# Lecture 3: January 21, 2015
# cs 573: Probabilistic Reasoning
# Professor Nevatia
# Spring 2015

# Review

- Enrollment: on-campus section is at full seating capacity
  - New students can be added only if and when some current ones drop
- Assignment #1 due Jan 26, class time
- Last lecture:
  - Conditional Independence: $P \models (\mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z})$
  - Expected Values
  - Entropy
  - Graph Terminology
- Today's objective
  - Bayesian Network Representation

# Joint Distribution

- Consider a joint distribution, *P*, over a set of random variables $\{X_1, X_2, \dots X_n\}$
- If variables are discrete, distribution can be represented in an n-dimensional array, each dimension has $m_i$ number of entries.
  - For binary-valued variables, the number of parameters needed to specify *P* is $2^n - 1$.
  - Many problems of interest may have tens or hundreds of variables.
- Given a joint distribution, we can perform desired queries easily
  - Marginal distributions: Sum over variables to be eliminated
  - Compute $P(\mathbf{Y} | \boldsymbol{E} = \boldsymbol{e})$, where $\mathbf{Y}$ is set of query variables; set evidence variables to the given values and read the results directly
  - MAP query: $MAP(\mathbf{W} | \boldsymbol{e}) = \arg\max_w P(\mathbf{w}, \mathbf{e})$; again search over all assignments of **W** and given **e** and read off directly.

# Representing the Joint Distribution

- The number of parameters in joint distribution can be reduced if each variable is not dependent on every other variable
- In an extreme case, any pair of disjoint subsets, say **X** and **Y**, are independent of each other (*i.e.* $\mathbf{X} \perp \mathbf{Y}$). Then,
  - $P(X_1, X_2, \ldots X_n) = P(X_1)\, P(X_2) \ldots P(X_n) = \Pi_i P(X_i)$
  - If $X_i$ are binary valued (such as coin toss results), each variable's distribution is specified by a single parameter, say $\theta_i$ , and $P(x_1, x_2, \ldots x_n) = \Pi_i P(\theta_i)$
  - Only *n* parameters needed to specify the distribution
- Complete independence is usually not of much interest as we want to reason about related entities; we study conditional independence next.

# Conditional Parameterization

- An alternative method to represent the joint distribution
- Consider the example of relation between "Intelligence (I)" and "SAT" scores (both binary valued for this example)
- Joint distribution

| $I$ | $S$ | $P(I,S)$ |
|---|---|---|
| $i^0$ | $s^0$ | 0.665 |
| $i^0$ | $s^1$ | 0.035 |
| $i^1$ | $s^0$ | 0.06 |
| $i^1$ | $s^1$ | 0.24. |

- Alternatively: P(I,S) = P(I) P(S|I). Specify the two terms separately:

| $i^0$ | $i^1$ |
|---|---|
| 0.7 | 0.3 |

P(I), prior distribution over I

| $I$ | $s^0$ | $s^1$ |
|---|---|---|
| $i^0$ | 0.95 | 0.05 |
| $i^1$ | 0.2 | 0.8 |

P(S|I): conditional probability distribution (CPD)

- Note total of 3 independent parameters in both cases but CPD representation is more "natural"

# Naïve Bayes Model: Example

- Consider example with three variables: I, S and G (for grade). Let P |= (S ⊥ G | I).
  - Then, P(I,S,G) = P(I) P(S,G|I) = P(I) P(S|I) P(G|I)

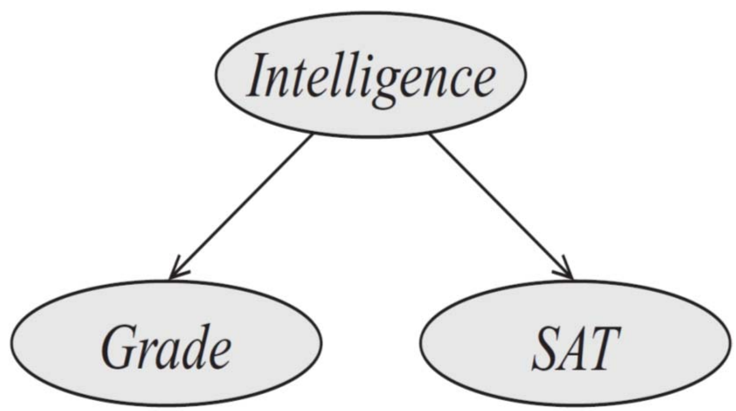
| $i^0$ | $i^1$ |
|---|---|
| 0.7 | 0.3 |

| $I$ | $s^0$ | $s^1$ |
|---|---|---|
| $i^0$ | 0.95 | 0.05 |
| $i^1$ | 0.2 | 0.8 |

| $I$ | $g^1$ | $g^2$ | $g^3$ |
|---|---|---|---|
| $i^0$ | 0.2 | 0.34 | 0.46 |
| $i^1$ | 0.74 | 0.17 | 0.09 |

$$\begin{aligned} P(i^1, s^1, g^2) &= P(i^1)P(s^1 \mid i^1)P(g^2 \mid i^1) \\ &= 0.3 \cdot 0.8 \cdot 0.17 = 0.0408. \end{aligned}$$

- Note, reduction of number of parameters (7 vs 12)
- Graphical model (Bayesian Network)

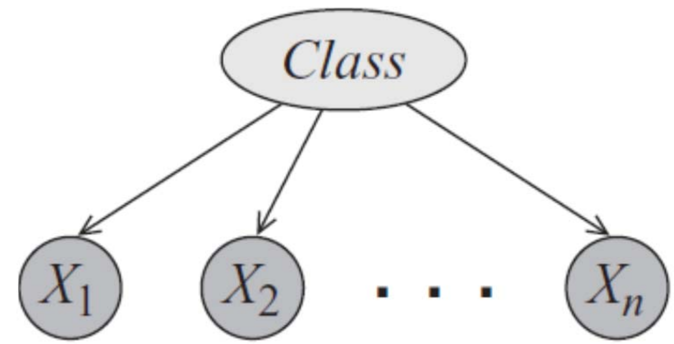


# Naïve Bayes: General

- One variable, called class (or cause) variable, C;
  Several feature (effect) variables, $X_i$;
- Naïve Bayes assumption is that all feature variables are conditional independent given the class

$$(X_i \perp \boldsymbol{X}_{-i} \mid C) \quad \text{for all } i,$$

$$\text{where } \boldsymbol{X}_{-i} = \{X_1, \ldots, X_n\} - \{X_i\}$$

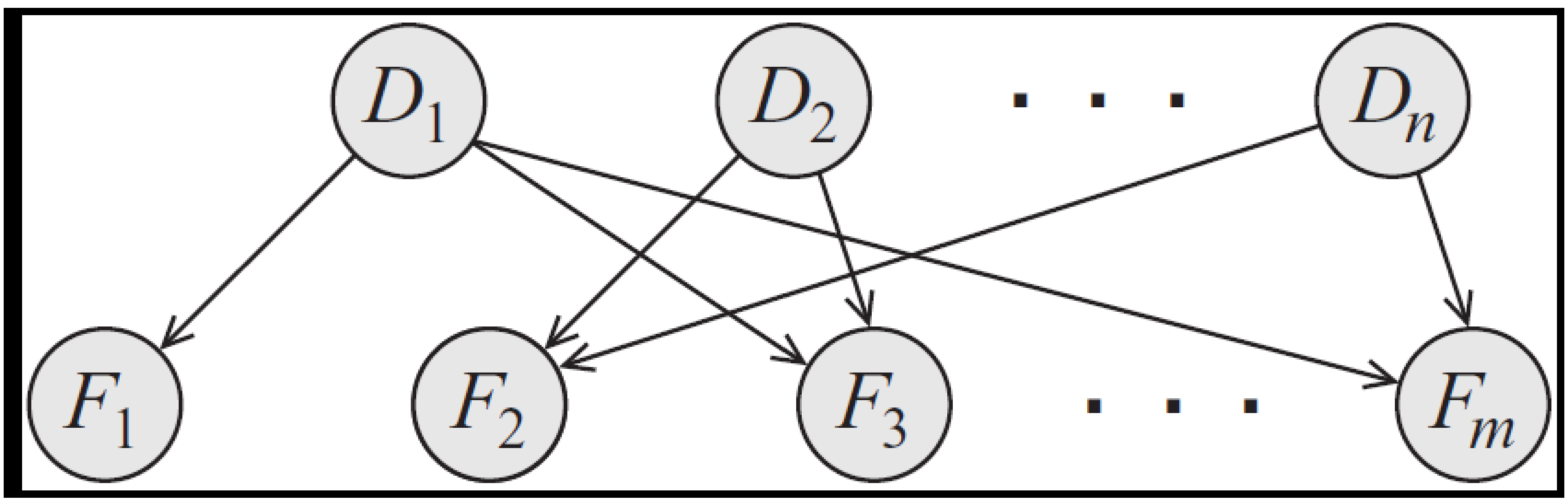



$$P(C, X_1, \ldots, X_n) = P(C) \prod_{i=1}^{n} P(X_i \mid C)$$

Major reduction in number of needed parameters. How much?

# Alternate Representation

- Figure 5.C
- Note: allows multiple, simultaneous disease diagnosis, may be more realistic than naïve Bayes.



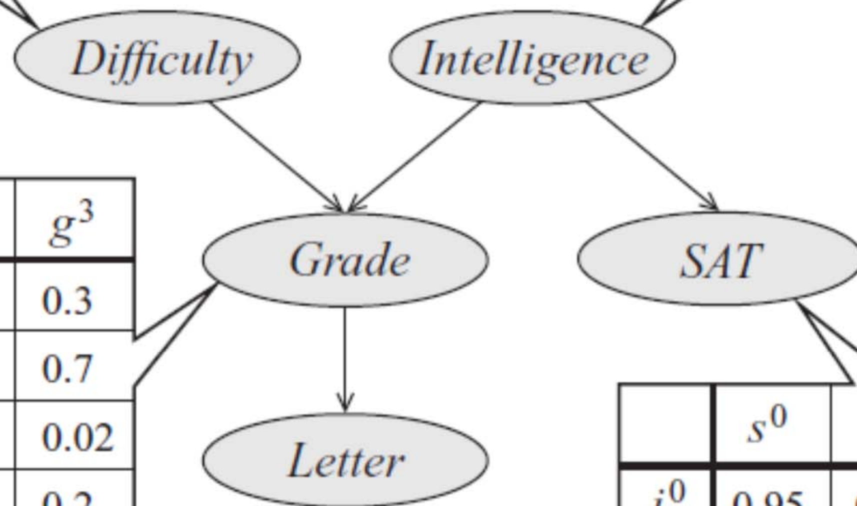
# Naïve Bayes Discussion

- Used for classification tasks such as disease based on symptoms, object based on some features etc
- Compute $P(C| X_1,X_2,\ldots X_n)$
- How to make a decision based on this calculation?
  - Need for prior probabilities of class and evidence
  - Take ratios to remove need for evidence probability
  - Account for cost/utility
- Naïve Bayes assumption is typically not valid but many examples show very good results, often hard to beat by including some dependencies
  - Possible reasons: easier to estimate the smaller number of parameters and to elicit them from an expert.
  - Errors in estimation may dominate the more accurate models constructed by more complex Bayes networks.

# Bayes Network: More Complex 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 |
| $i^0, d^0$ | 0.9 | 0.08 | 0.02 |
| $i^0, d^1$ | 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 |

# Bayes Network Comments

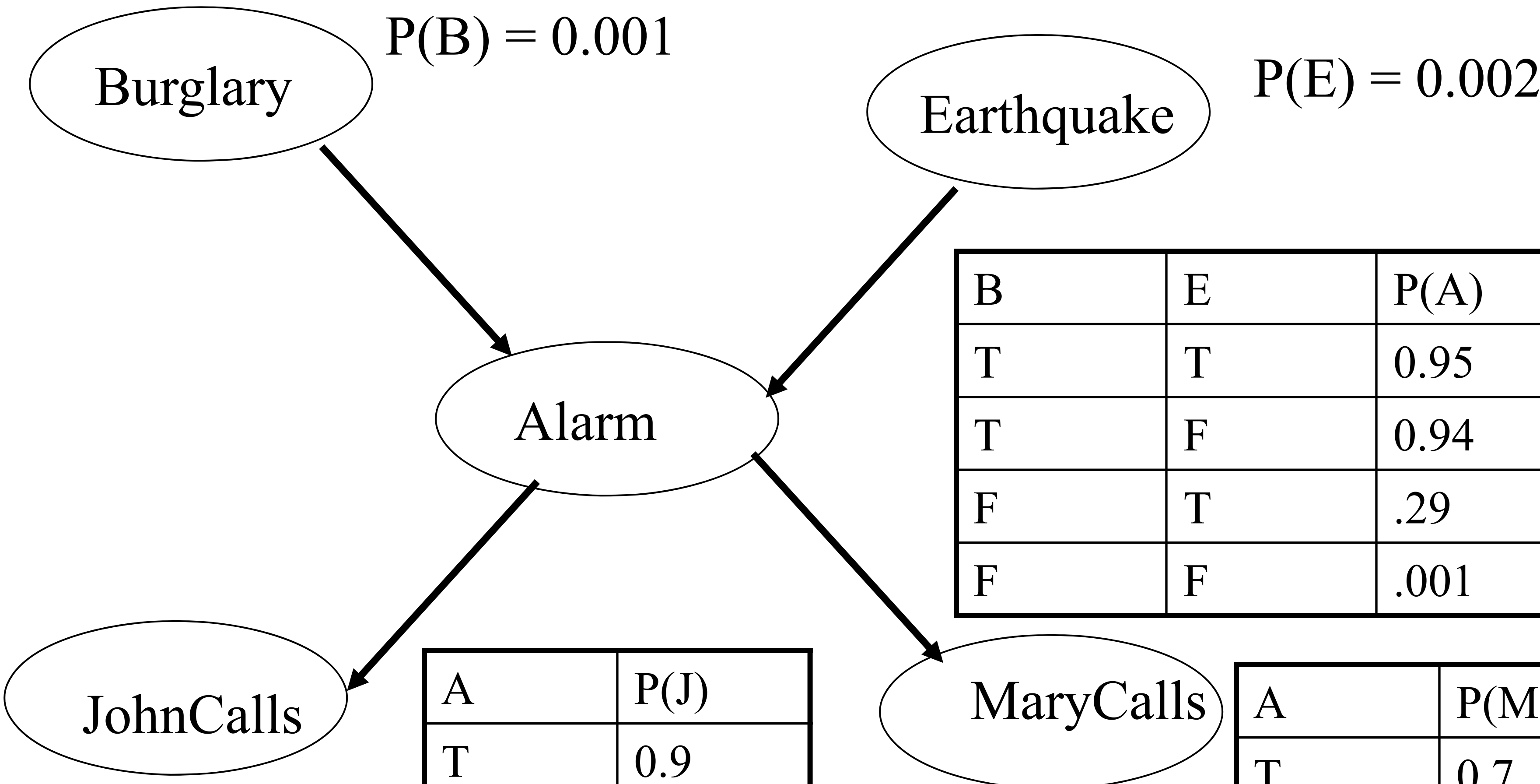
- Note: local probability models, smaller distributions
  - How many parameters vs complete joint distribution?
- Conditional probability distributions (CPDs), also called Conditional probability tables (CPTs) for discrete-valued variables
- Complete joint distribution can be computed from the network
- P (I,D,G,L,S) = P(I)P(D)P(G|I,D)P(S|I)P(L|G)

$$P(i^1,d^0, g^2, s^1, l^0) = P(i^1)\ P(d^0)\ P(g^2|i^1,d^0)\ P(s^1,|i^1)\ P(l^0\,|g^2)$$

$$= .3 \times .6 \times .08 \times .8 \times .4 = .004608$$

- Note: Marginal and conditional distributions can be computed from the joint distribution (e.g. $P(g^2|s^1)$) as before.

# Another Popular Example




| B | E | P(A) |
|---|---|---|
| T | T | 0.95 |
| T | F | 0.94 |
| F | T | .29 |
| F | F | .001 |

| A | P(J) |
|---|---|
| T | 0.9 |
| F | 0.05 |

| A | P(M) |
|---|---|
| T | 0.7 |
| F | 0.01 |

# Reasoning Patterns (Queries)

- BN allows us to compute the complete joint distribution.
  - Joint distribution allows us to answer any query (compute any desired marginal or conditional probabilities).
- **Causal reasoning**:
  - Example: probability that a student gets a strong letter (of recommendation).
  - We can predict without any knowledge of the student (prior probability) or based on some evidence (such as the intelligence of the student) and/or his/her grade etc.
  - Process may be tedious as we may need to compute several terms in the joint distribution and sum them, more efficient algorithms to be studied later.

# Reasoning Patterns (Cont’d)

- **Evidential reasoning**:
  - Reason about intelligence from knowledge of student grade and/or the recommendation letter.
  - Knowledge of one cause may change probability of another cause.
- **Explaining away** (*inter-causal* reasoning):
  - Poor grade may be due to difficulty of course rather than low intelligence

# Independences in BNs

- BN structure implies some conditional independencies.
- For the student example:

$$(L \perp I, D, S \mid G)$$

$$(S \perp D, G, L \mid I)$$

$$(G \perp S \mid I, D)$$

$$(I \perp D).$$

$$(D \perp I, S)$$

How to find these independences in the original distribution?

- *Local* Independencies: a variable is conditional independent of its ancestors (non-descendants), given the parents (all ancestor influence flows through the parents)
- Note: a variable is *not* independent of its descendants, given the parents
- Formal definition in Def 3.1 (next slide)
- There may also exist additional *global* independencies.

# BN Semantics: Local Independencies

*A* Bayesian network structure $\mathcal{G}$ *is a directed acyclic graph whose nodes represent random variables* $X_1, \ldots, X_n$. *Let* $\mathrm{Pa}^{\mathcal{G}}_{X_i}$ *denote the parents of* $X_i$ *in* $\mathcal{G}$, *and* $\mathrm{NonDescendants}_{X_i}$ *denote the variables in the graph that are not descendants of* $X_i$. *Then* $\mathcal{G}$ *encodes the following set of conditional independence assumptions, called the* local independencies, *and denoted by* $\mathcal{I}_\ell(\mathcal{G})$:

*For each variable* $X_i$: $(X_i \perp \mathrm{NonDescendants}_{X_i} \mid \mathrm{Pa}^{\mathcal{G}}_{X_i})$. ■

# I-Maps

- Two ways to specify independence relations: assertions and BN structure. What are the relations between the two?
- We introduce some formal notation and assertions but skip the proofs
- I-Maps
  - I (P): set of independence assertions that hold in P
  - $I_L(G)$: set of local independences implied in graph G
  - If $I_L(G) \subseteq I(P)$, G is said to be an I-map of P
  - *Note*: G may assert *fewer* independencies than P.
  - A fully connected graph is an I-map of any P.
  - A more general definition in Def 3.3
- D-Map (not in book): all dependencies in graph also present in P
  - Graph with no edges is a trivial example
- P-map is both an I-map and a D-map (more on P-map later)

# Formal Definitions

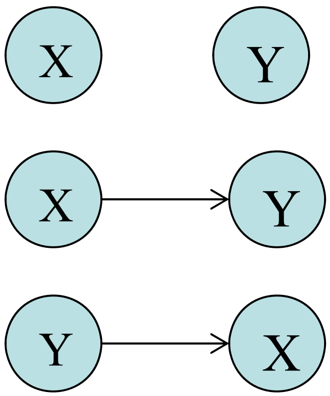
*Let $P$ be a distribution over $\mathcal{X}$. We define $\mathcal{I}(P)$ to be the set of* independence assertions *of the form $(\boldsymbol{X} \perp \boldsymbol{Y} \mid \boldsymbol{Z})$ that hold in $P$.* ∎

We can now rewrite the statement that "$P$ satisfies the local independencies associated with $\mathcal{G}$" simply as $\mathcal{I}_\ell(\mathcal{G}) \subseteq \mathcal{I}(P)$. In this case, we say that $\mathcal{G}$ is an *I-map* (independency map) for $P$. However, it is useful to define this concept more broadly, since different variants of it will be used throughout the book.

*Let $\mathcal{K}$ be any graph object associated with a set of independencies $\mathcal{I}(\mathcal{K})$. We say that $\mathcal{K}$ is an* I-map *for a set of independencies $\mathcal{I}$ if $\mathcal{I}(\mathcal{K}) \subseteq \mathcal{I}$.* ∎

# Independent Variables

- Given distribution, is it decomposable? Is P(X,Y) = P(X) P(Y) ?
- How to establish this? Compute P(X) and P (Y).



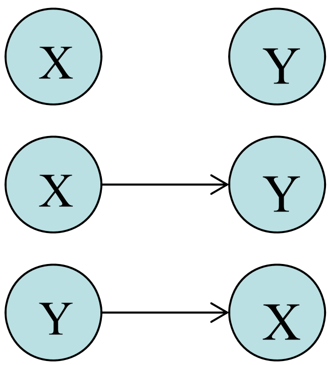

| X | Y | P(X,Y) |
|---|---|---|
| 0 | 0 | .08 |
| 0 | 1 | .32 |
| 1 | 0 | .12 |
| 1 | 1 | .48 |

P(X,Y) = P(X) P(Y)

Are any of the three graphs are I-maps for this example?

Note: a fully connected graph is an I-map for any distribution

# Non-independent Variables



| X | Y | P(X,Y) |
| --- | --- | --- |
| 0 | 0 | .4 |
| 0 | 1 | .3 |
| 1 | 0 | .2 |
| 1 | 1 | .1 |

$$P(X,Y) \neq P(X)\,P(Y)$$

All three graphs are I-maps for this example

# Factorization

- Given five variables, I,D,G,L,S joint can always be represented as $P(I,D,G,L,S) = P(I)P(D|I)P(G|I,D)P(L|I,D,G)P(S|I,D,G,L)$
- If we know (assume) that a graph is an I-map of the distribution P, we can simplify the terms based on independence assertions. For example:

  Given that $(I \perp D) \in I(P)$, we can write $P(D|I) = P(D)$

  Given $(L \perp I, D|G) \in I(P)$, we can write $P(L|I,D,G) = P(L|G)$

  Given $(S \perp D,|G,L|I) \in I(P)$, we can write $P(S|I,D,G,L)= P(S|I)$

- $P(I,D,G,L,S) = P(I)P(D|I)P(G|I,D)P(L|G)P(S|I)$

# Factorization Theorems

*Let $\mathcal{G}$ be a BN graph over the variables $X_1, \ldots, X_n$. We say that a distribution $P$ over the same space* factorizes *according to $\mathcal{G}$ if $P$ can be expressed as a product*

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \mid \mathrm{Pa}^{\mathcal{G}}_{X_i}). \tag{3.17}$$

Thm 3.1

*Let $\mathcal{G}$ be a BN structure over a set of random variables $\mathcal{X}$, and let $P$ be a joint distribution over the same space. If $\mathcal{G}$ is an I-map for $P$, then $P$ factorizes according to $\mathcal{G}$.*

Thm 3.2

*Let $\mathcal{G}$ be a BN structure over a set of random variables $\mathcal{X}$ and let $P$ be a joint distribution over the same space. If $P$ factorizes according to $\mathcal{G}$, then $\mathcal{G}$ is an I-map for $P$.*

# Knowledge Engineering

- How to pick variables?
  - Variables that we observe and whose distributions are important
  - Sometimes, "hidden" variables may also be important as they may provide a simpler model
- How to pick structure?
  - Causal structure usually leads to simpler graphs
- How to pick probabilities?
  - Elicit from experts
  - Learn from data
  - Avoid zero values!

# Next Class

- Read sections 3.3, 3.4.1 and 3.4.2 of the KF book