CS 2750 Machine Learning
Lecture 18

Bayesian belief networks IV

Milos Hauskrecht
milos@pitt.edu
5329 Sennott Square

Inference in Bayesian network

- **Exact inference algorithms:**
  - **Variable elimination**
  - Recursive decomposition (Cooper, Darwiche)
  - Symbolic inference (D'Ambrosio)
  - Belief propagation algorithm (Pearl)
  - Clustering and joint tree approach (Lauritzen, Spiegelhalter)
  - Arc reversal (Olmsted, Schachter)

- **Approximate inference algorithms:**
  - **Monte Carlo methods:**
    - Forward sampling, Likelihood sampling
  - Variational methods

Monte Carlo approaches

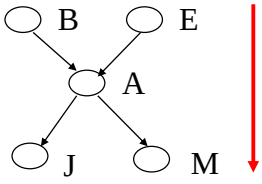
- **MC approximation**:
  - The probability is approximated using sample frequencies
  - **Example:**

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

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

N ← *total # samples*

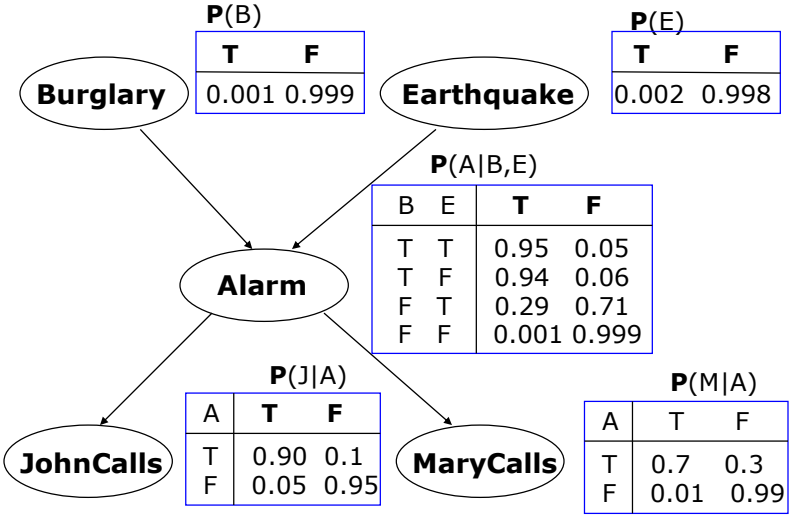
- **Sample generation: BBN sampling of the joint is easy**



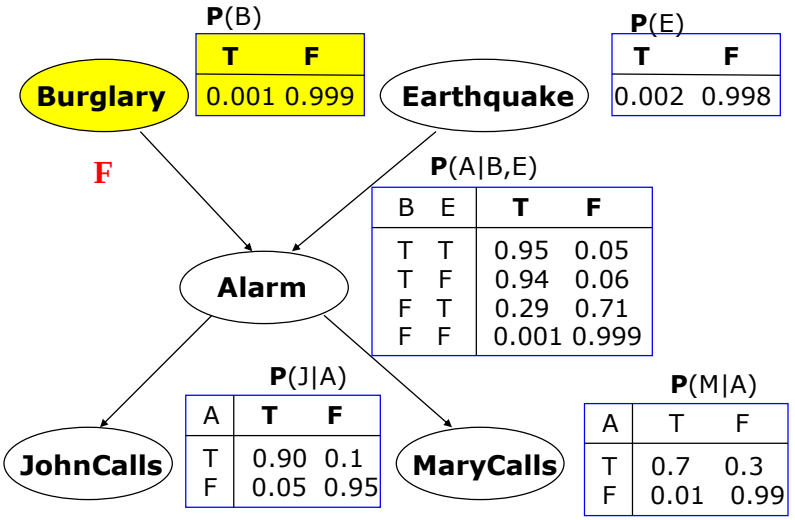
Generate sample in a top down manner, following the links

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

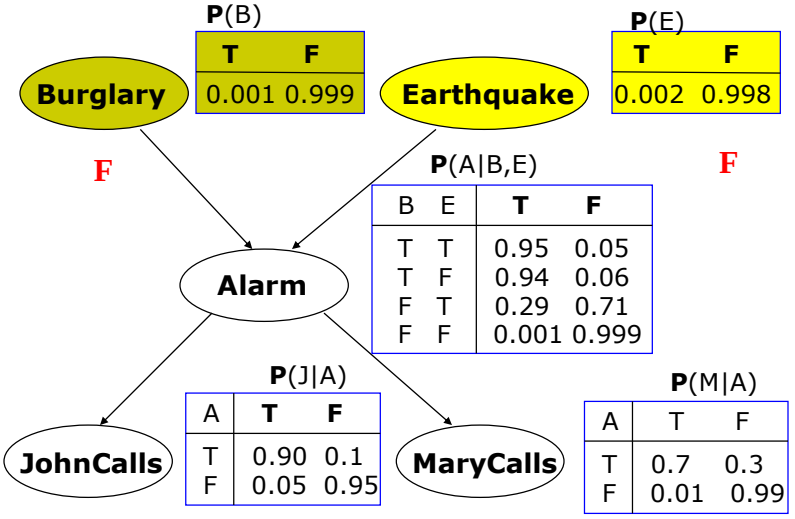
BBN sampling example



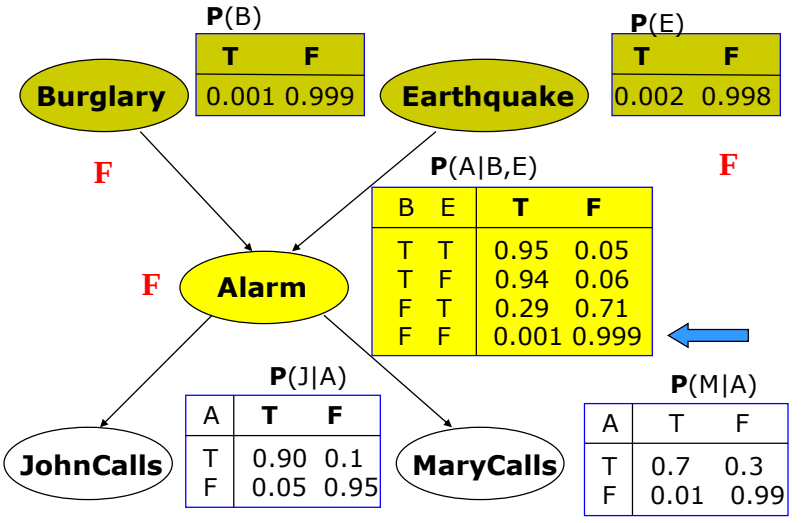
BBN sampling example



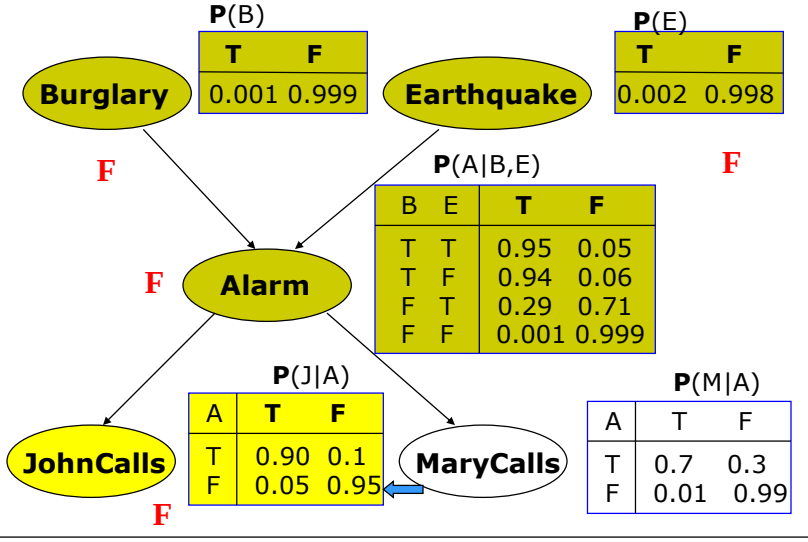
BBN sampling example



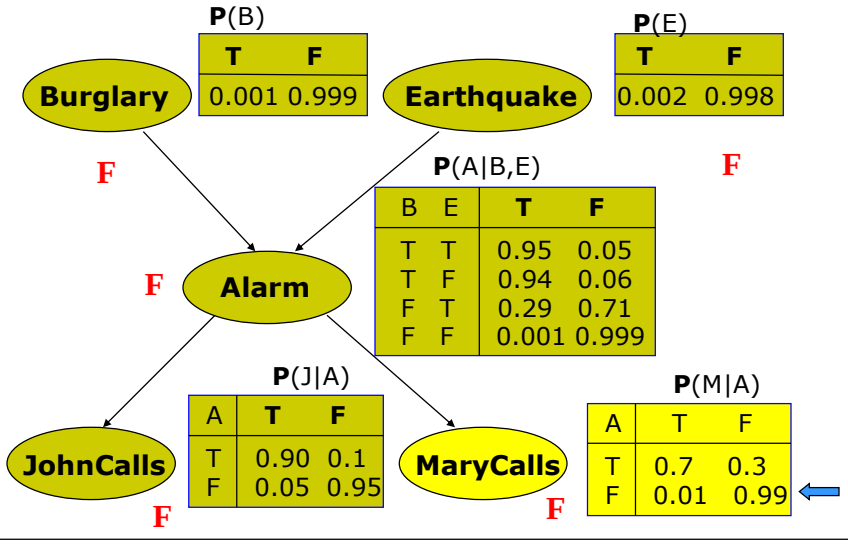
BBN sampling example



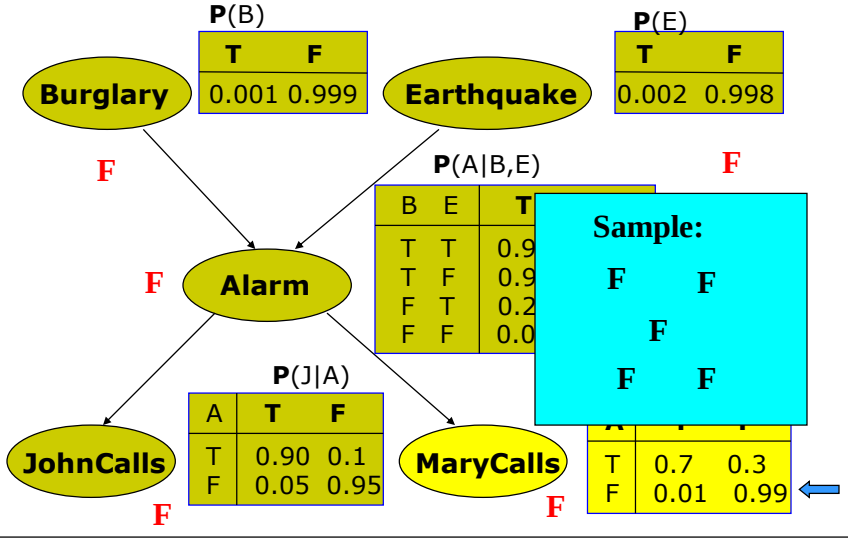
BBN sampling example



BBN sampling example



BBN sampling example



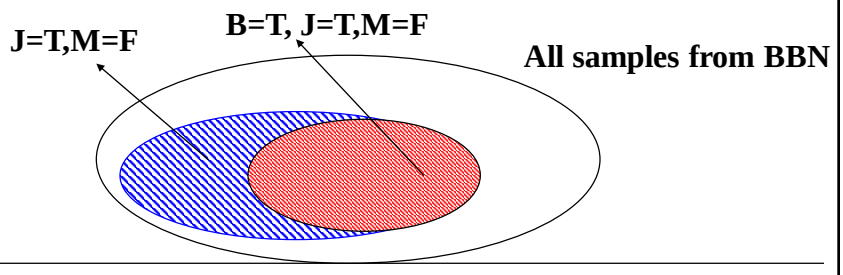
Monte Carlo approaches

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

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

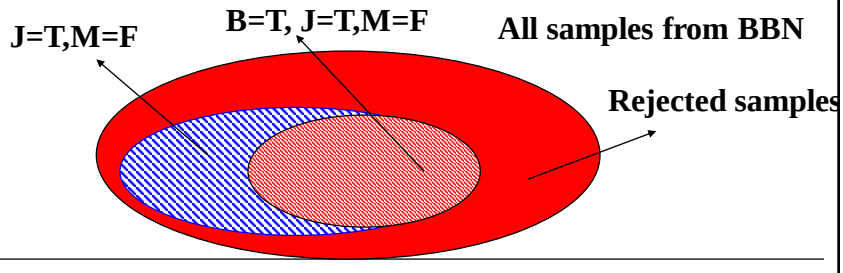
$\#\,samples\ with\ B=T, J=T, M=F$

$\#\,samples\ with\ J=T, M=F$



Monte Carlo approaches

- **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

Idea: generate only samples consistent with an evidence (or conditioning event)

– Benefit: Avoids inefficiencies of rejection sampling

Problem:

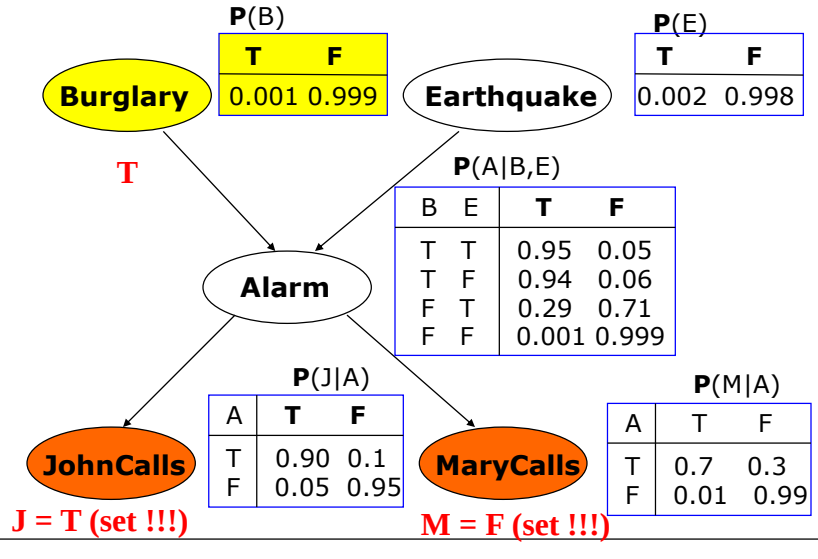
- **the distribution generated by enforcing the conditioning variables to set values is biased**
- simple counts are not sufficient to estimate the probabilities

Solution:

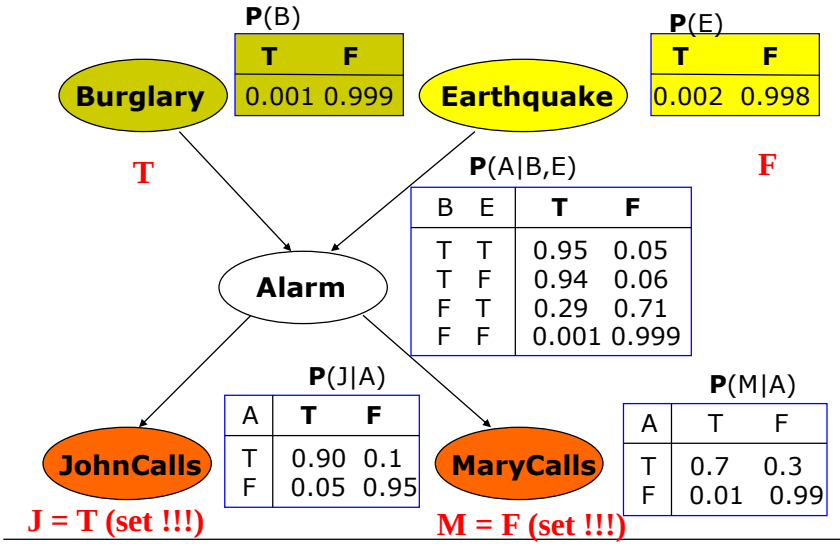
- **With every sample keep a weight with which it should count towards the estimate**

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

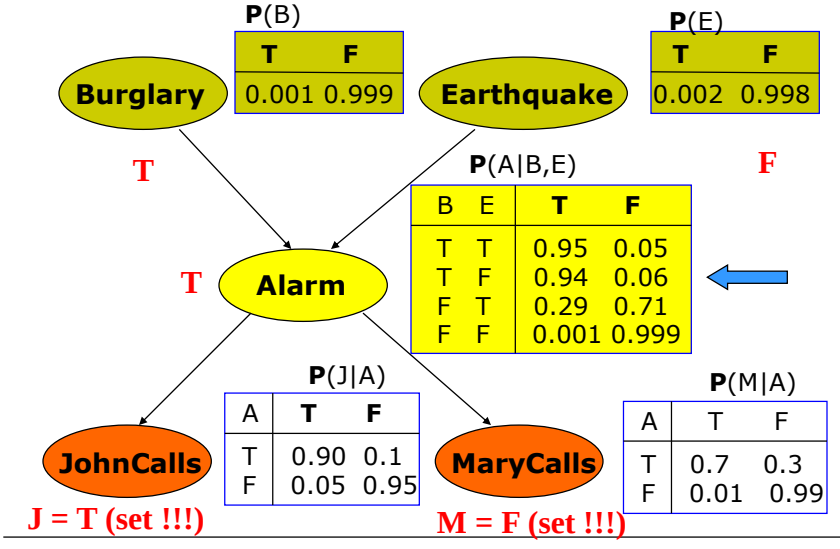
BBN likelihood weighting example



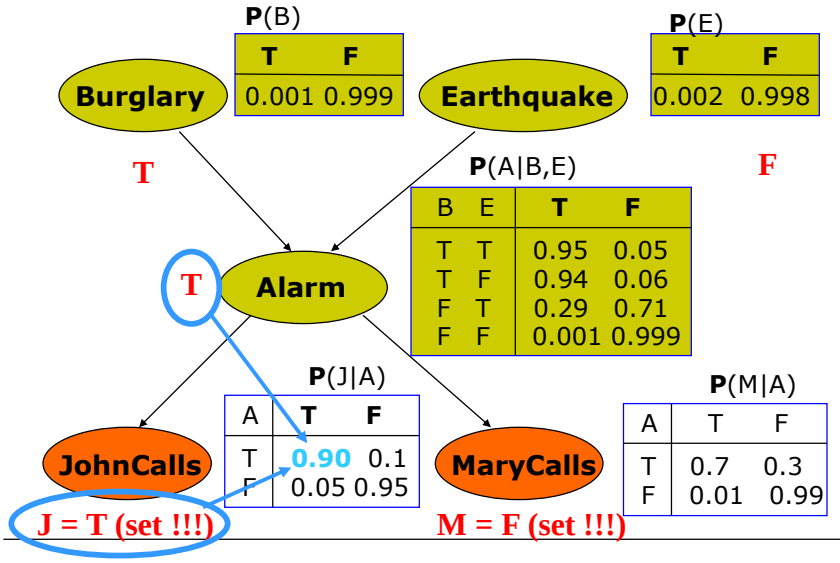
BBN likelihood weighting example



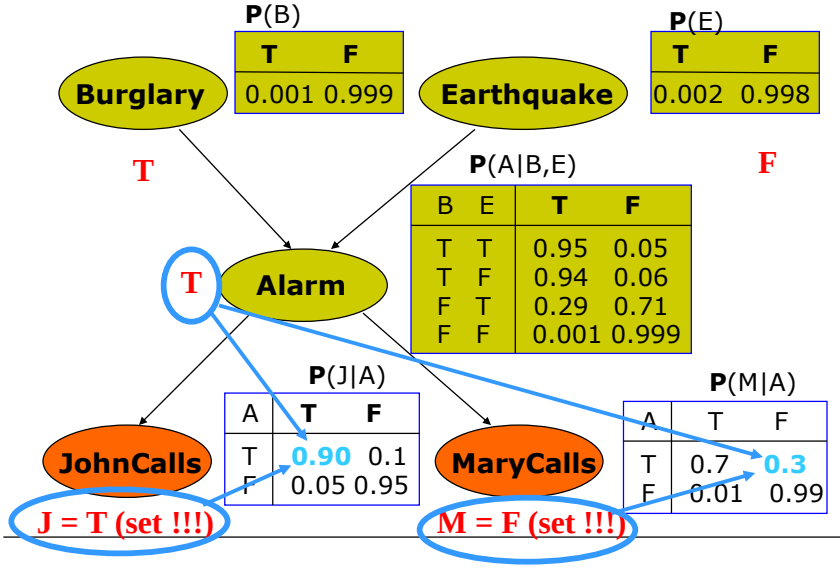
BBN likelihood weighting example



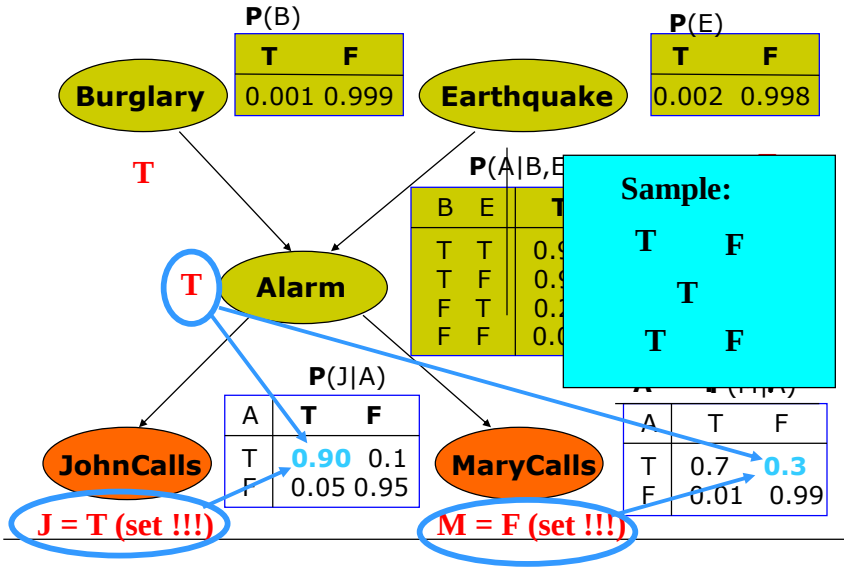
BBN likelihood weighting example



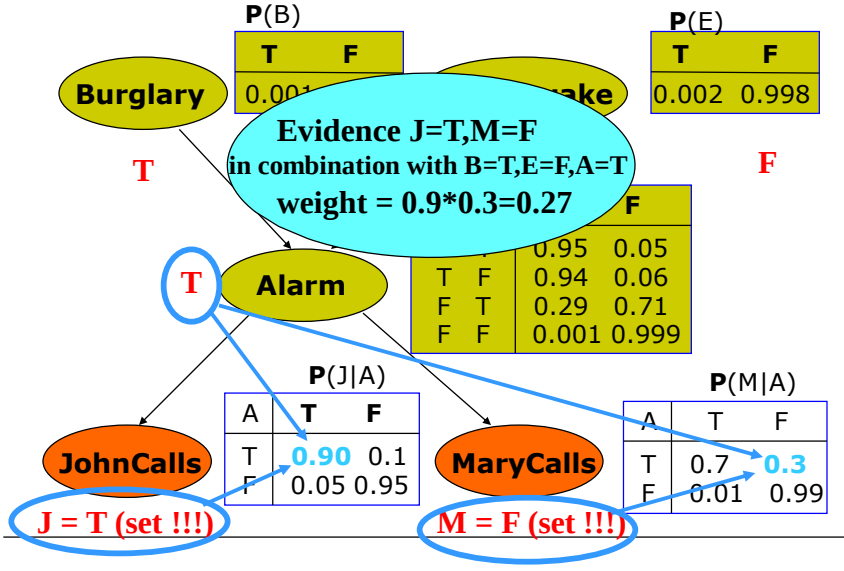
BBN likelihood weighting example



BBN likelihood weighting example

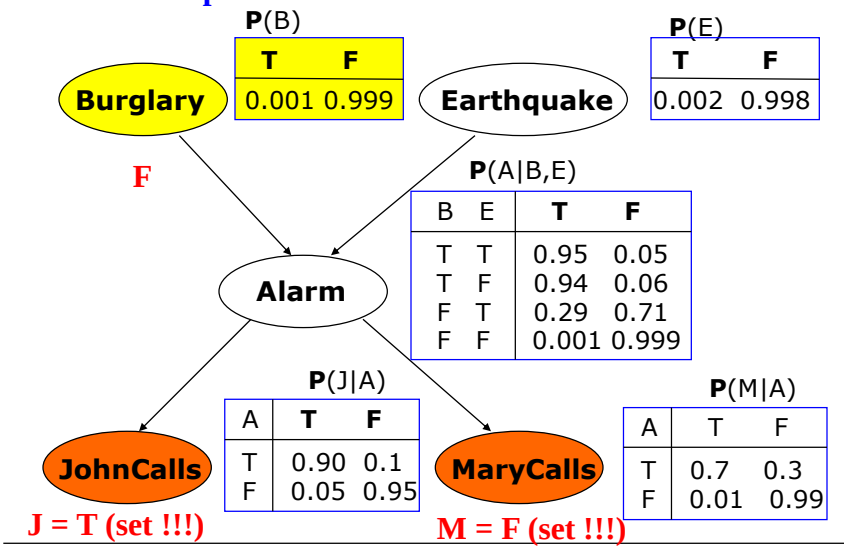


BBN likelihood weighting example



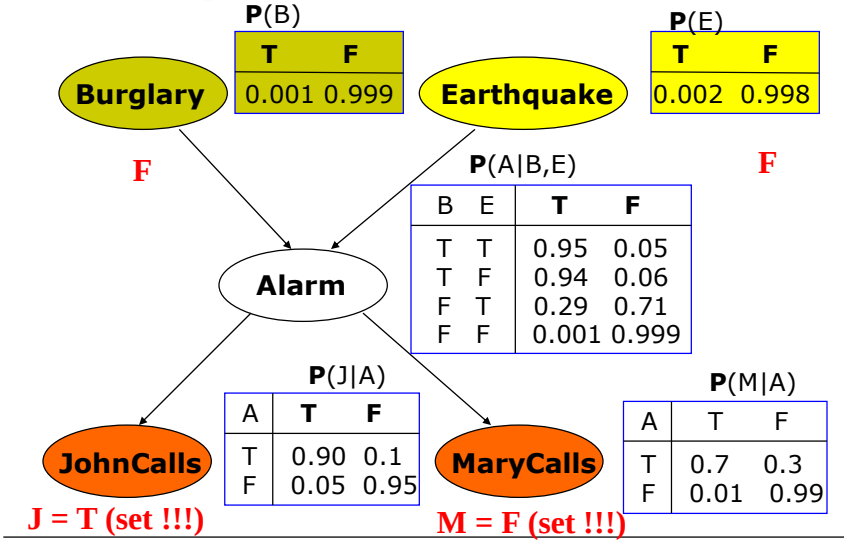
BBN likelihood weighting example

Second sample



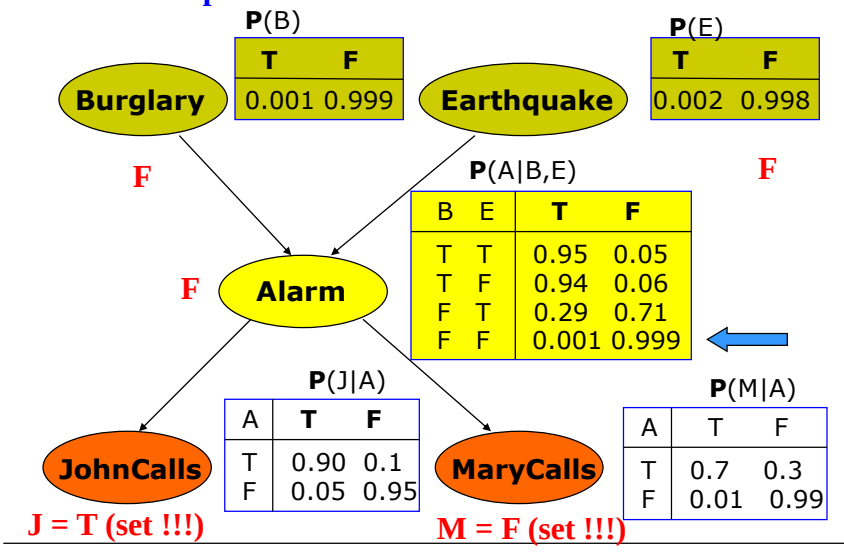
BBN likelihood weighting example

Second sample



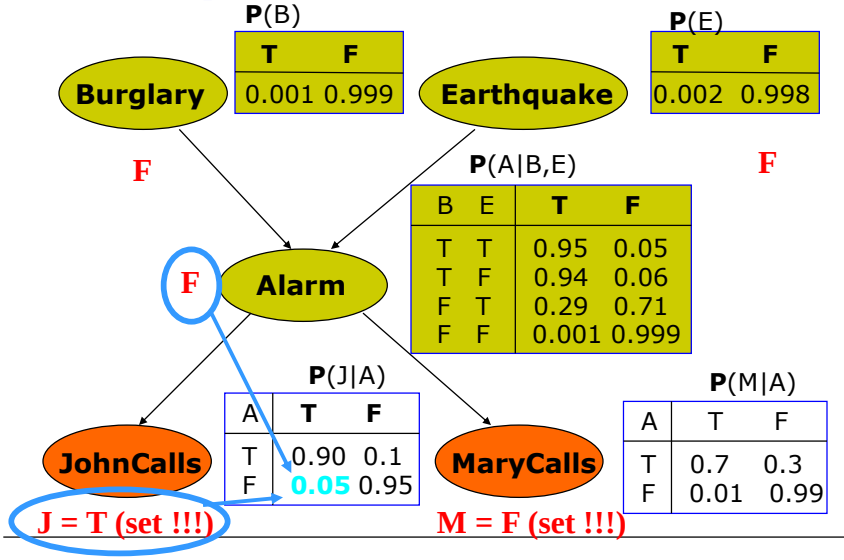
BBN likelihood weighting example

Second sample



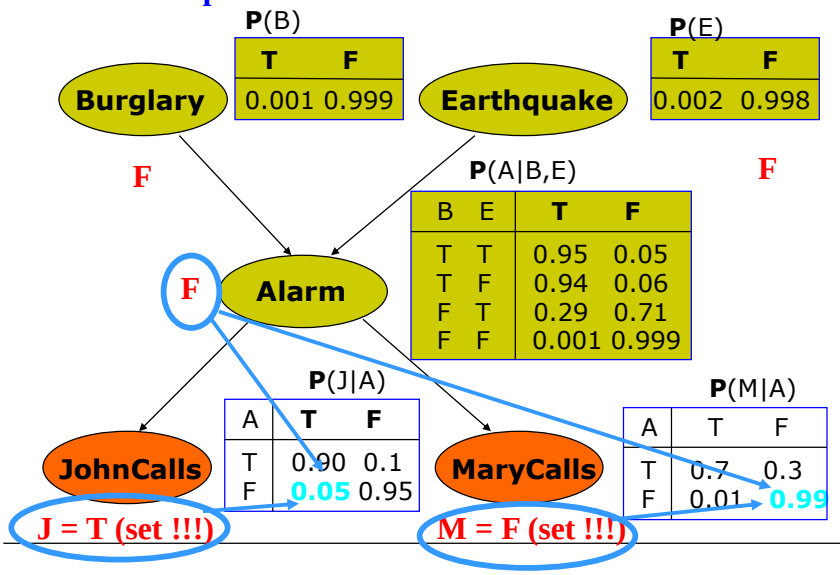
BBN likelihood weighting example

Second sample



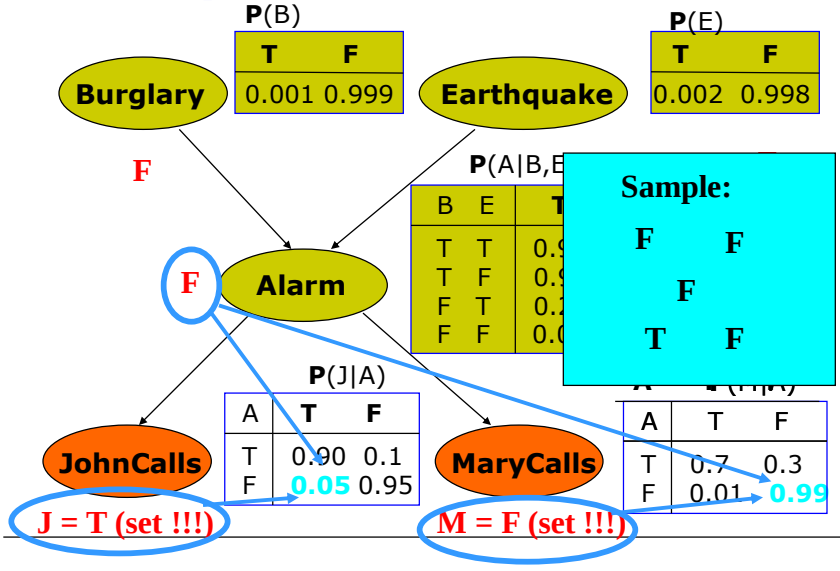
BBN likelihood weighting example

Second sample



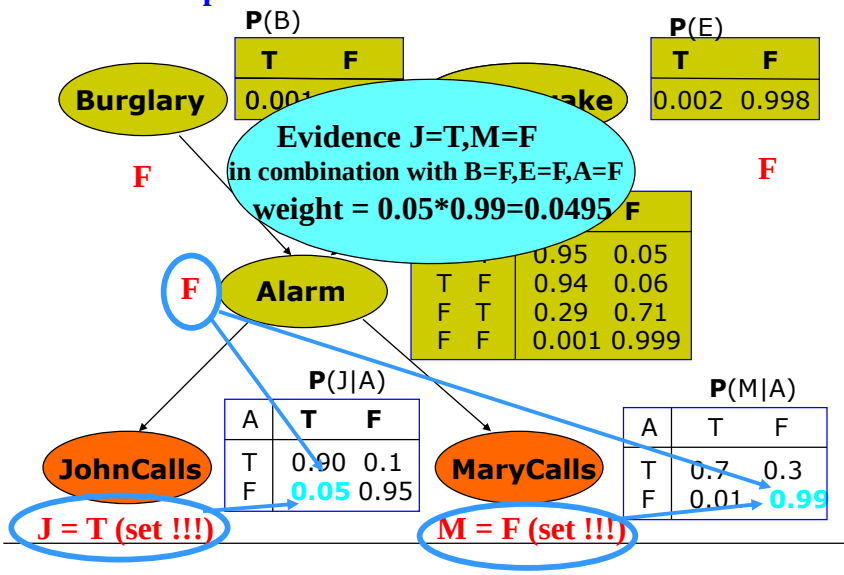
BBN likelihood weighting example

Second sample



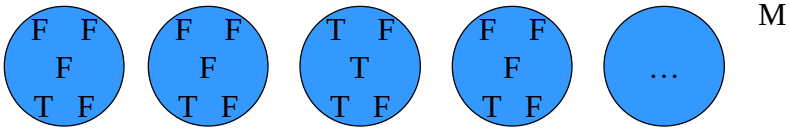
BBN likelihood weighting example

Second sample



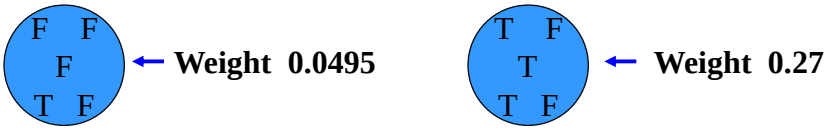
Likelihood weighting

- Assume we have generated the following M samples:



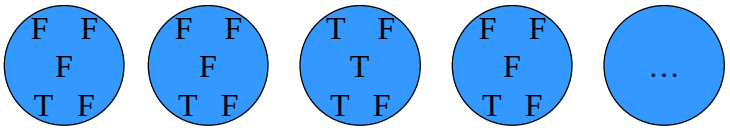
How to make the samples consistent?

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



Likelihood weighting

- Assume we have generated the following M samples:



M

0.0495 **0.0495** **0.27** **0.0495**

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

Expectation Maximization (EM)

Learning probability distribution

Basic learning settings:

- A set of random variables $\mathbf{X} = \{X_1, X_2, \ldots, X_n\}$
- **A model of the distribution** over variables in $\boldsymbol{X}$ with parameters Θ
- **Data** $D = \{D_1, D_2, .., D_N\}$

  s.t. $D_i = (x_1^i, x_2^i, \ldots x_n^i)$

Objective: find parameters $\hat{\Theta}$ that describe the data

Assumptions considered so far:

- Known parameterizations
- No hidden variables
- No-missing values

Hidden variables

Modeling assumption:

Variables $\mathbf{X} = \{X_1, X_2, \ldots, X_n\}$

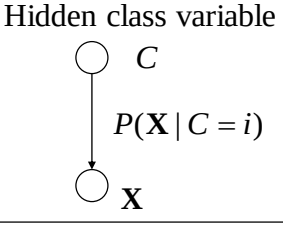
- Additional variables are hidden – never observed in data

Why to add hidden variables?

- **More flexibility in describing the distribution** $P(\mathbf{X})$
- **Smaller parameterization of** $P(\mathbf{X})$
  - **New independences can be introduced via hidden variables**

Example:

- Latent variable models
  - hidden classes (categories)



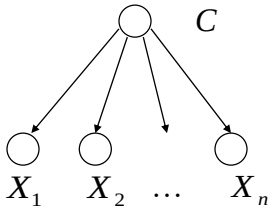
Naïve Bayes with a hidden class variable

Introduction of a hidden variable can reduce the number of parameters defining $P(\mathbf{X})$

Example:

- Naïve Bayes model with a hidden class variable

Hidden class variable



Attributes are independent given the class

- **Useful in customer profiles**
  - Class value = type of customers

Missing values

A set of random variables $\mathbf{X} = \{X_1, X_2, \ldots, X_n\}$

- **Data** $D = \{D_1, D_2, .., D_N\}$
- **But some values are missing**

$D_i = (x_1^i, x_3^i, \ldots x_n^i)$

Missing value of x_2^i

$D_{i+1} = (x_3^{i+1}, \ldots x_n^{i+1})$

Missing values of x_1^{i+1}, x_2^{i+1}

Etc.

- **Example: medical records**
- **We still want to estimate parameters of** $P(\mathbf{X})$

Learning with hidden variables and missing values

Problem: we want to learn $P(\mathbf{X})$ from data D

- **Data consists of examples with values for variables**
- **But sometimes the values are missing**

Data

x_1	x_2			x_5
T	T	F	T	T
T	-	T	F	F
F	T	-	F	T

Missing values

Extended Data

x_1	x_2			x_5	h
T	T	F	T	T	-
T	-	T	F	F	-
F	T	-	F	T	-

Hidden variable h (all values missing)

Density estimation

Goal: Find the set of parameters $\hat{\Theta}$

Estimation criteria:

- **ML** $\max_{\Theta} p(D \mid \boldsymbol{\Theta}, \xi)$
- **Bayesian** $p(\boldsymbol{\Theta} \mid D, \xi)$

Optimization methods for ML: gradient-ascent, conjugate gradient, Newton-Rhapson, etc.

Problem: No or very small advantage from the structure of the corresponding belief network when there are unobserved values

Expectation-maximization (EM) method

- An alternative optimization method
- Suitable when there are missing or hidden values
- **Takes advantage of the structure of the belief network**

General EM

The key idea of a method:

Compute the parameter estimates iteratively by performing the following two steps:

Two steps of the EM:

1. **Expectation step**. For all hidden and missing variables (and their possible value assignments) calculate their expectations for the current set of parameters Θ'
2. **Maximization step**. Compute the new estimates of Θ by considering the expectations of the different value completions

Stop when no improvement possible

General EM

Current parameters Θ'

(1) Expectation step. For all hidden and missing variables (and their possible value assignments) calculate their expectations for the current set of parameters

x_1	x_2			x_5	h
T	T	F	T	T	-
T	-	T	F	F	-
F	T	-	F	T	-

Calculate expectations for all missing value assignments from Θ'

$$P(h^{(1)} = T \mid D^{(1)}, \Theta')$$
$$P(h^{(1)} = F \mid D^{(1)}, \Theta')$$

$$P(x_2^{(2)} = T, h^{(2)} = T \mid D^{(2)}, \Theta')$$
$$P(x_2^{(2)} = T, h^{(2)} = F \mid D^{(2)}, \Theta')$$
$$P(x_2^{(2)} = F, h^{(2)} = T \mid D^{(2)}, \Theta')$$
$$P(x_2^{(2)} = F, h^{(2)} = F \mid D^{(2)}, \Theta')$$

General EM

Current parameters Θ'

(2) Maximization step. Compute the new estimates of parameters Θ by considering the expectations of the different value completions

x_1	x_2			x_5	h
T	T	F	T	T	-
T	-	T	F	F	-
F	T	-	F	T	-

Compute the new estimates of Θ

$$P(h^{(1)} = T \mid D^{(1)}, \Theta')$$
$$P(h^{(1)} = F \mid D^{(1)}, \Theta')$$

$\Rightarrow \Theta$

$$P(x_2^{(2)} = T, h^{(2)} = T \mid D^{(2)}, \Theta')$$
$$P(x_2^{(2)} = T, h^{(2)} = F \mid D^{(2)}, \Theta')$$
$$P(x_2^{(2)} = F, h^{(2)} = T \mid D^{(2)}, \Theta')$$
$$P(x_2^{(2)} = F, h^{(2)} = F \mid D^{(2)}, \Theta')$$