

CS 583: PROBABILISTIC GRAPHICAL MODELS

BAYESIAN NETWORKS



Mustafa Bilgic



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



<https://twitter.com/bilgicm>

MOTIVATION

- We would like to represent a joint distribution P over $\mathcal{X} = \{X_1, X_2, \dots, X_n\}$
- Why is such a P useful?
- The naïve representation \Rightarrow Specify a value for each possible combination
- If all X_i are binary, how many numbers are needed to represent P with 1000 variables?
- How many atoms in the observable universe?

WHY IS 2^N BAD?

○ Computational challenges

- Answering queries requires manipulating exponential number of entries
- Storing exponential number of entries is almost always impossible

○ Cognitive challenges

- How can we wrap our minds around about a specific assignment and its corresponding, extremely small, probability?
- How can an expert provide those numbers or even verify they are correct?

○ Statistical challenges

- We often would like to estimate probabilities from data; estimation requires repetition. How can we have a dataset where an exponential number of events are present and repeated multiple times?

WE WOULD LIKE TO HAVE A REPRESENTATION THAT IS

- Compact
 - Easy to store, manipulate, understand, and estimate
- Intuitive
 - Easy to understand, verify, and construct
- Modular
 - Easy to add and remove variables
- Declarative
 - Separates representation and reasoning

NUMBER OF PARAMETERS

- Please see OneNote

COMPACTNESS

- A joint distribution P over $\mathcal{X} = \{X_1, X_2, \dots, X_n\}$ requires exponential number of numbers
- Reduce the number of parameters through independence
 1. Marginal independence
 2. Conditional independence

MARGINAL INDEPENDENCE

- Represent a joint distribution P over $\mathcal{X} = \{X_1, X_2, \dots, X_n\}$, where all X_i are binary
- How many independent parameters? $2^n - 1$
- Assume for $\forall i \neq j, X_i \perp X_j$
- $P(\mathcal{X}) = P(X_1)P(X_2)\dots P(X_n)$
- Now, how many independent parameters? n

CONDITIONAL PROBABILITIES

- Two variables, Industrious (I) and SAT score (S)
- Industrious: false (i^0), true (i^1)
- SAT score: low (s^0), high (s^1)

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

What's the number of independent parameters needed?

3

CONDITIONAL PROBABILITIES

- $P(I, S) = P(I)P(S | I)$

$P(I)$

i^0	i^1
0.7	0.3


$P(S | I)$

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

What's the number of independent parameters needed?

3

A NEW VARIABLE

- Let's add the variable grade (G), with three possible values, A (g^1), B (g^2), and C (g^3).
- Can we assume that G is independent of I or S in real life?
- A more reasonable assumption is to assume that I determines S and G ; that is, S and G are conditionally independent 
 - This is not totally true either, but then, in the real-world, we cannot really assume anything is independent of anything
 - Butterfly effect?

$$P(I) P(S|I) P(G|S, I)$$

CONDITIONAL INDEPENDENCE

- $P(I, S, G) = \underline{P(S, G | I)} P(I) = \underline{P(S | I)} P(G | I) P(I)$
- $P(I, S, G) = P(I) P(S | I) P(G | S, I) = P(I) P(S | I) P(G | I)$
- We already have $P(I)$ and $P(S | I)$. We need to specify $P(G | I)$

$P(G | I)$

$$2 \times (3 - 1) = 4$$

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

$$2 \times 2 \times 3 - 1$$

What's the number of independent parameters needed for $P(I, S, G)$ if we use the full joint table?

What if we use the factorization $P(I, S, G) = P(I) P(S | I) P(G | I)$?

$$1 + 2 + 4 = 7$$

CONDITIONAL INDEPENDENCIES

- Compactness

- Fewer parameters to specify

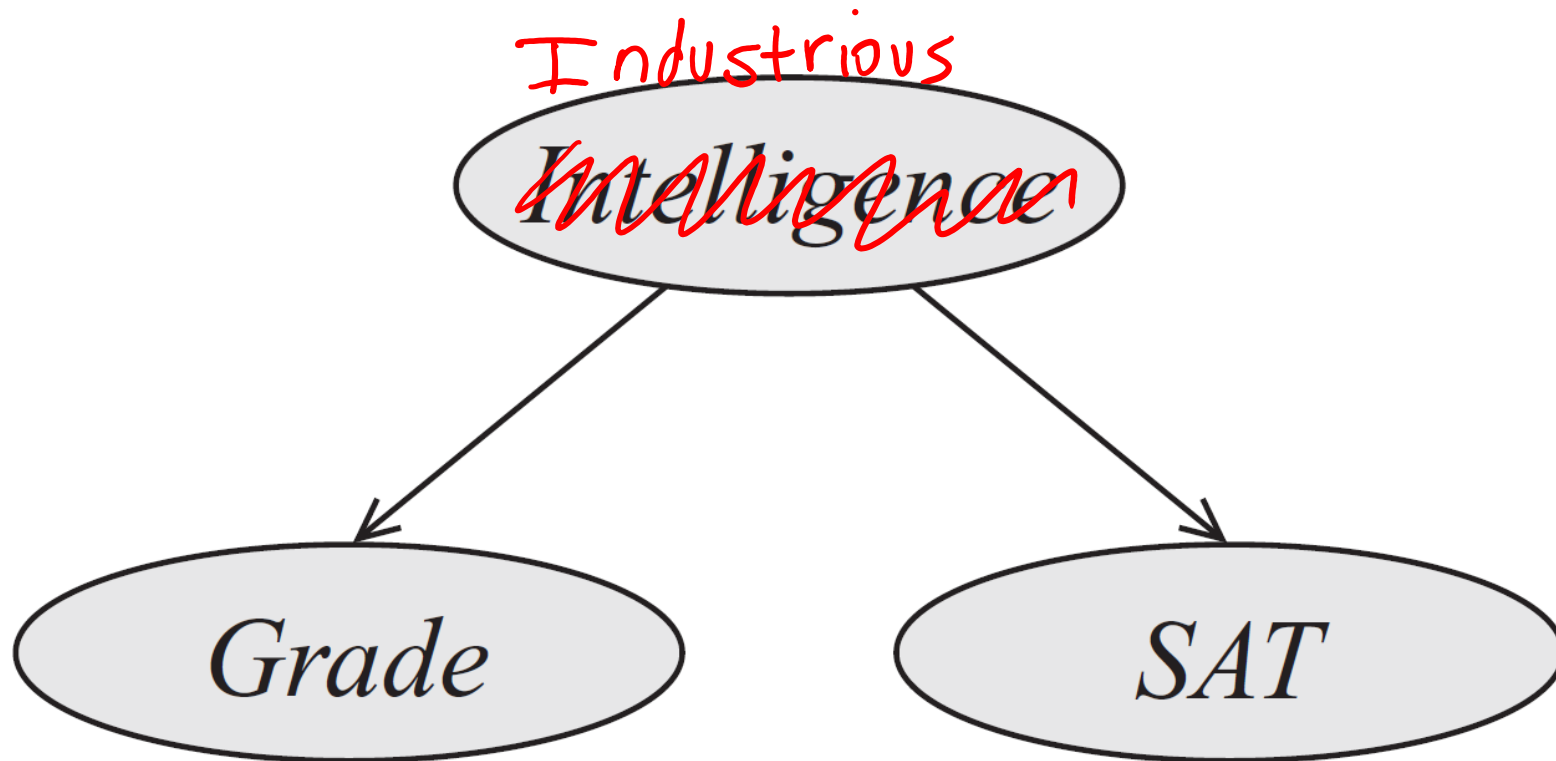
- Intuition

- Easier to specify

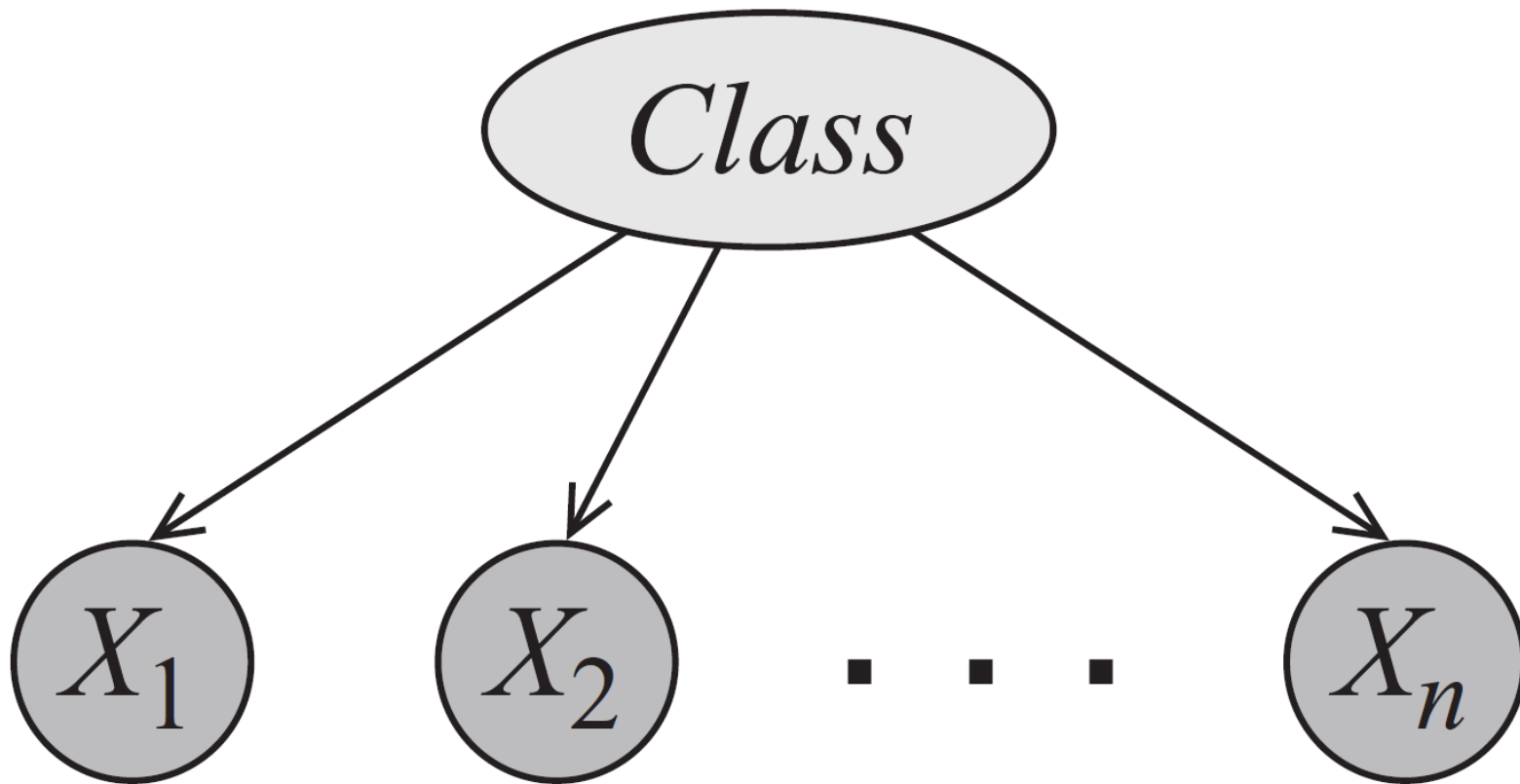
- Modularity

- Adding a new variable does not cause us to change all the entries in the joint table

BAYESIAN NETWORK REPRESENTATION



NAÏVE BAYES



NAÏVE BAYES

- How do we write the joint $P(C, X_1, X_2, \dots, X_n)$?
- How many independent parameters are needed if C and X_i are all binary?
- Naïve Bayes is used for **classification**: given attributes of an object (X_i), classify it into one of pre-given categories (C) (i.e., $P(C | X_1, \dots, X_n)$).
- Read Box 3.A in the textbook for more details

BAYESIAN NETWORKS

DAGs

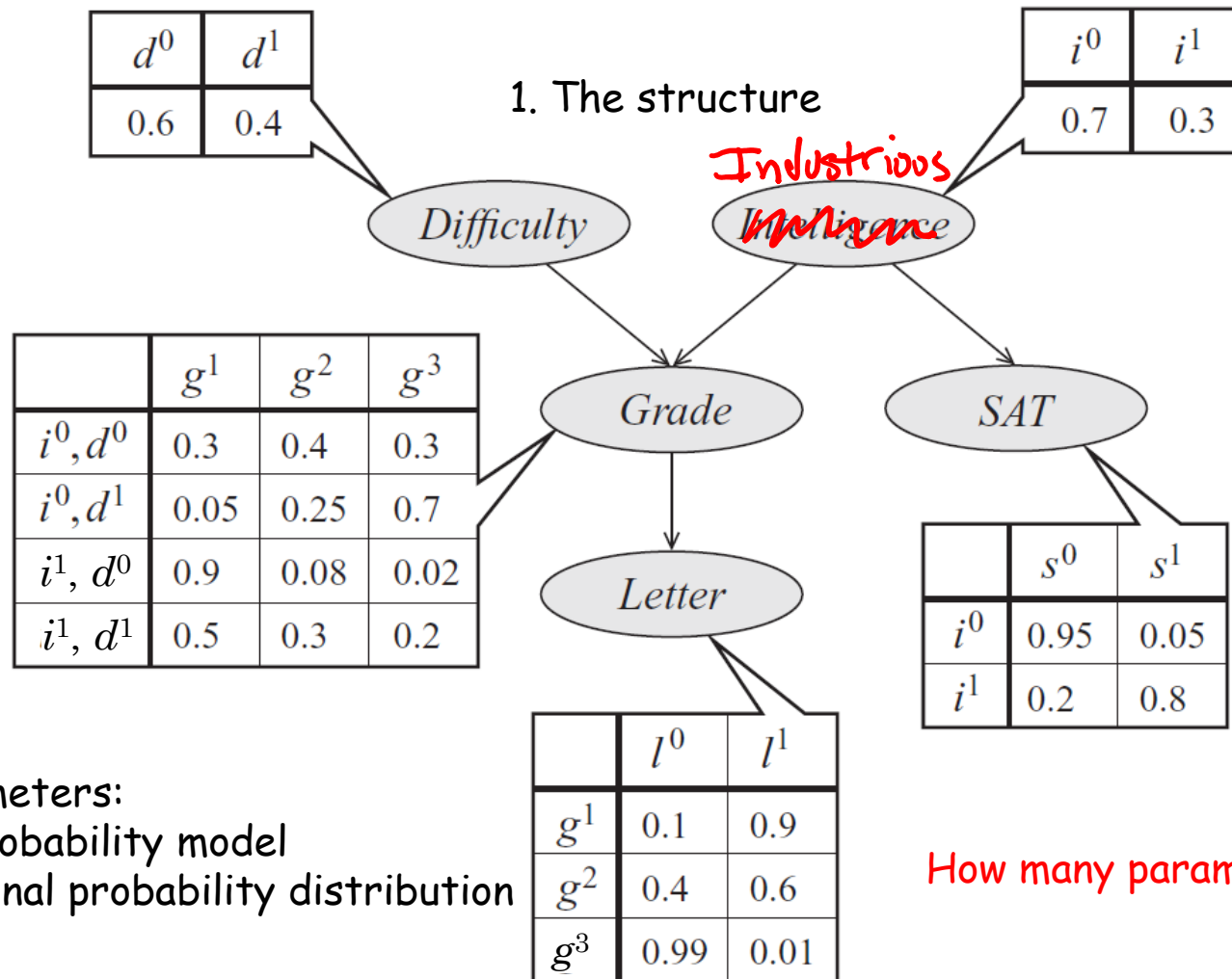
- *A Bayesian Network is a directed acyclic graph whose nodes are random variables and edges represent, intuitively, the direct influence of one node on another*
- Naïve Bayes is a special Bayesian network
- Bayesian networks is
 - A data structure that provides the skeleton for representing a joint distribution compactly in a factorized way
 - A compact representation for a set of conditional independence assumptions about a distribution

THE STUDENT EXAMPLE

- So far we have I, S, G
- We add two more random variables
 - Student's grade also depends on the difficulty (D) of the class: $\text{Val}(D) = \{\text{easy}(d^0), \text{hard}(d^1)\}$
 - Student's professor writes a recommendation letter (L), where $\text{Val}(L) = \{\text{weak}(l^0), \text{strong}(l^1)\}$
 - Professor writes the letter based only the grade and it is a stochastic function of the grade

$$P(I, S, G, D, L)$$

THE STUDENT NETWORK



2. Parameters:

Local probability model

Conditional probability distribution

How many parameters?

THE JOINT?

- What is the meaning of $P(i^1, d^0, g^2, s^1, l^0)$?
- Probability that
 - The student is industrious
 - The class is easy
 - The smart student gets a B in an easy class
 - The smart students get a high score in SAT
 - The student who got a B in the class gets a weak letter
 - $= P(i^1) P(d^0) P(g^2 | i^1, d^0) P(s^1 | i^1) P(l^0 | g^2)$

REASONING PATTERNS

- Causal reasoning
 - Causes to effects
- Evidential reasoning
 - Effects to causes
- Intercausal reasoning
 - Explaining away

CAUSAL REASONING

- Causes to effects
- Don't know anything. Probability of a strong letter
 - $P(l^1) = 0.502$
- Learn that the student is not industrious
 - $P(l^1 | i^0) = 0.389$
- Additionally, learn the class is easy
 - $P(l^1 | i^0, d^0) = 0.513$

EVIDENTIAL REASONING

- Effects to causes
- Don't know anything. Probability of a student being industrious
 - $P(i^1) = 0.3$
- Learn that the student got a C in a class
 - $P(i^1 | g^3) = 0.079$
- Or, learn that the student received a weak letter
 - $P(i^1 | l^0) = 0.14$
- Learn both
 - $P(i^1 | g^3, l^0) = 0.079$

INTERCAUSAL REASONING

- Different causes of the same effect interact
- Don't know anything. Probability of a student being industrious
 - $P(i^1) = 0.3$
- Learn that the student got a B in a class
 - $P(i^1 | g^2) = 0.175$
- Learn that the class was difficult
 - $P(i^1 | g^2, d^1) = 0.34$
- Student's B is *explained away* with the other cause

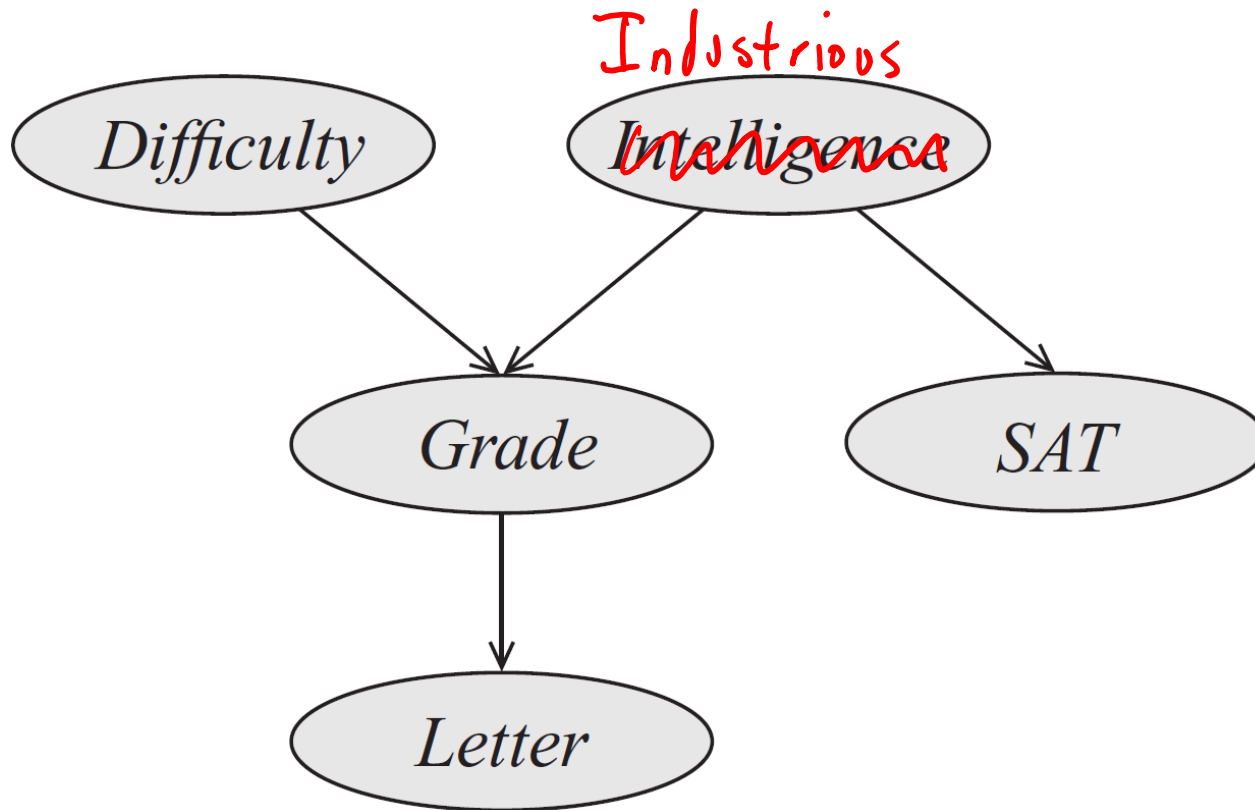
HUGIN LITE

- Download Hugin Lite
 - <https://www.hugin.com/index.php/hugin-lite/>
- Get the student example from GitHub
- Try a few causal, evidential, and intercausal queries
- Extra fun
 - Get more .net files from <https://www.bnlearn.com/bnrepository/> and load them to Hugin and play with them
- Double-extra fun
 - Install pgmpy Python package <https://github.com/pgmpy/pgmpy>
 - Check out the examples at <http://pgmpy.org/>

BAYESIAN NETWORK STRUCTURE

- A *Bayesian network structure* \mathcal{G} is a directed acyclic graph whose nodes represent random variables X_1, \dots, X_n . Let $\text{Pa}(X_i)$ denote the parents of X_i , and $\text{ND}(X_i)$ denote the variables that are not descendants of X_i . Then \mathcal{G} encodes the following set of conditional independence assumptions:
 - $X_i \perp \text{ND}(X_i) \mid \text{Pa}(X_i)$
- These independencies are called the *local independencies*
- Clarification. A node itself and its parents are part of non-descendants according the definition of ND. A clearer statement would be, in my opinion:
 - $X_i \perp \text{ND}(X_i) \setminus \{X_i \cup \text{Pa}(X_i)\} \mid \text{Pa}(X_i)$

LOCAL INDEPENDENCIES EXAMPLE



1. $D \perp I, S$
2. $I \perp D$
3. $S \perp D, G, L \mid I$
4. $G \perp S \mid D, I$
5. $L \perp D, I, S \mid G$

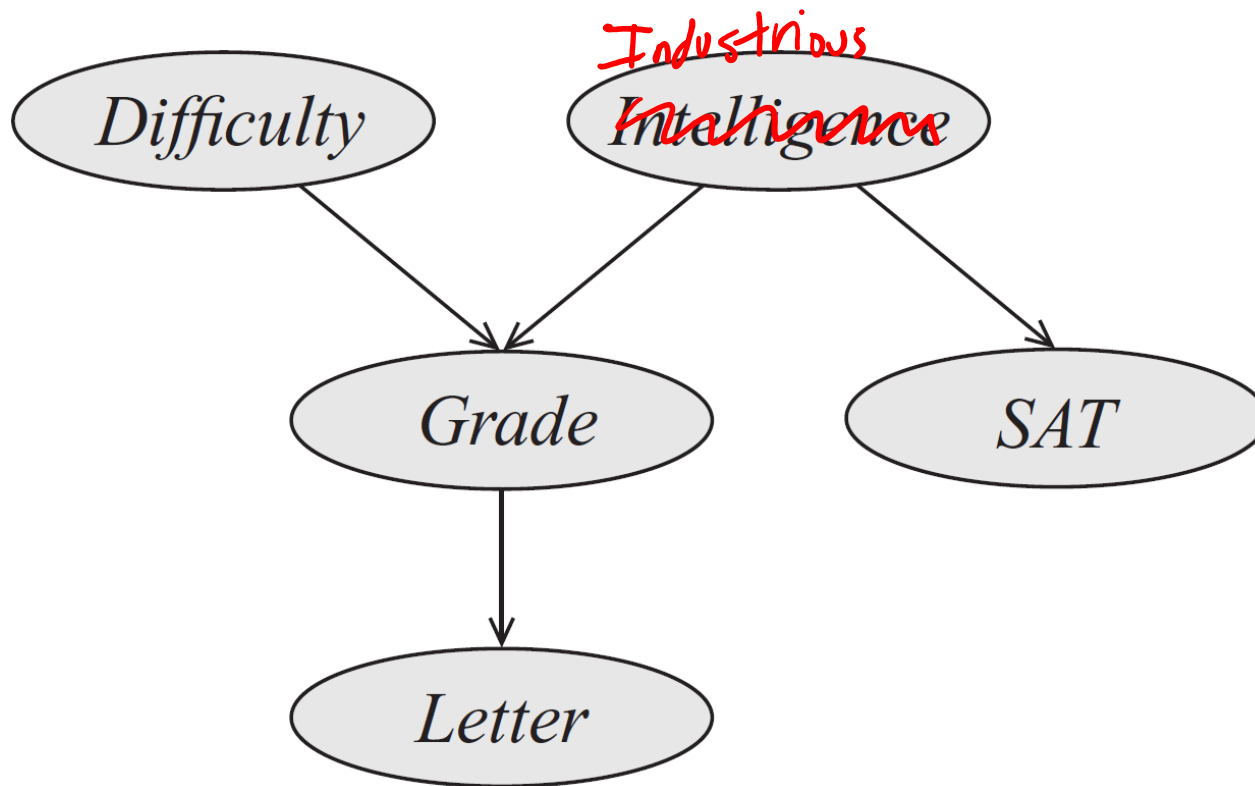
BAYESIAN NETWORK FACTORIZATION

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

Why is this factorization useful?

How can you prove this factorization holds?

FACTORIZATION EXAMPLE



$$\begin{aligned} P(I, D, S, G, L) = & \\ & P(I)^* \\ & P(D)^* \\ & P(S \mid I)^* \\ & P(G \mid I, D)^* \\ & P(L \mid G) \end{aligned}$$

WE'LL PROVE

1. Local conditional independence assertions \Rightarrow Bayesian network factorization
2. Bayesian network factorization \Rightarrow Local conditional independence assertions

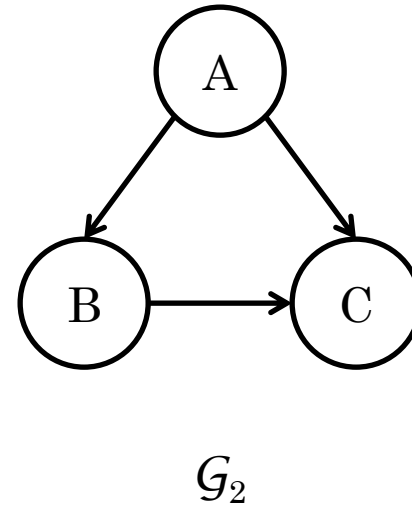
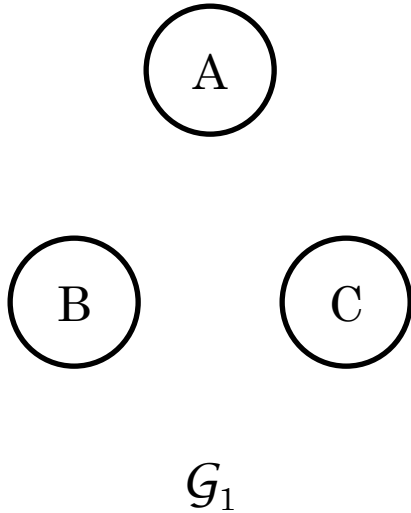
I-MAP

- Let P be a distribution over \mathcal{X} . We define $I(P)$ to be the set of independencies of the form $\mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z}$ that hold in P .
- “ P satisfies independencies of the structure \mathcal{K} ” $\equiv I(\mathcal{K}) \subseteq I(P)$
- If $I(\mathcal{K}) \subseteq I(P) \Rightarrow \mathcal{K}$ is an I-Map for P
- For \mathcal{K} to be an I-Map of P , any independence assertion made by \mathcal{K} has to be true in P . However, it is possible that P can contain additional independencies
 - That is, whatever \mathcal{K} says independent is independent in P , but \mathcal{K} may not know all the independencies in P

P-MAP

- **Definition:** *Perfect Map*: A graph structure \mathcal{G} is a perfect map (P-Map) for a distribution P , if $I(\mathcal{G}) = I(P)$

TWO GRAPHS



What are the local independence assertions made by G_1 ? G_2 ?

WE'LL PROVE

1. Local conditional independence assertions \Rightarrow Bayesian network factorization
 - I-Map to Factorization
2. Bayesian network factorization \Rightarrow Local conditional independence assertions
 - Factorization to I-Map

I-MAP TO FACTORIZATION

- **Theorem**: Let G be a BN structure over \mathcal{X} , and let P be joint distribution over the same space. If G is an I-Map for P , then P factorizes according to G .
- **Proof?**

FACTORIZATION TO I-MAP

- Theorem: Let G be a BN structure over X and let P be a joint distribution over the same space. If P factorizes over G , then G is an I-Map for P .
- Proof?

INDEPENDENCIES IN GRAPHS

- We have already seen the local independence assertions
- Are there additional independence statements we can make simply based on the graph structure?
 - Yes.
- Will these additional independencies help us reduce the number parameters further?
 - No.
- Why are they useful then?

DEPENDENCE \equiv INFORMATION FLOW

- X and Y are independent if the information cannot flow from one to the other
 1. No trail between X and Y
 - Information cannot flow; X and Y are independent
 2. X and Y are connected by an edge
 - Information can flow; X and Y are not independent
 3. X and Y are not connected by an edge, but there is a trail between them
 - Depends...

TRAILS 1-2

○ Causal trail

- $X \rightarrow Z \rightarrow Y$
- Information can flow between X and Y if Z is not observed; if Z is observed, it blocks the information flow. $X \perp Y \mid Z$
- E.g, $I \rightarrow G \rightarrow L$
 - If we don't know the student's grade, then knowing her intelligence tells us something about the strength of the letter. But, once we know her grade, I and L are independent

○ Evidential trail

- $X \leftarrow Z \leftarrow Y$
- Information can flow between X and Y if Z is not observed; if Z is observed, it blocks the information flow. $X \perp Y \mid Z$

TRAILS 3

- Common cause

- $X \leftarrow Z \rightarrow Y$
- Information can flow between X and Y if Z is not observed; if Z is observed, it blocks the information flow.
 $X \perp Y \mid Z$
- E.g, $G \leftarrow I \rightarrow S$
 - If we don't know the student's intelligence, then knowing his SAT score tells us something about his grade. If we know his intelligence, then his grade and SAT score are independent

TRAILS 4

- Common effect
 - $X \rightarrow Z \leftarrow Y$ (v-structure)
 - Information can flow between X and Y only if Z or at least one of Z 's descendants is observed
 - E.g. $D \rightarrow G \leftarrow I$
 - If we do not know the grade or the letter quality, then D and I are independent. If we know, however, the grade or the letter quality, then D and I interact

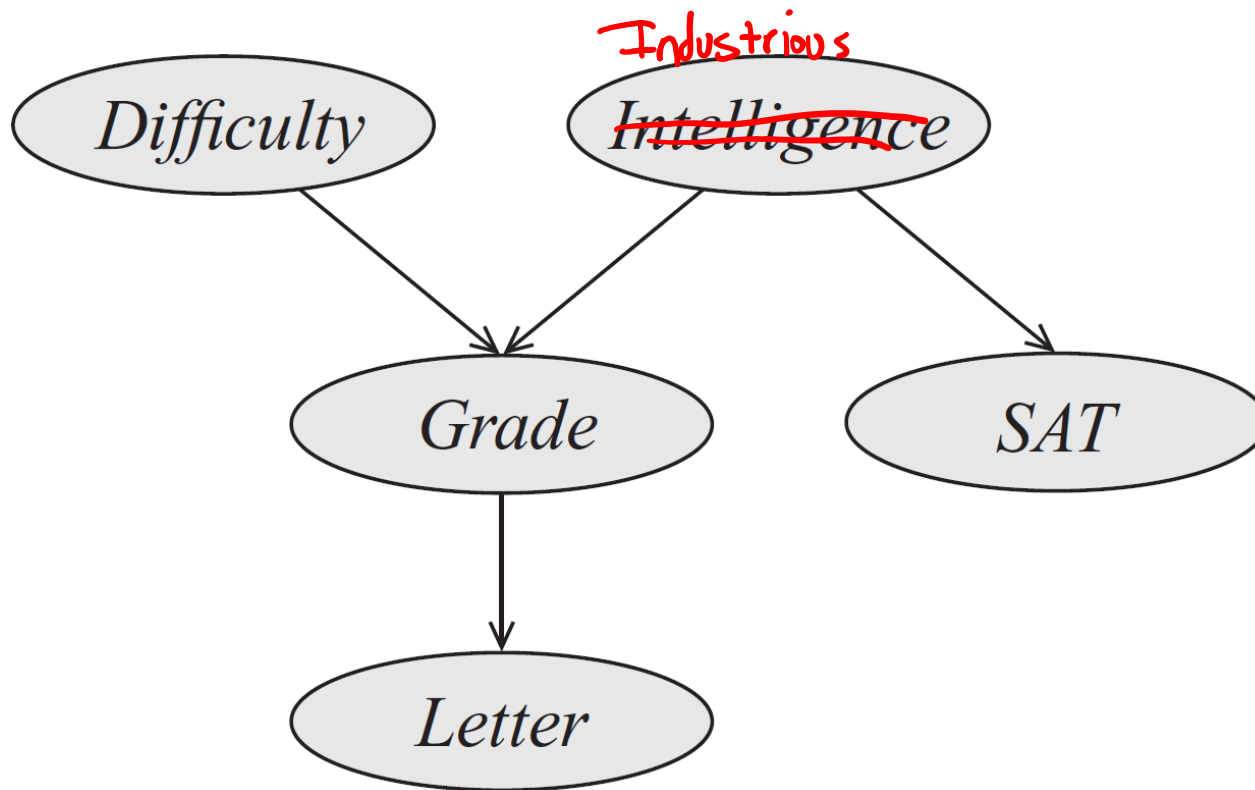
D-SEPARATION

- To answer whether X_i and X_j are independent given \mathbf{E}
 - Find all trails between X_i and X_j
 - If information can flow through at least one trail, then X_i and X_j are dependent given \mathbf{E} ; otherwise they are independent given \mathbf{E}

MORE FORMALLY

- Let \mathcal{G} be a BN structure and $X_1 \leftrightarrow \dots \leftrightarrow X_n$ a trail in \mathcal{G} . Let \mathbf{E} be the observed set of variables. The information can flow along the trail $X_1 \leftrightarrow \dots \leftrightarrow X_n$ if
 - Whenever we have a v-structure $X_{i-1} \rightarrow X_i \leftarrow X_{i+1}$, then X_i or one of its descendants is in \mathbf{E} ; and
 - No other node along the trail is in \mathbf{E}

D-SEPARATION EXAMPLE



Are the following statements true?

1. $D \perp G ?$
2. $D \perp L ?$
3. $D \perp I ?$
4. $D \perp I | G ?$
5. $D \perp I | L ?$
6. $D \perp S | L ?$

I-EQUIVALENCE

- If two BN structures G_i and G_j encode exactly the same independencies, i.e., if $I(G_i) = I(G_j)$, then G_i and G_j are I-equivalent
- Why is this an important concept?
- A given distribution can be represented with one of I-equivalent structures and it might be impossible to identify a unique structure
 - E.g., $X \rightarrow Y$ and $X \leftarrow Y$ are I-equivalent. Then, we cannot readily argue whether X causes Y or Y causes X.

I-EQUIVALENCE

- *Definitions:*

- *Skeleton* of Bayesian network G is the undirected version of G
 - That is, G and its skeleton have the same node and edge set, except edges in G are directed, whereas in its skeleton, the edges are undirected
- A v-structure $X \rightarrow Z \leftarrow Y$ is an *immorality* if there is no direct edge between X and Y (informally, it is an immorality if the parents are not married)
- *Theorem:* Two Bayesian networks are I-Equivalent if and only if they have the same skeleton and the same set of immoralities

FROM DISTRIBUTIONS TO GRAPHS

- For a given P , we would like to find a structure that represents P
- One approach is to take any graph that is an I-Map for P
 - This is not a good idea. Why?
- **Definition: Minimal I-Map:** A graph G is a minimal I-Map for a P , if it is an I-Map for P , and removal of a single edge from G renders it to be not an I-Map

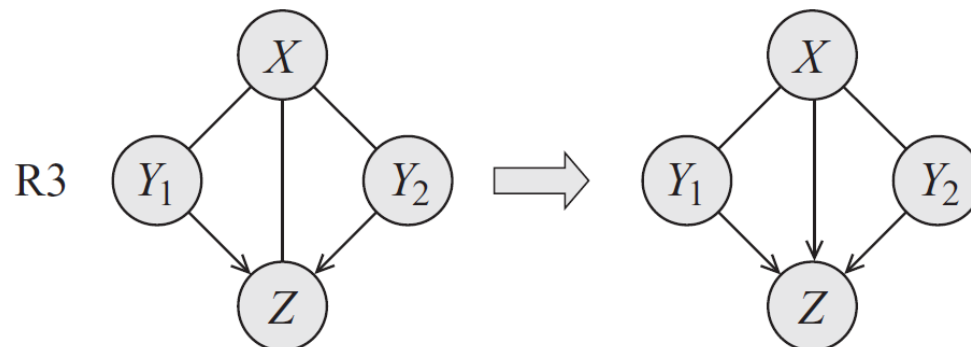
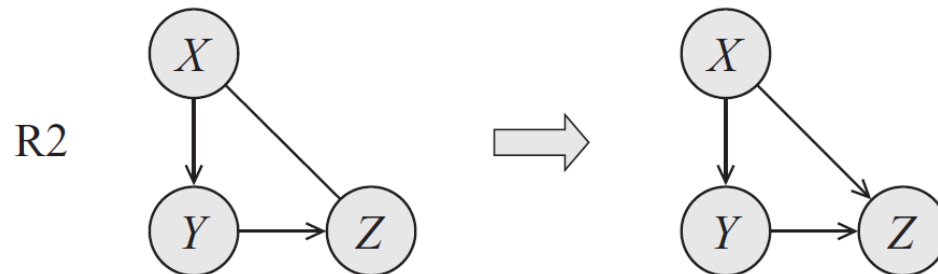
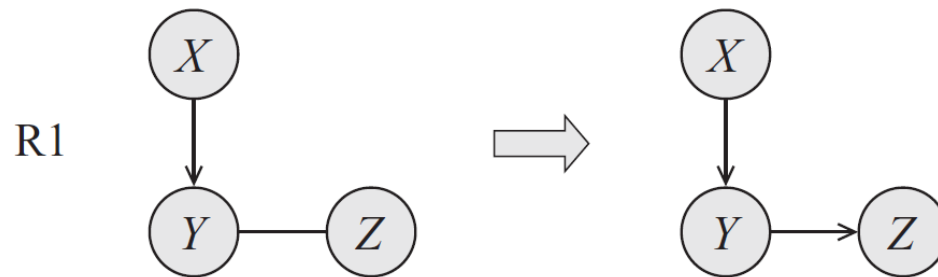
FINDING MINIMAL I-MAPS

- Given a distribution P over \mathcal{X} , how can we find a structure G that is a minimal I-Map for P ?
- The procedure is given on page 79
 - Pick an ordering of the variables, X_1, X_2, \dots, X_n
 - For each X_i , find minimal subset \mathbf{U} of $\{X_1, X_2, \dots, X_{i-1}\}$ such that $X_i \perp \{X_1, X_2, \dots, X_{i-1}\} \setminus \mathbf{U} \mid \mathbf{U}$
 - Set \mathbf{U} to be the parents of X_i
- Assume $I(P^{\text{student}}) = I(G^{\text{student}})$. Construct a minimal network using the order
 - D, I, S, G, L
 - L, S, G, I, D
 - L, D, S, I, G

FINDING I-EQUIVALENT STRUCTURES

- Start with a fully connected undirected graph
- Assume a max indegree of d
- For all pairs X - Y
 - Search for a set U where $X \perp Y \mid U$ and $|U| \leq d$ (U has to be a subset of neighbors of X or neighbors of Y)
 - If such a U cannot be found, then X and Y are connected, otherwise,
 - Remove the edge X - Y and record U as the witness set for X and Y
- Find all immoralities; For all X - Z - Y where X and Y are not directly connected
 - If $Z \in U$, then X - Z - Y is not an immorality, otherwise
 - It is an immorality, orient the edges as $X \rightarrow Z \leftarrow Y$
- Orient any other edges if necessary by applying three rules

RULES FOR ORIENTING THE EDGES



NEXT

- Markov networks