# Fundamentals of AI and KR - Module 3

## 2. Bayesian network representation

Paolo Torroni

Fall 2023

# Notice

## Credits

The present slides are largely an adaptation of existing material, including:

- slides from Russel & Norvig
- slides by Daphne Koller on Probabilistic Graphical Models
- slides by Fabrizio Riguzzi on Data Mining and Analytics

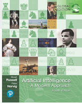
I am especially grateful to these authors.

## Downloading and ~~sharing~~

# Table of Contents


- Independence
- Bayesian network representation

# $\boxed{\text{Independence}}$




# Independence

$P(A)$

$P(A|B) = \frac{P(A,B)}{P(B)}$

$A$ and $B$ are independent, denoted $\mathbf{P} \models (A \perp B)$, iff

$\mathbf{P}(A|B) = \mathbf{P}(A)$ or $\mathbf{P}(B|A) = \mathbf{P}(B)$ or $\mathbf{P}(A,B) = \mathbf{P}(A)\mathbf{P}(B)$

1 2 3

Cavity
Toothache Catch
Weather

decomposes into $\Rightarrow$

Cavity
Toothache Catch

Weather

$P \models (W \perp T, Cat, Cav)$

$\mathbf{P}(Toothache, Catch, Cavity, Weather)$
$= \mathbf{P}(Toothache, Catch, Cavity)\mathbf{P}(Weather)$

32 entries reduced to 12; for $n$ independent biased coins, $2^n \to n$

Absolute (marginal) independence powerful but rare.
Dentistry is a large field with hundreds of variables, none of which are independent. What to do?

# Conditional independence

$\mathbf{P}(Toothache, Cavity, Catch)$ has $2^3 - 1 = 7$ independent entries

If I have a cavity, the probability that the probe catches in it doesn't depend on whether I have a toothache:

(1) $P(catch|toothache, \boxed{cavity}) = P(catch|\boxed{cavity})$

The same independence holds if I haven't got a cavity:

(2) $P(catch|toothache, \neg cavity) = P(catch|\neg cavity)$

*Catch* is conditionally independent of *Toothache* given *Cavity*:

$\mathbf{P}(Catch|Toothache, Cavity) = \mathbf{P}(Catch|Cavity)$

Equivalent statements:

- $\mathbf{P}(Toothache|Catch, Cavity) = \mathbf{P}(Toothache|Cavity)$
- $\mathbf{P}(Toothache, Catch|Cavity) = \mathbf{P}(Toothache|Cavity)\mathbf{P}(Catch|Cavity)$

Notation: $\mathbf{P} \models (Catch \perp Toothache | Cavity)$

# Conditional independence

Write out full joint distribution using chain rule:

$\mathbf{P}(Toothache, Catch, Cavity)$ ← 7 indep. values

$= \mathbf{P}(Toothache|Catch, Cavity)\mathbf{P}(Catch|Cavity)\mathbf{P}(Cavity)$

$= \mathbf{P}(Toothache|Cavity)\mathbf{P}(Catch|Cavity)\mathbf{P}(Cavity)$

2

i.e., $2 + 2 + 1 = 5$ independent numbers

In most cases, the use of conditional independence reduces the size of the representation of the joint distribution from exponential in $n$ to linear in $n$.

**Conditional independence is our most basic and robust form of knowledge about uncertain environments.**

# Bayes' Rule

Product rule $P(a \wedge b) = P(a|b)P(b) = P(b|a)P(a)$

$$\Rightarrow \text{ Bayes' rule } P(a|b) = \frac{P(b|a)P(a)}{P(b)}$$

or in distribution form

$$\mathbf{P}(Y|X) = \frac{\mathbf{P}(X|Y)\mathbf{P}(Y)}{\mathbf{P}(X)} = \alpha\mathbf{P}(X|Y)\mathbf{P}(Y)$$

Useful for assessing diagnostic probability from causal probability:

$$P(Cause|Effect) = \frac{P(Effect|Cause)P(Cause)}{P(Effect)}$$

?

# Example of diagnosis using Bayes' Rule

$$P(Cause|Effect) = \frac{P(Effect|Cause)P(Cause)}{P(Effect)}$$

Say 1 individual in 50,000 suffers from meningitis, 1% from a stiff neck, and 70% of the times meningitis causes a stiff neck. *What is the probability that an individual with a stiff neck has meningitis?*

# Example of diagnosis using Bayes' Rule

$$P(Cause|Effect) = \frac{P(Effect|Cause)P(Cause)}{P(Effect)}$$

Say 1 individual in 50,000 suffers from meningitis, 1% from a stiff neck, and 70% of the times meningitis causes a stiff neck. *What is the probability that an individual with a stiff neck has meningitis?*

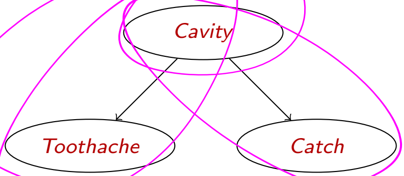
Let $M$ be meningitis and $S$ be stiff neck.
$P(m) = 1/50,000$, $P(s) = 0.01$, $P(s|m) = 0.7$.

$$P(m|s) = \frac{P(s|m)P(m)}{P(s)} = \frac{0.7 \times 1/50,000}{0.01} = 0.0014$$

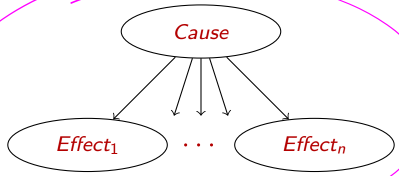
Note: posterior probability of meningitis still very small!

# Bayes' Rule and conditional independence

$$
\begin{aligned}
&\mathbf{P}(Cavity \mid toothache \wedge catch) \\
&\quad = \alpha\, \mathbf{P}(toothache \wedge catch \mid Cavity)\mathbf{P}(Cavity) \\
&\quad = \alpha\, \mathbf{P}(toothache \mid Cavity)\mathbf{P}(catch \mid Cavity)\mathbf{P}(Cavity)
\end{aligned}
$$



# Bayes' Rule and conditional independence




This is an example of a naive Bayes model:

$$\underline{\mathbf{P}(Cause, Effect_1, \ldots, Effect_n)} = \mathbf{P}(Cause) \prod_i \mathbf{P}(\underline{Effect_i} | \underline{Cause})$$

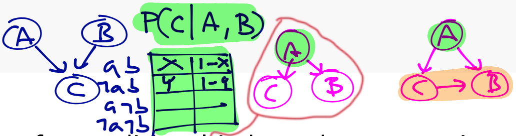
Total number of parameters is **linear** in $n$

# Summary so far

- Probability is a rigorous formalism for uncertain knowledge
- Joint probability distribution specifies probability of every atomic event
- Queries can be answered by summing over atomic events
- For nontrivial domains, we must find a way to reduce the joint size
- Independence and conditional independence provide the tools

# Bayesian network representation

# Bayesian networks




A simple, graphical notation for conditional independence assertions and hence for compact specification of full joint distributions.

KNOWN A, C and B are independent

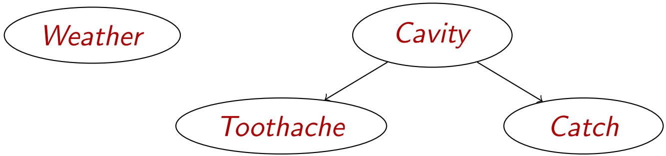

Syntax:

- a set of nodes, one per variable
- a directed, acyclic graph (link ≈ "directly influences")
- a conditional distribution for each node given its parents: $\mathbf{P}(X_i|Parents(X_i))$

In the simplest case, conditional distribution represented as a conditional probability table (CPT) giving the distribution over $X_i$ for each combination of parent values

# Example

Topology of network encodes conditional independence assertions:




- *Weather* is independent of the other variables
- *Toothache* and *Catch* are conditionally independent given *Cavity*

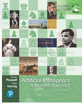
# Example

I'm at work, neighbor John calls to say my alarm is ringing, but neighbor Mary doesn't call. Sometimes it's set off by minor earthquakes. Is there a burglar?

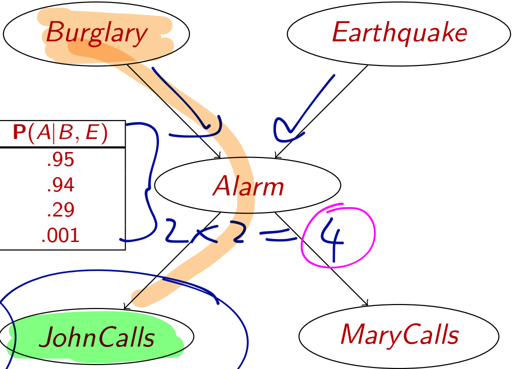
Variables: *Burglar*, *Earthquake*, *Alarm*, *JohnCalls*, *MaryCalls*

Network topology reflects "causal" knowledge:

- A burglar can set the alarm off
- An earthquake can set the alarm off
- The alarm can cause Mary to call
- The alarm can cause John to call



# Example




1

| **P**(*B*) |
|---|
| 0.001 |

2 2

| **P**(*E*) | P(¬e) |
|---|---|
| .002 | .998 |

1

| *B* | *E* | **P**(*A*\|*B*, *E*) |
|---|---|---|
| *T* | *T* | .95 |
| *T* | *F* | .94 |
| *F* | *T* | .29 |
| *F* | *F* | .001 |

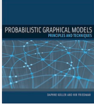
$2 \times 2 = 4$

| *A* | **P**(*J*\|*A*) |
|---|---|
| *T* | .90 |
| *F* | .05 |

2

| *A* | **P**(*M*\|*A*) |
|---|---|
| *T* | .70 |
| *F* | .01 |

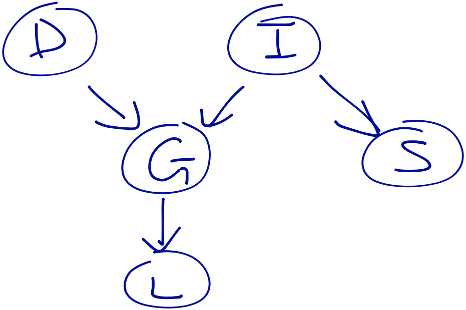
2




# Reasoning patterns

**The student network**

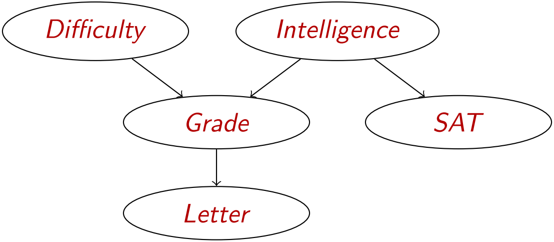
A student's grade depends on intelligence and on the difficulty of the course. SAT scores are correlated with intelligence. A professor writes recommendation letters by only looking at grades.



# Reasoning patterns

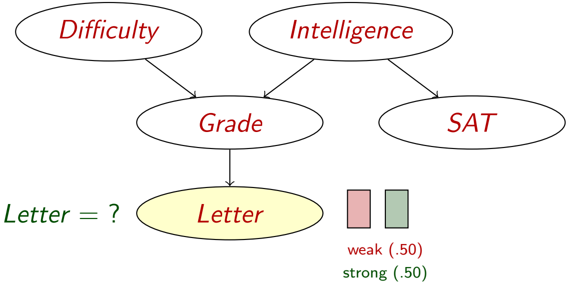
### The student network

A student's grade depends on intelligence and on the difficulty of the course. SAT scores are correlated with intelligence. A professor writes recommendation letters by only looking at grades.



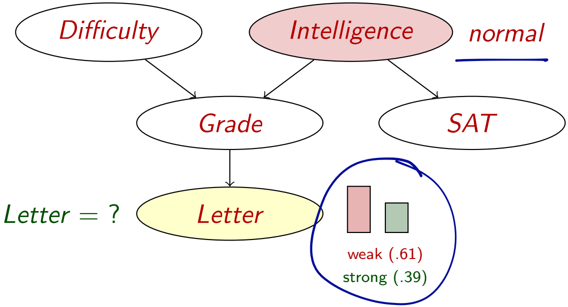
# "Reasoning patterns"

- Causal: will George get a strong reference letter? (prediction)



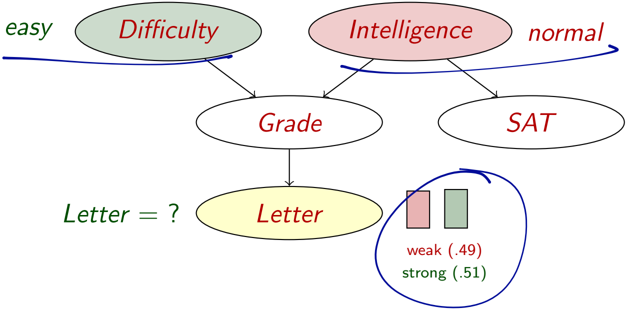
# Reasoning patterns

- Causal: will George get a strong reference letter? (prediction)



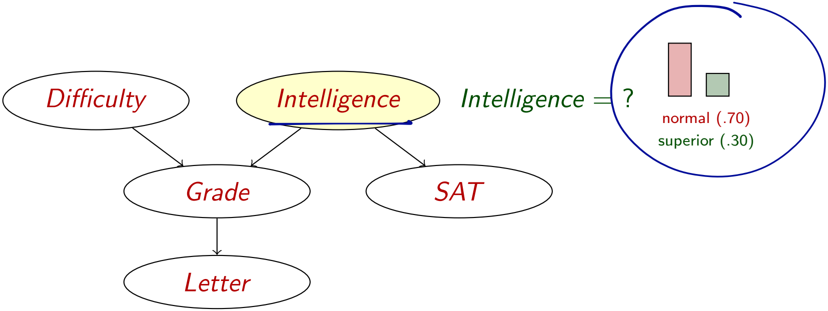
# Reasoning patterns

- Causal: will George get a strong reference letter? (prediction)



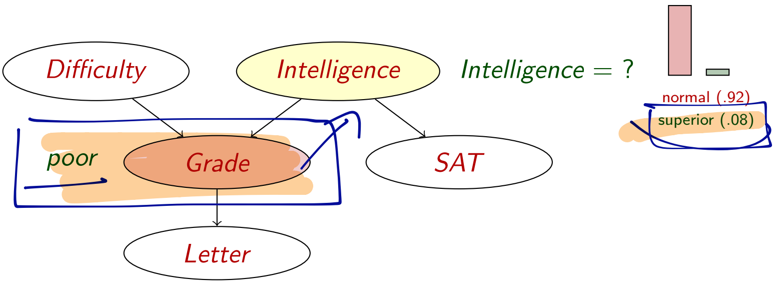
# Reasoning patterns

- Causal: will George get a strong reference letter? (prediction)
- Evidential: is George a good potential recruit? (explanation)



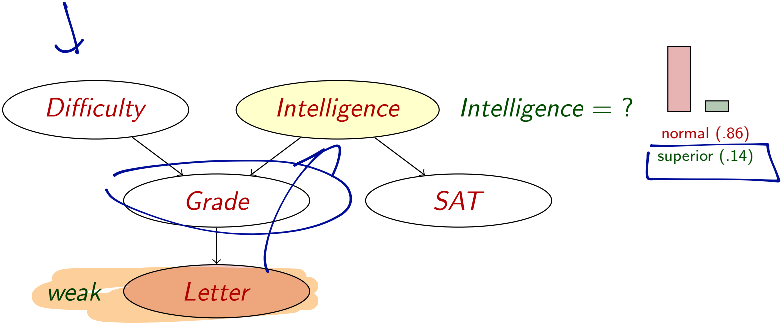
# Reasoning patterns

- Causal: will George get a strong reference letter? (prediction)
- Evidential: is George a good potential recruit? (explanation)



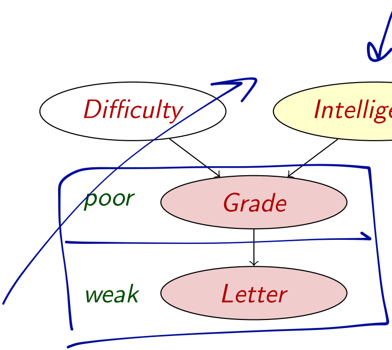
# Reasoning patterns

- Causal: will George get a strong reference letter? (prediction)
- Evidential: is George a good potential recruit? (explanation)

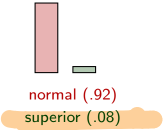


# Reasoning patterns

- Causal: will George get a strong reference letter? (prediction)
- Evidential: is George a good potential recruit? (explanation)



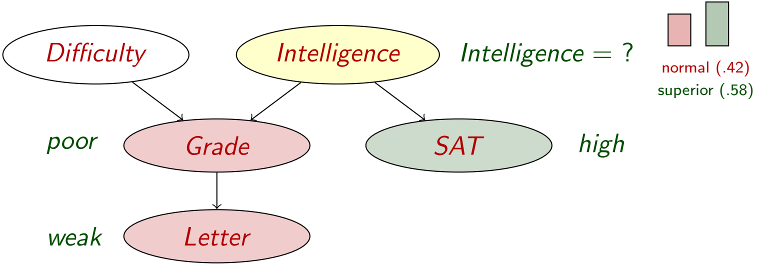

*Intelligence* = ?




PF ( I ⊥ L | G )

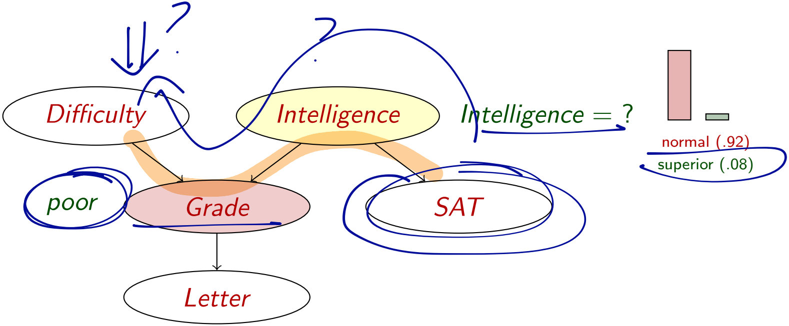
# Reasoning patterns

- Causal: will George get a strong reference letter? (prediction)
- Evidential: is George a good potential recruit? (explanation)



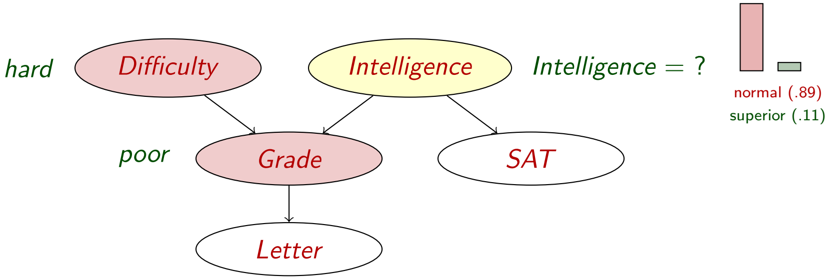
# Reasoning patterns

- Causal: will George get a strong reference letter? (prediction)
- Evidential: is George a good potential recruit? (explanation)
- Intercausal: why did George score low/high? (explaining away)



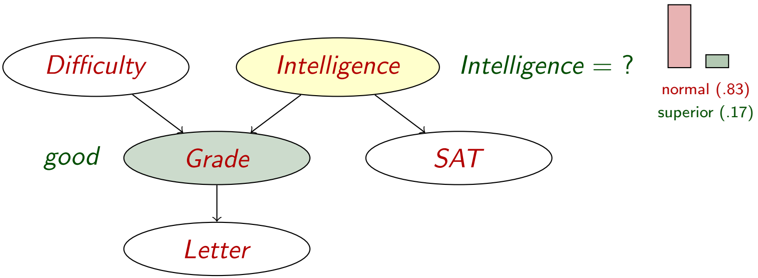
# Reasoning patterns

- Causal: will George get a strong reference letter? (prediction)
- Evidential: is George a good potential recruit? (explanation)
- Intercausal: why did George score low/high? (explaining away)



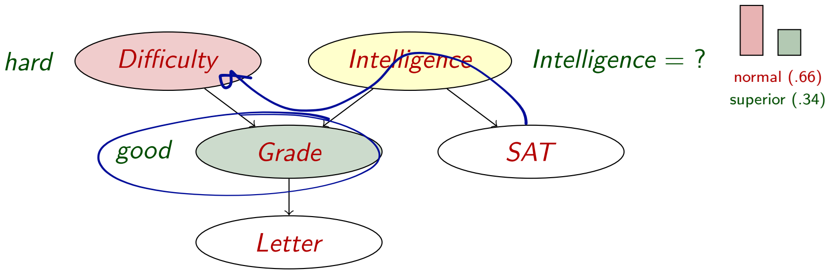
# Reasoning patterns

- Causal: will George get a strong reference letter? (prediction)
- Evidential: is George a good potential recruit? (explanation)
- Intercausal: why did George score low/high? (explaining away)



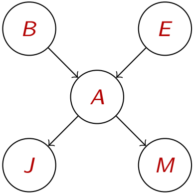
# Reasoning patterns

- Causal: will George get a strong reference letter? (prediction)
- Evidential: is George a good potential recruit? (explanation)
- Intercausal: why did George score low/high? (explaining away)



# Compactness

A CPT for Boolean $X_i$ with $k$ Boolean parents has $2^k$ rows for the combinations of parent values
Each row requires one number $p$ for $X_i = true$
(the number for $X_i = false$ is just $1 - p$)




If each variable has no more than $k$ parents, the complete network requires $O(n \cdot 2^k)$ numbers

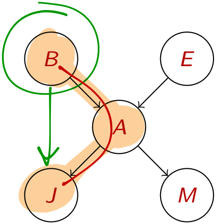
I.e., grows linearly with $n$, vs. $O(2^n)$ for the full joint distribution

For burglary net, $1 + 1 + 4 + 2 + 2 = 10$ numbers (vs. $2^5 - 1 = 31$)

For student net, $1 + 1 + 8 + 2 + 3 = 15$ numbers (vs. $2^4 \times 3 - 1 = 47$)

# Global semantics

Global semantics defines the full joint distribution as the product of the local conditional distributions:




$$P(x_1, \ldots, x_n) = \prod_{i=1}^{n} P(x_i | parents(X_i))$$

e.g., $P(j \wedge m \wedge a \wedge \neg b \wedge \neg e)$ →

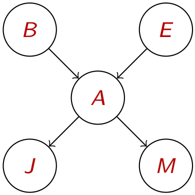
→ $P(j|a) P(m|a) P(a|\neg b, e) P(\neg b) P(e)$

$P(j|a, m, \neg b, e) P(a|m, \neg b, e) P(\ldots$

# Global semantics

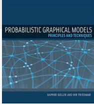
Global semantics defines the full joint distribution as the product of the local conditional distributions:

$$P(x_1, \ldots, x_n) = \prod_{i=1}^{n} P(x_i | parents(X_i))$$




e.g., $P(j \wedge m \wedge a \wedge \neg b \wedge \neg e)$

$$\begin{aligned} &= P(j|a)P(m|a)P(a|\neg b, \neg e)P(\neg b)P(\neg e) \\ &= 0.9 \times 0.7 \times 0.001 \times 0.999 \times 0.998 \\ &\approx 0.00063 \end{aligned}$$

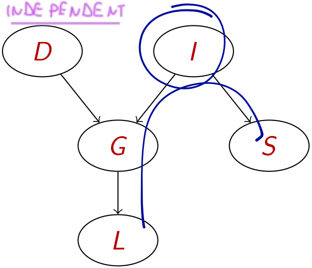



# Basic independences in the Student network

IF YOU DON'T KNOW THE GRADE, POINTS AND DIFFICULTY ARE INDEPENDENT

$P \models (S \perp D)$

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




What independences?

- $\mathbf{P} \models (L \perp \ldots ?$
- $\mathbf{P} \models (S \perp \ldots ?$
- $\mathbf{P} \models (G \perp \ldots ?$
- $\mathbf{P} \models (I \perp \ldots ?$
- $\mathbf{P} \models (D \perp \ldots ?$

$P \models (S \perp L \mid G)$

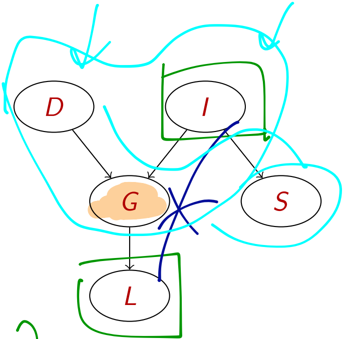
$P \models (S \not\perp D \mid G)$

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

$P \models (\underbrace{L \perp S}_{(S \perp L \mid I)} \mid I)$

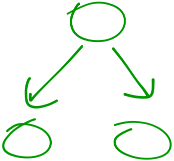
IF YOU KNOW THE GRADE, POINTS AND DIFFICULTY ARE DEPENDENT

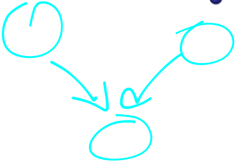
# Basic independences in the Student network




What independences?

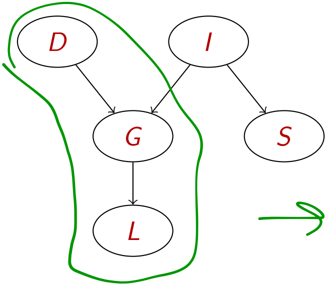
- $\mathbf{P} \models (L \perp I, D, S \mid G)$
- $\mathbf{P} \models (S \perp G, D, L \mid I)$
- $\mathbf{P} \models (G \perp S \mid I)$
- $\mathbf{P} \models (I \perp D)$
- $\mathbf{P} \models (D \perp I, S)$
- ...





V - structure

# Flow of probabilistic influence




Or else, when could $X$ influence $Y$?

- $X \rightarrow Y$ (direct cause)
- $X \leftarrow Y$ (direct effect)
- $X \rightarrow Z \rightarrow Y$ (causal trail)
- $X \leftarrow Z \leftarrow Y$ (evidential trail)
- $X \leftarrow Z \rightarrow Y$ (common cause)
- $X \rightarrow Z \leftarrow Y$ (common effect)

**Definition (active two-edge trail)**

If influence can flow from $X$ to $Y$ via $Z$, the trail $X \rightleftharpoons Z \rightleftharpoons Y$ is active

# Flow of probabilistic influence: active trails

Consider a longer trail $X_1 \rightleftharpoons \cdots \rightleftharpoons X_n$.

For influence to flow from $X_1$ to $X_n$, it needs to flow through every single node on the trail

This is true if and only if every two-edge trail $X_{i-1} \rightleftharpoons X_i \rightleftharpoons X_{i+1}$ along the trail allows influence to flow

**Definition (active trail)**

Let $\mathbf{Z}$ be a subset of observed variables.
The trail $X_{i-1} \rightleftharpoons X_i \rightleftharpoons X_{i+1}$ is active given $\mathbf{Z}$ if

- $\forall X_{i-1} \rightarrow X_i \leftarrow X_{i+1}$, $X_i$ or one of its descendants are in $\mathbf{Z}$
- no other node along the trail is in $\mathbf{Z}$

# Flow of probabilistic influence: direct separation

**Definition (d-separation)**

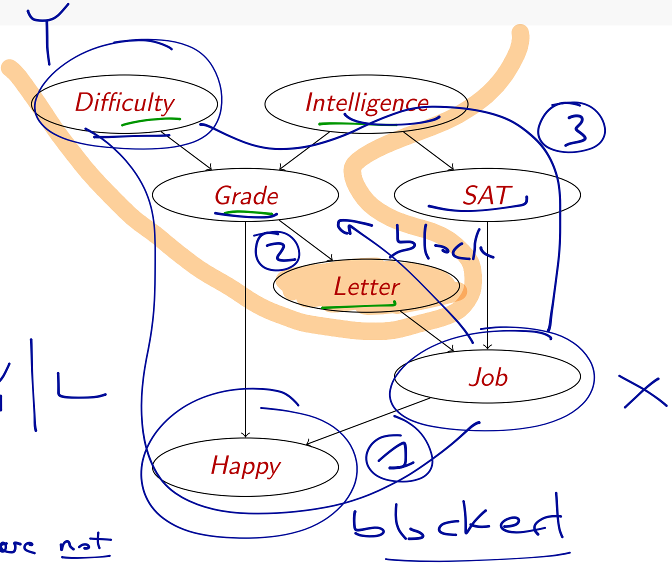
Two sets of nodes $\mathbf{X}$, $\mathbf{Y}$ are d-separated given $\mathbf{Z}$ if there is no active trail between any $X \in \mathbf{X}$ and $Y \in \mathbf{Y}$ given $\mathbf{Z}$

To determine if $X$ and $Y$ are **independent** given $\mathbf{Z}$:

1. traverse the graph bottom-up marking all nodes in $\mathbf{Z}$ or having descendants in given $\mathbf{Z}$
2. traverse the graph from $X$ to $Y$, stopping if we get to a blocked node
3. if we can't reach $Y$, then $X$ and $Y$ are independent

A node is blocked if either the middle of an unmarked v-structure, or in $\mathbf{Z}$ (not both)

# Example

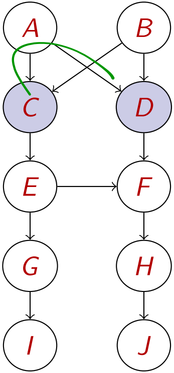



$X \not\perp Y \mid L$

$\Updownarrow$

X and Y are not d. separated given L

# Example

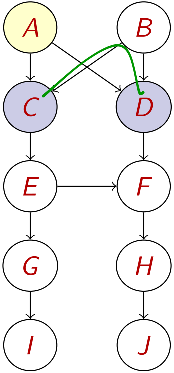



What independences?

- $\mathbf{P} \models (C \perp D)$? no

THERE IS AN ACTIVE TRAIL

# Example

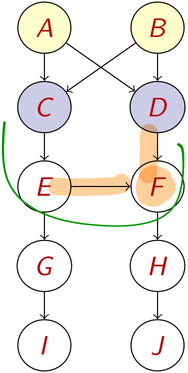



What independences?

- $\mathbf{P} \models (C \perp D)$?
- $\mathbf{P} \models (C \perp D | A)$?

ONE IS ACTIVE BUT THE OTHER ONE NOT

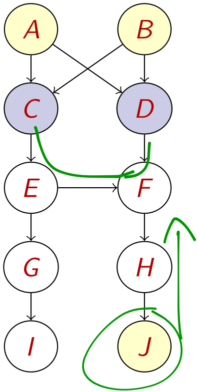
# Example




What independences?

- $\mathbf{P} \models (C \perp D)$?
- $\mathbf{P} \models (C \perp D | A)$?
- $\mathbf{P} \models (C \perp D | A, B)$? yes

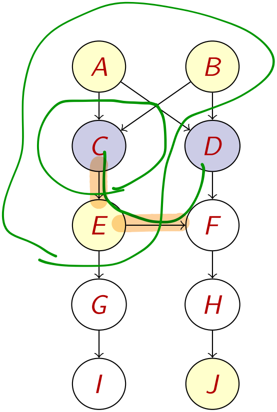
# Example




What independences?

- $\mathbf{P} \models (C \perp D)$?
- $\mathbf{P} \models (C \perp D | A)$?
- $\mathbf{P} \models (C \perp D | A, B)$?
- $\mathbf{P} \models (C \perp D | \underline{A, B, J})$? no

## Example




What independences?

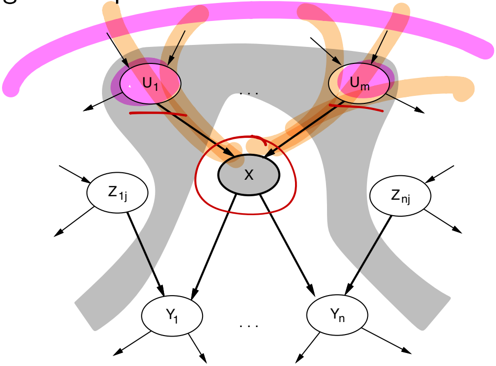
- $\mathbf{P} \models (C \perp D)$?
- $\mathbf{P} \models (C \perp D | A)$?
- $\mathbf{P} \models (C \perp D | A, B)$?
- $\mathbf{P} \models (C \perp D | A, B, J)$?
- $\mathbf{P} \models (C \perp D | A, B, E, J)$? yes

$\boxed{C \perp \text{all other nodes} \mid A, B, E}$

"markov blanket"

# Local semantics

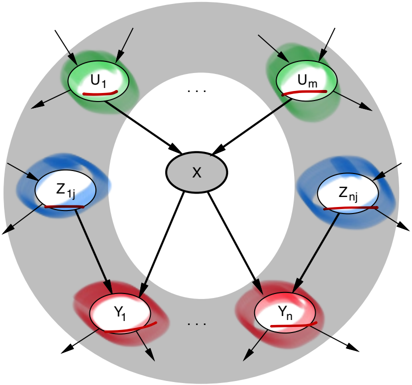
Local semantics: each node is conditionally independent of its nondescendants given its parents




Theorem: Local semantics $\Leftrightarrow$ global semantics

# Markov blanket

Each node is conditionally independent of all others given its Markov blanket: parents + children + children's parents



# Questions?