Universität Stuttgart
IPVS – Institute for Parallel and Distributed Systems

Analytic Computing

Advanced Topics in Machine Learning
3 Representation: Bayesian Networks
Part 2

Prof. Dr. Steffen Staab
Dr. Rafika Boutalbi
Zihao Wang
https://www.ipvs.uni-stuttgart.de/departments/ac/

Learning Objectives

Given: Formal Definition of Bayesian Network (last week)

- Abstract syntax and semantics

- How does information flow in a Bayesian Network
  - Put this formally
    - d-separation
    - I-Map
    - Equivalence of factorization using chain rule and representation of conditional independencies through the Bayesian Network Graph
- How to efficiently write down large networks
  - Template notation

Meaning of boxes

explains the slide content

important take away

side note: nice to know

Disclaimer

Figures and examples are taken from the book by Koller & Friedman

1 Intermezzo about random variables and factors

Remember: Random Variable

- Discrete Random Variable
  - $X = \left(D_X, P_X\right)$
  - D_X is the domain;
    a (possible infinite set)
  - $P_X : D_X \to \left[0,1\right]$, such that
    $$\sum_{x \in D_X} P(X = x) = 1$$
- There are all kind of shorthand/alternative notations that mean the same, e.g.:
  $$\sum_{x \in X} P(x) = 1$$
- Though in a Java program this might be considered a type mismatch, when reading books and research papers you must juggle the different kind of notations

Factors

A *factor* is a function:

$\phi: \mathrm{D}_\phi \rightarrow \mathbb{R}_0^+$

- D_ϕ is a set
- D_ϕ is the domain of ϕ
- D_ϕ is called *scope* of factor ϕ

- Examples:
  1. $P(I, D, G)$, with scope $\mathrm{val}(I) \times \mathrm{val}(D) \times \mathrm{val}(G) = \mathrm{val}(I \times \mathrm{D}$
  2. $P(I, D, g^1)$, with scope $\mathrm{val}(I) \times \mathrm{val}(D)$, because g^1 is a constant
  3. $P(G \mid I, D)$, with scope $\mathrm{val}(I \times \mathrm{D} \times \mathrm{G})$
  4. $\phi: \{A, B, C\} \rightarrow \mathbb{R}_0^+$, with

You may think of a ***factor*** as a random variable whose probability distribution is not normalized – and therefore cannot be called a random variable!

=

Factors will come handy all over the place

Operations on Factors: 1. Reduction

Reduction corresponds to Conditioning on a Value

Example

a^1	b^1	c^1	0.25
a^1	b^1	c^2	0.35
a^1	b^2	c^1	0.08
a^1	b^2	c^2	0.16
a^2	b^1	c^1	0.05
a^2	b^1	c^2	0.07
a^2	b^2	c^1	0
a^2	b^2	c^2	0
a^3	b^1	c^1	0.15
a^3	b^1	c^2	0.21
a^3	b^2	c^1	0.09
a^3	b^2	c^2	0.18

$\sigma_{C=c^1}(\text{TABLE})$

a^1	b^1	c^1	0.25
a^1	b^2	c^1	0.08
a^2	b^1	c^1	0.05
a^2	b^2	c^1	0
a^3	b^1	c^1	0.15
a^3	b^2	c^1	0.09

Not a random variable,

not a probability distribution,

but a factor

Comparison with SQL:

Selection on $C = c^1$

Operations on Factors: 2. Marginalization

Example

```
SELECT A,C,sum(*)
FROM TABLE
GROUP BY A,C
```

a^1	b^1	c^1	0.25
a^1	b^1	c^2	0.35
a^1	b^2	c^1	0.08
a^1	b^2	c^2	0.16
a^2	b^1	c^1	0.05
a^2	b^1	c^2	0.07
a^2	b^2	c^1	0
a^2	b^2	c^2	0
a^3	b^1	c^1	0.15
a^3	b^1	c^2	0.21
a^3	b^2	c^1	0.09
a^3	b^2	c^2	0.18

a^1	c^1	0.33
a^1	c^2	0.51
a^2	c^1	0.05
a^2	c^2	0.07
a^3	c^1	0.24
a^3	c^2	0.39

Comparison with SQL:

- Group by values from $A \times C$
- Aggregate $\phi(\,.\,, B, .\,)$

Operations on Factors: 3. Factor product

Example

a^1	b^1	0.5
a^1	b^2	0.8
a^2	b^1	0.1
a^2	b^2	0
a^3	b^1	0.3
a^3	b^2	0.9

ϕ_1

b^1	c^1	0.5
b^1	c^2	0.7
b^2	c^1	0.1
b^2	c^2	0.2

ϕ_2

$\Rightarrow$

a^1	b^1	c^1	$0.5 \cdot 0.5 = 0.25$
a^1	b^1	c^2	$0.5 \cdot 0.7 = 0.35$
a^1	b^2	c^1	$0.8 \cdot 0.1 = 0.08$
a^1	b^2	c^2	$0.8 \cdot 0.2 = 0.16$
a^2	b^1	c^1	$0.1 \cdot 0.5 = 0.05$
a^2	b^1	c^2	$0.1 \cdot 0.7 = 0.07$
a^2	b^2	c^1	$0 \cdot 0.1 = 0$
a^2	b^2	c^2	$0 \cdot 0.2 = 0$
a^3	b^1	c^1	$0.3 \cdot 0.5 = 0.15$
a^3	b^1	c^2	$0.3 \cdot 0.7 = 0.21$
a^3	b^2	c^1	$0.9 \cdot 0.1 = 0.09$
a^3	b^2	c^2	$0.9 \cdot 0.2 = 0.18$

$$\underbrace{\qquad\qquad}_{\mathrm{dom}(\phi_1) \bowtie \mathrm{dom}(\phi_2)}$$

Question 1:

Random Variables vs Factors

2 Bayesian Networks: The story so far

What do we know about Bayesian Networks?

- Definition using graphs and conditional probability tables
  - Model (conditional) probabilistic independencies
    - which exactly?
  - Use probabilistic chain rule to factorize joint probability distribution
    - how exactly?
  - How are the two connected?

Joint Probability Distribution: Running Example

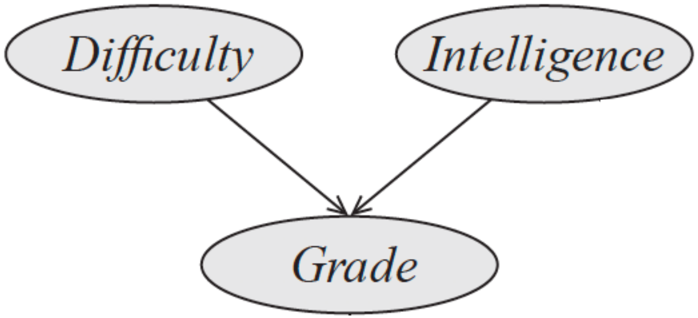
RANDOM VARIABLES: I, D, G

- Intelligence of student:
  $\text{val}(I) = \{i^0, i^1\}$
  {low, high}
- Difficulty of exam:
  $\text{val}(D) = \{d^0, d^1\}$
  {easy, hard}
- Grade achieved:
  $\text{val}(G) = \{g^1, g^2, g^3\}$
  {A,B,C}

I	D	G	Prob.
i^0	d^0	g^1	0.126
i^0	d^0	g^2	0.168
i^0	d^0	g^3	0.126
i^0	d^1	g^1	0.009
i^0	d^1	g^2	0.045
i^0	d^1	g^3	0.126
i^1	d^0	g^1	0.252
i^1	d^0	g^2	0.0224
i^1	d^0	g^3	0.0056
i^1	d^1	g^1	0.06
i^1	d^1	g^2	0.036
i^1	d^1	g^3	0.024

Joint Probability Distribution: Running Example

RANDOM VARIABLES: I, D, G

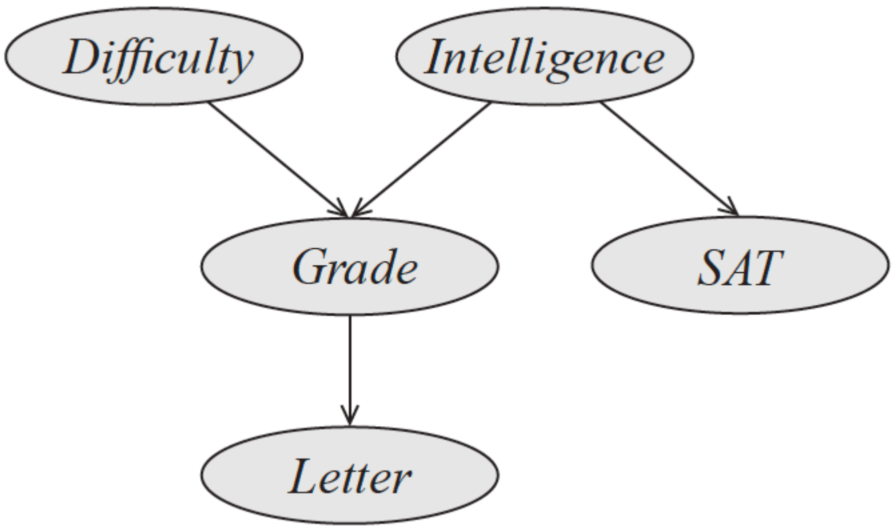


When/how do random variables influence each other?

Under which conditions do we know that the graph and the joint probability distribution match?

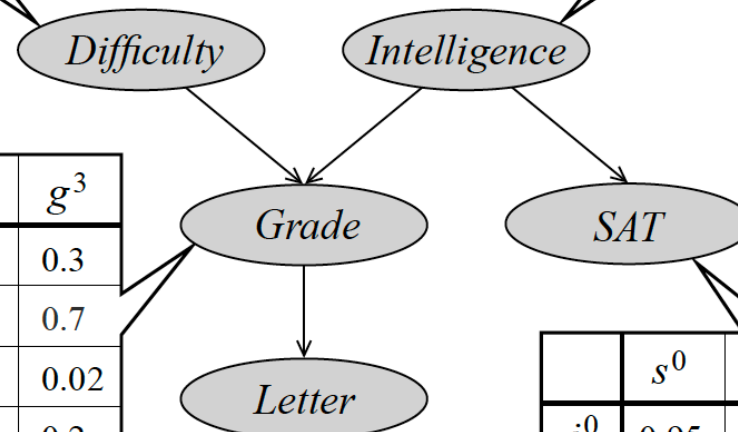
I	D	G	Prob.
i^0	d^0	g^1	0.126
i^0	d^0	g^2	0.168
i^0	d^0	g^3	0.126
i^0	d^1	g^1	0.009
i^0	d^1	g^2	0.045
i^0	d^1	g^3	0.126
i^1	d^0	g^1	0.252
i^1	d^0	g^2	0.0224
i^1	d^0	g^3	0.0056
i^1	d^1	g^1	0.06
i^1	d^1	g^2	0.036
i^1	d^1	g^3	0.024

Running Example Extended



- Intelligence: $\mathrm{val}(I) = \{i^0, i^1\}$
- Difficulty: $\mathrm{val}(D) = \{d^0, d^1\}$
- Grade:
  $\mathrm{val}(G) = \{g^1, g^2, g^3\}$
- Scholastic aptitude test (SAT):
  $\mathrm{val}(S) = \{s^0, s^1\}$
  {below average, above average}
- Recommendation Letter:
  $\mathrm{val}(L) = \{l^0, l^1\}$
  {weak, strong}

Bayesian Network Graph and Conditional Probability Tables: $\mathcal{B}^{student}$



d^0	d^1
0.6	0.4

i^0	i^1
0.7	0.3

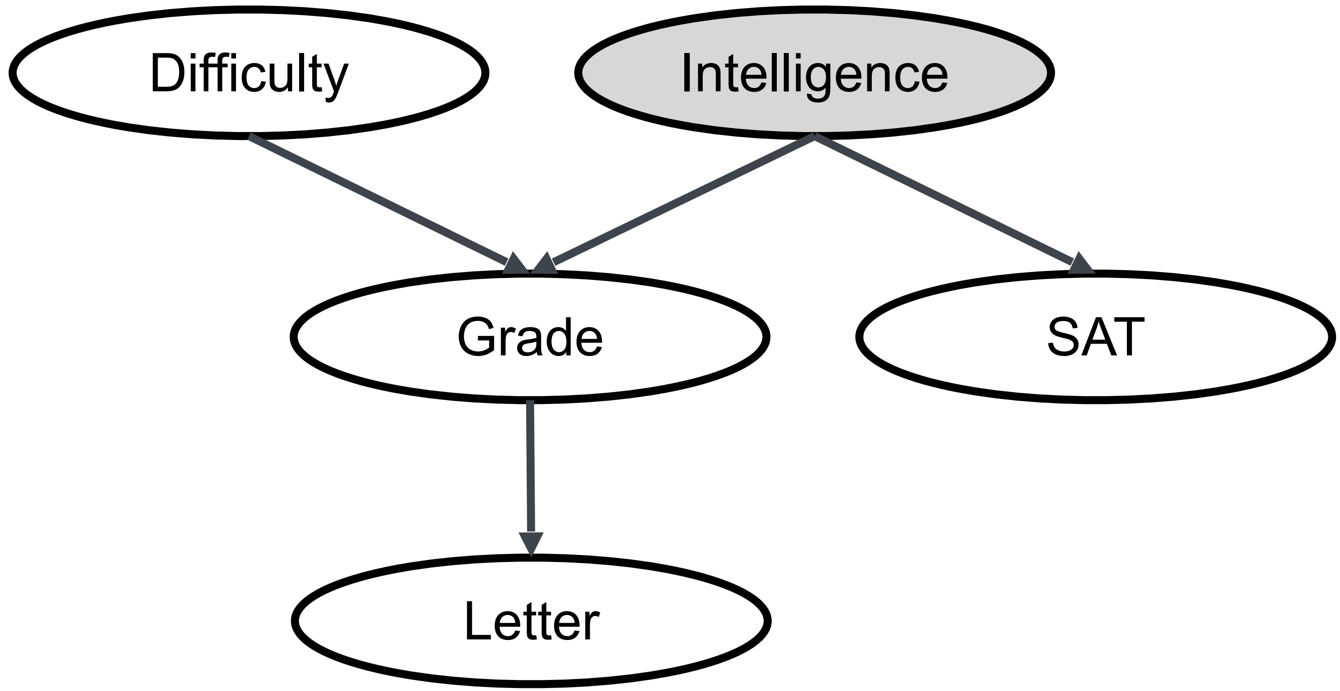
	g^1	g^2	g^3
i^0, d^0	0.3	0.4	0.3
i^0, d^1	0.05	0.25	0.7
i^1, d^0	0.9	0.08	0.02
i^1, d^1	0.5	0.3	0.2

	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

3 Reasoning Patterns

Causal Reasoning

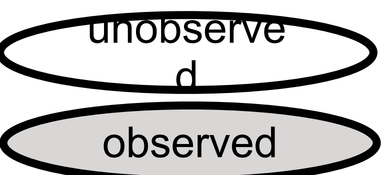


Without observation:

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

With observation:

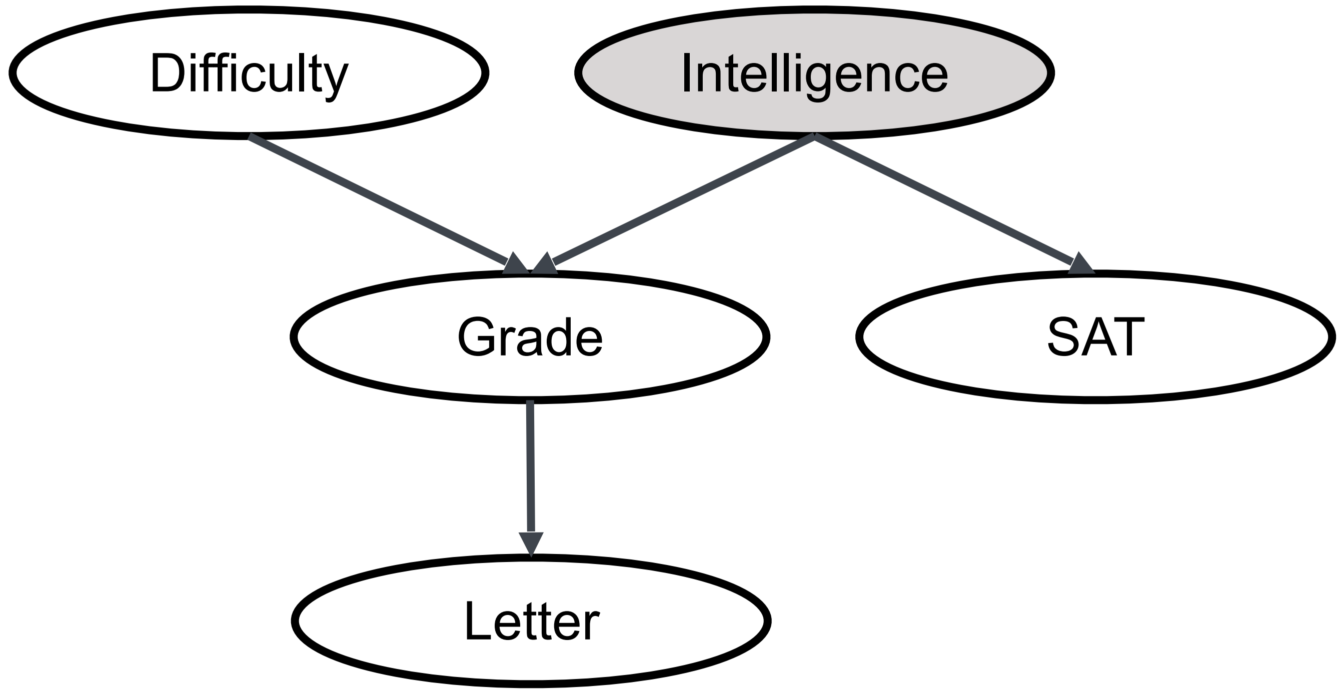
$$P(l^1 \mid i^0) \approx 0.39$$



Often, we model causality by descendant/ancestor relationship in Bayesian networks.

In general: not all descendant/ancestor relationships in Bayesian networks constitute causality!!!

Causal Reasoning

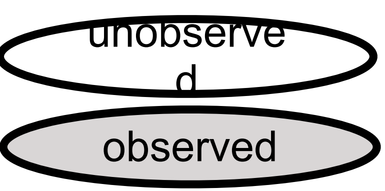


Without observation:

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

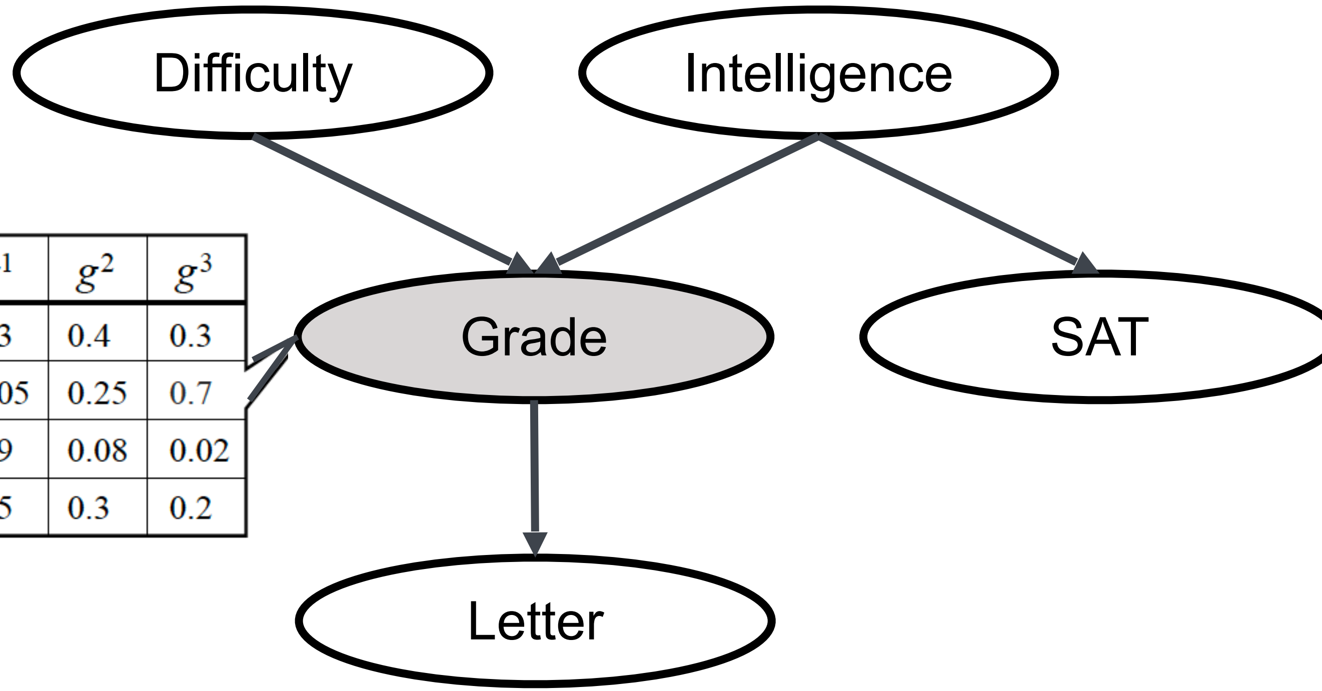
With observation:

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



If a distinction is made by shading nodes, then shaded nodes are commonly the observed ones.

Evidential Reasoning



	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

Without observation:

$$P\left(d^1\right) = 0.4$$

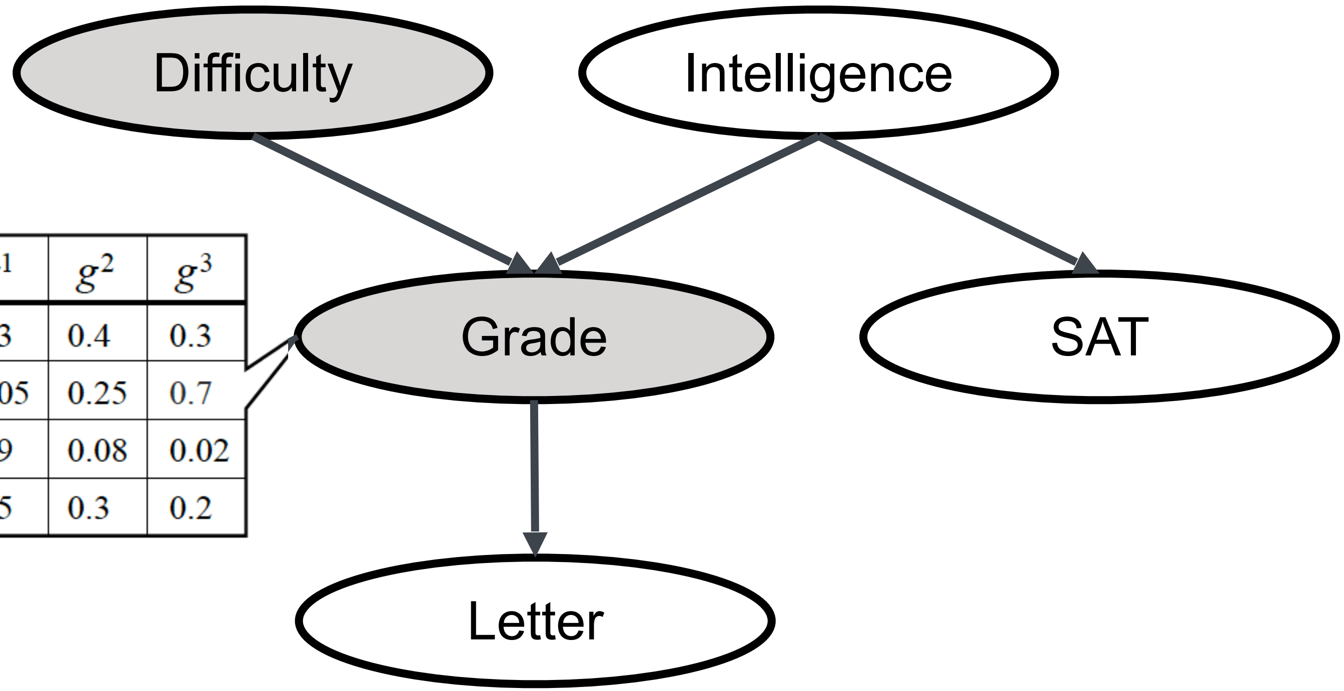
$$P\left(i^1\right) = 0.3$$

With observation: Student gets C

$$P\left(d^1 \middle| g^3\right) \approx 0.63$$

$$P(i^1 | g^3) \approx 0.08$$

Intercausal Reasoning



	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

Without observation:

$$P(d^1) = 0.4$$

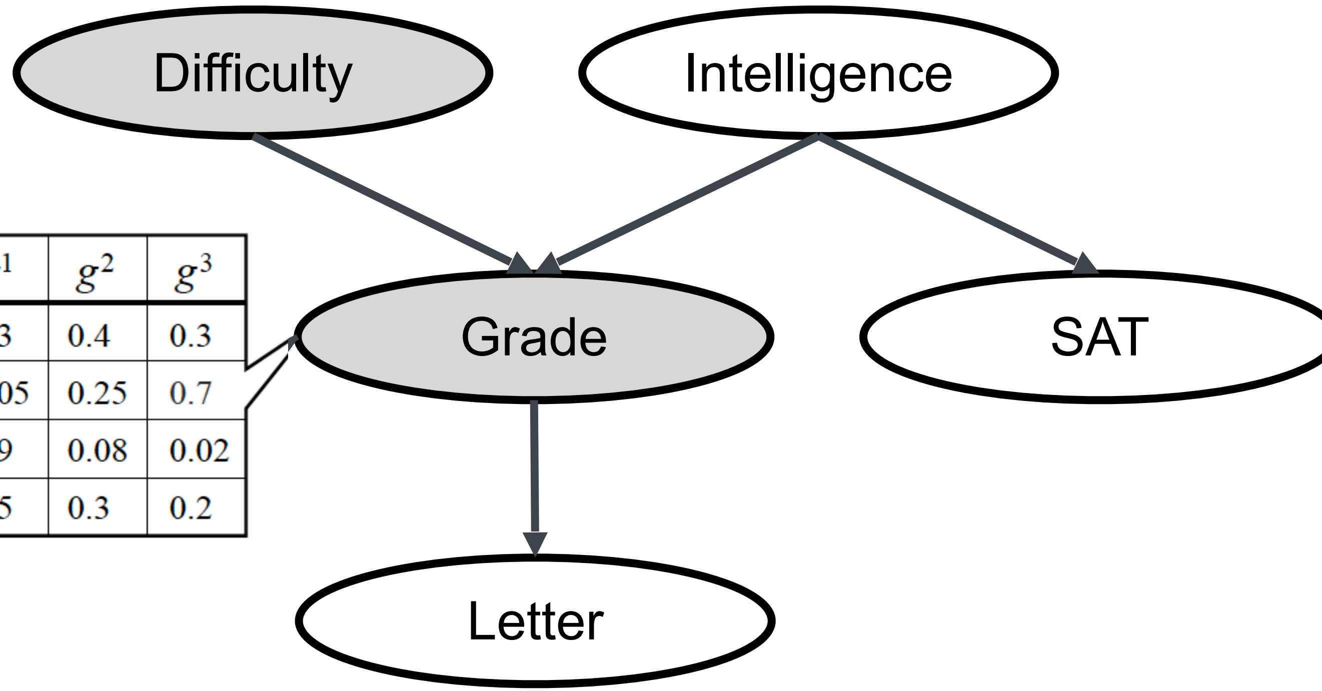
$$P(i^1) = 0.3$$

With observations: Student gets C & class is hard

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

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

Intercausal Reasoning II



	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

Without observation:

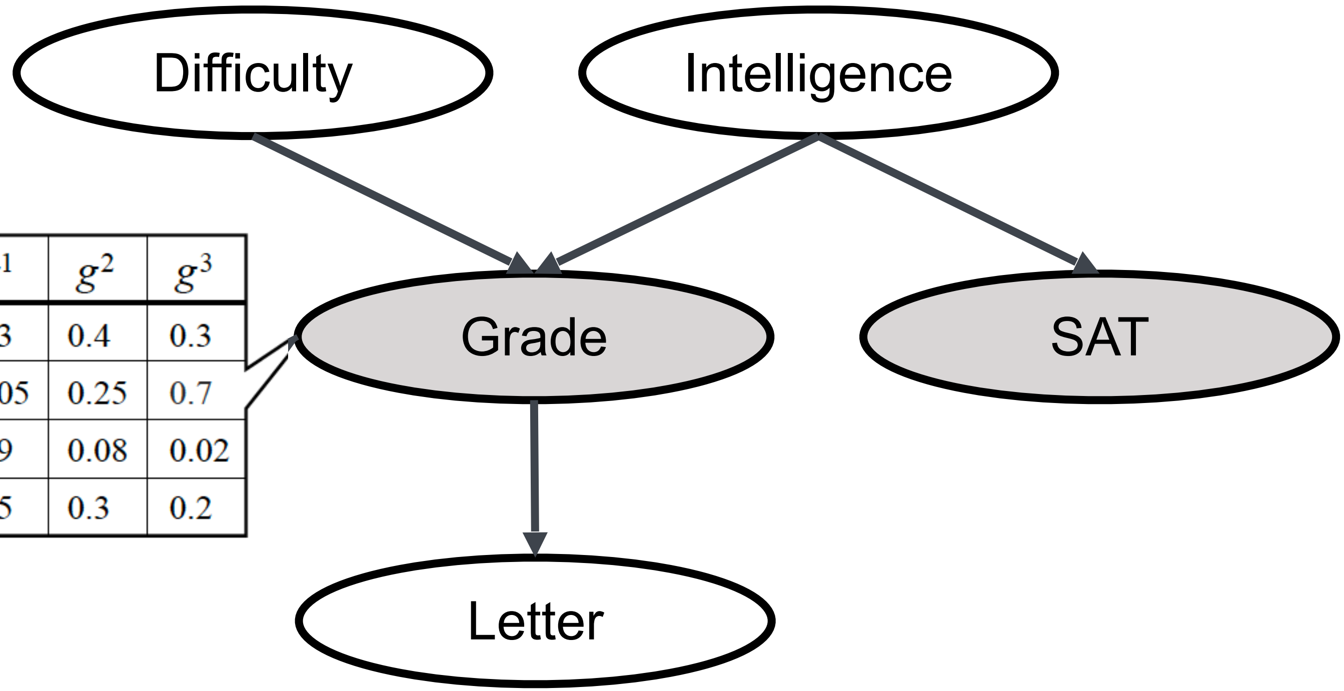
$$P(i^1) = 0.3$$

With observations: Student gets B & class is hard

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

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

Explaining Away: Student Aces the SAT



	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

Without observation:

$$P(d^1) = 0.4$$

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

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

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

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

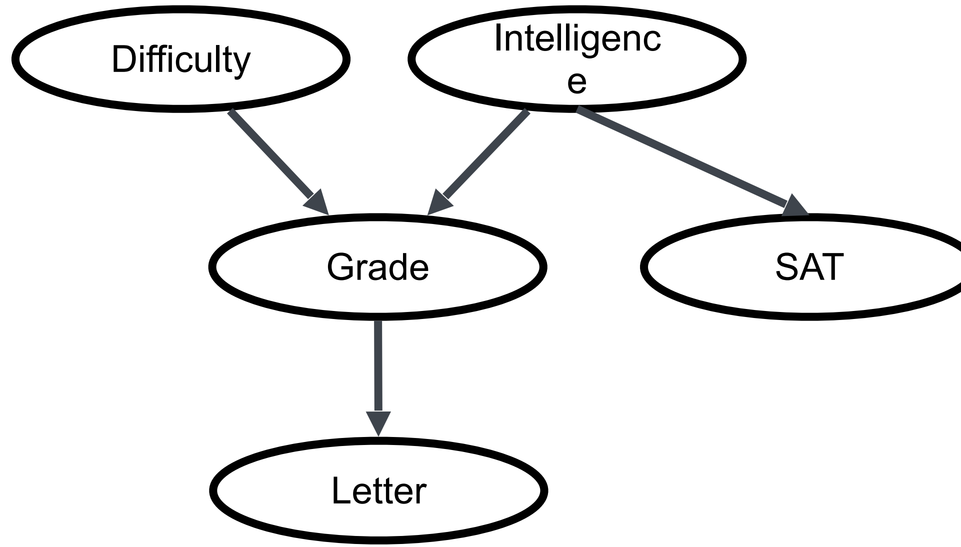
Question 3: Reasoning Patterns

4 Flow of Probabilistic Inference & D-Separation

When can X influence Y?

- $X \rightarrow Y$
- $X \leftarrow Y$
- $X \rightarrow W \rightarrow Y$
- $X \leftarrow W \leftarrow Y$
- $X \leftarrow W \rightarrow Y$
- $X \rightarrow W \leftarrow Y$

  v-structure

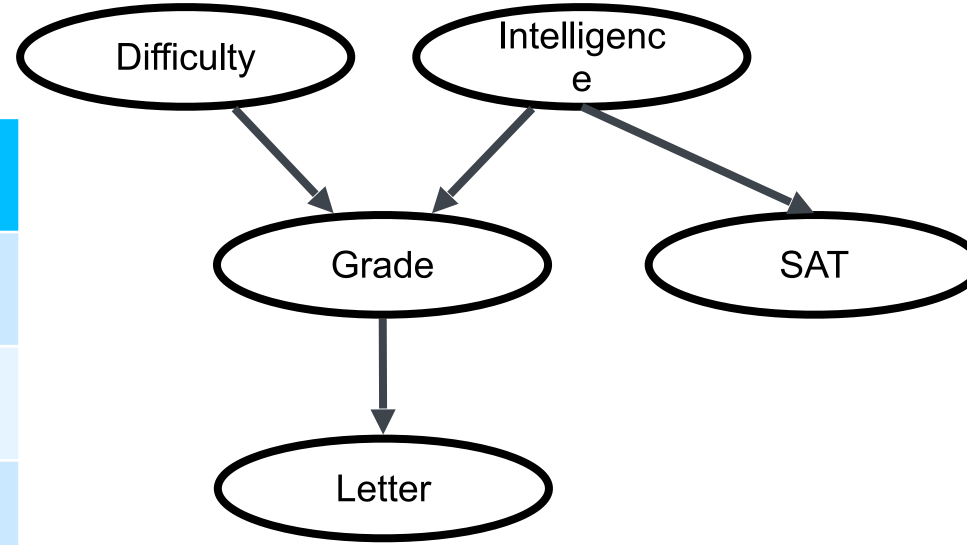


Active Trail

- A trail $X_1 - \ldots - X_k$ is active if
  - it has no v-structures $X_{i-1} \rightarrow X_i \leftarrow X_{i+1}$

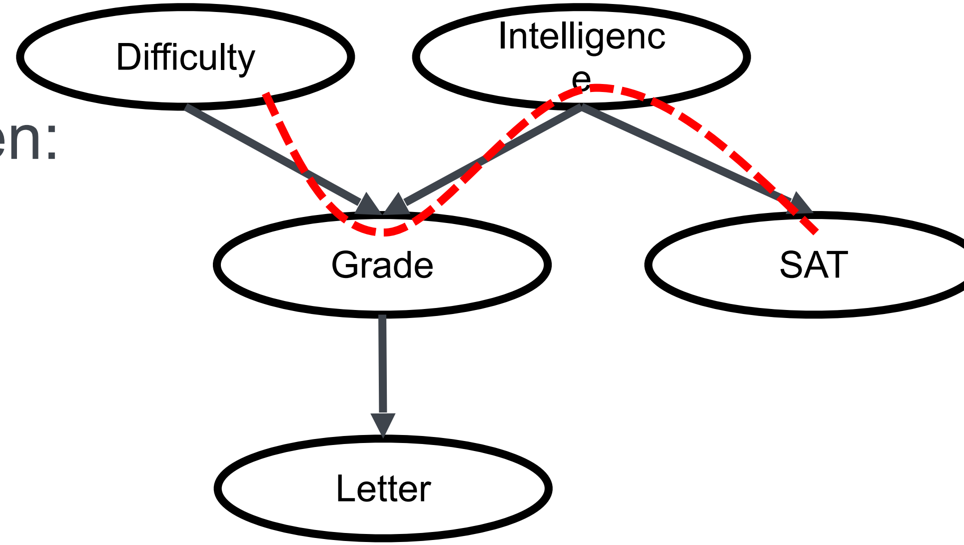
When can X influence Y given evidence about Z?

- $X \rightarrow Y$
- $X \leftarrow Y$
- $X \rightarrow W \rightarrow Y$
- $X \leftarrow W \leftarrow Y$
- $X \leftarrow W \rightarrow Y$
- $X \rightarrow W \leftarrow Y$

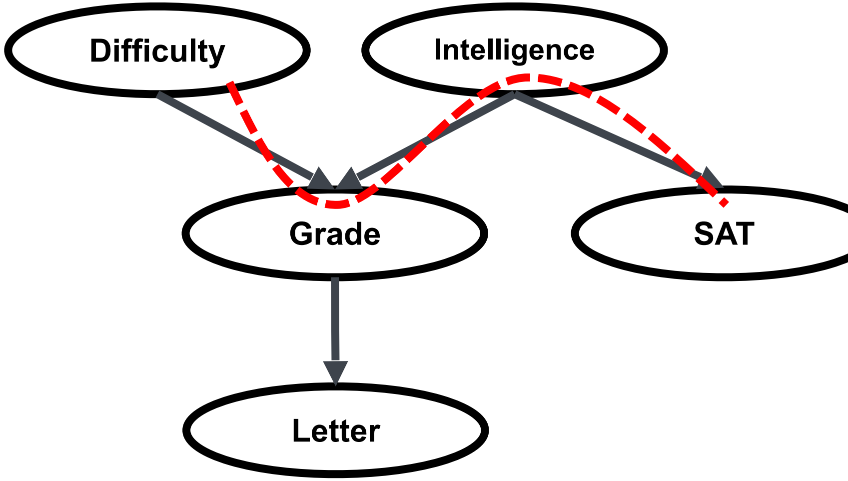


When can S influence D?

- $S - I - G - D$
  allows influence to flow when:
- I is not observed,
  G is observed

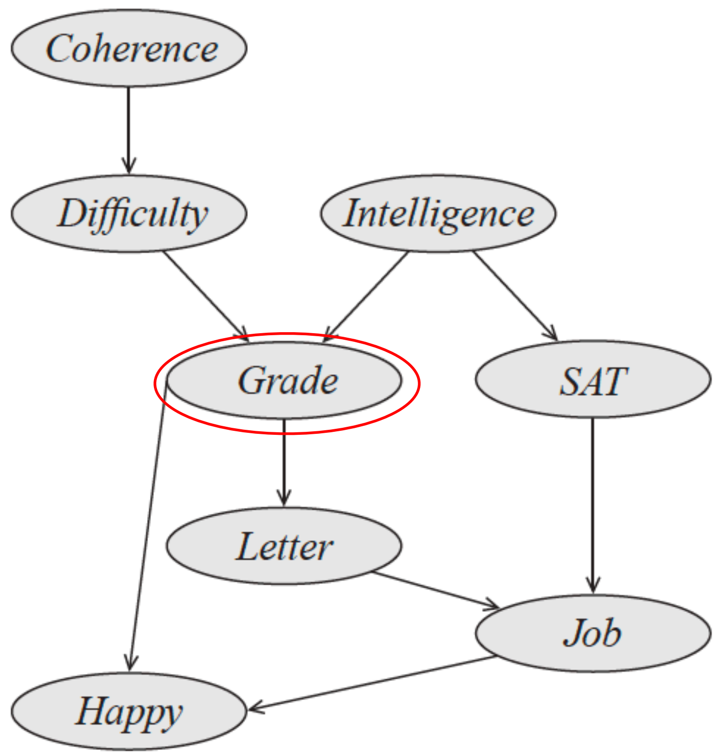


Active Trail and D-Separation



- Definition: A trail $X_1 - \ldots - X_k$ is active if
  - for any v-structure $X_{i-1} \rightarrow X_i \leftarrow X_{i+1}$,
    we have that X_i or one of its descendants is in Z
  - no other X_i is in Z
- Definition: X and Y are d-separated in G given Z if there is no active trail in G between X and Y given Z.
- Notation: $\mathrm{d\text{-}sep}_G(X, Y \mid Z)$

Any node is d-separated from its non-descendants given its parents



Question 4: D-Separation

5 Two Views of BN Graphs

Probability Distribution $P(\mathscr{X})$ Factorizes Over G

Definition („factorizes over")

- Let G be a graph over $\mathscr{X} = \{X_1, \ldots, X_n\}$.
- $P(\mathscr{X})$ factorizes over G if forall $x_i \in X_i,\ i = 1\ldots n$:

$$P(x_1, \ldots, x_n) = \prod_{i=1\ldots n} P(x_i \mid Parents_G(x_i))$$

Is $\mathscr{B}^{student}$ a Probability Distribution? (just considering our specific example)

Is $$\sum_{d \in D, i \in I, g \in G, s \in S, l \in L} P(d, i, g, s, l) = 1 \ ?$$

We write: $$\sum_{D,I,G,S,L} P(D, I, G, S, L) = \sum_{d \in D, i \in I, g \in G, s \in S, l \in L} P(d, i, g, s, l)$$

$$\sum_{D,I,G,S,L} P(D, I, G, S, L) = \sum_{D,I,G,S,L} P(D)P(I)P(G \mid I, D)\underbrace{P(S \mid I)P(L \mid G)}_{1} =$$

$$= \sum_{D,I,G,S} P(D)P(I)\underbrace{P(G \mid I, D)P(S \mid I)\sum_{L} P(L \mid G)} =$$

1

$$= \sum_{D,I,G} P(D)P(I)P(G \mid I, D)\sum_{S} P(S \mid I) =$$

Is $\mathscr{B}^{student}$ a Probability Distribution? (just considering our specific example)

$$\sum_{D,I,G,S,L} P(D, I, G, S, L) =$$

$$= \sum_{D,I,G} P(D)P(I)P(G \mid I, D) \sum_{S} P(S \mid I) =$$

$$= \sum_{D,I} P(D)P(I) \sum_{G} P(G \mid I, D) =$$

$$= \sum_{D,I} P(D)P(I) = 1$$

Typical method of proving statements

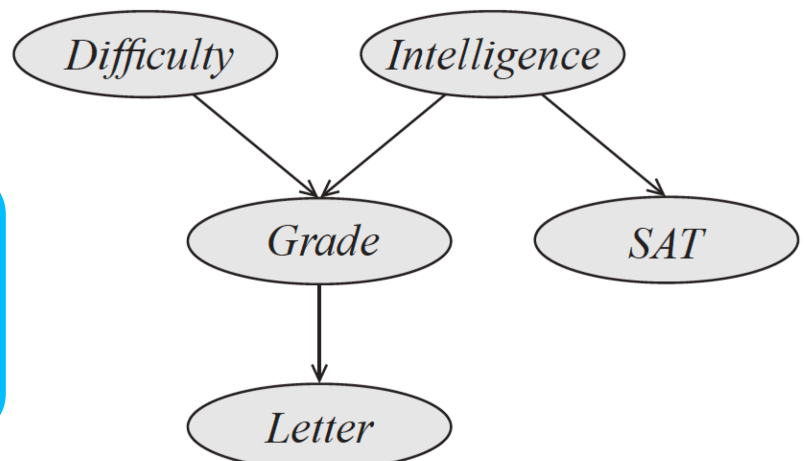
Bayesian Network Graph

A BAYESIAN NETWORK GRAPH IS

a data structure that provides the skeleton for representing a joint distribution compactly in a factorized way.

A BAYESIAN NETWORK GRAPH IS

a compact representation for a set of conditional independence assumptions about a distribution.



Factorization implies independences

Read conditional indpendences from G

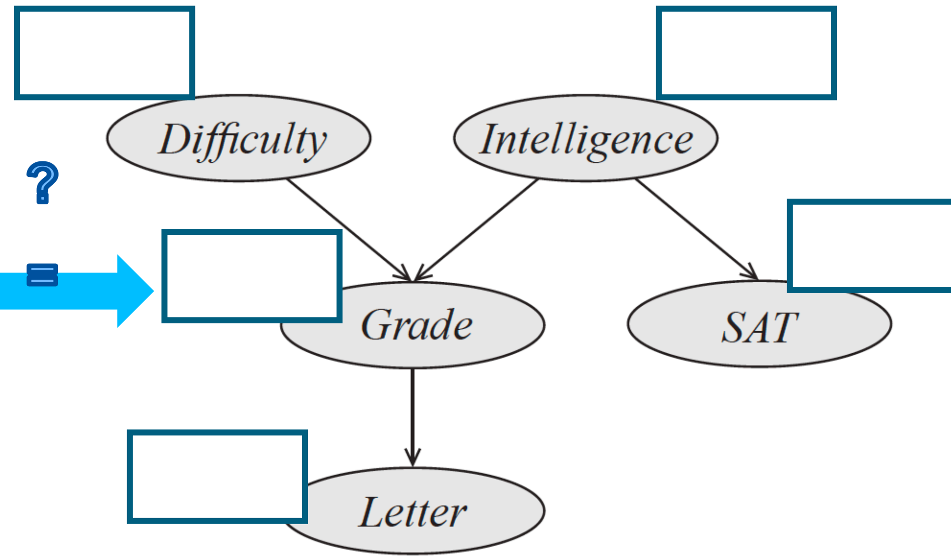
Challenge of Soundness

PROBABILISTIC MODEL

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

$$(G \perp S \mid I) \quad \ldots \quad (D \perp I) \quad (L \perp S \mid G)$$

BAYESIAN NETWORK GRAPH



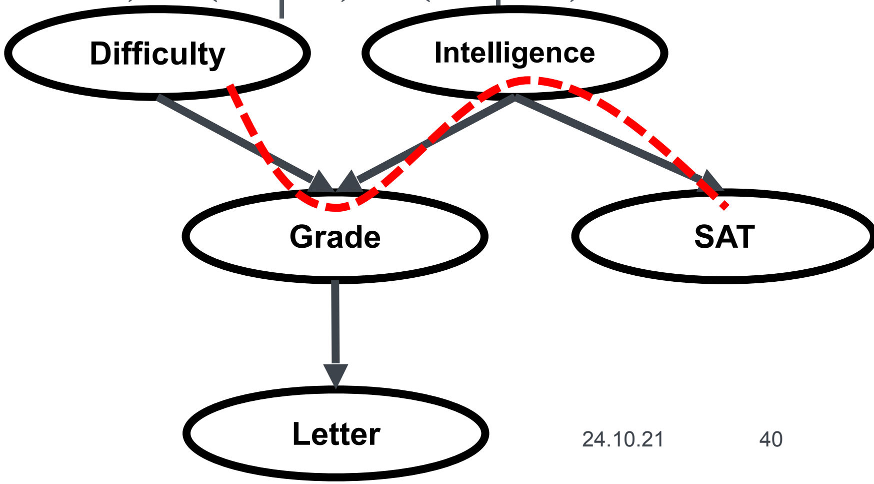
From Factorization to Independence (I)
Illustrated by example

Theorem: If $P(\mathscr{X})$ factorizes over G and $\text{d-sep}_G(X, Y \mid Z)$ then $P(\mathscr{X})$ satisfies $(X \perp Y \mid Z)$.

Chain rule of BN:

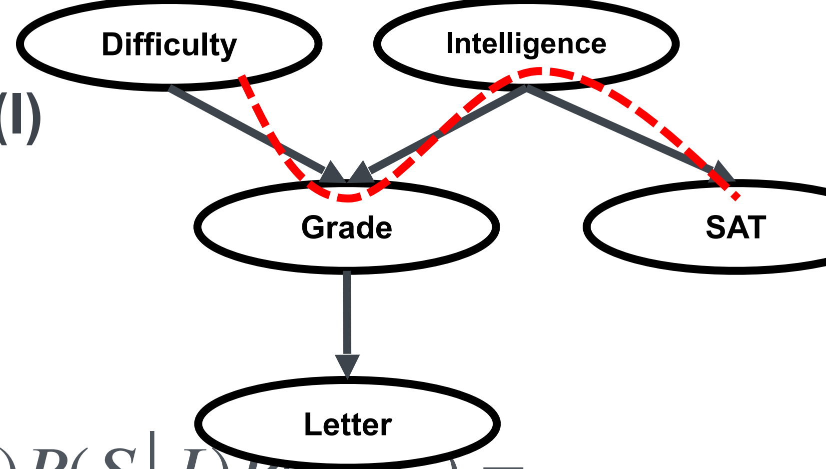
$$P(D, I, G, S, L) = P(D)P(I)P(G \mid D, I)P(S \mid I)P(L \mid G)$$

Is $\text{d-sep}_G(D, S)$?



From Factorization to Independence (I)

Illustrated by example



Is $\text{d-sep}_G(D, S)$?

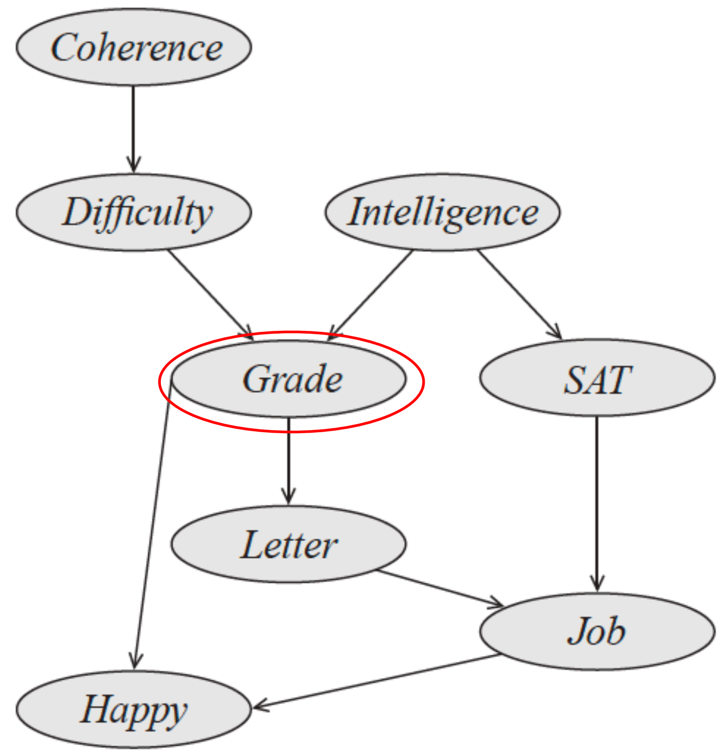
$$P(D, S) = \sum_{G,L,I} P(D)P(I)P(G \mid D, I)P(S \mid I)P(L \mid G) =$$

$$= \sum_{I} P(D)P(I)P(S \mid I) \sum_{G} P(G \mid D, I) \sum_{L} P(L \mid G) =$$

$$= P(D) \sum_{I} P(I)P(S \mid I) = P(D)P(S)$$

Given: Any node is d-separated from its non-descendants given its parents

If $P(\mathscr{X})$ factorizes over G, then in $P(\mathscr{X})$ any random variable is independent of its non-descendants given its parents.



Independency Map: I-Map $\mathscr{I}(G)$

- d-separation in G implies that $\boldsymbol{P}(\mathscr{X})$ satisfies corresponding independence statement

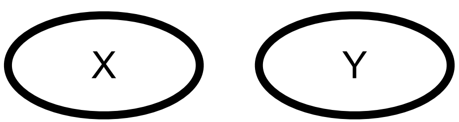
$$\mathscr{I}(G) = \{(\boldsymbol{X} \perp \boldsymbol{Y} \,|\, \boldsymbol{Z}) \colon \text{d--sep}_G(\boldsymbol{X}, \boldsymbol{Y} \,|\, \boldsymbol{Z})\}$$

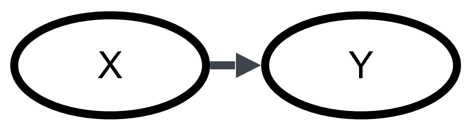
- Definition: If $\boldsymbol{P}(\mathscr{X})$ satisfies $\mathscr{I}(G)$, we say that G is an I-map (independency map) of $\boldsymbol{P}(\mathscr{X})$

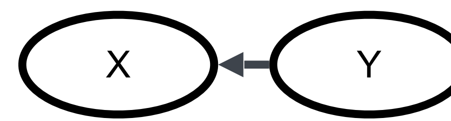
What are the I-Maps of these two distributions $P(X, Y)$?

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	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

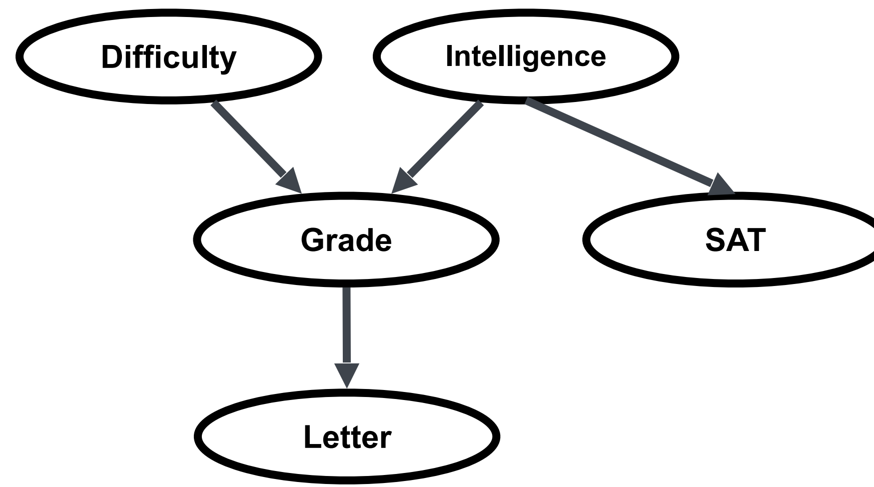






From Factorization to Independence (II) and Vice Versa

Theorem:
If $P(\mathscr{X})$ factorizes over G,
then G is an I-map for $P(\mathscr{X})$.



Theorem:
If G is an I-map for $P(\mathscr{X})$,
then $P(\mathscr{X})$ factorizes over G.

$$P(D, I, G, S, L) = P(D)P(I \mid \cancel{D})P(G \mid D, I)P(S \mid \cancel{D}, I, \cancel{G})P(L \mid \cancel{D}, \cancel{I}, G, \cancel{S})$$

Summary

Two equivalent views of graph structure:

- Factorization: G allows $P(\mathcal{X})$ to be represented
- I-Map: Independencies encoded by G hold in $P(\mathcal{X})$

Question 5: Equivalent Views

6 Knowledge Engineering

Picking Random Variables

- Precision of definition
  - What does „fever“ mean?
    - How reported? (which measurement)
    - When reported? (may fluctuate over day)
    - Observed for how long? Single reading or over protracted time period?
  - What does „sunny“ mean?

Can an objective third party determine the value of a random variable?

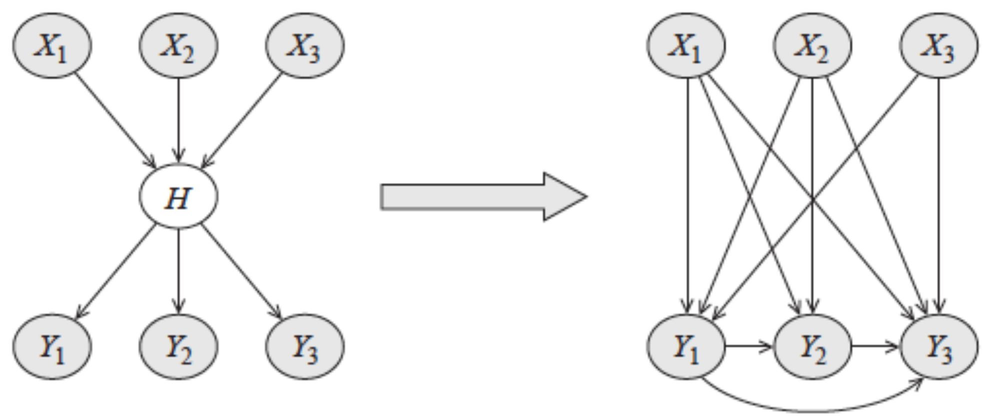
Picking Random Variables

- Parsimony when picking random variables
  - Usually not necessary to model every meal of a patient
  - Usually not necessary to model „fever“ up to precision 0.1 degree

Picking Random Variables

- Latent random variables
  - Most random variables should be observable
  - BUT: adding unobservable latent variables can
    - simplify the model
    - make the model more expressive
    - model properties you cannot/must not/do not want to observe

p. 714

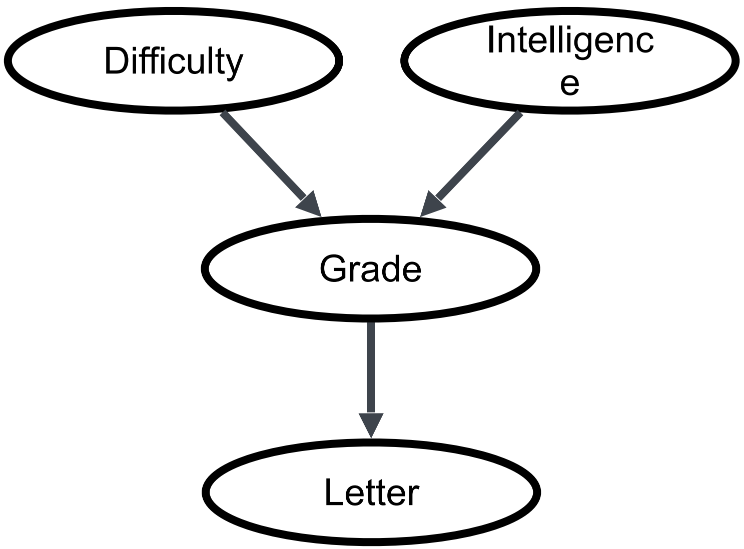


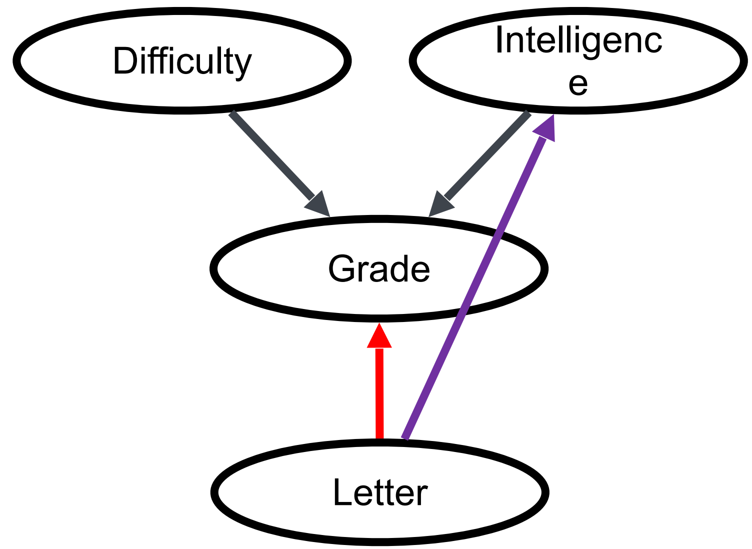
17 parameters   59 parameters

Picking Structures

- Pick structure that reflects causal order
  - causes are parents of the effect
  - causality is in the world – not in the inference process
  - example:
    - insurance company: hadPreviousAccident → badDriver
    - realWorld: badDriver → hadPreviousAccident
- Trade-offs
  - dense connections: computationally difficult, many parameters
  - sparse connections: approximate real world

Picking Structures





Aim at modeling causality (if known)

Turning around some conditional probabilities (without violating independence assumptions!)

Additional (possibly unnecessary) dependencies

Interesting challenges

- When can two graphs G, G' represent the same $P(\mathcal{X})$?
- When are two graphs G, G' equivalent?
- When is a graph G that represents $P(\mathcal{X})$ minimal?

See Section 3.4 in Book Koller & Friedman

Picking Probabilities

- Asking people
  - Try to ask common vocabulary: „common“, „rare“
  - Relatively small differences do not change result much
    - 0.7 vs 0.75
  - Orders of magnitude play a major role:
    - 10^{-4} vs. 10^{-5}
- Zero probabilities:
  - unlikely events need to be smoothened otherwise irrecoverable errors are produced
    - cf. Laplace smoothing in ML lecture

Question 6: Knowledge Engineering

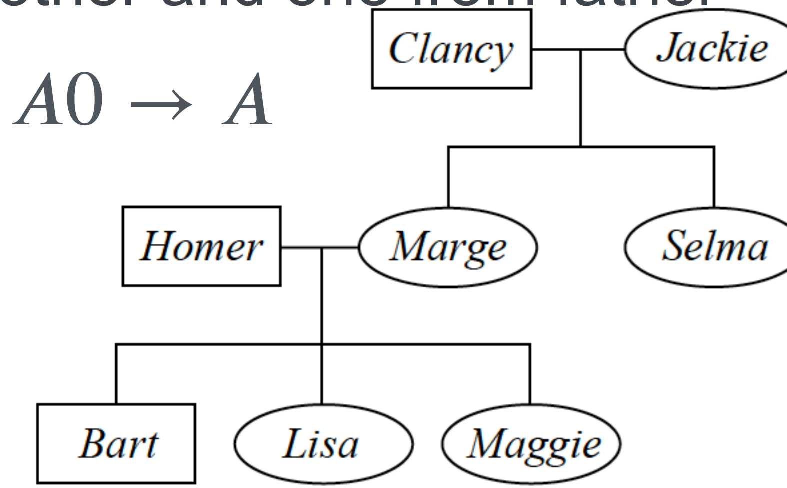
7 Template Variables

Case Study: Genetics in Blood

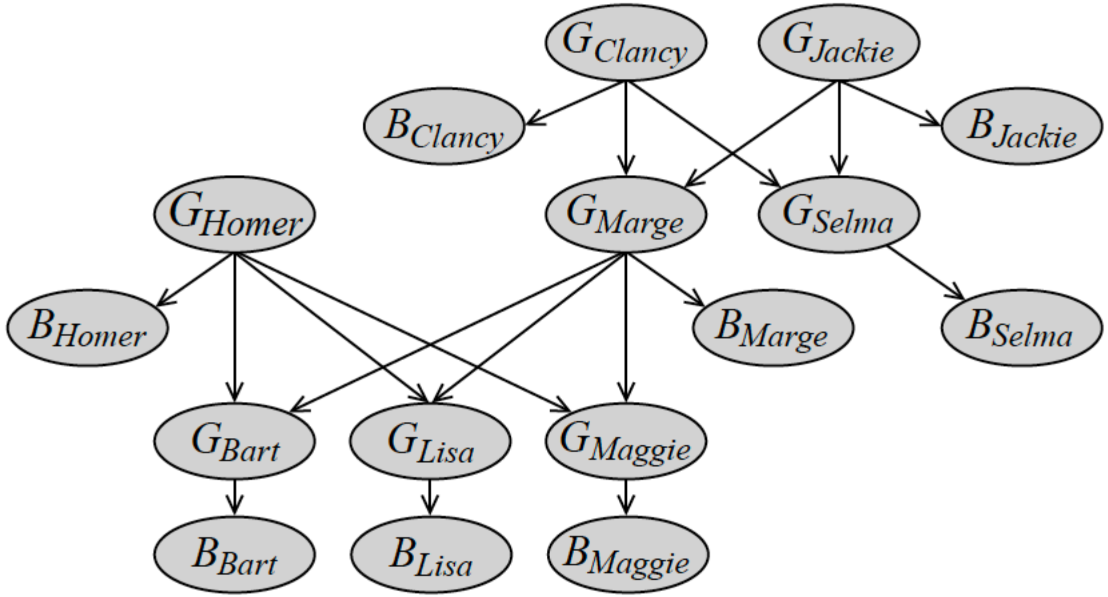
- Genotype: unordered pair of alleles of $\{A, B, 0\} \times \{A, B, 0\}$
- Everyone gets one allele from mother and one from father
- Phenotype: $\{A, B, AB, 0\}$, e.g. $A0 \rightarrow A$

Phenotype is easy to observe
Genotype is not

Interesting questions:
Given phenotype of C,J,H
what phenotype has B?
Can Lisa be Homer‘s daughter?



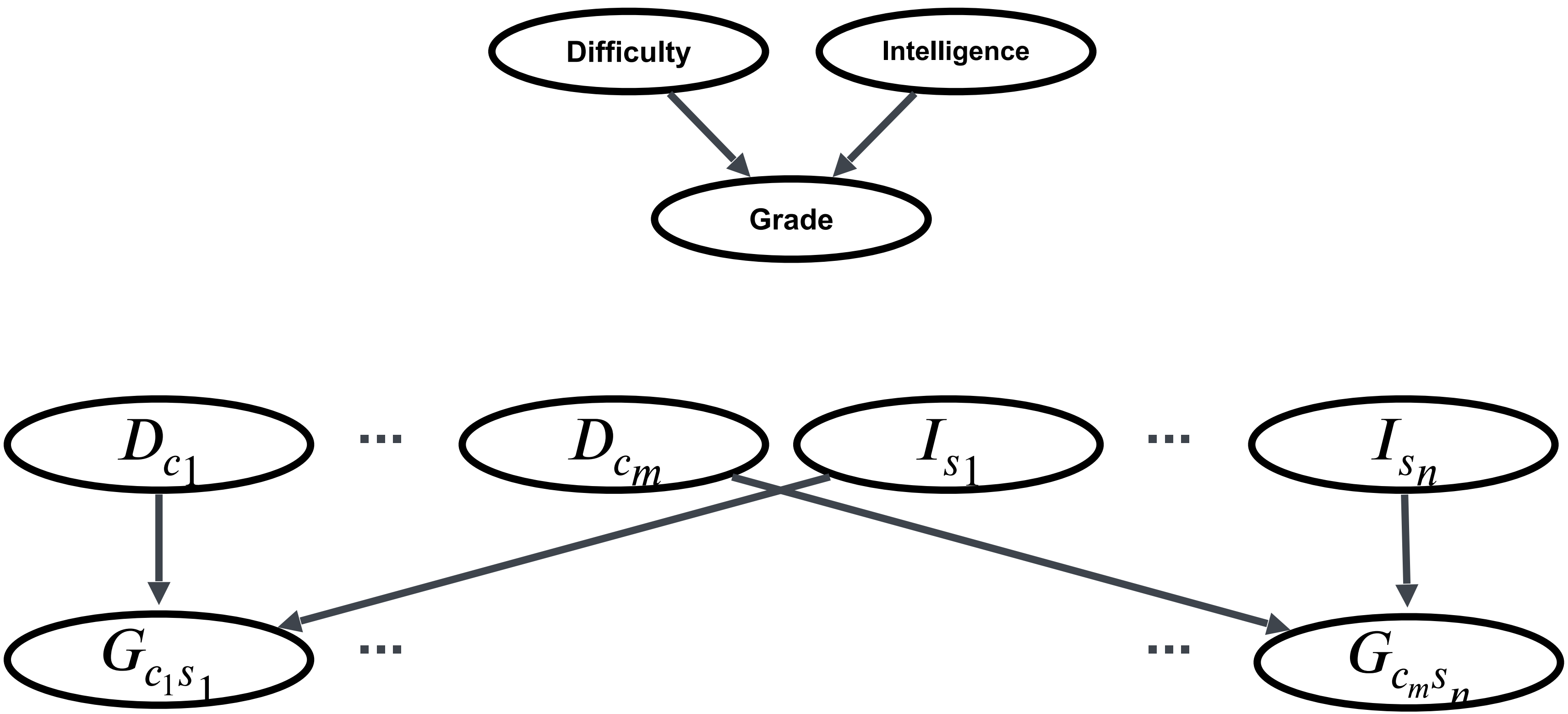
Bayesian Network



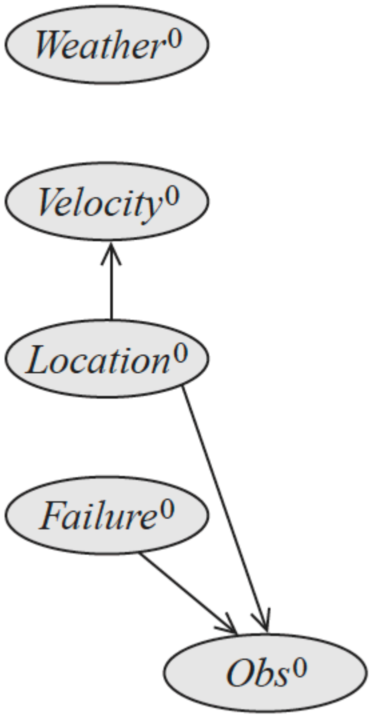
Tedious!!!

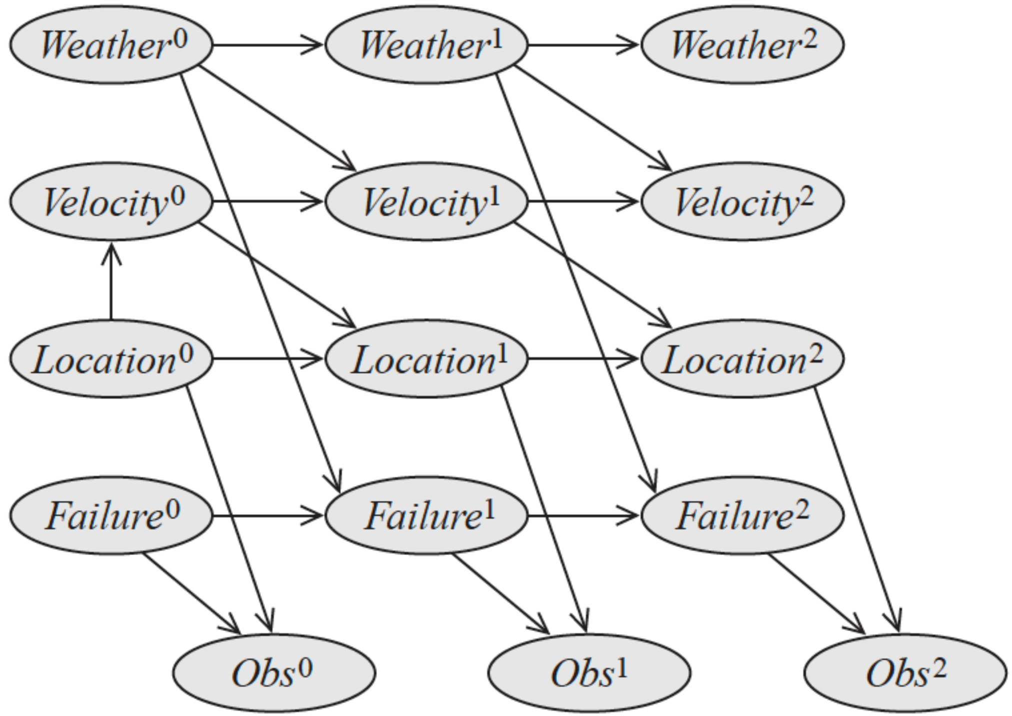
Sharing within and between models!

Recurring Structures for Courses and Students



Temporal Models





- One time slice
- Three out of (arbitrarily many) time slices

Template Variables

- Template variable $X(U_1, \ldots, U_k)$ is instantiated multiple times
  - Genotype(person), Phenotype(person)
  - Difficulty(course), Intelligence(student), Grade(course, student)
  - Obs(time), Failure(time), Location(time),...

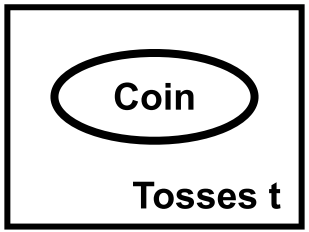
Template Models

Languages that specify how *ground variables* inherit dependency model from template

- Temporal Models: Dynamic Bayesian Networks
- Object-relational models
  - Directed: Plate models
  - Undirected

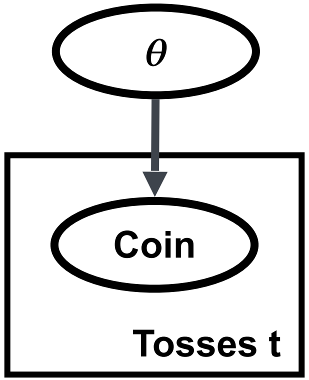
8 Plate Notation

Modeling Repetition



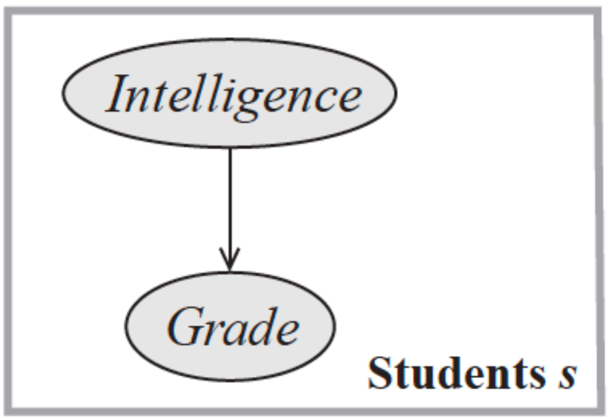
The same (conditional) probability table is instantiated for all Coin_t

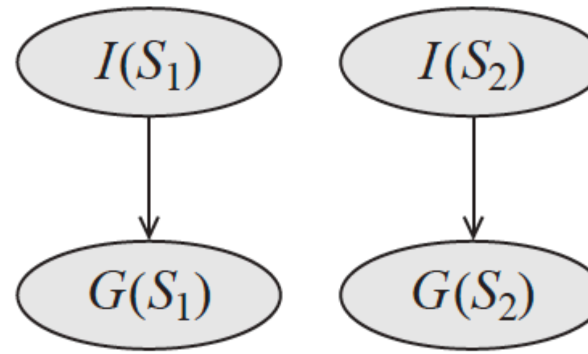
Modeling Repetition



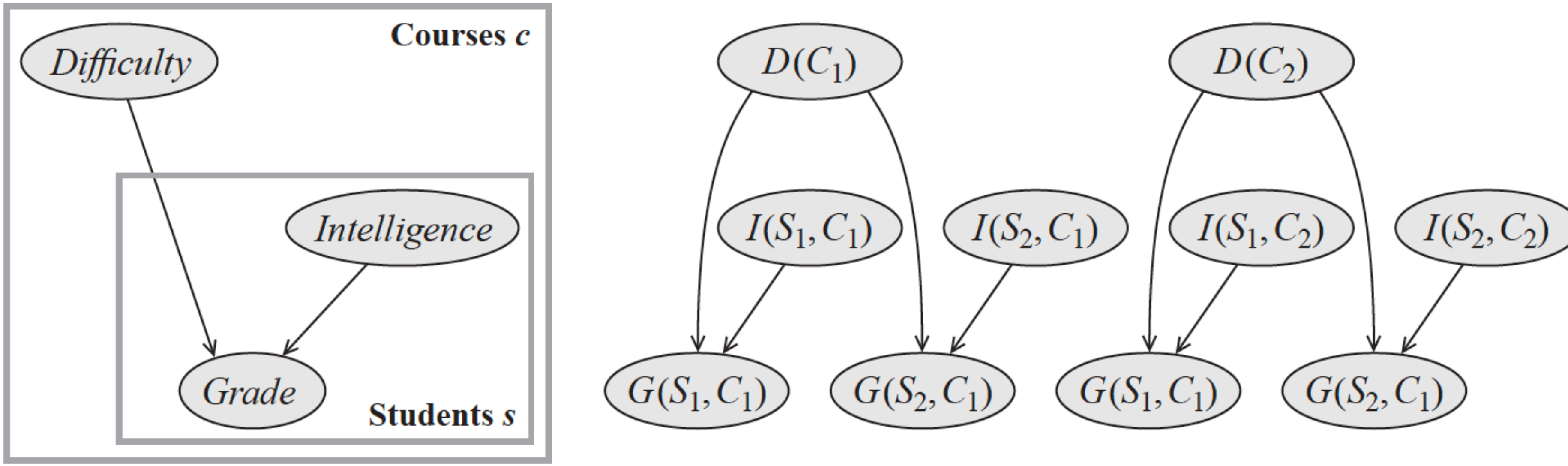
Parameter (representing CPD) is outside the box denoting explicitly that it is constant

Further Example

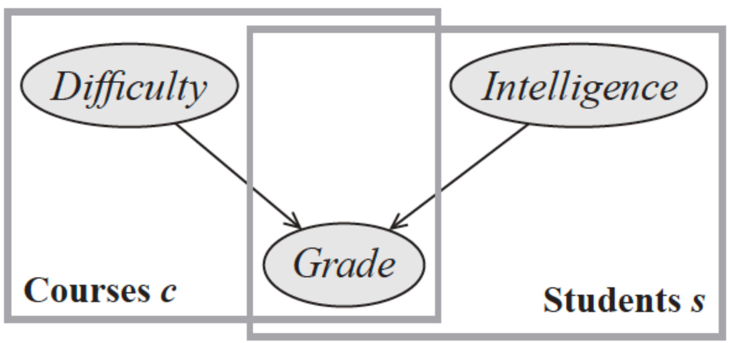


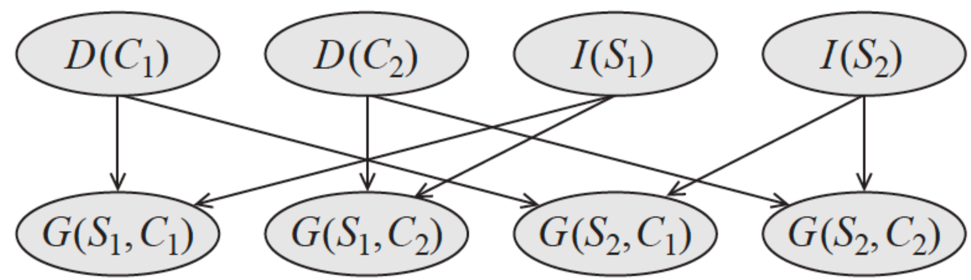


Nested Plates



Nested Plates: Parameter Sharing





Using Parameter Sharing for Collective Inference

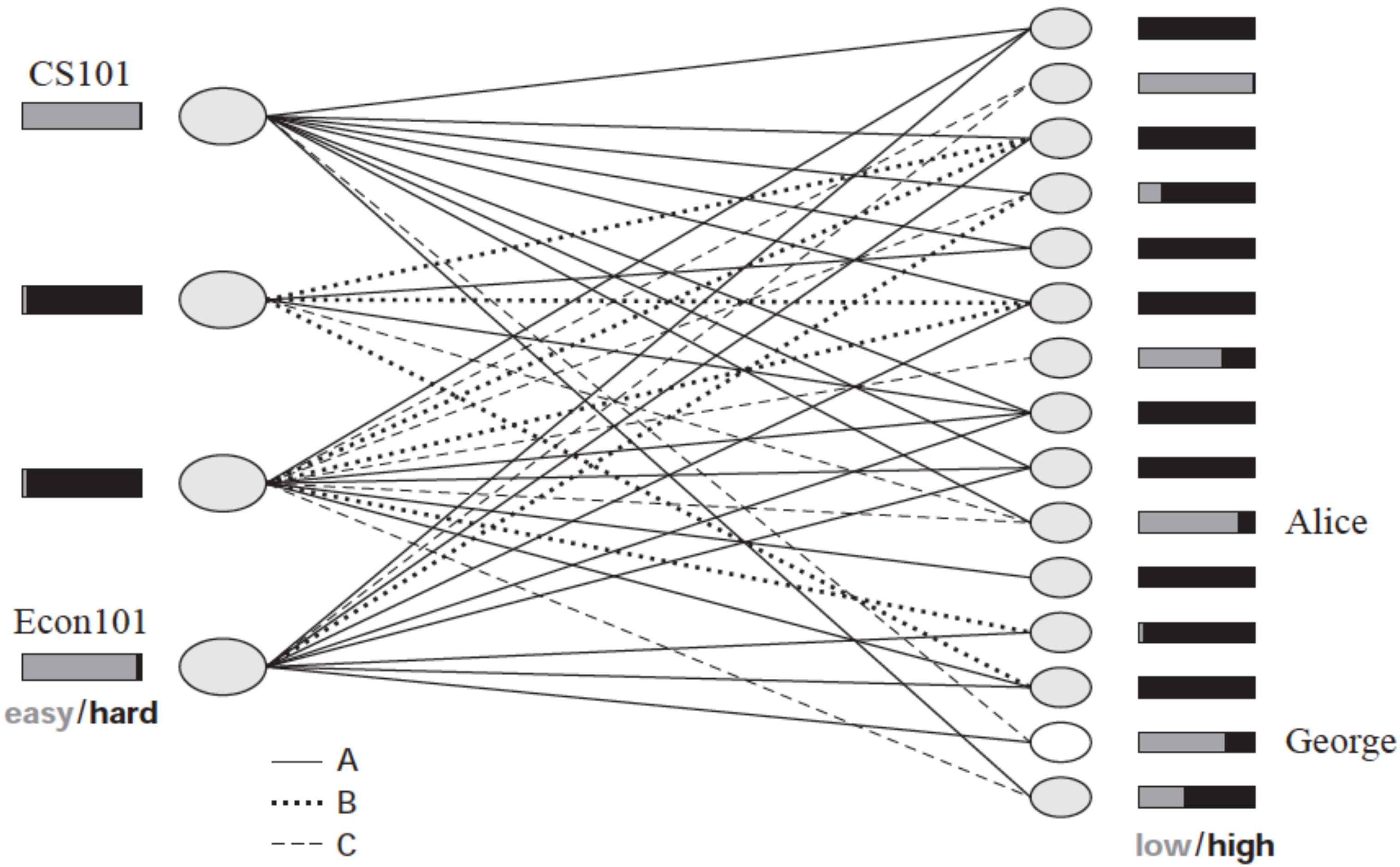


Plate Dependency Model (I)

- Template variable
  $A(U_1, \ldots, U_k)$
  - Template parents
    $B_1(U_1), \ldots B_m(U_m)$
  - CPD $P\left(A \middle| B_1, \ldots, B_m\right)$
- Example:
  - $G(c, s),\ I(c, s)$
  - Template parents: $D(c)$
  - CPD: $P\left(G(c, s) \middle| D(c), I(c, s)\right)$

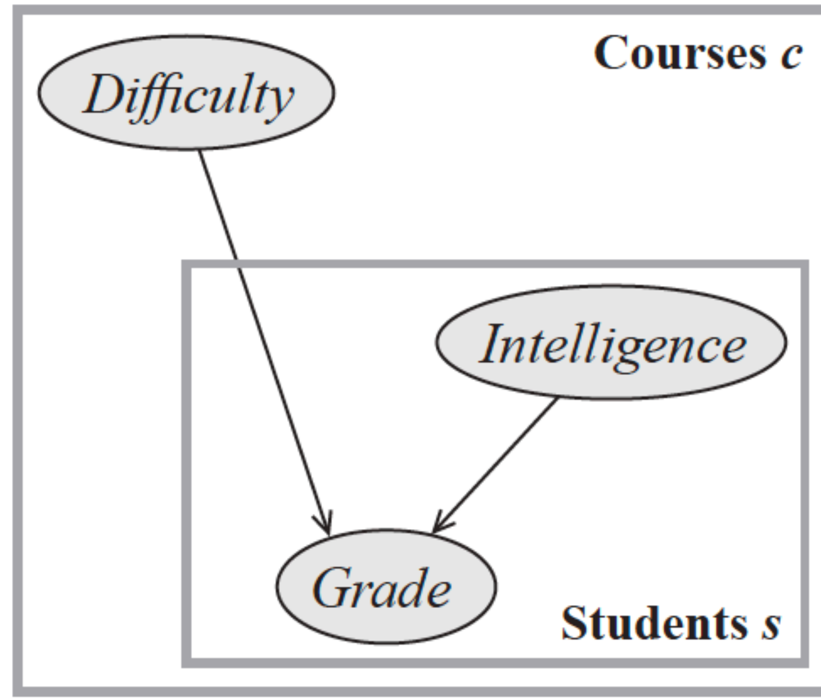
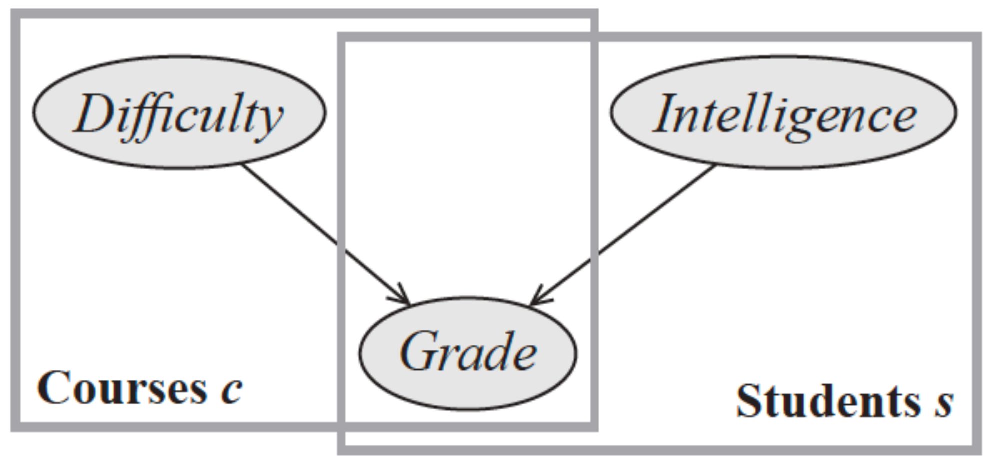


Plate Dependency Model (II)

- Template variable
  $A(U_1, \ldots, U_k)$
  - Template parents
    $B_1(U_1), \ldots B_m(U_m)$
  - CPD $P\left(A \middle| B_1, \ldots, B_m\right)$



- Example:
  - $G(c, s)$
  - Template parents: $D(c), I(s)$
  - CPD: $P\left(G(c, s) \middle| D(c), I(s)\right)$

Plate Dependency Model (III)

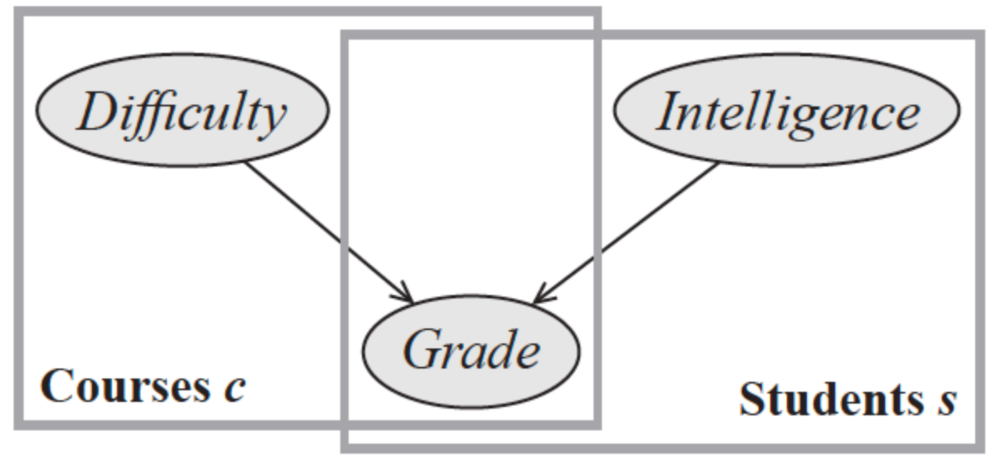
- Template variable

  $A(U_1, \ldots, U_k)$

  - Template parents

    $B_1(U_1), \ldots B_m(U_m)$

  - CPD $P\left(A \mid B_1, \ldots, B_m\right)$



No indices in parent that are not in child

Large variety of language extensions exists

Summary on Plate Notation

- Templates allow for denoting an infinite set of Bayesian Networks, each induced by a different set of domain objects
- Parameters and structure are reused within a BN and across different BNs
- Models encode correlations across multiple objects allowing collective inference

Thank you!

Steffen Staab

E-Mail Steffen.staab@ipvs.uni-stuttgart.de

Telefon +49 (0) 711 685- To be defined

www. ipvs.uni-stuttgart.de/departments/ac/

Universität Stuttgart
Analytic Computing, IPVS
Universitätsstraße 32, 50569 Stuttgart