Bayesian Network

Il-Chul Moon
Dept. of Industrial and Systems Engineering
KAIST
icmoon@kaist.ac.kr

Weekly Objectives

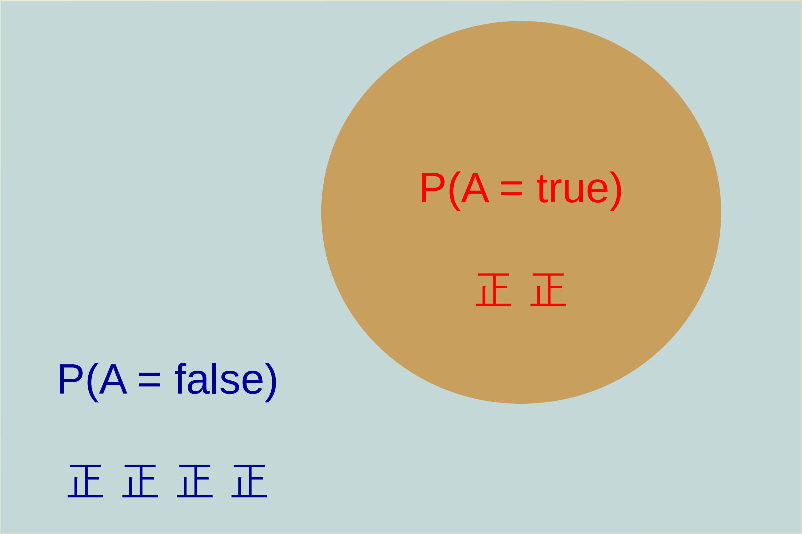
- Know and memorize the theorems of probability
  - Recover the probability concepts
  - Recover the probability theorems
  - Recover the concepts of the marginal and the conditional independencies
- Understand Bayesian networks
  - Know the syntax and the semantics of Bayesian networks
  - Know how to factorize Bayesian networks
  - Able to calculate a probability with given conditions
- Understand the inference of Bayesian networks
  - Able to calculate parameters of Bayesian networks
  - Able to list the exact inference of Bayesian networks

RECAP ON PROBABILITY

Probabilities

- We will write P(A = true) to mean the probability that A = true.
- What is probability?
  - It is the relative frequency with which an outcome would be obtained if the process were repeated a large number of times under similar conditions*

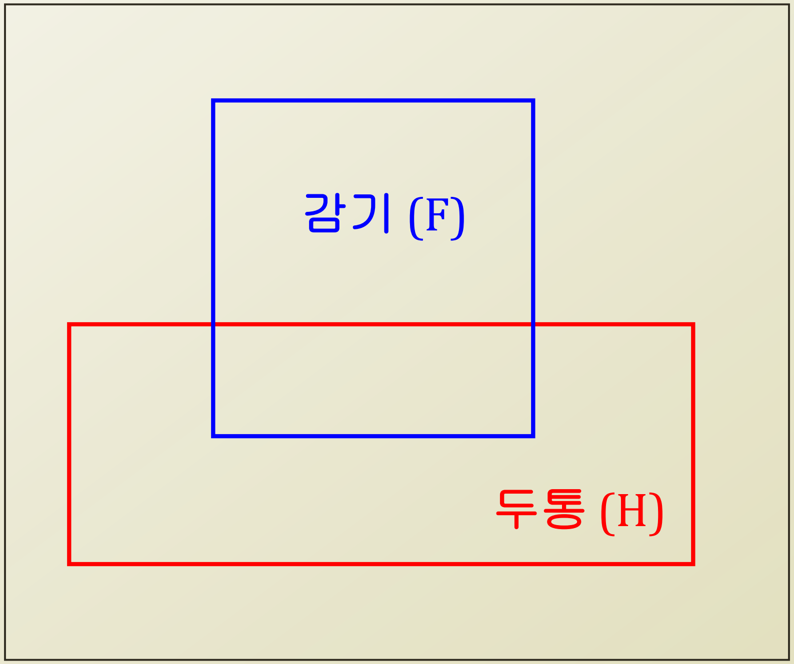
The sum of the red and blue areas is 1



*Ahem…there's also the Bayesian definition which says probability is your degree of belief in an outcome

Conditional Probability

- P(A = *true* | B = *true*)
  - Out of all the outcomes in which B is true, how many also have A equal to true
- Read this as:
  - "Probability of A conditioned on B" or "Probability of A given B"



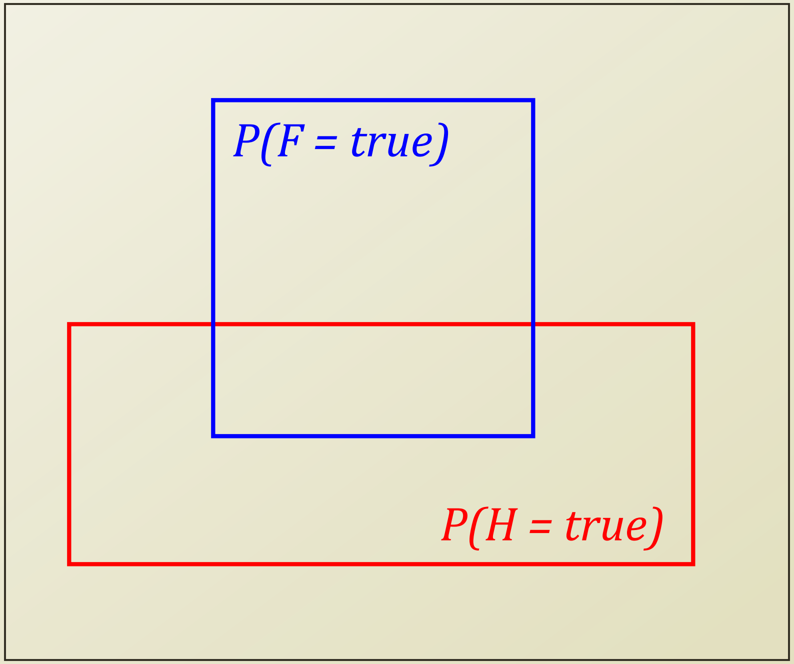
H = "Have a headache"
F = "Coming down with Flu"

P(H = *true*) = 1/10
P(F = *true*) = 1/40
P(H = *true* | F = *true*) = 1/2

"Headaches are rare and flu is rarer, but if you're coming down with flu there's a 50-50 chance you'll have a headache."

Joint Probability

- We will write *P(A = true, B = true)* to mean
  - the probability of *A = true* **and** *B = true*



$$P(H=true \mid F=true)$$
$$= \frac{\text{Area of "H and F" region}}{\text{Area of "F" region}}$$
$$= \frac{P(H=true, F=true)}{P(F=true)}$$

In general, $P(X|Y)=P(X,Y)/P(Y)$

Computing with Probabilities: Law of Total Probability

- Law of Total Probability
  - a.k.a "summing out" or marginalization
  - $P(a) = \sum_b P(a, b) = \sum_b P(a \mid b) P(b)$
  - When B is any random variable
- Consider this case
  - given a joint distribution (e.g., P(a,b,c,d))
  - We can obtain any "marginal" probability (e.g., P(b)) by summing out the other variables
  - $P(b) = \sum_a \sum_c \sum_d P(a, b, c, d)$
- Also, consider this case
  - given a joint distribution (e.g., P(a,b,c,d))
  - We can obtain any conditional probability of interest
  - $P(c \mid b) = \sum_a \sum_d P(a, c, d \mid b) = 1/P(b) \sum_a \sum_d P(a, c, d, b)$
  - Where 1 / P(b) is just a normalization constant
- **Joint distribution contains the information we need to compute any probability of interest.**

Computing with Probabilities: Chain Rule or Factorization

- We can always write
  - P(a, b, c, … z) = P(a | b, c, …. z) P(b, c, … z)
  - by definition of joint probability
- Repeatedly applying this idea, we can write
  - P(a, b, c, … z) = P(a | b, c, …. z) P(b | c,.. z) P(c| .. z)..P(z)
- This factorization holds for any ordering of the variables
  - Chain rule for probabilities
  - Any joint probability → Can be factorized into a series of multiplication

Joint Probability Distribution

- Joint probabilities can be between any number of variables
  - *P(Int = true, Effort = true, GPA = true)*
- For each combination of variables, we need to say how probable that combination is
- The probabilities of these combinations need to sum to 1

Int.	Effort	GPA	P(I,E,G)
false	false	false	0.1
false	false	true	0.2
false	true	false	0.05
false	true	true	0.05
true	false	false	0.3
true	false	true	0.1
true	true	false	0.05
true	true	true	0.15

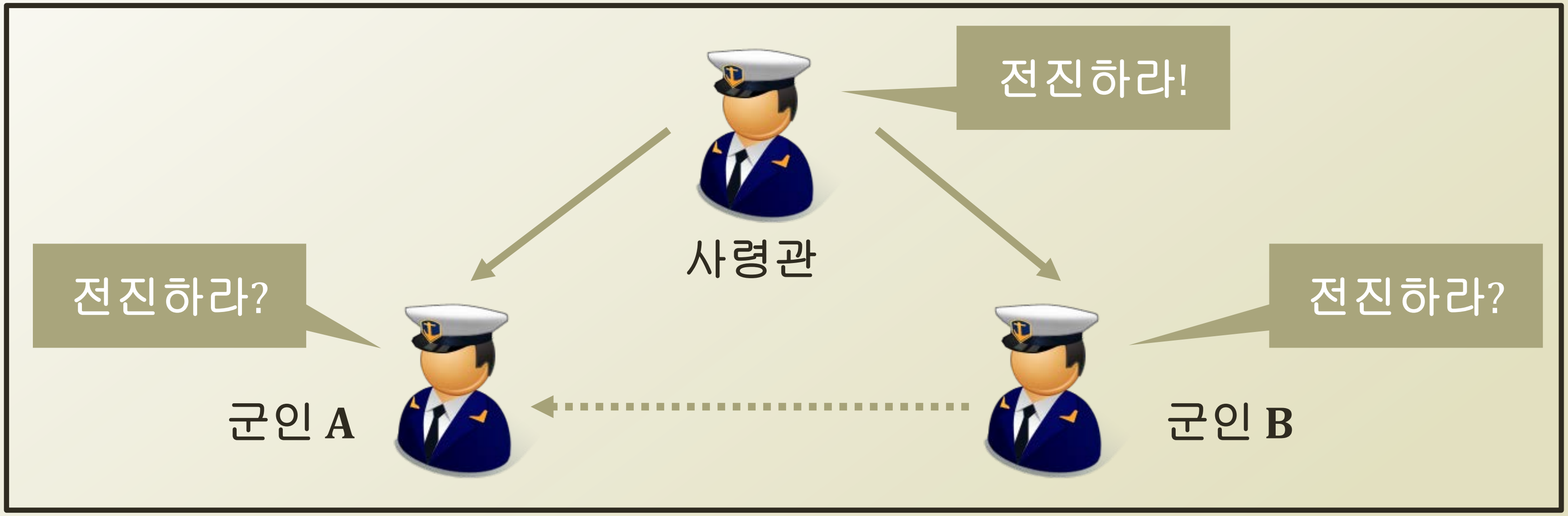
Sums to 1

- P(I=true)=sum of P(I,E,G) in rows with Int.=true
- $P(I=true, E = true \mid G=true)=\frac{P(I = true, E = true, G = true)}{P(G = true)}$
- **Any problem in this statistical model?**

Independence

- Recall the naïve Bayes classifier
  - Why introduce naïve assumption?
- Variables *A* and *B* are independent if any of the following hold:
  - P(A | B) = P(A)
    - P(A,B) = P(A)P(B)
    - P(B|A) = P(B)
  - This says that knowing the outcome of A does not tell me anything new about the outcome of B.
- Example
  - Suppose you have n coin flips and you want to calculate the joint distribution $P(C_1, \ldots, C_n)$
  - If the coin flips are not independent, you need 2^n-1 values in the table
  - If the coin flips are independent, then you need only one value
    - $P(C_1, .. C_n) = \prod_{i=1}^{n} P(C_i)$

Conditional vs. Marginal Independence

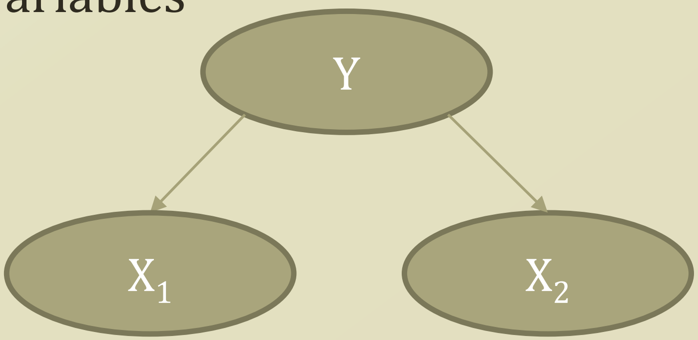


- Marginal independence
  - P(OfficerA=Go|OfficerB=Go) > P(OfficerA=Go)
  - **This is not marginally independent!**
    - X and Y are independent if and only if P(X)=P(X|Y)
    - Consequently, P(X,Y)=P(X)P(Y)
- Conditional independence
  - P(OfficerA=Go|OfficerB=Go,Commander=Go) =P(OfficerA=Go|Commander=Go)
  - **This is conditionally independent!**

BAYESIAN NETWORK

Detour: Naïve Bayes Classifier

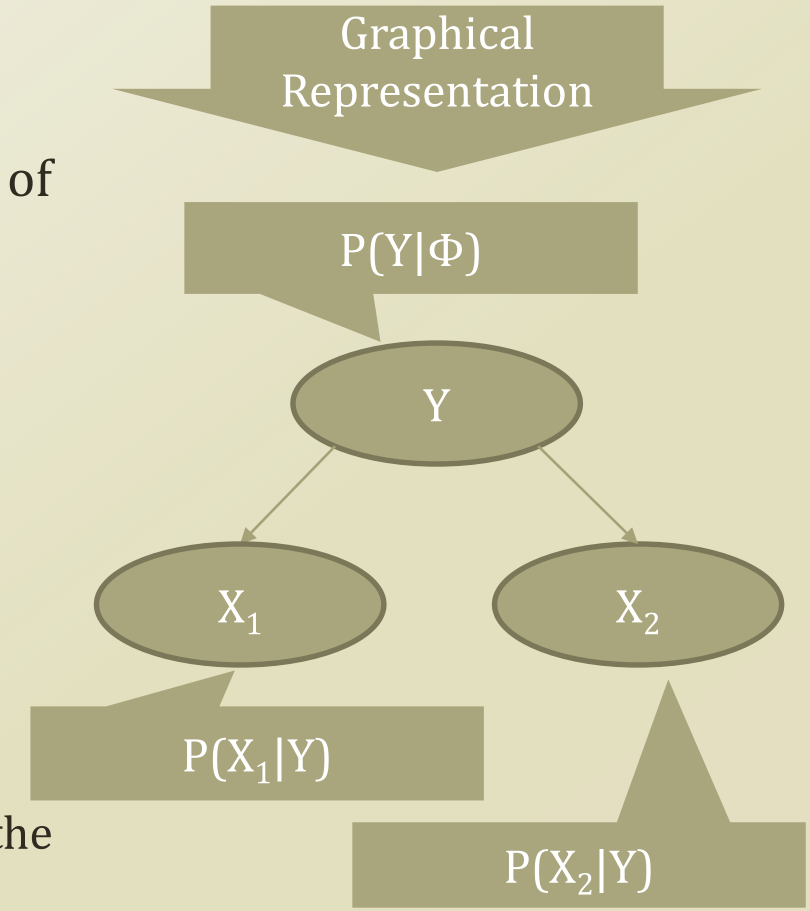
- Given:
  - Class Prior $\mathbf{P(Y)}$
  - $\boldsymbol{d}$ conditionally independent features $\boldsymbol{X}$ given the class $\boldsymbol{Y}$
  - For each $\boldsymbol{X_i}$, we have the likelihood of $\boldsymbol{P(X_i|Y)}$
- **Naïve Bayes Classifier Function**
  - $\boldsymbol{f_{NB}(x) = argmax_{Y=y} P(Y=y) \prod_{1 \le i \le d} P(X_i = x_i | Y = y)}$
- Essential information is modeled by
  - Random variables
  - Probability distribution of the random variables
  - Independence
- Any way to represent the model
  - Other than the formula?
  - i.e. graphical notation



Bayesian Network

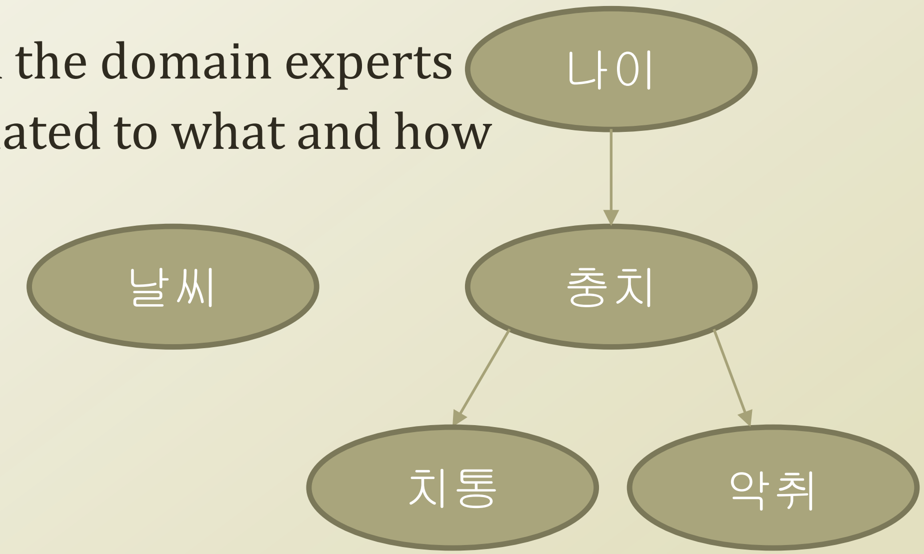
$$P(Y=y)\prod_{1\le i\le d} P(X_i = x_i | Y=y)$$

- A graphical notation of
  - Random variables
  - Conditional independence
  - To obtain a compact representation of the full joint distributions
- Syntax
  - A acyclic and directed graph
  - A set of nodes
    - A random variable
    - A conditional distribution given its parents
    - P $(X_i|\text{Parents}(X_i))$
  - A set of links
    - Direct influence from the parent to the child



Interpretation of Bayesian Network

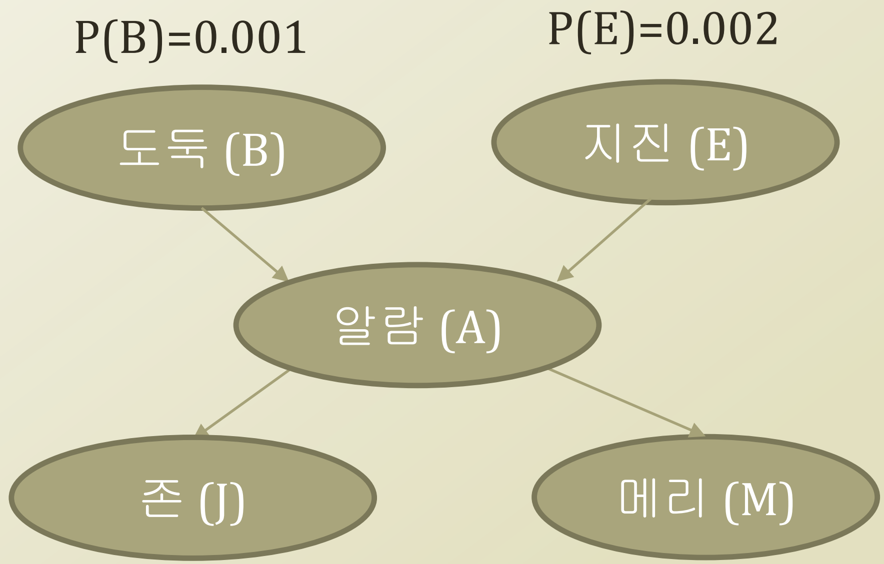
- Topology of network encodes conditional independence assertions
  - Often from the domain experts
  - What is related to what and how



- Interpretation
  - *Weather* is independent of the other variables
  - *Toothache* and *Stench* are conditionally independent given *Cavity*
  - *Cavity* influences the probability of *toothache and stench*

Another Example

- Scenario
  - I'm at work
  - Neighbor John calls to say my alarm is ringing
  - Neighbor Mary doesn't call.
  - Sometimes it's set off by minor earthquakes.
  - Is there a burglar?
- Variables
  - *Burglary, Earthquake, Alarm, JohnCalls, MaryCalls*
- Network topology reflects "causal" knowledge:
  - A burglar can set the alarm off
  - An earthquake can set the alarm off
  - The alarm can cause Mary to call
  - The alarm can cause John to call



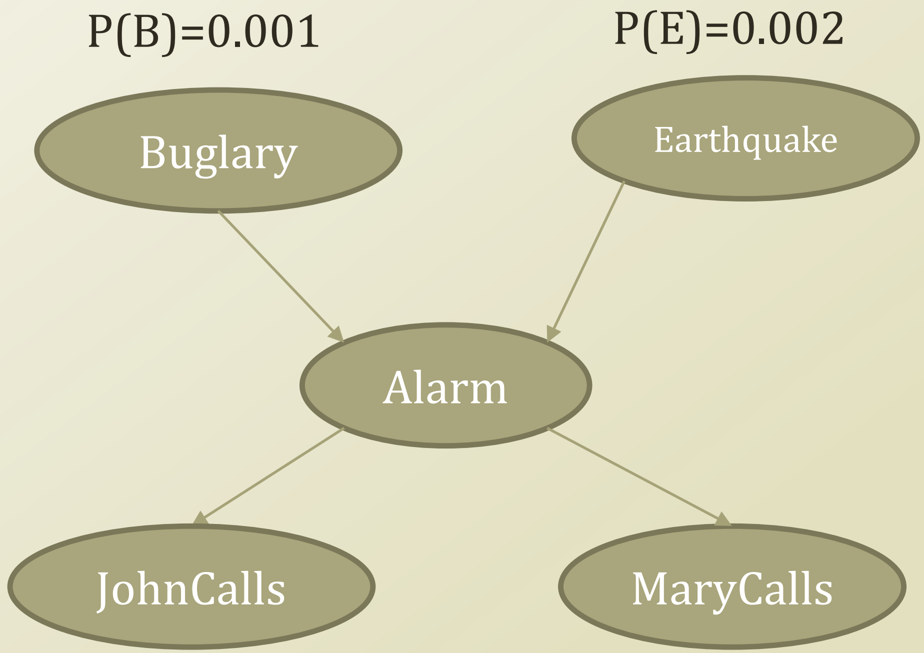
B	E	P(A|B,E)
T	T	0.95
T	F	0.94
F	T	0.29
F	F	0.001

A	P(J|A)
T	0.90
F	0.05

A	P(M|A)
T	0.70
F	0.01

Components of Bayesian Network

- Qualitative components
  - Prior knowledge of causal relations
  - Learning from data
  - Frequently used structures
  - Structural aspects
- Quantitative components
  - Conditional probability tables
  - Probability distribution assigned to nodes
- Probability computing is related to both
  - Quantitative and Qualitative



B	E	P(A|B,E)
T	T	0.95
T	F	0.94
F	T	0.29
F	F	0.001

A	P(J|A)
T	0.90
F	0.05

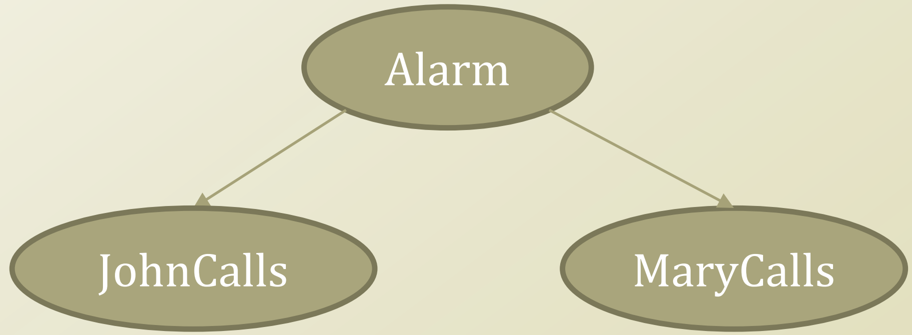
A	P(M|A)
T	0.70
F	0.01

$P(A,B)=P(A)P(B)$

$P(A|B)=P(A)$

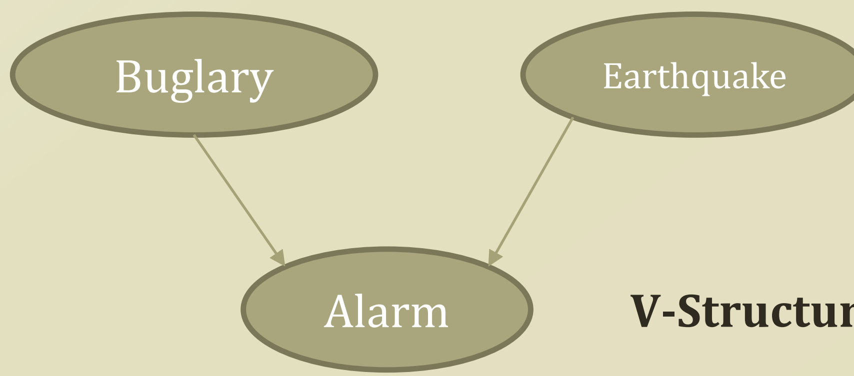
Typical Local Structures

- Common parent
  - Fixing "alarm" decouples "JohnCalls" and "MaryCalls"
  - $J \perp M|A$
    - $P(J,M|A)=P(J|A)P(M|A)$
- Cascading
  - Fixing "alarm" decouples "Buglary" and "MaryCalls"
  - $B \perp M|A$
    - $P(M|B,A)=P(M|A)$
- V-Structure
  - Fixing "alarm" couples "Buglary" and "Earthquake"
  - $\sim (B \perp E|A)$
    - $P(B,E,A)=P(B)P(E)P(A|B,E)$
- Any algorithm for complex graph?



Common Parent

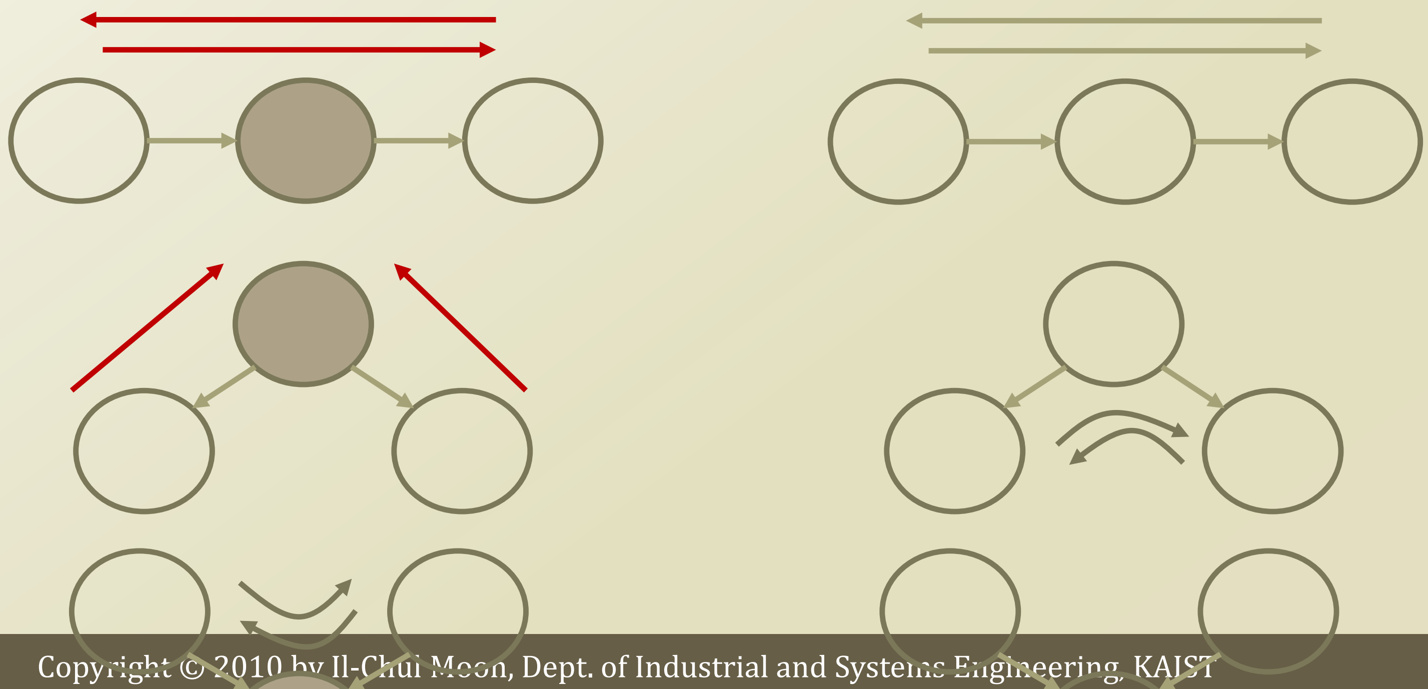
Cascading



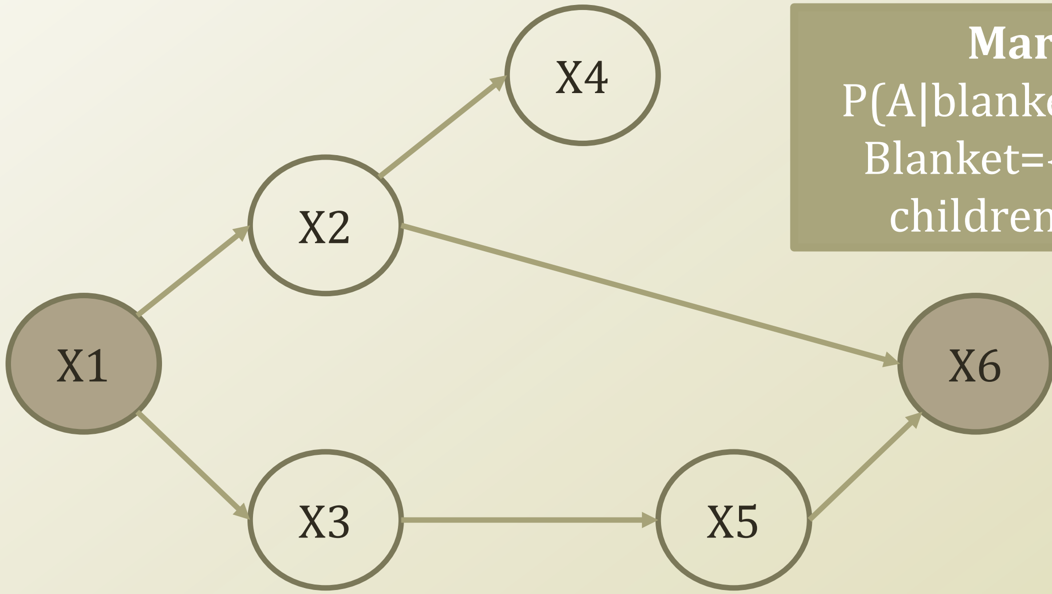
V-Structure

Bayes Ball Algorithm

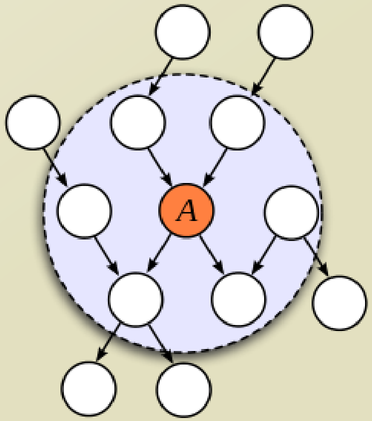
- Purpose: checking $X_A \perp X_B | X_c$
  - Shade all nodes in X_c
  - Place balls at each node in X_A
  - Let the ball rolling on the graph by Bayes ball rules
  - Then, ask whether there is any ball reaching X_B



Exercise of Bayes Ball Algorithm



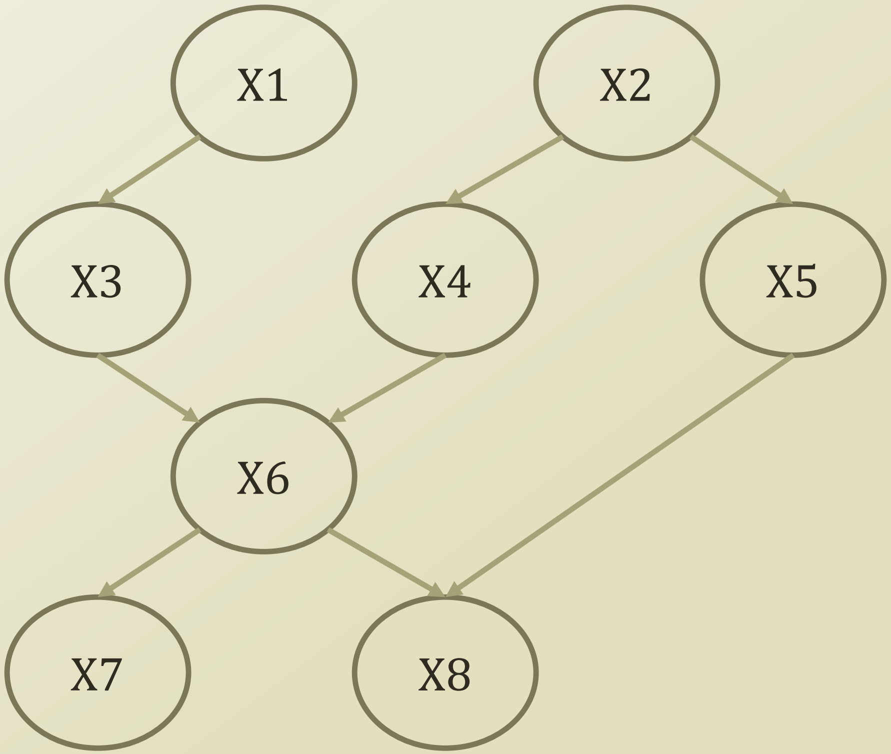
Markov Blanket
P(A|blanket, B)=P(A|blanket)
Blanket={parents, children, children's other parents}



- Answer the below case
  - $X_1 \perp X_4 | \{X_2\}$
  - $X_2 \perp X_5 | \{X_1\}$
  - $X_1 \perp X_6 | \{X_2, X_3\}$
  - $X_2 \perp X_3 | \{X_1, X_6\}$
- D-Seperation
  - X is d-separated (directly-separated) from Z given Y if we cannot send a ball from any node in X to any node in Z using the Bayes ball algorithm

Factorization of Bayesian Network

- Factorization theorem
  - Given a Bayesian network
  - The most general form of the probability distribution
    - that is consistent with the probabilistic independencies encoded in the network
  - Factorizes according to the node given its parents
    - $P(X) = \prod_i P(X_i|X_{\pi_i})$
  - X_{π_i} is the set of parent nodes of X_i
- The most general form?
  - What are the not-most general form???
  - More discussions of d-separation, not going to be in this classroom

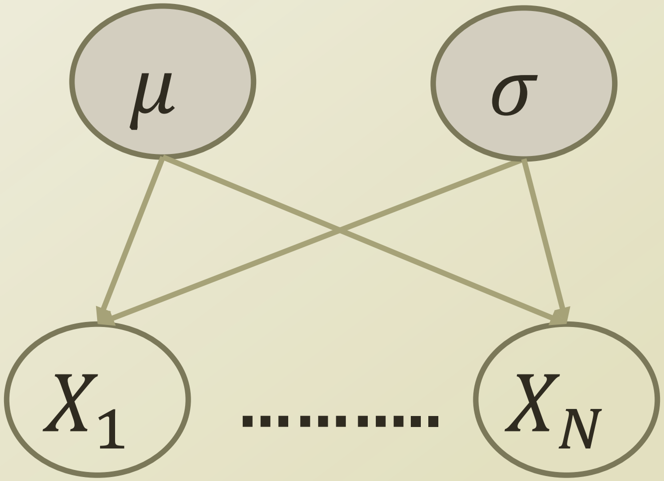


P(X1,X2,X3,X4,X5,X6,X7,X8) =P(X1)P(X2)P(X3|X1)P(X4|X2)P(X5|X2) P(X6|X3,X4)P(X7|X6)P(X8|X5,X6)

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)}$$

$$Posterior = \frac{Likelihood \times Prior\ Knowledge}{Normalizing\ Constant}$$

Plate Notation



- Let's consider a certain Gaussian model
  - Many Xs
  - Depend upon the same parameter
    - Mean and variance
  - Independent between Xs
- Dealing with many random variables
  - Simplify the graphical notation with a box
  - Works like a for-loop
- $P(D|\theta) = P(X_1, \ldots, X_N|\mu, \sigma)$
  $$= \prod_N P(X_n|\mu, \sigma)$$
  - Naïve assumption
- Likelihood function
  - $L(\theta|D) = P(D|\theta)$

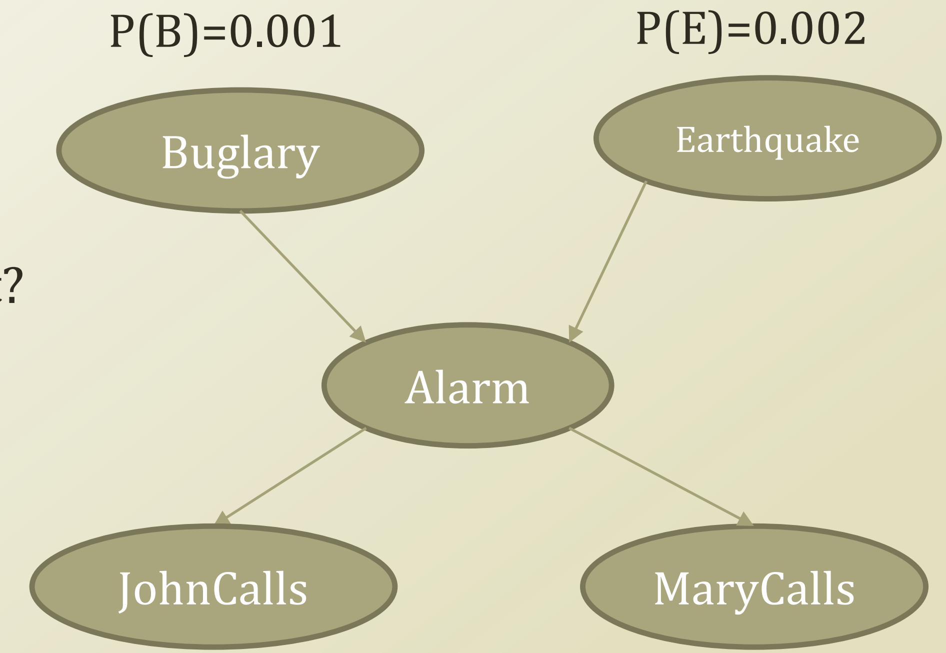
Inference Question 1: Likelihood

- Given a set of evidence, what is the likelihood of the evidence set?
  - $X = \{X_1 \dots X_N\}$: all random variables
  - $X_V = \{X_{k+1} \dots X_N\}$: evidence variables
    - x_V: evidence values
  - $X_H = X - X_V = \{X_1 \dots X_k\}$: hidden variables
- General form
  - $P(x_V) = \sum_{X_H} P(X_H, X_V)$

$$= \sum_{x_1} \dots \sum_{x_k} P(x_1 \dots x_k, x_V)$$

  - Likelihood of x_V

P(B=true, MC=true)=?



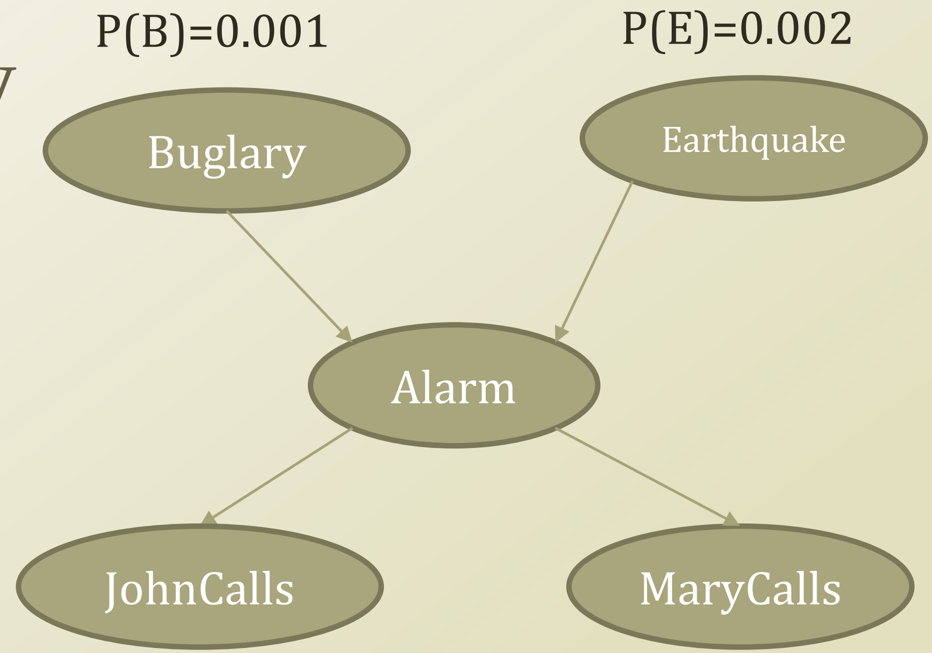
B	E	P(A|B,E)
T	T	0.95
T	F	0.94
F	T	0.29
F	F	0.001

A	P(J|A)
T	0.90
F	0.05

A	P(M|A)
T	0.70
F	0.01

Inference Question 2: Conditional Probability

P(A|B=true, MC=true)=?



- Given a set of evidence, what is the conditional probability of interested hidden variables?
  - $X_H = \{Y, Z\}$
    - Y: interested hidden variables
    - Z: uninterested hidden variables
- General form
  - $P(Y|x_V) = \sum_z P(Y, Z = z|x_V)$
    $$= \sum_z \frac{P(Y, Z, x_V)}{P(x_V)}$$
    $$= \sum_z \frac{P(Y, Z, x_V)}{\sum_{y,z} P(Y = y, Z = z, x_V)}$$
  - Conditional probability of Y given x_V

B	E	P(A|B,E)
T	T	0.95
T	F	0.94
F	T	0.29
F	F	0.001

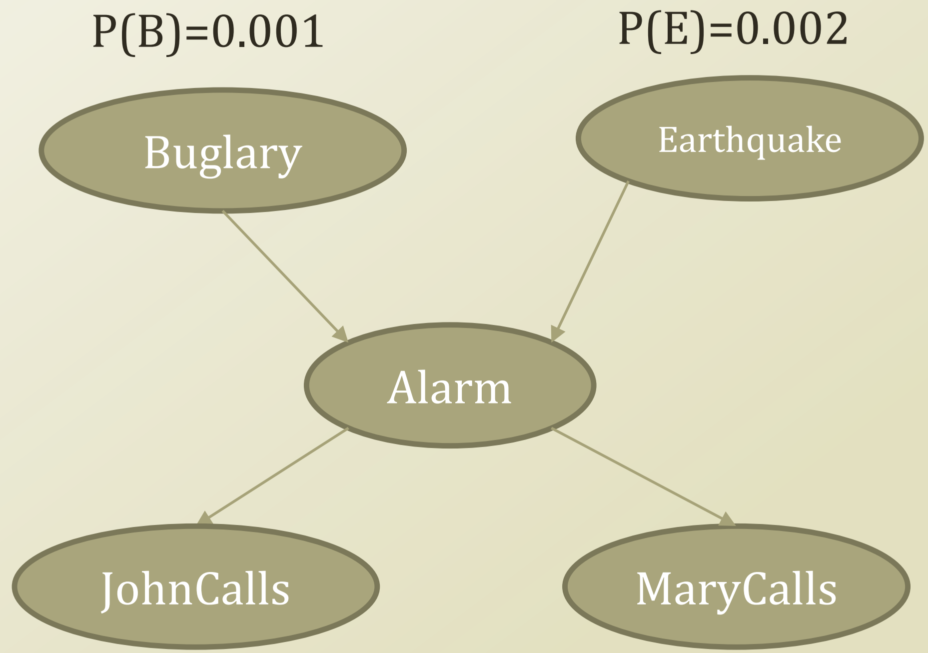
A	P(J|A)
T	0.90
F	0.05

A	P(M|A)
T	0.70
F	0.01

$\text{argmax}_a P(A|B=true, MC=true)=?$

Inference Question 3: Most Probable Assignment

- Given a set of evidence, what is the most probable assignment, or explanation, given the evidence?
  - Some variables of interests
  - Need to utilize the inference question 2
    - Conditional probability
  - Maximum a posteriori configuration of *Y*
- Applications of *a posteriori*
  - Prediction
    - B, E → A
  - Diagnosis
    - A → B, E



B	E	P(A|B,E)
T	T	0.95
T	F	0.94
F	T	0.29
F	F	0.001

A	P(J|A)
T	0.90
F	0.05

A	P(M|A)
T	0.70
F	0.01

Marginalization and Elimination

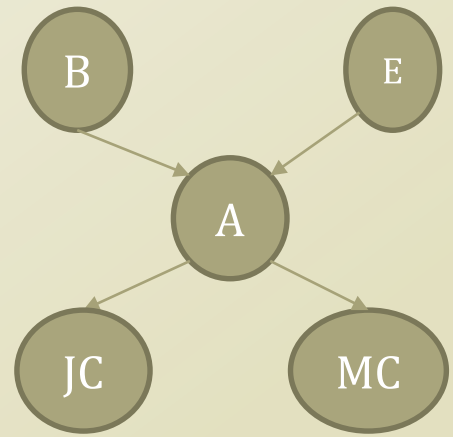
- Computing joint probabilities is a key
  - How to compute them?
  - Many, many, many multiplications and summations
    - P(a=true,b=true,mc=true)=$\sum_{JC}\sum_{E} P(a,b,E,JC,mc)$

$$=\sum_{JC}\sum_{E} P(JC|a)P(mc|a)P(a|b,E)P(E)P(b)$$

  - In big Oh notation?
- Is there any better method?
  - What-if we move around the summation?
    - P(a,b,mc)=$\sum_{JC}\sum_{E} P(a,b,E,JC,mc)$

$$=P(b)P(mc|a)\sum_{JC} P(JC|a)\sum_{E} P(a|b,E)P(E)$$

  - Did we reduced the computation complexity?



P(E)=0.002

P(B)=0.001

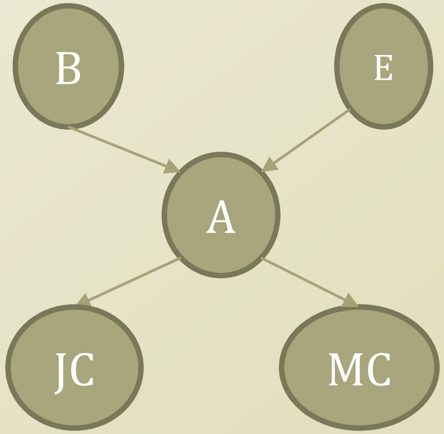
B	E	P(A|B,E)
T	T	0.95
T	F	0.94
F	T	0.29
F	F	0.001

A	P(J|A)
T	0.90
F	0.05

A	P(M|A)
T	0.70
F	0.01

Variable Elimination

- Preliminary
  - $P(e|jc,mc) = \alpha P(e,jc,mc)$
- Joint probability (e=jc=mc=true)
  - $P(e,jc,mc,B,A) =$
    $\alpha P(e)\sum_B P(b)\sum_A P(a|b,e)\, P(jc|a)P(mc|a)$
    - Line up the terms by the topological order
    - Consider a probability distribution as a function
      - $f_E(E=t) = 0.002$
  - $= \alpha f_E(e)\sum_B f_B(b)\sum_A f_A(a,b,e)\, f_J(a)f_M(a)$



A	$f_{JM}(A)$
T	0.63
F	0.0005

A	$f_J(A)$
T	0.90
F	0.05

×

A	$f_M(A)$
T	0.70
F	0.01

P(B)=0.001

B	E	P(A|B,E)
T	T	0.95
T	F	0.94
F	T	0.29
F	F	0.001

P(E)=0.002

A	P(J|A)
T	0.90
F	0.05

A	P(M|A)
T	0.70
F	0.01

  - $= \alpha f_E(e)\sum_B f_B(b)\sum_A f_A(a,b,e)\, f_{JM}(a)$

Variable Elimination cont.

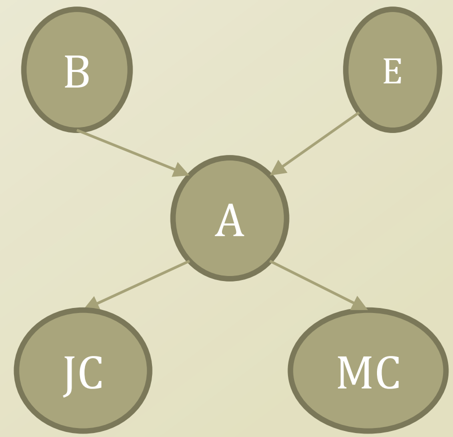
- $= \alpha f_E(e) \sum_B f_B(b) \underbrace{\sum_A f_A(a,b,e)\, f_{JM}(a)}$

A	B	E	$f_{AJM}(A,B,E)$
T	T	T	0.95*0.63
T	T	F	0.94*0.63
T	F	T	0.29*0.63
T	F	F	0.001*0.63
F	T	T	0.05*0.0005
F	T	F	0.06*0.0005
F	F	T	0.71*0.0005
F	F	F	0.999*0.0005

A	B	E	$f_A(A,B,E)$
T	T	T	0.95
T	T	F	0.94
T	F	T	0.29
T	F	F	0.001
F	T	T	0.05
F	T	F	0.06
F	F	T	0.71
F	F	F	0.999

✖

A	$f_{JM}(A)$
T	0.63
F	0.0005



P(B)=0.001

P(E)=0.002

B	E	P(A|B,E)
T	T	0.95
T	F	0.94
F	T	0.29
F	F	0.001

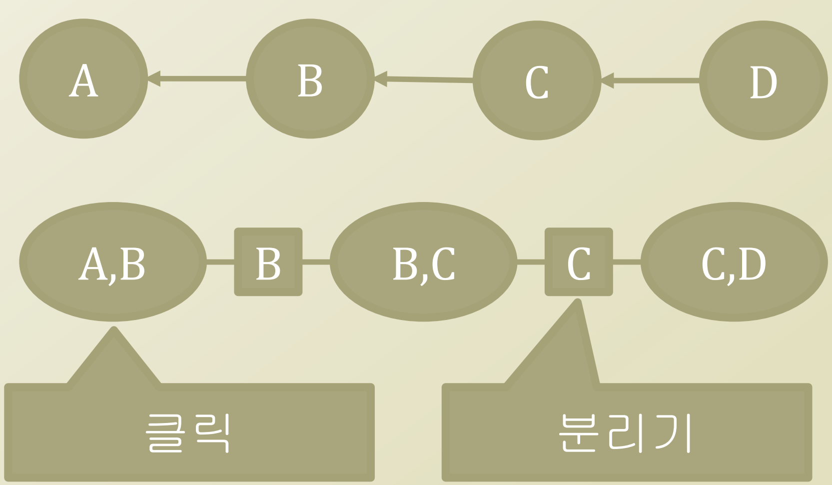
A	P(J|A)
T	0.90
F	0.05

A	P(M|A)
T	0.70
F	0.01

- $= \alpha f_E(e) \sum_B f_B(b) \sum_A f_{AJM}(a,b,e)$
- $= \alpha f_E(e) \sum_B f_B(b)\, f_{\bar{A}JM}(b,e)$
- $= \alpha f_E(e) \sum_B f_{B\bar{A}JM}(b,e)$
- $= \alpha f_E(e) f_{\bar{B}\bar{A}JM}(e)$
- $= \alpha f_{E\bar{B}\bar{A}JM}(e)$

B	E	$f_{\bar{A}JM}(B,E)$
T	T	0.95*0.63+0.05*0.0005
T	F	0.94*0.63+0.06*0.0005
F	T	0.29*0.63+0.71*0.0005
F	F	0.001*0.63+0.999*0.0005

Potential Functions



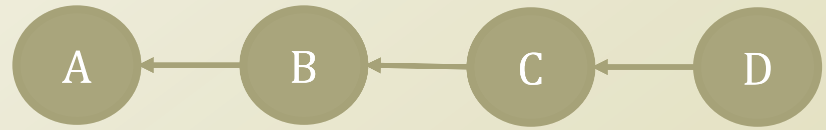
- $P(A,B,C,D)$
- $= P(A|B)P(B|C)P(C|D)P(D)$
- Let's define a potential function
  - Potential function:
    a function which is not a probability function yet, but once normalized it can be a probability distribution function
  - Potential function on nodes
    - $\psi(a,b), \psi(b,c), \psi(c,d)$
  - Potential function on links
    - $\phi(b), \phi(c)$
- How to setup the function?
  - $P(A,B,C,D) = P(U) = \frac{\prod_N \psi(N)}{\prod_L \phi(L)} = \frac{\psi(a,b)\psi(b,c)\psi(c,d)}{\phi(b)\phi(c)}$
    - $\psi(a,b) = P(A|B), \psi(b,c) = P(B|C), \psi(c,d) = P(C|D)P(D)$
    - $\phi(b) = 1, \phi(c) = 1$
  - $P(A,B,C,D) = P(U) = \frac{\prod_N \psi(N)}{\prod_L \phi(L)} = \frac{\psi^*(a,b)\psi^*(b,c)\psi^*(c,d)}{\phi^*(b)\phi^*(c)}$
    - $\psi^*(a,b) = P(A,B), \psi^*(b,c) = P(B,C), \psi^*(c,d) = P(C,D)$
    - $\phi^*(b) = P(B), \phi^*(c)$ =P(C)

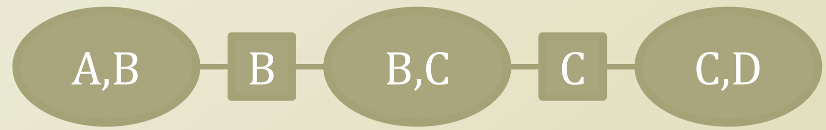
Marginalization is also applicable:

$$\psi(w) = \sum_{v-w} \psi(v)$$

Constructing a potential of a subset (w) of all variables (v)

Absorption in Clique Graph





- Only applicable to the tree structure of clique graph
- Let's assume
  - $P(B) = \sum_A \psi(A,B)$
  - $P(B) = \sum_C \psi(B,C)$
  - $P(B) = \phi(B)$
  - How to find out the ψs and the ϕs?
    - When the ψs change by the observations: P(A,B) → P(A=1,B)
    - A single ψ change can result in the change of multiple ψs
    - The effect of the observation propagates through the clique graph
    - Belief propagation!
- How to propagate the belief?
  - Absorption (update) rule
  - Assume $\psi^*(A,B)$,$\psi(B,C)$, and $\phi(B)$
  - Define the update rule for separators
    - $\phi^*(B) = \sum_A \psi^*(A,B)$
  - Define the update rule for cliques
    - $\psi^*(B,C)$=$\psi(B,C)\frac{\phi^*(B)}{\phi(B)}$

Why does this work?

$$\sum_C \psi^*(B,C) = \sum_C \psi(B,C)\frac{\phi^*(B)}{\phi(B)}$$

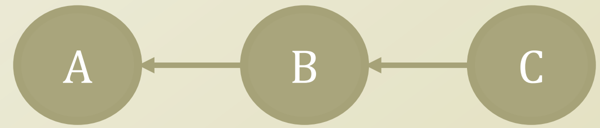
$$= \frac{\phi^*(B)}{\phi(B)}\sum_C \psi(B,C) = \frac{\phi^*(B)}{\phi(B)}\phi(B) = \sum_A \psi^*(A,B)$$

Guarantees the local consistency

→ Global consistency after iterations

Simple Example of Belief Propagation
신뢰 전파

- Initialized the potential functions
  - $\psi(a,b) = P(a|b), \psi(b,c) = P(b|c)P(c)$
  - $\phi(b) = 1$
- Example 1. P(b)=?
  - $\phi^*(b) = \sum_a \psi(a,b) = 1$
  - $\psi^*(b,c) = \psi(b,c)\frac{\phi^*(b)}{\phi(b)} = P(b|c)P(c) = P(b,c)$
  - $\phi^{**}(b) = \sum_c \psi(b,c) = \sum_c P(b,c) \; = P(b)$
  - $\psi^*(a,b) = \psi(a,b)\frac{\phi^{**}(b)}{\phi^*(b)} = \frac{P(a|b)P(b)}{1} = P(a,b)$
  - $\phi^{***}(b) = \sum_a \psi^*(a,b) = P(b)$
- Example 2. P(b|a=1,c=1)=?
  - $\phi^*(b) = \sum_a \psi(a,b)\delta(a=1) = P(a=1|b)$
  - $\psi^*(b,c) = \psi(b,c)\frac{\phi^*(b)}{\phi(b)} = P(b|c=1)P(c=1)\frac{P(a=1|b)}{1}$
  - $\phi^{**}(b) = \sum_c \psi(b,c)\,\delta(c=1) = P(b|c=1)P(c=1)P(a=1|b)$
  - $\psi^*(a,b) = \psi(a,b)\frac{\phi^{**}(b)}{\phi^*(b)} = \mathrm{P(a=1|b)}\frac{P(b|c=1)P(c=1)P(a=1|b)}{P(a=1|b)} = P(b|c=1)P(c=1)P(a=1|b)$
  - $\phi^{***}(b) = \sum_a \psi^*(a,b)\,\delta(a=1) = P(b|c=1)P(c=1)P(a=1|b)$

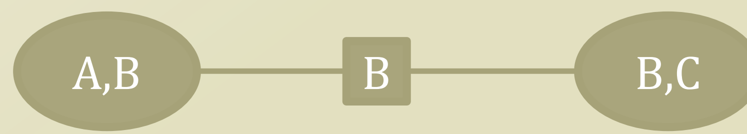


베이지안 네트워크

1) $\phi^*(b) \rightarrow$　2) $\psi^*(b,c) \rightarrow$

$\leftarrow$ 4) $\psi^*(a,b)$　$\leftarrow$ 3) $\phi^{**}(b)$

5) $\phi^{***}(b) \rightarrow$



클릭 그래프

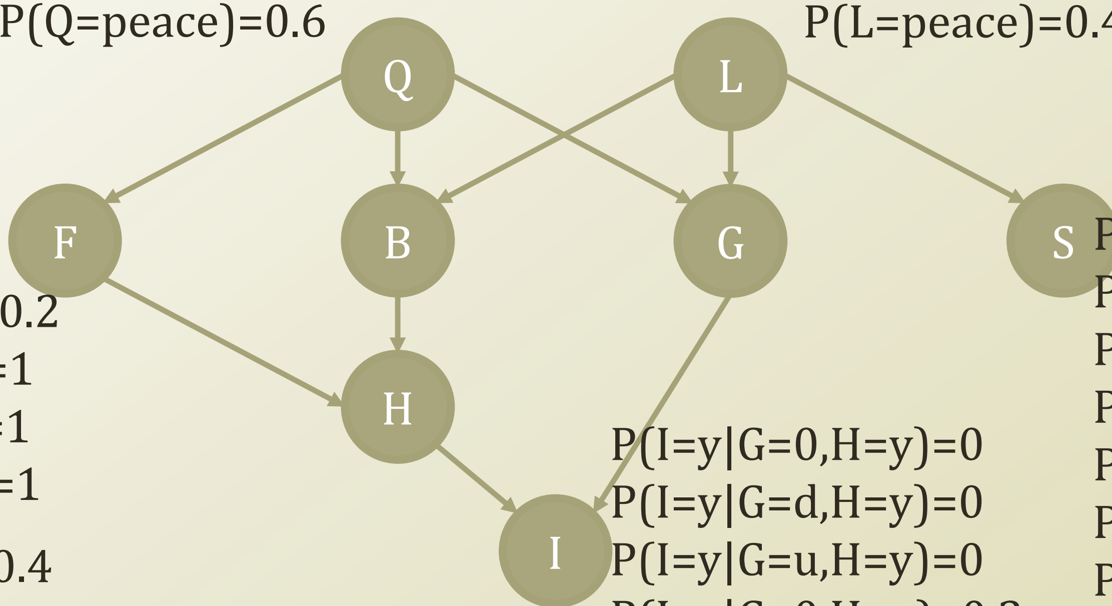
Another Example of Bayes Net.

$P(Q=peace)=0.6$

$P(L=peace)=0.4$

$P(F=y|Q=p)=0.8$
$P(F=y|Q=w)=0.1$

$P(S=y|L=p)=0.8$
$P(S=y|L=w)=0.3$



$P(G=u|L=p,Q=p)=0.3$
$P(G=u|L=p,Q=w)=0.8$
$P(G=u|L=w,Q=p)=0.1$
$P(G=u|L=w,Q=w)=0.2$
$P(G=0|L=p,Q=p)=0.6$
$P(G=0|L=p,Q=w)=0.1$
$P(G=0|L=w,Q=p)=0.2$
$P(G=0|L=w,Q=w)=0.2$

$P(B=r|L=p,Q=p)=0.2$
$P(B=r|L=p,Q=w)=1$
$P(B=r|L=w,Q=p)=1$
$P(B=r|L=w,Q=w)=1$

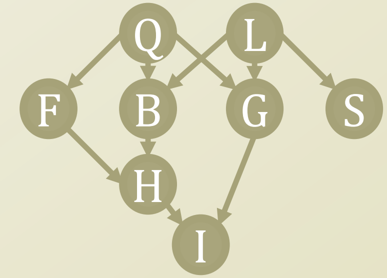
$P(H=y|B=r,F=y)=0.4$
$P(H=y|B=r,F=n)=1$
$P(H=y|B=s,F=n)=0.5$
$P(H=y|B=s,F=y)=0.1$

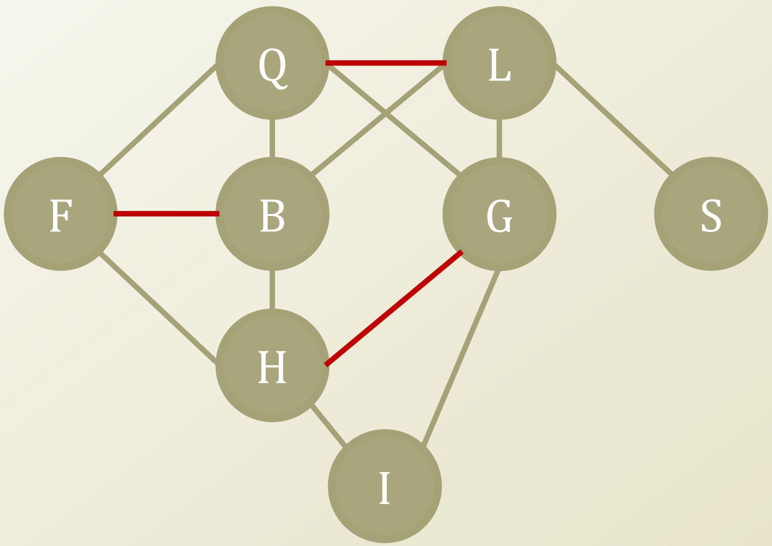
$P(I=y|G=0,H=y)=0$
$P(I=y|G=d,H=y)=0$
$P(I=y|G=u,H=y)=0$
$P(I=y|G=0,H=n)=0.3$
$P(I=y|G=d,H=n)=0.1$
$P(I=y|G=u,H=n)=1$

- L={peace,war}, Q={peace, war}, F={yes,no},B={run, stay}, G={up,0,down},S={yes,no},H={yes,no},I={yes,no}
- My clan (Q) vs. Opponent Clan (L) dispute over Goats (G)
- Smoke (S) lets know the opponent strategy
- Farming (F) and Hunting (B) are the major sources of resources
- Hunger (H) should be avoided, and the improvement of life quality (I) should be pursued.

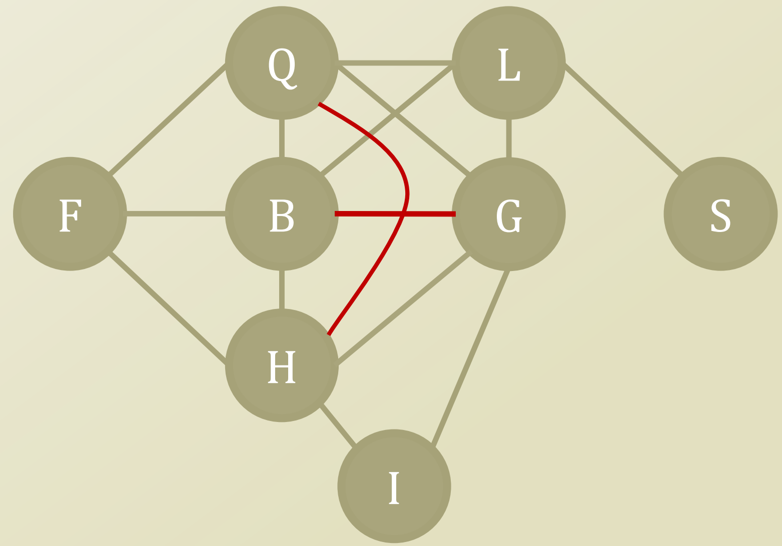
Given that 1) my clan is in war, 2) goat gain is steady, and 3) farming goes on, will I face hunger?

Moralization and Triangulation





Moralized Network

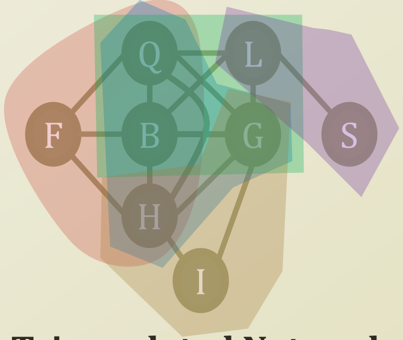


Triangulated Network

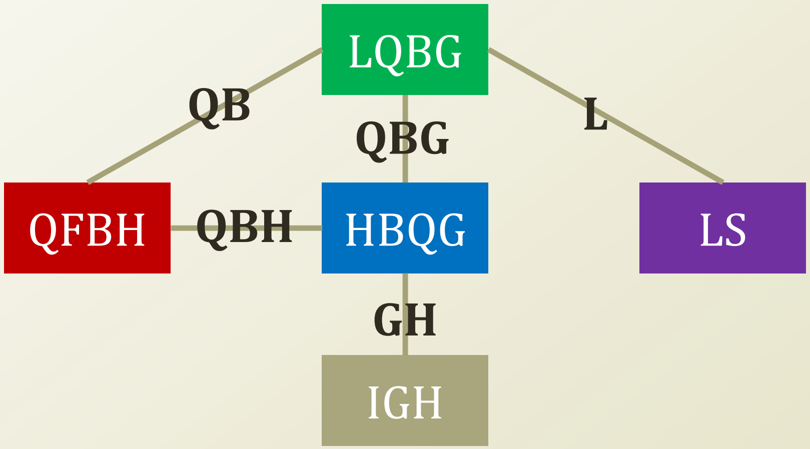
Loop:
Q→B→H→G→Q
L→B→H→G→L
and one more?

- Moralization
  - Marry the parents
  - Drop the directionality
- Triangulation 삼각화
  - Add a chord link for a loop whose length is over four
  - Chord: a link joining two non-consecutive nodes of a loop

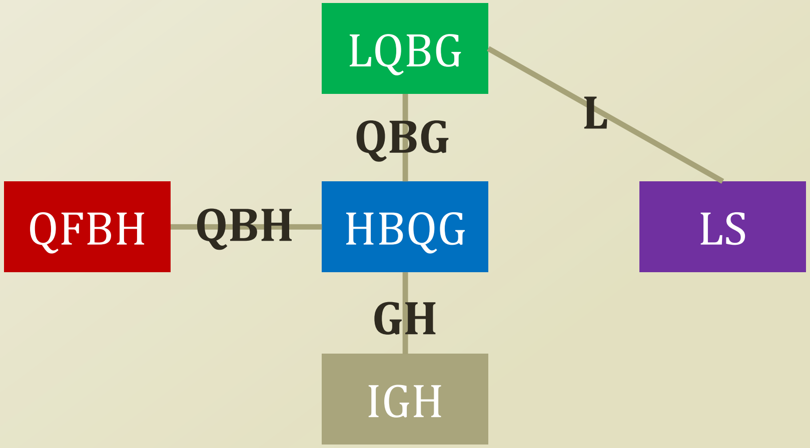
Clique Graph and Junction Tree



Triangulated Network



Clique Graph



Junction Tree

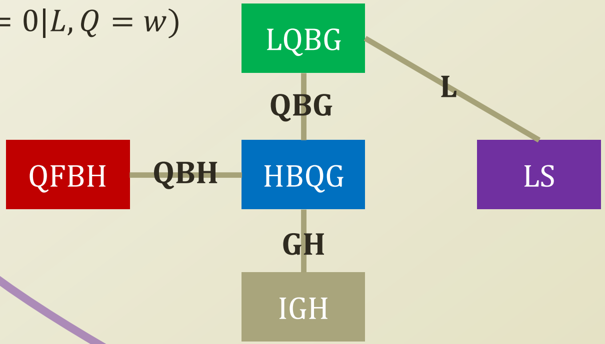
- Clique graph
  - Clique
    - Maximal cliques in the triangulated network
  - Separator
    - Intersecting nodes of two identified cliques in the triangulated network

- Junction Tree
  - Sub-graph of clique graph
  - Tree structured
  - Contains all nodes in clique graph
  - Weight of an edge in the clique graph == # of random variables associated with the edge
  - Maximum weight spanning tree == junction tree

Define the update rule for separators: $\phi^*(B) = \sum_A \psi^*(A,B)$

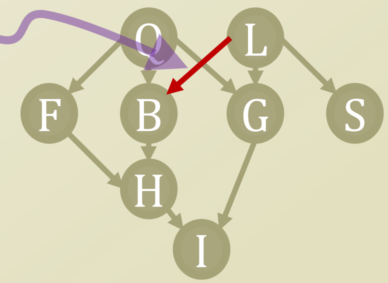
Define the update rule for cliques: $\psi^*(B,C) = \psi(B,C)\frac{\phi^*(B)}{\phi(B)}$

$$m_L(B) = \sum_L P(L)P(B|L,Q=w)P(G=0|L,Q=w)$$



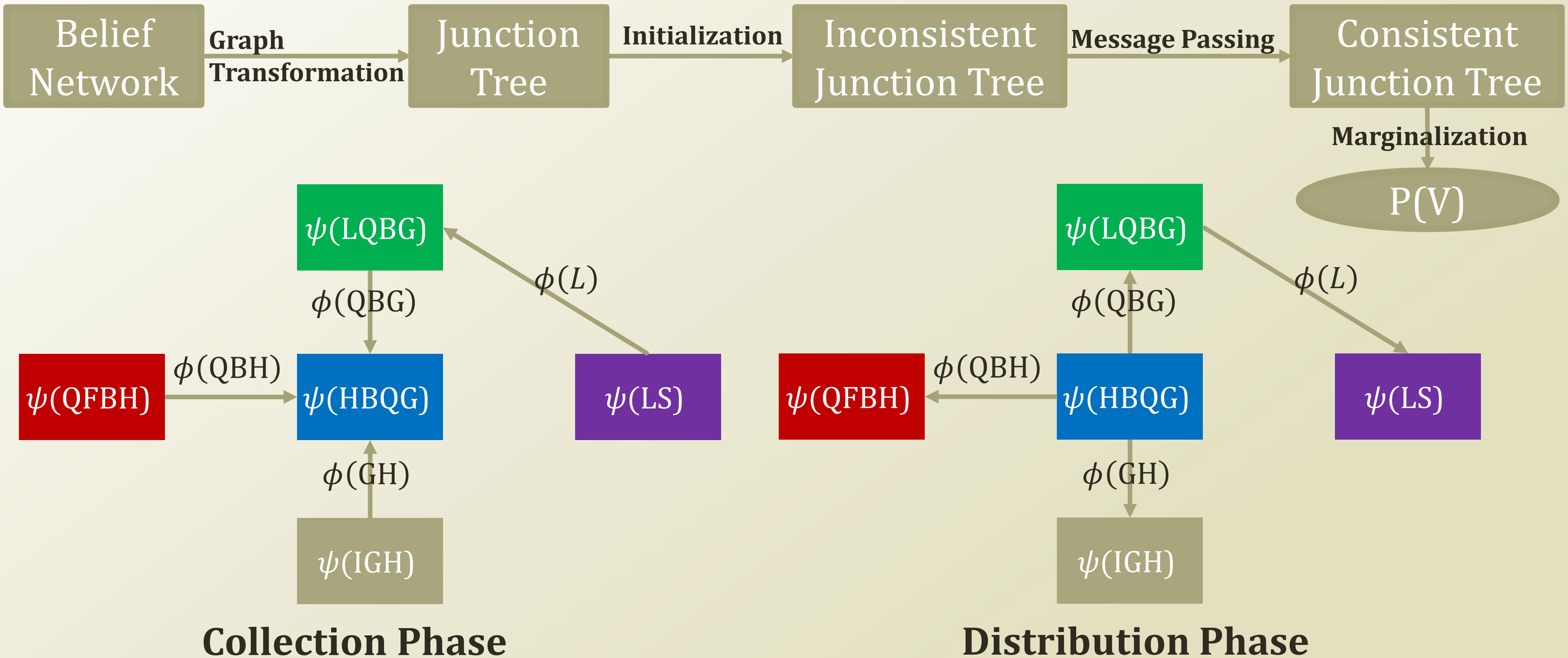
Junction Tree Algorithm Example

- Calculate P(H=y|Q=w,G=0,F=y)
- Initialize
  - $\phi(L) = \phi(BHQ) = \phi(QBG) = \phi(GH) = 1$
  - Choose HBQG as the root node because it has H
  - Assign potential functions from the joint probability
    - $\psi(SL) = P(L)P(S|L), \psi(LQBG) = P(B|L,Q)P(G|L,Q),$
      $\psi(HBQG) = P(Q), \psi(QFBH) = P(F|Q)P(H|B,F), \psi(IGH) = P(I|H,G)$
- Absorption
  - $\phi^*(GH) = \sum_I \psi(IGH)\delta(G=0) = 1$
  - $\phi^*(QBH) = \sum_F \psi(QFBH)\delta(F=y,Q=w) = P(F=y|Q=w)P(H|B,F=y)$
  - $\phi^*(QBG) = \sum_L \psi^*(LQBG)\delta(Q=w,G=0) = \sum_L P(L)P(B|L,Q=w)P(G=0|L,Q=w)$
    - $\psi^*(LQBG) = \frac{\sum_S \psi(SL)}{\phi(L)}\psi(LQBG)\delta(Q=w,G=0) = \sum_S P(L)P(S|L)\,P(B|L,Q=w)P(G=0|L,Q=w)$



- Update of the root node, HBQG
  - $\psi^*(HBQG) = \frac{\phi^*(BHQ)\phi^*(QBG)\phi^*(GH)}{\phi(BHQ)\phi(QBG)\phi(GH)}\psi(HBQG)\delta(Q=w,G=0)$
    $= P(Q)P(F=y|Q=w)P(H|B,F$
    $= y)\sum_L P(L)P(B|L,Q=w)P(G=0|L,Q=w)$ =P(H,B,F=y,G=0,Q=w) Q=w)
- Answer
  - P(H=y|F=y,G=0, Q=w)= $\frac{\sum_B \psi^*(HBQG)(H=y)}{\sum_{H,B}\psi^*(HBQG)} =$
    $\frac{\sum_B P(Q)P(F=y|Q=w)P(H=y|B,F=y)\sum_L P(L)P(B|L,Q=w)P(G=0|L,Q=w)}{\sum_{H,B} P(Q)P(F=y|Q=w)P(H|B,F=y)\sum_L P(L)P(B|L,Q=w)P(G=0|L,Q=w)} = \frac{\sum_B P(H=y|B,F=y)m_L(B)}{\sum_{H,B} P(H|B,F=y)m_L(B)}$

Junction Tree Algorithm in Diagram



- Iterative run of
  - Collection phase: going up to the root
  - Distribution phase: going down to the leaf
  - Messages are absorbed in the process
- Stop when the queried probability is answered
  - You may have to perform summation on the joint probability obtained by the chosen potential function

Acknowledgement

- This slideset is greatly influenced
  - By Prof. Eric Xing at CMU
  - By Prof. Chris Williams at Univ. of Edinburgh

Further Readings

- Bishop Chapter 8
- Murphy Chapter 20
- Barber Chapter 4, 5, 6