CS480 – Artificial Intelligence Fall 2015

Topic: Bayesian Networks
Chapter: 14
Date: 10/26

Mustafa Bilgic

http://www.cs.iit.edu/~mbilgic

Joint Distribution

- We have n random variables, $V_1, V_2, \ldots, V_n$
- We are interested in the probability of a possible world, where
  - $V_1=v_1, V_2=v_2, \ldots, V_n = v_n$
- $P(V_1, V_2, \ldots, V_n)$ associates a probability for each possible world ≡ the **joint distribution**
- How many independent parameters are needed, if V_i are all binary?

Joint Distribution

- Extremely useful
  - Can answer any type of query
- Extremely inefficient
  - Requires exponential size memory
  - Inference using an exponential-size table requires exponential time
- Chapter 14 $\Rightarrow$ Efficient representation and inference

Chain Rule

- $P(V_1, V_2, \ldots, V_n) =$
  - $P(V_1)P(V_2 | V_1)P(V_3 | V_1,V_2) \ldots P(V_n | V_1, V_2, \ldots, V_{n-1})$
  - $P(V_2)P(V_1 | V_2)P(V_3 | V_2,V_1) \ldots P(V_n | V_2, V_1, \ldots, V_{n-1})$
- $P(V_1, V_2, \ldots, V_n)$ requires 2^n-1 independent parameters
- $P(V_1)$: How many?
- $P(V_2 | V_1)$: How many?
- $P(V_3 | V_1, V_2)$: How many?
- …
- $P(V_n | V_1, V_2, \ldots, V_{n-1})$: How many?
- How many in total?

Marginal Independence

- Two random variables A and B are **marginally independent** if and only if
  - P(A, B) = P(A)*P(B)
- Two random variables A and B are **marginally independent** if and only if
  - P(A, B) = P(A)*P(B), equivalently
  - P(A | B) = P(A), equivalently
  - P(B | A) = P(B)

The Joint revisited

- $P(V_1, V_2, \ldots, V_n) =$
  - $P(V_1)P(V_2 | V_1)P(V_3 | V_1,V_2) \ldots P(V_n | V_1, V_2, \ldots, V_{n-1})$
- If $V_i \perp V_j$ for all $i \neq j$
  - $P(V_1, V_2, \ldots, V_n) =$
    - $P(V_1)P(V_2 | V_1)P(V_3 | V_1,V_2) \ldots P(V_n | V_1, V_2, \ldots, V_{n-1})$
    - $P(V_1)P(V_2)P(V_3) \ldots P(V_n)$
    - How many independent parameters now?

CONDITIONAL INDEPENDENCE

- Marginal independence is not very common
- Two random variables A and B are conditionally independent given C if and only if
  - $P(A, B|C) = P(A|C) * P(B|C)$, equivalently
  - $P(A|B,C) = P(A|C)$, equivalently
  - $P(B|A,C) = P(B|C)$

Why Independence?

- The joint distribution for n binary random variables
  - $2^n - 1$ independent entries; exponential
- If the variables were all
  - Marginally independent, then
    - $1 + 1 + \ldots + 1 = n$ independent parameters; polynomial
  - Conditionally independent given one of them, then
    - $1 + 2 + 2 + \ldots + 2 = 1 + 2(n-1) = 2n - 1$ independent parameters; polynomial

Advantages of more compact representation

- Fewer parameters
  - Makes learning and reasoning easier
- Consider asking an expert the probability of specific entry in a huge probability table

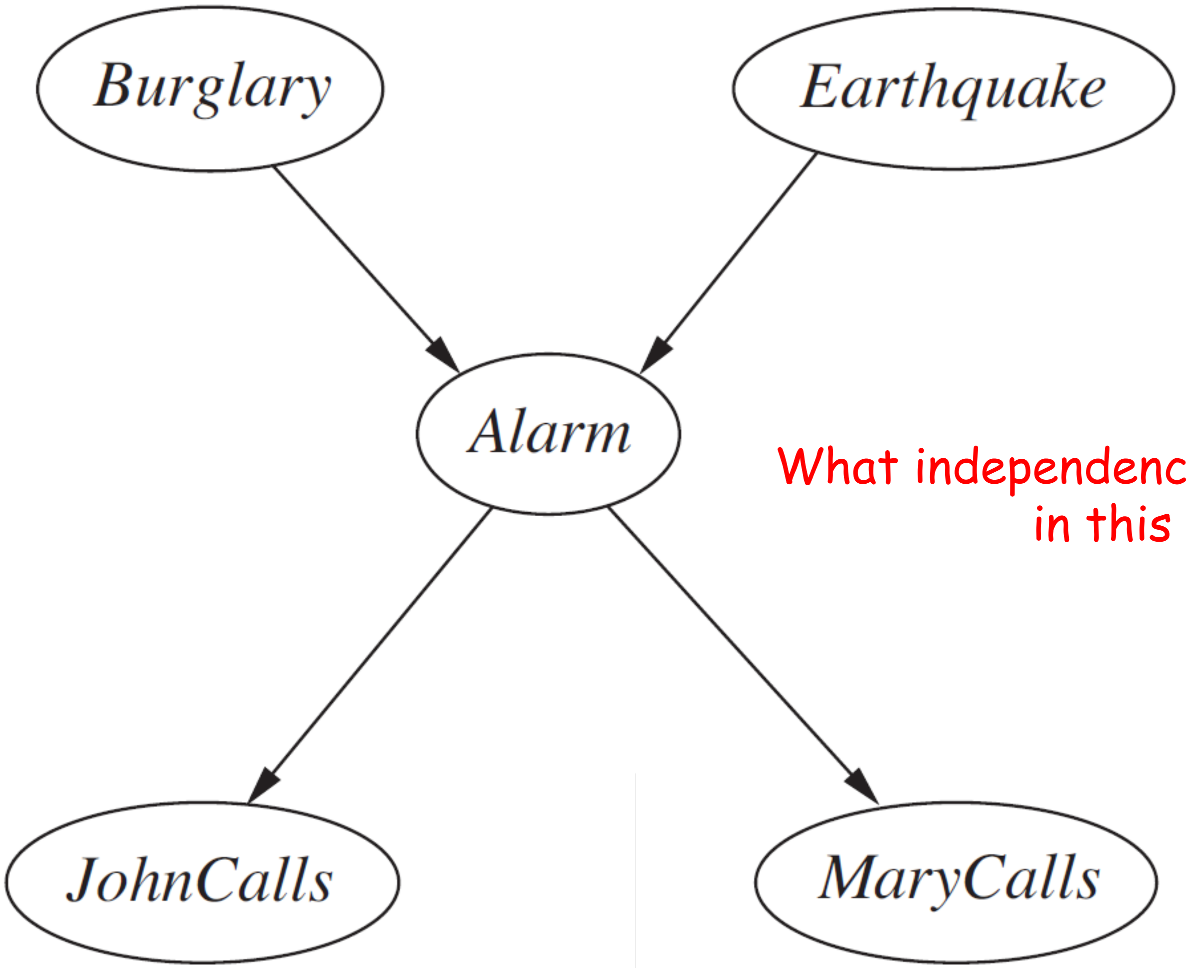
Bayesian Networks

- Random variables = nodes
- Direct relationships = directed edges
- BNs capture independencies
  - More compact than full joint representation
- Graphs provide
  - Graph theory / efficient reasoning
  - Intuition

EXAMPLES

- X causes Y and Y causes Z; no direct relationship between X and Z
  - $X \rightarrow Y \rightarrow Z$
  - Nothing is marginally independent of each other
  - $Z \perp X \mid Y$
- Y causes both X and Z; no direct relationship between X and Z
  - $X \leftarrow Y \rightarrow Z$
  - Nothing is marginally independent of each other
  - $Z \perp X \mid Y$
- Both X and Z cause Y; no direct relationship between X and Z
  - $X \rightarrow Y \leftarrow Z$
  - X and Z are marginally independent
  - X and Z become dependent when the value of Y is known

Burglary Example



What independencies are encoded in this BN?

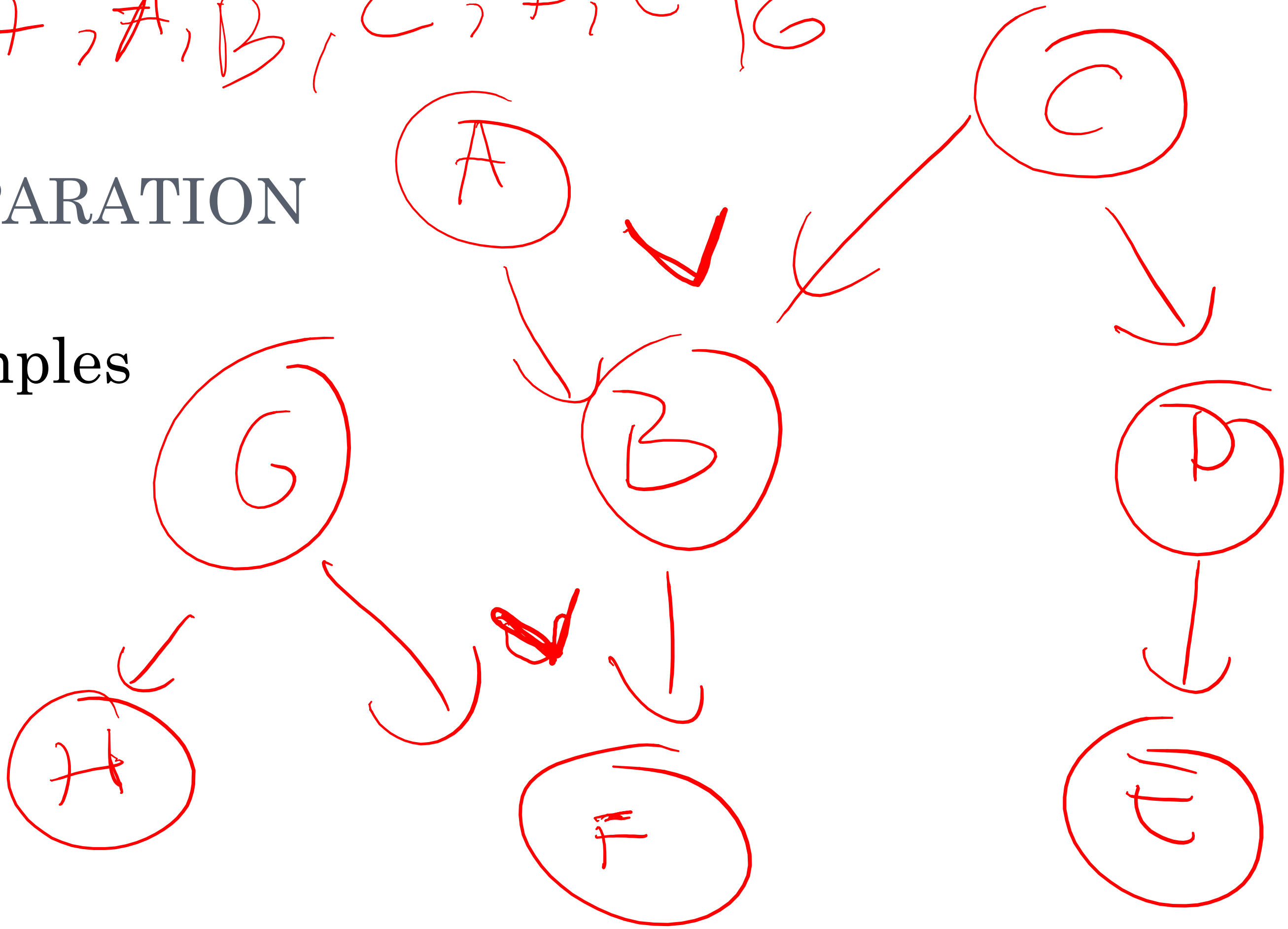
INDEPENDENCIES – D-SEPARATION

- Definition: Observed ≡ It's value is known
- Causal trail
  - $X \rightarrow Y \rightarrow Z$; E.g., Burglary → Alarm → MaryCalls
  - X and Z are independent if Y is observed
- Evidential trail
  - $X \leftarrow Y \leftarrow Z$; E.g., MaryCalls ← Alarm ← Burglary
  - X and Z are independent if Y is observed
- Common cause
  - $X \leftarrow Y \rightarrow Z$; E.g., JohnCalls ← Alarm → MaryCalls
  - X and Z are independent if Y is observed
- Common effect
  - $X \rightarrow Y \leftarrow Z$; E.g., Burglary → Alarm ← Earthquake
  - X and Z are marginally independent but they become dependent if Y or any of Y's descendants are observed

$H \perp F, A, B, C, D, E \mid G$

D-SEPARATION

- Examples



Path is blocked if 1) not V and observed

2) V and not observed itself and descendants not observed

$A \perp G, H, C, D, E$

Independencies - Parents

- X is independent of its non-descendants given its parents
  - X $\perp$ Non-descendants(X) | Parents(X)
- What's a non-descendant?
- What are the independencies in the burglary example?

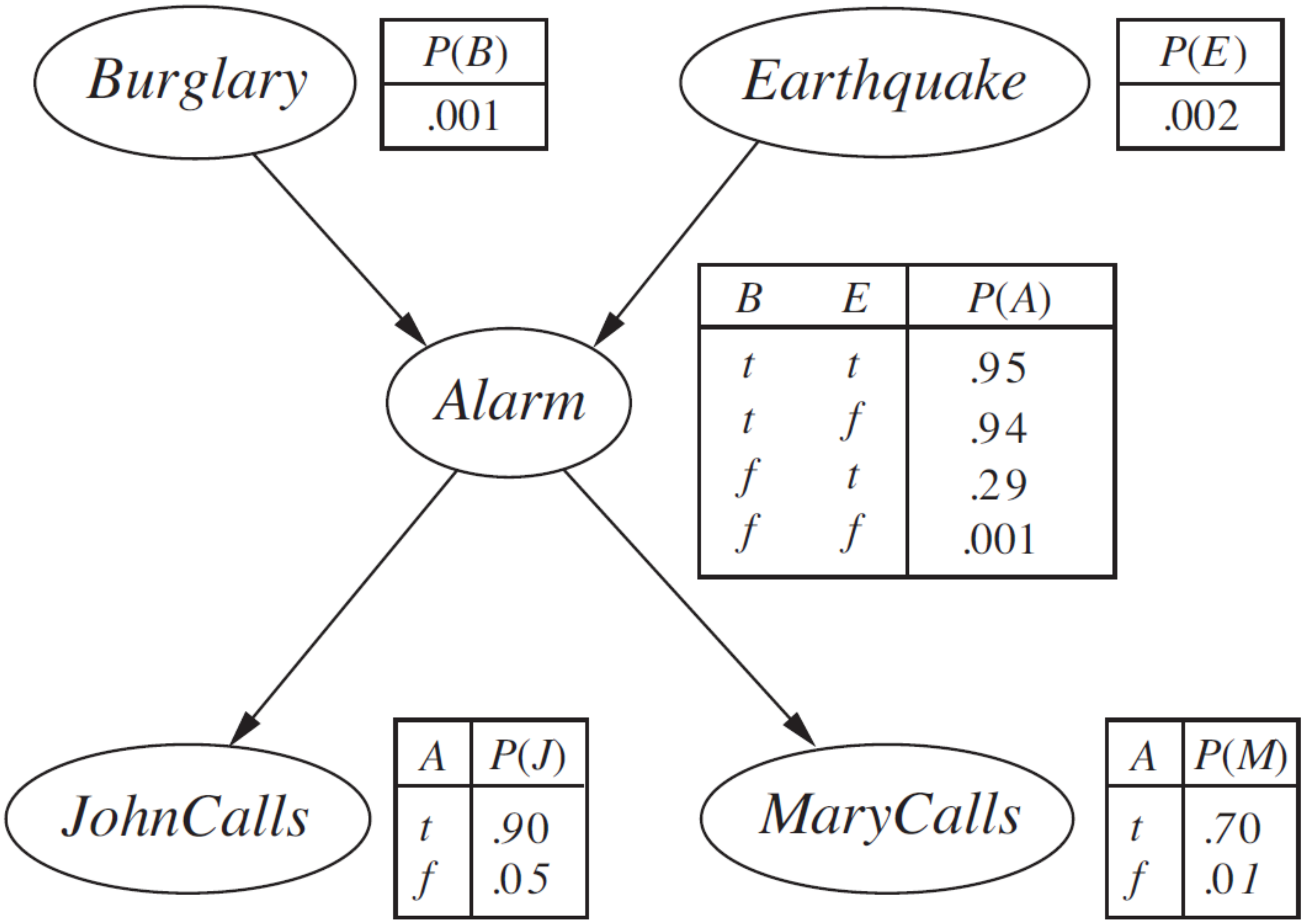
Parameterization

Given the indecencies encoded in a BN, what are the parameters needed to capture the joint representation efficiently?

Bayesian network parameterization

$$P(\mathbf{V}) = \prod_i P(V_i \mid Pa(V_i))$$

Burglary Example



$P(B)$
.001

$P(E)$
.002

B	E	$P(A)$
t	t	.95
t	f	.94
f	t	.29
f	f	.001

A	$P(J)$
t	.90
f	.05

A	$P(M)$
t	.70
f	.01

Theorems

- **Theorem 1:** If a probability distribution P holds the independencies encoded in G, then P factorizes according to G
- **Theorem 2:** If P factorizes according to G, then it holds the independencies encoded in G
- Let's see a constructive proof for Theorem 1; we'll not prove Theorem 2

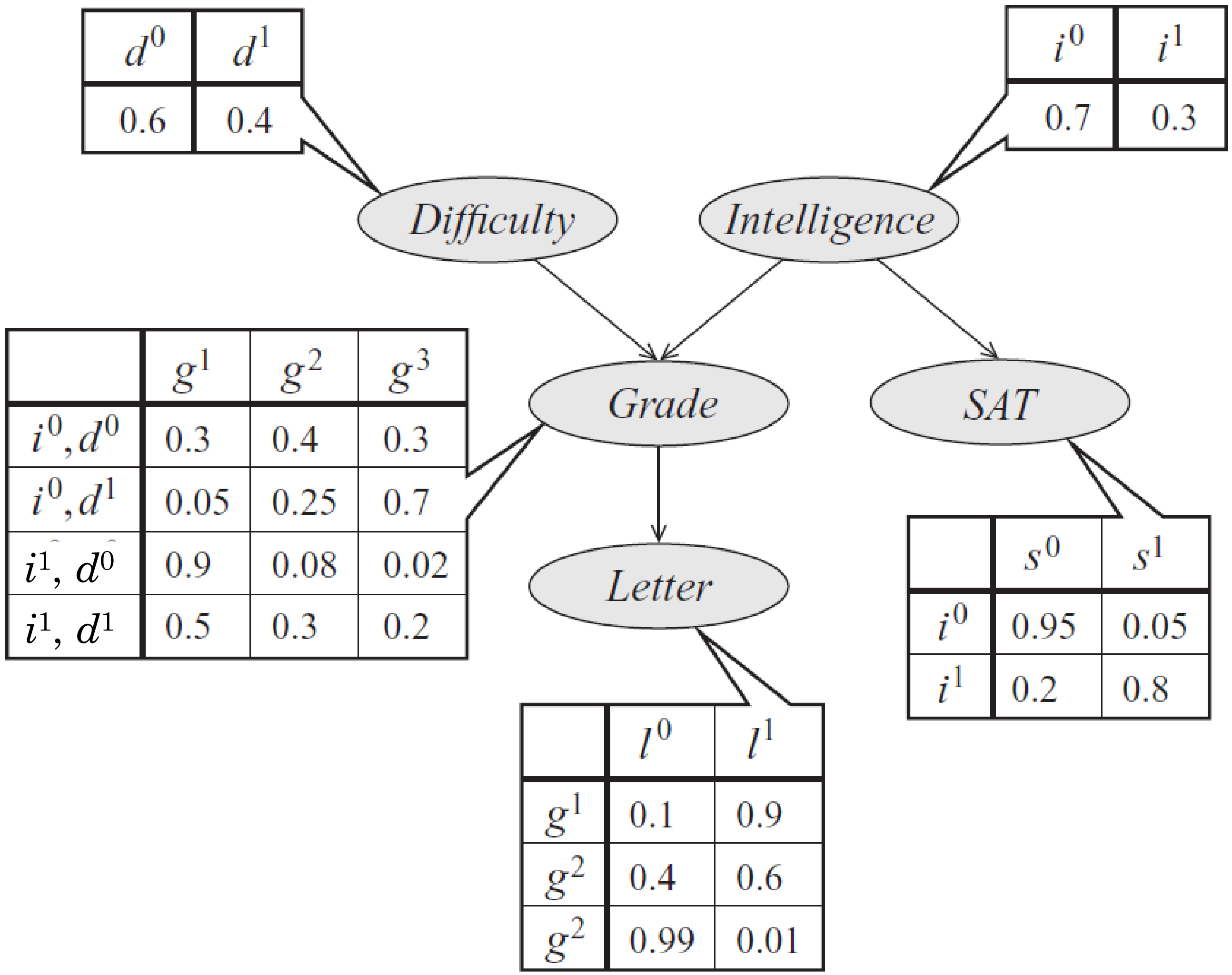
From independence to factorization

- Linear chain example
  - $A \rightarrow B \rightarrow C \rightarrow D \rightarrow E$
- Burglary example

Burglary Example

- The joint representation
  - Equation
- Contrast number of parameters for
  - Probability table for joint
  - Bayesian network

Student Example



d^0	d^1
0.6	0.4

i^0	i^1
0.7	0.3

	g^1	g^2	g^3
i^0, d^0	0.3	0.4	0.3
i^0, d^1	0.05	0.25	0.7
i^1, d^0	0.9	0.08	0.02
i^1, d^1	0.5	0.3	0.2

	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^2	0.99	0.01

Student Example

- The joint representation
  - Equation
- Contrast number of parameters for
  - Probability table for joint
  - Bayesian network

So far

- We've discussed the representation
- Now, it's time for inference

Reasoning Patterns

- **Causal reasoning**
  - From causes to effects
    - E.g., Burglary to Alarm to MaryCalls
    - E.g., Intelligence to Grade to Letter
- **Evidential reasoning**
  - From effects to the causes
    - E.g., JohnCalls to Alarm to Earthquake
    - E.g, Letter to Grade to Difficulty
- **Explaining away/inter-causal reasoning**
  - Causes of a common effect interact
    - E.g., Earthquake, Burglary, and Alarm (and Alarm's descendants)
    - E.g., Difficulty, Intelligence, and Grade (and Grade's descendants)

Inference in Bayesian Networks

- Variable elimination
  - Without evidence
  - With evidence
- Sampling
  - Without evidence
  - With evidence

Variable Elimination

- Let
  - **V** be the set of all variables, **Q** be the set of query variables, **E** be the set of evidence variables
  - P(**Q** | **E**) be the query

1. Write down the joint dist. using the Bayesian network structure
2. Set the variables in **E** to their respective values
3. Sum over all variables in $\mathbf{V} \setminus (\mathbf{Q} \cup \mathbf{E})$
   a) Pick an order for variables in $\mathbf{V} \setminus (\mathbf{Q} \cup \mathbf{E})$
   b) For each variable V_i in $\mathbf{V} \setminus (\mathbf{Q} \cup \mathbf{E})$, create a new factor by
      - Multiplying all the factors that contains V_i, and
      - Summing over possible values of V_i
4. Normalize the last remaining factor (this step is unnecessary if **E** is empty)

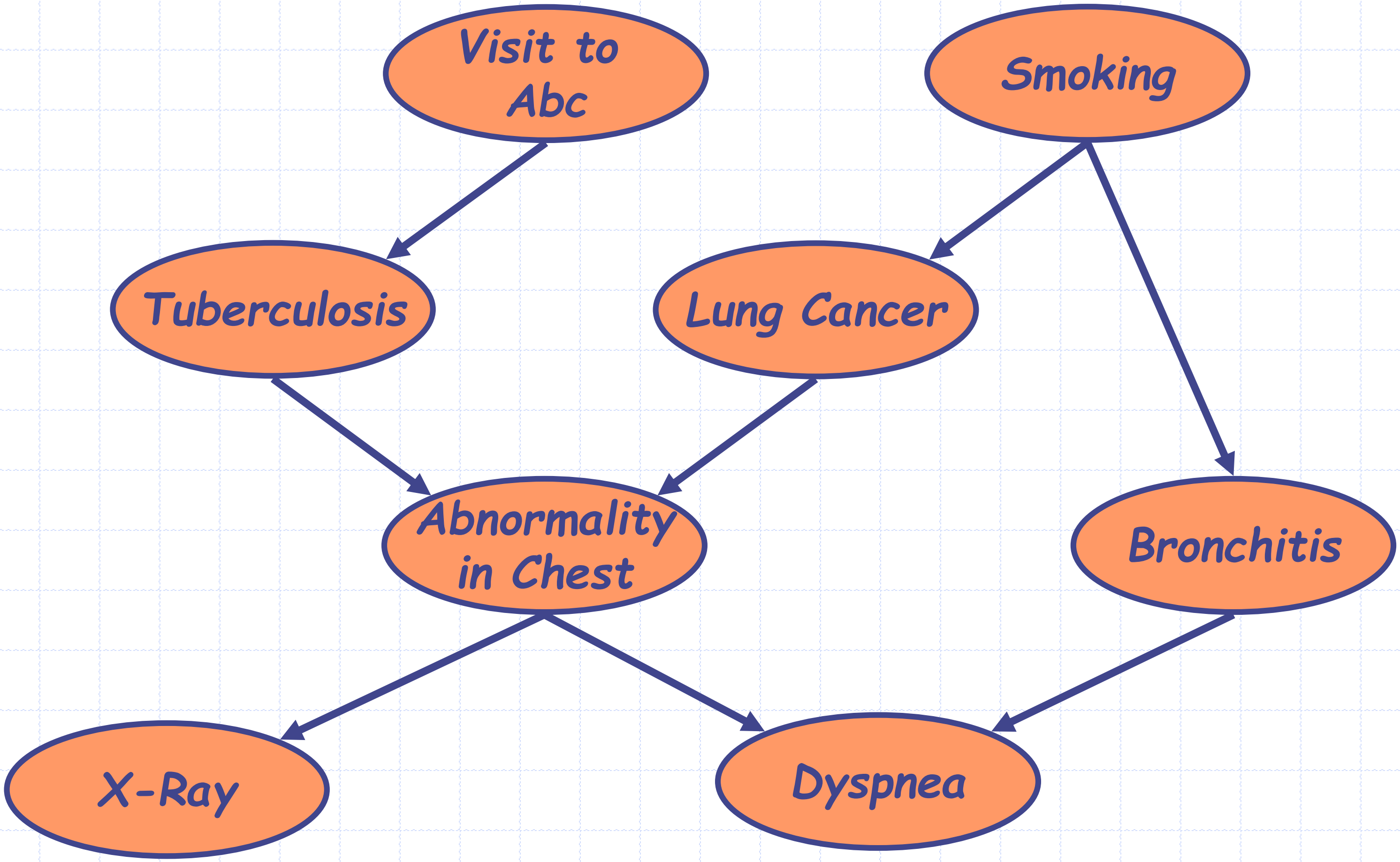
Examples

- Given the following BNs, compute the requested probabilities efficiently (without computing the full joint)
  - $A \rightarrow B \rightarrow C$;
    - P(A) = <0.6, 0.4>,
    - P(B | A=t) = <0.8, 0.2>, P(B | A=f) = <0.1, 0.9>
    - P(C | B=t) = <0.7, 0.3>, P(C | B=f) = <0.4, 0.6>
  - Compute P(A), P(B), P(C), P(C | A=t), P(A | C=t)

ACKNOWLEDGEMENT

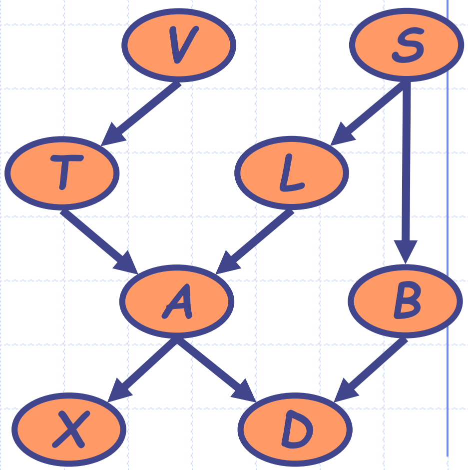
- The following slides are courtesy of Dr. Lise Getoor
- I have modified them to fit to our needs

Abc Network



Visit to Abc
Smoking
Tuberculosis
Lung Cancer
Abnormality in Chest
Bronchitis
X-Ray
Dyspnea

- We want to compute *P(D)*
- Need to eliminate: *V,S,X,T,L,A,B*



The joint distribution

$$P(V)P(S)P(T|V)P(L|S)P(B|S)P(A|T,L)P(X|A)P(D|A,B)$$

Eliminate: *V*

Compute:

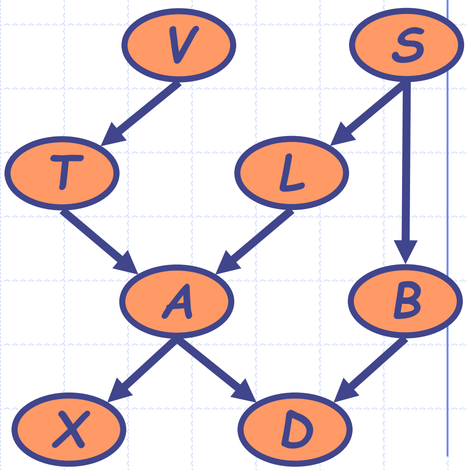
$$f_V(T) = \sum_V P(V)P(T|V)$$

$$\Rightarrow f_V(T)P(S)P(L|S)P(B|S)P(A|T,L)P(X|A)P(D|A,B)$$

Note: $f_V(T) = P(T)$

In general, result of elimination is not necessarily a probability term

- We want to compute $P(D)$
- Need to eliminate: S, X, T, L, A, B



The joint distribution

$$P(V)P(S)P(T|V)P(L|S)P(B|S)P(A|T,L)P(X|A)P(D|A,B)$$

$$\Rightarrow f_V(T)\underline{P(S)}\,\underline{P(L|S)}\,\underline{P(B|S)}P(A|T,L)P(X|A)P(D|A,B)$$

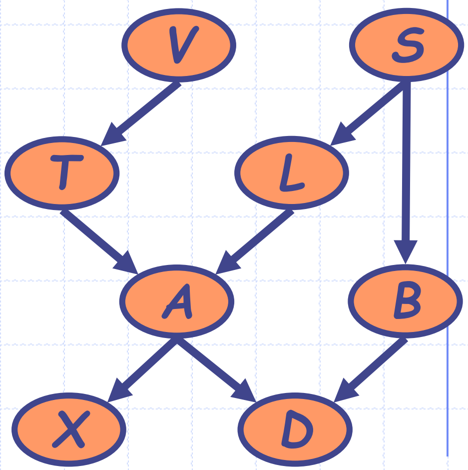
Eliminate: S

Compute: $$f_S(B,L) = \sum_S P(S)P(B|S)P(L|S)$$

$$\Rightarrow f_V(T)\underline{f_S(B,L)}P(A|T,L)P(X|A)P(D|A,B)$$

Summing on S results in a factor with two arguments $f_S(B,L)$
In general, result of elimination may be a function of several variables

- We want to compute *P(D)*
- Need to eliminate: *X,T,L,A,B*



The joint distribution

$$P(V)P(S)P(T|V)P(L|S)P(B|S)P(A|T,L)P(X|A)P(D|A,B)$$

$$\Rightarrow f_V(T)P(S)P(L|S)P(B|S)P(A|T,L)P(X|A)P(D|A,B)$$

$$\Rightarrow f_V(T)f_S(B,L)P(A|T,L)\underline{P(X|A)}P(D|A,B)$$

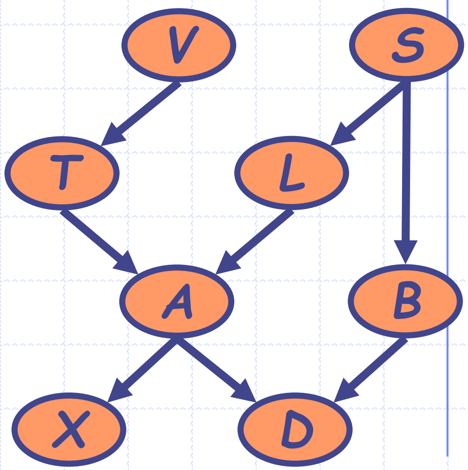
Eliminate: *X*

Compute: $$f_X(A) = \sum_X P(X|A)$$

$$\Rightarrow f_V(T)f_S(B,L)\underline{f_X(A)}P(A|T,L)P(D|A,B)$$

Note: $f_X(A) = 1$ for all values of *A* !!

- We want to compute $P(D)$
- Need to eliminate: T, L, A, B



The joint distribution

$$P(V)P(S)P(T|V)P(L|S)P(B|S)P(A|T,L)P(X|A)P(D|A,B)$$

$$\Rightarrow f_V(T)P(S)P(L|S)P(B|S)P(A|T,L)P(X|A)P(D|A,B)$$

$$\Rightarrow f_V(T)f_S(B,L)P(A|T,L)P(X|A)P(D|A,B)$$

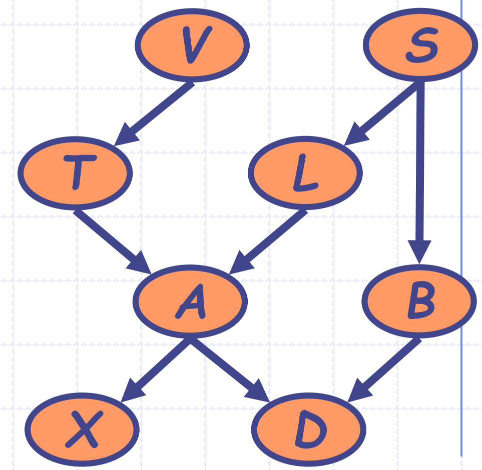
$$\Rightarrow \underline{f_V(T)}f_S(B,L)f_X(A)\underline{P(A|T,L)}P(D|A,B)$$

Eliminate: T

Compute: $$f_T(A,L) = \sum_T f_V(T)P(A|T,L)$$

$$\Rightarrow f_S(B,L)f_X(A)\underline{f_T(A,L)}P(D|A,B)$$

- We want to compute *P(D)*
- Need to eliminate: *L*, *A*, *B*



The joint distribution

$$P(V)P(S)P(T|V)P(L|S)P(B|S)P(A|T,L)P(X|A)P(D|A,B)$$

$$\Rightarrow f_V(T)P(S)P(L|S)P(B|S)P(A|T,L)P(X|A)P(D|A,B)$$

$$\Rightarrow f_V(T)f_S(B,L)P(A|T,L)P(X|A)P(D|A,B)$$

$$\Rightarrow f_V(T)f_S(B,L)f_X(A)P(A|T,L)P(D|A,B)$$

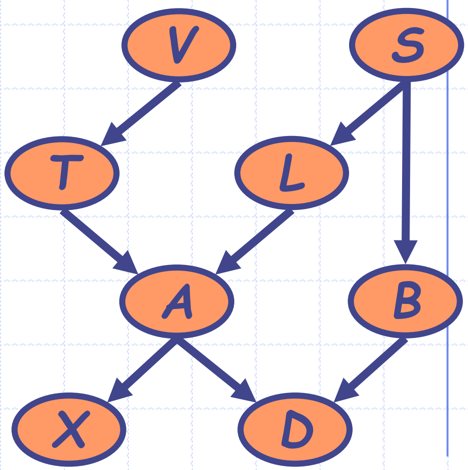
$$\Rightarrow \underline{f_S(B,L)}f_X(A)\underline{f_T(A,L)}P(D|A,B)$$

Eliminate: *L*

Compute: $$f_L(A,B) = \sum_L f_S(B,L)f_T(A,L)$$

$$\Rightarrow f_X(A)\underline{f_L(A,B)}P(D|A,B)$$

- We want to compute *P(D)*
- Need to eliminate: *A*,*B*



The joint distribution

$$P(V)P(S)P(T|V)P(L|S)P(B|S)P(A|T,L)P(X|A)P(D|A,B)$$

$$\Rightarrow f_V(T)P(S)P(L|S)P(B|S)P(A|T,L)P(X|A)P(D|A,B)$$

$$\Rightarrow f_V(T)f_S(B,L)P(A|T,L)P(X|A)P(D|A,B)$$

$$\Rightarrow f_V(T)f_S(B,L)f_X(A)P(A|T,L)P(D|A,B)$$

$$\Rightarrow f_S(B,L)f_X(A)f_T(A,L)P(D|A,B)$$

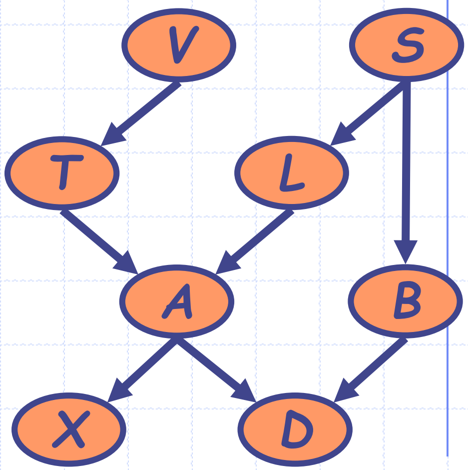
$$\Rightarrow \underline{f_X(A)}\,\underline{f_L(A,B)}\,\underline{P(D|A,B)}$$

Eliminate: *A*

Compute: $f_A(B,D) = \sum_A f_X(A) f_L(A,B) P(D|A,B)$

$$\Rightarrow \underline{f_A(B,D)}$$

- We want to compute $P(D)$
- Need to eliminate: B



The joint distribution

$$P(V)P(S)P(T|V)P(L|S)P(B|S)P(A|T,L)P(X|A)P(D|A,B)$$

$$\Rightarrow f_V(T)P(S)P(L|S)P(B|S)P(A|T,L)P(X|A)P(D|A,B)$$

$$\Rightarrow f_V(T)f_S(B,L)P(A|T,L)P(X|A)P(D|A,B)$$

$$\Rightarrow f_V(T)f_S(B,L)f_X(A)P(A|T,L)P(D|A,B)$$

$$\Rightarrow f_S(B,L)f_X(A)f_T(A,L)P(D|A,B)$$

$$\Rightarrow f_X(A)f_L(A,B)P(D|A,B)$$

$$\Rightarrow \underline{f_A(B,D)}$$

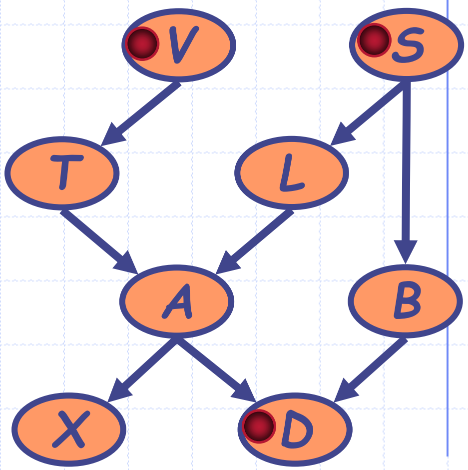
Eliminate: B
Compute: $$f_B(D)=\sum_B f_A(B,D)$$

$$\Rightarrow \underline{f_B(D)}=P(D)$$

Why Variable Elimination?

- We could compute P(D) by
  - Computing the full joint table, and then
  - Summing over the remaining variables
- Variable elimination, with a *good* ordering, can
  - Save memory, and
  - Save time

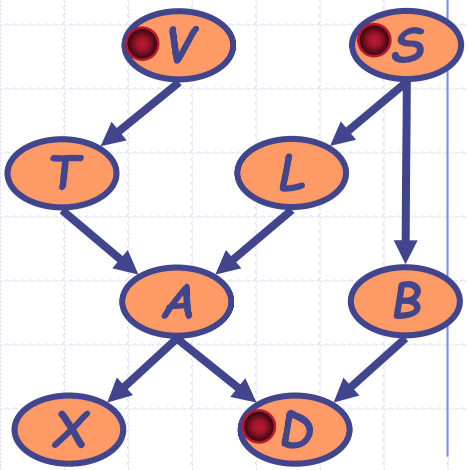
Dealing with Evidence



- How do we deal with evidence?
- Suppose get evidence $V = t, S = f, D = t$
- We want to compute $P(L \mid V = t, S = f, D = t)$

$P(L \mid V = t, S = f, D = t) = ?$

Dealing with Evidence



- We start by writing the joint:

$$P(V)P(S)P(T \mid V)P(L \mid S)P(B \mid S)P(A \mid T, L)P(X \mid A)P(D \mid A, B)$$

- We do not need to eliminate variables whose values are known; just set those variables to their known values.

$$P(V = t)P(S = f)P(T \mid V = t)P(L \mid S = f)$$

$$P(B \mid S = f)P(A \mid T, L)P(X \mid A)P(D = t \mid A, B)$$

- Eliminate all but the query variables (*L*) and the evidence variables (*V, S, D*). i.e, eliminate T, A, B, and X
- Doing so gives us *P(L, V=t, S=f, D=t)*
- We need *P(L | V=t, S=f, D=t)*
- *P(L | V=t, S=f, D=t) = P(L, V=t, S=f, D=t) / P(V=t, S=f, D=t)*
- How can we compute *P(V=t, S=f, D=t)* efficiently, given all the computations we have done so far?

Bayesian Network Inference

- We have seen variable elimination
- We have not seen
  - Junction-tree and message passing (very similar)
  - Approximate inference techniques – where approximate probabilities are computed for efficiency reasons

Applications of Bayesian Networks

- Too many to list
- Here is a book about it: http://www.wiley.com/WileyCDA/WileyTitle/productCd-0470060301.html
- Chapters include:
  - Medical diagnosis
  - Complex genetic models
  - Crime risk factors analysis
  - Inference problems in forensic science
  - Classifiers for modeling of mineral potential
  - Reliability analysis of systems
  - Credit-rating of companies
  - Classification of Chilean wines
  - Complex industrial process operation
  - Probability of default for large corporates
  - Risk management in robotics