Particle-Based Approximate Inference using Random Sampling

Presented by Hua Ai

09/26/2005

Modified by milos 10/15/05

Particles

- Particles: a set of instantiations of joint distribution to all or some of the variables in the network

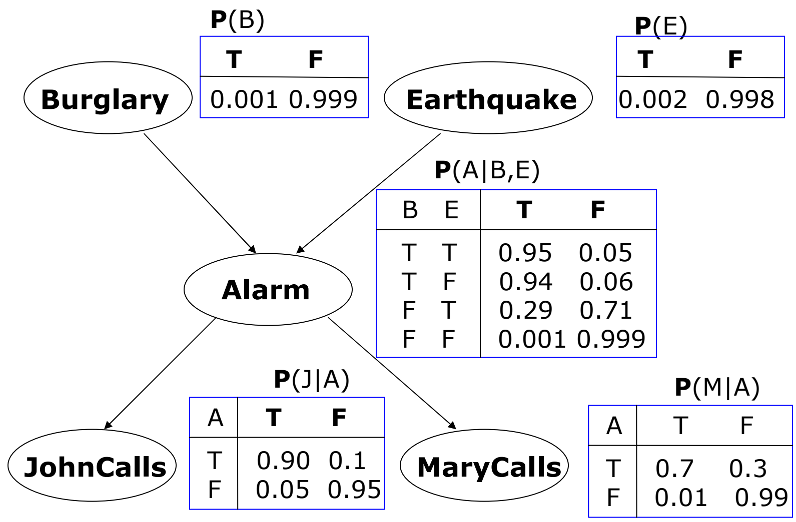
Outline

- Forward Sampling
- Rejection Sampling
- Likelihood Weighting Sampling
- Importance Sampling

Forward Sampling

- Sample the nodes in some order consistent with the partial order of the BN, so that by the time we sample a node, we have values for all its parents.

BBN sampling example



P(B)

T	F
0.001	0.999

P(E)

T	F
0.002	0.998

P(A|B,E)

B	E	T	F
T	T	0.95	0.05
T	F	0.94	0.06
F	T	0.29	0.71
F	F	0.001	0.999

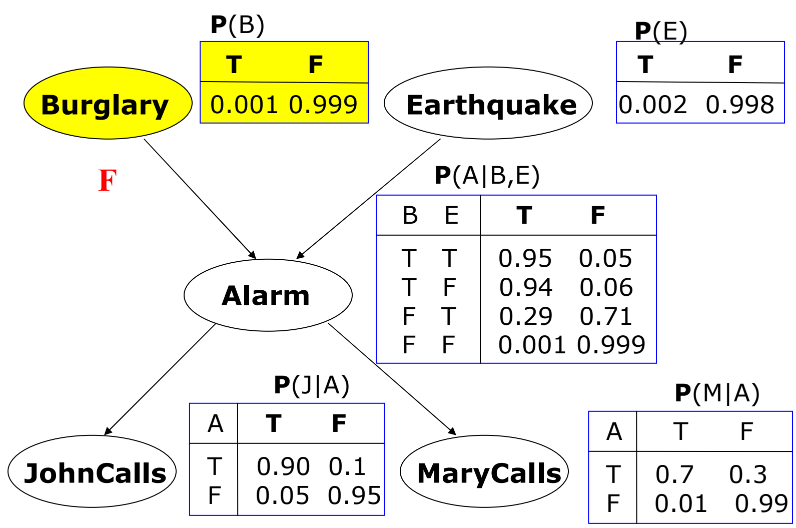
P(J|A)

A	T	F
T	0.90	0.1
F	0.05	0.95

P(M|A)

A	T	F
T	0.7	0.3
F	0.01	0.99

BBN sampling example



P(B)

T	F
0.001	0.999

P(E)

T	F
0.002	0.998

P(A|B,E)

B	E	T	F
T	T	0.95	0.05
T	F	0.94	0.06
F	T	0.29	0.71
F	F	0.001	0.999

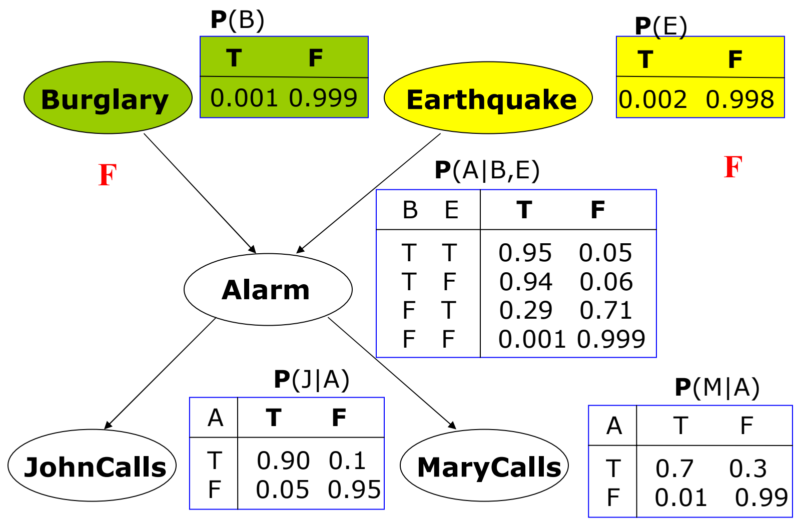
P(J|A)

A	T	F
T	0.90	0.1
F	0.05	0.95

P(M|A)

A	T	F
T	0.7	0.3
F	0.01	0.99

BBN sampling example



P(B)

T	**F**
0.001	0.999

P(E)

T	**F**
0.002	0.998

P(A|B,E)

B	E	**T**	**F**
T	T	0.95	0.05
T	F	0.94	0.06
F	T	0.29	0.71
F	F	0.001	0.999

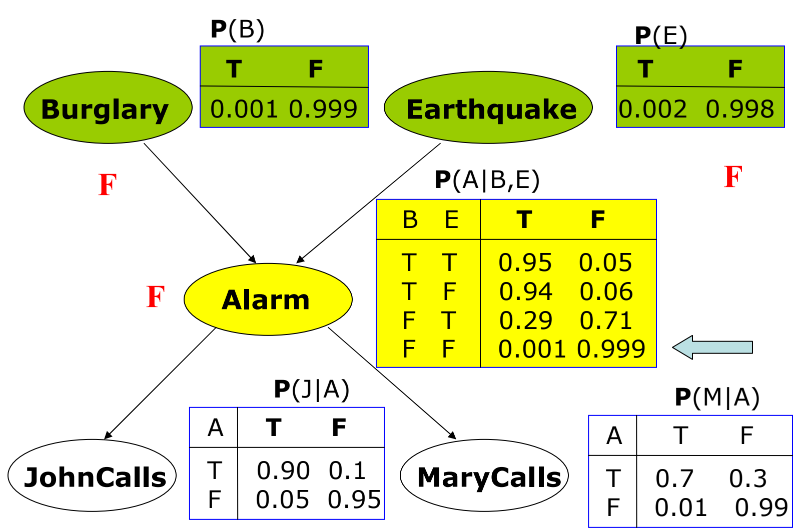
P(J|A)

A	**T**	**F**
T	0.90	0.1
F	0.05	0.95

P(M|A)

A	T	F
T	0.7	0.3
F	0.01	0.99

BBN sampling example



P(B)

T	**F**
0.001	0.999

P(E)

T	**F**
0.002	0.998

P(A|B,E)

B	E	**T**	**F**
T	T	0.95	0.05
T	F	0.94	0.06
F	T	0.29	0.71
F	F	0.001	0.999

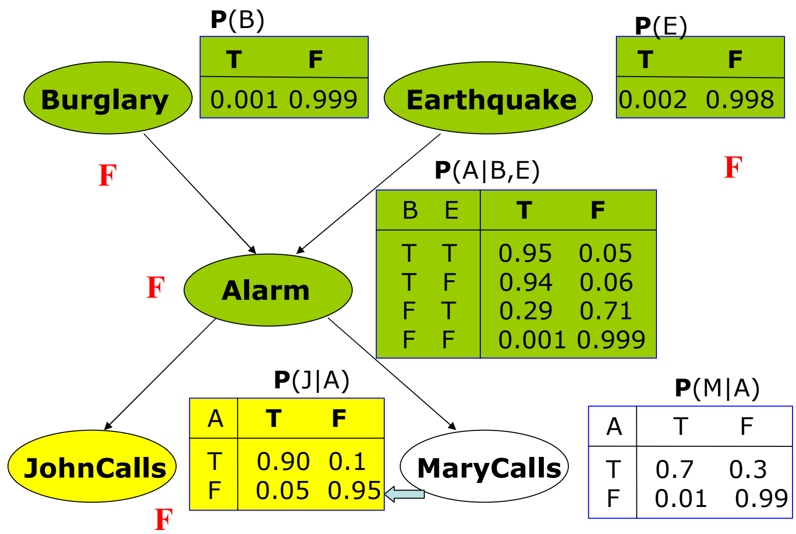
P(J|A)

A	**T**	**F**
T	0.90	0.1
F	0.05	0.95

P(M|A)

A	T	F
T	0.7	0.3
F	0.01	0.99

BBN sampling example



P(B)

T	F
0.001	0.999

P(E)

T	F
0.002	0.998

P(A|B,E)

B	E	T	F
T	T	0.95	0.05
T	F	0.94	0.06
F	T	0.29	0.71
F	F	0.001	0.999

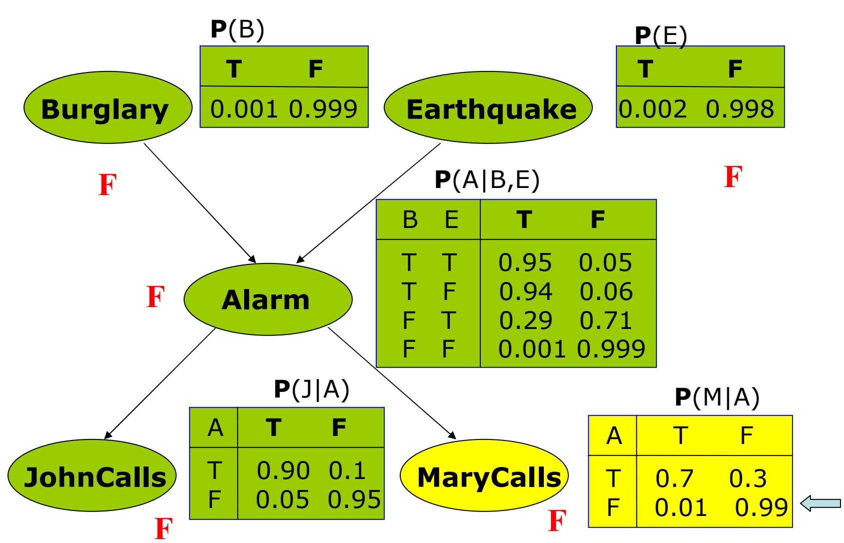
P(J|A)

A	T	F
T	0.90	0.1
F	0.05	0.95

P(M|A)

A	T	F
T	0.7	0.3
F	0.01	0.99

BBN sampling example



P(B)

T	F
0.001	0.999

P(E)

T	F
0.002	0.998

P(A|B,E)

B	E	T	F
T	T	0.95	0.05
T	F	0.94	0.06
F	T	0.29	0.71
F	F	0.001	0.999

P(J|A)

A	T	F
T	0.90	0.1
F	0.05	0.95

P(M|A)

A	T	F
T	0.7	0.3
F	0.01	0.99

Forward Sampling in a Bayesian network

Procedure Forward-Sample (B)

```
Let X_1,...,X_n be a topological ordering of X
For i=1,…,n
u_i ← x<Pa_{X_i}>    //Assignment to Pa_{X_i} in x_1,...,x_{i-1}
    Sample x_i from P(X_i | u_i)
return ( x_1,...,x_{i-1} )
```


Absolute Error Bound

- Apply **Hoeffding's bound** to estimate how many samples are required to achieve an estimate whose error is bounded by ε, with probability at least 1- δ

$$P_D(\hat{P}_D(y) \notin [P(y)-\varepsilon, P(y)+\varepsilon]) \leq 2e^{-2M\varepsilon^2} \leq \delta$$

Gives sample complexity:

$$M \geq \frac{\ln(2/\delta)}{2\varepsilon^2}$$

Relative Error Bound

- By applying **Chernoff's bound** to conclude that Is also within a relative error ε of the true value $P(y)$ $\hat{P}_D(y)$, within high probability. Specifically, we have that:

$$P_D(\hat{P}_D(y) \notin P(y)(1+\in)) \leq 2e^{-MP(y)\varepsilon^2/3}$$

So that:

$$M \geq 3\frac{\ln(2/\delta)}{P(y)\varepsilon^2}$$

Rejection Sampling

To generate samples from P(x|e), we can:

1. generate samples x from P(X),
2. reject any sample which is not compatible with e.

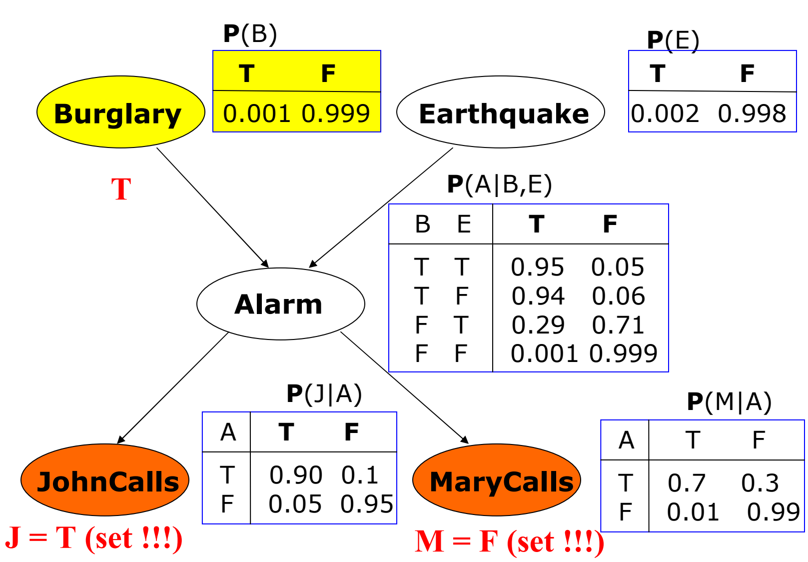
Problem: the number of accepted particles can be quite small. The expected number is MP(e).

- The number of samples required to achieve a low relative error grows linearly with 1/P(e)

Likelihood Weighting

- **Idea:** Instead of generating samples that are rejected, simply force the samples to take on the appropriate values at observed nodes.
-
- **Problem:** particles are generated with probability that is different from P(x)
- **Solution:** each particle generated is assigned a weight that represents P(e) for that sample

BBN likelihood weighting example



P(B)

T	F
0.001	0.999

P(E)

T	F
0.002	0.998

P(A|B,E)

B	E	T	F
T	T	0.95	0.05
T	F	0.94	0.06
F	T	0.29	0.71
F	F	0.001	0.999

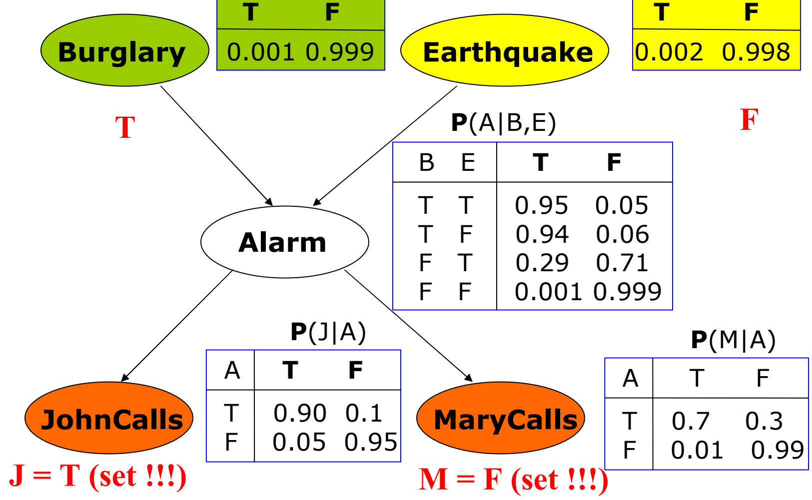
P(J|A)

A	T	F
T	0.90	0.1
F	0.05	0.95

P(M|A)

A	T	F
T	0.7	0.3
F	0.01	0.99

BBN likelihood weighting example



P(B)

T	F
0.001	0.999

P(E)

T	F
0.002	0.998

P(A|B,E)

B	E	T	F
T	T	0.95	0.05
T	F	0.94	0.06
F	T	0.29	0.71
F	F	0.001	0.999

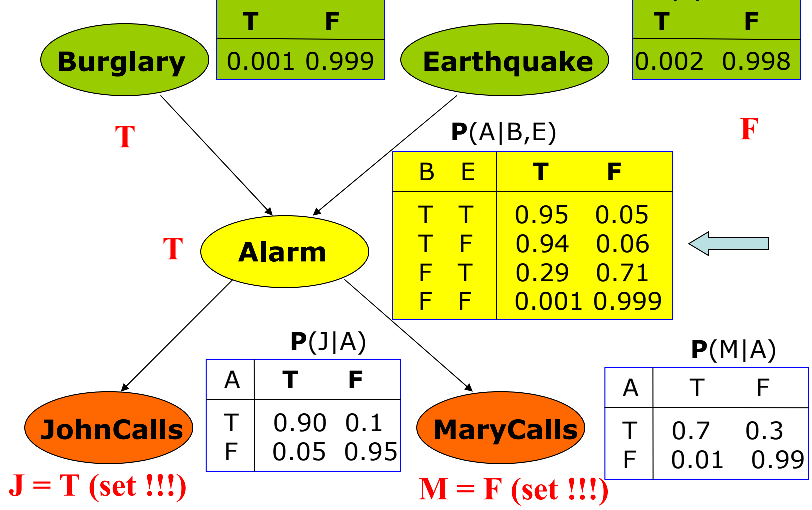
P(J|A)

A	T	F
T	0.90	0.1
F	0.05	0.95

P(M|A)

A	T	F
T	0.7	0.3
F	0.01	0.99

BBN likelihood weighting example



P(B)

T	F
0.001	0.999

P(E)

T	F
0.002	0.998

P(A|B,E)

B	E	T	F
T	T	0.95	0.05
T	F	0.94	0.06
F	T	0.29	0.71
F	F	0.001	0.999

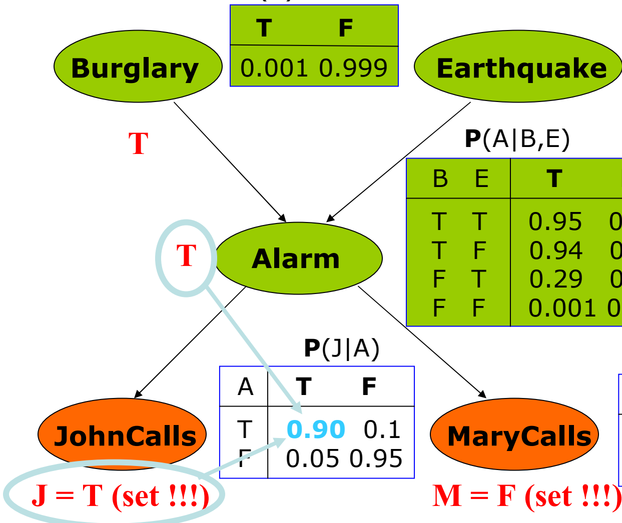
P(J|A)

A	T	F
T	0.90	0.1
F	0.05	0.95

P(M|A)

A	T	F
T	0.7	0.3
F	0.01	0.99

BBN likelihood weighting example



P(B)

T	F
0.001	0.999

P(E)

T	F
0.002	0.998

P(A|B,E)

B	E	T	F
T	T	0.95	0.05
T	F	0.94	0.06
F	T	0.29	0.71
F	F	0.001	0.999

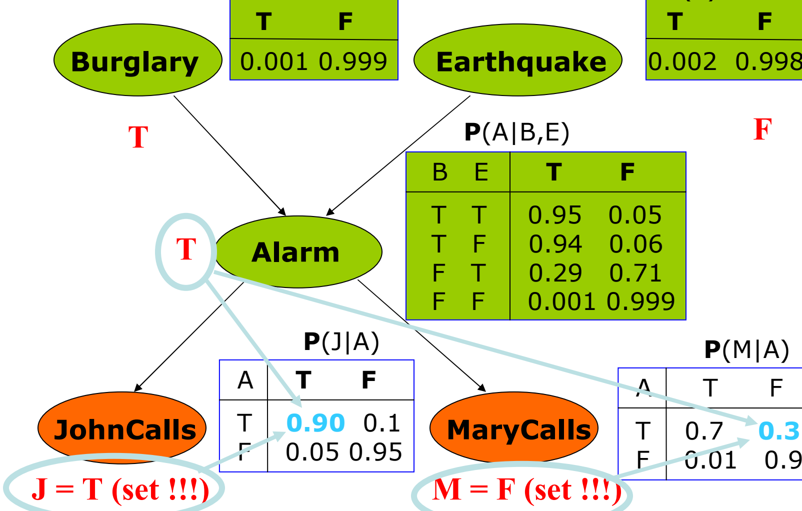
P(J|A)

A	T	F
T	0.90	0.1
F	0.05	0.95

P(M|A)

A	T	F
T	0.7	0.3
F	0.01	0.99

BBN likelihood weighting example



P(B)

T	F
0.001	0.999

P(E)

T	F
0.002	0.998

P(A|B,E)

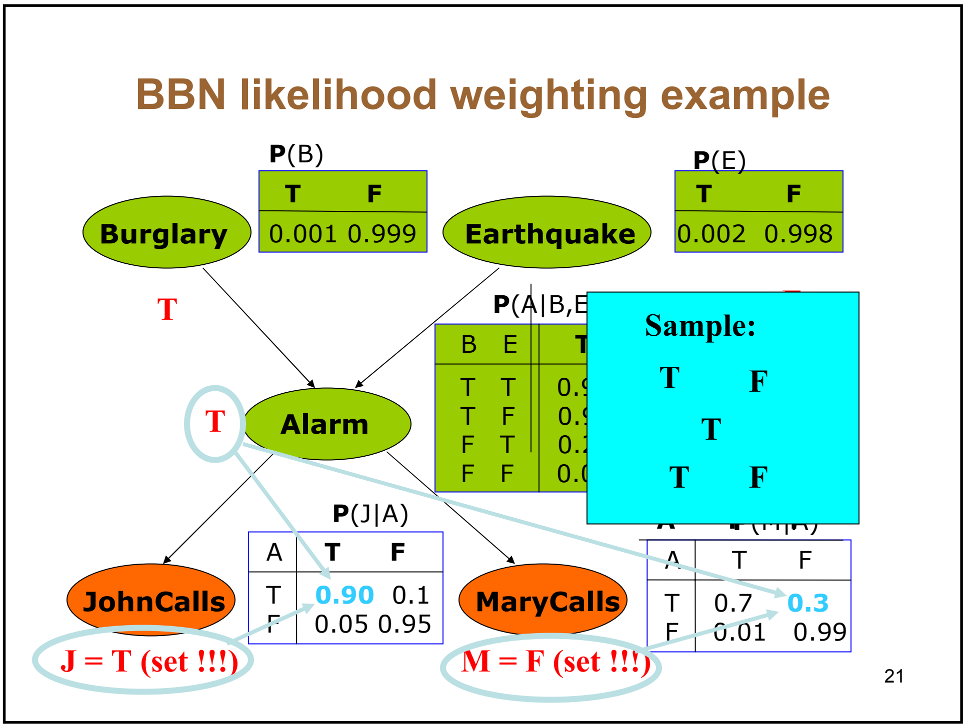
B	E	T	F
T	T	0.95	0.05
T	F	0.94	0.06
F	T	0.29	0.71
F	F	0.001	0.999

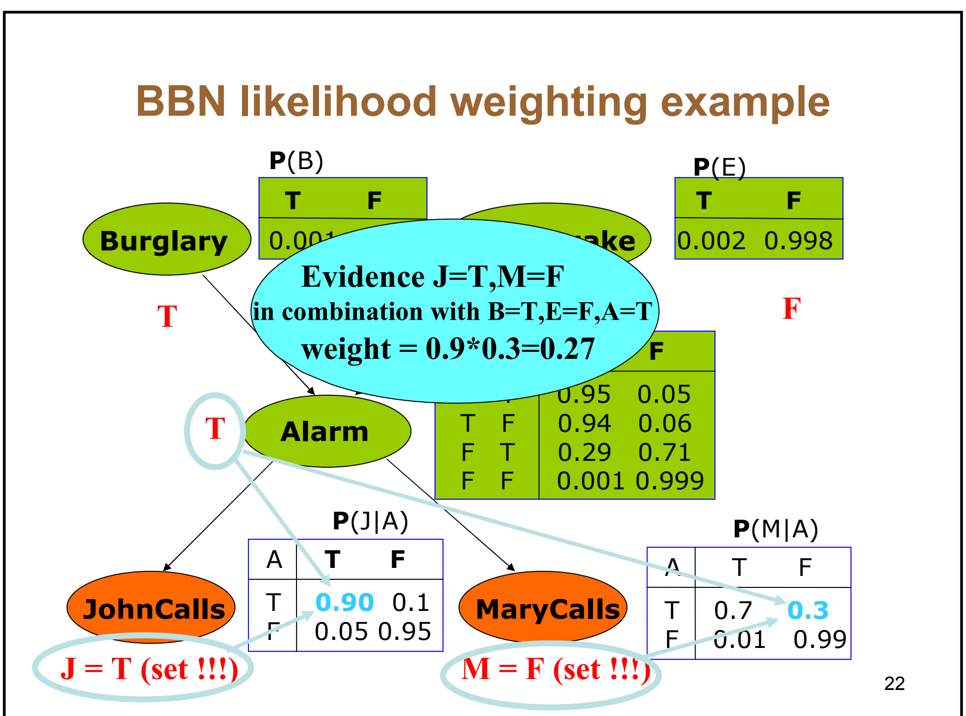
P(J|A)

A	T	F
T	0.90	0.1
F	0.05	0.95

P(M|A)

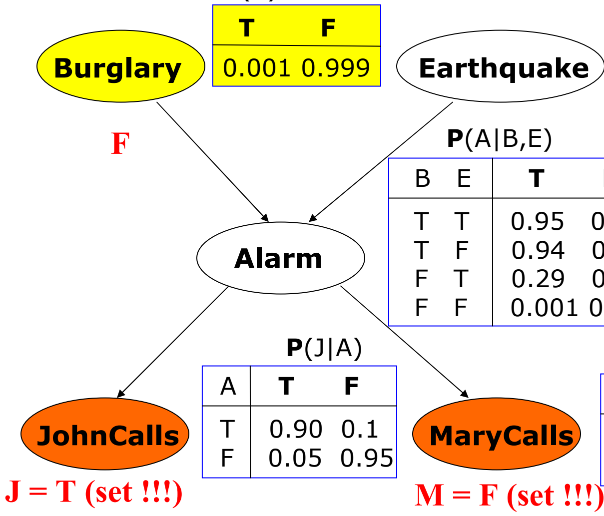
A	T	F
T	0.7	0.3
F	0.01	0.99


BBN likelihood weighting example
P(B)
T F
0.001 0.999
Burglary
Earthquake
P(E)
T F
0.002 0.998
T
P(A|B,E
B E T
T T 0.
T F 0.
F T 0.
F F 0.
Sample:
T F
T
T F
T
Alarm
P(J|A)
A T F
T 0.90 0.1
F 0.05 0.95
JohnCalls
MaryCalls
A T F
T 0.7 0.3
F 0.01 0.99
J = T (set !!!)
M = F (set !!!)
21


BBN likelihood weighting example
P(B)
T F
0.00
Burglary
P(E)
T F
0.002 0.998
Evidence J=T,M=F
in combination with B=T,E=F,A=T
weight = 0.9*0.3=0.27
T
F
F
0.95 0.05
T F 0.94 0.06
F T 0.29 0.71
F F 0.001 0.999
T
Alarm
P(J|A)
A T F
T 0.90 0.1
F 0.05 0.95
JohnCalls
MaryCalls
P(M|A)
A T F
T 0.7 0.3
F 0.01 0.99
J = T (set !!!)
M = F (set !!!)
22

BBN likelihood weighting example

Second sample



P(B)

T	F
0.001	0.999

P(E)

T	F
0.002	0.998

P(A|B,E)

B	E	T	F
T	T	0.95	0.05
T	F	0.94	0.06
F	T	0.29	0.71
F	F	0.001	0.999

P(J|A)

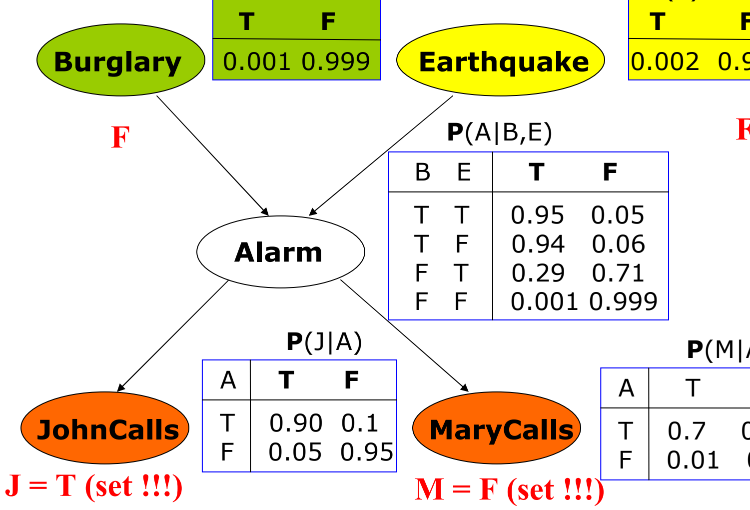
A	T	F
T	0.90	0.1
F	0.05	0.95

P(M|A)

A	T	F
T	0.7	0.3
F	0.01	0.99

BBN likelihood weighting example

Second sample



P(B)

T	F
0.001	0.999

P(E)

T	F
0.002	0.998

P(A|B,E)

B	E	T	F
T	T	0.95	0.05
T	F	0.94	0.06
F	T	0.29	0.71
F	F	0.001	0.999

P(J|A)

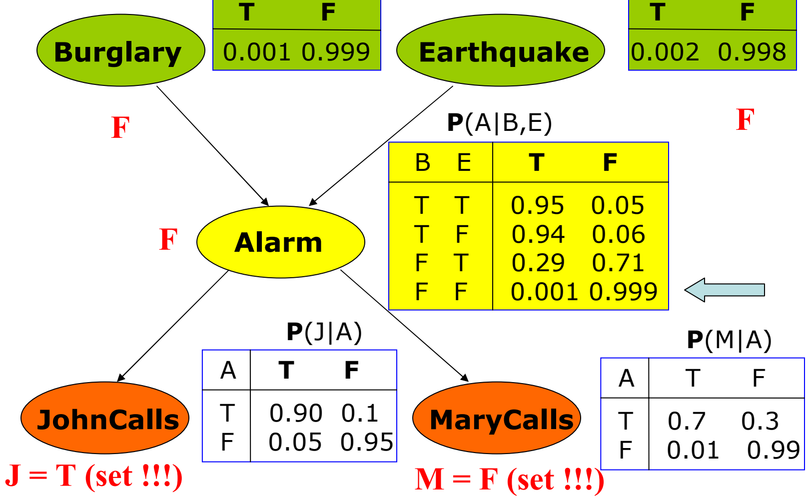
A	T	F
T	0.90	0.1
F	0.05	0.95

P(M|A)

A	T	F
T	0.7	0.3
F	0.01	0.99

BBN likelihood weighting example

Second sample



P(B)

T	F
0.001	0.999

P(E)

T	F
0.002	0.998

P(A|B,E)

B	E	T	F
T	T	0.95	0.05
T	F	0.94	0.06
F	T	0.29	0.71
F	F	0.001	0.999

P(J|A)

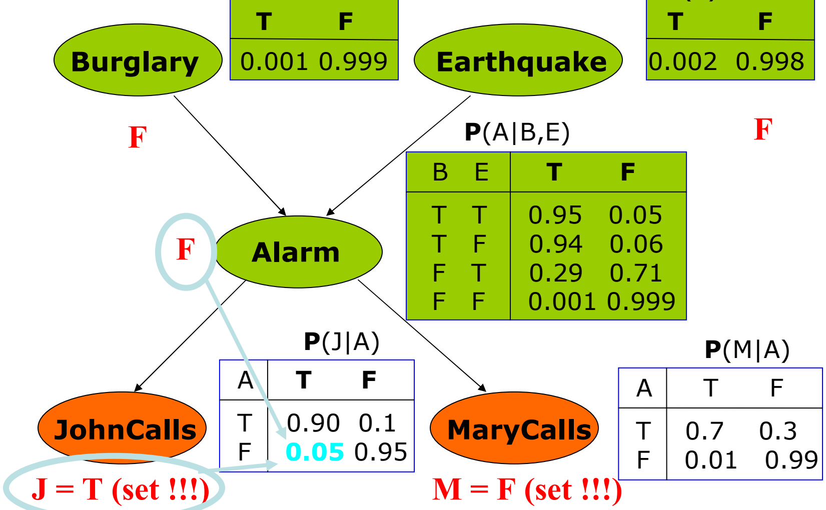
A	T	F
T	0.90	0.1
F	0.05	0.95

P(M|A)

A	T	F
T	0.7	0.3
F	0.01	0.99

BBN likelihood weighting example

Second sample



P(B)

T	F
0.001	0.999

P(E)

T	F
0.002	0.998

P(A|B,E)

B	E	T	F
T	T	0.95	0.05
T	F	0.94	0.06
F	T	0.29	0.71
F	F	0.001	0.999

P(J|A)

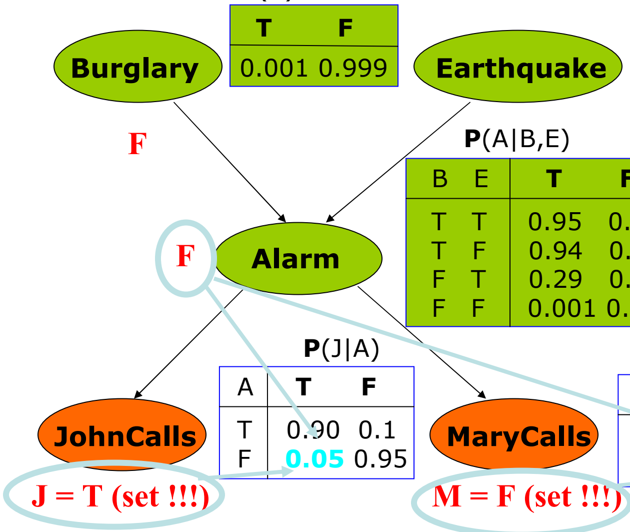
A	T	F
T	0.90	0.1
F	0.05	0.95

P(M|A)

A	T	F
T	0.7	0.3
F	0.01	0.99

BBN likelihood weighting example

Second sample



P(B)

T	F
0.001	0.999

P(E)

T	F
0.002	0.998

P(A|B,E)

B	E	T	F
T	T	0.95	0.05
T	F	0.94	0.06
F	T	0.29	0.71
F	F	0.001	0.999

P(J|A)

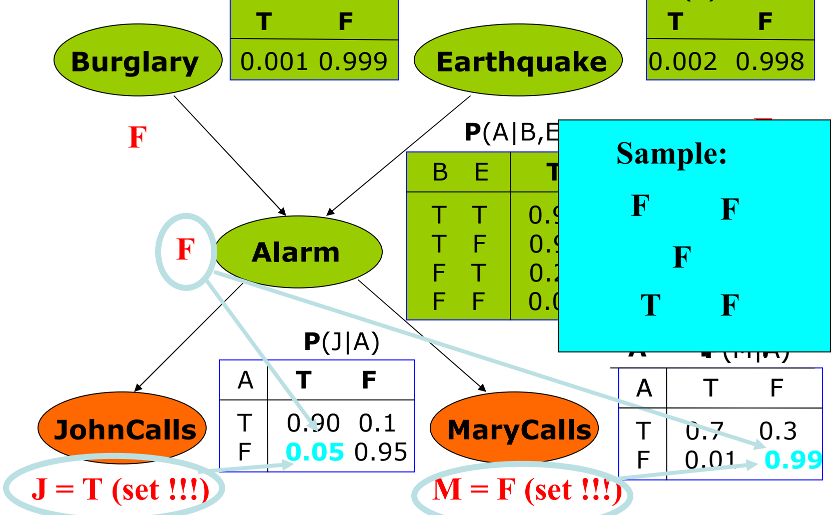
A	T	F
T	0.90	0.1
F	0.05	0.95

P(M|A)

A	T	F
T	0.7	0.3
F	0.01	0.99

BBN likelihood weighting example

Second sample



P(B)

T	F
0.001	0.999

P(E)

T	F
0.002	0.998

P(A|B,E

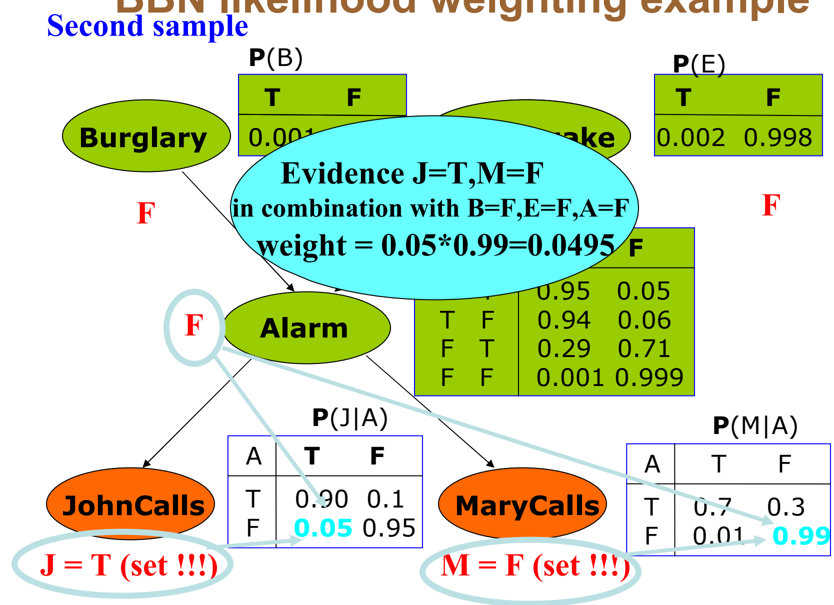
B	E	T
T	T	0.9
T	F	0.9
F	T	0.2
F	F	0.0

P(J|A)

A	T	F
T	0.90	0.1
F	0.05	0.95

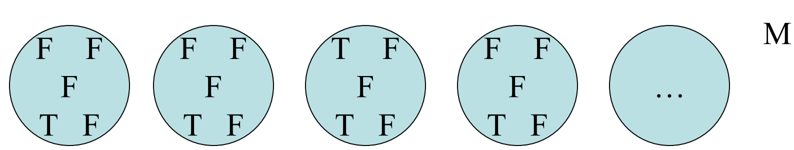
A	T	F
T	0.7	0.3
F	0.01	0.99

BBN likelihood weighting example



Likelihood weighting

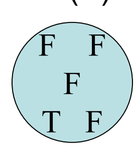
- Assume we have generated the following M samples:



- If we calculate the estimate:

$$P(B=T \mid J=T, M=F) = \frac{\#sample\_with(B=T)}{\#total\_sample}$$

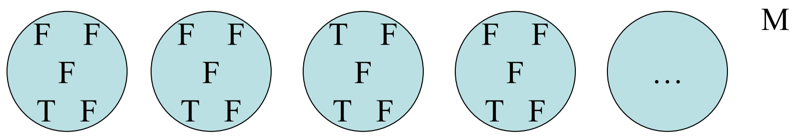
a less likely sample from P(X) may be generated more often.

- For example, sample  is generated more often than in P(X)

- So the samples are not consistent with P(X).

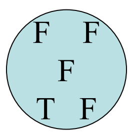
Likelihood weighting

- Assume we have generated the following M samples:

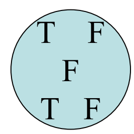


How to make the samples consistent?

Weight each sample by probability with which it agrees with the conditioning evidence P(e).



← **Weight 0.0495**



← **Weight 0.27**

Likelihood weighting

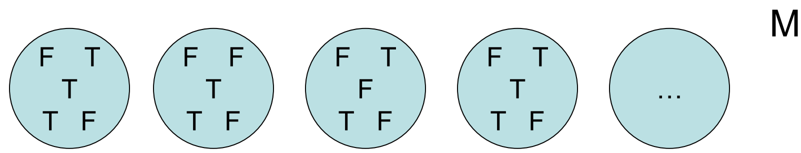
- How to compute weights for the sample?
- Assume the query $P(B=T \mid J=T, M=F)$
- Likelihood weighting:
  - **With every sample keep a weight with which it should count towards the estimate**

$$\tilde{P}(B=T \mid J=T, M=F) = \frac{\sum_{i=1}^{M} 1\{B^{(i)}=T\} w^{(i)}}{\sum_{i=1}^{M} w^{(i)}}$$

$$\tilde{P}(B=T \mid J=T, M=F) = \frac{\sum_{\text{samples with } B=T \text{ and } J=T, M=F} w_{B=T}}{\sum_{\text{samples with any value of } B \text{ and } J=T, M=F} w_{B=x}}$$

Likelihood weighting

- Assume we have generated the following M samples:



- If we calculate the estimate:

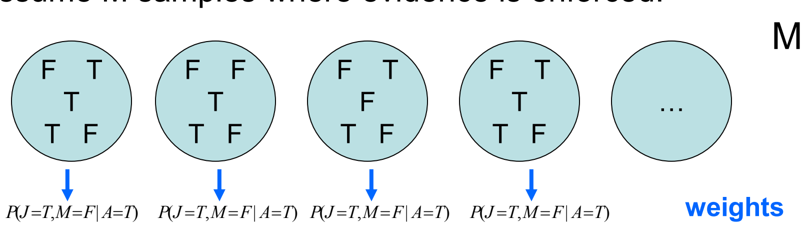
$$P(A=T \mid J=T, M=F) = \frac{\#sample\_with(A=T)}{\#total\_sample}$$

a less likely sample from P(x) may be generated more often. So the samples are not consistent with P(x).

How to make the samples consistent? The probability of the evidence P(e) for the sample tells us how likely the evidence is in the sample. So we can use P(e) to weight each sample and correct the bias.

Likelihood weighting

- Assume M samples where evidence is enforced:



- We can use P(e) to weight each sample and correct the bias.
- The correct estimate is then:

$$\tilde{P}(A=T \mid J=T, M=F) = \frac{\sum_{i=1}^{M} 1\{A^{(i)}=T\} w^{(i)}}{\sum_{i=1}^{M} w^{(i)}}$$

Likelihood weighted Particle Generation

```
Procedure LW-sample (B, Z=z)
//B – Bayesian network over X, Z– event in the network
Let X_1,..., X_n be a topological ordering of X
w ← 1
for i=1,..., n
    u_i ← x<Pa_{X_i}> //Assignment to Pa_{X_i} in x_1,...,x_{i-1}
    If X_i ∉ Z then
        Sample x_i from P(X_i | u_i)
    else
        x_i ← z<X_i> //Assignment to X_i in z
        w ← wP(x_i | u_i) //Multiply weight by probability of desired value
return (x_1,...x_n), w
```


Likelihood Weighting

Summary

Likelihood Weighting

- generates M weighted particles $<\xi[1],w[1]>,\ldots,<\xi[M],w[M]>$ using LW Sample procedure.
- Estimates the conditional probability P(y|e) using M samples as :

$$\hat{P}(y \mid e) = \frac{\sum_{m=1}^{M} w[m] 1\{y[m] = y\}}{\sum_{m=1}^{M} w[m]}$$

Importance Sampling

- Importance Sampling is a general approach for estimating the expectation of a function f(x) relative to some distribution P(X) (target distribution):

$$E_p[f] = \sum_{\{x\}} P(x) f(x) \quad \text{or} \quad E_p[f] = \int_x p(x) f(x) dx$$

- Generally, we can estimate this expectation by generating samples x[1], …, x[M] from P, and then estimating

$$\tilde{E}_p[f] = \frac{1}{M} \sum_{m=1}^{M} f(x[m])$$

Importance Sampling

- Estimate of $\tilde{E}_p[f]$ requires to sample P(x)
- It might be impossible or computationally very expensive to generate samples directly from P.
- Because of that we might prefer to generate samples from a different distribution Q (**a proposal or sampling distribution**) instead

- A **proposal distribution Q** can be arbitrary, but it should satisfy:

  Q(x)>0 whenever P(x)>0

Unnormalized Importance Sampling
(P is Known)

- Since we generate samples from Q instead of P, we must adjust our estimator to compensate for the incorrect sampling distribution.

$$E_{p(X)}[f(X)] = E_{Q(x)}[f(x)\frac{P(x)}{Q(x)}] = E_{Q(x)}[f(x)w(x)]$$

- We use standard estimator for expectations relative to Q. We generate a set of samples D={x[1],…,x[M]} from Q, and estimate:

$$\hat{E}_D(f) = \frac{1}{M}\sum_{m=1}^{M} f(x[m])\frac{P(x[m])}{Q(x[m])}$$

Unnormalized Importance Sampling
(P is Known)

- This is an unbiased estimator: its mean for any data set is precisely the desired value

- We can estimate the distribution of this estimator around its mean: as M $\rightarrow \infty$

$$E_{Q(X)}[f(X)w(X)] - E_p[f(X)] \propto N(0; \sigma_Q^{\ 2}/M)$$

$$w(x) = P(x)/Q(x)$$

where $\sigma_Q^{\ 2} = [E_{Q(X)}[(f(X)w(X))^2]] - (E_{P(X)}[f(X)])^2$

Unnormalized Importance Sampling
(P is Known)

- The variance of this estimator decreases linearly with the number of samples.
- When f(X)=1, the variance is simply the weighting function P(X)/Q(X). Thus the more different Q is from P, the higher the variance will be.
- The lowest variance is achieved when

$$Q(X) \propto |f(X)| P(X)$$

- We should avoid cases where our sampling probability Q(X)<<P(X)f(X) in any part of the space, as these cases can lead to very large or even infinite variance.

Normalized Importance Sampling
(P is known up to a normalizing constant)

- When P is only known up to a normalizing constant α, but we have access to a function P'(X), such that P' is not a normalized distribution, but P' (X)= α P(x)
- In this context, we cannot define the weights relative to P, so we define:

$$w(X) = \frac{P'(X)}{Q(X)}$$

$$E_{P(X)}[f(X)] = \sum_x P(x) f(x) = \sum_x Q(x) f(x) \frac{P(X)}{Q(x)} = \frac{1}{\alpha} \sum_x Q(x) f(x) \frac{P'(x)}{Q(x)}$$

$$= \frac{1}{\alpha} E_{Q(x)}[f(X) w(X)] = \frac{E_{Q(X)}[f(X) w(X)]}{E_{Q(X)}[w(X)]}$$

Normalized Importance Sampling
(P is known up to a normalizing constant)

- Using an empirical estimator for both the numerator and denominator, we can estimate:

$$\hat{E}_D(f) = \frac{\sum_{m=1}^{M} f(x[m])w(x[m])}{\sum_{m=1}^{M} w(x[m])}$$

- Although the normalized estimator is biased, its variance is typically lower than that of the unnormalized estimator. This reduction in variance often outweighs the bias term.
- Normalized estimator is often used in place of the unnormalized estimator, even in cases where P is known and we can sample from it effectively.

Proposal Distribution
based on the Mutilated Belief network

Assume a Bayesian Network

- We want to calculate P(x|e)
- This is hard if we need to go opposite the links and account for the effect of evidence on nondescendants

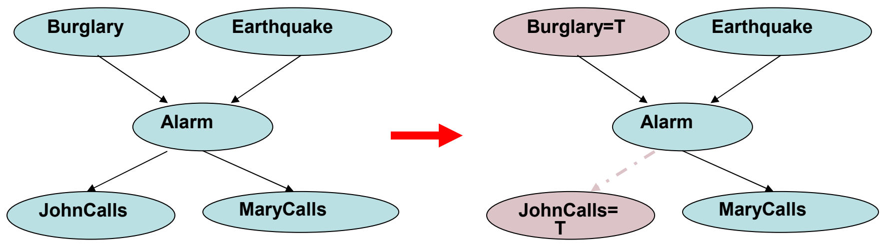
Objective: generate particles efficiently using a simpler proposal distribution Q(x)

Solution: **a mutilated belief network**

- Idea:
  - Avoid propagation of evidence effects to nondescendants;
  - Disconnect all variables in the evidence from their parents

Mutilated Belief network

- Assume we want to calculate P(x|B=T,J=T) in the Alarm network
- Use B=T and J=T to build a mutilated network



Original network

Mutilated network

Mutilated Belief network

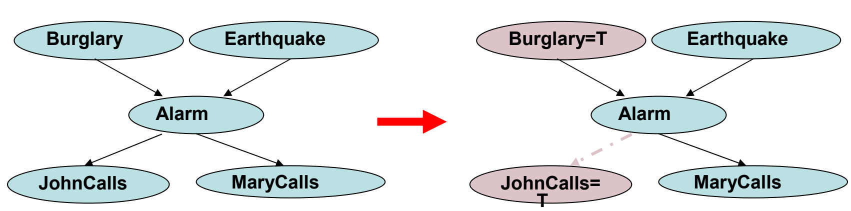
- Assume the evidence is J=j* and B=b*
- Original network:

$$P(E=e, A=a, M=m, J=j^*, B=b^*) = P(b^*)P(e)P(a \mid b^*, e)P(j^* \mid a)P(m \mid a)$$

- Mutilated network:

$$Q(E=e, A=a, M=m, J=j^*, B=b^*) = P(e)P(a \mid b^*, e)P(m \mid a)$$

- Note that $w(x) = \frac{P(x)}{Q(x)} = P(b^*)P(j^* \mid a)$



Original network

Mutilated network

Mutilated Belief network

- Assume the evidence is J=j* and B=b*
- Original network:

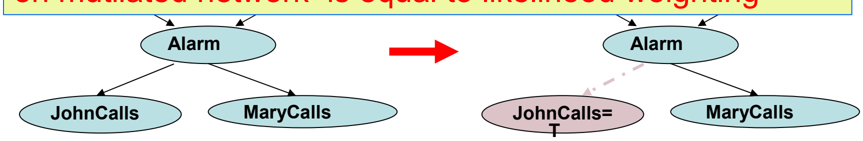
$$P(E=e, A=a, M=m, J=j^*, B=b^*) = P(b^*)P(e)P(a \mid b^*, e)P(j^* \mid a)P(m \mid a)$$

- Mutilated network:

$$Q(E=e, A=a, M=m, J=j^*, B=b^*) = P(e)P(a \mid b^*, e)P(m \mid a)$$

- Note that $w(x) = \frac{P(x)}{Q(x)} = P(b^*)P(j^* \mid a)$

So importance sampling with a proposal distribution based on mutilated network is equal to likelihood weighting



Original network

Mutilated network

Data-Dependent Likelihood Weighting

- **Question:** When to stop? How many samples do we need to see?
- **Intuition:** not every samples contribute equally to the quality of the estimate. A sample with high weight is more compatible with the evidence e, and may provide us with more information.
- **Solution:** We stop sampling when the total weight of the generated particles reaches a pre-defined value.
- **Benefits:** It allows early stopping in cases where we were lucky in our random choice of samples.

Ratio Likelihood Weighting

- Estimate the conditional probability P(y|e) in two phases: use likelihood weighing to estimate P(e) and P(y,e) separately.

- Use LW M times with the argument E=e to generate a set D of weighted samples $(\xi[1], w[1]), \ldots, (\xi[M], w[M])$
  use the same algorithm M' times with argument Y=y, E=e to generate another set D' of weighted samples
  $(\xi'[1], w'[1]), \ldots, (\xi'[M], w'[M])$

- Then we can estimate:

$$\hat{P}_D(y \mid e) = \frac{\hat{P}_{D'}(y, e)}{\hat{P}_D(e)} = \frac{1/M' \sum_{m=1}^{M'} w'[m]}{1/M \sum_{m=1}^{M} w[m]}$$

Q&A

- Thank you!