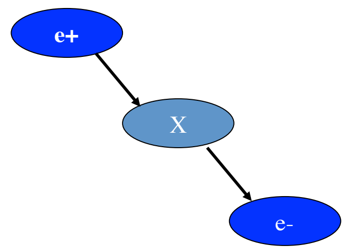
- Our next method exploits the structure of some Bayes nets….
  - Very similar to the way we exploited tree structures in CSP!

Simple Case

13

We want:

$$P(x|e^+, e^-)$$



$$P(x|e^+, e^-) = \frac{P(e^+, e^-|x)P(x)}{P(e^+, e^-)}$$

$$P(x|e^+, e^-) = \frac{P(e^+|x)P(e^-|x)P(x)}{P(e^+, e^-)}$$

$$P(x|e^+, e^-) = \frac{P(x|e^+)P(e^+)}{P(x)} \frac{P(e^+|x)P(e^-|x)P(x)}{P(e^+, e^-)}$$

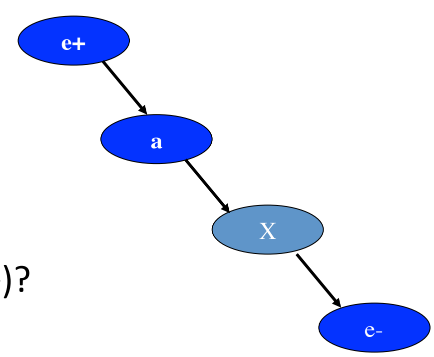
$$P(x|e^+, e^-) \propto P(x|e^+)P(e^-|x)$$

Interpretation: message passing

What if it's a bit more complex?

14

We (still) want:



$$P(x|e^+, e^-)$$

$$P(x|e^+, e^-) \propto P(x|e^+)P(e^-|x)$$

How do we compute P(x|e+)?

$$P(x|e^+) = \sum_A P(x, a|e^+)$$

$$P(x|e^+) = \sum_A P(a|e^+)P(x|a, e^+)$$

$$P(x|e^+) = \sum_A P(a|e^+)P(x|a)$$

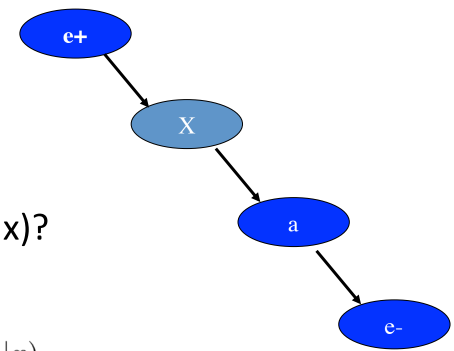
Another example

15

We (still) want:

$$P(x|e^+, e^-)$$

$$P(x|e^+, e^-) \propto P(x|e^+)P(e^-|x)$$



How do we compute P(e-|x)?

$$P(e^-|x) = \sum_A P(e^-, a|x)$$

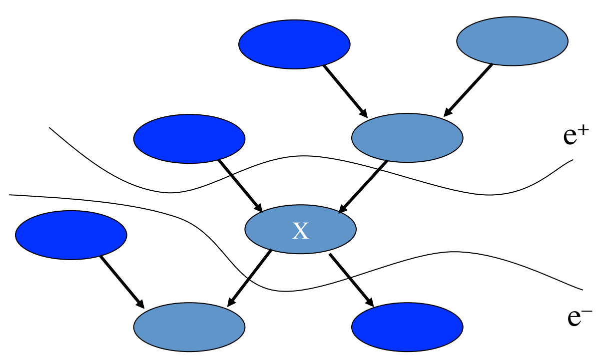
$$P(e^-|x) = \sum_A P(e^-|a, x)P(a|x)$$

$$P(e^-|x) = \sum_A P(e^-|a)P(a|x)$$

General Case

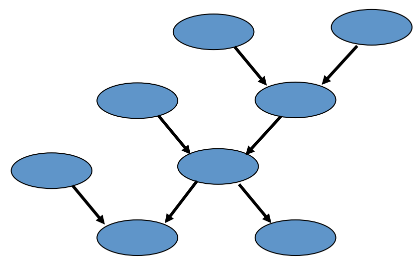
- Can partition evidence into causal and evidential support
- Pr(X | e) = Pr(X | e^+,e^-)

Local message-passing algorithm implements recursive computation of evidence contribution in linear time



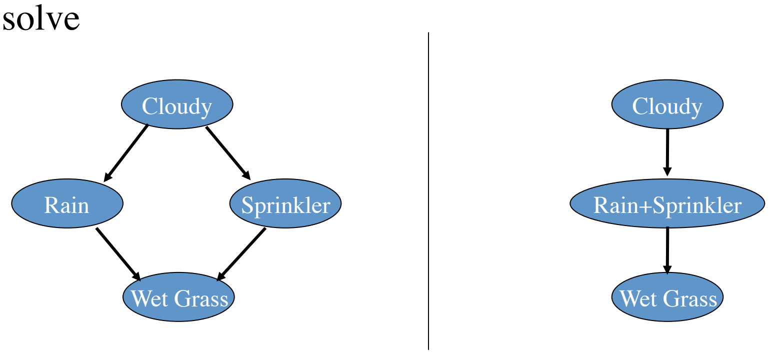
General Case: Pearl's Algorithm

- At most one *undirected* path between any pair of nodes
  - Why are loops bad?
- Can pass messages for inference O(N) time
- Turbo decoding & UTSM



Clustering

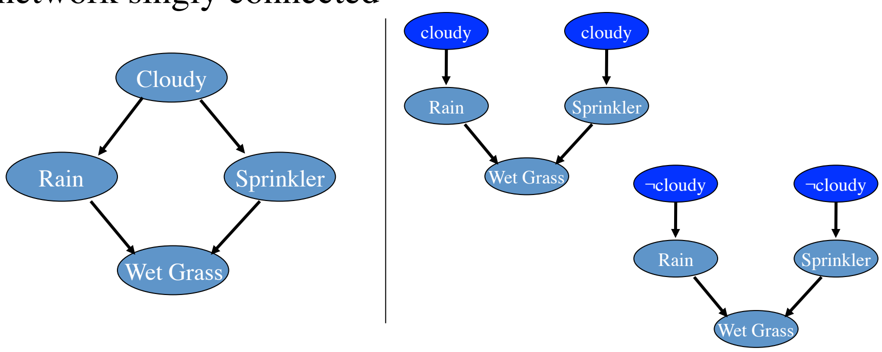
Convert multiply connected network to polytree, then solve



- May entail exponential blowup

Cutset Conditioning

Identify set of variables (cutset) that would render network singly connected



- May entail exponential time for conditioning

Complexity of BN Algorithms

Method	Applicability	Space	Time
Enumeration	general	O(n)	O($n2^n$)
Variable elimination	general	O(2^n)	O(2^n)
Local propagation	polytrees	O(n)	O(n)
Clustering	general	O(2^n)	O(2^n)
Conditioning	general	O(n)	O(2^n)

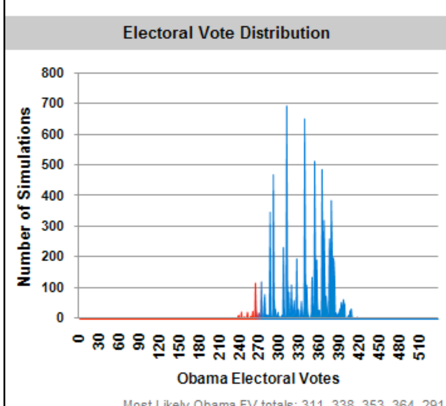
Approximate Inference

21

- So far, we've dealt with *exact* inference methods.
  - Don't always need *exact*!
- Approximate inference methods can quickly yield useful and interesting results!

Approximate Inference: Sampling

- Generate scenarios according to joint distribution
- Answer queries according to frequency in sample



from FiveThirtyEight.com, 3 Nov 08

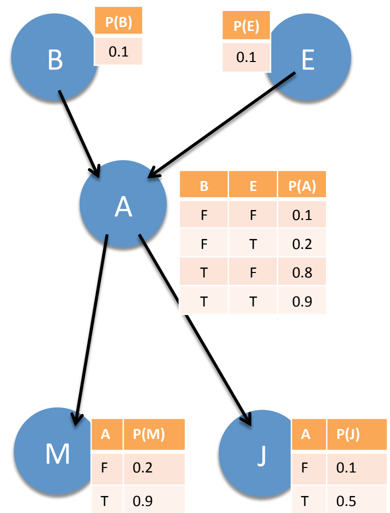
Scenario Analysis

Electoral College **TIE**	0.21%	(21 of 10000)
Recount (one or more decisive states <=0.5%)	2.46%	(246 of 10000)
Obama wins Popular Vote	96.40%	(9640 of 10000)
McCain wins Popular Vote	3.60%	(360 of 10000)
Obama loses Popular Vote, wins Electoral Vote	1.02%	(102 of 10000)
McCain loses Popular Vote, wins Electoral Vote	1.11%	(111 of 10000)
Obama loses Popular Vote >=3%, wins election	0.00%	(0 of 10000)
McCain loses Popular Vote >=3%, wins election	0.00%	(0 of 10000)
Obama landslide (375+ EV)	23.68%	(2368 of 10000)
McCain landslide (375+ EV)	0.00%	(0 of 10000)

Direct Sampling

23

- Work from top to bottom
  - Sample according to distribution, given previously sampled parents.
- Suppose I have 5 random numbers [0,9]: 7,0,1,4,9
  - Sample **b**: 7 < 1 ? False
  - Sample **e**: 0 < 1 ? True
  - Sample **a | -B, E:** 1 < 2 ? True
  - Sample **m | A**: 4 < 9 ? True
  - Sample **j | A**: 9 < 5 ? False
- Thus, my sample is:
  - -B, E, A, M, -J
- Repeat this process many times
  - P(M): How often does M occur?



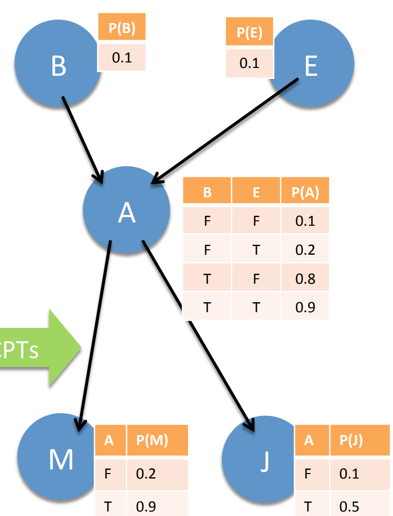
Direct Sampling: Your turn

24

Last 5 digits of your phone number

- Sample **b**: ___ < ___ ?
- Sample **e**: ___ < ___ ?
- Sample **a | b, e**: ___ < ___ ?
- Sample **m | a**: ___ < ___ ?
- Sample **j | a**: ___ < ___ ?

From CPTs



- Write down your sample
  - Give thumbs up when you've got it!

Direct Sampling

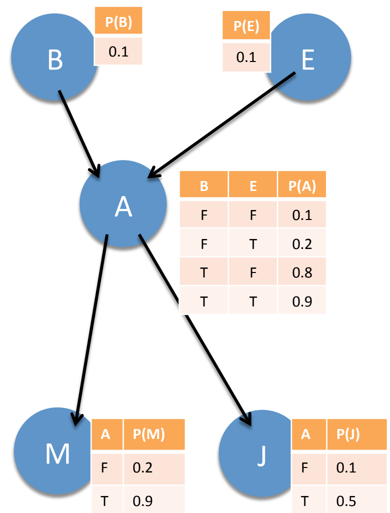
25

- How often does M occur?

$$P(M) \approx \frac{N_M}{N_{total}}$$

- How do we incorporate evidence?

$$P(M|\neg E)$$



Rejection Sampling

26

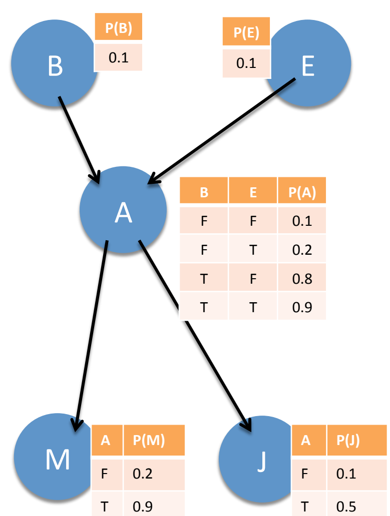
- How do we answer:

$$P(M|\neg E)$$

- Idea: discard samples where E=true, then compute statistics.
  - How many samples will be rejected?

$$P(M|\neg E) = \frac{P(M, \neg E)}{P(\neg E)}$$

- How many of you have –E?
- How many of you have M^-E?



Rejection Sampling (2)

27

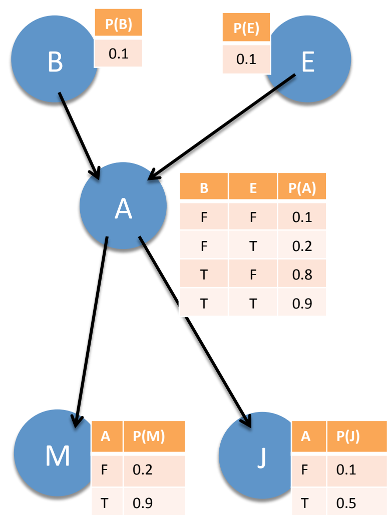
- Okay, how about:

$$P(J|B, M)$$

- How do we compute this?

$$P(J|B, M) = \frac{P(J, B, M)}{P(B, M)}$$

- How many of you have B^M?
- How many of you have JBM?
- What happened?



Rejection Sampling

Suppose we want to estimate Pr(X|e)?

Scenario Analysis		
Electoral College **TIE**	0.21%	(21 of 10000)
Recount (one or more decisive states <=0.5%)	2.46%	(246 of 10000)
Obama wins Popular Vote	96.40%	(9640 of 10000)
McCain wins Popular Vote	3.60%	(360 of 10000)
Obama loses Popular Vote, wins Electoral Vote	1.02%	(102 of 10000)
McCain loses Popular Vote, wins Electoral Vote	1.11%	(111 of 10000)
Obama loses Popular Vote >=3%, wins election	0.00%	(0 of 10000)
McCain loses Popular Vote >=3%, wins election	0.00%	(0 of 10000)
Obama landslide (375+ EV)	23.68%	(2368 of 10000)
McCain landslide (375+ EV)	0.00%	(0 of 10000)
Obama loses **OH**, wins election	81.91%	(1639 of 2001)
McCain loses **OH**, wins election	0.01%	(1 of 7999)
Obama loses **OH/FL**, wins election	79.15%	(1374 of 1736)
McCain loses **OH/FL**, wins election	0.00%	(0 of 7994)
Obama loses **OH/FL/PA**, wins election	7.94%	(15 of 189)
McCain loses **OH/FL/PA**, wins election	0.00%	(0 of 6183)
Obama wins all Kerry states	97.38%	(9738 of 10000)
McCain wins all Bush states	0.01%	(1 of 10000)
Obama wins **VA** when losing **OH**	71.26%	(1426 of 2001)
Obama wins **FL** when losing **OH**	13.24%	(265 of 2001)
Obama wins **CO** when losing **OH**	81.26%	(1626 of 2001)
Obama wins **OH** when losing **PA**	2.56%	(5 of 195)

Rejection Sampling: Summary

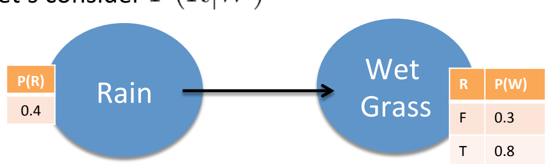
29

- Rejection sampling is an easy way to do inference, however:
  - As conditional becomes more rare, accuracy rapidly falls.
- Is there a better way?
  - Yes! Likelihood weighting!

Likelihood Weighting

30

- Idea: ensure that evidence values are satisfied during the sampling process.
  - If evidence node has no parents, just set the value.
  - What do we do if an evidence node *has* parents?
    - Let's consider $P(R|W)$



- To draw samples:
  - Sample from Rain as usual (suppose we sample 'true').
  - Now, must force W=true. How likely was this outcome?
    - Count this sample as 0.8 of a sample.

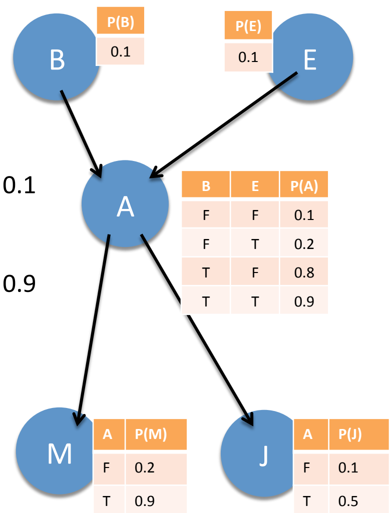
Likelihood Weighting

31

- Our pesky example again:

$$P(J|B,M)$$

- I need 3 random numbers [0,9]: 2,4,9
  - Initialize weight .... w = 1
  - Sample **b**: ***given*** True w *= 0.1
  - Sample **e**: 2 <1 ? False
  - Sample **a | B, -E:** 4 < 8 ? True
  - Sample **m | A**: ***given*** True w *= 0.9
  - Sample **j | A**: 9 < 5 ? False
- Our sample is: [B,-E,A,M,-J]
  - Satisfies conditionals by construction
  - Likelihood = 0.09



Likelihood Sampling: Your turn

32

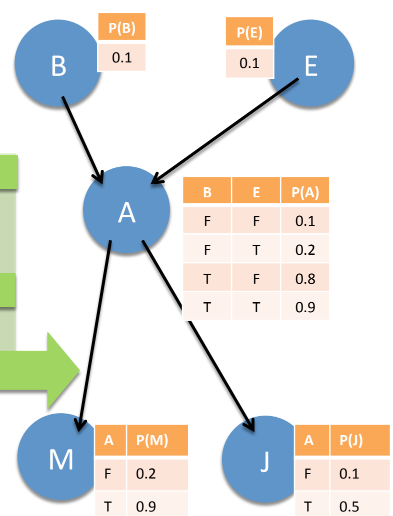
$$P(J|B,M)$$

Last 3 digits of your phone number

- Initialize w = 1
- Sample **b**: **TRUE** w*=
- Sample **e**: __ < __ ?
- Sample **a | b, e:** __ < __ ?
- Sample **m | a**: **TRUE** w*=
- Sample **j | a**: __ < __ ?

From CPTs

- What outcome did you sample?
  - P(J|B,M) or P(-J|B,M)?
  - What's your likelihood?



Likelihood Weighting: Your turn

33

$$P(J|B,M)$$

- Did we get a better estimate?

Probability	Sum of likelihoods
$P(J\|B,M)$	
$P(\neg J\|B,M)$	

How many samples do I need?

34

- Or equivalently, what's the variance associated with my estimate?
- Let's consider rejection sampling (simpler math)
  - N *accepted* samples
  - Each has some unknown variance
    - If p is actual answer, variance of samples is p(1-p)
    - Maximum possible value when p=0.5: σ^2=0.25
  - We're summing N of them
    - Variance of sum: $N\sigma^2$
  - Then we divide by N (which scales variance as $1/N^2$)
    - Variance of estimated probability goes as 1/(4N)
    - Which means that standard deviation goes as $\frac{1}{2\sqrt{N}}$

Sampling Review

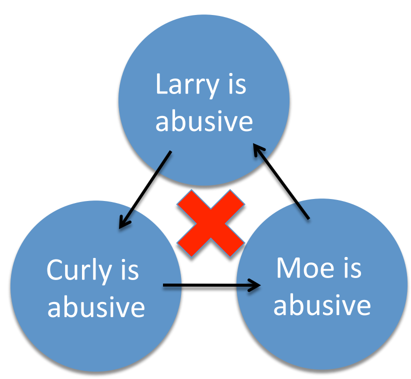
35

- Time Complexity?
  - O(S): Linear in the number of samples
  - Need more samples when evidence is rare!
    - Likelihood sampling helps, but doesn't solve, the problem.
- Space Complexity?
  - O(1)
- Simple and easy-to-implement methods
  - If you don't need exact answers, a very good thing to try!
- Also read in the book about Markov Chain Monte Carlo approach!

Limitations (?) of Bayes Nets

36

- Directional edges
  - Joint distribution = product of conditional distributions
- Cycles not permitted
  - Not required for *expressivity*
  - But sometimes it'd be more natural…
- Computing dependencies is a bit tricky

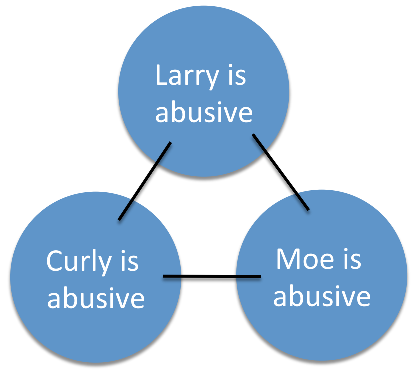


Cycles are not permitted in Bayes nets

Markov Random Fields

37

- A different approach
  - Makes some problems easier to specify
- Undirected edges
  - Joint distribution = product of *potential* functions
    - Potential functions: functions of one or more variables.
    - We'll get more specific later.
  - Cycles permitted
- Dependencies are easy…



Cycles are permitted in Markov Random Fields