Graphical Model II: Inference

Wang Houfeng
Institute of Computational Linguistics
Peking University

Outline

- **Overview**
- Inference in GM
- Exact inference

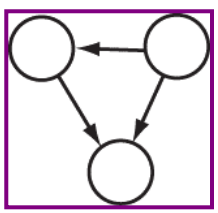
Bayesian Networks (BN)

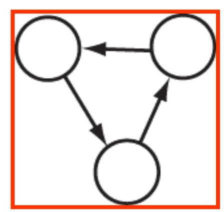
- Graph $G = (V, E)$, $V = \{X_1, X_2, \ldots, X_N\}$ (sometimes we write $V = \{v_1, \ldots, v_N\}$).

- $E = \{(X_i, X_j) : i \neq j\}$: set of directed edges.

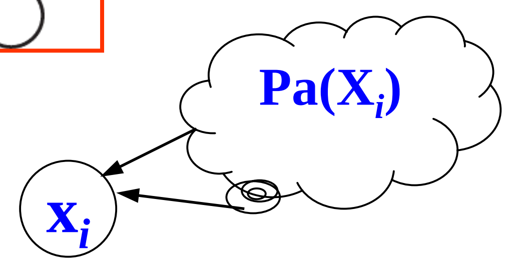
Also, $(X_i, X_j) \in E$ means $X_i \to X_j$

- G is acyclic(a DAG), so there are no directed cycles.

valid: 

invalid: 

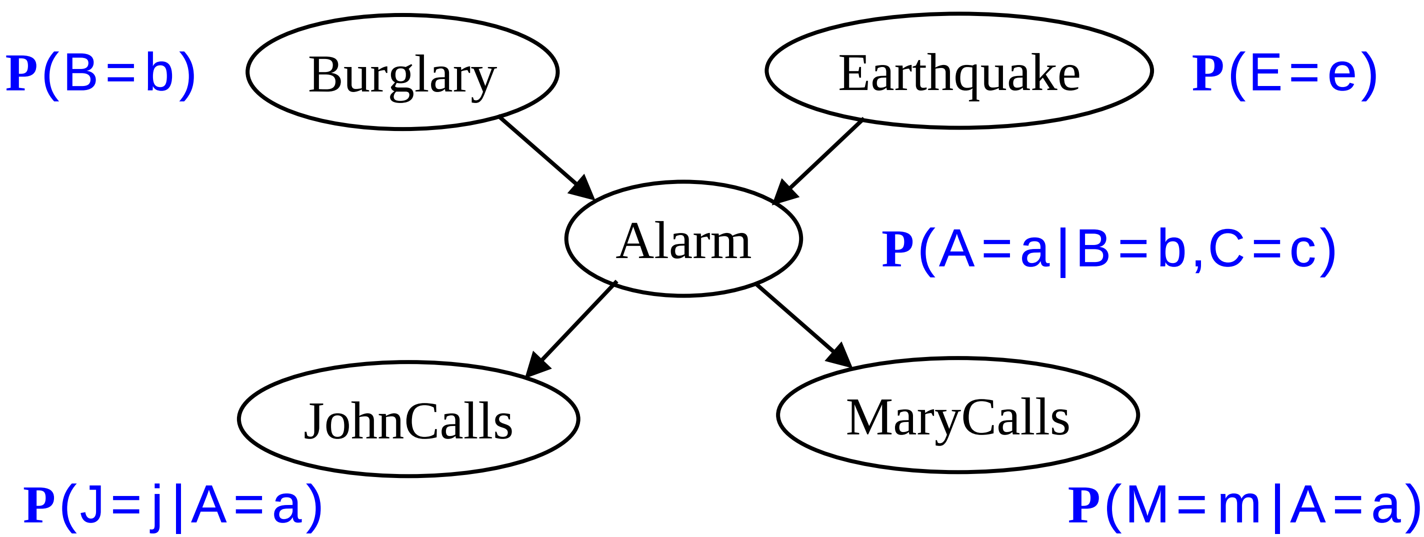
- Each node X_i has a set of parents $Pa(X_i)$ (arrows points to X_i)



An Alarm Network Example

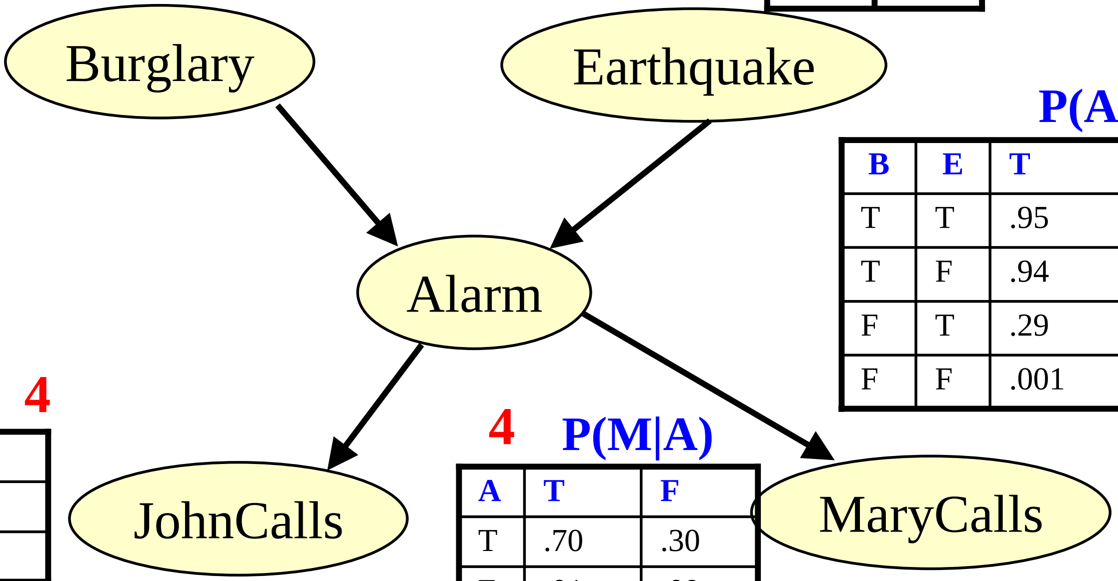
BN: Directed acyclic graph (DAG)

- Nodes - random variables
- Edges - direct "causal influences" : the chance of Alarm being is influenced by Earthquake, the chance of John calling is affected by the Alarm



Conditional Probability Tables

- Each node has a **conditional probability table** (**CPT**) that gives the probability of each of its values given every possible combination of values for its parents (conditioning case).



P(B) **2**

T	F
.001	.999

P(E) **2**

T	F
.002	.998

P(A|B,E) **8**

B	E	T	F
T	T	.95	.05
T	F	.94	.06
F	T	.29	.71
F	F	.001	.999

P(J|A) **4**

A	T	F
T	.90	.10
F	.05	.05

4 **P(M|A)**

A	T	F
T	.70	.30
F	.01	.09

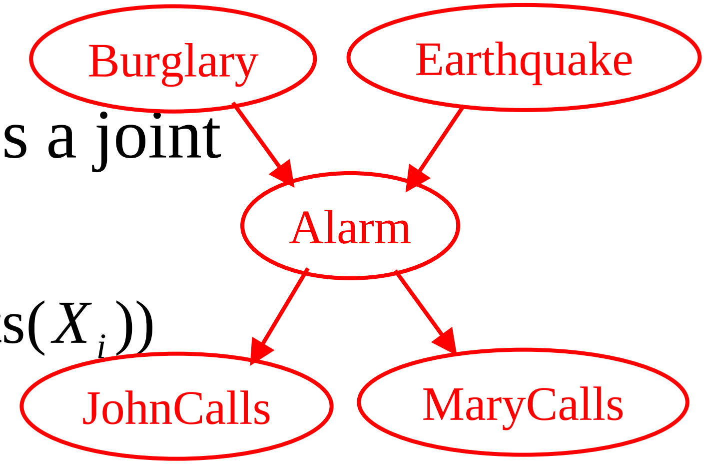
CPT Comments

- **Setting: 5 binary (True, False) variables**
- **Number of parameters of the full joint:** $2^5 - 1 = 31$

$$P(x_1, x_2, \ldots x_n) = \prod_{i=1}^{n} P(x_i \mid pa_i)$$

- **Number of parameters of the BN:20 (see CPT)**
- For complete joint probability distribution(JPD) over all variables, the number is $O(2^n)$ if each node has 2 state (such as true or false)
- So, BN: from $O(2^n)$ to $O(n*2^k)$, where, k is the maximal number of parents of a node

Joint Distributions



- A Bayesian Network implicitly defines a joint distribution.

$$P(x_1, x_2, \ldots x_n) = \prod_{i=1}^{n} P(x_i \mid \text{Parents}(X_i))$$

- Example

$$P(J \wedge M \wedge A \wedge \neg B \wedge \neg E)$$

$$= P(J \mid A) P(M \mid A) P(A \mid \neg B \wedge \neg E) P(\neg B) P(\neg E)$$

$$= 0.9 \times 0.7 \times 0.001 \times 0.999 \times 0.998 = 0.00062$$

- Therefore an **inefficient approach** to inference is:
  - 1) Compute the joint distribution using this equation.
  - 2) Compute any desired conditional probability using the joint distribution.

Outline

- ✓ Overview
- ➢ **Inference in GM**
- Exact inference

Inference in GM

- Input:
  - Evidence :Observed values of some nodes;
- Goal:
  - Compute the posterior distributions of one or more subsets of other nodes

BN Inference

- Problem Definition
  - Given
    - Bayesian network with specified CPTs
    - Observed values for some nodes in network(**evidence**)
  - Return: inferred values (posterior distribution of some other nodes)
- Implementation
  - Bayesian network contains all information needed for this inference
  - can succeed in many cases

Various Inference tasks

- Diagnostic task. (from effect to cause: Diagnostic Inference)
- Prediction task. (causal Inference : from cause to effect)
- Other probabilistic queries (queries on joint distributions).

Query:Inference

This query is useful in many cases:

- **Prediction(casual inference)**: what is the probability of an outcome given the starting condition.
  - Target is a **descendent** of the evidence in Bayesian Net
- **Diagnosis(diagnosis inference)**: what is the probability of disease/fault given symptoms.
  - Target is an **ancestor** of the evidence
- the direction between variables does not restrict the directions of the queries:
  - Probabilistic inference can combine evidence form all parts of the network

Inference in Belief Networks

- Find $P(Q=q|\boldsymbol{E}=\boldsymbol{e})$
  - Q the query variable
  - $\boldsymbol{E}$ set of evidence variables

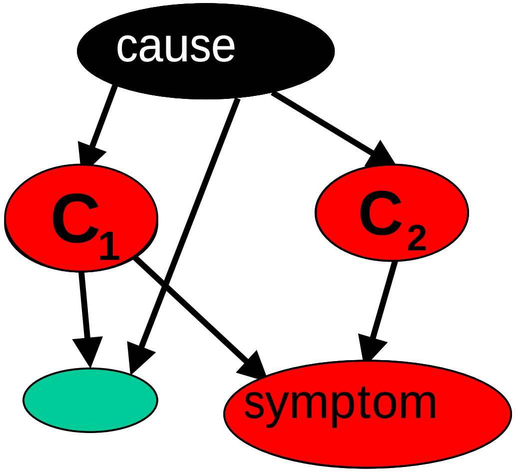
$$P(q \mid \mathbf{e}) = \frac{P(q, \mathbf{e})}{P(\mathbf{e})}$$

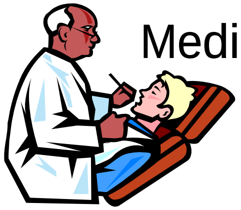
$X_1, \ldots, X_n$ are network variables except Q, $\boldsymbol{E}$

$$P(q, \mathbf{e}) = \sum_{x_1, x_2, \ldots x_n} P(q, \mathbf{e}, x_1, x_2, \ldots x_n)$$

Different Application

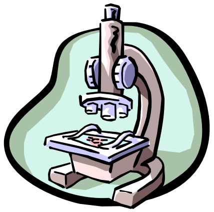
- Diagnosis: P(cause|symptom)=?
- Prediction: P(symptom|cause)=?
- Classification: $\max_{class} P(class|data)$
- ……



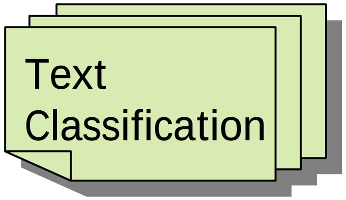


Medicine

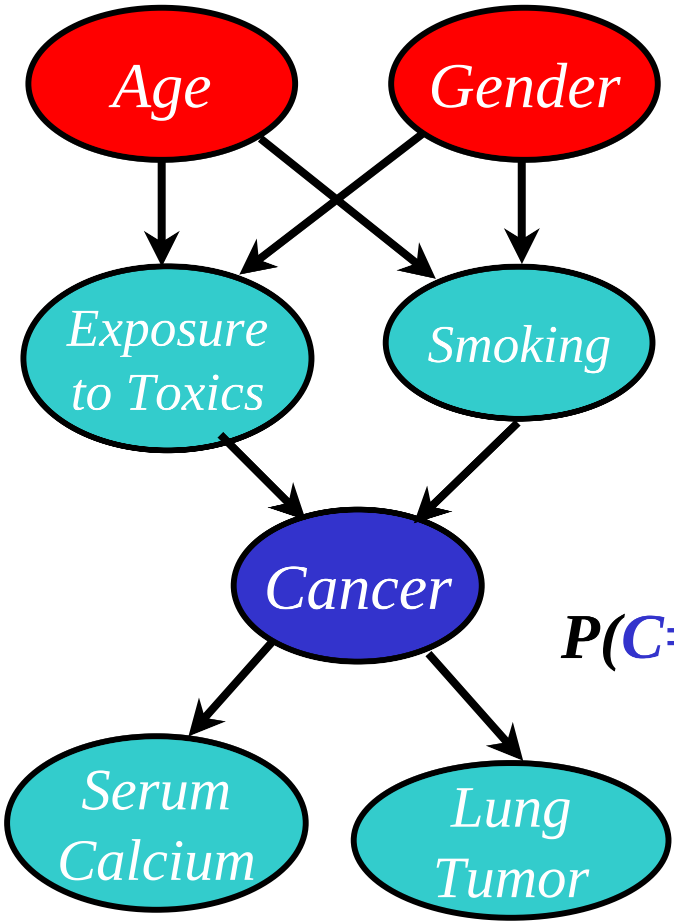
Speech recognition



Bio-informatics



Predictive(Causal) Inference

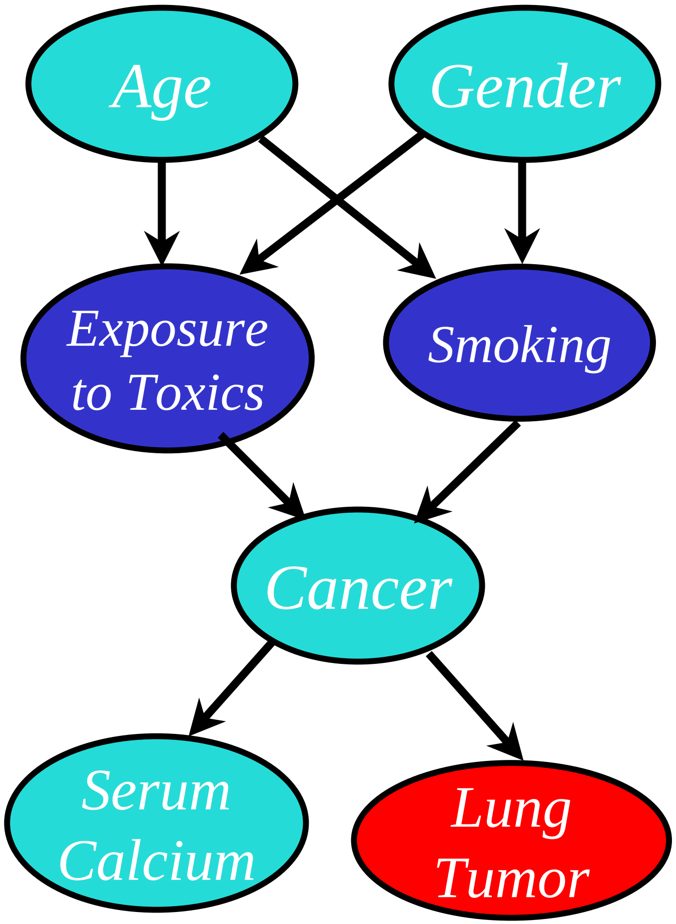


Observed data: elder+males;

Goal: find the probability of malignant cancer?

P(C=malignant | Age>60, Gender= male)

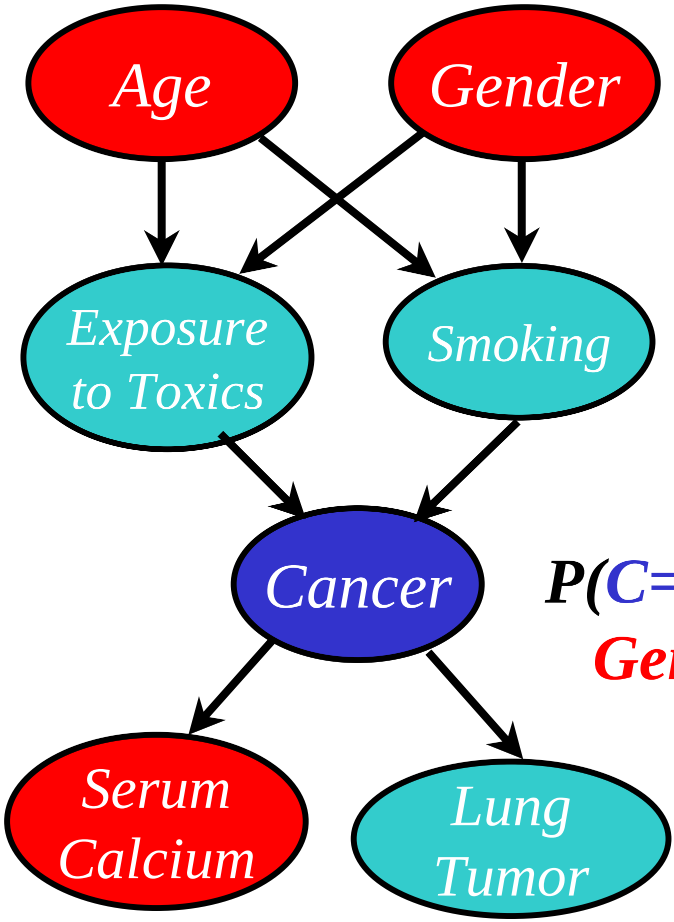
Diagnosis



Observed data: lung tumor,

Goal: find the probability of heavy smoking and of exposure to toxics both go up.

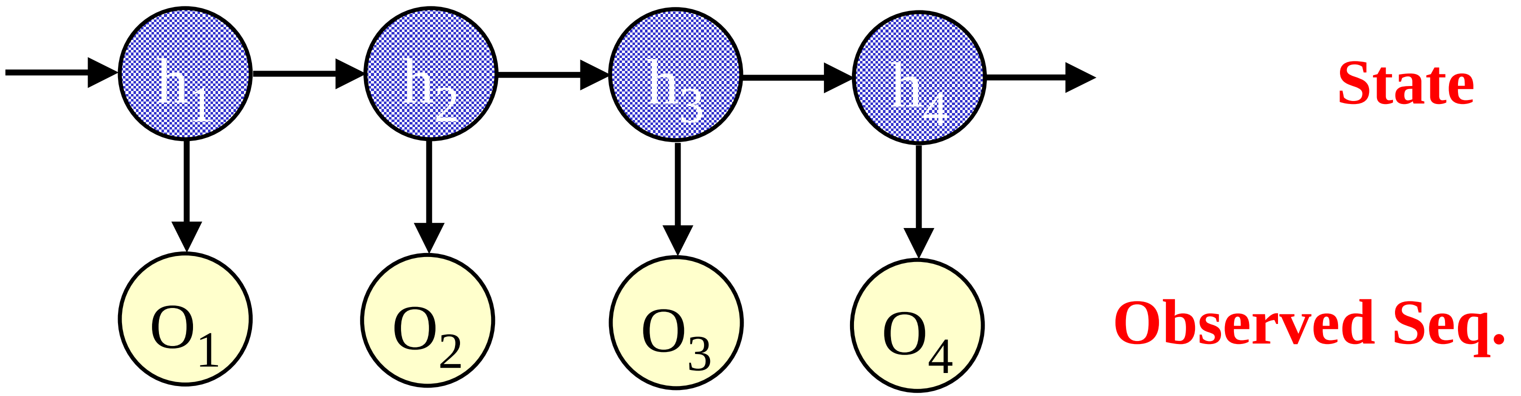
Combined



Observed data:
elderly+male+Serum Calcium;
Goal: the probability of malignant cancer?

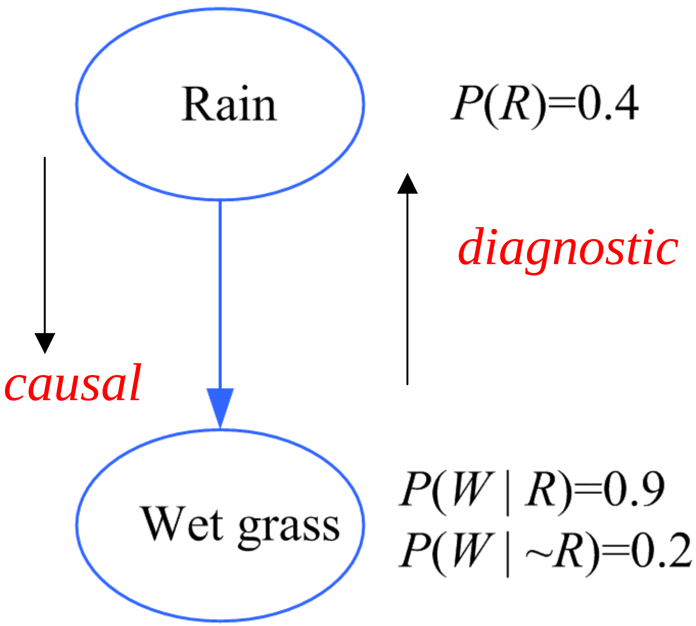
P(C=malignant | Age>60, Gender= male, Serum Calcium = high)

NLP Seq. Tag



- $P(h_1 h_2 \ldots h_n | o_1 o_2 \ldots o_n)$

Causal vs Diagnostic Inference



$$P(R|W) = \frac{P(R,W)}{P(W)} = \frac{P(W|R)P(R)}{P(W)}$$

$$= \frac{P(W|R)P(R)}{P(W|R)P(R) + P(W|\sim R)P(\sim R)}$$

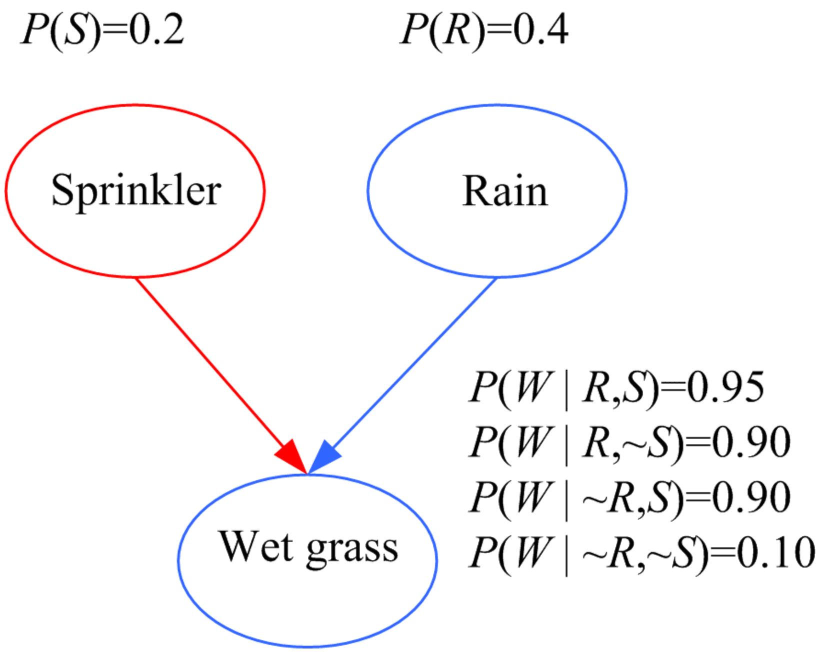
$$= \frac{0.9 \times 0.4}{0.9 \times 0.4 + 0.2 \times 0.6} = 0.75$$

Diagnostic inference:

Knowing that the **grass is wet**,

what is the probability that **rain** *is the cause?*

Causal vs Diagnostic Inference



Causal inference: *If the* **sprinkler is on**, *what is the probability that the* **grass is wet**?

$$P(W|S) = P(W|R,S)\,P(R|S) + P(W|{\sim}R,S)\,P({\sim}R|S)$$

$$= P(W|R,S)\,P(R) + P(W|{\sim}R,S)\,P({\sim}R)$$

$$= 0.95 \times 0.4 + 0.9 \times 0.6 = 0.92$$

Diagnostic inference: *If the* **grass is wet**, *what is the probability that the* **sprinkler is on**? $P(S|W) = 0.184/(0.184+0.336) = 0.35 > 0.2 = P(S)$

$P(S|R,W) = 0.21$ ***Explaining away:*** **Knowing that it has rained decreases the probability that the sprinkler is on.**

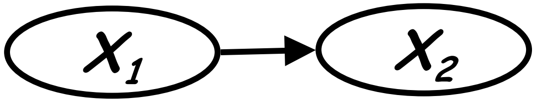
Outline

- ✓ Overview
- ✓ Inference in GM
- ➢ **Exact inference**

Approaches to exact inference

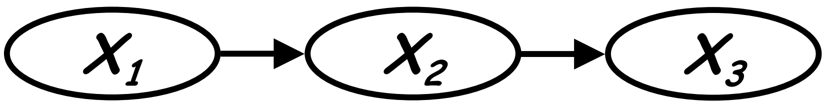
- **Variable elimination**
- Join tree algorithms

Inference in Simple Chains



Marginal probability

$$P(x_2) = \sum_{x_1} P(x_1, x_2) = \sum_{x_1} P(x_1) P(x_2 \mid x_1)$$

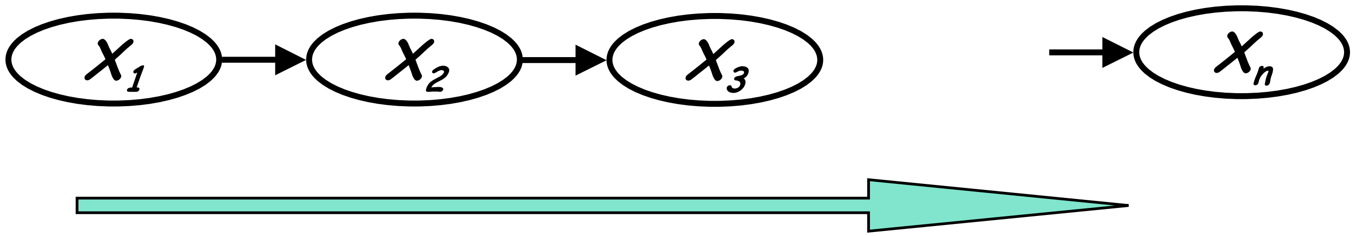


$$P(x_3) = \sum_{x_2} P(x_2, x_3) = \sum_{x_2} P(x_2) P(x_3 \mid x_2)$$

So, $$P(x_3) = \sum_{x_2} \sum_{x_1} P(x_2 \mid x_1) P(x_1) P(x_3 \mid x_2)$$

$$= \sum_{x_2} P(x_3 \mid x_2) \sum_{x_1} P(x_2 \mid x_1) P(x_1)$$

Forward Inference in Chains



$$P(x_{i+1}) = \sum_{x_i} P(x_i) P(x_{i+1} \mid x_i)$$

Marginal probability

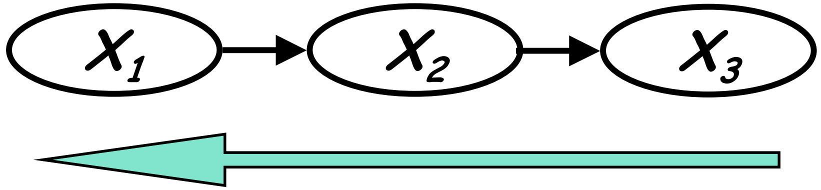
So,

$$P(x_k) = \sum_{x_{k-1}} \sum_{x_{k-2}} \dots \sum_{x_1} P(x_k \mid x_{k-1}) P(x_{k-1} \mid x_{k-2}) \dots P(x_2 \mid x_1) P(x_1)$$

Complexity for $P(X_n)$:

- $O(|Val(X_1)|*|Val(X_2)|*|Val(X_n)|)$ for all X_n

Backward Inference in Chains



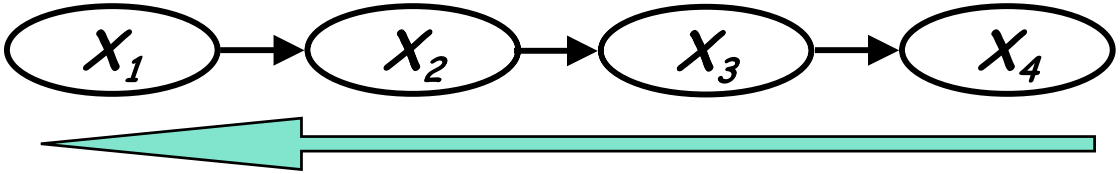
- Observed value: $X_3 = x_3$

$$P(x_3 \mid x_1) = \frac{P(x_1, x_3)}{P(x_1)} = \frac{\sum_{x_2} P(x_1, x_2, x_3)}{P(x_1)}$$

$$= \frac{\sum_{x_2} P(x_1) P(x_2 \mid x_1) P(x_3 \mid x_1)}{P(x_1)}$$

$$= \sum_{x_2} P(x_2 \mid x_1) P(x_3 \mid x_2)$$

Backward Inference in Chains



- Observed value: $X_4 = x_4$

$$P(x_4 \mid x_2) = \sum_{x_3} P(x_3, x_4 \mid x_2) = \sum_{x_3} P(x_3 \mid x_2) P(x_4 \mid x_3)$$

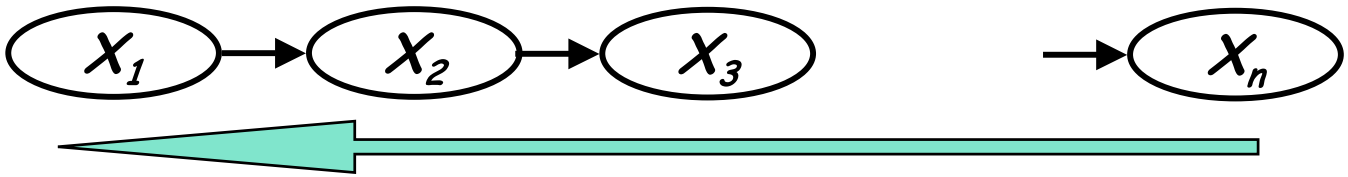
$$P(x_4 \mid x_1) = \frac{P(x_1, x_4)}{P(x_1)} = \frac{\sum_{x_2, x_3} P(x_1, x_2, x_3, x_4)}{P(x_1)}$$

$$= \sum_{x_2, x_3} P(x_2 \mid x_1) P(x_3 \mid x_2) P(x_4 \mid x_3)$$

$$= \sum_{x_2} P(x_2 \mid x_1) \sum_{x_3} P(x_3 \mid x_2) P(x_4 \mid x_3)$$

$$= \sum_{x_2} P(x_2 \mid x_1) P(x_4 \mid x_2)$$

Backward Inference in Chains



- Generally, for observed value: $X_n = x_n$

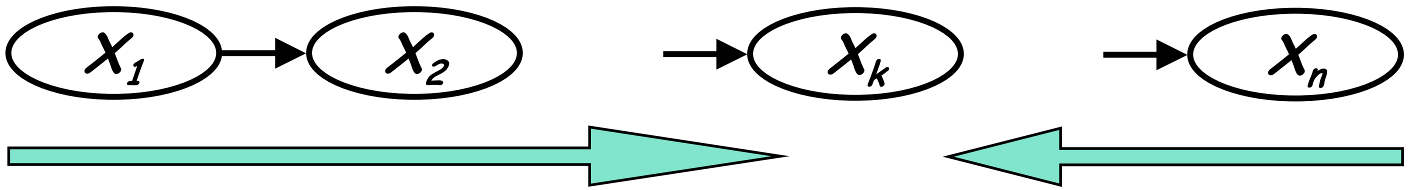
From: $P(x_n \mid x_{i+1}) = \sum_{x_{i+2}} P(x_{i+2} \mid x_{i+1}) P(x_n \mid x_{i+2})$

We can get that:

$$P(x_n \mid x_i) = \sum_{x_{i+1}} P(x_{i+1} \mid x_i) P(x_n \mid x_{i+1})$$

How to compute $P(x_n/x_{n-1}), P(x_n/x_{n-2}), \ldots P(x_n/x_i)$ iteratively

Inference in Simple Chains



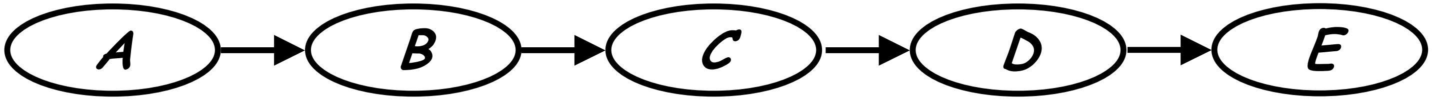
- Observed value: $X_n = x_n$
- How to find $P(X_k, x_n)$?

$$P(x_k, x_n) = P(x_k)P(x_n \mid x_k)$$

- Compute $P(X_k)$ by **forward iterations**
- Compute $P(x_n / X_k)$ by **backward iterations**

- How to find $P(X_k/x_n)$? $P(X_k/x_n) = P(X_k, x_n) / P(x_n)$

Elimination in Chains: Forward



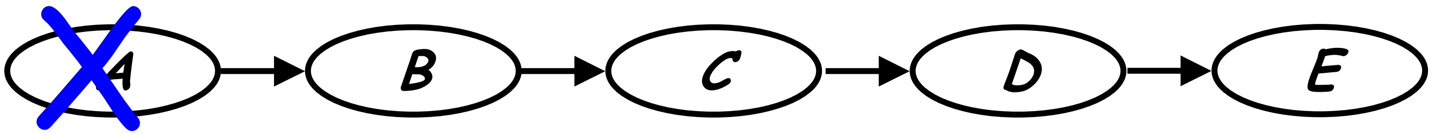
$$P(e) = \sum_d \sum_c \sum_b \sum_a P(a,b,c,d,e)$$
$$= \sum_d \sum_c \sum_b \sum_a P(a)P(b \mid a)P(c \mid b)P(d \mid c)P(e \mid d)$$

- By chain decomposition, we get

$$P(e) = \sum_d \sum_c \sum_b \boxed{\sum_a P(a)P(b \mid a)} P(c \mid b)P(d \mid c)P(e \mid d)$$
$$= \sum_d \sum_c \sum_b P(c \mid b)P(d \mid c)P(e \mid d) \boxed{\sum_a P(a)P(b \mid a)}$$

P(b)

Elimination in Chains : Forward

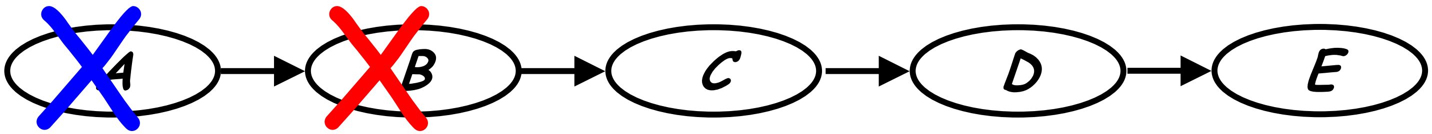


$$P(e)=\sum_d\sum_c\sum_b P(c\mid b)P(d\mid c)P(e\mid d)\sum_a P(a)P(b\mid a)$$

$$=\sum_d\sum_c\sum_b P(c\mid b)P(d\mid c)P(e\mid d)p(b)$$

- This summation, is exactly the first step in the **forward** iteration.

Elimination in Chains : Forward



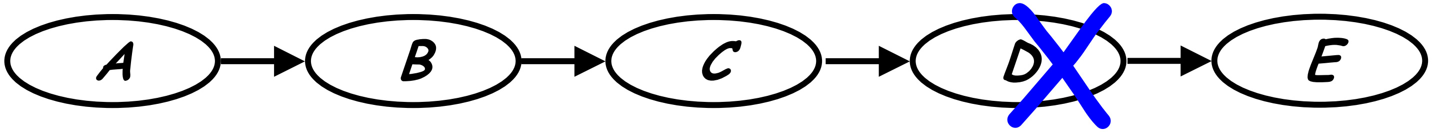
- Rearranging and then summing again:

$$P(e) = \sum_d \sum_c \sum_b P(c \mid b) P(d \mid c) P(e \mid d) p(b)$$

$$= \sum_d \sum_c P(d \mid c) P(e \mid d) \sum_b P(c \mid b) p(b)$$

$$= \sum_d \sum_c P(d \mid c) P(e \mid d) p(c)$$

Elimination with evidence:Backward



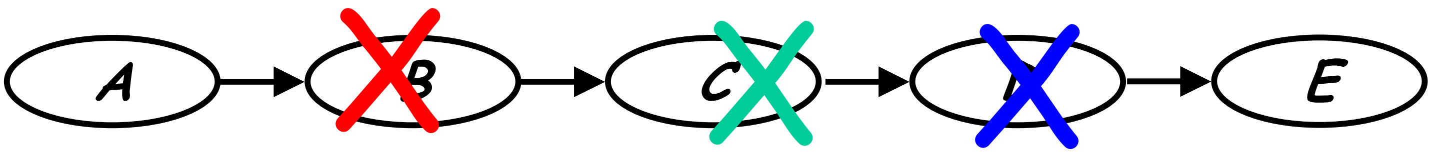
- Observed value: $A=a,\ E=e,$

$$P(a,e)=\sum_b\sum_c\sum_d P(a,b,c,d,e)$$
$$=\sum_b\sum_c\sum_d P(a)P(b\mid a)P(c\mid b)P(d\mid c)P(e\mid d)$$

- Eliminating d,

$$P(a,e)=\sum_b\sum_c\sum_d P(a)P(b\mid a)P(c\mid b)P(d\mid c)P(e\mid d)$$
$$=\sum_b\sum_c P(a)P(b\mid a)P(c\mid b)\sum_d P(d\mid c)P(e\mid d)$$
$$=\sum_b\sum_c P(a)P(b\mid a)P(c\mid b)P(e\mid c)$$

Elimination with evidence:Backward



- Observed value: $A=a,\ E=e,$

$$P(a,e)=\sum_b\sum_c P(a)P(b\mid a)P(c\mid b)P(e\mid c)$$

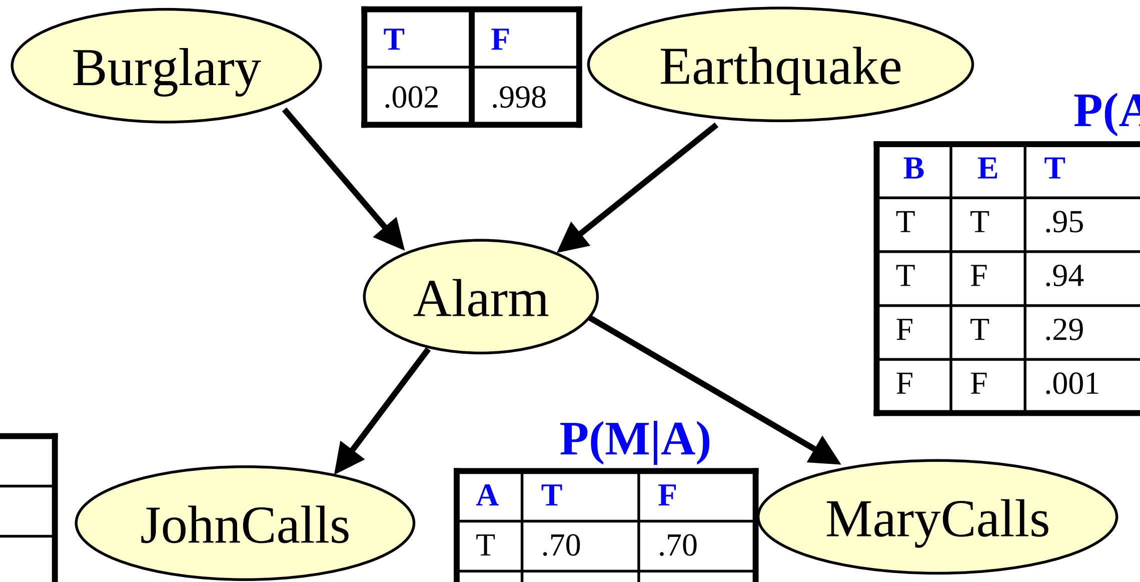
$$=\sum_b P(a)P(b\mid a)\sum_c P(c\mid b)P(e\mid c)$$

$$=\sum_b P(a)P(b\mid a)p(e\mid b)$$

$$=P(a)\sum_b P(b\mid a)P(e\mid b)$$

$$=P(a)P(e\mid a)$$

Complex Inference: back to JohnCall



P(B)

T	F
.001	.999

P(E)

T	F
.002	.998

P(A|B,E)

B	E	T	F
T	T	.95	.05
T	F	.94	.06
F	T	.29	.71
F	F	.001	.999

P(J|A)

A	T	F
T	.90	.90
F	.05	.05

P(M|A)

A	T	F
T	.70	.70
F	.01	.01

- **Computing:** P(J=T)

Blind approach

- Sum out all variables from the full joint
- Express the joint distribution as a product of conditionals

$$P(J=T)$$

$$=\sum_{b\in\{T,F\}}\sum_{e\in\{T,F\}}\sum_{a\in\{T,F\}}\sum_{m\in\{T,F\}}P(B=b,E=e,A=a,J=T,M=m)$$

$$=\sum_{b\in\{T,F\}}\sum_{e\in\{T,F\}}\sum_{a\in\{T,F\}}\sum_{m\in\{T,F\}}P(J=T|A=a)P(M=m|A=a)P(A=a|B=b,E=e)P(B=b)P(E=e)$$

- **Computational Cost**
  - Number of additions: **15**
  - Number of products: **16*4=64**

B	E	A	M
F	**F**	**F**	**F**
F	**F**	**F**	**T**
…	…	…	…
T	**T**	**T**	**F**
T	**T**	**T**	**T**

16 groups

Interleave sums and products

- Combines sums and product in a smart way (multiplications by constants can be taken out of the sum)

$$P(J=\mathrm{T}) = \sum_{b\in\{T,F\}}\sum_{e\in\{T,F\}}\sum_{a\in\{T,F\}}\sum_{m\in\{T,F\}} P(B=b,E=e,A=a,J=\mathrm{T},M=m)$$

$$= \sum_{b\in\{T,F\}}\sum_{e\in\{T,F\}}\sum_{a\in\{T,F\}}\sum_{m\in\{T,F\}} P(J=\mathrm{T}\mid A=a)P(M=m\mid A=a)P(A=a\mid B=b,E=e)P(B=b)P(E=e)$$

$$= \sum_{b\in\{T,F\}}\sum_{a\in\{T,F\}}\sum_{m\in\{T,F\}} P(J=\mathrm{T}\mid A=a)P(M=m\mid A=a)P(B=b)\left[\sum_{e\in\{T,F\}} P(A=a\mid B=b,E=e)P(E=e)\right]$$

$$= \sum_{a\in\{T,F\}}\sum_{m\in\{T,F\}} P(J=\mathrm{T}\mid A=a)P(M=m\mid A=a)\left[\sum_{b\in\{T,F\}} P(B=b)\left[\sum_{e\in\{T,F\}} P(A=a\mid B=b,E=e)P(E=e)\right]\right]$$

$$= \sum_{a\in\{T,F\}} P(J=\mathrm{T}\mid A=a)\left[\sum_{m\in\{T,F\}} P(M=m\mid A=a)\right]\left[\sum_{b\in\{T,F\}} P(B=b)\left[\sum_{e\in\{T,F\}} P(A=a\mid B=b,E=e)P(E=e)\right]\right]$$

- **Computational Cost**
  - Number of additions: 1+ 2*(1)+2*(1+2*(1))=**9**
  - Number of products: 2*(2+2*(1)+2*(2*(1)))=**16**

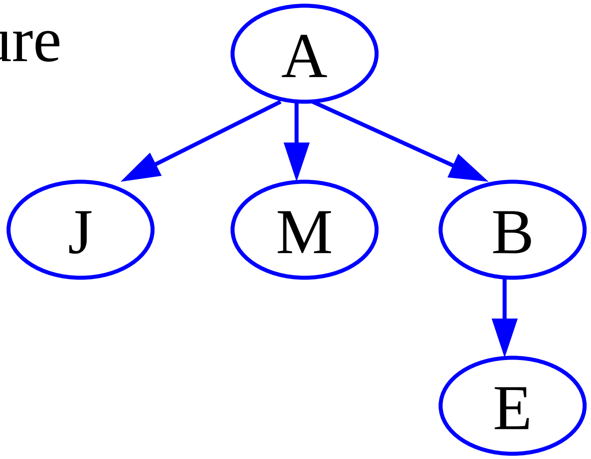
- General technique: **Variable elimination**

Variable Elimination

$$P(True) = 1 = \sum_{b\in\{T,F\}} \sum_{e\in\{T,F\}} \sum_{j\in\{T,F\}} \sum_{m\in\{T,F\}} P(B=b, E=e, A=a, J=j, M=m)$$

$$= \sum_{a\in\{T,F\}} [\underbrace{\sum_{j\in\{T,F\}} P(J=j \mid A=a)]}_{f_J(a)} \underbrace{[\sum_{m\in\{T,F\}} P(M=m \mid A=a)]}_{f_M(a)} \underbrace{[\sum_{b\in\{T,F\}} P(B=b) \underbrace{[\sum_{e\in\{T,F\}} P(A=a \mid B=b, E=e) P(E=e)]}_{f_E(a,b)}]}_{f_B(a)}$$

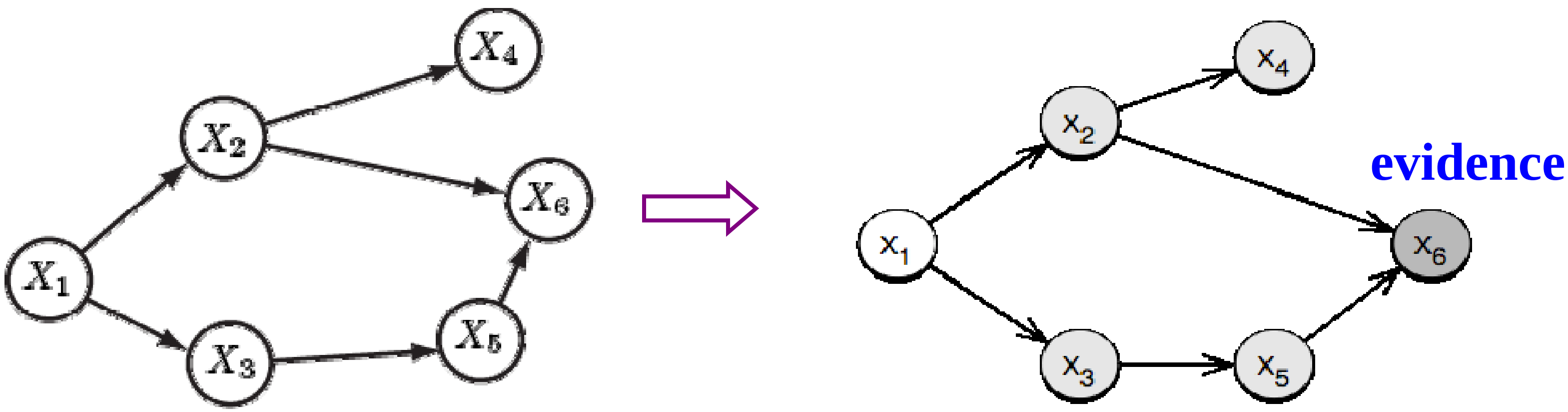
- Results cashed in the tree structure



Basic idea of variable elimination

- **Basic Idea**
  - **Sort**: *impose* an ordering over the variables, with the query variable coming *last* ;
  - **Factorize**: *maintain* a list of “factors”, which depend on given variables(parents);
  - **Marginalize**: *sum* over the variables in the order in which they appear in the list;
  - **Temporary Results**: *memorize* the result of intermediate computations
- **Method**: *dynamic programming*

Another Complex Example



Want to calculate:

$$p(x_1|\underline{x}_6) = \frac{p(x_1,\underline{x}_6)}{\sum_{x_1} p(x_1,\underline{x}_6)}$$

marginalize (sum) out **X**$_{2,3,4,5}$

$$
\begin{aligned}
p(x_1,\underline{x}_6) &= \sum_{x_2}\sum_{x_3}\sum_{x_4}\sum_{x_5} p(x_1,x_2,x_3,x_4,x_5,\underline{x}_6) \\
&= \sum_{x_2}\sum_{x_3}\sum_{x_4}\sum_{x_5} p(x_1)p(x_2|x_1)p(x_3|x_1)p(x_4|x_2)p(x_5|x_3)p(\underline{x}_6|x_2,x_5) \\
&= p(x_1)\sum_{x_2} p(x_2|x_1)\sum_{x_3} p(x_3|x_1)\sum_{x_4} p(x_4|x_2)\sum_{x_5} p(x_5|x_3)p(\underline{x}_6|x_2,x_5)
\end{aligned}
$$

Inference

$$p(x_1,\underline{x}_6)=p(x_1)\sum_{x_2}p(x_2|x_1)\sum_{x_3}p(x_3|x_1)\sum_{x_4}p(x_4|x_2)\sum_{x_5}p(x_5|x_3)p(\underline{x}_6|x_2,x_5)$$
$$=p(x_1)\sum_{x_2}p(x_2|x_1)\sum_{x_3}p(x_3|x_1)\sum_{x_4}p(x_4|x_2)m_5(x_2,x_3)$$

We define: $m_5(x_2,x_3)\triangleq\sum_{x_5}p(x_5|x_3)p(\underline{x}_6|x_2,x_5)$

$$p(x_1,\underline{x}_6)=p(x_1)\sum_{x_2}p(x_2|x_1)\sum_{x_3}p(x_3|x_1)m_5(x_2,x_3)\sum_{x_4}p(x_4|x_2)$$
$$=p(x_1)\sum_{x_2}p(x_2|x_1)\sum_{x_3}p(x_3|x_1)m_4(x_2)m_5(x_2,x_3)$$

Where, $m_4(x_2)\triangleq\sum_{x_4}p(x_4|x_2)$

$$p(x_1,\underline{x}_6)=\cdots=p(x_1)m_2(x_1)$$

Where, $m_3(x_1, x_2)$ and $m_2(x_1)$ are introduced.

Inference process

Finally, $$p(x_1|x_6) = \frac{p(x_1)m_2(x_1)}{\sum_{x_1} p(x_1)m_2(x_1)}$$

- The m_2, m_3, m_4, m_5 values store the partial sums, which can each be computed easily.
- Each time we introduce an 'm' function, we are *eliminating* variables from the equation.
- Calculation of probabilities after observing the values of some variables and marginalising out others.

✓ **Disadvantage:** The algorithm needs to be run many times to calculate the marginals for all nodes in the graph. **The cost is also big !!**

Approaches to exact inference

- ✓Variable elimination
- **Join tree algorithms**

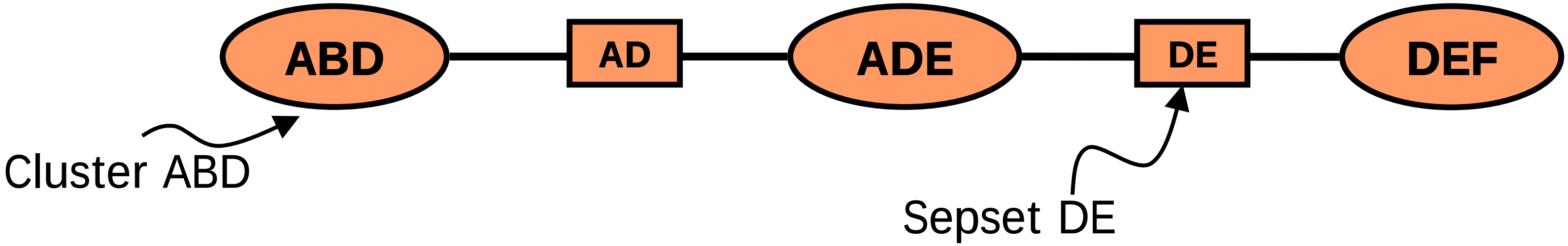
Junction Tree Propagation

(Lauritzen and Spiegelhalter, 1988)

The general idea is that the propagation of evidence through the network can be carried out more efficiently by representing the joint probability distribution on an *undirected* graph called the Junction tree (or Join tree).

Properties of Junction Tree

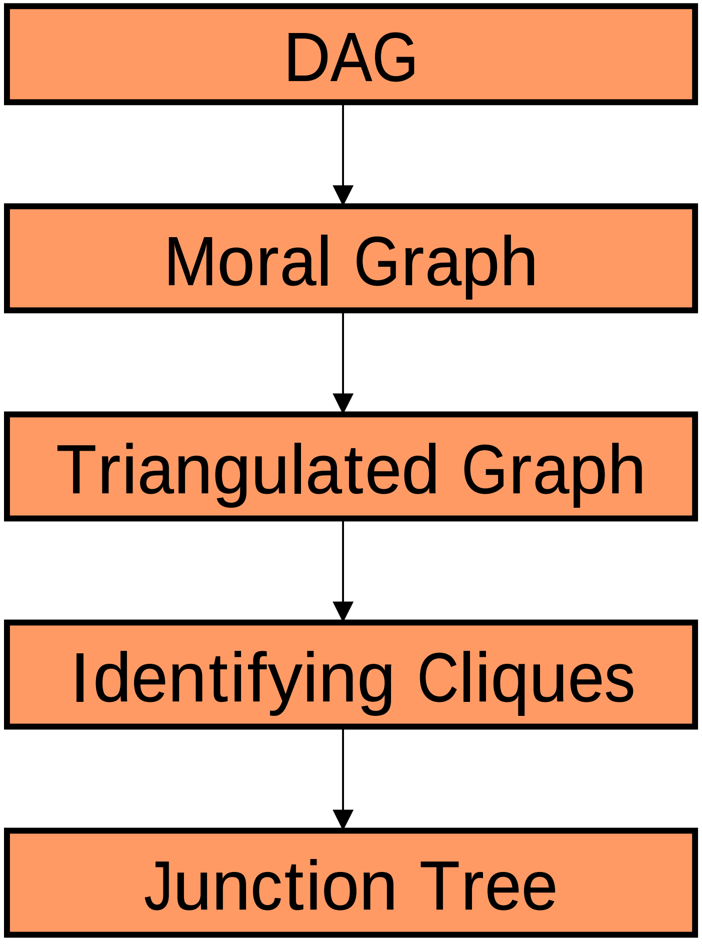
- An undirected tree
- Each node is a cluster (nonempty set) of variables
- Running intersection property:
  - Given two clusters X and Y, all clusters on the path between X and Y contain $X \cap Y$
- Separator sets (sepsets):
  - Intersection of the adjacent cluster



Junction Tree

- Why junction tree?
  - Variable elimination is inefficient if the undirected graph underlying the Bayesian Nets contains cycles
  - We can avoid cycles if we turn highly-interconnected subsets of the nodes into “supernodes” → cluster
- Objective
  - Compute $P(V = v \mid \mathbf{E} = \mathbf{e})$
    - v is a value of a variable V and $\mathbf{e}$ is evidence for a set of variable $\mathbf{E}$

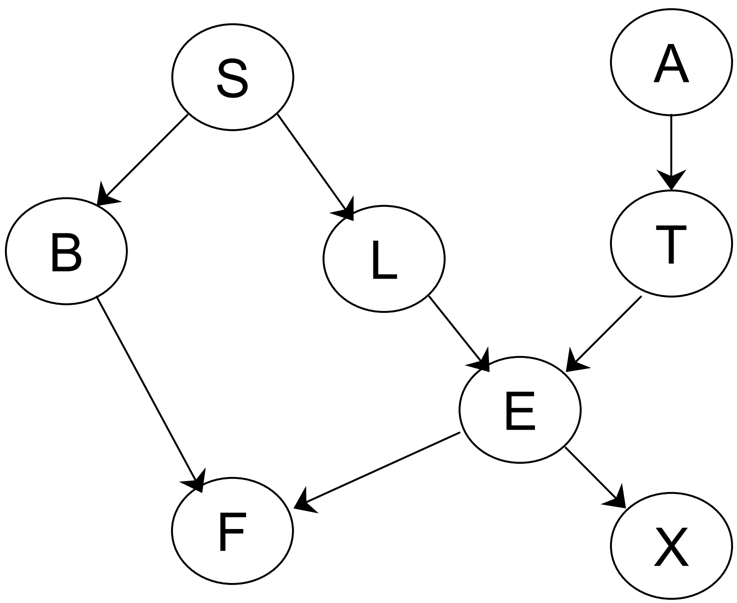
Building Junction Trees



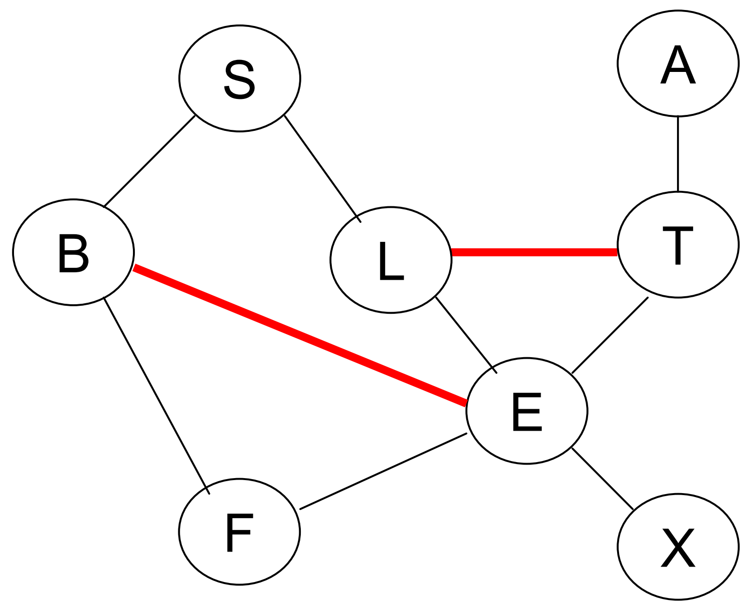
Constructing the Junction Tree

Step 1. Form the moral graph from the DAG

Consider the Bayesian network



DAG

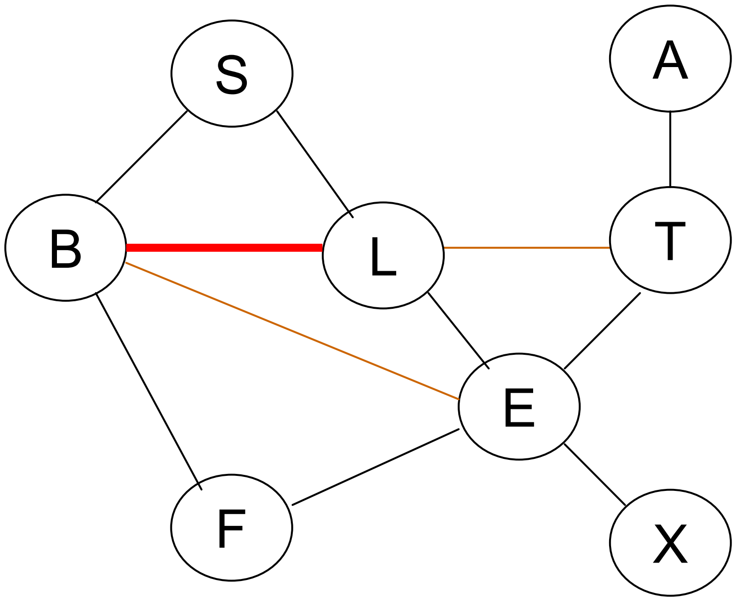


Moral Graph – marry parents and remove arrows

Constructing the Junction Tree

Step 2. **Triangulate** the moral graph

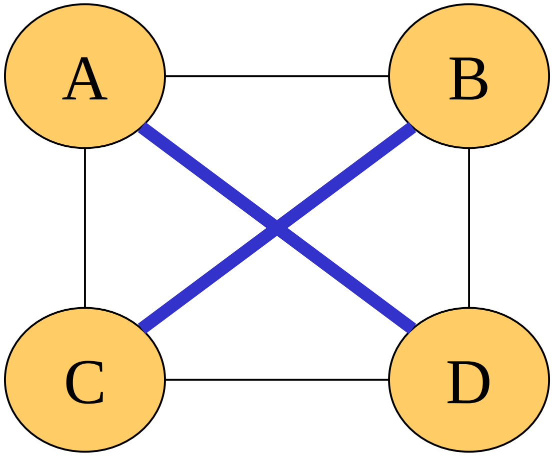
An undirected graph is triangulated if every cycle of length greater than 3 contains an edge to connects two nonadjacent nodes



Some Notes on Triangulation

Is Triangulation unique? No.

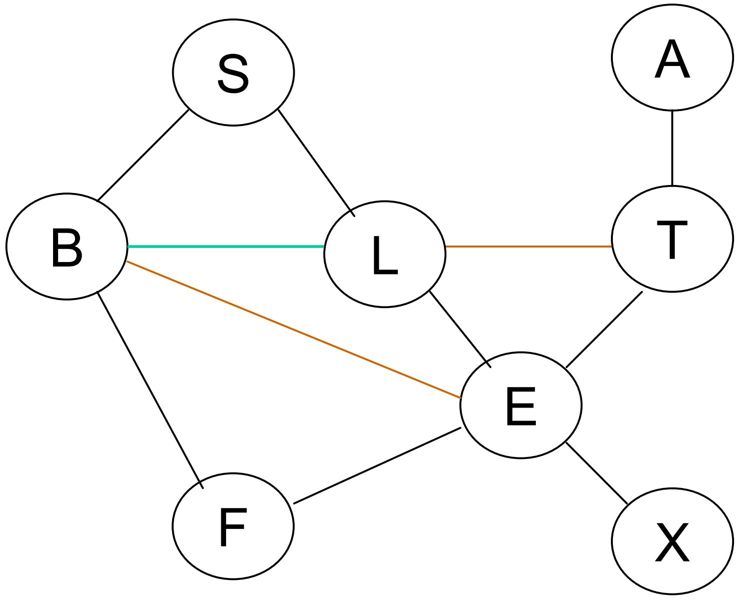
Could find the best triangulation? Sadly, that's NP hard.



Constructing the Junction Tree

Step 3. Identify the Cliques

A clique is a subset of nodes in which each pair is connected and *maximal*.



$\Rightarrow$

Cliques

{B,S,L}
{B,L,E}
{B,E,F}
{L,E,T}
{A,T}
{E,X}

Constructing the Junction Tree

Step 4. Build Junction Tree

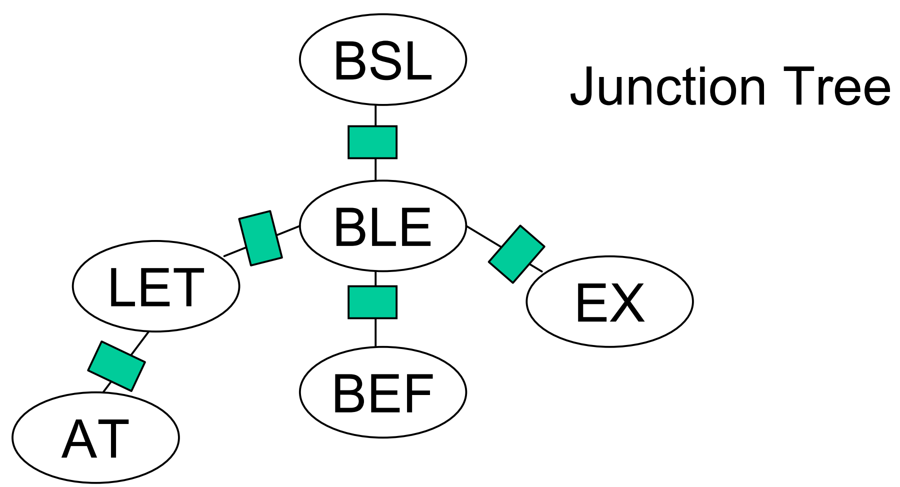
The cliques should be ordered $(C_1, C_2, \ldots, C_k)$ so they possess the running intersection property: for all $1 < j \leq k$, there is an $i < j$ such that $C_j \cap (C_1 \cup \ldots \cup C_{j-1}) \subseteq C_i$.

To build the junction tree choose one such i for each j and add an edge between C_j and C_i.

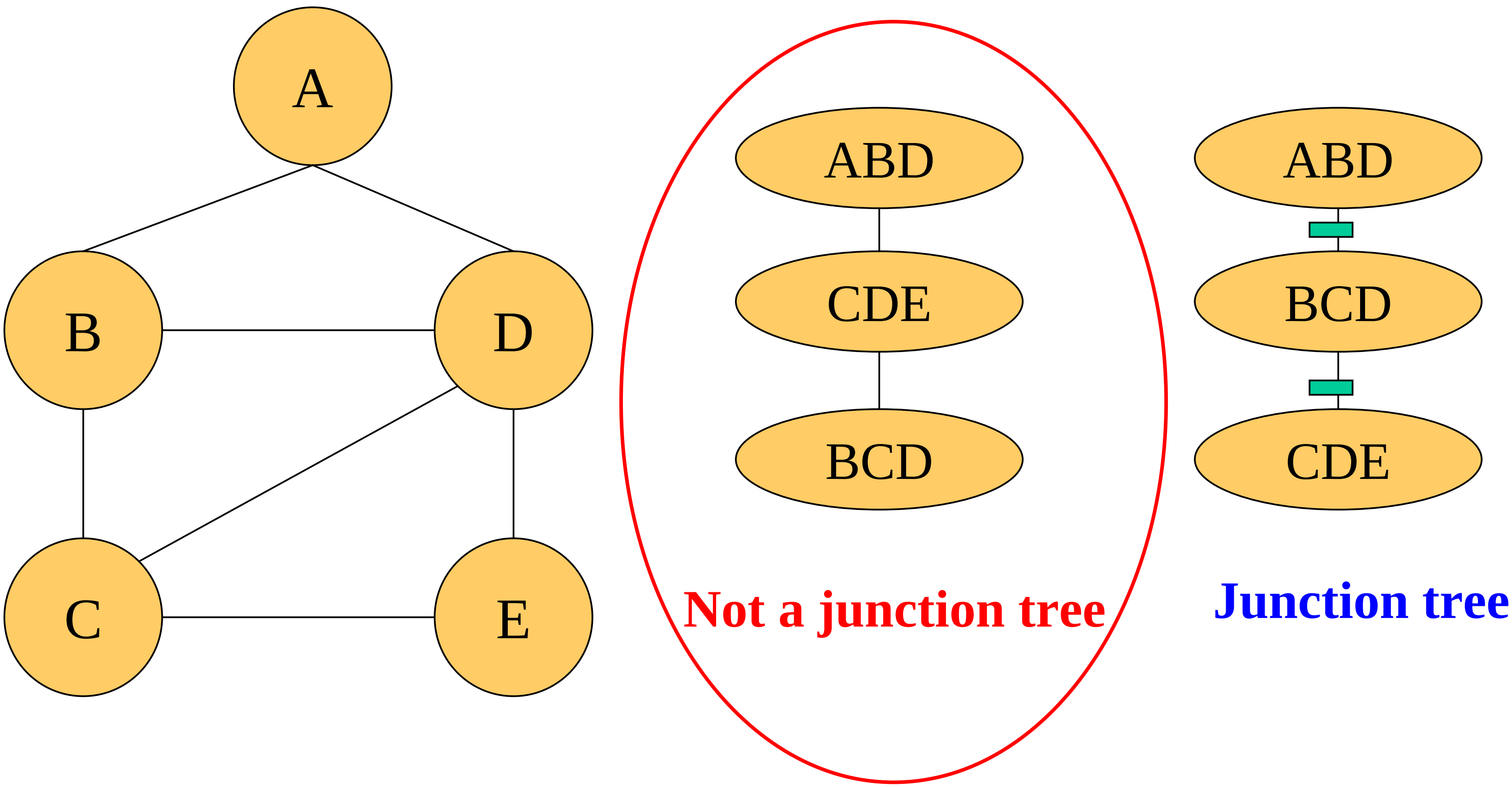
Cliques

{B,S,L}
{B,L,E}
{B,E,F}
{L,E,T}
{A,T}
{E,X}

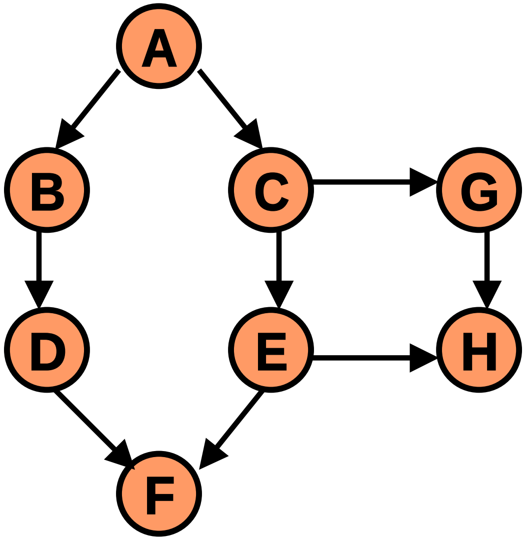
$\Rightarrow$



An Example


A
B
D
C
E
ABD
CDE
BCD
Not a junction tree
ABD
BCD
CDE
Junction tree

How to Build Junction Tree?



Potential Representation

- Potentials: $\Psi : \mathbf{X} \rightarrow R^{+} \cup \{0\}$

  Denoted by $\psi_{\mathrm{x}}(\mathrm{x})$ or ψ_{x}

- Marginalization :

  X $\subseteq$ Y, the marginalization of ψ_{Y} into X:

$$\psi_{\mathrm{X}} = \sum_{Y \setminus X} \psi_{\mathrm{Y}}$$

Potential Representation

Constraint: The joint probability distribution of DAG equals to the junction tree:

$$P(U) = \frac{\prod_{c \in C} \psi_c(x_c)}{\prod_{s \in S} \psi_s(x_s)}$$

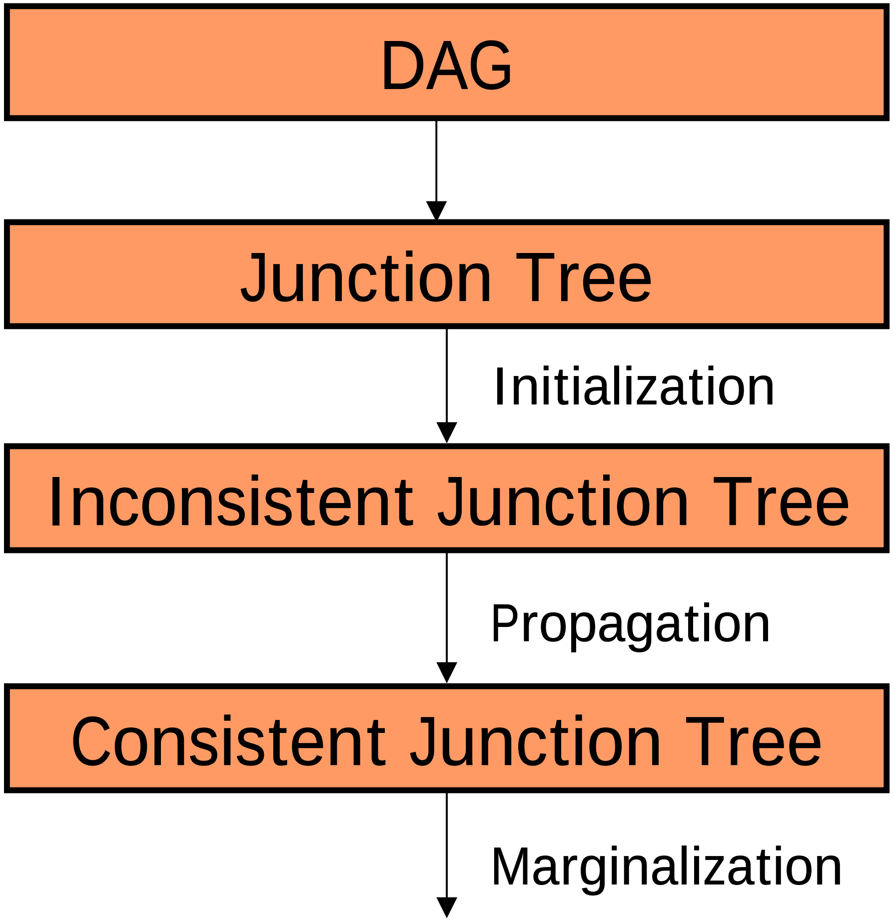
Basic idea (equality):

- transform one representation of the joint distribution to another in which for each clique, C, the potential function gives the marginal distribution for the variables in C:

$$\psi_c(x_c) = P(x_c)$$

This will also apply for the separators, S.

Principle of Inference

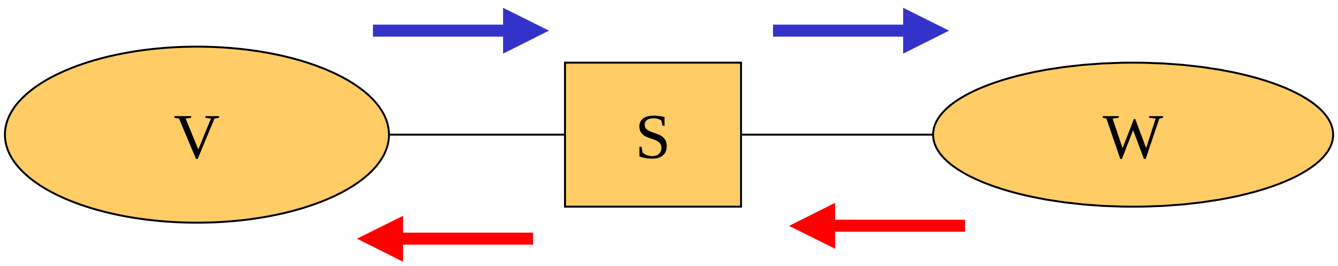


$$P(V = v \mid \mathbf{E} = \mathbf{e})$$

Inference

- Modify potentials
- Ensure joint probability is consistent
- Ensure consistency between neighbouring cliques
- Ensure clique potentials = clique marginals
- Ensure separator potentials = separator marginals

Inference



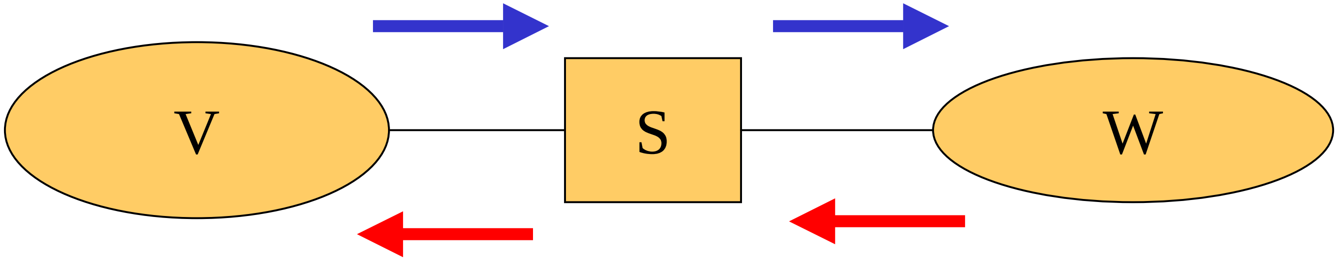
1. $\psi^*(S) = \sum_{V \backslash S} \psi(V)$
2. $\psi^*(W) = \psi(W)\ \psi^*(S) / \psi(S)$
3. $\psi^{**}(S) = \sum_{W \backslash S} \psi^*(W)$
4. $\psi^*(V) = \psi(V)\ \psi^{**}(S) / \psi^*(S)$

$$\sum_{V \backslash S} \psi^*(V) = \psi^{**}(S)$$

$$= \sum_{W \backslash S} \psi^*(W)$$

Consistency

Inference



1. $\psi^*(S) = \sum_{V\backslash S} \psi(V)$

2. $\psi^*(W) = \psi(W)\ \psi^*(S)\ /\ \psi(S)$

3. $\psi^{**}(S) = \sum_{W\backslash S} \psi^*(W)$

4. $\psi^*(V) = \psi(V)\ \psi^{**}(S)\ /\ \psi^*(S)$

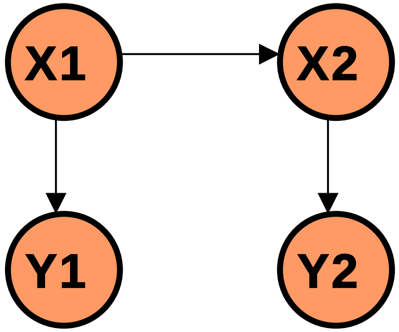
$\psi^*(V)\ \psi^*(W)\ /\ \psi^{**}(S)$

$= \psi(V)\ \psi(W)\ /\ \psi(S)$

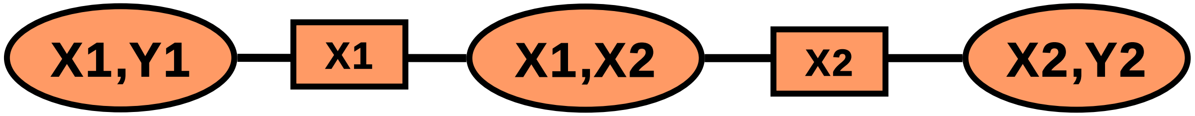
Joint probability remains same

Example: Create Junction Tree

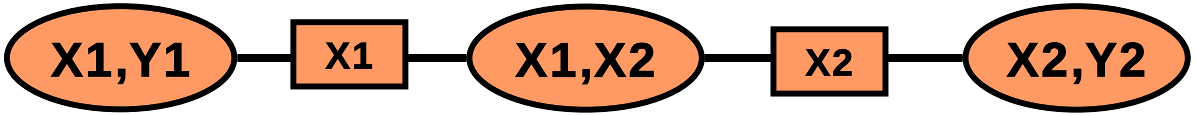
HMM with 2 time steps:



Junction Tree:



Initialization



$\psi_{X1} = 1$

$\psi_{X2} = 1$

Variable	Associated Cluster	Potential function
X1	X1,Y1	$\psi_{X1} = P(X1)$
Y1	X1,Y1	$\psi_{X1,Y1} = P(X1)P(Y1 \mid X1)$
X2	X1,X2	$\psi_{X1,X2} = P(X2 \mid X1)$
Y2	X2,Y2	$\psi_{X2,Y2} = P(Y2 \mid X2)$

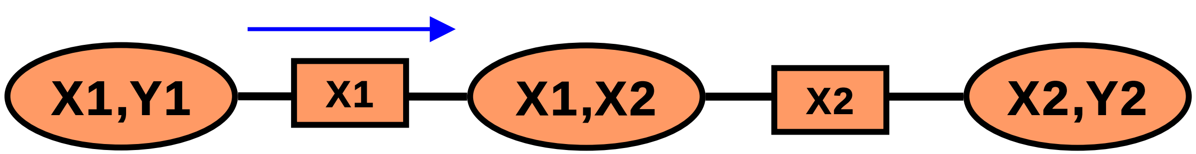
Collect Evidence

- Choose arbitrary clique, e.g. X1,X2, where all potential functions will be collected.
- Call recursively neighboring cliques for messages:

1. Call X1,Y1.
   - Projection: $$\psi_{X1} = \sum_{\{X1,Y1\}-X1} \psi_{X1,Y1} = \sum_{Y1} P(X1,Y1) = P(X1)$$
   - Absorption:

$$\psi_{X1,X2} \leftarrow \psi_{X1,X2} \frac{\psi_{X1}}{\psi_{X1}^{old}} = P(X2 \mid X1)P(X1) = P(X1,X2)$$

$$\psi_{X1}^{old} = 1$$



Collect Evidence (cont.)

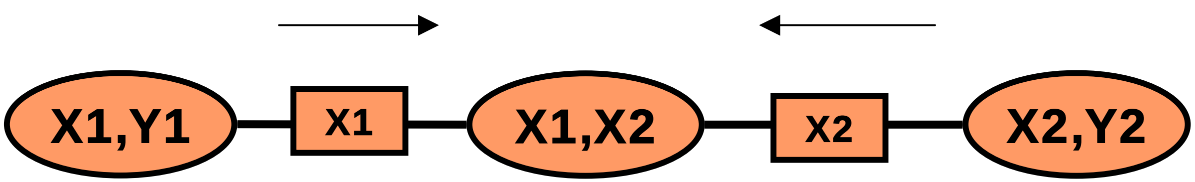
- 2. Call X2,Y2:
  - Projection:

$$\psi_{X2} = \sum_{\{X2,Y2\}-X2} \psi_{X2,Y2} = \sum_{Y2} P(Y2 \mid X2) = 1$$

  - Absorption:

$$\psi_{X1,X2} \leftarrow \psi_{X1,X2} \frac{\psi_{X2}}{\psi_{X2}^{old}} = P(X1, X2)$$

$$\psi_{X2}^{old} = 1$$



Distribute Evidence

- Pass messages recursively to neighboring nodes
- Pass message from X1,X2 to X1,Y1:
  - Projection:

$$\psi_{X1} = \sum_{\{X1,X2\}-X1} \psi_{X1,X2} = \sum_{X2} P(X1, X2) = P(X1)$$

  - Absorption:

$$\psi_{X1,Y1} \leftarrow \psi_{X1,Y1} \frac{\psi_{X1}}{\psi_{X1}^{old}} = P(X1,Y1)\frac{P(X1)}{P(X1)}$$

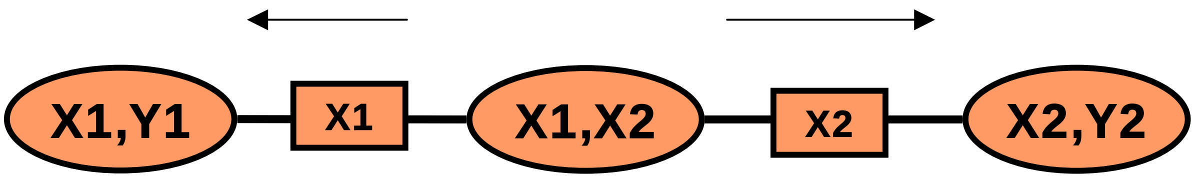
Distribute Evidence (cont.)

- Pass message from X1,X2 to X2,Y2:
  - Projection:

$$\psi_{X2} = \sum_{\{X1,X2\}-X2} \psi_{X1,X2} = \sum_{X1} P(X1, X2) = P(X2)$$

  - Absorption:

$$\psi_{X2,Y2} \leftarrow \psi_{X2,Y2} \frac{\psi_{X2}}{\psi_{X2}^{old}} = P(Y2 \mid X2) \frac{P(X2)}{1} = P(Y2, X2)$$

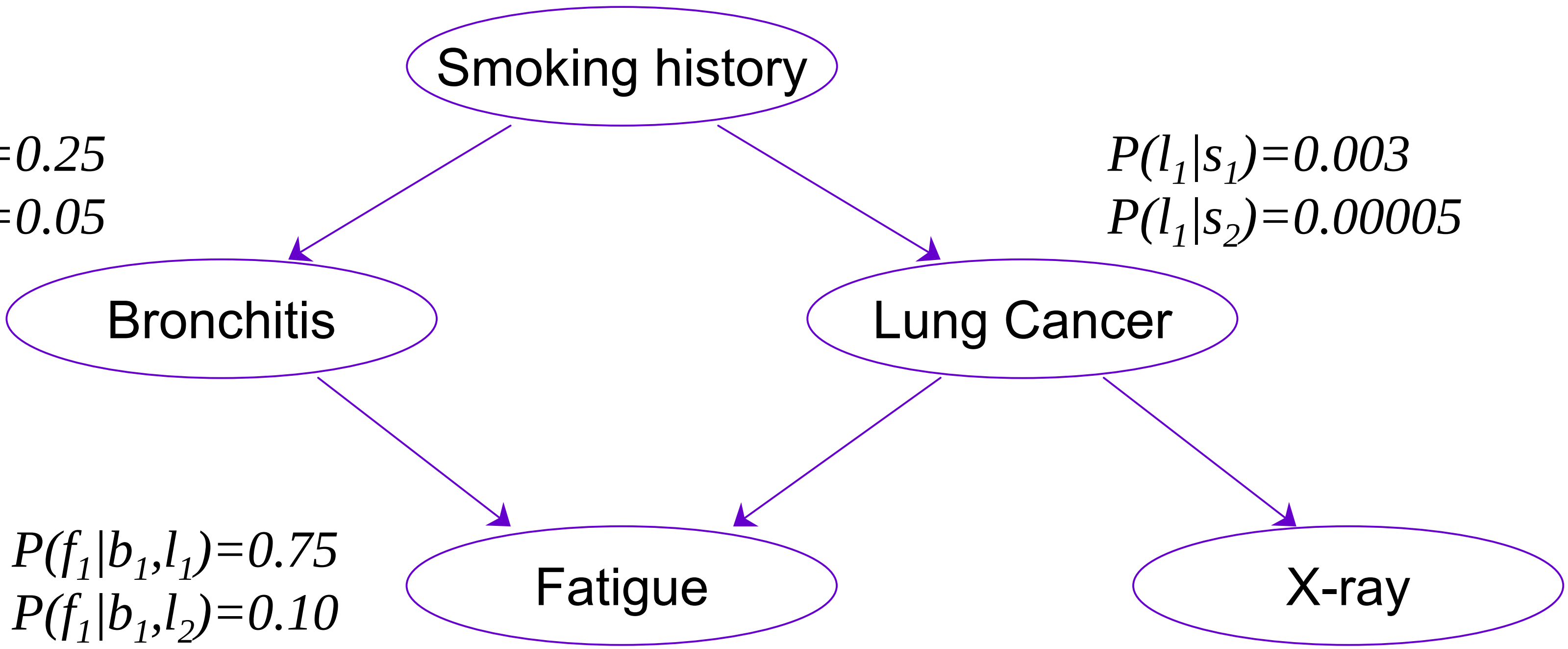


Another Example

$P(s_1)=0.2$

$P(b_1|s_1)=0.25$
$P(b_1|s_2)=0.05$

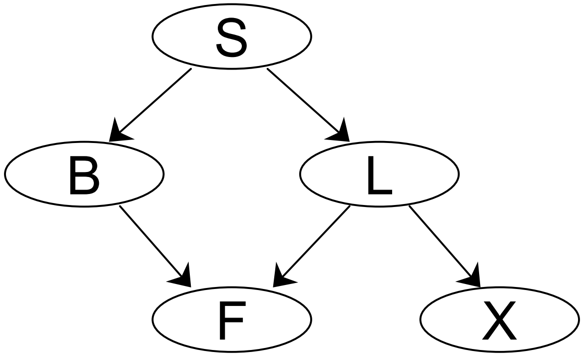
$P(l_1|s_1)=0.003$
$P(l_1|s_2)=0.00005$

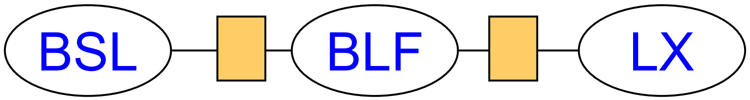


$P(f_1|b_1,l_1)=0.75$
$P(f_1|b_1,l_2)=0.10$
$P(f1|b_2,l_1)=0.5$
$P(f1|b_2,l_2)=0.05$

$P(x_1|l_1)=0.6$
$P(x_1|l_2)=0.02$

Build Junction Tree





Initialization

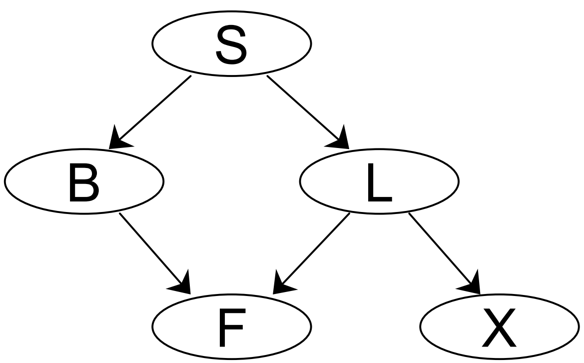
To initialize the potential functions:

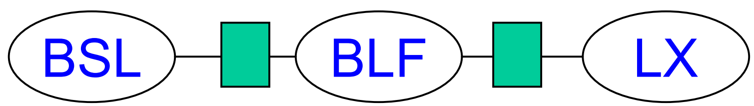
1. set all potentials to unity

2. for each variable X_i, select one node in the junction tree (one clique) containing both the variable and its parents, $pa(X_i)$, in the original DAG

3. multiply the potential by $P(x_i|pa(x_i))$

Example. The original BN can be represented as





$$\psi_{BSL} = P(b|s)P(b|s)P(s)$$

$$\psi_{BLF} = P(f|b,l)$$

$$\psi_{LX} = P(x|l)$$

For each separater, S:

$$\psi_S = 1$$

Propagating Information

Passing information from one clique C_1 to another C_2 via the separator between them, S_0, requires two steps:

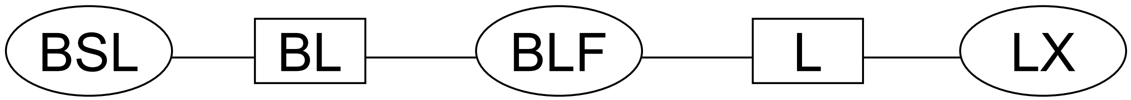
1. Obtain a new potential for S_0 by marginalizing out the variables in C_1 that are not in S_0:

$$\psi_{S_0} = \sum_{C_1 \setminus S_0} \psi_{C_1}$$

2. Obtain a new potential for C_2:

$$\psi^*_{C_2} = \psi_{C_2} \frac{\psi^*_{S_0}}{\psi_{S_0}}$$

An Example



Consider a flow from the clique {B,S,L} to {B,L,F}

Initial representation

$\psi_{BSL} = P(B \mid S)P(L \mid S)P(S)$

	l_1	l_2
s_1,b_1	0.00015	0.04985
s_1,b_2	0.00045	0.14955
s_2,b_1	0.000002	0.039998
s_2,b_2	0.000038	0.759962

$\psi_{BL} = 1$

	l_1	l_2
b_1	1	1
b_2	1	1

$\psi_{BLF} = P(F \mid B,L)$

	l_1	l_2
f_1,b_1	0.75	0.1
f_1,b_2	0.5	0.05
f_2,b_1	0.25	0.9
f_2,b_2	0.5	0.95

After Flow

ψ_{BSL}

	l_1	l_2
s_1,b_1	0.00015	0.04985
s_1,b_2	0.00045	0.14955
s_2,b_1	0.000002	0.039998
s_2,b_2	0.000038	0.759962

ψ_{BL}

	l_1	l_2
b_1	0.000152	0.089848
b_2	0.000488	0.909512

ψ_{BLF}

	l_1	l_2
f_1,b_1	0.000114	0.0089848
f_1,b_2	0.000244	0.0454756
f_2,b_1	0.000038	0.0808632
f_2,b_2	0.000244	0.8640364

Consider a flow from the clique {B,S,L} to {B,L,F}, but this time we include the information that Joe is a smoker, $S = s_1$.

Incorporation of Evidence

$\psi_{BSL} = P(B \mid S)P(L \mid S)P(S)$

	l_1	l_2
s_1,b_1	0.00015	0.04985
s_1,b_2	0.00045	0.14955
s_2,b_1	0	0
s_2,b_2	0	0

$\psi_{BL} = 1$

	l_1	l_2
b_1	1	1
b_2	1	1

$\psi_{BLF} = P(F \mid B,L)$

	l_1	l_2
f_1,b_1	0.75	0.1
f_1,b_2	0.5	0.05
f_2,b_1	0.25	0.9
f_2,b_2	0.5	0.95

After Flow

ψ_{BSL}

	l_1	l_2
s_1,b_1	0.00015	0.04985
s_1,b_2	0.00045	0.14955
s_2,b_1	0	0
s_2,b_2	0	0

ψ_{BL}

	l_1	l_2
b_1	0.00015	0.04985
b_2	0.00045	0.14955

ψ_{BLF}

	l_1	l_2
f_1,b_1	0.0001125	0.004985
f_1,b_2	0.000225	0.0074775
f_2,b_1	0.0000375	0.044865
f_2,b_2	0.000225	0.1420725

The Full Propagation (1)

Two phase propagation (Jensen et al, 1990)

1. Select an arbitrary clique, C_0

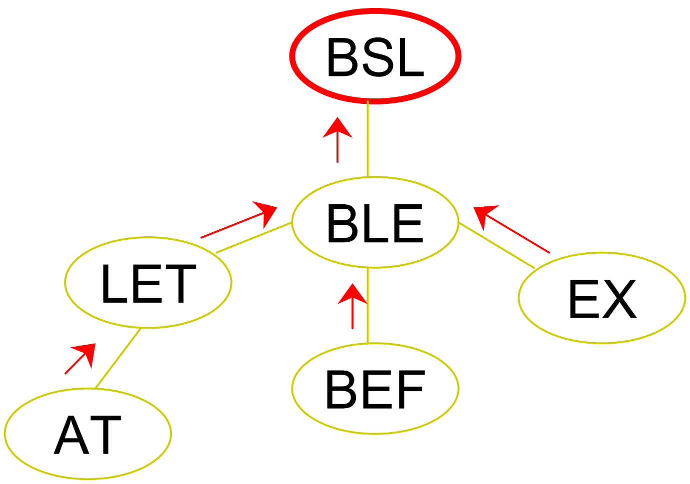
2. Collection Phase – flows passed from periphery to C_0

3. Distribution Phase – flows passed from C_0 to periphery

Eg

Collection

Distribution



Collection

The Full Propagation (2)

After the two propagation phases have been carried out the Junction tree will be in equilibrium with each clique containing the joint probability distribution for the variables it contains. Marginal probabilities for individual variables can then be obtained from the cliques.

Evidence, E, can be included before propagation by selecting a clique for each variable for which evidence is available. The potential for the clique is then set to 0 for any configuration which differs from the evidence. After propagation the result will be

$$P(x,E)=\frac{\prod_{c\in C}\psi_c(x_c,E)}{\prod_{s\in S}\psi_s(x_s,E)}$$

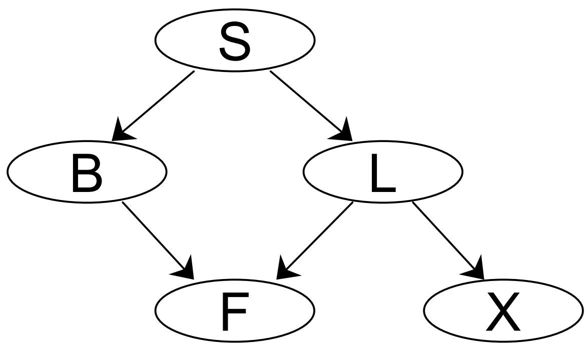
Normalizing gives

$$P(x\mid E)=\frac{\prod_{c\in C}\psi_c(x_c\mid E)}{\prod_{s\in S}\psi_s(x_s\mid E)}$$

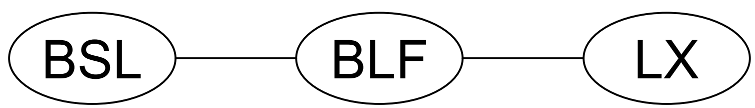
A Final Example (1)

What is the probability that someone has bronchitis given that they smoke, have fatigue and have received a positive Xray result?

Recall that the BN



can be represented by the junction tree



On entering evidence $S = s_1$, $F = f_1$ and $X = x_1$, we obtain …

ψ_{BSL} = P(B | S)P(L | S)P(S)

	l_1	l_2
s_1,b_1	0.00015	0.04985
s_1,b_2	0.00045	0.14955
s_2,b_1	0	0
s_2,b_2	0	0

ψ_{BL} = 1

	l_1	l_2
b_1	1	1
b_2	1	1

ψ_{BLF} = P(F | B,L)

	l_1	l_2
f_1,b_1	0.75	0.1
f_1,b_2	0.5	0.05
f_2,b_1	0	0
f_2,b_2	0	0

ψ_L = 1

l_1	l_2
1	1

ψ_{LX} = 1

	l_1	l_2
x_1	0.6	0.02
x_2	0	0

ψ_{BSL}

	l_1	l_2
s_1,b_1	0.0000675	0.0000997
s_1,b_2	0.000135	0.00014955
s_2,b_1	0	0
s_2,b_2	0	0

ψ_{BL}

	l_1	l_2
b_1	0.45	0.002
b_2	0.3	0.001

ψ_{BLF}

	l_1	l_2
f_1,b_1	0.45	0.002
f_1,b_2	0.3	0.001
f_2,b_1	0	0
f_2,b_2	0	0

ψ_L

l_1	l_2
0.6	0.02

ψ_{LX}

	l_1	l_2
x_1	0.6	0.02
x_2	0	0

After collection phase ψ_{BSL} is in final state.

To obtain P(b_1,E) marginalize out L,

0.0000675 +0.0000997 = 0.0001672

Normalizing for P(b_1|E) gives 0.37.

If we also observe L=l_1 then P(b_1|E,l_2) = 0.33