

CS 480 – INTRODUCTION TO ARTIFICIAL INTELLIGENCE

TOPIC: BAYESIAN NETWORKS
CHAPTER: 13



Mustafa Bilgic



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



<https://twitter.com/bilgicm>

JOINT DISTRIBUTION

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

of atoms in universe
 $10^{80} \approx 2^{300}$

$V_1 \dots V_n$
 $T \dots T$
 $T \dots T$
 \vdots
 $\left. \begin{array}{c} T \dots T \\ T \dots T \\ \vdots \end{array} \right\} 2^n$

JOINT DISTRIBUTION

$$P(\underbrace{\dots}_D \mid \underbrace{\dots}_S) = \frac{P(D, S)}{P(S)}$$

Handwritten red formula showing the relationship between joint and conditional probabilities. The numerator is $P(D, S)$ with a red arrow pointing down to it. The denominator is $P(S)$. The entire expression is enclosed in large parentheses with a vertical line separating the two parts, and the variables D and S are underlined below the dots.

- Extremely useful
 - Can answer any type of query
- Extremely inefficient
 - Requires exponential size memory
 - Inference using an exponential-size table requires exponential time
- Chapter 13 \Rightarrow Efficient representation and inference

CHAIN RULE

- $P(V_1, V_2, \dots, V_n) =$
 - $P(V_1)P(V_2|V_1)P(V_3|V_1, V_2) \dots P(V_n|V_1, V_2, \dots, V_{n-1})$
 - $P(V_2)P(V_1|V_2)P(V_3|V_2, V_1) \dots P(V_n|V_2, V_1, \dots, V_{n-1})$
 - ...
- If all V_i are binary, $P(V_1, V_2, \dots, V_n)$ 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?
- ...
- $P(V_n|V_1, V_2, \dots, V_{n-1})$: How many?
- How many in total?

MARGINAL INDEPENDENCE

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

THE JOINT REVISITED

○ $P(V_1, V_2, \dots, V_n) =$

• $P(V_1)P(V_2|V_1)P(V_3|V_1, V_2) \dots P(V_n|V_1, V_2, \dots, V_{n-1})$

○ If $V_i \perp V_j$ for all $i \neq j$

• $P(V_1, V_2, \dots, V_n) =$

○ $P(V_1)P(V_2|V_1)P(V_3|V_1, V_2) \dots P(V_n|V_1, V_2, \dots, V_{n-1})$

○ $P(V_1)P(V_2)P(V_3) \dots P(V_n)$

○ How many independent parameters now?

$2^n - 1$

← chain rule

\prod

Compare $2^n - 1$ vs \prod
→ ↑ exponential ↑ linear

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
 - $1 + 1 + \dots + 1 = n$ independent parameters; polynomial
 - Conditionally independent given one of them, then
 - $1 + 2 + 2 + \dots + 2 = 1 + 2(n-1) = 2n - 1$ independent parameters; polynomial
- naive Bayes*

ADVANTAGES OF MORE COMPACT REPRESENTATION

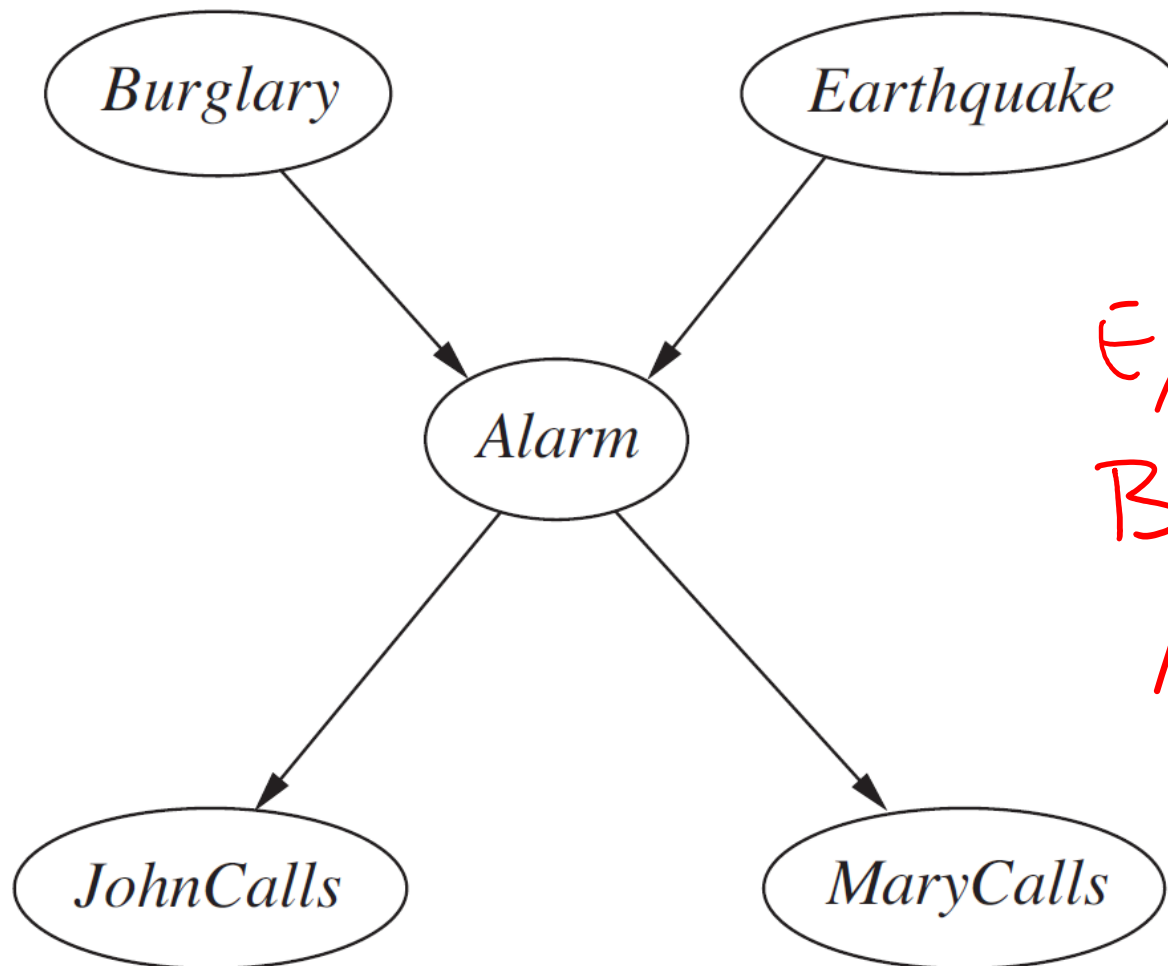
- Fewer parameters
 - Makes learning and reasoning easier
- Consider asking an expert the probability of specific entry in a huge probability table

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

BURGLARY EXAMPLE

5 variables
2⁵ - 1 incl



~~E~~XM
~~B~~XM
~~E~~XJ
~~A~~Aj
JXM

DIRECTED GRAPHS

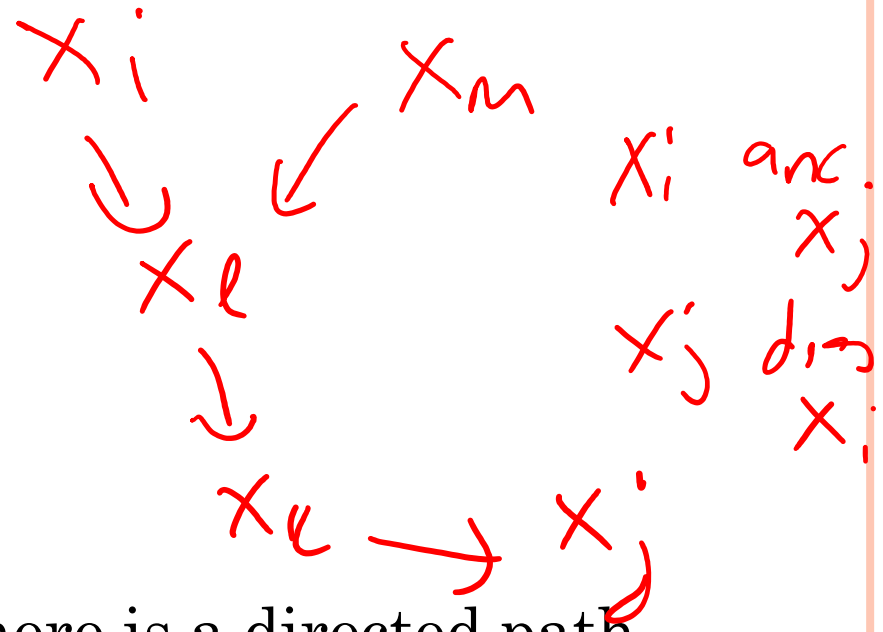
- A **graph** consists of **nodes** and **edges**
- **Nodes:** $\mathcal{X} = \{X_1, X_2, \dots, X_n\}$
- **Undirected Edge:** $X_i - X_j$
- **Directed Edge:** $X_i \rightarrow X_j$
- A graph is **directed** if its *all* edges are directed

BN directed acyclic graph

DAG

RELATIONSHIPS

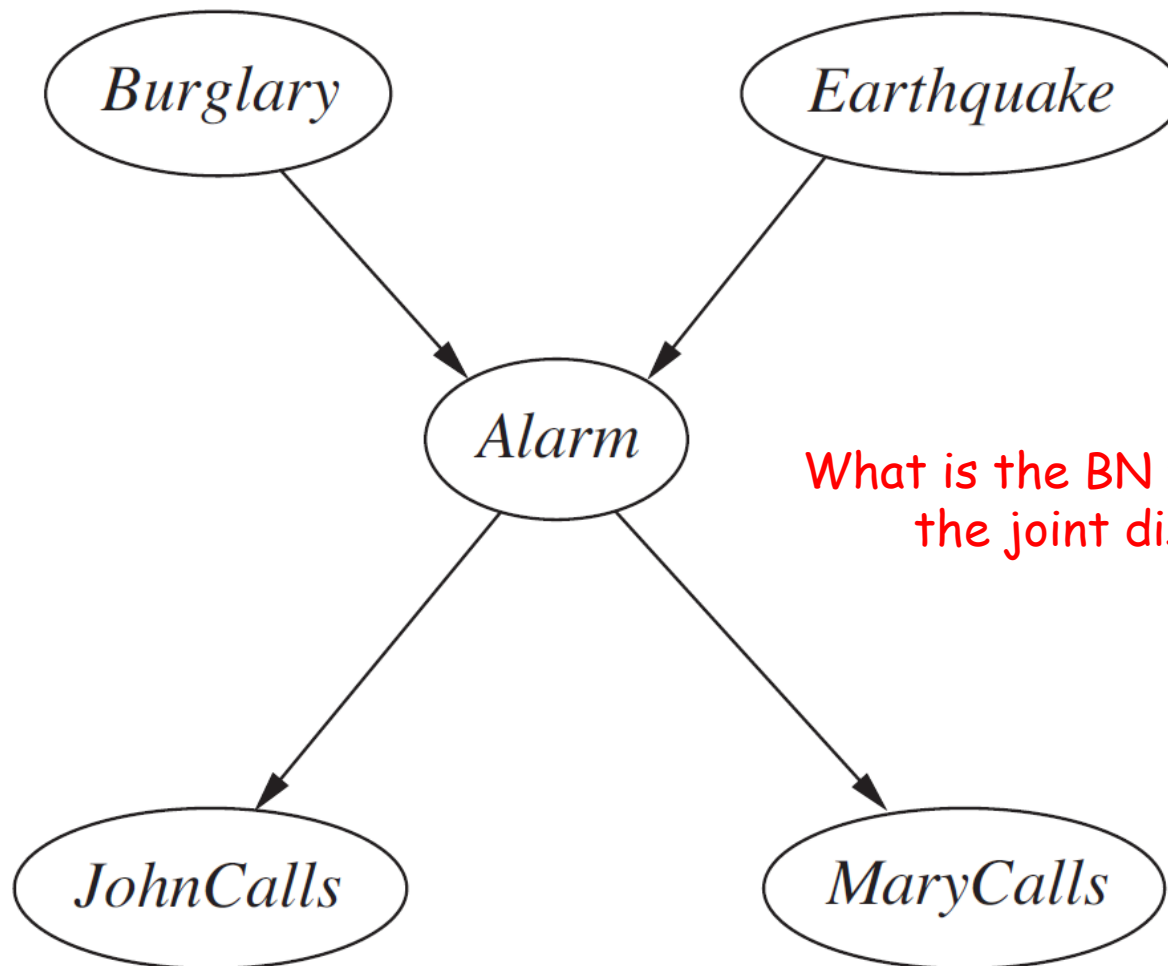
- $X_i \rightarrow X_j$
 - X_i is the **parent**
 - X_j is the **child**
- X_i is an **ancestor** of X_j if there is a directed path from X_i to X_j
- X_i is a **descendant** of X_j if there is a directed path from X_j to X_i
- **Nondescendants**(X_i) $\equiv \mathcal{X} \setminus \text{Descendants}(X_i)$



BAYESIAN NETWORK FACTORIZATION

$$P(X_1, \dots, X_n) = \prod_i P(X_i \mid Pa(X_i))$$

BURGLARY EXAMPLE



What is the BN factorization of the joint distribution?

INDEPENDENCIES - PARENTS

- X is independent of its non-descendants given its parents
 - $X \perp \text{Non-descendants}(X) \mid \text{Parents}(X)$
- 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?

THEOREMS

Ind \Leftrightarrow Factorization

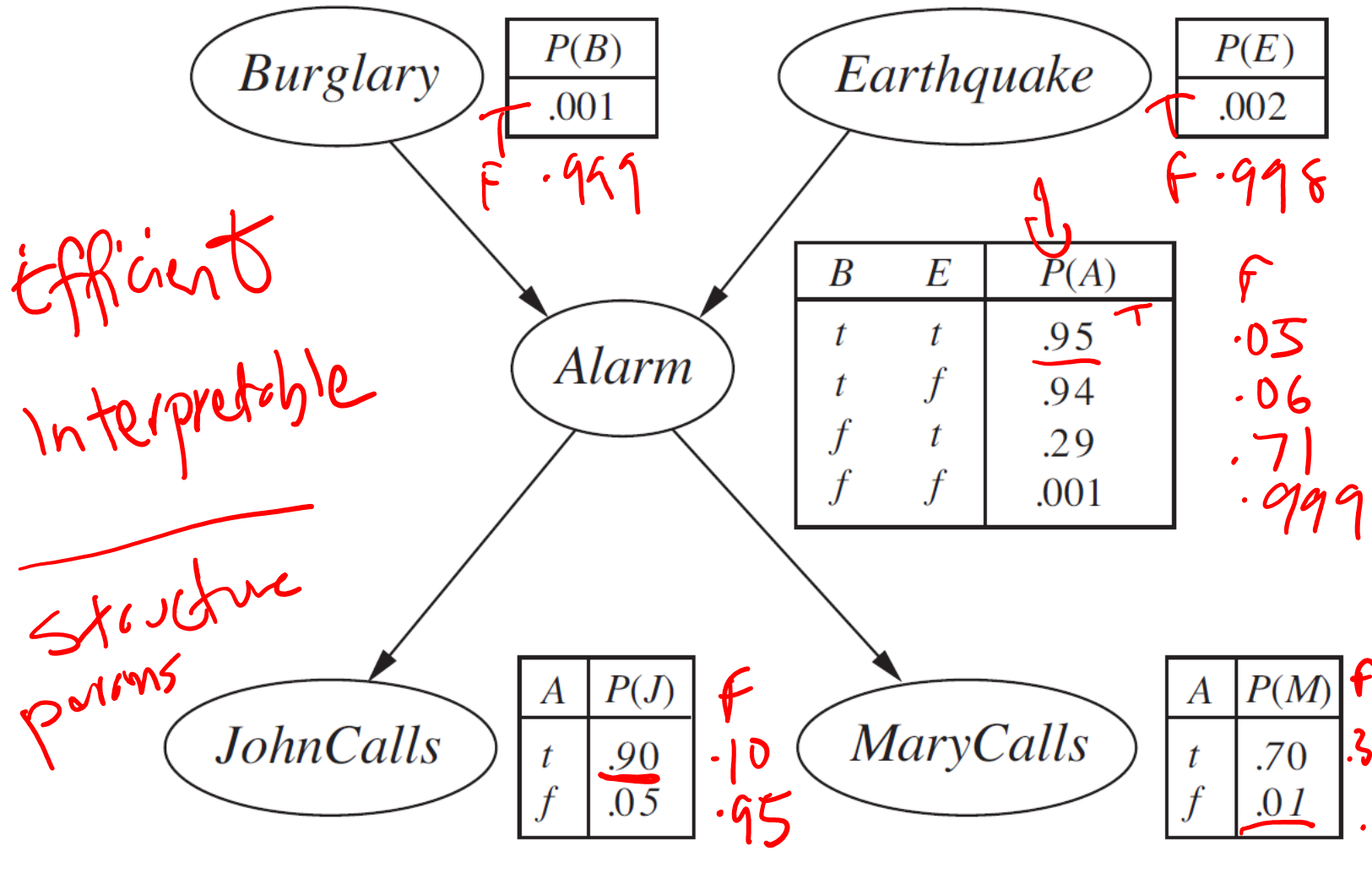
- **Theorem 1:** If a probability distribution P holds the independencies encoded in G , then P factorizes according to G
- **Theorem 2:** If P factorizes according to G , then it holds the independencies encoded in G
- Let's see a constructive proof for Theorem 1; we'll not prove Theorem 2

FROM INDEPENDENCE TO FACTORIZATION

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

$$P(B, E, A, J, M) = P(B) P(E) P(A|B, E) P(J|A) P(M|A)$$

BURGLARY EXAMPLE



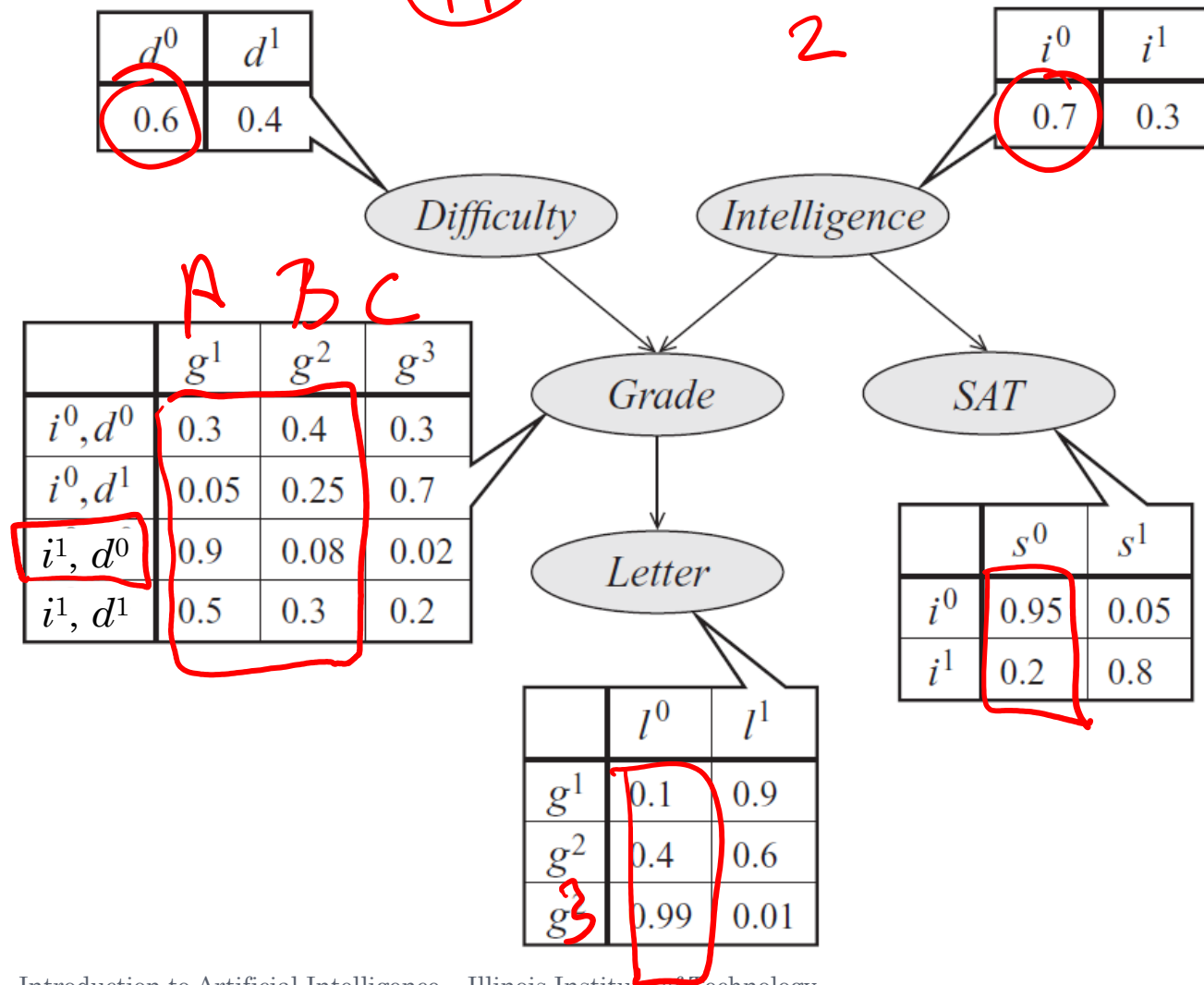
BURGLARY EXAMPLE

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

$P(D, I, S, G, L) = P(D) \cdot P(I) \cdot P(S|I) \cdot P(G|D, I) \cdot P(L|G)$
 $2 \times 2 \times 2 \times 3 \times 2 = 48$

$P(D) = 1$
 $P(I) = 1$
 $P(S|I) = 2$
 $P(G|D, I) = 8$
 $P(L|G) = 15$

STUDENT EXAMPLE



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

$\nabla \rightarrow$ descendants

- From causes to effects
 - E.g., Burglary to Alarm to MaryCalls
 - E.g., Intelligence to Grade to Letter

○ Evidential reasoning

$\nabla \rightarrow$ ancestors

- 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

- There are several methods, some are exact and some are approximate
- We will study only one in this class
- *Variable Elimination*

WHY VARIABLE ELIMINATION?

- Assume

- $D_x = \{2, 3, 4\}$
- $D_y = \{5, 6, 7, 8\}$

- Calculate

- $\sum_x \sum_y x \times y$

VARIABLE ELIMINATION

- Let B, E, A, J, m m $B = t$
 - 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 $P(B=t)P(E)P(A|B=t,E)P(J|A)P(m|A)$
- 2. Set the variables in E to their respective values
- 3. Sum over all variables in $V \setminus (Q \cup E)$ E, J, A
 - a) Pick an order for variables in $V \setminus (Q \cup E)$
 - b) For each variable V_i in $V \setminus (Q \cup E)$, create a new factor by
 - Multiplying all the factors that contains V_i , and ψ
 - Summing over possible values of V_i γ
- 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)
 - $A \rightarrow B \rightarrow C$;
 - $P(A) = \langle 0.6, 0.4 \rangle$,
 - $P(B | A=t) = \langle 0.8, 0.2 \rangle$, $P(B | A=f) = \langle 0.1, 0.9 \rangle$
 - $P(C | B=t) = \langle 0.7, 0.3 \rangle$, $P(C | B=f) = \langle 0.4, 0.6 \rangle$
 - Compute $P(A)$, $P(B)$, $P(C)$, $P(C | A=t)$, $P(A | C=t)$

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