ECE/CS/ISYE 8803

Probabilistic Graphical Models

Module 2 (Part A)
Bayesian Network Representation

Faramarz Fekri

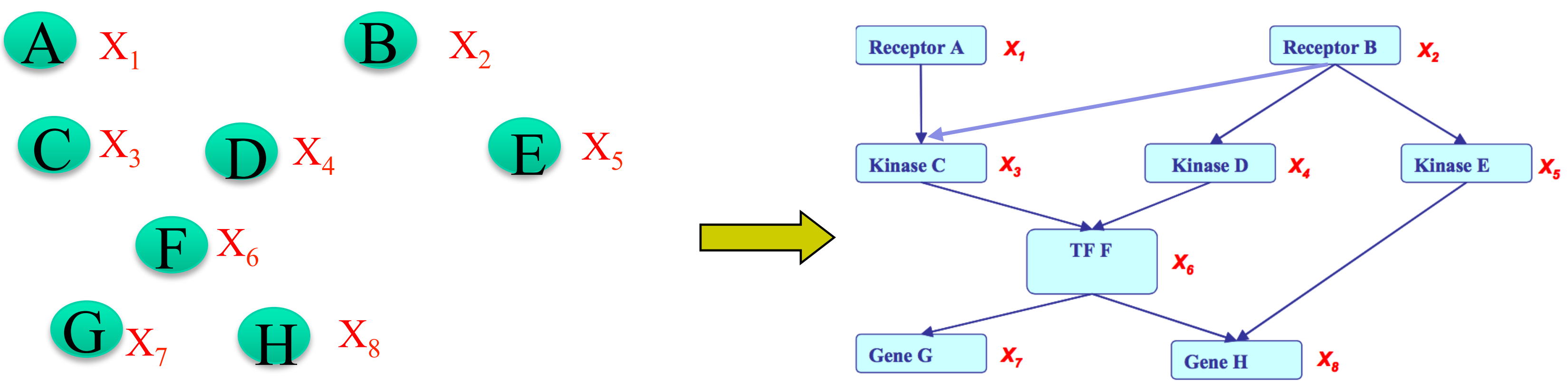
Center for Signal and Information Processing

Overview

- Bayesian Networks (BNs)
- Various forms of reasoning in BNs
- I-Map
- Representation Theorem
- Independencies Encoded by BN?

What is a Graphical Model?

- The informal blurb:
  - It is a smart way to write/specify/compose/design exponentially-large probability distributions without paying an exponential cost, and at the same time endow the distributions with ***structured semantics***



$$P(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8)$$

$$P(X_{1:8}) = P(X_1)P(X_2)P(X_3 \mid X_1 X_2)P(X_4 \mid X_2)P(X_5 \mid X_2)$$
$$P(X_6 \mid X_3, X_4)P(X_7 \mid X_6)P(X_8 \mid X_5, X_6)$$

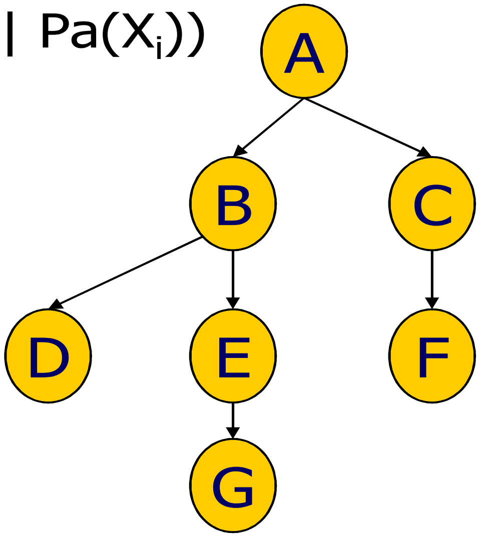
- A more formal description:
  - It refers to a family of distributions on a set of random variables that are compatible with all the probabilistic independence propositions encoded by a graph that connects these variables

-

Bayesian Networks (BNs)

Bayesian Networks (BNs)

- Directed acyclic graph G
  - Nodes $X_1,\ldots,X_n$ represent random variables
- G encodes local Markov assumptions
  - X_i is independent of its non-descendants given its parents
  - Formally: $(X_i \perp \mathrm{NonDesc}(X_i) \mid \mathrm{Pa}(X_i))$



Example:

$E \perp \{A,C,D,F\} \mid B$

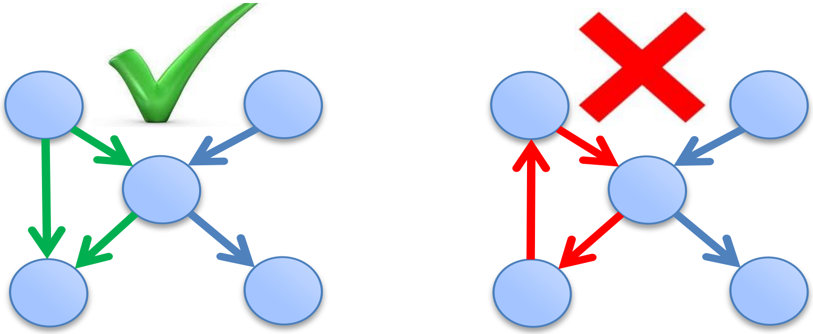
We will show that:

- Joint distribution:

$$P(X_1,\ldots,X_n) = \prod_{i=1}^{n} P\left(X_i \mid \mathbf{Pa}_{X_i}\right)$$

Directed Acyclic Graph

- Directed acyclic graph (DAG)
  - Loops are ok
  - But no directed cycle



Bayesian Network Definition

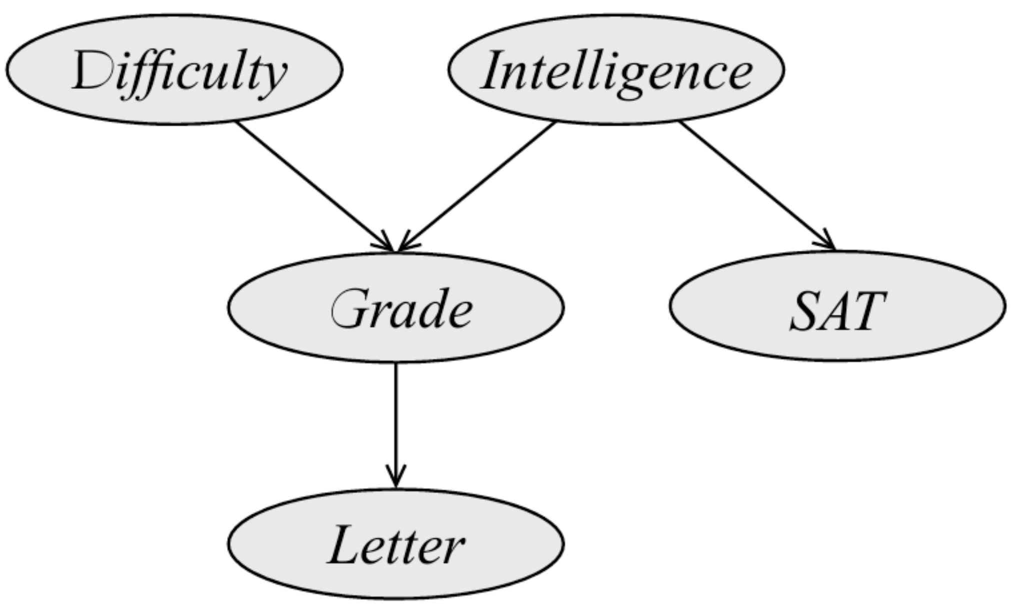
- A Bayesian network is a pair (G,P)
  - P factorizes over G
  - P is specified as set of CPDs associated with G's nodes

- Parameters
  - Joint distribution: 2^n
  - Bayesian network (bounded in-degree k): $n2^k$

Example: Quality of Rec. Letter (I)

- Course difficulty (**D**), Val(D) = {easy, hard}
- Intelligence (**I**) , Val (I) = {high, low}
- Grade (**G**), Val (G) = {A, B, C}
- Quality of the rec. letter (**L**), Val(L) = {strong, weak}
- SAT (**S**), Val (S) = {high, low}

Graph $G_{student}$

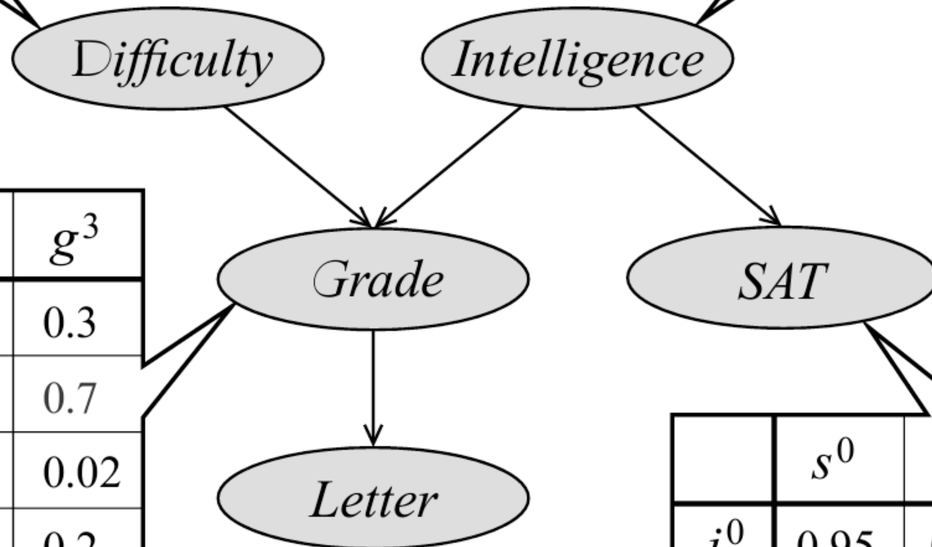


Local independencies $I_L(G_{student})$

Quality of Recommendation Letter (II)

d^0	d^1
0.6	0.4

i^0	i^1
0.7	0.3



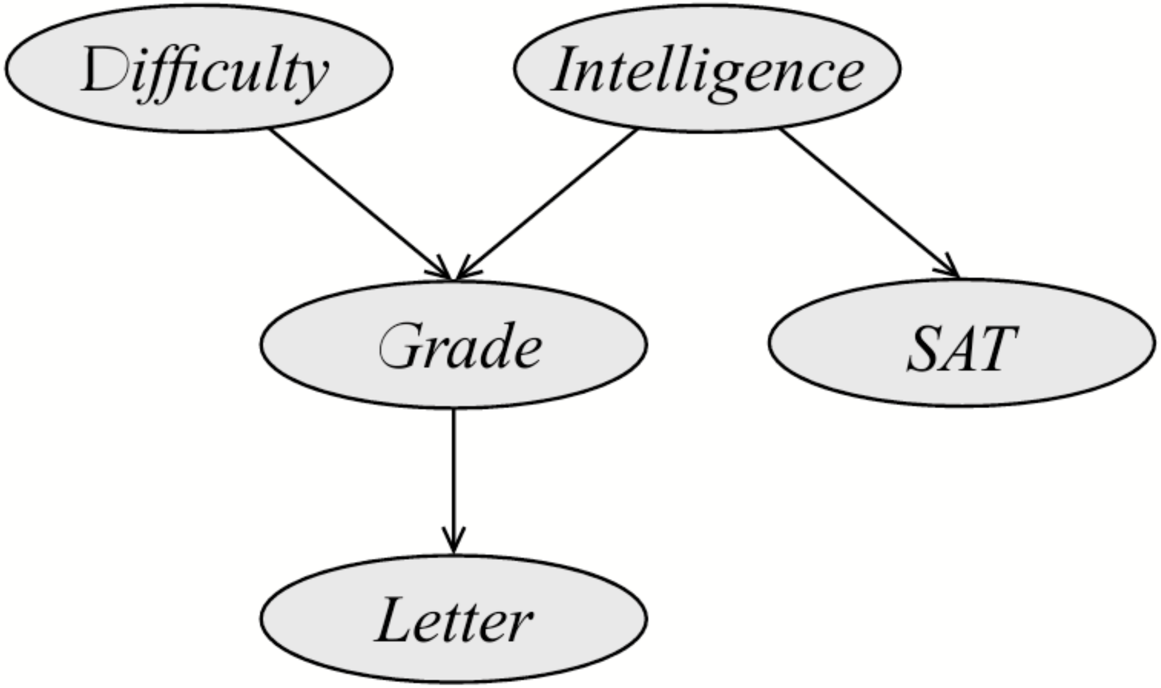
	g^1	g^2	g^3
i^0, d^0	0.3	0.4	0.3
i^0, d^1	0.05	0.25	0.7
i^1, d^0	0.9	0.08	0.02
i^1, d^1	0.5	0.3	0.2

	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^3	0.99	0.01

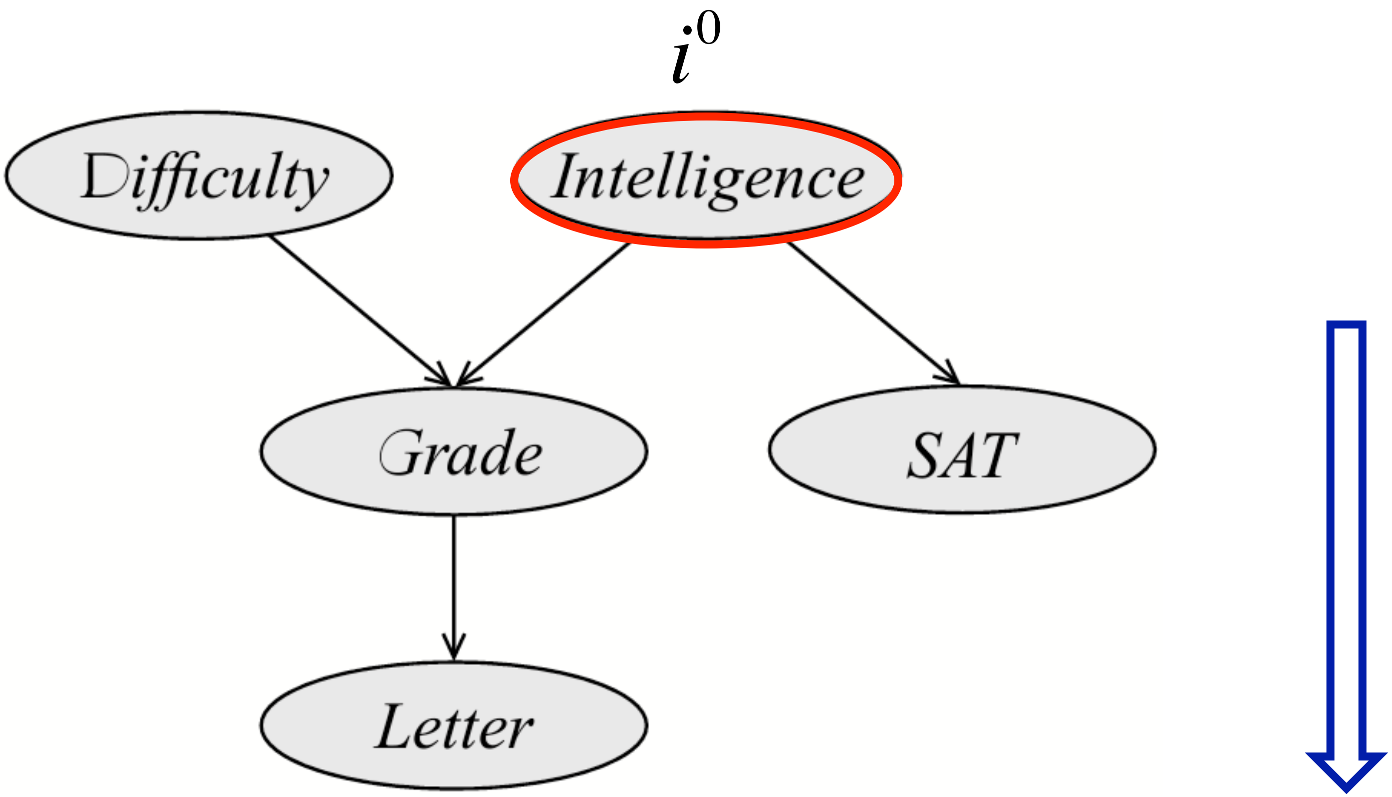
Table entries are CPD

Causal Reasoning (I)



$$P(l^1) \approx 0.5$$

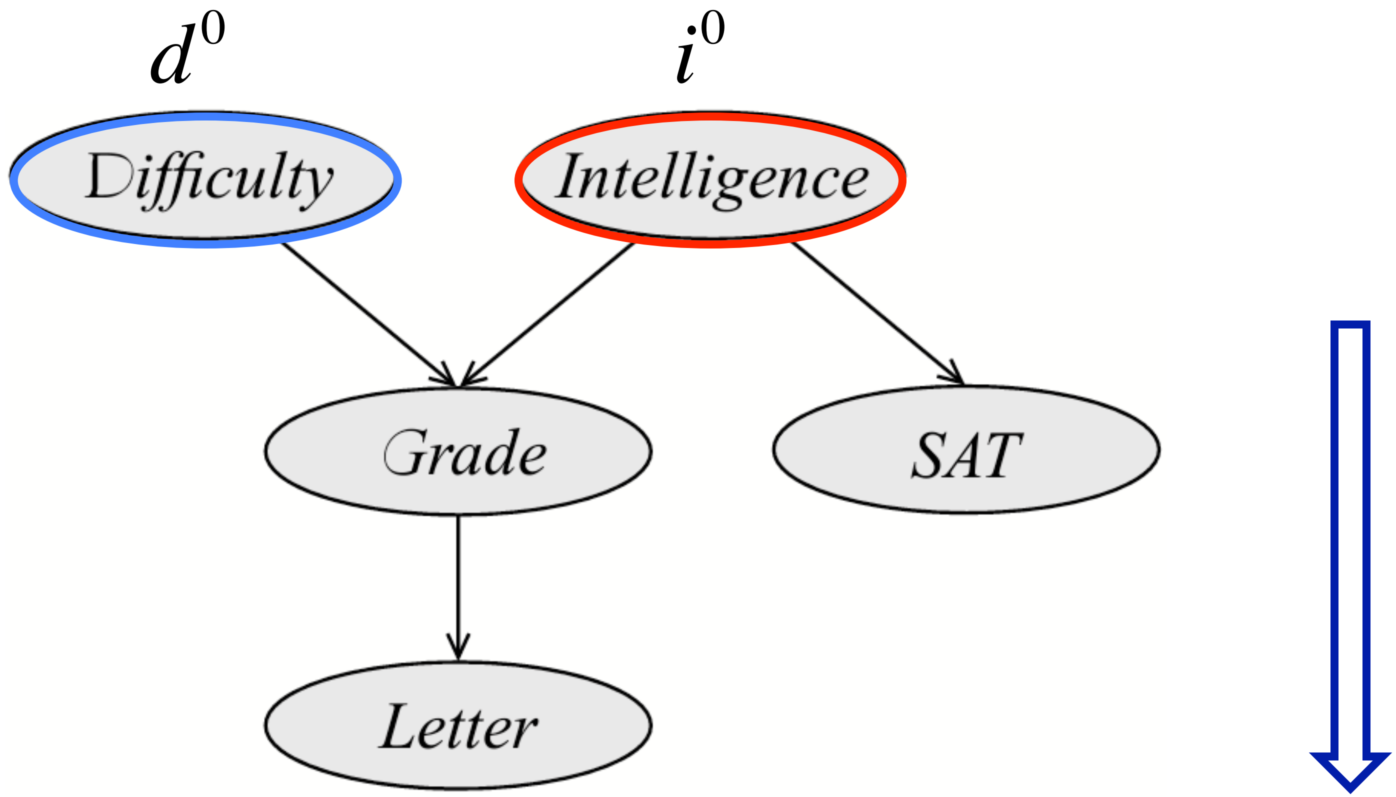
Causal Reasoning (II)



$$P(l^1) \approx 0.5$$

$$P(l^1 \mid i^0) \approx 0.39$$

Causal Reasoning (III)



$$P(l^1) \approx 0.5$$

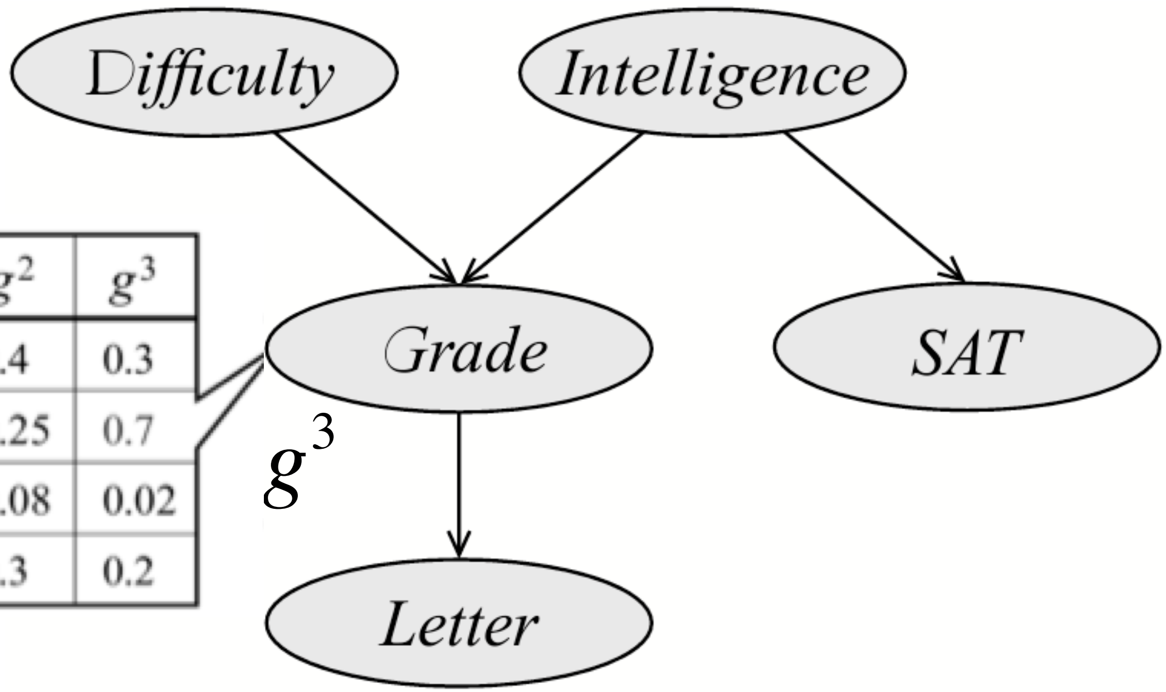
$$P(l^1 \mid i^0) \approx 0.39$$

$$P(l^1 \mid i^0, d^0) \approx 0.51$$

Evidential Reasoning (I)

$P(d^1) = 0.4 \qquad P(i^1) = 0.3$

	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

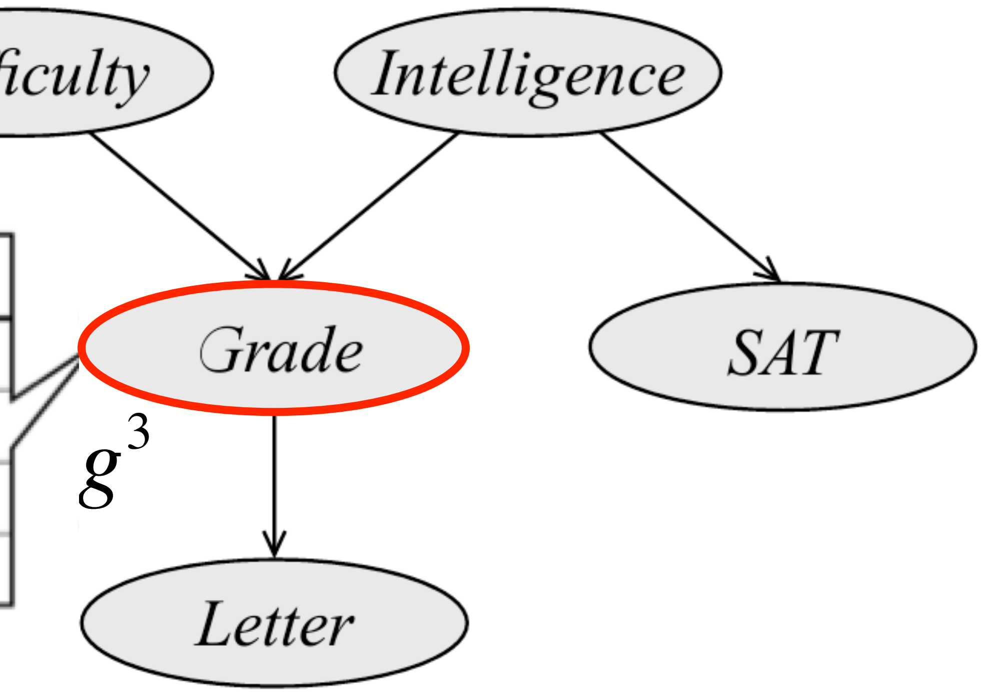


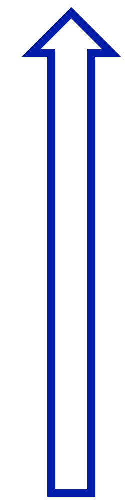
Evidential Reasoning (II)

$$P(d^1) = 0.4 \qquad P(i^1) = 0.3$$

$$P(d^1 \mid g^3) \approx 0.63 \qquad P(i^1 \mid g^3) \approx 0.08$$

	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



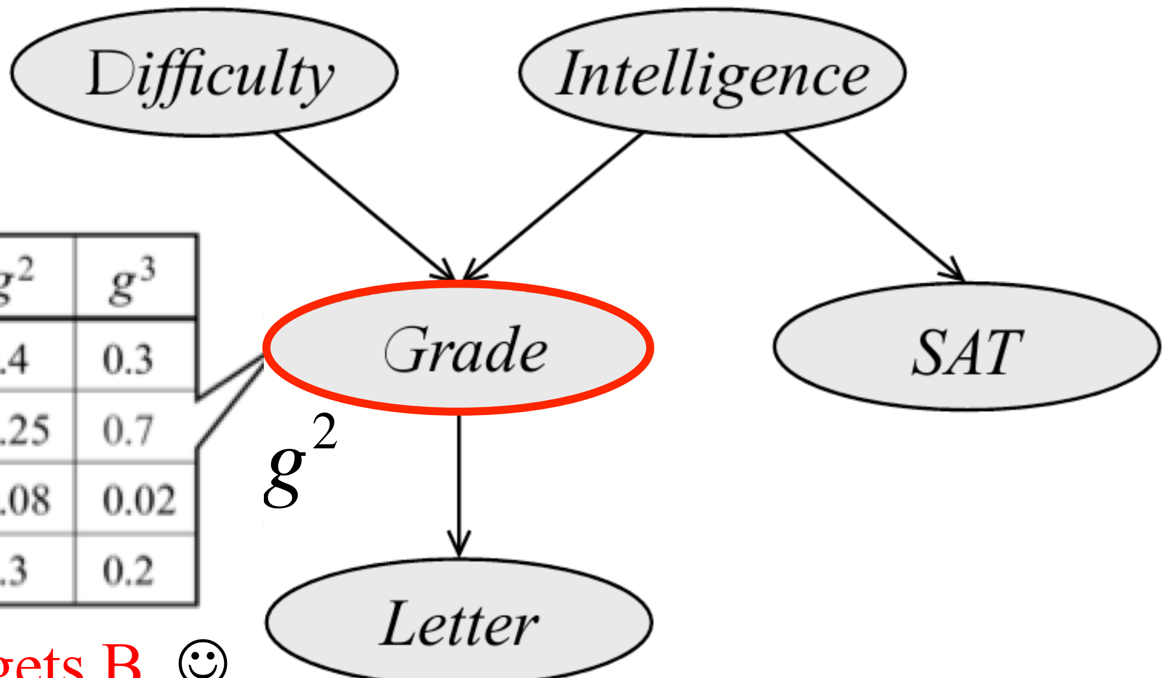


A student gets C ☹

Intercausal Reasoning (I)

$$P(i^1) = 0.3$$

$$P(i^1 \mid g^2) \approx 0.175$$



	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

A student gets B ☺

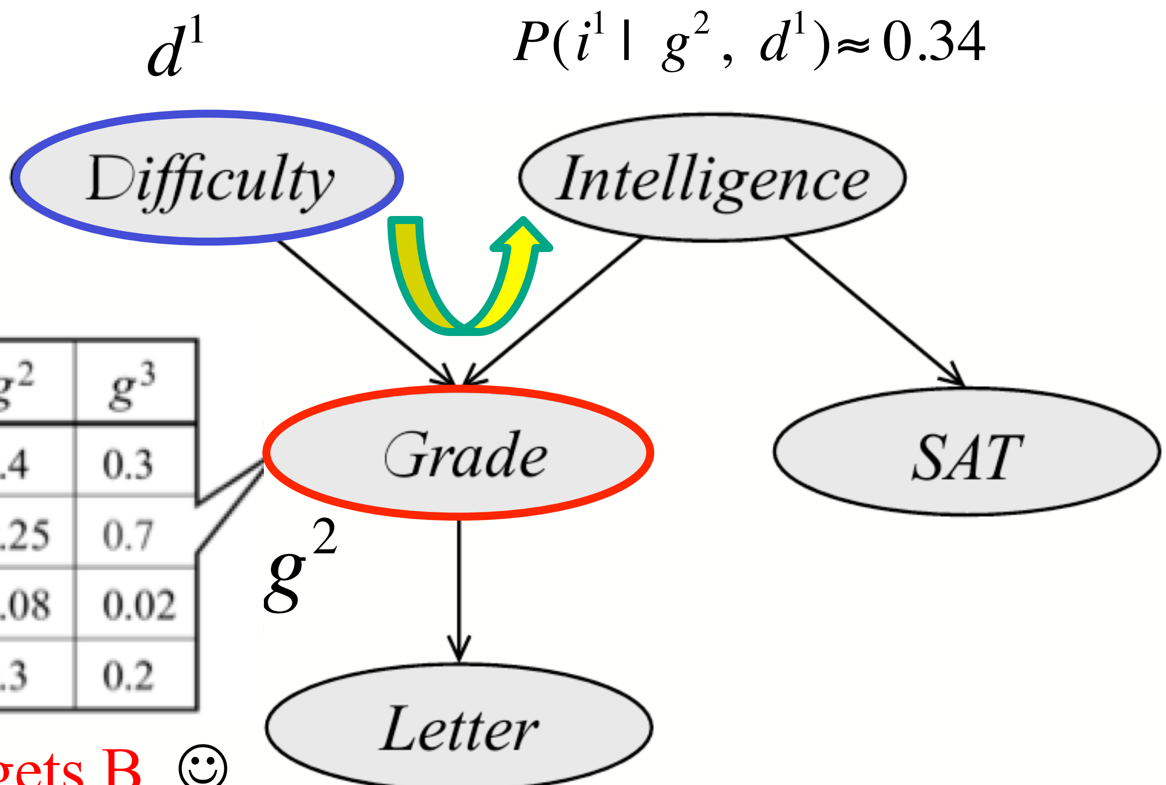
Intercausal Reasoning (II)

We were told that the Class was difficult, D= d^1

$$P(i^1) = 0.3$$

$$P(i^1 \mid g^2) \approx 0.175$$

$$P(i^1 \mid g^2, d^1) \approx 0.34$$



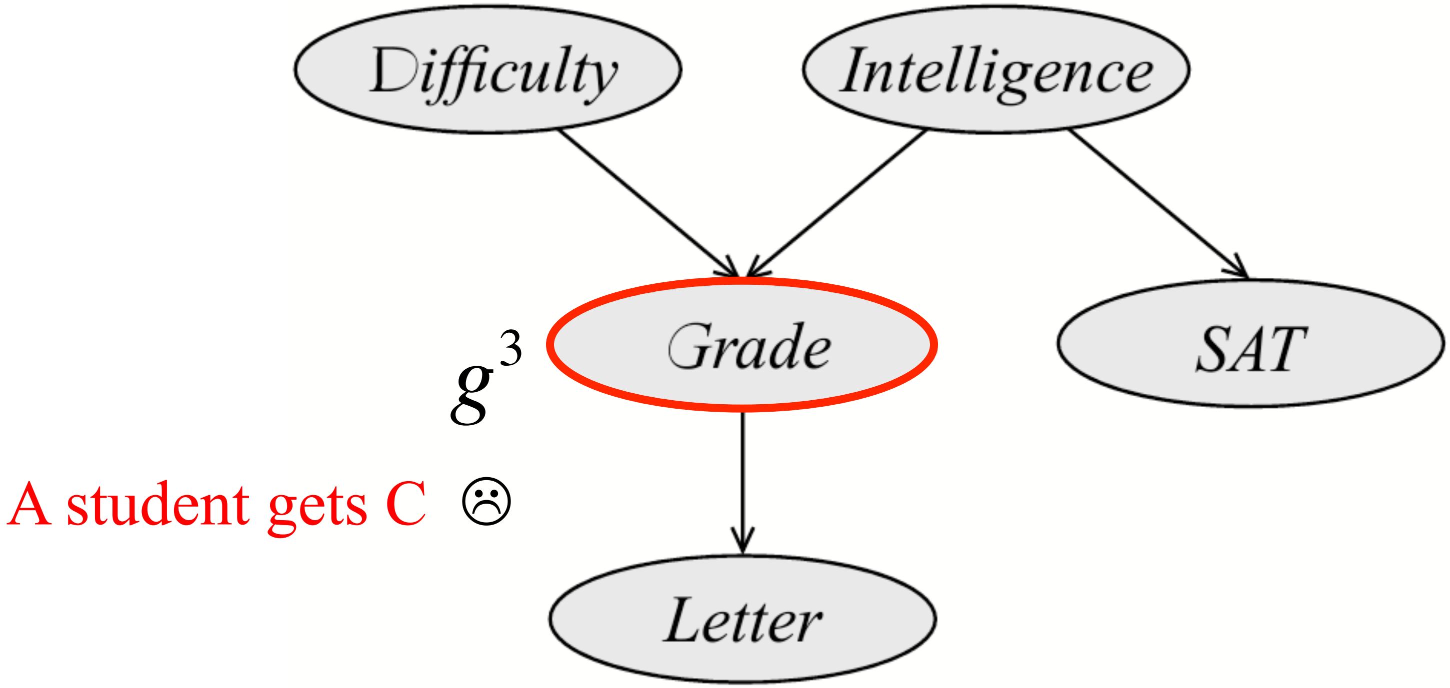
	g^1	g^2	g^3
i^0,d^0	0.3	0.4	0.3
i^0,d^1	0.05	0.25	0.7
i^1,d^0	0.9	0.08	0.02
i^1,d^1	0.5	0.3	0.2

A student gets B ☺

Multiple Evidence (I)

$$P(d^1) = 0.4 \qquad P(i^1) = 0.3$$

$$P(d^1 \mid g^3) \approx 0.63 \qquad P(i^1 \mid g^3) \approx 0.08$$

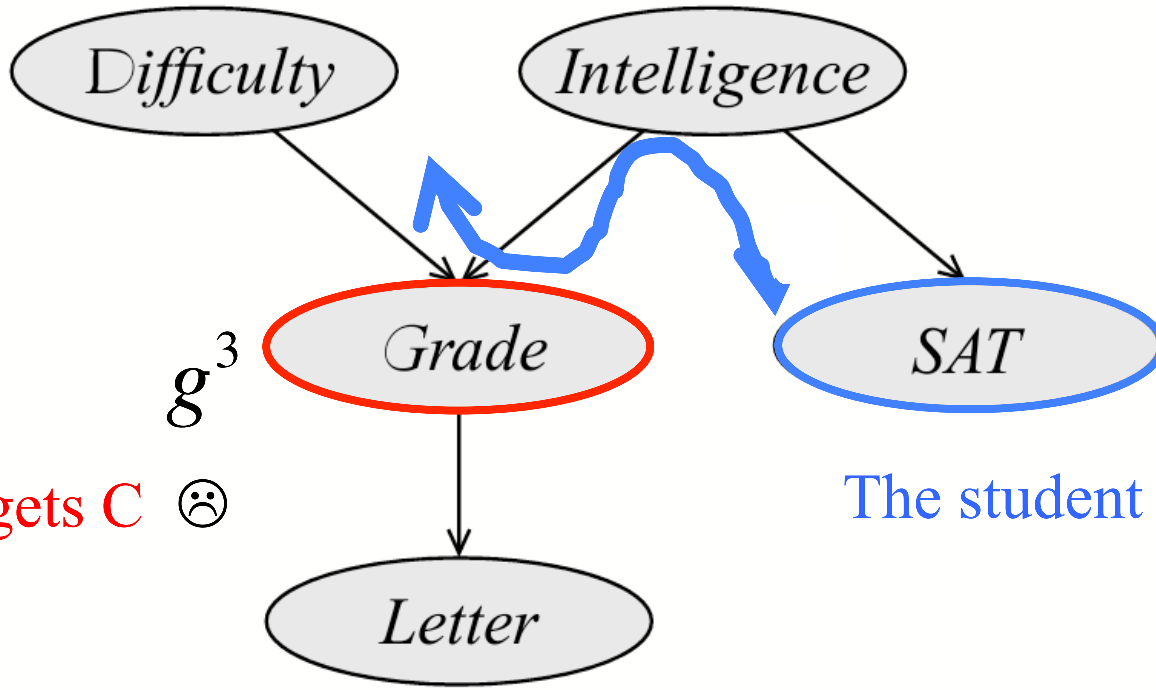


Multiple Evidence (II)

$$P(d^1) = 0.4 \qquad P(i^1) = 0.3$$

$$P(d^1 \mid g^3) \approx 0.63 \qquad P(i^1 \mid g^3) \approx 0.08$$

$$P(d^1 \mid g^3, s^1) \approx 0.76 \qquad P(i^1 \mid g^3, s^1) \approx 0.58$$



g^3

A student gets C ☹

The student gets s^1 ☺

Representation properties of BN

- What distributions can be represented by a Bayesian Networks?
- What Bayesian Networks can represent a distribution?
- What are the independence assumptions encoded by a BN?
  - In addition to the local Markov assumption

Independency Mapping (I-map)

Graph G encodes local independence assumptions $I_l(G)$

Example:

$\underline{I_L(G_{student})}$

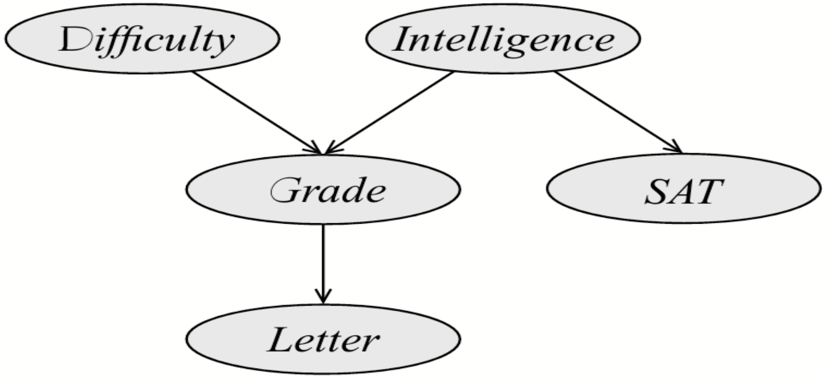
D $\perp$ I

D $\perp$ S

G $\perp$ S | D, I

L $\perp$ I,D,S | G

S $\perp$ D,G,L | I



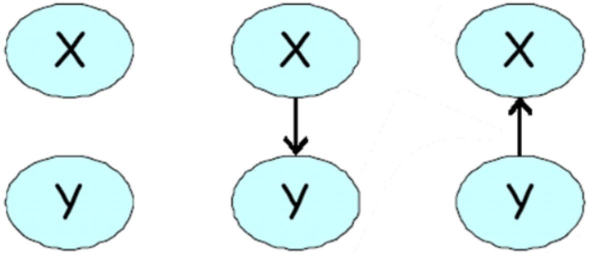
- Let P be a distribution over **X**
- Let I(P) be the independencies (**X** $\perp$ **Y** | **Z**) in P
- A Bayesian network structure is an I-map (independency mapping) of P if $I_l(G) \subseteq I(P)$

I-Map

- For G to be an I-map of P, it is necessary that G does not mislead us regarding independencies in P:

- any independence that G asserts must also hold in P. Conversely, P may have additional independencies that are not reflected in G

Example:



P_1

X	Y	$P(X,Y)$
x^0	y^0	0.08
x^0	y^1	0.32
x^1	y^0	0.12
x^1	y^1	0.48

X and Y independent

P_2

X	Y	$P(X,Y)$
x^0	y^0	0.4
x^0	y^1	0.3
x^1	y^0	0.2
x^1	y^1	0.1

X and Y aren't independent

Representation Theorem (I)

- If G is an I-Map of P, then P factorizes over G.

$$P(X_1, \dots, X_n) = \prod_i P(X_i \mid Pa_{X_i})$$

Sketch of Proof:

- wlog. (without loss of generality)

  $X_1, \dots, X_n$ is an ordering consistent with G (topological ordering)
- By chain rule: $P(X_1,\dots,X_n) = \prod_{i=1}^{n} P(X_i \mid X_1,\dots,X_{i-1})$
- From assumption: $Pa(X_i) \subseteq \{X_1, \dots, X_{i-1}\}$

  $\{X_1, \dots, X_{i-1}\} - Pa(X_i) \subseteq NonDesc(X_i)$
- Since G is an I-Map → $(X_i; NonDesc(X_i) \mid Pa(X_i)) \in I(P)$

Then we have:

$$P(X_i \mid X_1,\dots,X_{i-1}) = P(X_i \mid Pa(X_i))$$

Topological Ordering of a Directed Graph

- Topological ordering of $X_1, X_2, \dots X_n$:
  - Number variables such that
  - Parent has lower number than child
  - ie., $X_i \rightarrow X_j \Rightarrow i < j$
  - Variable has lower number than all its descendants

- A directed acyclic graph has always many topological orderings
  - Several algorithms to find topological ordering
    - , e.g., a depth first search alg.

Representation Theorem (II)

$$P(X_1, \dots, X_n) = \prod_i P(X_i \mid Pa_{X_i}) \Rightarrow \begin{cases} I_{l(G)} \subseteq I(P) \\ \text{G is an I-Map of P} \end{cases}$$

Hint of Proof:

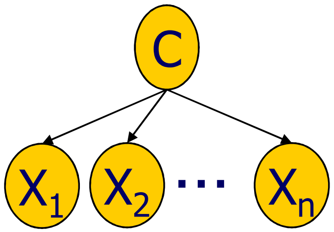
- Need to show $(X_i \perp NonDesc(X_i) \mid Pa(X_i)) \in I(P)$ or that $P(X_i \mid NonDesc(X_i)) = P(X_i \mid Pa(X_i))$

Proof of Theorem for Naïve Bayes

$$P(X_1, \dots, X_n) = \prod_i P(X_i \mid Pa_{X_i}) \overset{?}{\Rightarrow} I_{l(G)} \subseteq I(P)$$

Naïve Bayes:

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



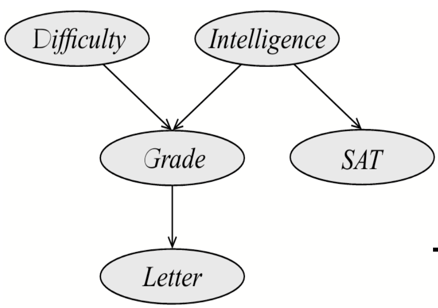
BN

Need to show that given C all X_i are independent of each other.

- Eg. $n = 4$, $P(X1, X\_2|C) = \frac{P(X_1,X_2,C)}{P(C)} = \frac{\sum_{x_3,x_4} P(X_1,X_2,X_3,X_4,C)}{P(C)}$
- $= \frac{1}{\cancel{P(C)}} \sum_{x_3,x_4} P(X_1|C)P(X_2|C)P(X_3|C)\,P(X_4|C)\cancel{P(C)}$
- $= P(X_1|C)P(X_2|C) \sum_{x_3} P(X_3|C) \sum_{x_4} P(X_4|C)$
- $= P(X_1|C)P(X_2|C) \cdot 1 \cdot 1$

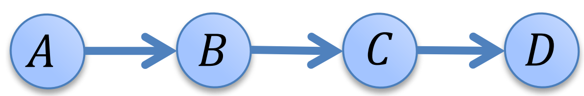
What are the Independencies Encoded by BN?

- We said: All you need is the local Markov assumption
  - $(X_i \perp \text{NonDescendants}_{X_i} \mid \mathbf{Pa}_{X_i})$
  - But many other can be derived using the algebra of conditional independencies!!!
    - Examples:



$(I \perp D)$

$\neg (I \perp D) \mid G$



Local Markov

$A \perp C \mid B$

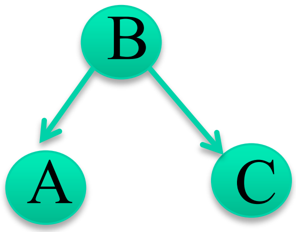
$D \perp AB \mid C$

Derived independence

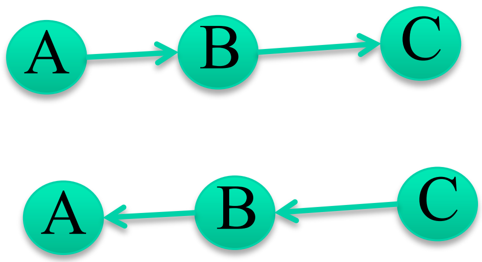
$A \perp D \mid B$

Independencies Encoded by BN

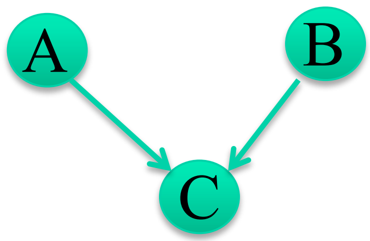
Common Cause



Indirect Causal Effect

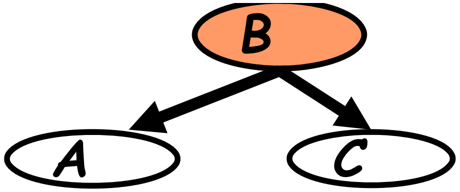


Common Effect



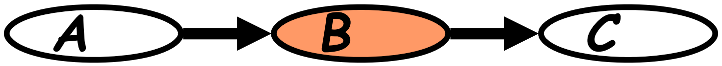
- Common parent
  - Fixing B decouples A and C

    "given the level of gene B, the levels of A and C are independent"



- Cascade
  - Knowing B decouples A and C

    "given the level of gene B, the level gene A provides no extra prediction value for the level of gene C"



- V-structure
  - Knowing C couples A and B

    because A can "explain away" B w.r.t. C

    "If A correlates to C, then chance for B to also correlate to B will decrease"