CS 2750 Machine Learning
Lecture 17

Probabilistic graphical models:
• BBN inference
• Markov Random Fields (MRFs)

Milos Hauskrecht
milos@pitt.edu
5329 Sennott Square

Modeling complex distributions

Question: How to model and learn complex multivariate distributions $\hat{p}(\mathbf{X})$ with a large number of variables?

- Represent the full joint distribution over the variables more compactly with a **smaller number of parameters**.
- Take advantage of **conditional and marginal independences** among random variables

Bayesian belief networks (BBNs)

Question: How to model and learn complex multivariate distributions with a large number of variables?

BBNs:

- Represent the full joint distribution over the variables more compactly with a **smaller number of parameters**.
- Take advantage of **conditional and marginal independences** among random variables
- **X and Y are independent** $P(X,Y) = P(X)P(Y)$
- **X and Y are conditionally independent given Z**

$$P(X,Y \mid Z) = P(X \mid Z)P(Y \mid Z)$$

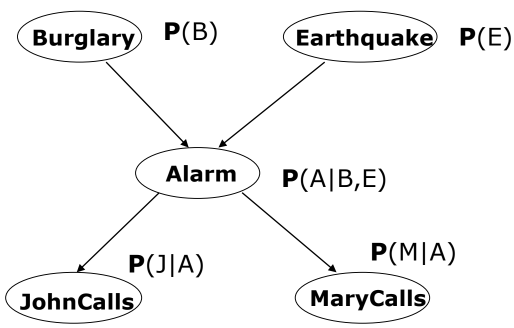
$$P(X \mid Y,Z) = P(X \mid Z)$$

Bayesian belief network

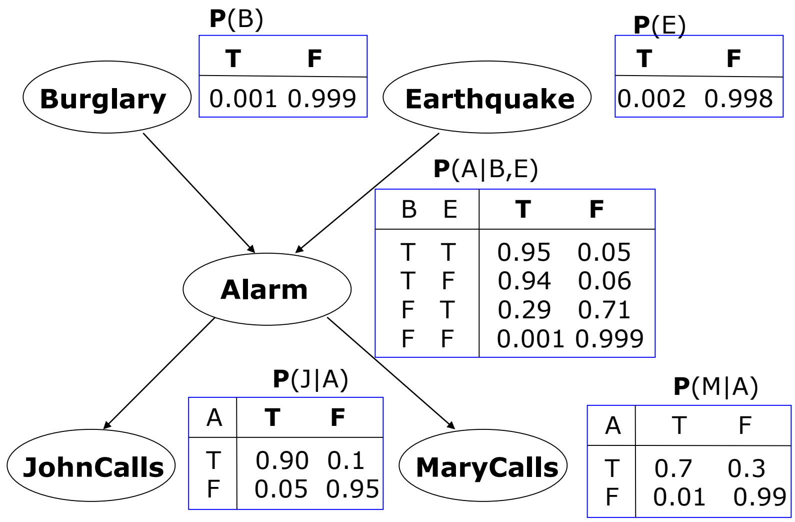
Belief network structure:

- **Nodes** = random variables
  Burglary, Earthquake, Alarm, Mary calls and John calls
- **Links** = direct (causal) dependencies between variables.

  The chance of Alarm being is influenced by Earthquake, The chance of John calling is affected by the Alarm



Bayesian belief network: parameters



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)

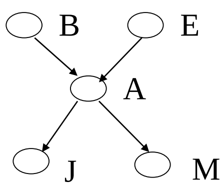
A	T	F
T	0.7	0.3
F	0.01	0.99

Full joint distribution in BBNs

Full joint distribution is defined in terms of local conditional distributions (obtained via the chain rule):

$$\mathbf{P}(X_1, X_2, .., X_n) = \prod_{i=1,..n} \mathbf{P}(X_i \mid pa(X_i))$$

Example:



Assume the following assignment of values to random variables

$B = T, E = T, A = T, J = T, M = F$

Then its probability is:

$$P(B = T, E = T, A = T, J = T, M = F) =$$
$$P(B = T)P(E = T)P(A = T \mid B = T, E = T)P(J = T \mid A = T)P(M = F \mid A = T)$$

Parameter complexity problem

- In the BBN the **full joint distribution** is defined as:

$$\mathbf{P}(X_1, X_2,.., X_n) = \prod_{i=1,..n} \mathbf{P}(X_i \mid pa(X_i))$$

- **What did we save?**

Alarm example: 5 binary (True, False) variables

# of parameters of the full joint:

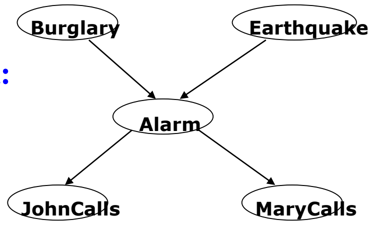
$$2^5 = 32$$

One parameter depends on the rest:

$$2^5 - 1 = 31$$

# of parameters of the BBN:

$$2^3 + 2(2^2) + 2(2) = 20$$

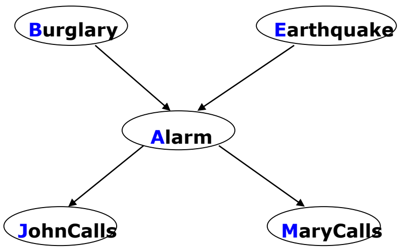


One parameter in every conditional depends on the rest:

$$2^2 + 2(2) + 2(1) = 10$$

Inference in Bayesian network

- **Bad news:**
  - Exact inference problem in BBNs is NP-hard (Cooper)
  - Approximate inference is NP-hard (Dagum, Luby)
- **But** very often we can achieve significant improvements
- Assume our Alarm network



- Assume we want to compute: $P(J = T)$

Inference in Bayesian networks

How to compute sums and products more efficiently?

$$\sum_x af(x) = a\sum_x f(x)$$

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

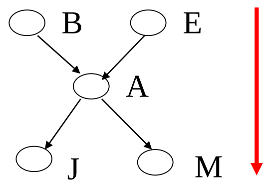
- **MC approximation**:
  - The probability is approximated using sample frequencies
  - **Example:**

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

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

N ← *total # samples*

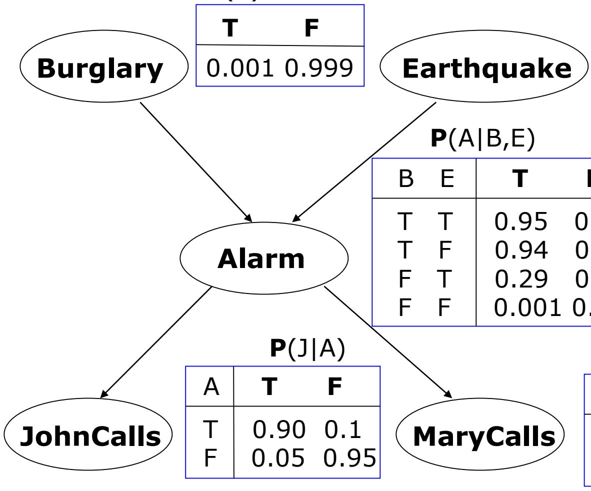
- **Sample generation: BBN sampling of the joint is easy**



Generate sample in a top down manner, following the links

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

BBN sampling example



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

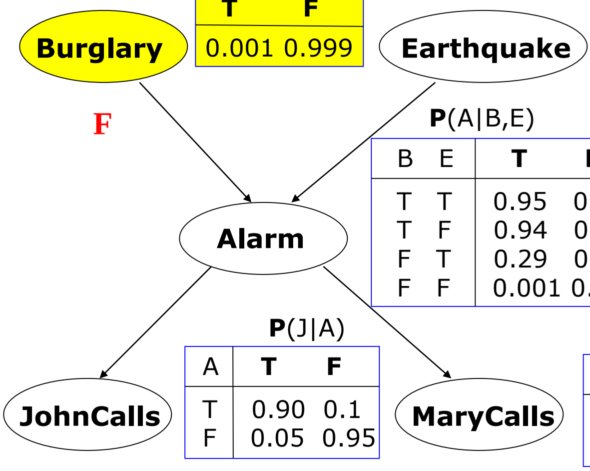
P(J|A)

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

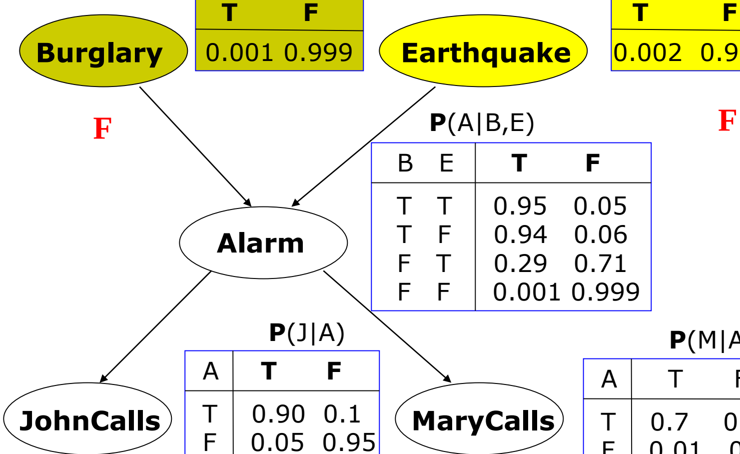
P(J|A)

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

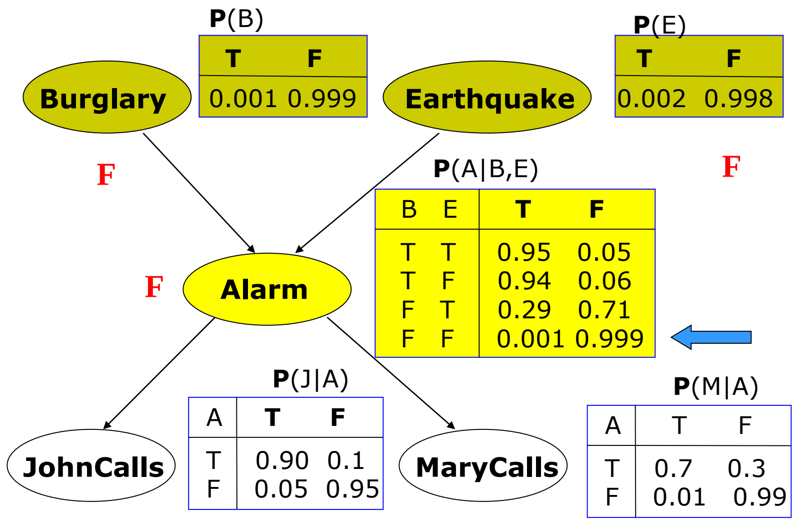
P(J|A)

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

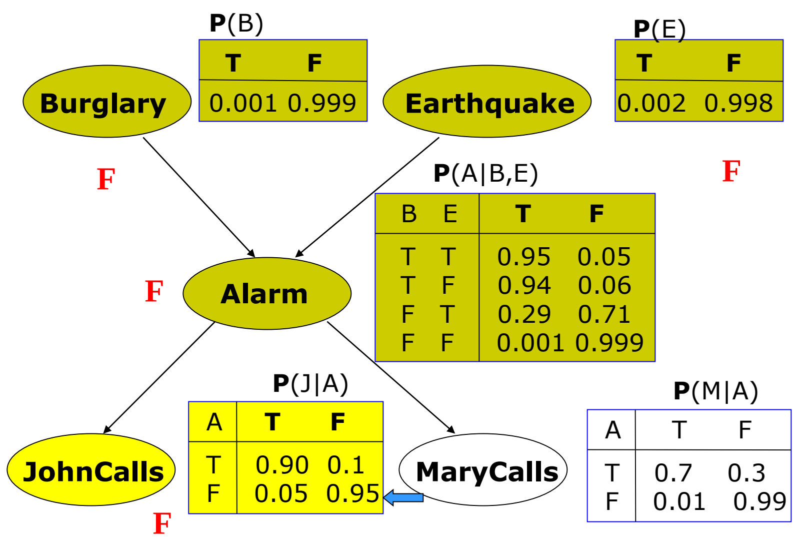
P(J|A)

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

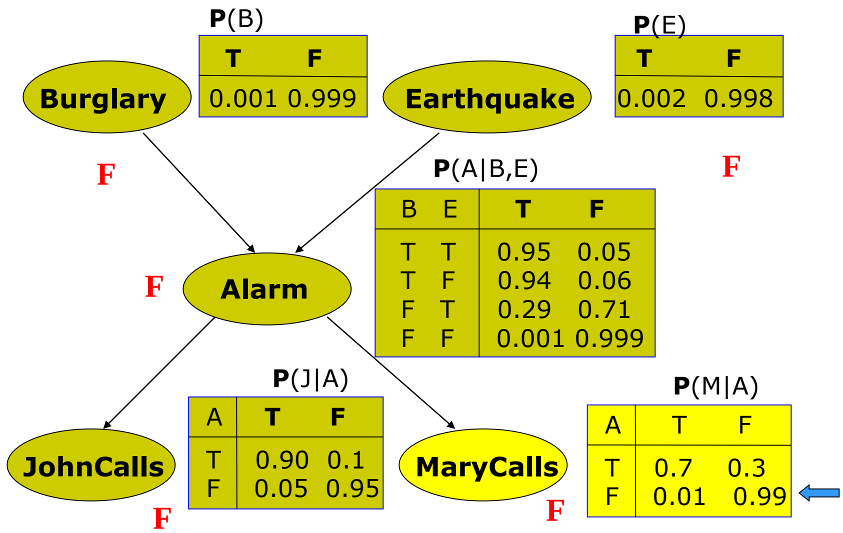
P(J|A)

A	T	F
T	0.90	0.1
F	0.05	0.95

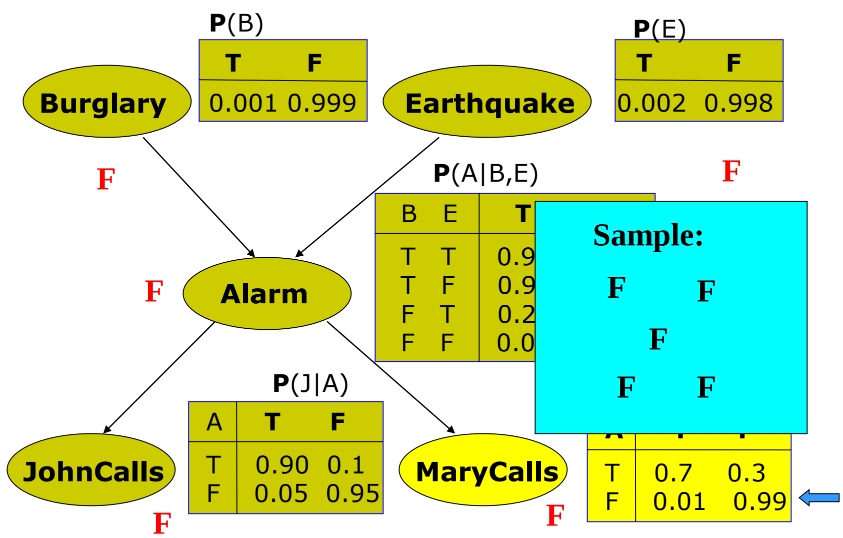
P(M|A)

A	T	F
T	0.7	0.3
F	0.01	0.99

BBN sampling example



BBN sampling example



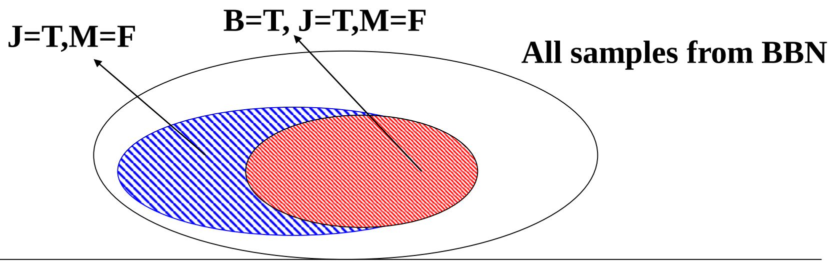
Monte Carlo approaches

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

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

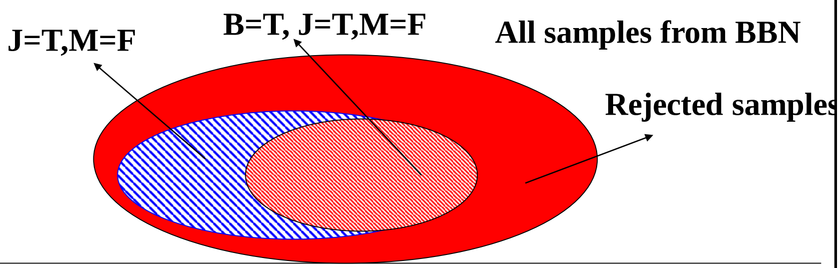
$N_{B=T,J=T,M=F}$: $\#\, samples\ with\ B=T, J=T, M=F$

$N_{J=T,M=F}$: $\#\, samples\ with\ J=T, M=F$



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



Importance sampling

Idea: generate only examples consistent with the evidence

- Avoids inefficiencies of rejection sampling

Problem:

- the distribution generated by enforcing the evidence is biased
- simple counts are not sufficient to estimate the probabilities

Solution: importance sampling

- Generate examples from the (sampling) distribution that is different from the target distribution.
- Give examples from the sample distribution a weight that reflects the consistency between the two distributions

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

Importance sampling

Solution: importance sampling /likelihood weighting

- Generate examples from the (sampling) distribution that is different from the target distribution.
- Give examples from the sample distribution a weight that reflects the consistency between the two distributions

Estimate based on the target distribution:

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

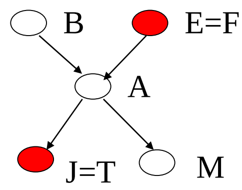
Estimate based on the sampling distribution:

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

Likelihood weighting

Consider the following evidence:

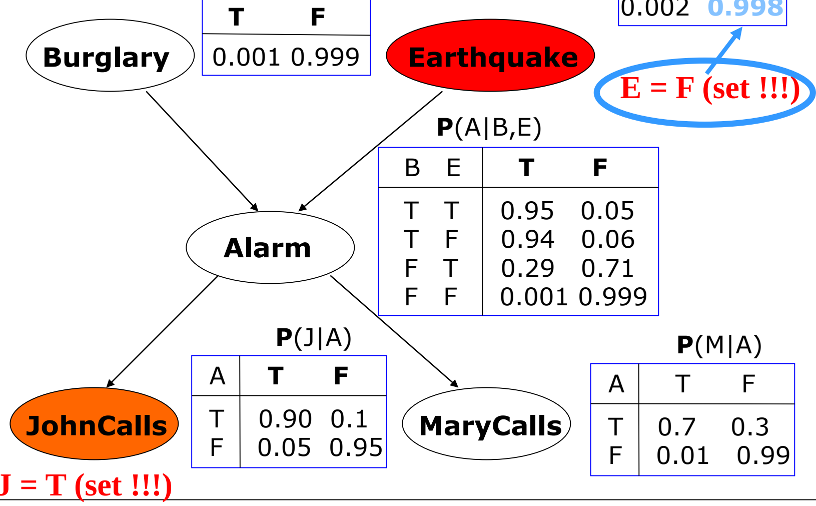
E=F and **J=T in the Alarm network**



Two questions:

- How to generate examples consistent with the evidence?
- How to de-bias (correct) the sample with a weight?

BBN likelihood weighting example



P(E)

T	F
0.002	0.998

P(B)

T	F
0.001	0.999

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

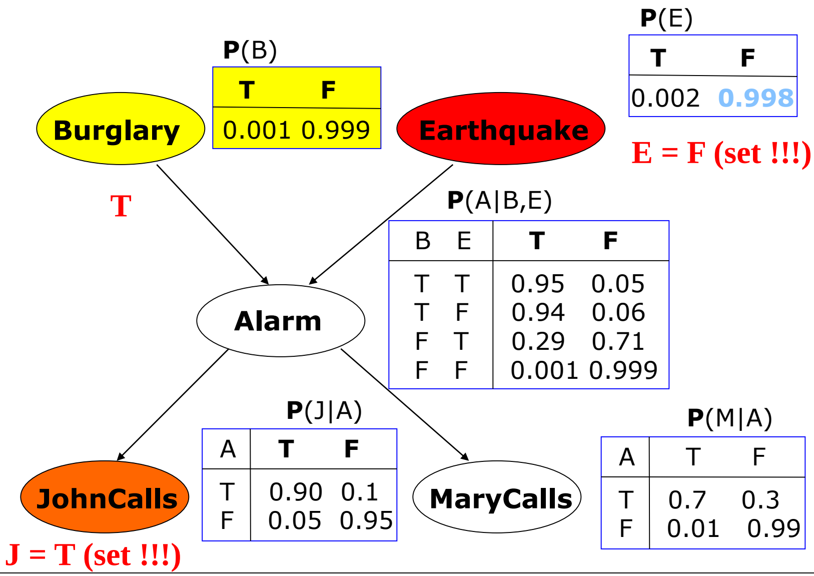
P(J|A)

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

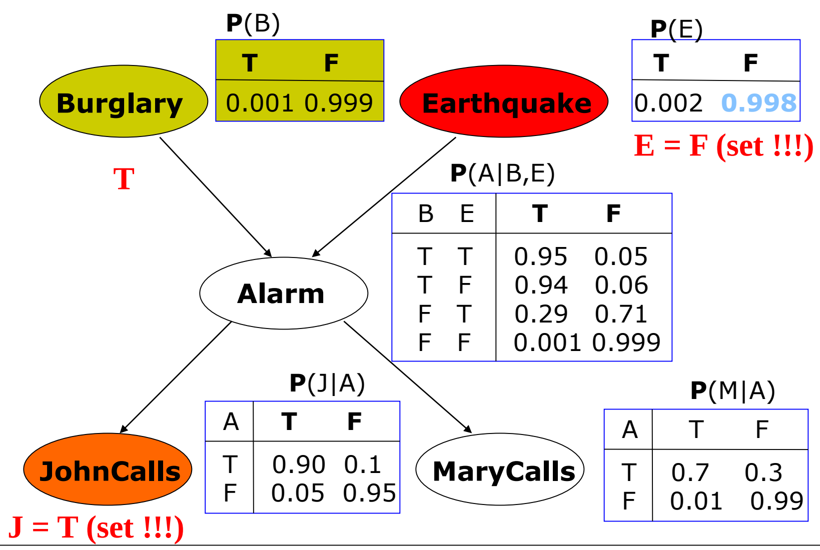
P(J|A)

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

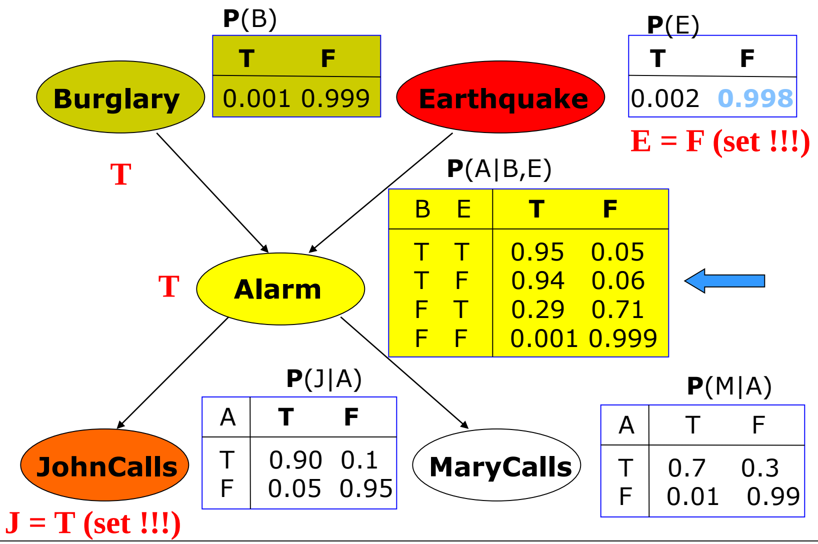
P(J|A)

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

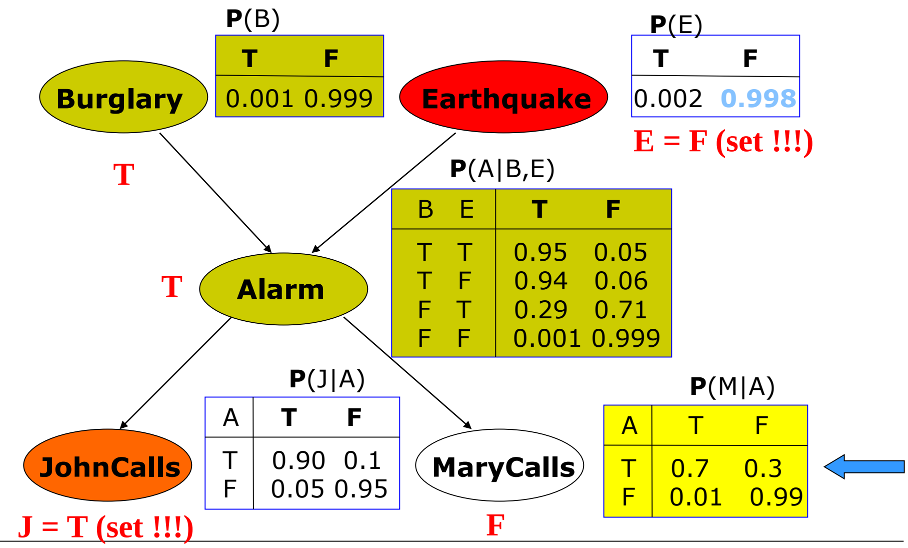
P(J|A)

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

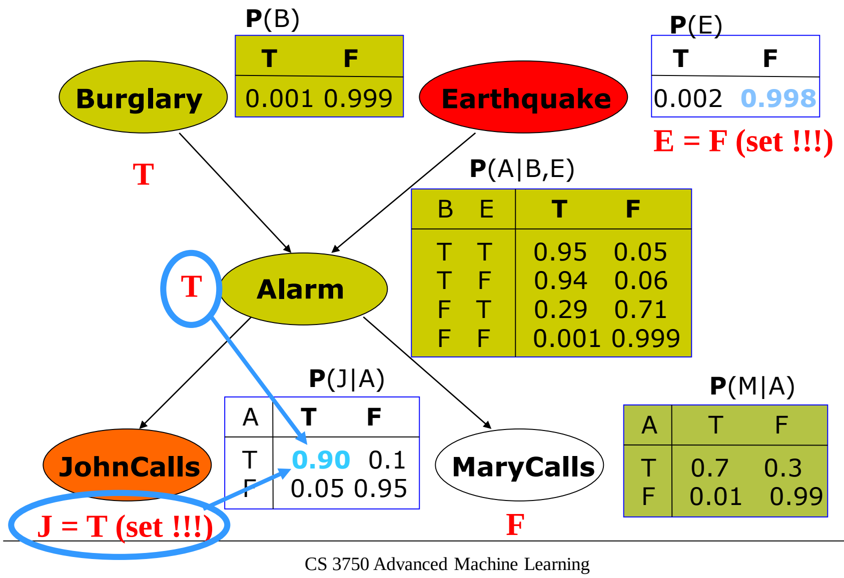
P(J|A)

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

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

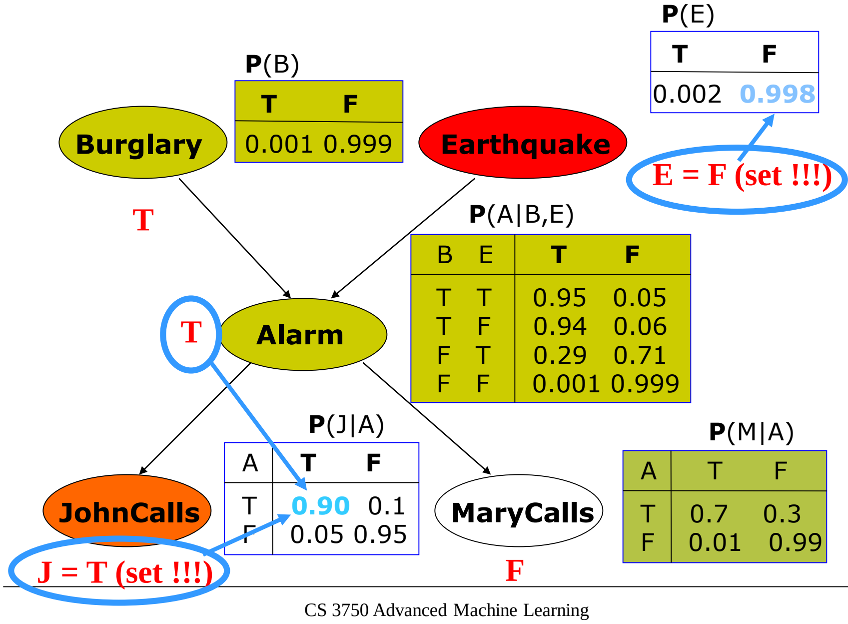
P(J|A)

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

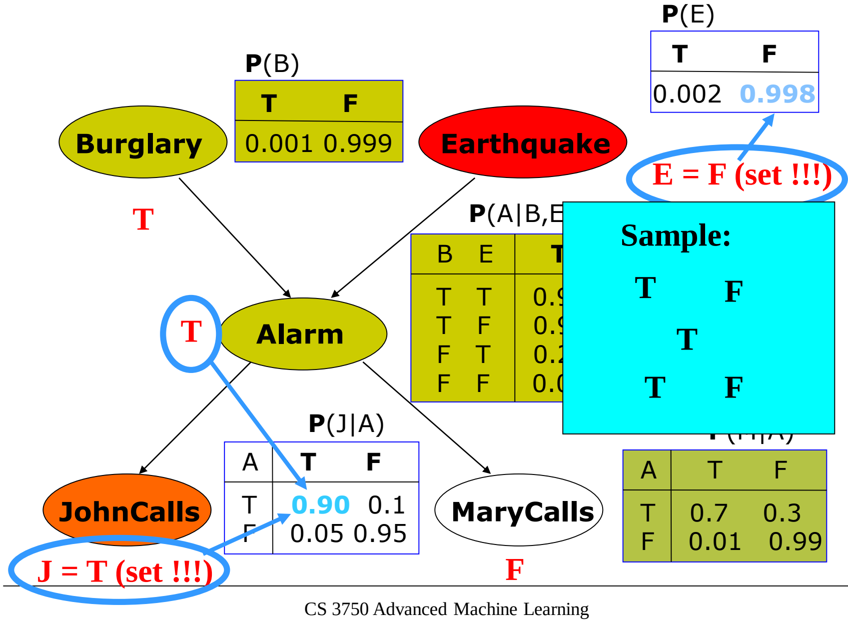
P(J|A)

A	T	F
T	0.90	0.1
F	0.05	0.95

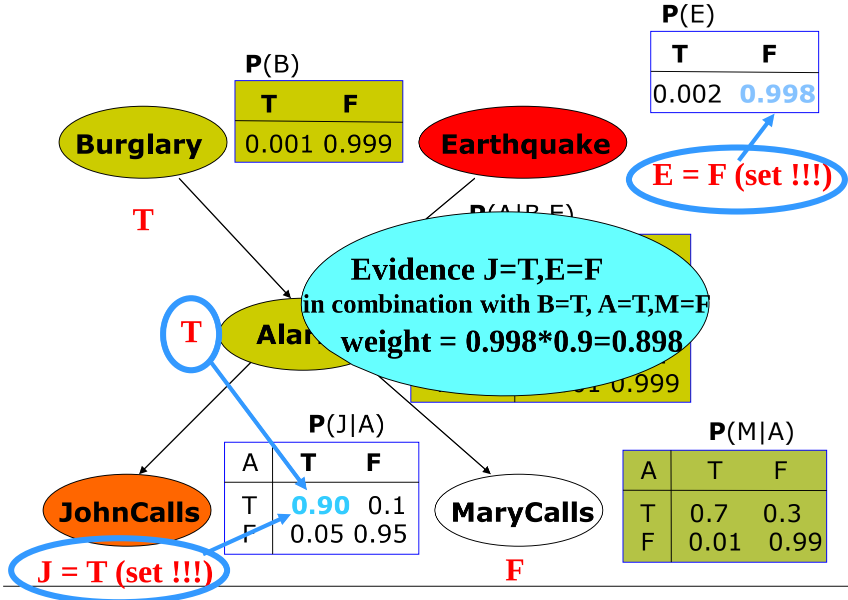
P(M|A)

A	T	F
T	0.7	0.3
F	0.01	0.99

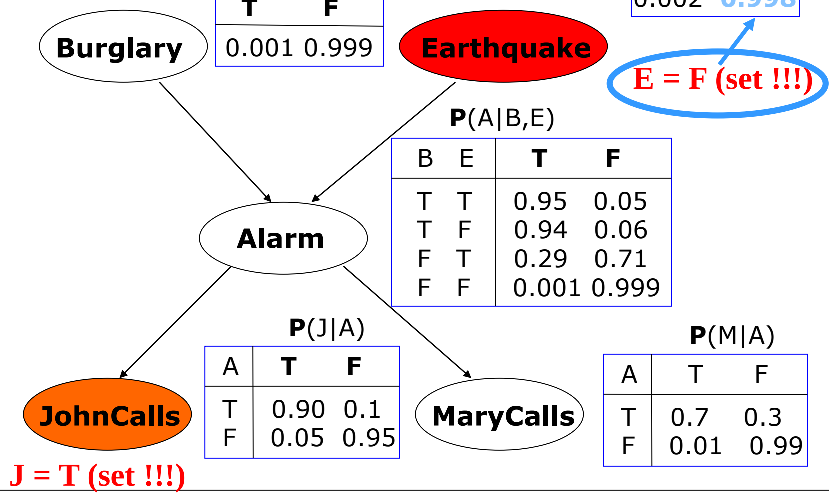
BBN likelihood weighting example



BBN likelihood weighting example



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)

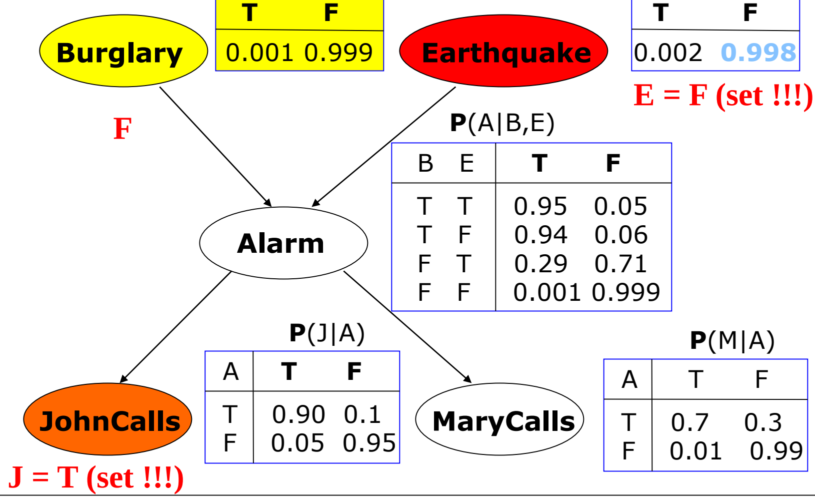
A	T	F
T	0.90	0.1
F	0.05	0.95

P(M|A)

A	T	F
T	0.7	0.3
F	0.01	0.99

BBN likelihood weighting example

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)

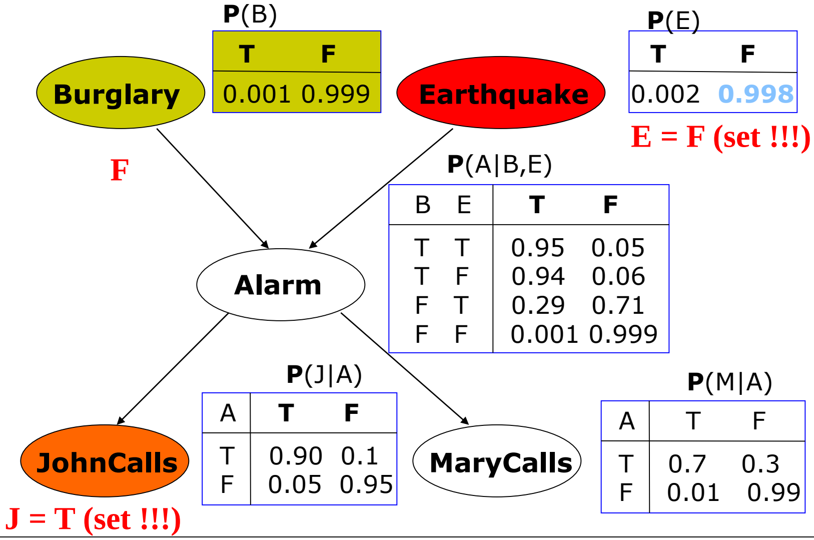
A	T	F
T	0.90	0.1
F	0.05	0.95

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)

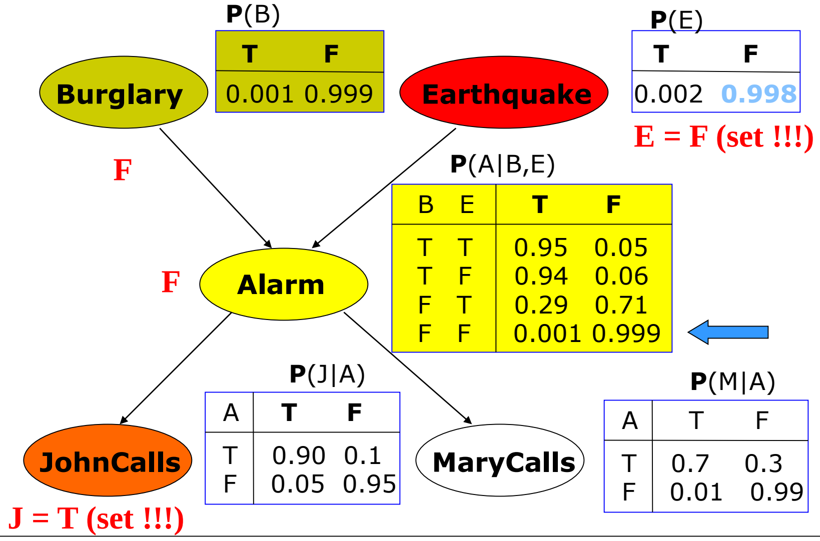
A	T	F
T	0.90	0.1
F	0.05	0.95

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)

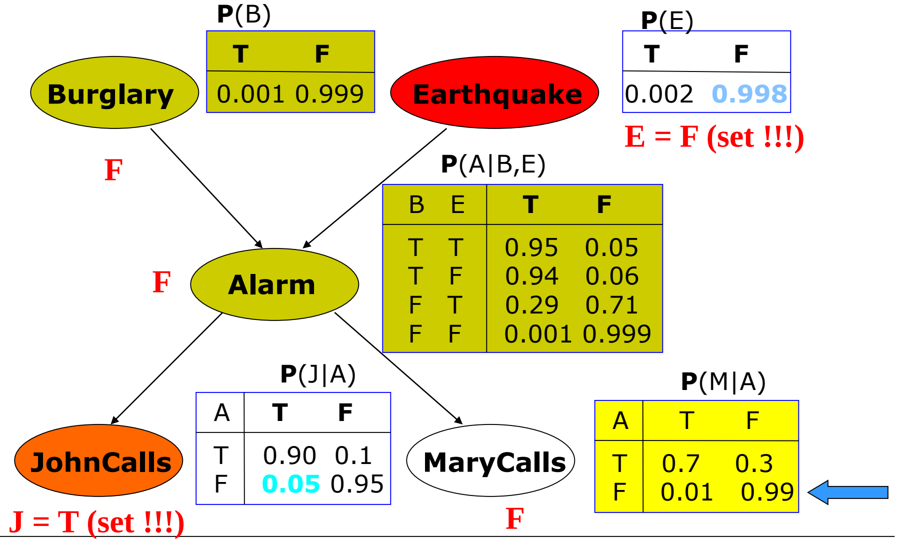
A	T	F
T	0.90	0.1
F	0.05	0.95

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)

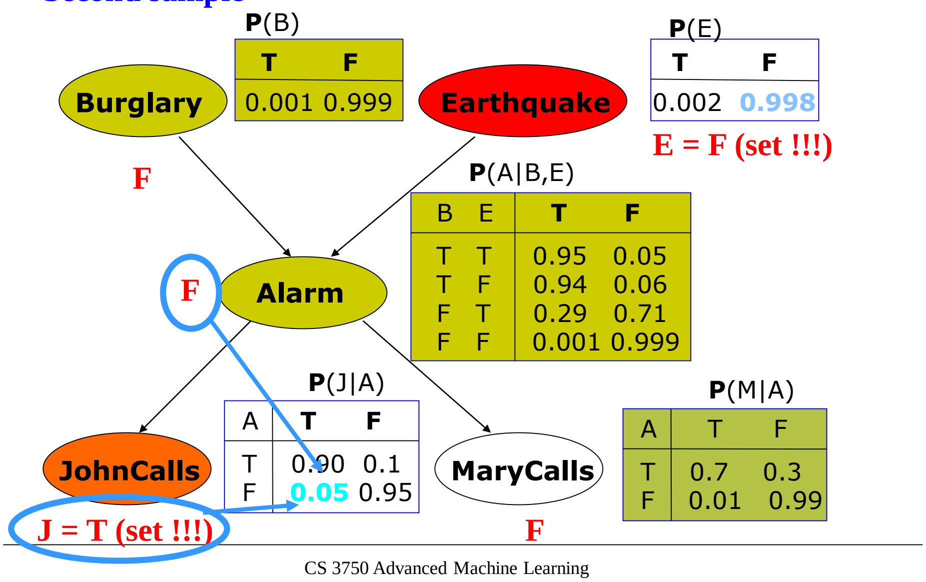
A	T	F
T	0.90	0.1
F	0.05	0.95

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)

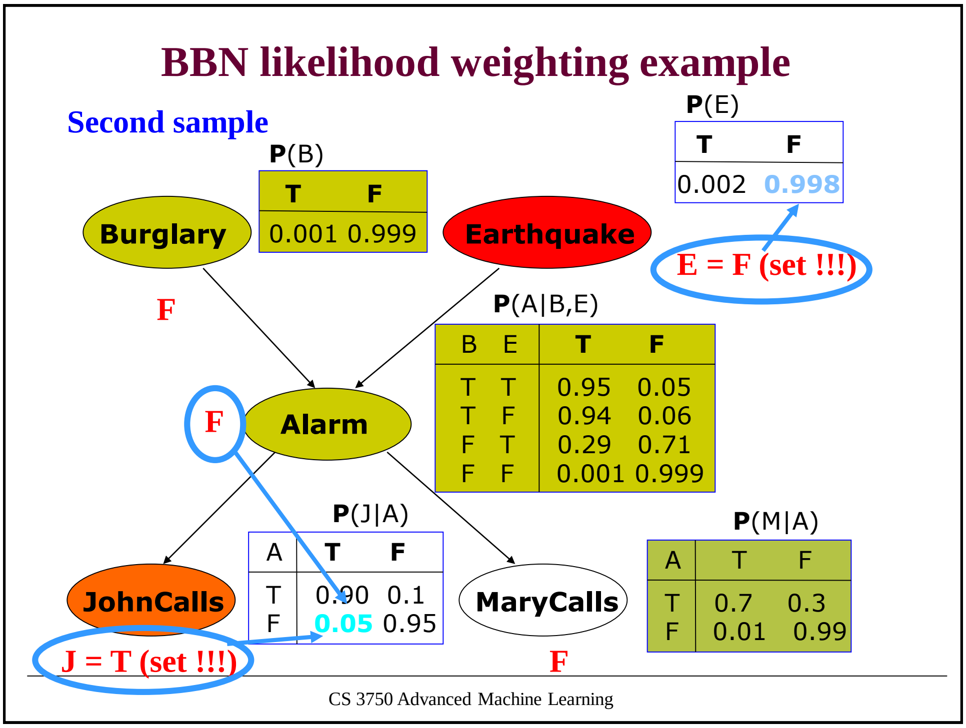
B	E	T	F
T	T	0.95	0.05
T	F	0.94	0.06
F	T	0.29	0.71
F	F	0.001	0.999

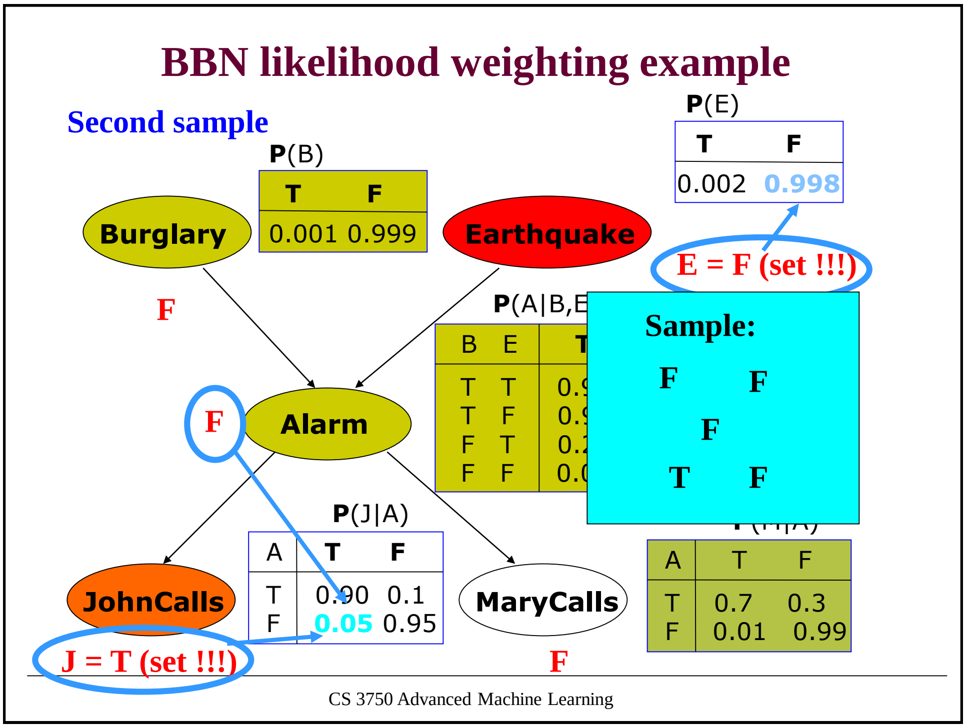
P(J|A)

A	T	F
T	0.90	0.1
F	0.05	0.95

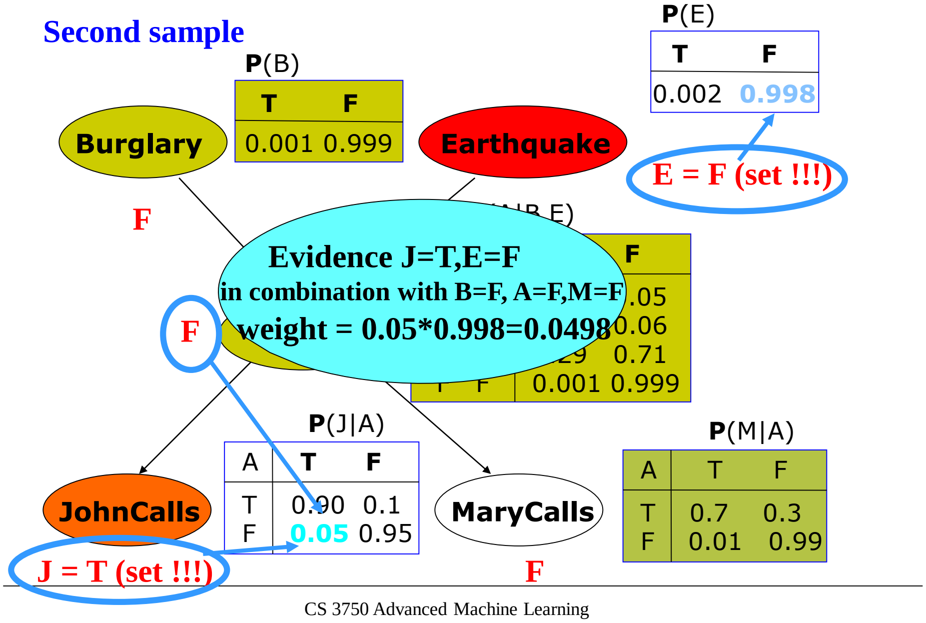
P(M|A)

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
Burglary
Earthquake
E = F (set !!!)
F
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
F
Alarm
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
JohnCalls
MaryCalls
J = T (set !!!)
F
CS 3750 Advanced Machine Learning

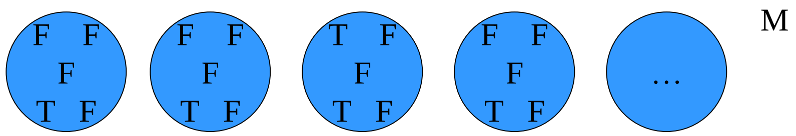

BBN likelihood weighting example
Second sample
P(B)
T F
0.001 0.999
P(E)
T F
0.002 0.998
Burglary
Earthquake
E = F (set !!!)
F
P(A|B,E
B E
T T
T F
F T
F F
Sample:
F F
F
T F
F
Alarm
P(J|A)
A T F
T 0.90 0.1
F 0.05 0.95
A T F
T 0.7 0.3
F 0.01 0.99
JohnCalls
MaryCalls
J = T (set !!!)
F
CS 3750 Advanced Machine Learning

BBN likelihood weighting example



Likelihood weighting

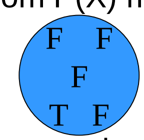
- Assume we have generated the following M samples:



- If we calculate the estimate:

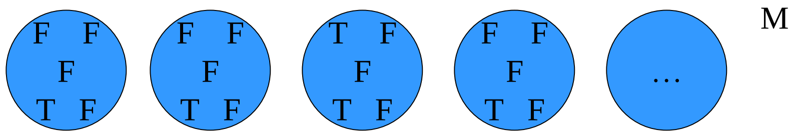
$$P(B=T \mid J=T, E=F) = \frac{\#sample\_with(B=T)}{\#total\_sample}$$

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).

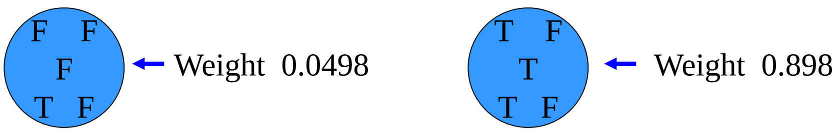
Likelihood weighting

- 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).



Likelihood weighting

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

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

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

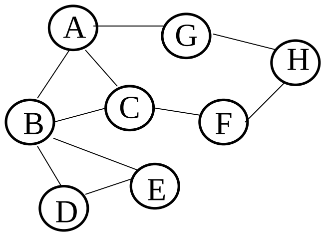
Markov random fields

An undirected network (also called independence graph)

- **Probabilistic models with symmetric dependences**
- G = (S, E)
  - S set of random variables
  - Undirected edges E that define dependences between pairs of variables

Example:

variables A,B ..H



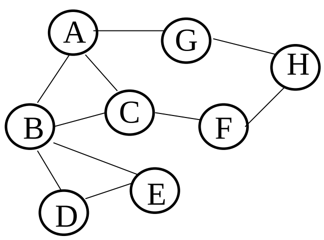
Markov random fields

The full joint of the MRF is defined

$$P(\mathbf{x}) \propto \prod_{c \in cl(x)} \phi_c(\mathbf{x}_c)$$

$\phi_c(x_c)$ - A potential function (defined over variables in cliques/factors)

Example:



Full joint:

$$P(A,B,\ldots H) \sim \phi_1(A,B,C)\phi_2(B,D,E)\phi_3(A,G)\phi_4(C,F)\phi_5(G,H)\phi_6(F,H)$$

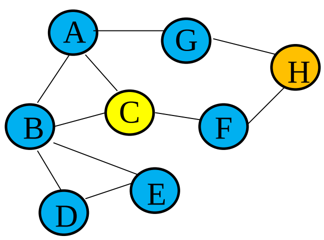
$\phi_c(x_c)$ - A potential function (defined over a clique of the graph)

Markov random fields: independence relations

- **Pairwise Markov property**
  - Two nodes in the network that are not directly connected can be made independent given all other nodes
- **Local Markov property**
  - A set of nodes (variables) can be made independent from the rest of nodes variables given its immediate neighbors
- **Global Markov property**
  - A vertex set A is independent of the vertex set B (A and B are disjoint) given set C if all chains in between elements in A and B intersect C

Markov random fields: independence relations

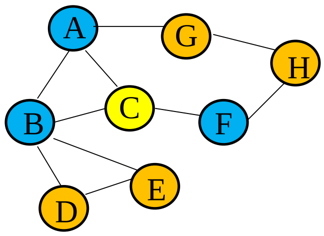
- **Pairwise Markov property**
  - Two nodes in the network that are not directly connected can be made independent given all other nodes
- **Example:**



C and H are independent given the rest of the nodes

Markov random fields: independence relations

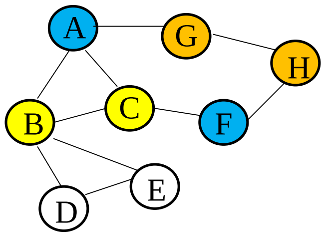
- **Local Markov property**
  - A set of nodes (variables) can be made independent from the rest of nodes variables given its immediate neighbors
- **Example:**



C is independent of {G,H,D,E} given the neighbors of C that is, variables {A,B,F}

Markov random fields: independence relations

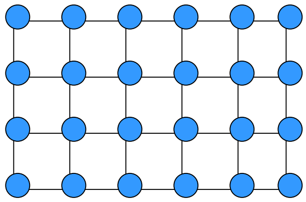
- **Global Markov property**
  - A vertex set A is independent of the vertex set B (A and B are disjoint) given set C if all chains in between elements in A and B intersect C
- **Example:**



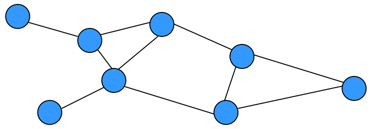
A set {B, C} is independent of {G,H} given the set{A,F}

Markov random fields

- regular lattice (Ising model)

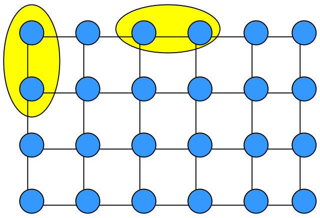


- Arbitrary graph

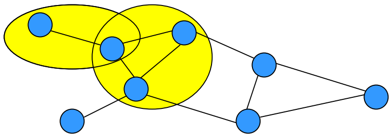


Markov random fields

- regular lattice (Ising model)



- Arbitrary graph



Markov random fields

- **Joint probability**

$$P(x) \approx \prod_{c \in cl(x)} \phi_c(x_c)$$

$\phi_c(x_c)$ - A potential function (defined over cliques/factors)

- **Typical condition on potential functions:**
  - If $\phi_c(x_c)$ is strictly positive we can rewrite the definition in terms of a log-linear model :

$$P(x) = \frac{1}{Z} \exp\left(- \sum_{c \in cl(x)} E_c(x_c) \right)$$

Energy function

- Gibbs (Boltzman) distribution

$$Z = \sum_{x \in \{x\}} \exp\left(- \sum_{c \in cl(x)} E_c(x_c) \right)$$

- A partition function

Are BBNs and MRFs different?

Both models represent independences that hold among variables or sets of variables?

- Are the two the same in terms of independences they can represent?
- Or, are they different?

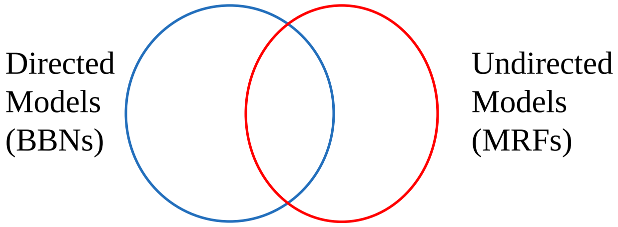
Are BBNs and MRFs different?

Both models represent independences that hold among variables or sets of variables?

- Are the two the same in terms of independences they can represent?
- Or, are they different?

Answer: MRFs are different from BBNs

- There are independences that can be represented by one model but not the other

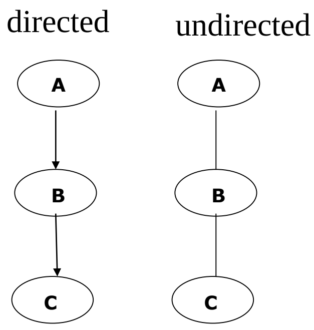


Are BBNs and MRFs different?

MRFs are different from BBNs

- There are independences that can be represented by one model but not the other

Analysis:



A is independent of C given B

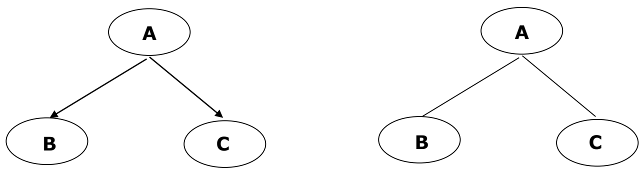
Are BBNs and MRFs different?

MRFs are different from BBNs

- There are independences that can be represented by one model but not the other

Analysis:

directed undirected



B is independent of C given A

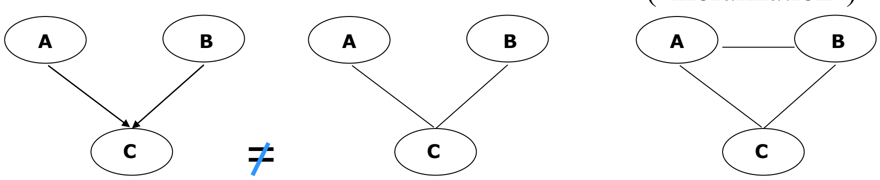
Are BBNs and MRFs different?

MRFs are different from BBNs

- There are independences that can be represented by one model but not the other

Analysis:

directed undirected Fix to undirected ("moralization")



A and B are marginally independent

A and B are independent given C

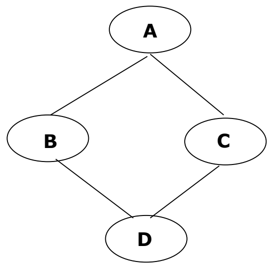
A, B, C are all dependent
No false independence

Are BBNs and MRFs different?

MRFs are different from BBNs

- There are independences that can be represented by one model but not the other

Analysis: undirected



No directed graph can represent the same set of independences

B and C are independent given A,D

A and D are independent given B,C

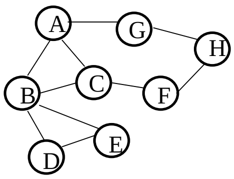
Markov random fields: inference

The full joint of the MRF is defined

$$P(\mathbf{x}) \propto \prod_{c \in cl(x)} \phi_c(\mathbf{x}_c)$$

$\phi_c(x_c)$ - A potential function (defined over variables in cliques/factors)

Example:



Full joint:

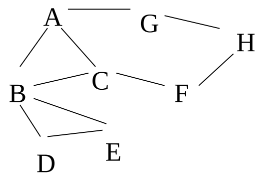
$$P(A, B, ...H) \sim \phi_1(A, B, C)\phi_2(B, D, E)\phi_3(A, G)\phi_4(C, F)\phi_5(G, H)\phi_6(F, H)$$

How to calculate probabilistic queries, such as *P(B)* ?

Next: Variable elimination

MRF variable elimination inference

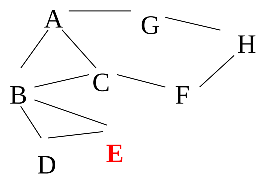
Example:



$$P(B) = \sum_{A,C,D,..H} P(A,B,...H)$$

$$= \frac{1}{Z} \sum_{A,C,D,..H} \phi_1(A,B,C)\phi_2(B,D,E)\phi_3(A,G)\phi_4(C,F)\phi_5(G,H)\phi_6(F,H)$$

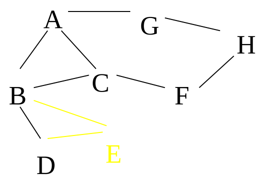
Eliminate E



$$= \frac{1}{Z} \sum_{A,C,D,F,G,H} \phi_1(A,B,C) \underbrace{\left[\sum_E \phi_2(B,D,E)\right]}_{\tau_1(B,D)} \phi_3(A,G)\phi_4(C,F)\phi_5(G,H)\phi_6(F,H$$

MRF variable elimination inference

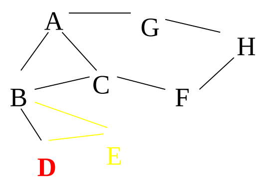
Example (cont):



$$P(B) = \sum_{A,C,D,..H} P(A,B,...H)$$

$$= \frac{1}{Z} \sum_{A,C,D,F,G,H} \phi_1(A,B,C)\tau_1(B,D)\phi_3(A,G)\phi_4(C,F)\phi_5(G,H)\phi_6(F,H)$$

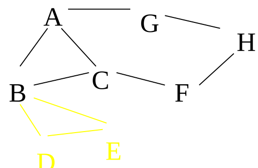
Eliminate D



$$= \frac{1}{Z} \sum_{A,C,F,G,H} \phi_1(A,B,C) \underbrace{\left[\sum_D \tau_1(B,D)\right]}_{\tau_2(B)} \phi_3(A,G)\phi_4(C,F)\phi_5(G,H)\phi_6(F,H)$$

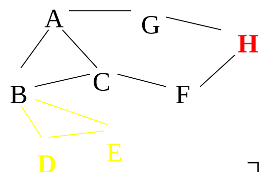
MRF variable elimination inference

Example (cont):



$$P(B) = \sum_{A,C,D,..H} P(A,B,...H)$$

$$= \frac{1}{Z} \sum_{A,C,F,G,H} \phi_1(A,B,C)\tau_2(B)\phi_3(A,G)\phi_4(C,F)\phi_5(G,H)\phi_6(F,H)$$

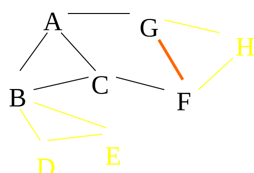


Eliminate H

$$= \frac{1}{Z} \sum_{A,C,F,G} \phi_1(A,B,C)\tau_2(B)\phi_3(A,G)\phi_4(C,F) \underbrace{\left[\sum_{H} \underbrace{\phi_5(G,H)\phi_6(F,H)}_{\tau_3(F,G,H)} \right]}_{\tau_4(F,G)}$$

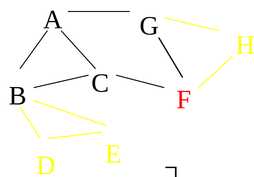
MRF variable elimination inference

Example (cont):



$$P(B) = \sum_{A,C,D,..H} P(A,B,...H)$$

$$= \frac{1}{Z} \sum_{..,C,F,G} \phi_1(A,B,C)\tau_2(B)\phi_3(A,G)\phi_4(C,F)\tau_4(F,G)$$



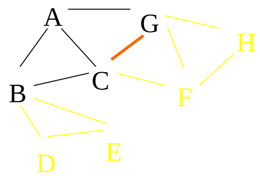
Eliminate F

$$= \frac{1}{Z} \sum_{A,C,G} \phi_1(A,B,C)\tau_2(B)\phi_3(A,G) \underbrace{\left[\sum_{F} \underbrace{\phi_4(C,F)\tau_4(F,G)}_{\tau_5(C,F,G)} \right]}_{\tau_6(G,C)}$$

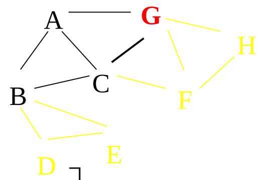
MRF variable elimination inference

Example (cont):

$$P(B) = \sum_{A,C,D,..H} P(A,B,...H)$$



$$= \frac{1}{Z} \phi_1(A,B,C)\tau_2(B)\phi_3(A,G)\tau_6(C,G)$$



Eliminate G

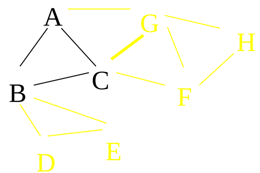
$$= \frac{1}{Z} \sum_{A,C} \phi_1(A,B,C)\tau_2(B)\left[\sum_{F} \underbrace{\phi_3(A,G)\tau_6(C,G)}_{\tau_7(A,C,G)}\right]$$

$$\tau_7(A,C,G) \rightarrow \tau_8(A,C)$$

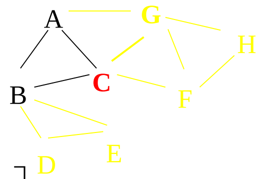
MRF variable elimination inference

Example (cont):

$$P(B) = \sum_{A,C,D,..H} P(A,B,...H)$$



$$= \frac{1}{Z} \sum_{A,C} \phi_1(A,B,C)\tau_2(B)\tau_8(A,C)$$



Eliminate C

$$= \frac{1}{Z} \sum_{A} \tau_2(B)\left[\sum_{C} \underbrace{\phi_1(A,B,C)\tau_8(A,C)}_{\tau_9(A,B,C)}\right]$$

$$\tau_9(A,B,C) \rightarrow \tau_{10}(A,B)$$

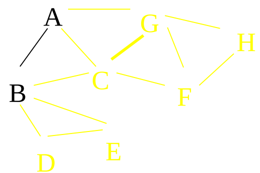
MRF variable elimination inference

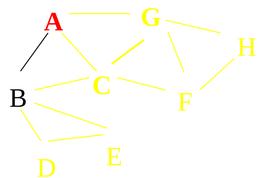
Example (cont):

$$P(B) = \sum_{A,C,D,..} P(A,B,...H)$$

$$= \frac{1}{Z} \sum_{A} \tau_2(B)\tau_{10}(A,B)$$

$$= \frac{1}{Z} \tau_2(B) \sum_{A} \tau_{10}(A,B)$$





Eliminate A

$$= \frac{1}{Z} \tau_2(B) \underbrace{\sum_{A} \tau_{10}(A,B)}_{\tau_{11}(B)}$$

$$= \frac{1}{Z} \tau_2(B)\tau_{11}(B)$$