CS 3750 Machine Learning
Lecture 9

Graphical models
Monte Carlo inference

Milos Hauskrecht
milos@cs.pitt.edu
5329 Sennott Square

Monte Carlo approaches

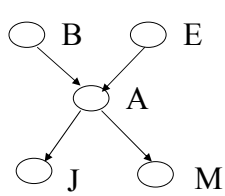
- **MC approximation**:
  - The probability is approximated using sample frequencies
  - **Example:**

$$\tilde{P}(B=T, J=T) = \frac{N_{B=T,J=T}}{N}$$

$\# samples \;\; with \;\; B=T, J=T$

$total \;\; \# samples$

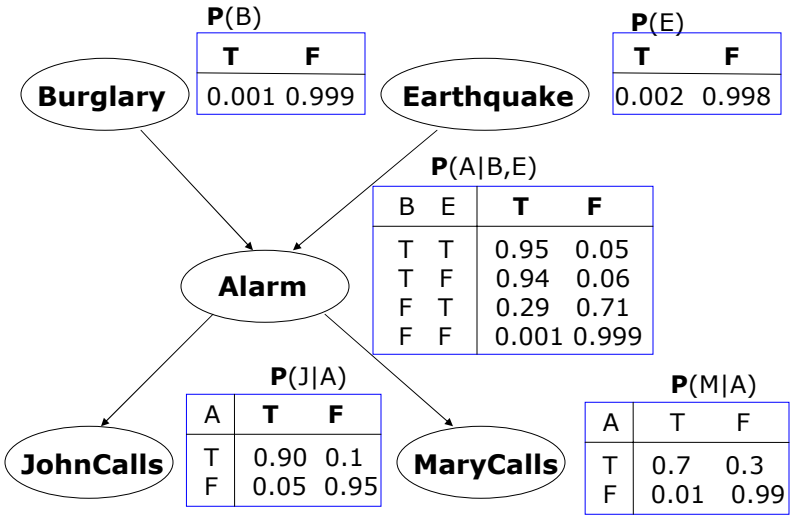
- **BBN sampling:**



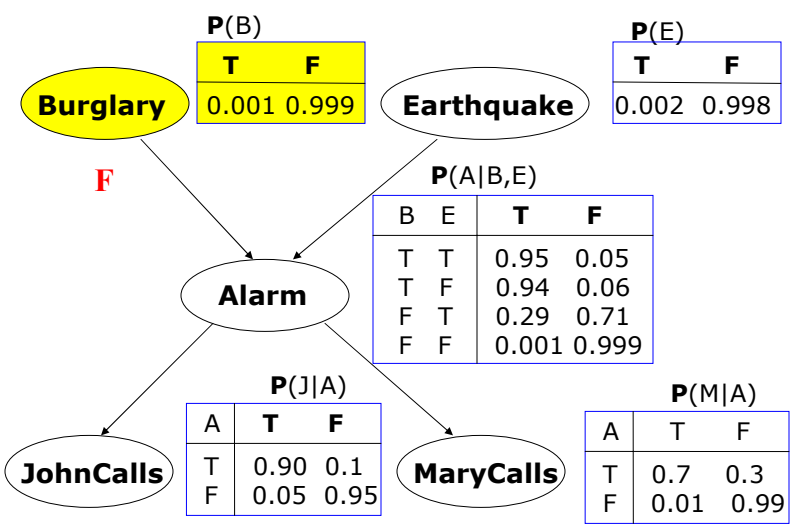
Generate sample in a top down manner, following the links

- **One sample gives one assignment of values to all variables**

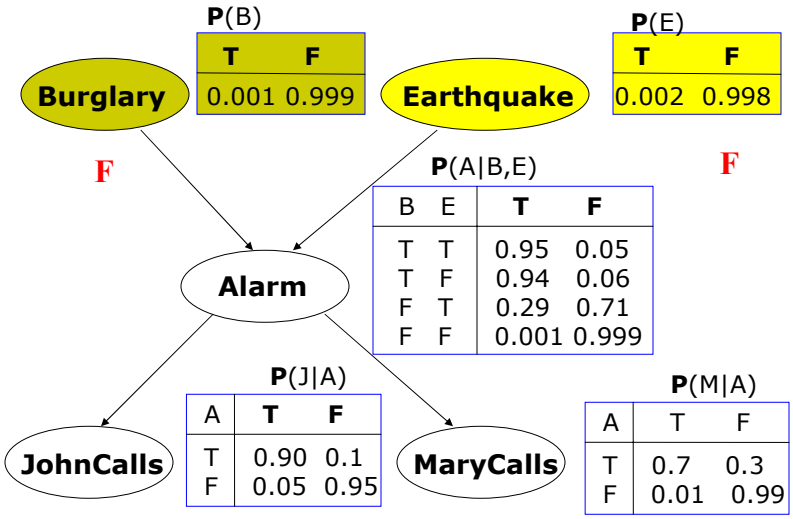
BBN sampling example



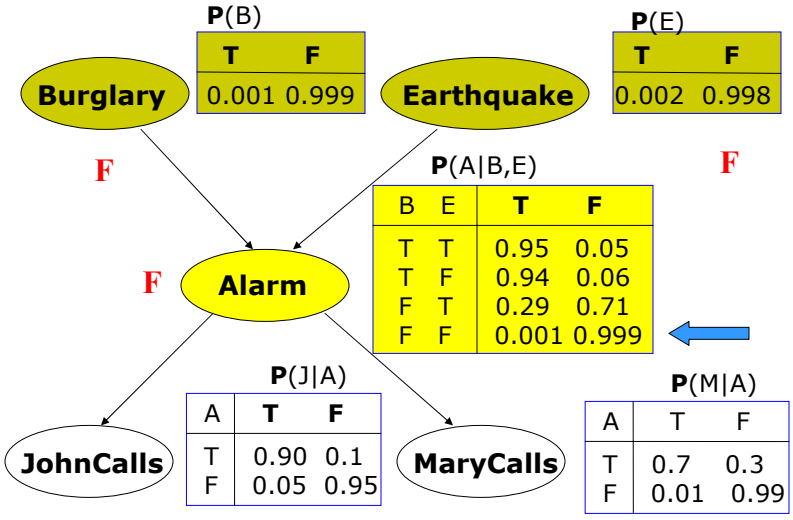
BBN sampling example



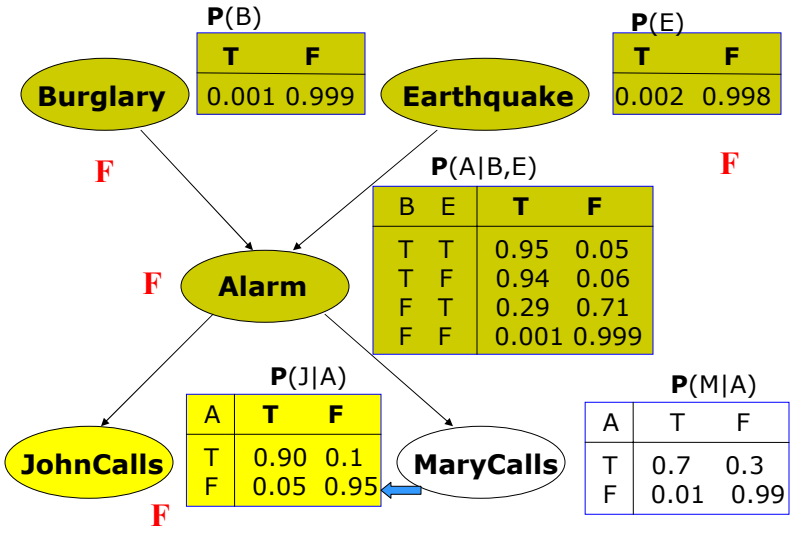
BBN sampling example



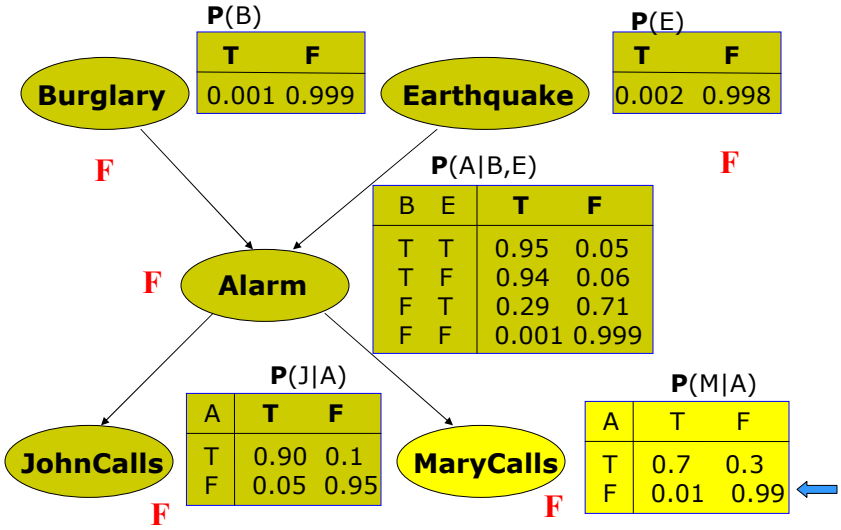
BBN sampling example



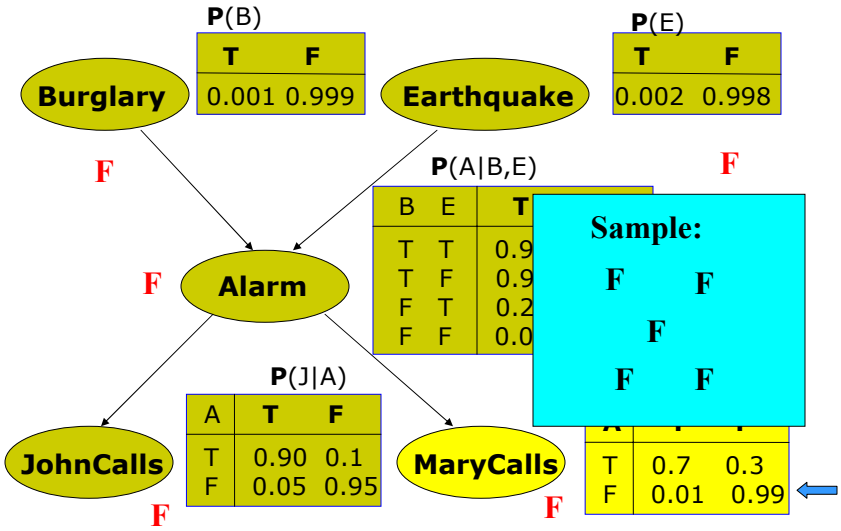
BBN sampling example



BBN sampling example



BBN sampling example



Monte Carlo approaches

- **MC approximation of conditional probabilities**:
  - The probability is approximated using sample frequencies
  - **Example:**

$$\tilde{P}(B=T \mid J=T) = \frac{N_{B=T,J=T}}{N_{J=T}}$$

$N_{B=T,J=T}$: *# samples with* $B=T, J=T$

$N_{J=T}$: *# samples with* $J=T$

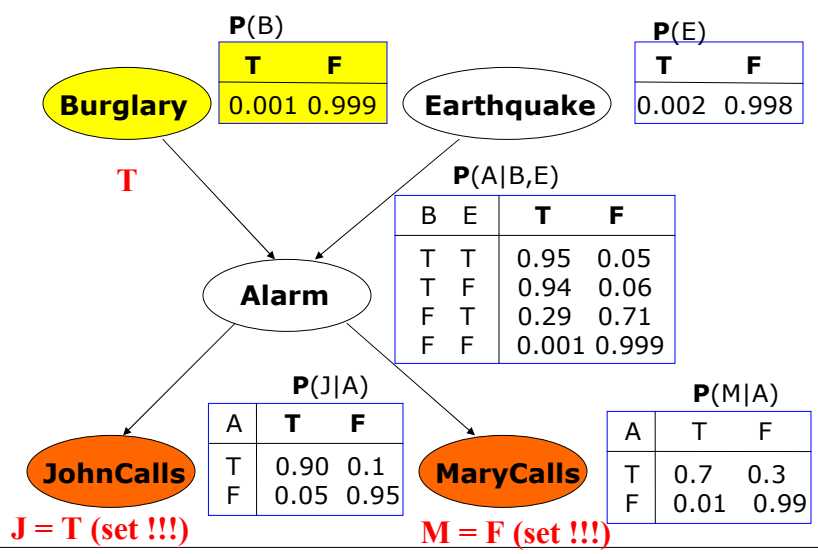
- **Rejection sampling:**
  - Generate samples from the full joint by sampling BBN
  - Use only samples that agree with the condition, the remaining samples are rejected
- **Problem:** many samples can be rejected

Likelihood weighting

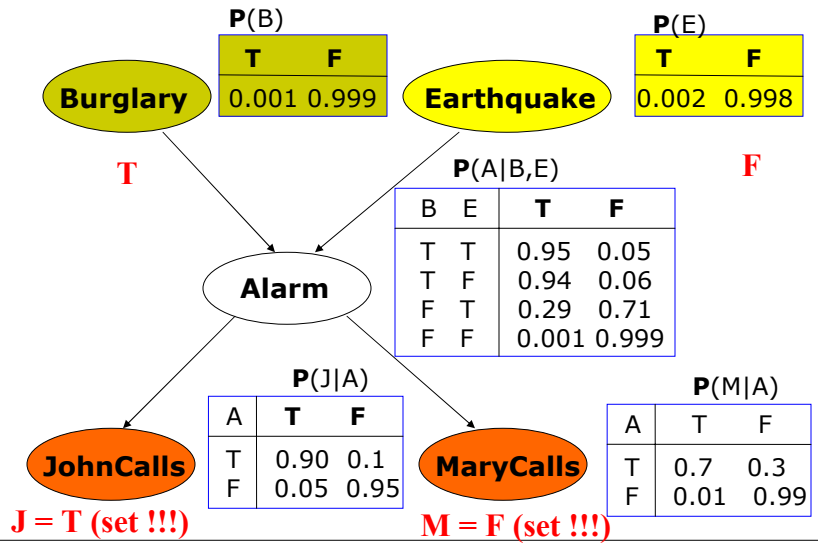
- **Avoids inefficiencies of rejection sampling**
  - **Idea:** generate only samples consistent with an evidence (or conditioning event)
  - **If the value is set no sampling**
- **Problem:** using simple counts is not enough since these may occur with different probabilities
- Likelihood weighting:
  - **With every sample keep a weight with which it should count towards the estimate**

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

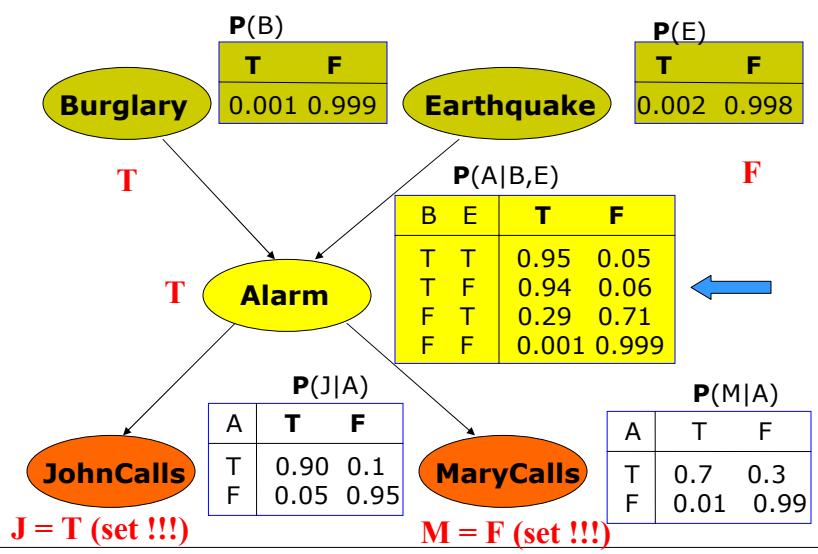
BBN likelihood weighting example



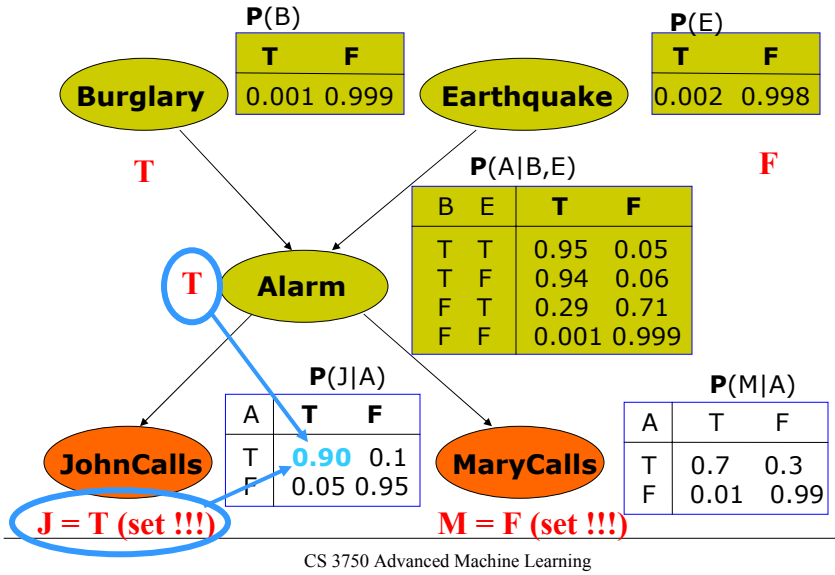
BBN likelihood weighting example



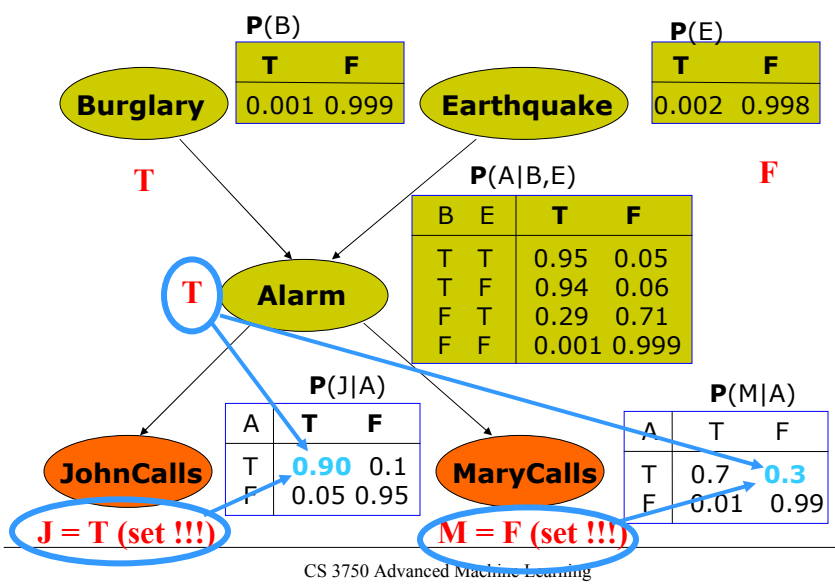
BBN likelihood weighting example



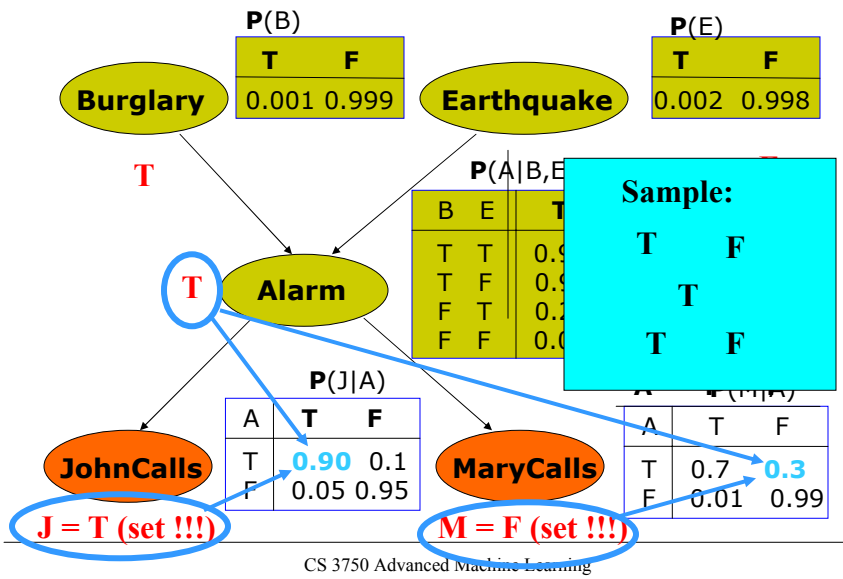
BBN likelihood weighting example



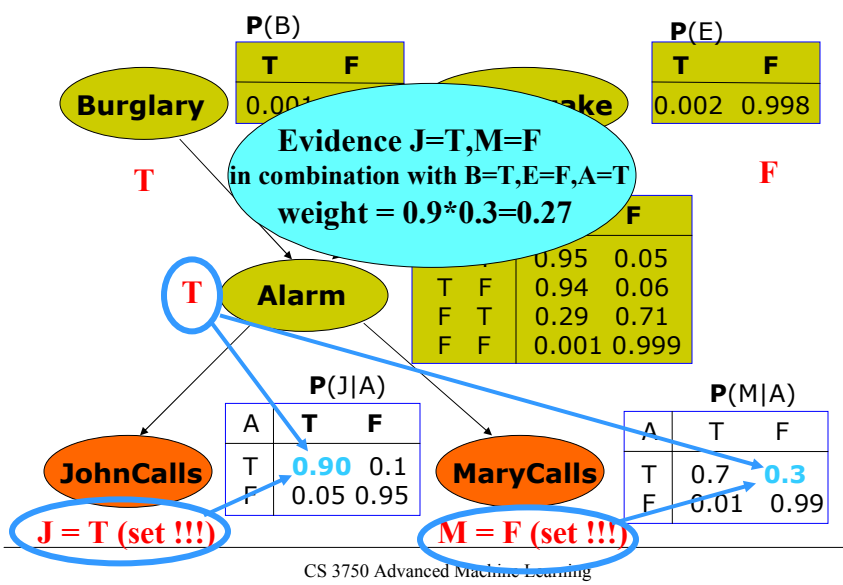
BBN likelihood weighting example



BBN likelihood weighting example

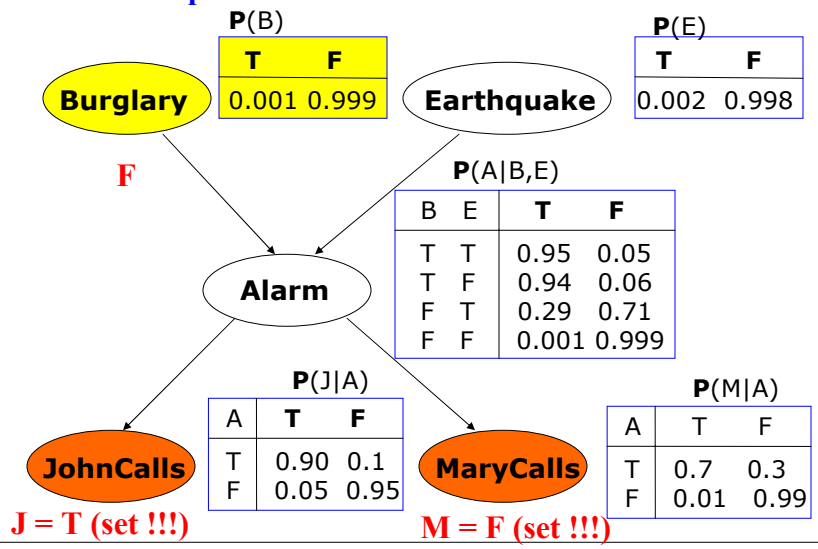


BBN likelihood weighting example



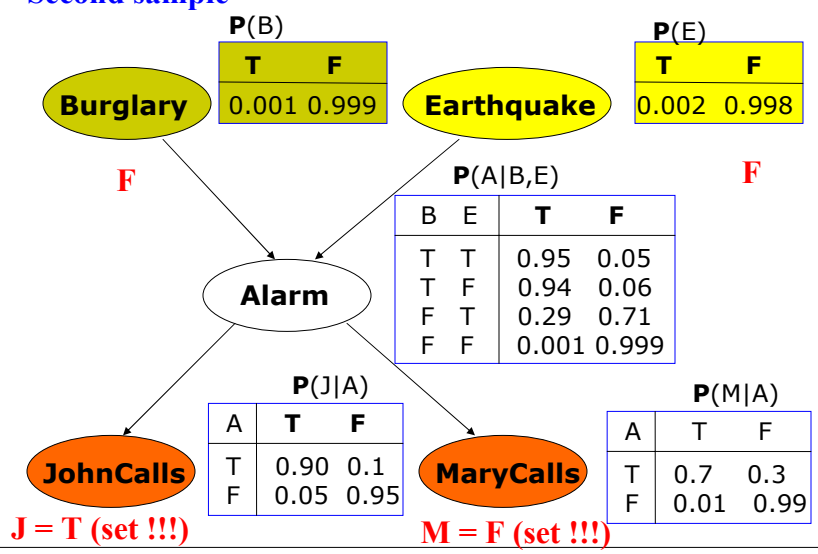
BBN likelihood weighting example

Second sample



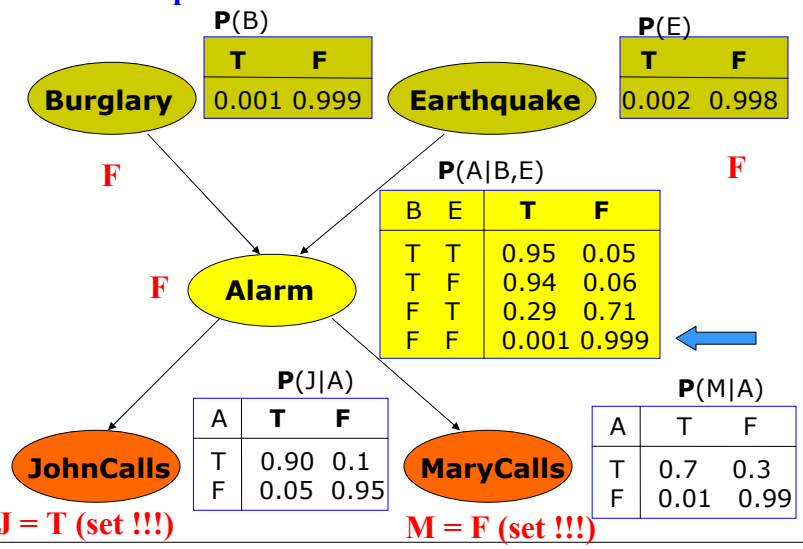
BBN likelihood weighting example

Second sample



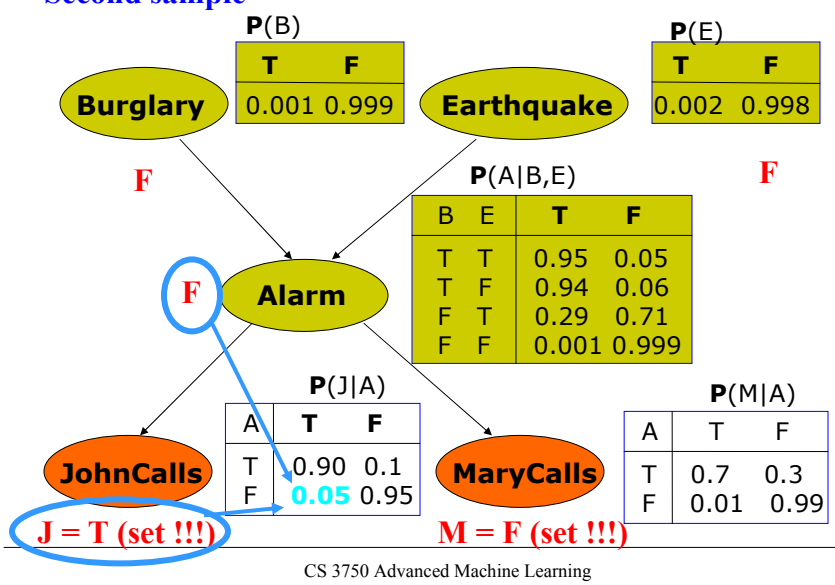
BBN likelihood weighting example

Second sample



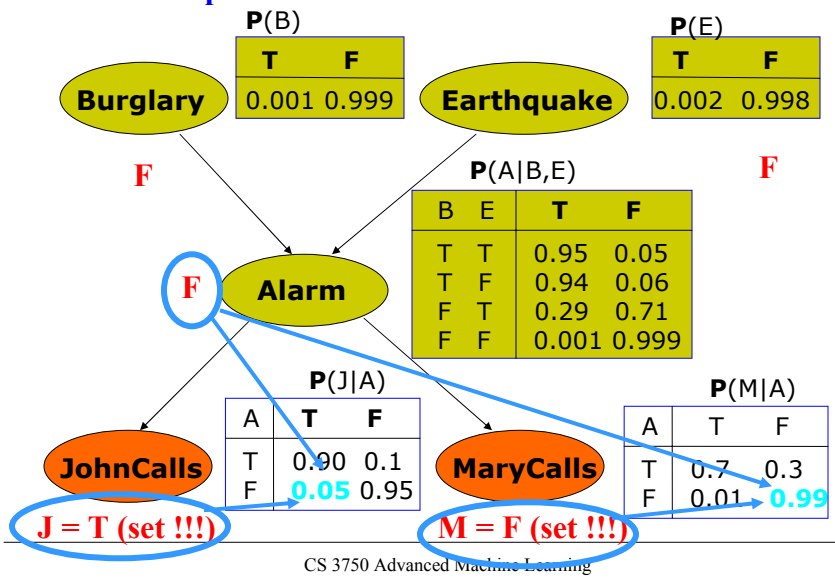
BBN likelihood weighting example

Second sample



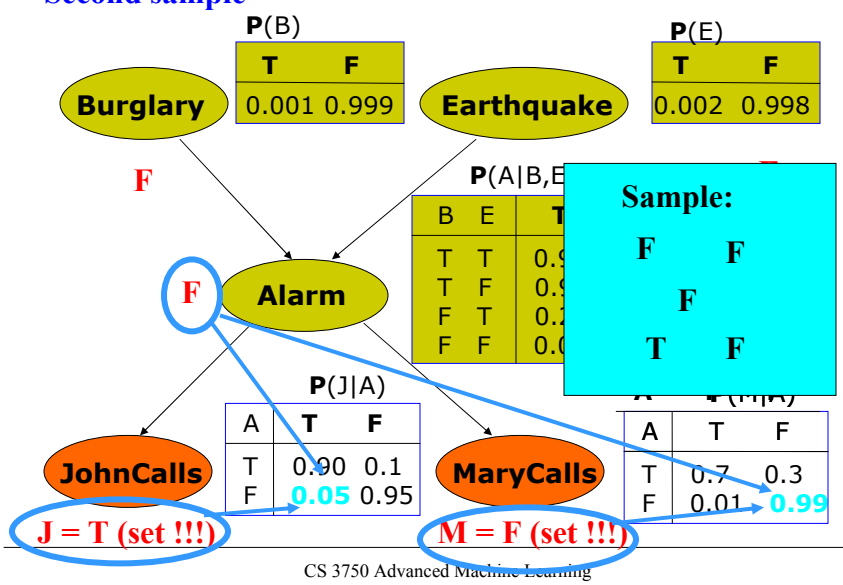
BBN likelihood weighting example

Second sample



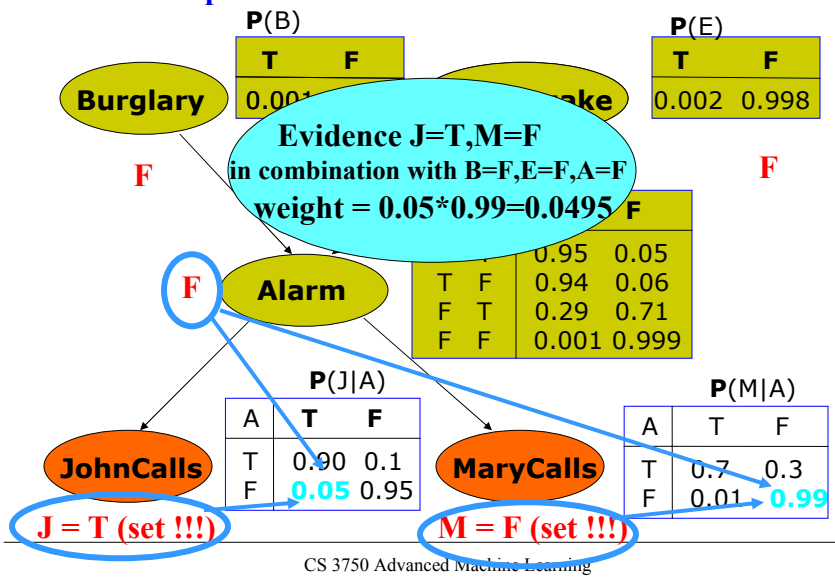
BBN likelihood weighting example

Second sample



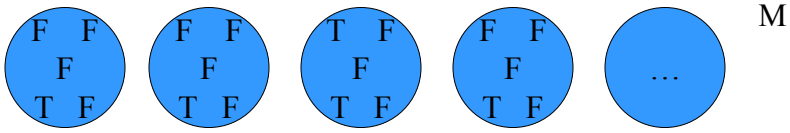
BBN likelihood weighting example

Second sample



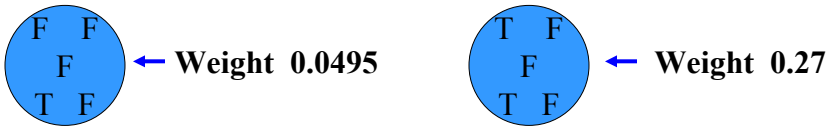
Likelihood weighting

- 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).



Importance Sampling

- an approach for estimating the expectation of a function f(x) relative to some distribution P(X) (target distribution)
- generally, we can estimate this expectation by generating samples x[1], …, x[M] from P, and then estimating

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

- However, we might prefer to generate samples from a different distribution Q (proposal or sampling distribution) instead, since it might be impossible or computationally very expensive to generate samples directly from P.

- Q can be arbitrary, but it should dominate P, i.e. Q(x)>0 whenever P(x)>0

Unnormalized Importance Sampling

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

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

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

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

Importance sampling

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

  $w(x) = P(x)/Q(x)$ - a weighting function, or a correction weight

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

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

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

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

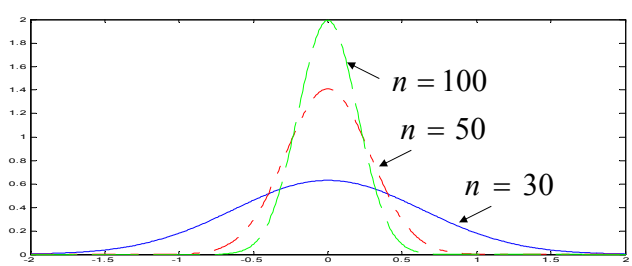
Central limit theorem

- **Central limit theorem:**

  Let random variables $X_1, X_2, \cdots X_n$ form a random sample from a distribution with mean μ and variance σ^2, then if the sample n is large, the distribution

$$\sum_{i=1}^{n} X_i \approx N(n\mu, n\sigma^2) \quad \text{or} \quad \frac{1}{n}\sum_{i=1}^{n} X_i \approx N(\mu, \sigma^2/n)$$

Effect of increasing the sample size *n* on the sample mean:



$\mu = 0$

$\sigma^2 = 4$

Importance sampling

- When f(X)=1, the variance is simply the variance of the weighting function P(X)/Q(X). Thus, the more different Q is from P, the higher is the variance of the estimator.
- In general, 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.

- Problem with unnormalized IS: P is assumed to be known