Graphical Models

Bayesian Networks

Siamak Ravanbakhsh

Fall 2019

Previously on Probabilistic Graphical Models

- Probability distribution and density functions
- Random variable
- Bayes' rule
- Conditional independence
- Expectation and Variance

Learning objectives

- what is a Bayesian network?
  - factorization
  - conditional independencies
    - how to read it from the graph

  how are they related?
- equivalence class of Bayesian networks

Representing distributions

give a number of random variables $X_1, \ldots, X_n$

how to represent $P(X_1, \ldots, X_n)$

- number of parameters exponential in n (curse of dimensionality)
- need to leverage some structure in **P**

Independence & representation

for **discrete** domains $Val(X_i) = \{1, \ldots, D\} \quad \forall i$

- representation of $P(\mathbf{X} = x_1, \ldots, x_n) = \theta_{i_1, \ldots, i_n}$
  - exponential in n: $\mathcal{O}(D^n\}$

Independence & representation

for **discrete** domains $Val(X_i) = \{1, \ldots, D\} \quad \forall i$

- representation of $P(\mathbf{X} = x_1, \ldots, x_n) = \theta_{i_1, \ldots, i_n}$
  - exponential in n: $\mathcal{O}(D^n\}$

assuming **independence** $X_i \perp X_j \quad \forall i, j$

- linear-sized representation:

$$P(\mathbf{X} = x_1^d, \ldots, x_n^d) = \prod_i P(X_i = x_i^d) = \prod_i \theta_{i,d}$$

a particular assignment (d) in discrete domain

Independence & representation

for **discrete** domains $Val(X_i) = \{1, \ldots, D\} \quad \forall i$

- representation of $P(\mathbf{X} = x_1, \ldots, x_n) = \theta_{i_1, \ldots, i_n}$
  - exponential in n: $\mathcal{O}(D^n\}$

assuming **independence** $X_i \perp X_j \quad \forall i, j$

- linear-sized representation:

$$P(\mathbf{X} = x_1^d, \ldots, x_n^d) = \prod_i P(X_i = x_i^d) = \prod_i \theta_{i,d}$$

a particular assignment (d) in discrete domain

independence assumption is too restrictive

Using the chain rule

- pick an *ordering* of the variables

$$P(\mathbf{X}) = P(X_1)P(X_2 \mid X_1) \ldots P(X_n \mid X_1, \ldots, X_{n-1})$$

Using the chain rule

- pick an *ordering* of the variables

$$P(\mathbf{X}) = P(X_1)P(X_2 \mid X_1)\dots P(X_n \mid X_1, \dots, X_{n-1})$$

- parameterize each term
- does it compress the representation?

  - original #params $D^n - 1$

Using the chain rule

- pick an *ordering* of the variables

$$P(\mathbf{X}) = P(X_1)P(X_2 \mid X_1)\ldots P(X_n \mid X_1, \ldots, X_{n-1})$$

- parameterize each term
- does it compress the representation?
  - original #params $D^n - 1$
  - new #params $\underbrace{(D-1)}_{P(X_1)} + \underbrace{(D^2 - D)}_{P(X_2 \mid X_1)} + \ldots + \underbrace{(D^n - D^{n-1})}_{P(X_n \mid X_1, \ldots, X_{n-1})} = D^n - 1$

Using the chain rule

$$P(\mathbf{X}) = P(X_1)P(X_2 \mid X_1)\ldots P(X_n \mid X_1, \ldots, X_{n-1})$$

simplify the conditionals

- flexible compression of P

Using the chain rule

$$P(\mathbf{X}) = P(X_1)P(X_2 \mid X_1)\ldots P(X_n \mid X_1, \ldots, X_{n-1})$$

simplify the conditionals

- flexible compression of P

A Bayesian network!

Chain rule: simplification

$$P(\mathbf{X}) = P(X_1)P(X_2 \mid X_1)P(X_3 \mid X_1, X_2) \dots P(X_n \mid X_1, \dots, X_{n-1})$$

an **extreme** form of simplification

$$P(\mathbf{X}) = P(X_1)P(X_2 \mid X_1)P(X_3 \mid X_1) \dots P(X_n \mid X_1)$$

Chain rule: simplification

$$P(\mathbf{X}) = P(X_1)P(X_2 \mid X_1)P(X_3 \mid X_1, X_2) \dots P(X_n \mid X_1, \dots, X_{n-1})$$

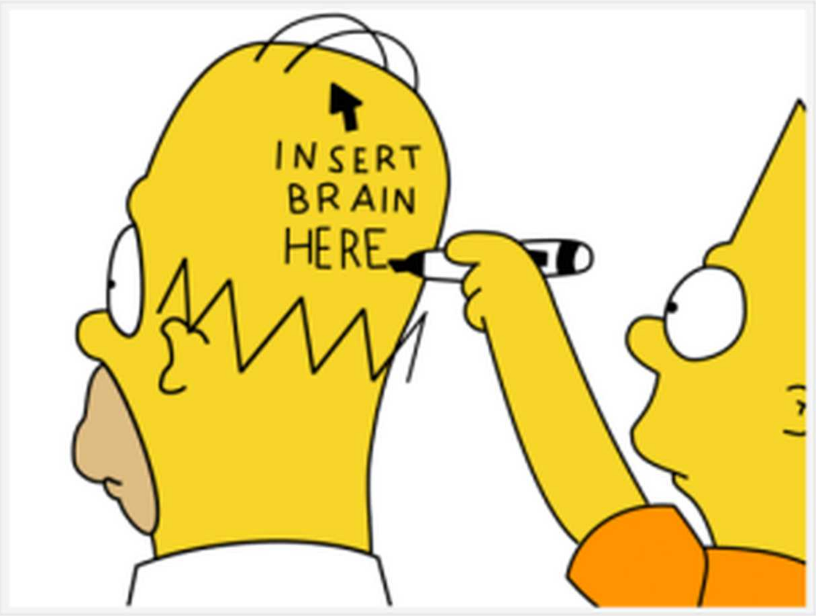
↓ an **extreme** form of simplification

$$P(\mathbf{X}) = P(X_1)P(X_2 \mid X_1)P(X_3 \mid X_1) \dots P(X_n \mid X_1)$$

\# params $(D-1) + (n-1)(D^2 - D)$

$\mathcal{O}(nD^2)$ instead of $\mathcal{O}(D^n)$

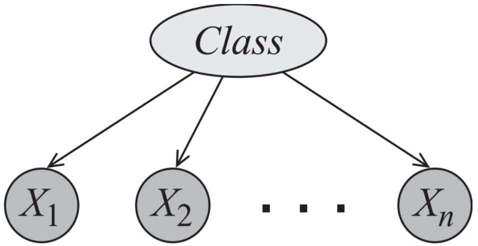
Idiot Bayes



...or naive Bayes

$$P(class, \mathbf{X}) = P(class)P(X_2 \mid class)P(X_3 \mid class)\dots P(X_n \mid class)$$

independence assumption: $X_i \perp \mathbf{X}_{-i} \mid class$



for classification (use Bayes rule)

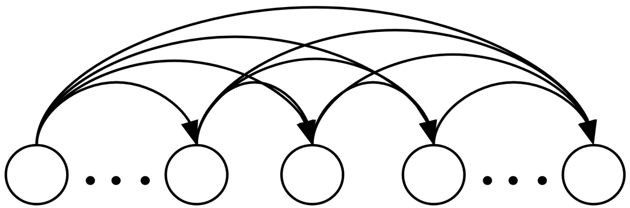
$$P(class \mid \mathbf{X}) \propto P(class)P(X_2 \mid class)P(X_3 \mid class)\dots P(X_n \mid class)$$

Example: medical diagnosis (what if two symptoms are correlated?)

Simplifying the chain rule: general case

simplify the full conditionals:

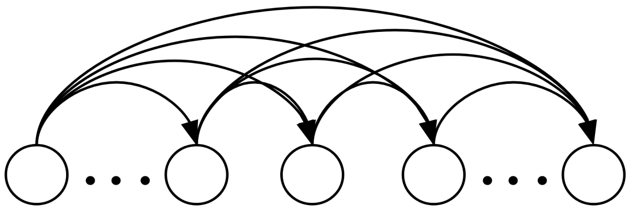
$$P(\mathbf{X}) = P(X_1)P(X_2 \mid X_1)\ldots P(X_n \mid X_1, \ldots, X_{n-1})$$



Simplifying the chain rule: general case

simplify the full conditionals:

$$P(\mathbf{X}) = P(X_1)P(X_2 \mid X_1)\ldots P(X_n \mid X_1, \ldots, X_{n-1})$$

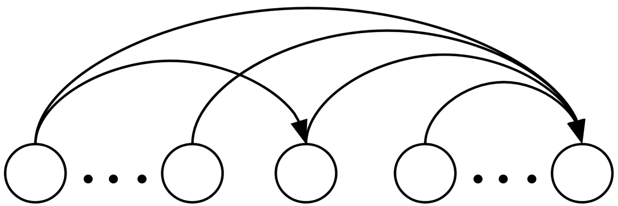


Bayesian network

represent it using a

Directed Acyclic Graph **(DAG)**

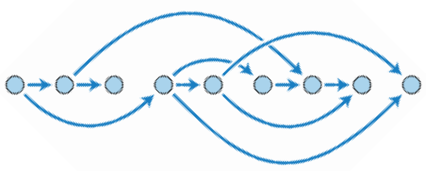
$$P(\mathbf{X}) = \prod_i P(X_i \mid Pa_{X_i})$$

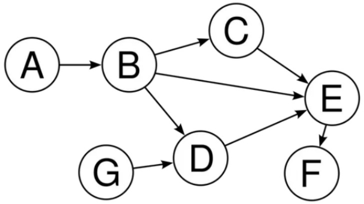


a topological ordering

DAG: identification

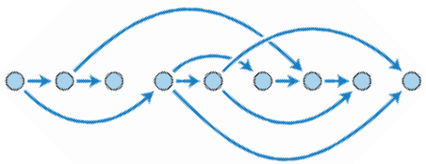
- identifying a DAG
  - has a topological ordering?
  - no directed path from a node to itself?





DAG: identification

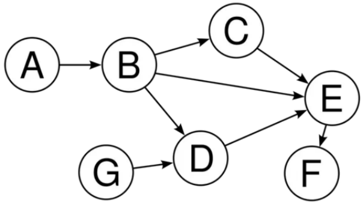
- identifying a DAG
  - has a topological ordering?
  - no directed path from a node to itself?



Example:

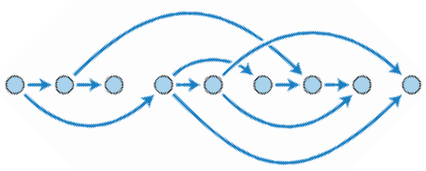
is this a DAG?

a topological ordering: *G,A,B,D,C,E,F*



DAG: identification

- identifying a DAG
  - has a topological ordering?
  - no directed path from a node to itself?

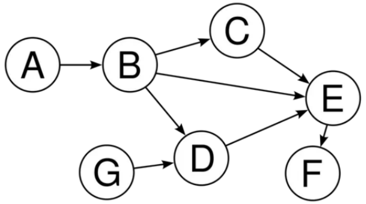


Example:

is this a DAG?

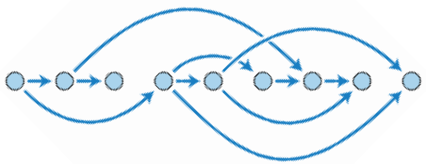
a topological ordering: *G,A,B,D,C,E,F*

A,B,C,G,D,E,F



DAG: identification

- identifying a DAG
  - has a topological ordering?
  - no directed path from a node to itself?

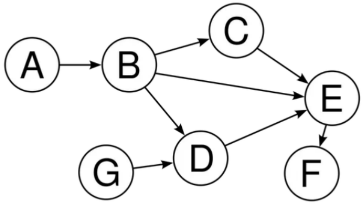


Example:

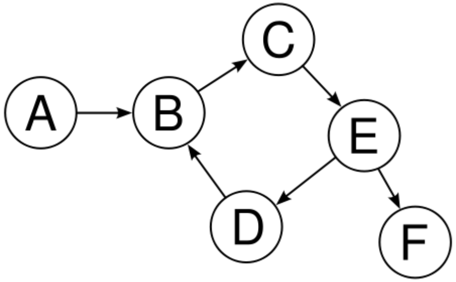
is this a DAG?

a topological ordering: *G,A,B,D,C,E,F*

A,B,C,G,D,E,F

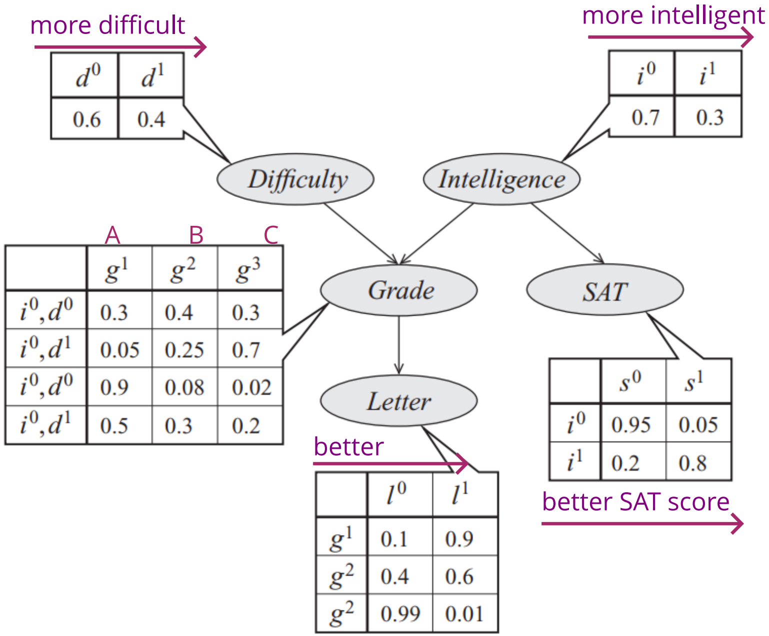


how about this?



Bayesian network (BN): running example

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



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

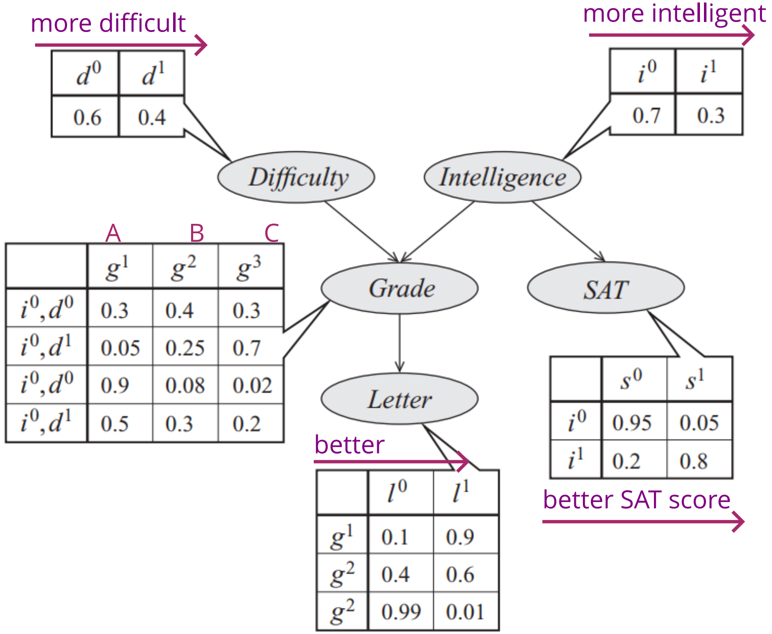
	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^2	0.99	0.01

Bayesian network (BN): running example

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

Conditional Probability Table (CPT)



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

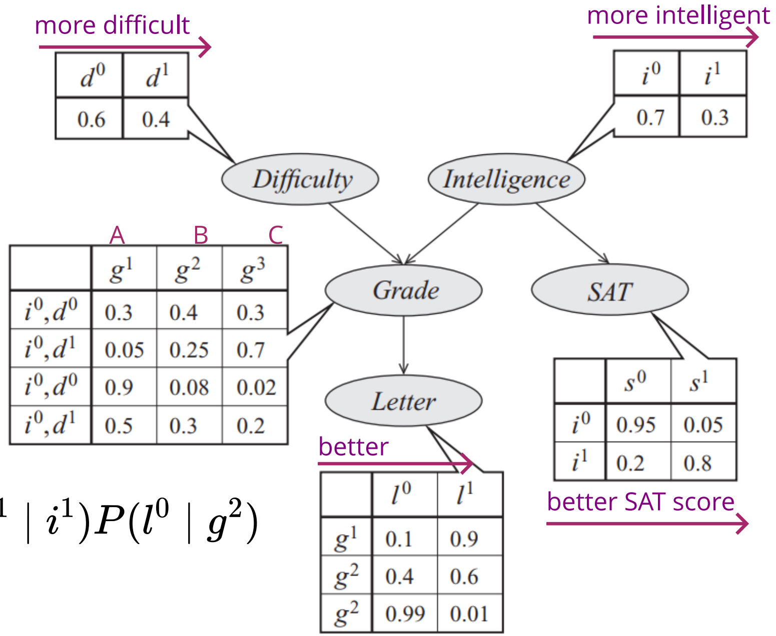
	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^2	0.99	0.01

Bayesian network (BN): running example

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

Conditional Probability Table (CPT)



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

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

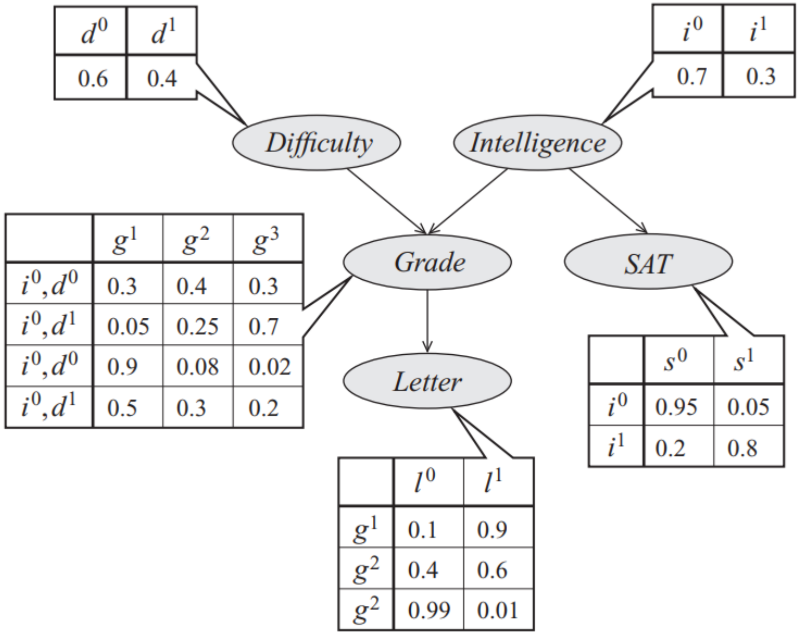
$$= .7 \times .6 \times .08 \times .8 \times .4 \approx .01$$

Intuition for reasoning in a BN

answering probabilistic queries

$$P(\mathbf{Y} = \mathbf{y} \mid \mathbf{E} = e) \quad ?$$

evidence



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

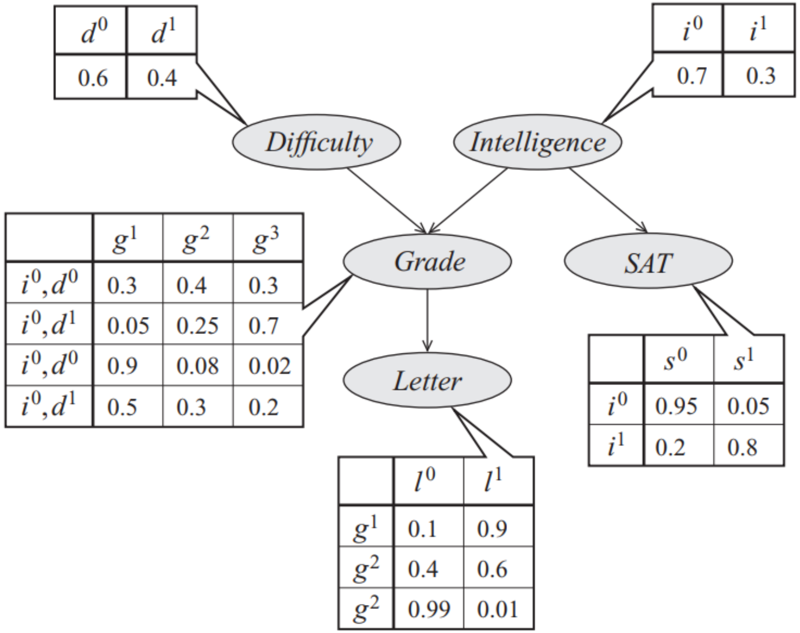
Intuition for reasoning in a BN

answering probabilistic queries

$$P(\mathbf{Y} = \mathbf{y} \mid \mathbf{E} = e) \quad ?$$

evidence

$$P(L = l^1 \mid S = s^1) = \frac{P(L=l^1, S=s^1)}{P(S=s^1)}$$



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

Intuition for reasoning in a BN

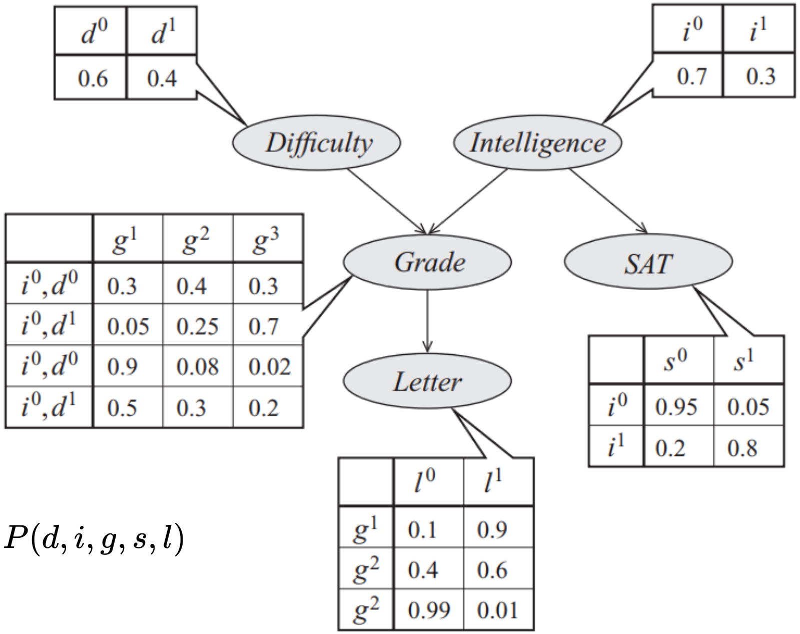
answering probabilistic queries

$$P(\mathbf{Y} = \mathbf{y} \mid \mathbf{E} = e) \quad ?$$

evidence

$$P(L = l^1 \mid S = s^1) = \frac{P(L=l^1, S=s^1)}{P(S=s^1)}$$

$$P(S = s^1) = \sum_{d,i,g,l} P(d, i, g, s, l)$$



an **inference** problem

- how to calculate? ... later

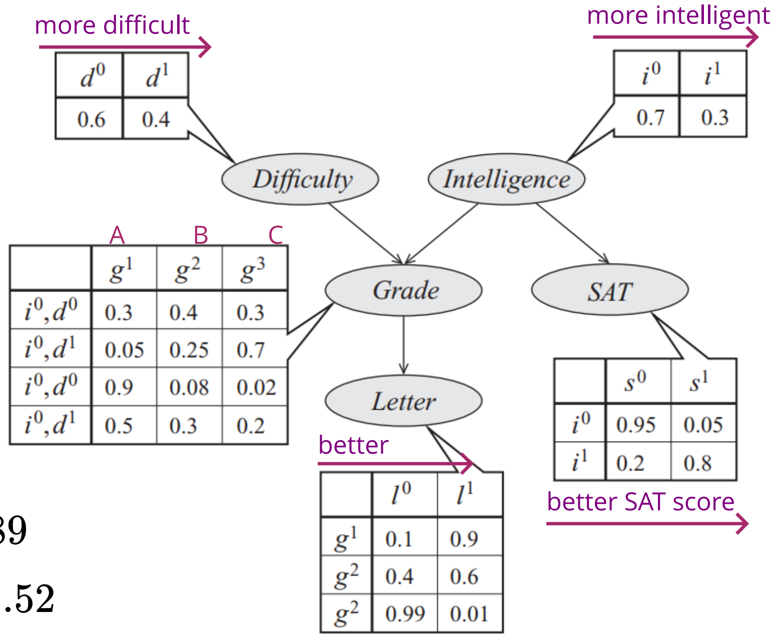
Intuition for reasoning in a BN

causal reasoning (top-down)

- marginal prior
  - of getting a good letter

$$P(l^1) \approx .50$$

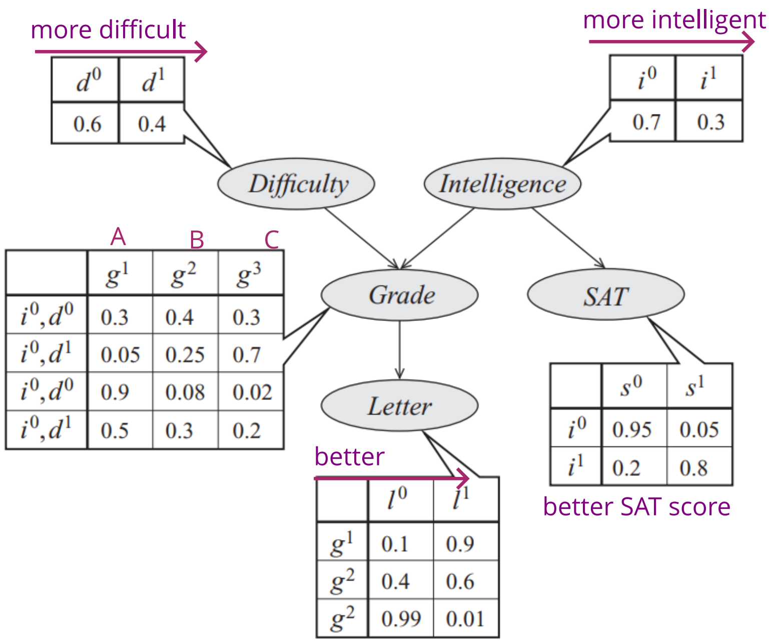
- marginal posterior
  - given low intelligence $P(l^1 \mid i^0) \approx .389$
  - ... and an easy exam $P(l^1 \mid i^0, d^0) \approx .52$



Intuition for reasoning in a BN

evidential reasoning (bottom-up)

- (marginal) prior
  - of a high intelligence $P(i^1) \approx .30$
- (marginal) posterior
  - given a bad letter $P(i^1 \mid l^0) \approx .14$
  - ... and a bad grade $P(i^1 \mid l^0, g^3) \approx .08$

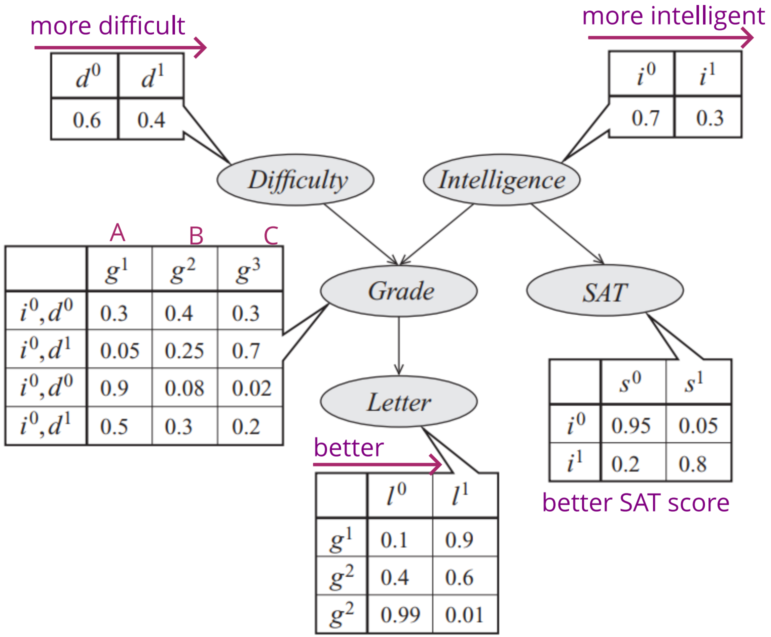


Intuition for Reasoning in BN

Explaining away (v-structure)

- prior
  - of a high intelligence $P(i^1) \approx .30$
- posterior
  - given a bad letter $P(i^1 \mid l^0) \approx .14$
  - ... and a bad grade $P(i^1 \mid l^0, g^3) \approx .08$
  - a difficult exam explains away the grade

$$P(i^1 \mid l^0, g^3, d^1) \approx .11$$



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

DAG: semantics

associating P with a DAG:

- ***factorization*** of the joint probability:

$$P(\mathbf{X}) = \prod_i P(X_i \mid Pa_{X_i})$$

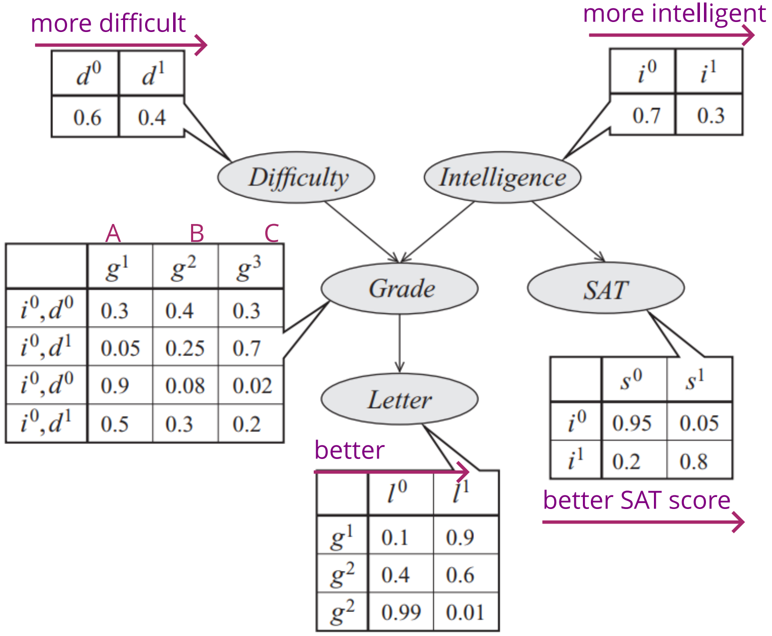
- ***conditional independencies*** in P from the DAG

Bayesian networks: factorization

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

In general

$$P(\mathbf{X}) = \prod_i P(X_i \mid Pa_{X_i})$$



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

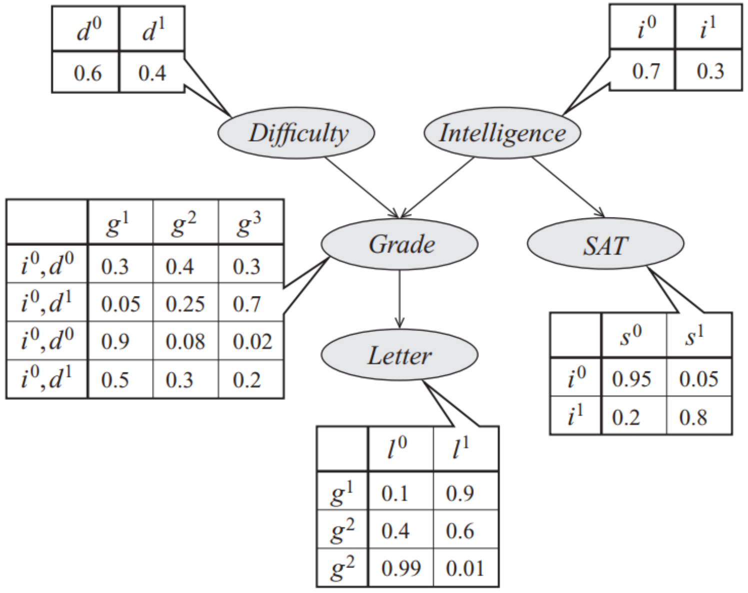
	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^2	0.99	0.01

Bayesian networks: conditioinal independencies

- quality of the letter (L) only depends on the grade (G)

$L \perp D, I, S \mid G$ ✔



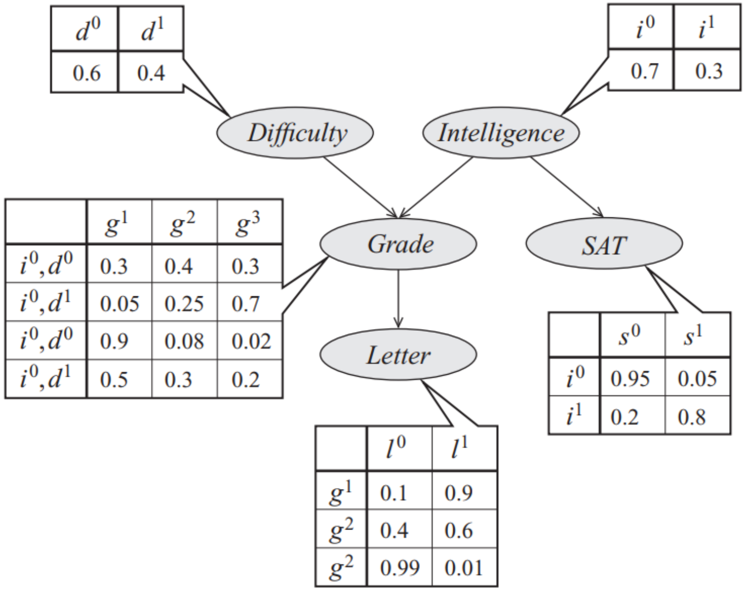
Bayesian networks: conditioinal independencies

- quality of the letter (L) only depends on the grade (G)

$$L \perp D, I, S \mid G$$ ✔

- How about the following assertions?

$$D \perp S \quad ?$$



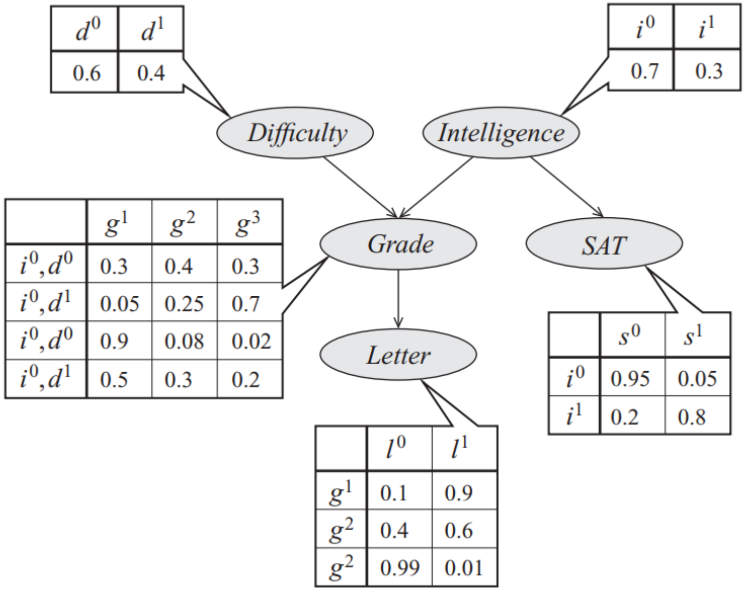
Bayesian networks: conditioinal independencies

- quality of the letter (L) only depends on the grade (G)

$$L \perp D, I, S \mid G$$ ✔

- How about the following assertions?

$$D \perp S \quad ?$$ ✔



Bayesian networks: conditioinal independencies

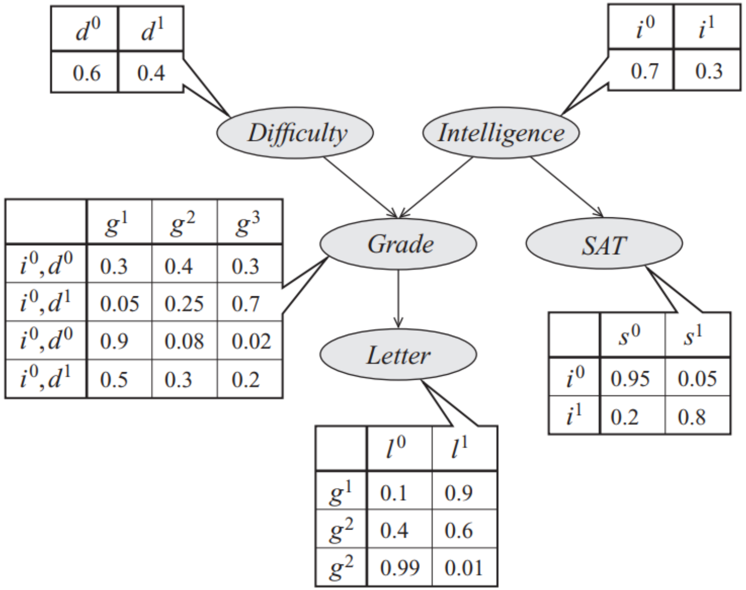
- quality of the letter (L) only depends on the grade (G)

$$L \perp D, I, S \mid G$$ ✔

- How about the following assertions?

$$D \perp S \quad ?$$ ✔

$$D \perp S \mid I \quad ?$$



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

Bayesian networks: conditioinal independencies

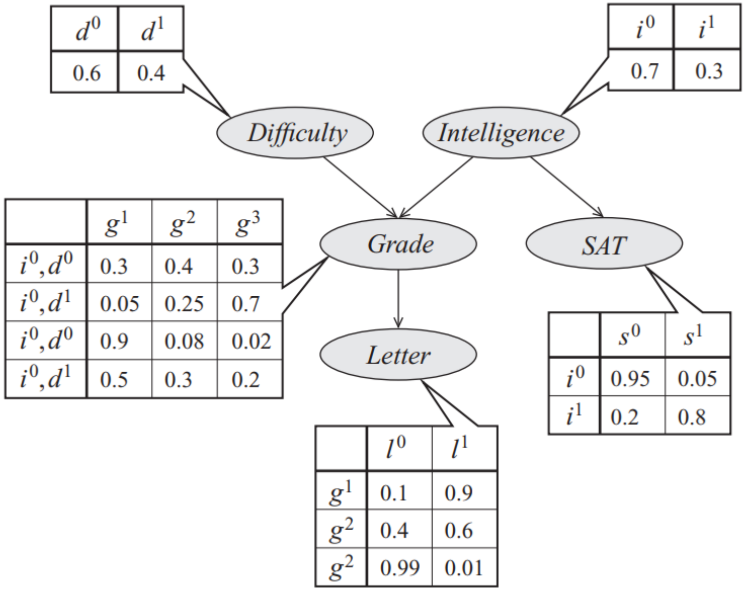
- quality of the letter (L) only depends on the grade (G)

$$L \perp D, I, S \mid G$$ ✔

- How about the following assertions?

$$D \perp S \quad ?$$ ✔

$$D \perp S \mid I \quad ?$$ ✔



Bayesian networks: conditioinal independencies

- quality of the letter (L) only depends on the grade (G)

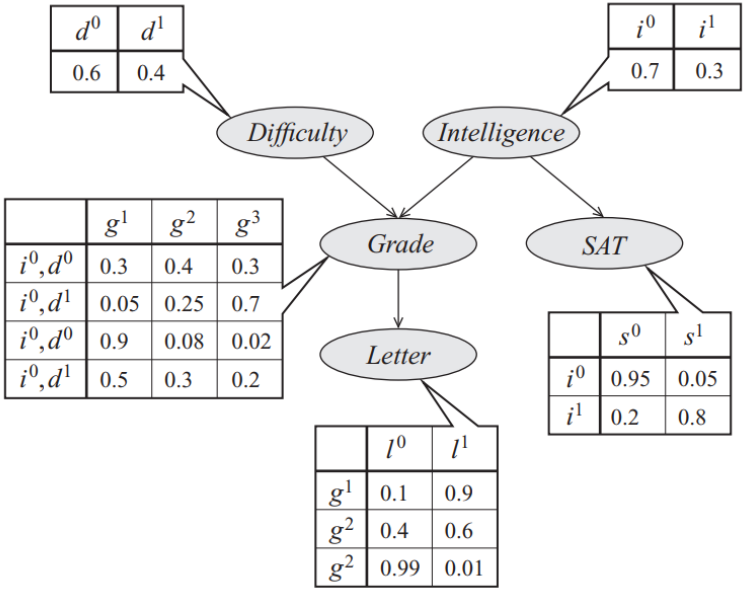
$L \perp D, I, S \mid G$ ✓

- How about the following assertions?

$D \perp S$? ✓

$D \perp S \mid I$? ✓

$D \perp S \mid L$?



Bayesian networks: conditioinal independencies

- quality of the letter (L) only depends on the grade (G)

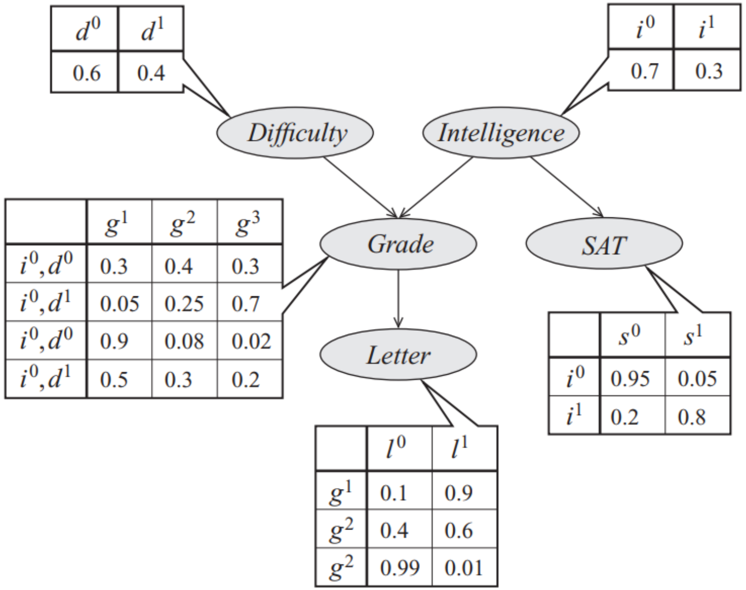
$L \perp D, I, S \mid G$ ✓

- How about the following assertions?

$D \perp S$? ✓

$D \perp S \mid I$? ✓

$D \perp S \mid L$? ✗ why?



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

Bayesian networks: conditioinal independencies

- quality of the letter (L) only depends on the grade (G)

$L \perp D, I, S \mid G$ ✓

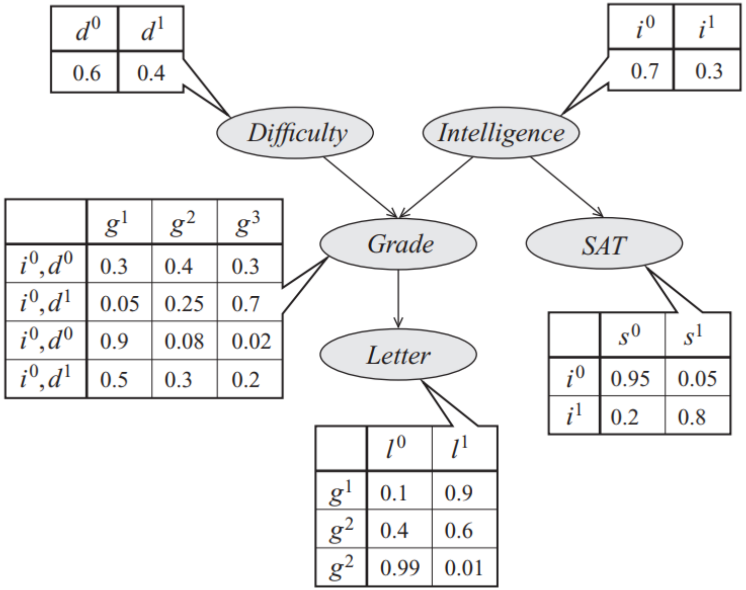
- How about the following assertions?

$D \perp S$? ✓

$D \perp S \mid I$? ✓

$D \perp S \mid L$? ✗ why?

- read from the graph?



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

Conditional independencies (CI): notation

1. set of all CIs of the **distribution** P $\mathcal{I}(P)$

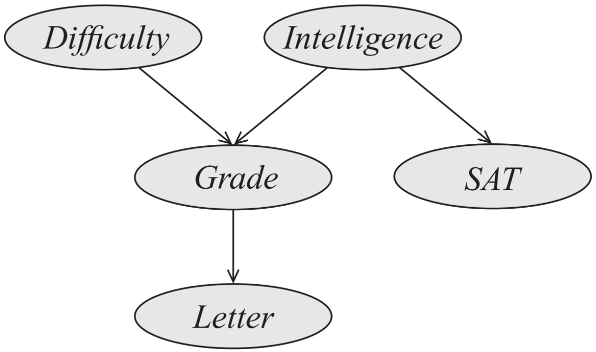
2. set of **local** CIs from the **graph** (DAG) $\mathcal{I}_{\ell}(\mathcal{G})$

3. set of all (**global**) CIs from the **graph** $\mathcal{I}(\mathcal{G})$

Local conditional independencies (CIs)

for any node X_i $\quad X_i \perp NonDescendents_{X_i} \mid Parents_{X_i}$

$$\mathcal{I}_\ell(\mathcal{G}) = \{ \begin{array}{l} D \perp I, S \\ I \perp D \\ G \perp S \mid I, D \\ S \perp G, L, D \mid I \\ L \perp D, I, S \mid G \end{array} \}$$



graph $\mathcal{G}$

Local CIs from factorization

use the **factorized form** $P(\mathbf{X}) = \prod_i P(X_i \mid Pa_{X_i})$

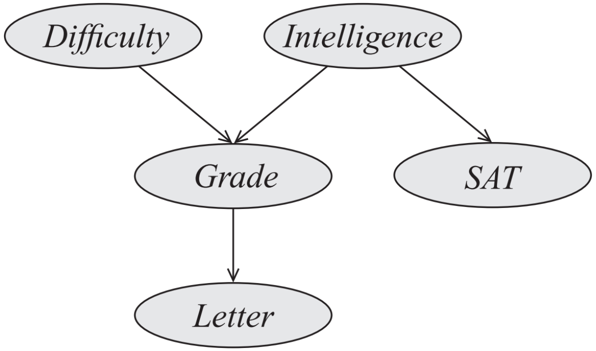
to show $\forall X_i$

$$P(X_i, NonDesc_{X_i} \mid Pa_{X_i}) = P(X_i \mid Pa_{X_i}) P(NonDesc_{X_i} \mid Pa_{X_i})$$

which means $X_i \perp NonDesc_{X_i} \mid Pa_{X_i}$

Local CIs from factorization: example

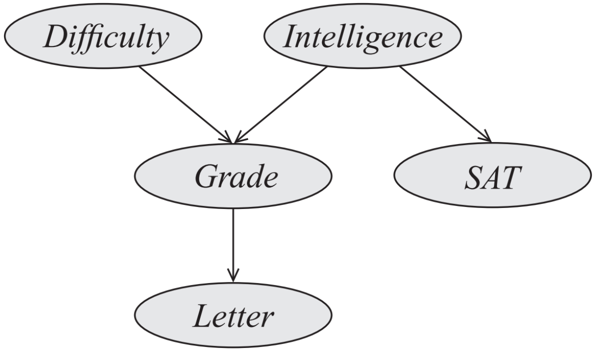
$S \perp G \mid I$ given $P(D, I, G, S, L) = P(D)P(I)P(G \mid D, I)P(S \mid I)P(L \mid G)$



Local CIs from factorization: example

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

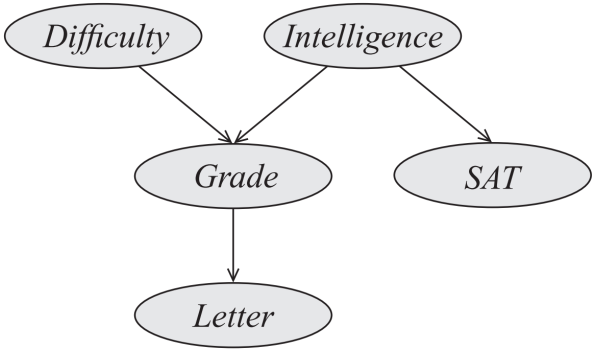
$$P(G, S \mid I) = \frac{\sum_{d,l} P(D,I,G,S,L)}{\sum_{d,g,s,l} P(D,I,G,S,L)} =$$



Local CIs from factorization: example

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

$$P(G, S \mid I) = \frac{\sum_{d,l} P(D,I,G,S,L)}{\sum_{d,g,s,l} P(D,I,G,S,L)} = \frac{\sum_{d,l} P(D)P(I)P(G|D,I)P(S|I)P(L|G)}{\sum_{d,g,s,l} P(D)P(I)P(G|D,I)P(S|I)P(L|G)} =$$

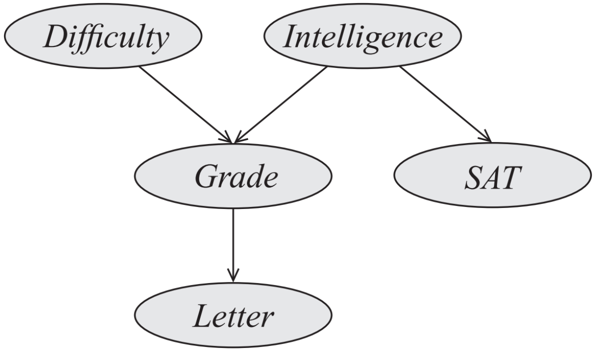


Local CIs from factorization: example

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

$$P(G,S \mid I) = \frac{\sum_{d,l} P(D,I,G,S,L)}{\sum_{d,g,s,l} P(D,I,G,S,L)} = \frac{\sum_{d,l} P(D)P(I)P(G|D,I)P(S|I)P(L|G)}{\sum_{d,g,s,l} P(D)P(I)P(G|D,I)P(S|I)P(L|G)} =$$

$$\frac{P(I)P(S|I)\sum_{d,l} P(D)P(G|D,I)P(L|G)}{P(I)\sum_{d,g,s,l} P(D)P(G|D,I)P(S|I)P(L|G)} =$$



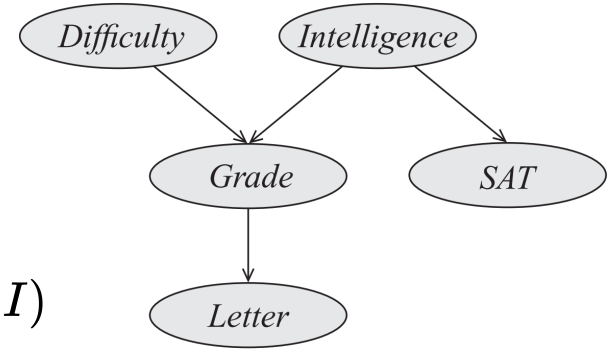
Local CIs from factorization: example

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

$$P(G, S \mid I) = \frac{\sum_{d,l} P(D,I,G,S,L)}{\sum_{d,g,s,l} P(D,I,G,S,L)} = \frac{\sum_{d,l} P(D)P(I)P(G|D,I)P(S|I)P(L|G)}{\sum_{d,g,s,l} P(D)P(I)P(G|D,I)P(S|I)P(L|G)} =$$

$$\frac{P(I)P(S|I)\sum_{d,l} P(D)P(G|D,I)P(L|G)}{P(I)\sum_{d,g,s,l} P(D)P(G|D,I)P(S|I)P(L|G)} =$$

$$\frac{P(S|I)\sum_{d,l} P(D)P(G|D,I)P(L|G)}{1} = P(S \mid I)P(G \mid I)$$



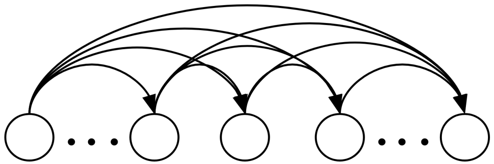
Factorization from local CIs

from **local CIs** $\mathcal{I}_\ell(\mathcal{G}) = \{X_i \perp NonDesc_{X_i} \mid Pa_{X_i} \mid i\}$

find a topological ordering *(parents before children):* $X_{i_1}, \ldots, X_{i_n}$

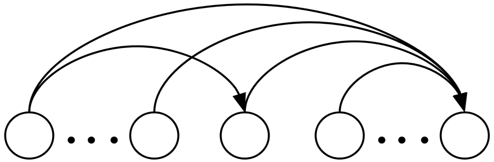
use the chain rule

$$P(\mathbf{X}) = P(X_{i_1}) \prod_{j=2}^{n} P(X_{i_j} \mid X_{i_1}, \ldots, X_{i_{j-1}})$$



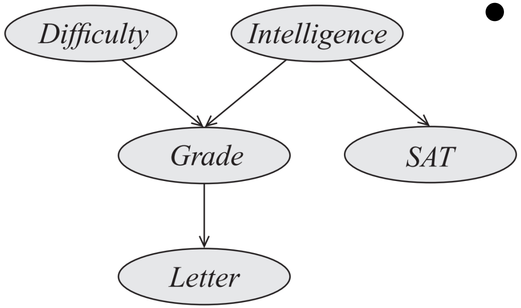
simplify using local CIs

$$P(\mathbf{X}) = P(X_{i_1}) \prod_{j=2}^{n} P(X_{i_j} \mid Pa_{X_{i_j}})$$



Factorization from local CIs: example

- local CIs $\mathcal{I}_\ell(\mathcal{G}) = \{\ (D \perp I, S),\ (I \perp D)\ , (G \perp S \mid I),$
  $(S \perp G, L, D \mid I),\ \ (L \perp D, I, S \mid G)\ \}$



- a topological ordering: D, I, G, L, S
  - use the chain rule

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

- simplify using $\mathcal{I}_\ell(\mathcal{G})$

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

Factorization $\Leftrightarrow$ local CIs

$P(\mathbf{X}) = \prod_i P(X_i \mid Pa_{X_i}^{\mathcal{G}})$ $\Leftrightarrow$ $\mathcal{I}_\ell(\mathcal{G})$ holds in P

P factorizes according to $\mathcal{G}$

Factorization $\Leftrightarrow$ **local CIs**

$P(\mathbf{X}) = \prod_i P(X_i \mid Pa^{\mathcal{G}}_{X_i})$ $\Leftrightarrow$ $\mathcal{I}_\ell(\mathcal{G})$ holds in P

P factorizes according to $\mathcal{G}$

$\mathcal{I}_\ell(\mathcal{G}) \subseteq \mathcal{I}(P)$

Factorization $\Leftrightarrow$ local CIs

$P(\mathbf{X}) = \prod_i P(X_i \mid Pa^{\mathcal{G}}_{X_i})$ $\Leftrightarrow$ $\mathcal{I}_\ell(\mathcal{G})$ holds in P

P factorizes according to $\mathcal{G}$

$\mathcal{I}_\ell(\mathcal{G}) \subseteq \mathcal{I}(P)$

$\mathcal{G}$ is an **I-map** for P

it does not mislead us
about independencies in P

Perfect map (P-map)

which graph G to use for P?

Perfect MAP: $\mathcal{I}(\mathcal{G}) = \mathcal{I}(P)$

P may not have a P-map in the form of BN

Perfect map (P-map)

which graph G to use for P?

Perfect MAP: $\mathcal{I}(\mathcal{G}) = \mathcal{I}(P)$

P may not have a P-map in the form of BN

Example:

$$p(x,y,z) = \begin{cases} 1/12, & \text{if } x \otimes y \otimes z = 0 \\ 1/6, & \text{if } x \otimes y \otimes z = 1 \end{cases}$$

$$(X \perp Y), (Y \perp Z), (X \perp Z) \in \mathcal{I}(P)$$

$$(X \perp Y \mid Z), (Y \perp Z \mid Z), (X \perp Z \mid Y) \notin \mathcal{I}(P)$$

Perfect map (P-map)

which graph G to use for P?

Perfect MAP: $\mathcal{I}(\mathcal{G}) = \mathcal{I}(P)$

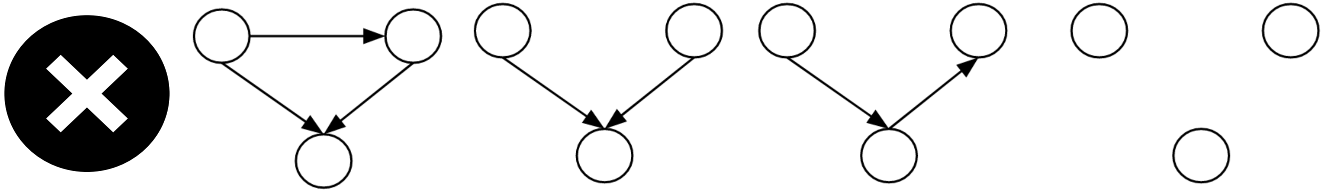
P may not have a P-map in the form of BN

Example:

$$p(x,y,z) = \begin{cases} 1/12, & \text{if } x \otimes y \otimes z = 0 \\ 1/6, & \text{if } x \otimes y \otimes z = 1 \end{cases}$$

$(X \perp Y), (Y \perp Z), (X \perp Z) \in \mathcal{I}(P)$

$(X \perp Y \mid Z), (Y \perp Z \mid Z), (X \perp Z \mid Y) \notin \mathcal{I}(P)$



Summary so far

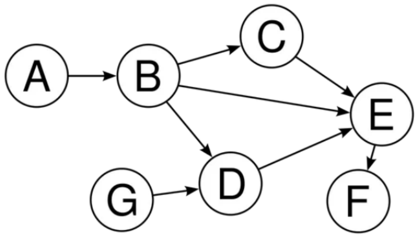
- simplification of the chain rule $P(\mathbf{X}) = \prod_i P(X_i \mid Pa_{X_i})$

- Bayes-net represented using a DAG
- naive Bayes
- local conditional independencies $\mathcal{I} = \{X_i \perp NonDesc_{X_i} \mid Pa_{X_i} \mid i\}$
  - hold in a Bayes-net
  - imply a Bayes-net
- Note: motivation is not just compressed representation, but faster inference and learning as well

Global CIs *from the graph*

for any subset of vars **X, Y** and **Z**, we can ask $\mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z}?$

global CI: the set of all such CIs

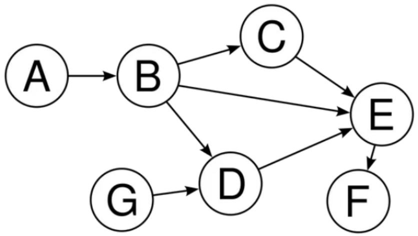


Global CIs *from the graph*

for any subset of vars **X, Y** and **Z,** we can ask $\mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z}$?

global CI: the set of all such CIs

factorized form of P $\Rightarrow$ **global** CIs $\mathcal{I}_\ell(\mathcal{G}) \subseteq \mathcal{I}(\mathcal{G}) \subseteq \mathcal{I}(P)$



Global CIs *from the graph*

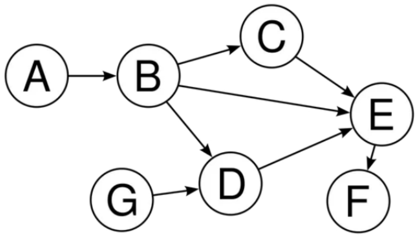
for any subset of vars **X, Y** and **Z**, we can ask $\mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z}?$

global CI: the set of all such CIs

factorized form of P $\Rightarrow$ **global** CIs $\mathcal{I}_\ell(\mathcal{G}) \subseteq \mathcal{I}(\mathcal{G}) \subseteq \mathcal{I}(P)$

Example:

$C \perp D \mid B, F \quad ?$

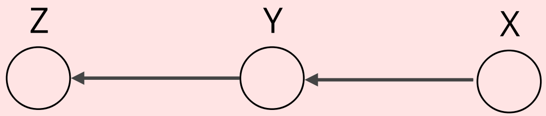


algorithm: directed separation (**D-separation**)

Three canonical settings

for three random variables

1. causal / evidence trail

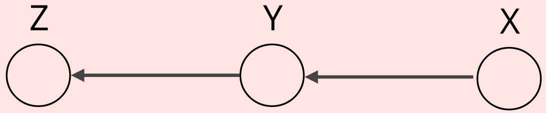


$$P(X, Y, Z) = P(X)P(Y|X)P(Z \mid Y)$$

Three canonical settings

for three random variables

1. causal / evidence trail



$$P(X,Y,Z) = P(X)P(Y|X)P(Z \mid Y)$$

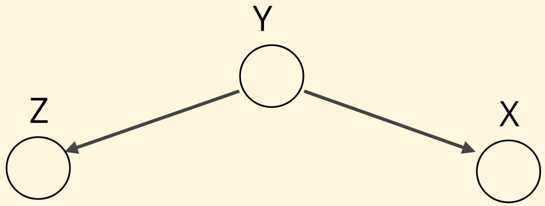
~~marginal independence:~~ $P(X,Z) \neq P(X)P(Z)$

conditional Independence:

$$P(Z \mid X,Y) = \frac{P(X,Y,Z)}{P(X,Y)} = \frac{P(X)P(Y|X)P(Z|Y)}{P(X)P(Y|X)} = P(Z \mid Y)$$

Three canonical settings

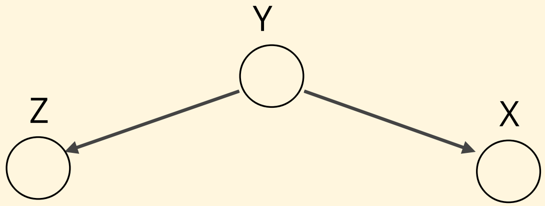
2. common cause



$$P(X, Y, Z) = P(Y)P(X \mid Y)P(Z \mid Y)$$

Three canonical settings

2. common cause



$$P(X, Y, Z) = P(Y)P(X \mid Y)P(Z \mid Y)$$

~~marginal independence:~~ $P(X, Z) \neq P(X)P(Z)$

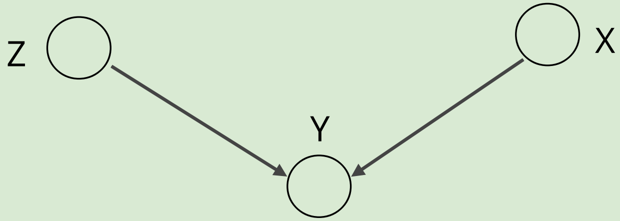
conditional independence:

$$P(X, Z \mid Y) = \frac{P(X,Y,Z)}{P(Y)} = P(X \mid Y)P(Z \mid Y)$$

Three canonical settings

3. common effect

a.k.a. *collider*, *v-structure*

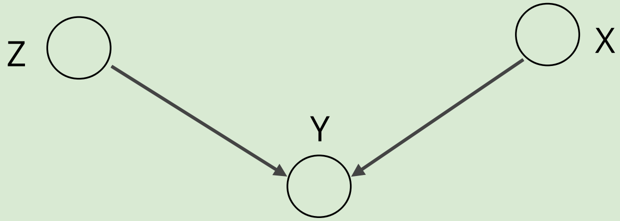


$$P(X, Y, Z) = P(X)P(Z)P(Y \mid X, Z)$$

Three canonical settings

3. common effect

a.k.a. *collider, v-structure*



$$P(X,Y,Z) = P(X)P(Z)P(Y \mid X,Z)$$

marginal independence:

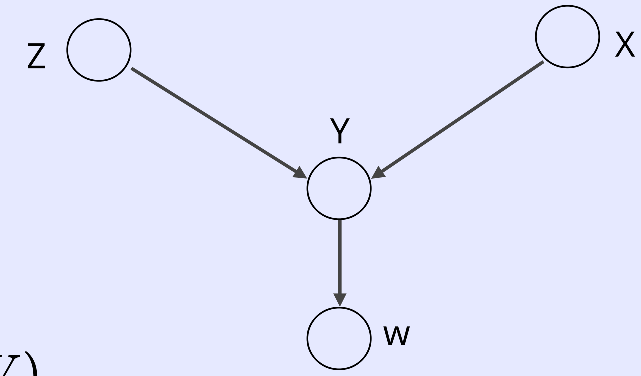
$$P(X,Z) = \sum_Y P(X,Y,Z) = P(X)P(Z)\sum_Y P(Y \mid X,Z) = P(X)P(Z)$$

~~conditional independence:~~

$$P(X,Z \mid Y) = \frac{P(X,Y,Z)}{P(Y)} \neq P(X \mid Y)P(Z \mid Y)$$

Three canonical settings

3. common effect



~~**conditional** Independence:~~

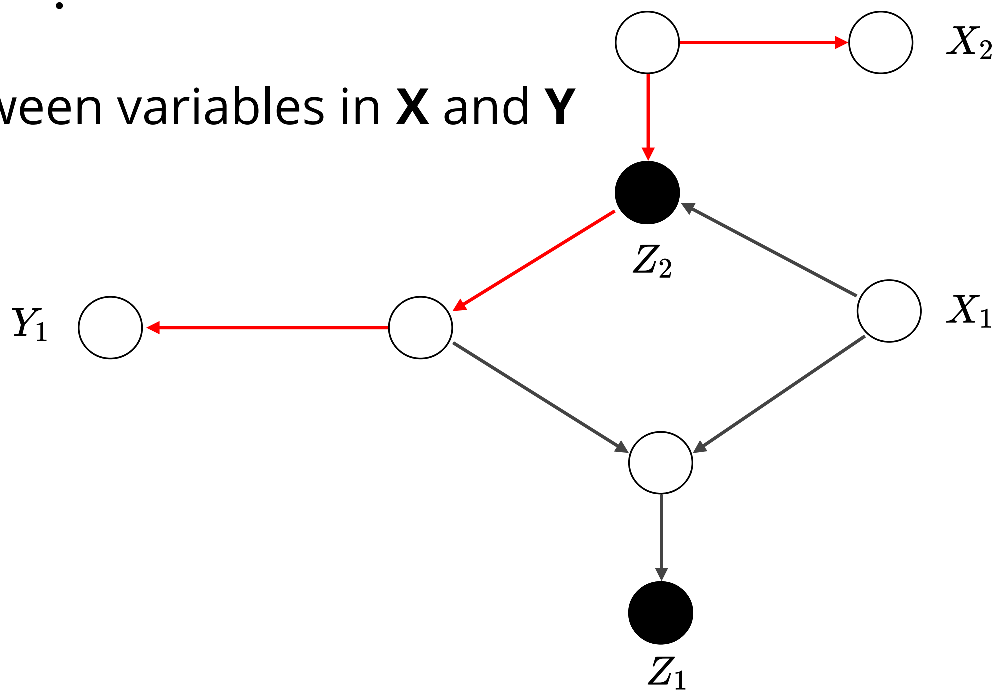
$$P(X, Z \mid W) \neq P(X \mid W)P(Z \mid W)$$

even observing a descendant of Y makes X, Z dependent

Putting the three cases together

$$X_1, X_2 \perp Y_1 \mid Z_1, Z_2 \quad ?$$

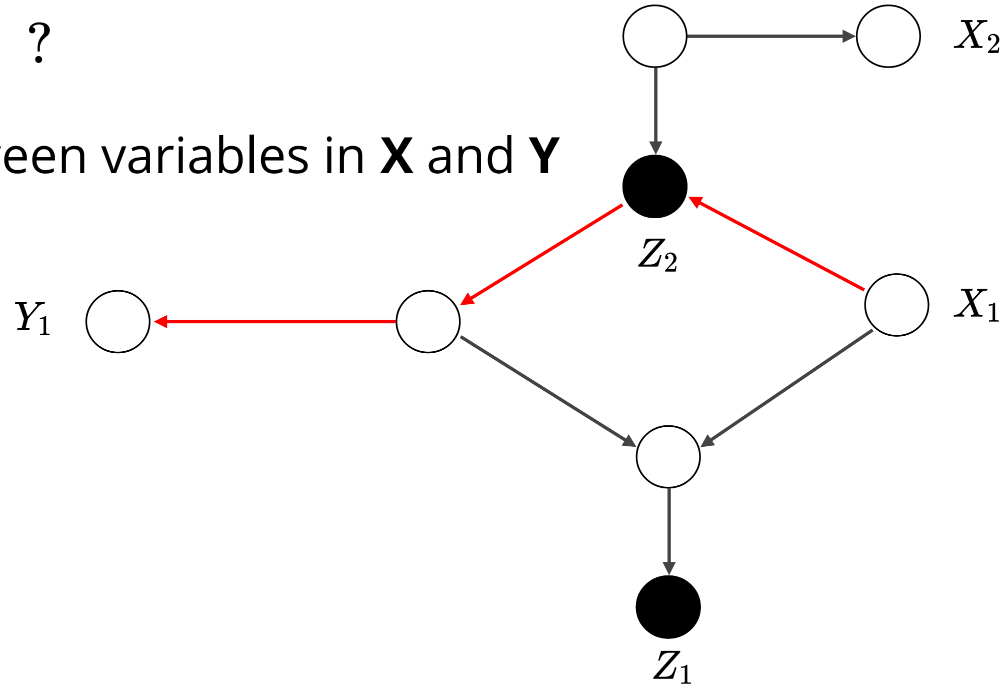
consider all paths between variables in **X** and **Y**



Putting the three cases together

$$X_1, X_2 \perp Y_1 \mid Z_1, Z_2 \quad ?$$

consider all paths between variables in **X** and **Y**



so far $\mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z}$

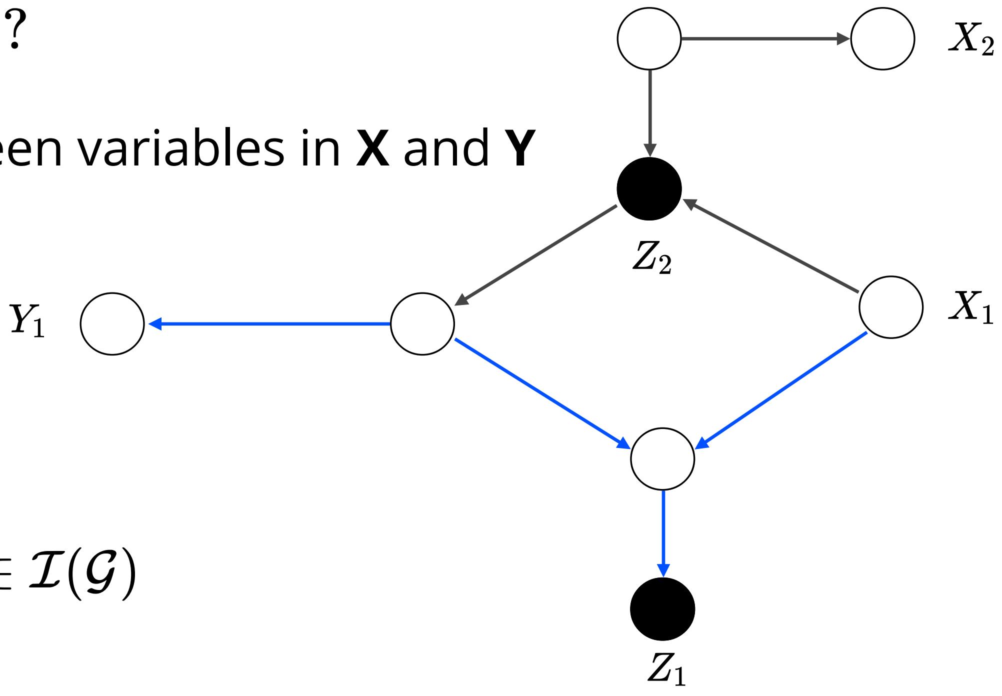
Putting the three cases together

$X_1, X_2 \not\perp Y_1 \mid Z_1, Z_2 \quad ?$

consider all paths between variables in **X** and **Y**

had we not observed Z_1

$$(X_1, X_2 \perp Y_1 \mid Z_2) \in \mathcal{I}(\mathcal{G})$$



D-seperation

(a.k.a. **Bayes-Ball** algorithm)

$\mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z} \quad ?$

See whether at least one ball from **X** reaches **Y**

Z is shaded

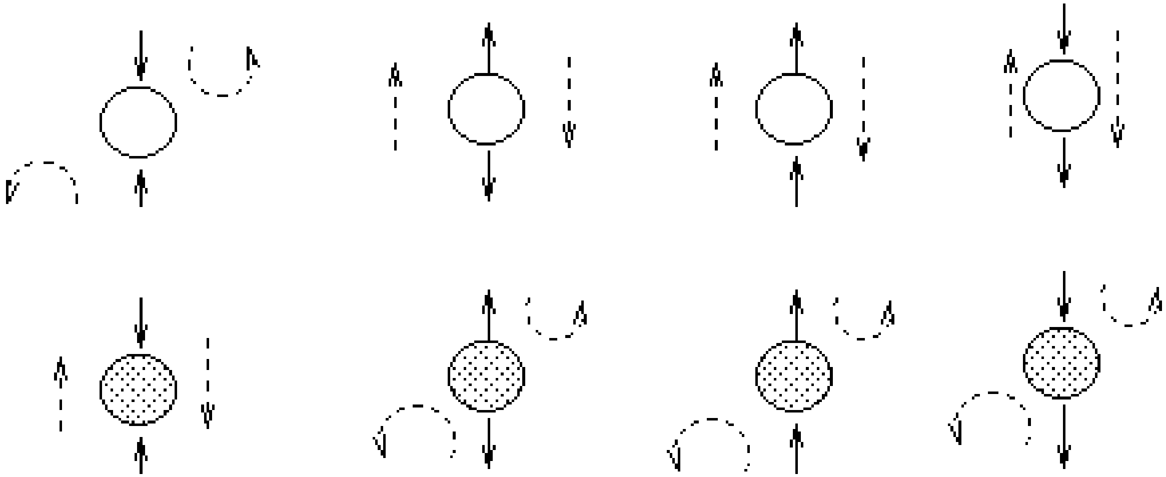


image from:https://www.cs.ubc.ca/~murphyk/Bayes/bnintro.html

D-separation: algorithm

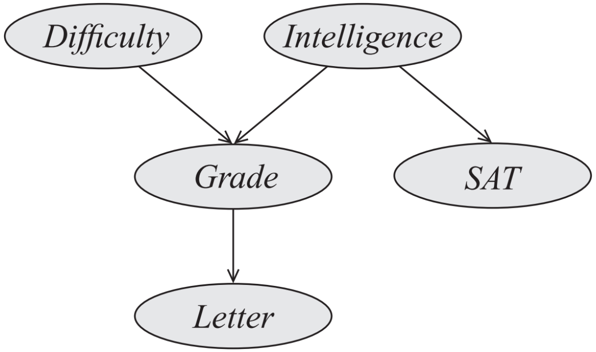
Linear time complexity

- **input:** graph G and **X, Y, Z**
- **output:** $\mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z}$?
- ***mark*** the variables in **Z** and all of their *ancestors* in G
- ***breadth-first-search*** starting from **X**
- stop any trail that reaches a **blocked node**
- a node in **Y** is reached?

Blocked node:

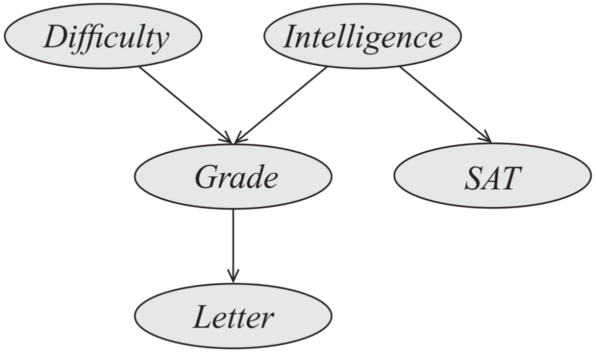
- **unmarked** middle of a collider (V-structure)
- in **Z** and not a collider

D-separation quiz



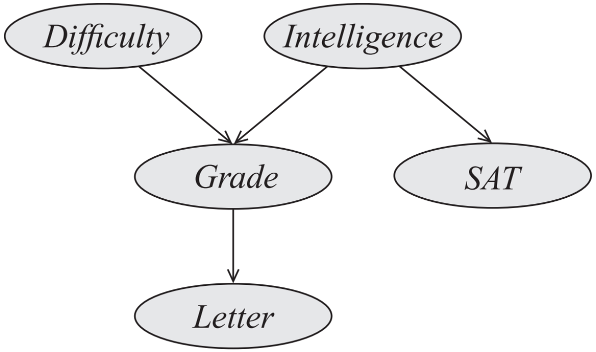
D-separation quiz

$G \perp S \mid \emptyset?$



D-separation quiz

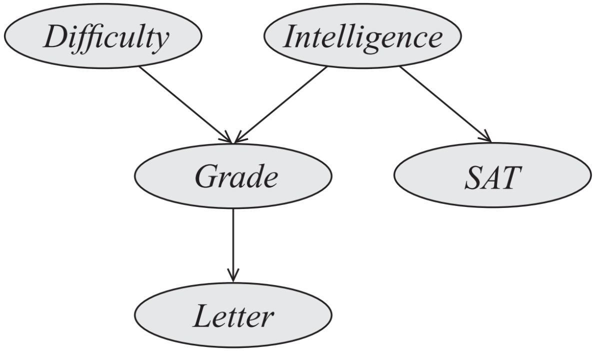
$G \perp S \mid \emptyset?$ ✖



D-separation quiz

$G \perp S \mid \emptyset?$ ✖

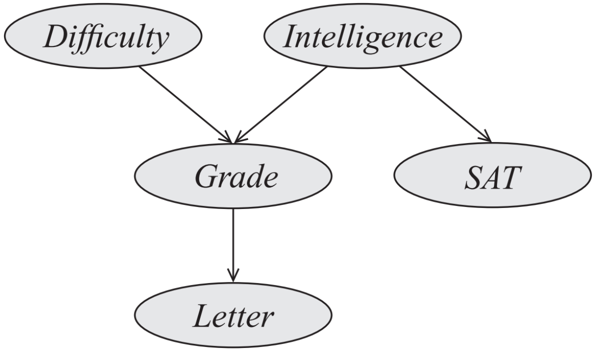
$D \perp L \mid G?$



D-separation **quiz**

$G \perp S \mid \emptyset?$

$D \perp L \mid G?$

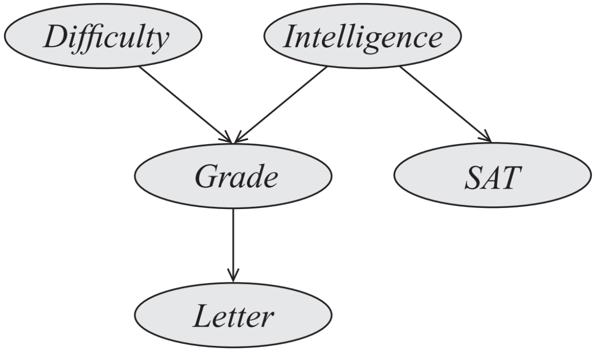


D-separation quiz

$G \perp S \mid \emptyset?$ 

$D \perp L \mid G?$ 

$D \perp I, S \mid \emptyset?$

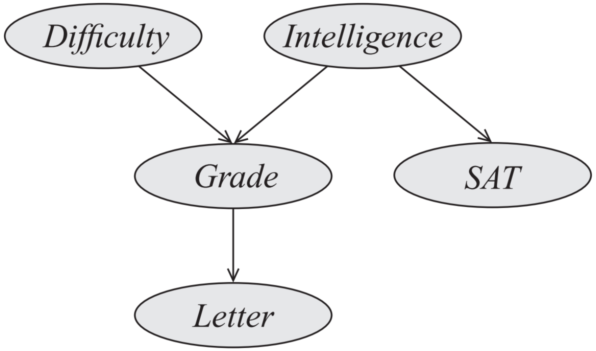


D-separation quiz

$G \perp S \mid \emptyset$? ✗

$D \perp L \mid G$? ✓

$D \perp I, S \mid \emptyset$? ✓



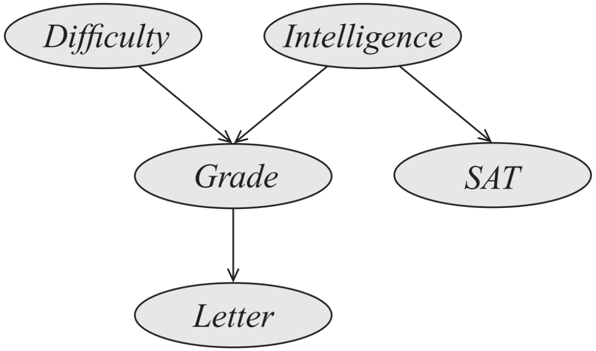
D-separation quiz

$G \perp S \mid \emptyset?$ 

$D \perp L \mid G?$ 

$D \perp I, S \mid \emptyset?$ 

$D, L \perp S \mid I, G?$



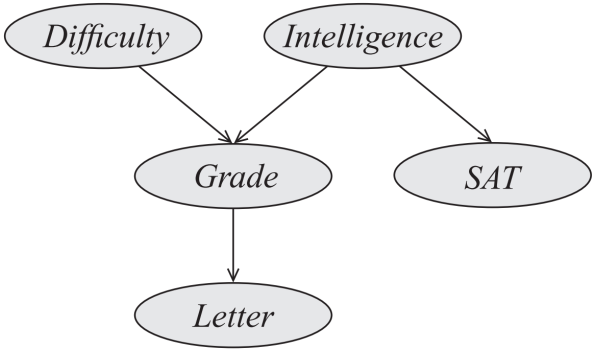
D-separation quiz

$G \perp S \mid \emptyset?$ ✗

$D \perp L \mid G?$ ✓

$D \perp I, S \mid \emptyset?$ ✓

$D, L \perp S \mid I, G?$ ✓



Summary

graph and **distribution** are combined:

- factorization of the **distribution**
  - according to the **graph** $P(\mathbf{X}) = \prod_i P(X_i \mid Pa^{\mathcal{G}}_{X_i})$

- conditional independencies of the **distribution**
  - inferred from the **graph**
    - local CI: $X_i \perp NonDescendents_{X_i} \mid Parents_{X_i}$
    - global CI: D-separation

Summary

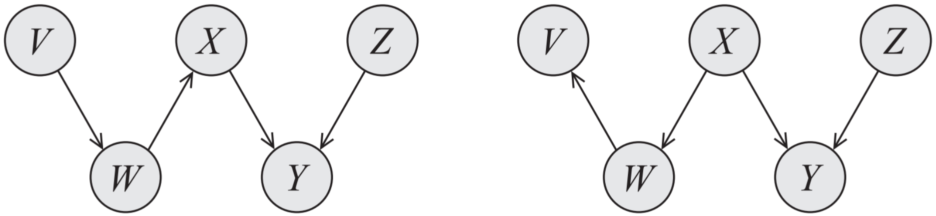
- factorization of the distribution
- local conditional independencies
- global conditional independencies

identify the same **family** of distributions

Bonus slides

Equivalence class of DAGs

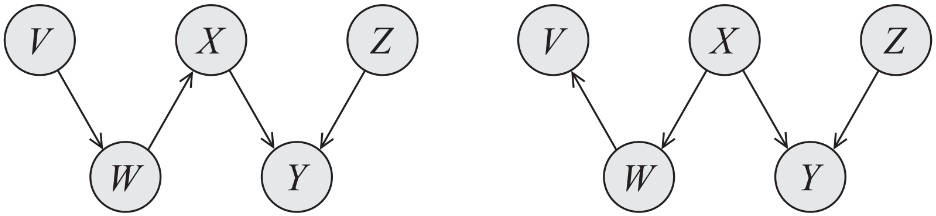
Two DAGs are I-equivalent if $\mathcal{I}(\mathcal{G}) = \mathcal{I}(\mathcal{G}')$



P factorizes on both of these graphs

Equivalence class of DAGs

Two DAGs are I-equivalent if $\mathcal{I}(\mathcal{G}) = \mathcal{I}(\mathcal{G}')$



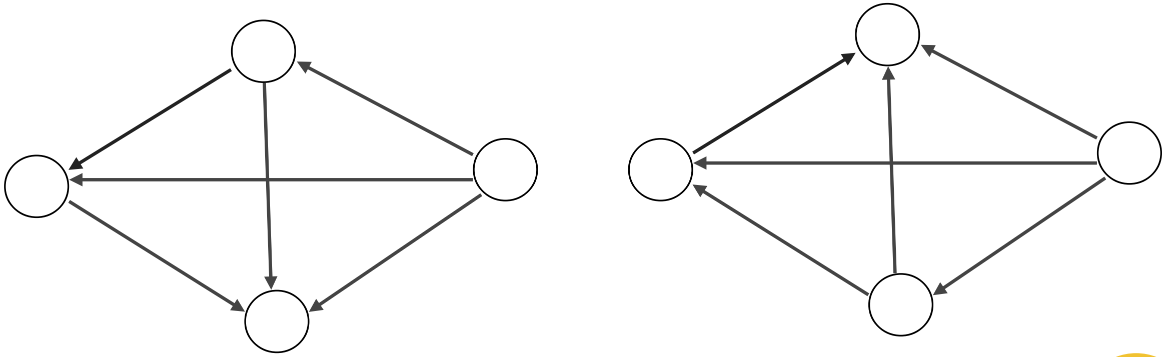
P factorizes on both of these graphs

From d-separation algorithm it is ***sufficient***

- same undirected skeleton
- same v-structures

Equivalence class of DAGs

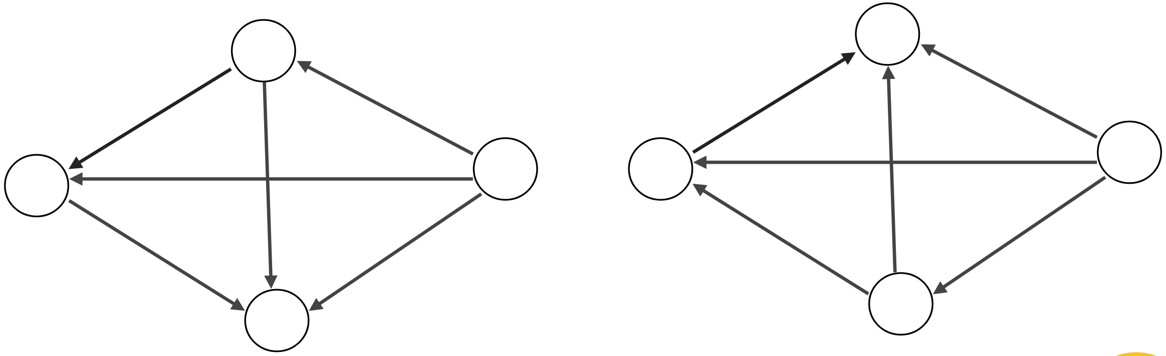
Two DAGs are I-equivalent if $\mathcal{I}(\mathcal{G}) = \mathcal{I}(\mathcal{G}')$



different v-structures, yet $\mathcal{I}(\mathcal{G}) = \mathcal{I}(\mathcal{G}') = \emptyset$

Equivalence class of DAGs

Two DAGs are I-equivalent if $\mathcal{I}(\mathcal{G}) = \mathcal{I}(\mathcal{G}')$



different v-structures, yet $\mathcal{I}(\mathcal{G}) = \mathcal{I}(\mathcal{G}') = \emptyset$ 

here, v-structures are irrelevant for I-equivalent because:

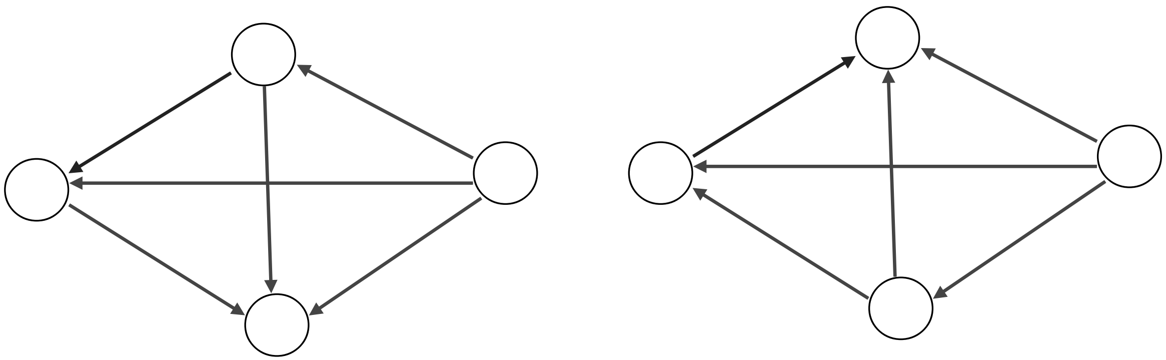
- parents are connected (**moral parents!**)

Equivalence class of DAGs

Two DAGs are I-equivalent if $\mathcal{I}(\mathcal{G}) = \mathcal{I}(\mathcal{G}')$

$\mathcal{I}(\mathcal{G}) = \mathcal{I}(\mathcal{G}') \quad \Leftrightarrow$

same undirected skeleton
same immoralities



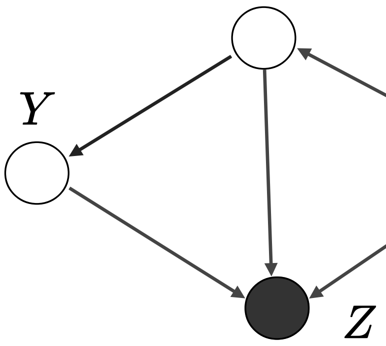
Equivalence class of DAGs

Two DAGs are I-equivalent if $\mathcal{I}(\mathcal{G}) = \mathcal{I}(\mathcal{G}')$

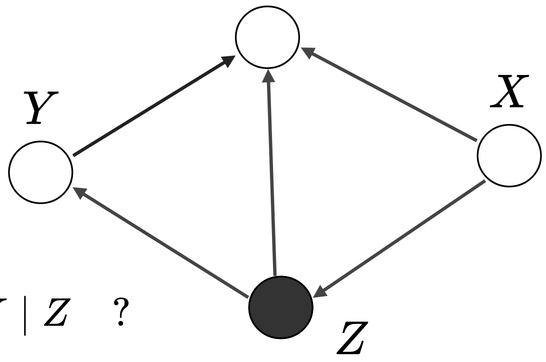
$\mathcal{I}(\mathcal{G}) = \mathcal{I}(\mathcal{G}') \quad \Leftrightarrow$

same undirected skeleton

same immoralities

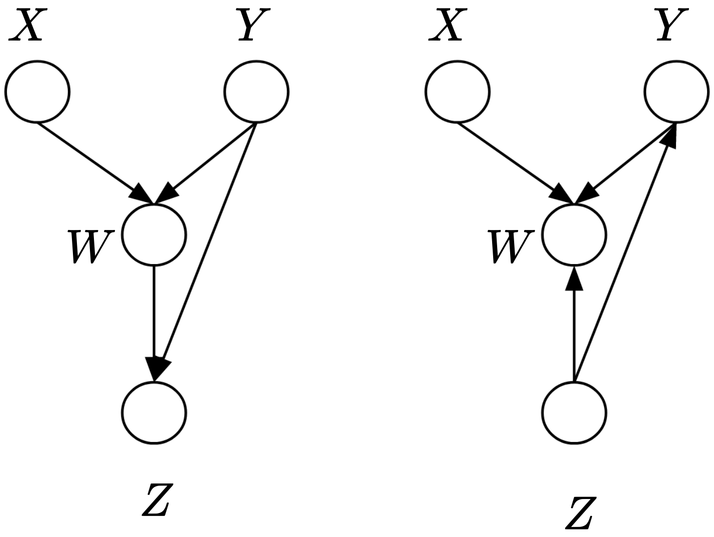


$X \perp Y \mid Z \quad ?$



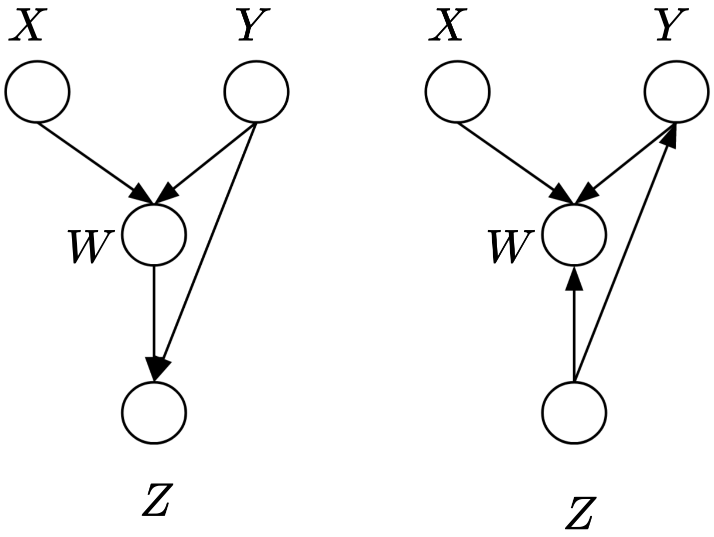
I-Equivalence quiz

do these DAGs have the same set of CIs?



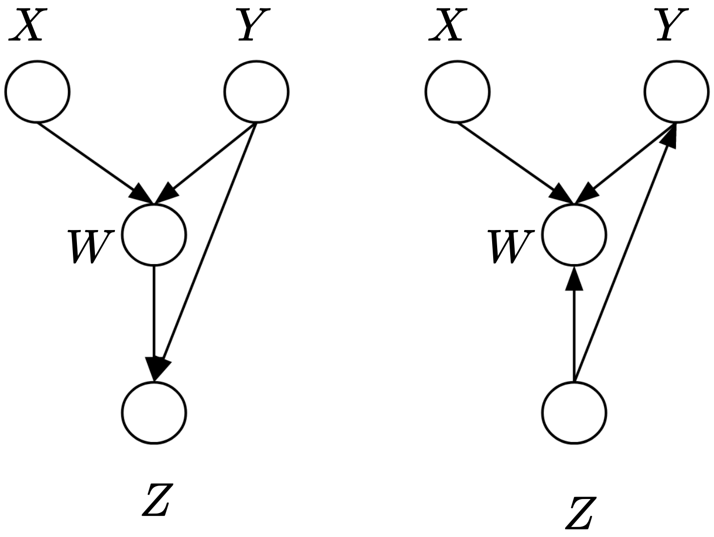
I-Equivalence quiz

do these DAGs have the same set of CIs? no!



I-Equivalence **quiz**

do these DAGs have the same set of CIs? no!



$X \perp Z \mid W$

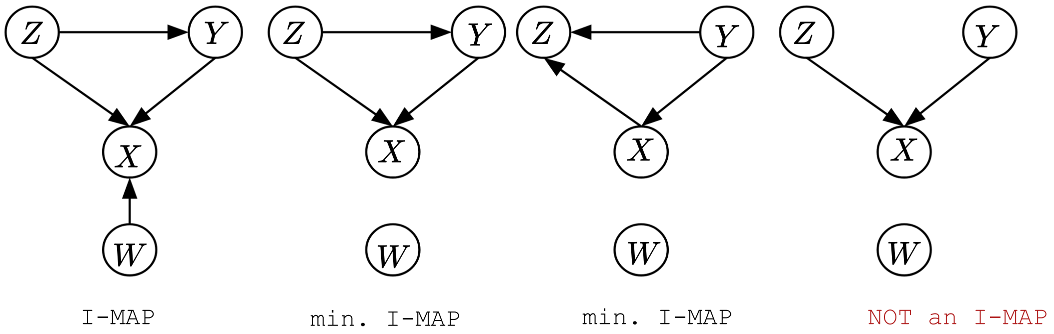
Minimal I-map

which graph G to use for P?

G is minimal I-map for P:

- G is an I-map for P: $\mathcal{I}(\mathcal{G}) \subseteq \mathcal{I}(P)$
- removing any edge destroys this property

Example: $P(X, Y, Z, W) = P(X \mid Y, Z)P(W)P(Y \mid Z)P(Z)$

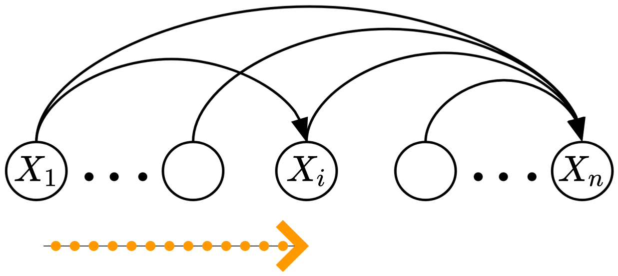


Minimal I-map from CI

which graph G to use for P?

input: $\mathcal{I}(P)$ **or** an oracle; an ordering $X_1, \ldots, X_n$

output: a minimal I-map G



for i=1...n

- find minimal $\mathbf{U} \subseteq \{X_1, \ldots, X_{i-1}\}$ s.t. $(X_i \perp X_1, \ldots, X_{i-1} - \mathbf{U} \mid \mathbf{U})$
- set $Pa_{X_i} \leftarrow \mathbf{U}$

$X_i \perp NonDesc_{X_i} \mid Pa_{X_i}$

Minimal I-map from CI

which graph G to use for P?

input: $\mathcal{I}(P)$ **or** an oracle; an ordering $X_1, \ldots, X_n$

output: a minimal I-map G

`different orderings give different graphs`

Minimal I-map from CI

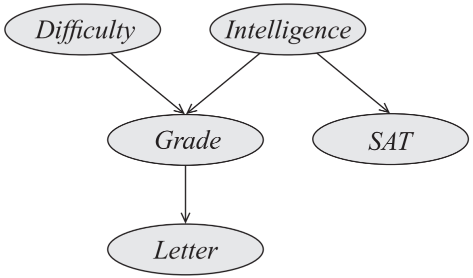
which graph G to use for P?

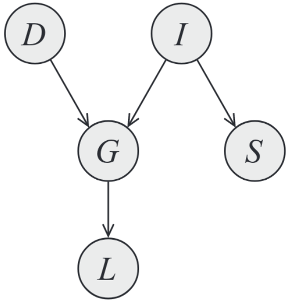
input: $\mathcal{I}(P)$ **or** an oracle; an ordering $X_1, \ldots, X_n$

output: a minimal I-map G

different orderings give different graphs

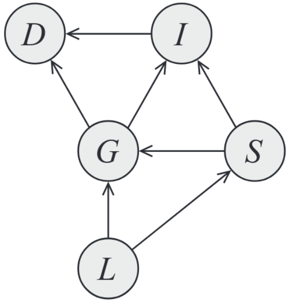
Example:



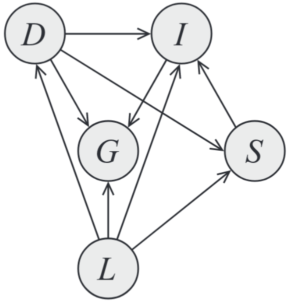


D,I,S,G,L

(a topological ordering)



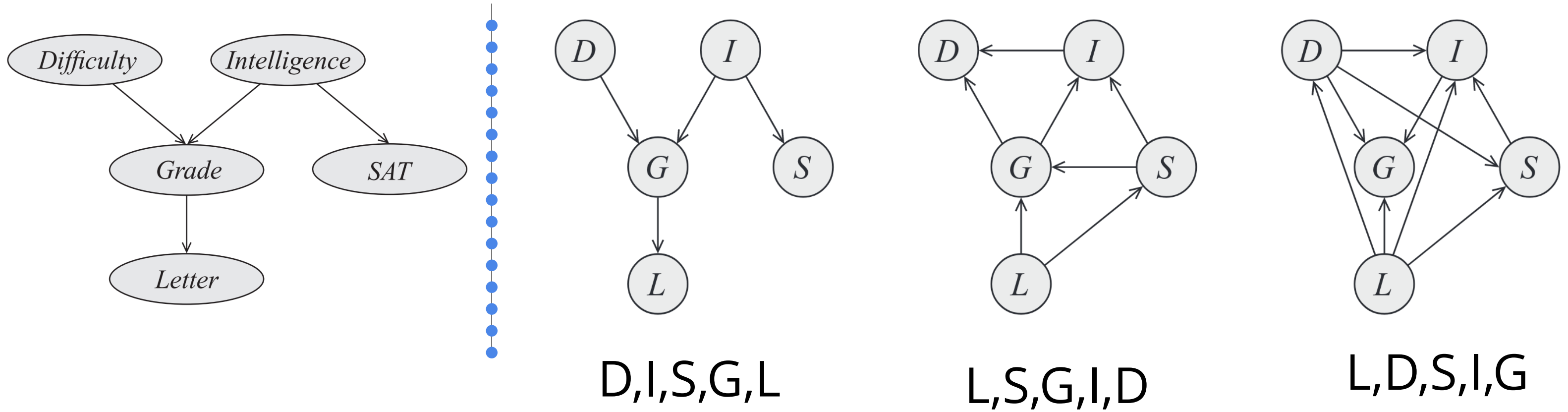
L,S,G,I,D



L,D,S,I,G

Perfect MAP (P-MAP)

which graph G to use for P?



all the graphs above are minimal I-MAPs $\mathcal{I}(\mathcal{G}) \subseteq \mathcal{I}(P)$

Perfect MAP: $\mathcal{I}(\mathcal{G}) = \mathcal{I}(P)$

Perfect map (P-map)

which graph G to use for P?

Perfect MAP: $\mathcal{I}(\mathcal{G}) = \mathcal{I}(P)$

P may not have a P-map in the form of BN

Perfect map (P-map)

which graph G to use for P?

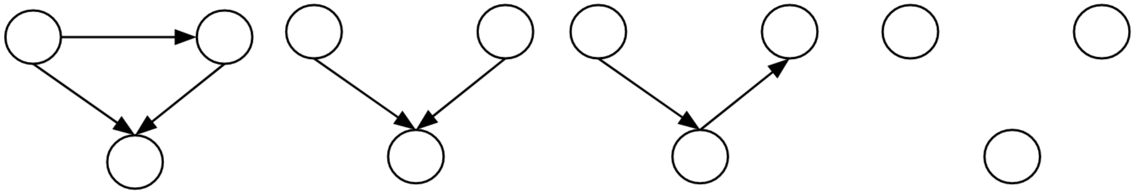
Perfect MAP: $\mathcal{I}(\mathcal{G}) = \mathcal{I}(P)$

P may not have a P-map in the form of BN

Example:

$$P(x, y, z) = \begin{cases} 1/12, & \text{if } x \otimes y \otimes z = 0 \\ 1/6, & \text{if } x \otimes y \otimes z = 1 \end{cases}$$

$$(X \perp Y), (Y \perp Z), (X \perp Z) \in \mathcal{I}(P)$$
$$(X \perp Y \mid Z), (Y \perp Z \mid Z), (X \perp Z \mid Y) \notin \mathcal{I}(P)$$



Perfect map (P-map)

which graph G to use for P?

Perfect MAP: $\mathcal{I}(\mathcal{G}) = \mathcal{I}(P)$

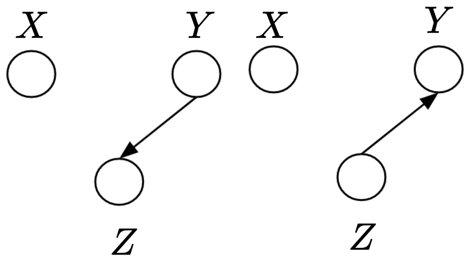
P may not have a P-map in the form of a BN

if P has a P-map: is it unique?

unique up to I-equivalence

Example:

$$\mathcal{I}(P) = \{(X \perp Y, Z \mid \emptyset), (X \perp Y \mid Z), (X \perp Z \mid Y)\}$$



Perfect map (P-map)

which graph G to use for P?

Perfect MAP: $\mathcal{I}(\mathcal{G}) = \mathcal{I}(P)$

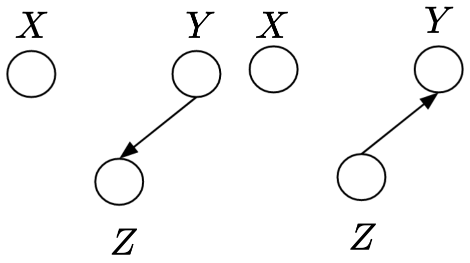
P may not have a P-map in the form of a BN

if P has a P-map: is it unique?

unique up to I-equivalence

Example:

$$\mathcal{I}(P) = \{(X \perp Y, Z \mid \emptyset), (X \perp Y \mid Z), (X \perp Z \mid Y)\}$$



How to find P-MAPs? discussed in learning BNs

Summary

- factorization of the dist.
- local CIs
- global CIs

identify the same family of distributions

can be represented using an equivalent class of graphs*:*

- alternative factorization
- different local CIs
- same global CIs