Probabilistic (Graphical) Models

and inference

Oliver Obst · Autumn 2024

WESTERN SYDNEY UNIVERSITY

Probabilistic (Graphical) Models and Inference

(PGM: Probabilistic Graphical Models: Principles and Techniques by Daphne Koller and Nir Friedman. MIT Press)
(PMLI: Probabilistic Machine Learning: An introduction by Kevin Murphy. MIT Press)

Week	Lecture	Required reading	Assessment
1 Monday, 4 March 2024	Introduction, Probability Theory	PGM Chapter 2, PMLI Chapter 6.1	
2 Monday, 11 March 2024	Directed and undirected networks introduction	PGM Chapter 3 & 4	Quiz 1
3 Monday, 18 March 2024	Variable elimination		
4 Monday, 25 March 2024	Belief propagation		Quiz 2
5 Monday, 1 April 2024	public holiday		5 April 2024: census date
6 Monday, 8 April 2024	Message passing / Graph neural networks		
7 Monday, 15 April 2024	Sampling		Quiz 3
8 Monday, 22 April 2024	Mid-term break		
9 Monday, 29 April 2024	Variational inference		Intra-session exam
10 Monday, 6 May 2024	Autoregressive models		Quiz 4
11 Monday, 13 May 2024	Variational Auto-Encoders		
12 Monday, 20 May 2024	GANs		Quiz 5
13 Monday, 27 May 2024	Energy-based models		
14 Monday, 3 June 2024	Evaluating generative models		Quiz 6
Monday, 17 June 2024			Project due

Graphical models

Graphs, formal definitions

A graph is a data structure $K = (V, E)$, where V is the set of vertices (nodes), and E is the set of edges.

Graphs, formal definitions

A graph is a data structure $K = (V, E)$, where V is the set of vertices (nodes), and E is the set of edges.

- $V = \{X_1, \ldots, X_n\}$.
  A pair of nodes, X_i, X_j can be connected by a directed edge $X_i \rightarrow X_j$ or an undirected edge $X_i - X_j$.

Graphs, formal definitions

A graph is a data structure $K = (V, E)$, where V is the set of vertices (nodes), and E is the set of edges.

- $V = \{X_1, \ldots, X_n\}$.
  A pair of nodes, X_i, X_j can be connected by a directed edge $X_i \rightarrow X_j$ or an undirected edge $X_i - X_j$.

- The set of edges E is a set of pairs; each pair is one of $X_i \rightarrow X_j$, $X_j \rightarrow X_i$, or $X_i - X_j$, for $X_i, X_j \in V, i < j$.
  To denote any connection (directed or undirected): $X_i \rightleftharpoons X_j$

Graphs, formal definitions

A graph is a data structure $K = (V, E)$, where V is the set of vertices (nodes), and E is the set of edges.

- $V = \{X_1, \ldots, X_n\}$.
  A pair of nodes, X_i, X_j can be connected by a directed edge $X_i \to X_j$ or an undirected edge $X_i - X_j$.

- The set of edges E is a set of pairs; each pair is one of $X_i \to X_j$, $X_j \to X_i$, or $X_i - X_j$, for $X_i, X_j \in V, i < j$.
  To denote any connection (directed or undirected): $X_i \rightleftharpoons X_j$

Graphs that have only directed edges are called *directed graphs*.
In the PGM book, directed graphs are usually written using the letter G.

Graphs, formal definitions

A graph is a data structure $K = (V, E)$, where V is the set of vertices (nodes), and E is the set of edges.

- $V = \{X_1, \ldots, X_n\}$.
  A pair of nodes, X_i, X_j can be connected by a directed edge $X_i \to X_j$ or an undirected edge $X_i - X_j$.

- The set of edges E is a set of pairs; each pair is one of $X_i \to X_j$, $X_j \to X_i$, or $X_i - X_j$, for $X_i, X_j \in V, i < j$.
  To denote any connection (directed or undirected): $X_i \rightleftharpoons X_j$

Graphs that have only directed edges are called *directed graphs*.
In the PGM book, directed graphs are usually written using the letter G.

Graphs with only undirected edges are *undirected graphs*, denoted with H.

Graphs, formal definitions

A graph is a data structure $K = (V, E)$, where V is the set of vertices (nodes), and E is the set of edges.

- $V = \{X_1, \ldots, X_n\}$.
  A pair of nodes, X_i, X_j can be connected by a directed edge $X_i \rightarrow X_j$ or an undirected edge $X_i - X_j$.

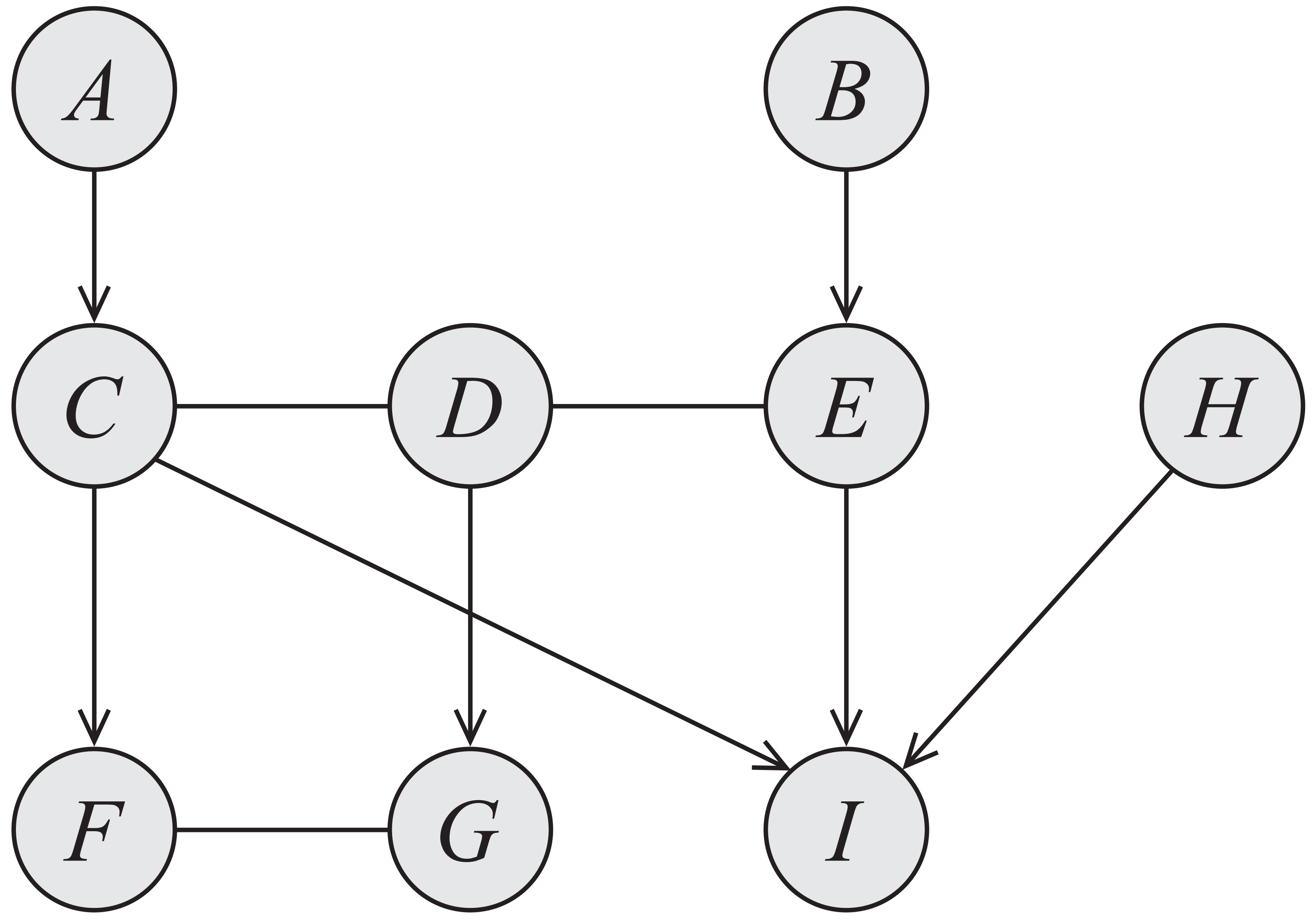
- The set of edges E is a set of pairs; each pair is one of $X_i \rightarrow X_j$, $X_j \rightarrow X_i$, or $X_i - X_j$, for $X_i, X_j \in V, i < j$.
  To denote any connection (directed or undirected): $X_i \rightleftharpoons X_j$

Graphs that have only directed edges are called *directed graphs*.
In the PGM book, directed graphs are usually written using the letter G.

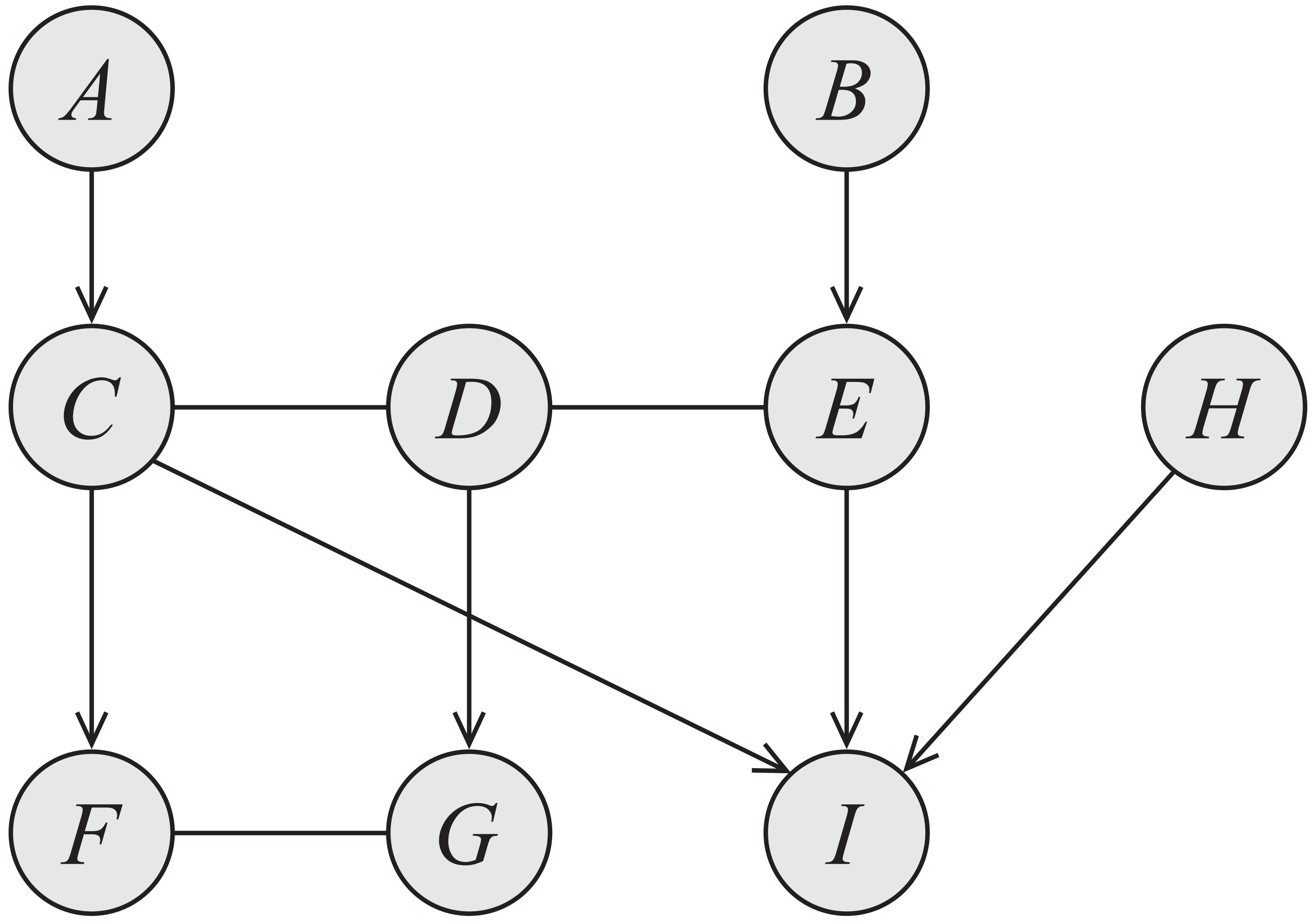
Graphs with only undirected edges are *undirected graphs*, denoted with H.

- We can get the undirected version of a graph $K = (V, E)$ by replacing all edges with undirected edges:
  $H = (V, E')$, where $E' = \{X - Y : X \leftrightharpoons Y \in E\}$.

Graphs, formal definitions



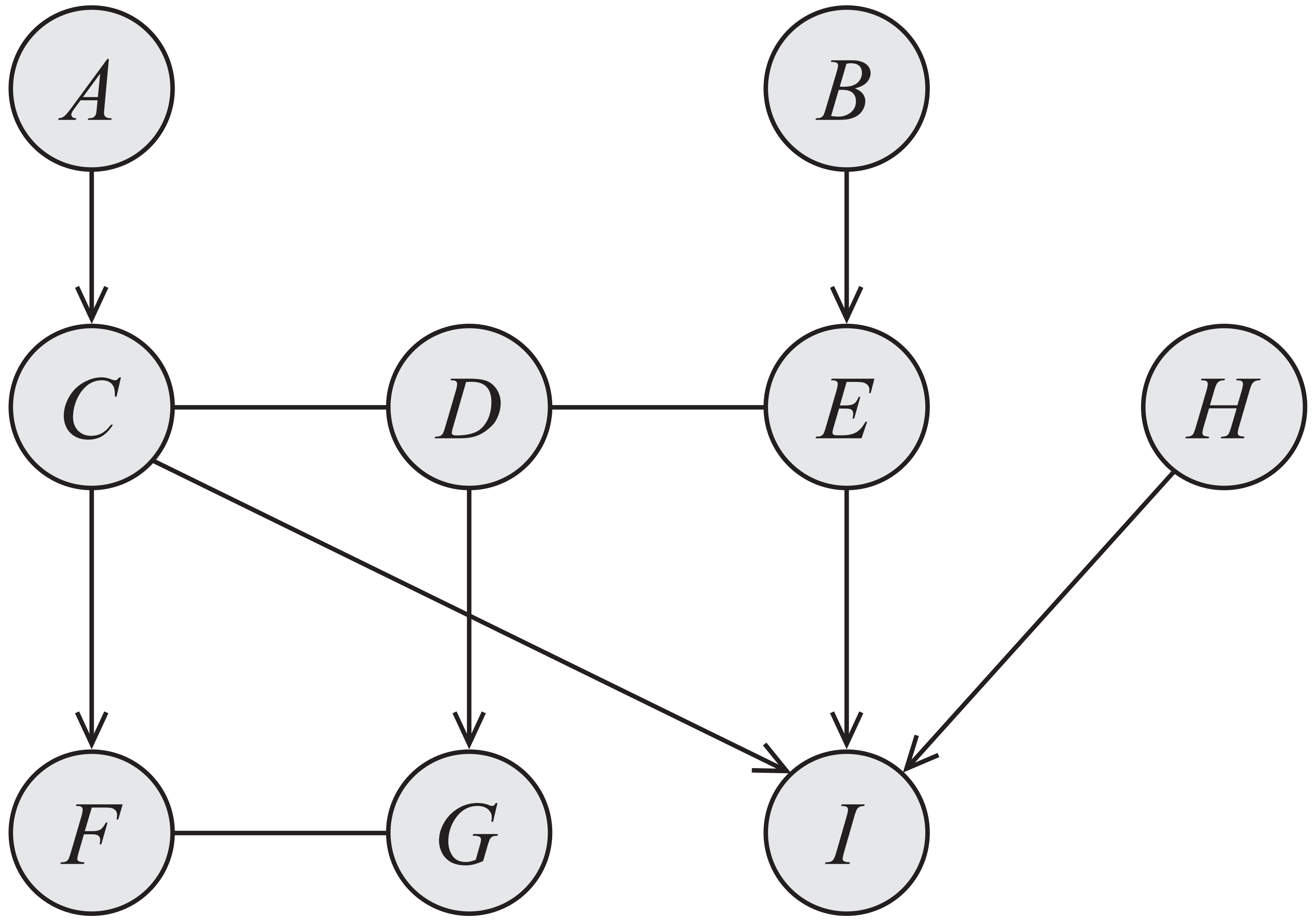
Graphs, formal definitions



$X_i \to X_j \in E$: X_j is a *child* of X_i in K.

Notation: P_{a_X}, Ch_X : parents, children of X

Graphs, formal definitions



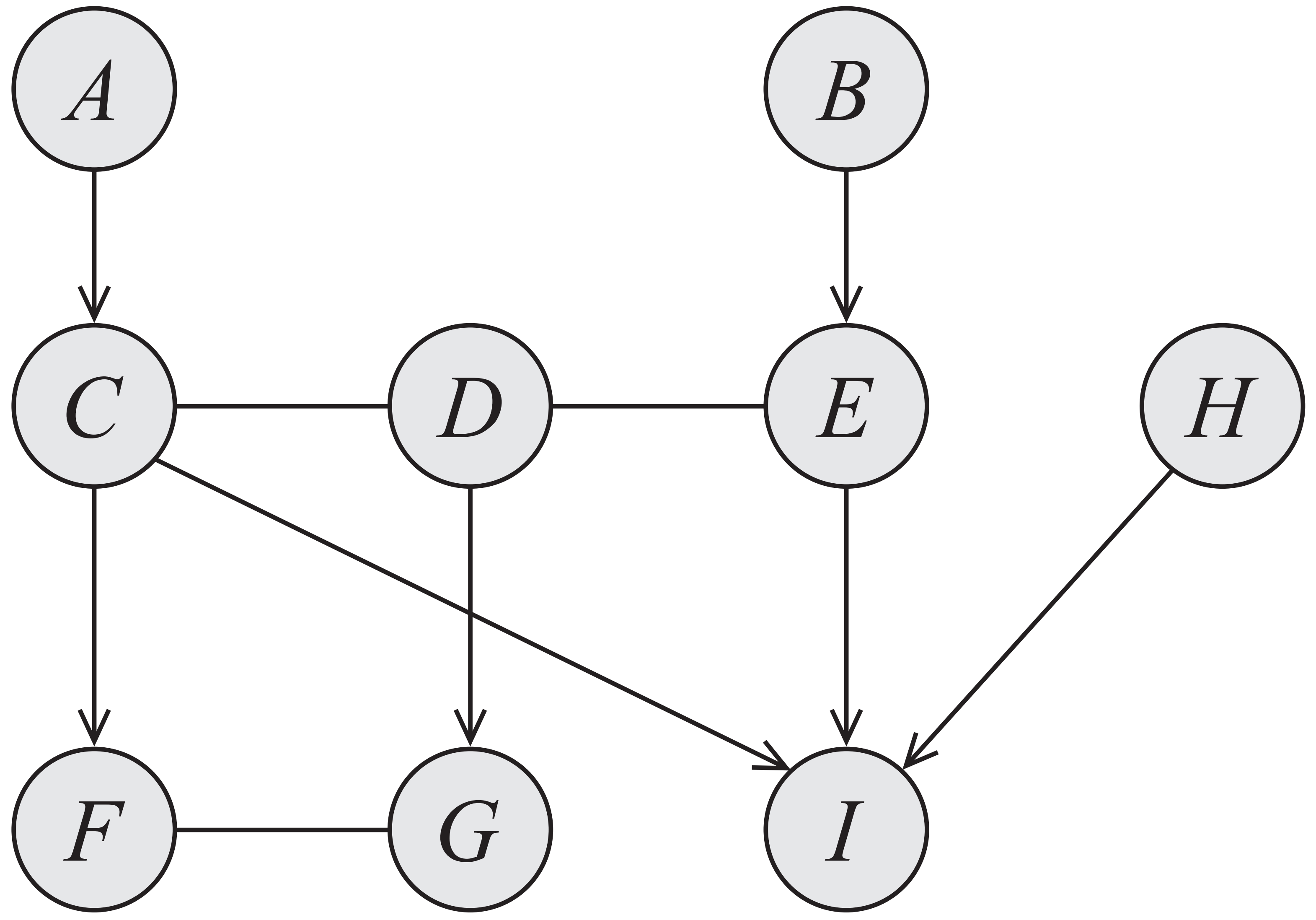
$X_i \to X_j \in E$: X_j is a *child* of X_i in K.

Notation: P_{a_X}, Ch_X : parents, children of X

$X_i - X_j \in E$: X_i, X_j are *neighbours*

Notation: Nb_X : neighbours of X

Graphs, formal definitions



$X_i \to X_j \in E$: X_j is a *child* of X_i in K.

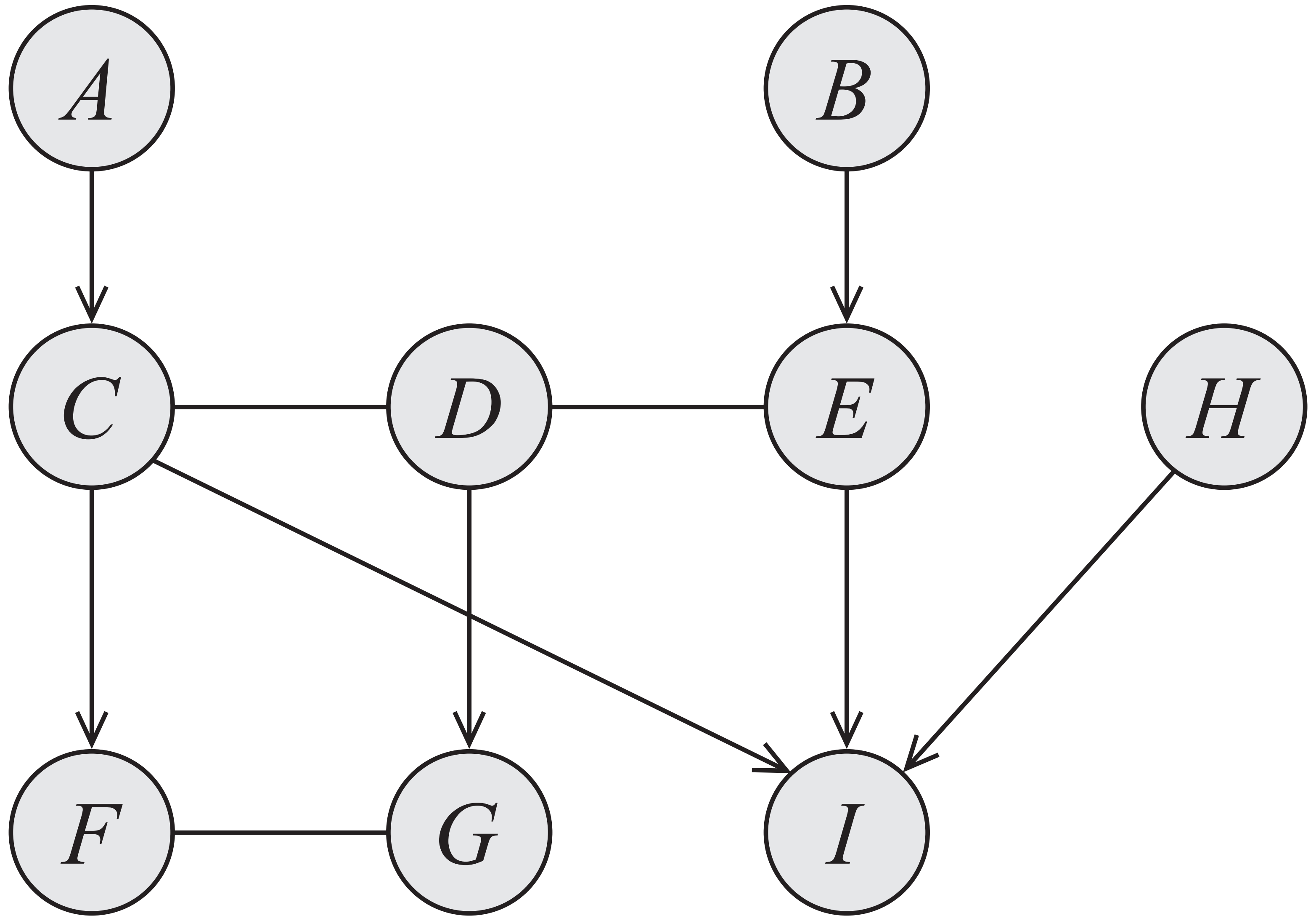
Notation: P_{a_X}, Ch_X : parents, children of X

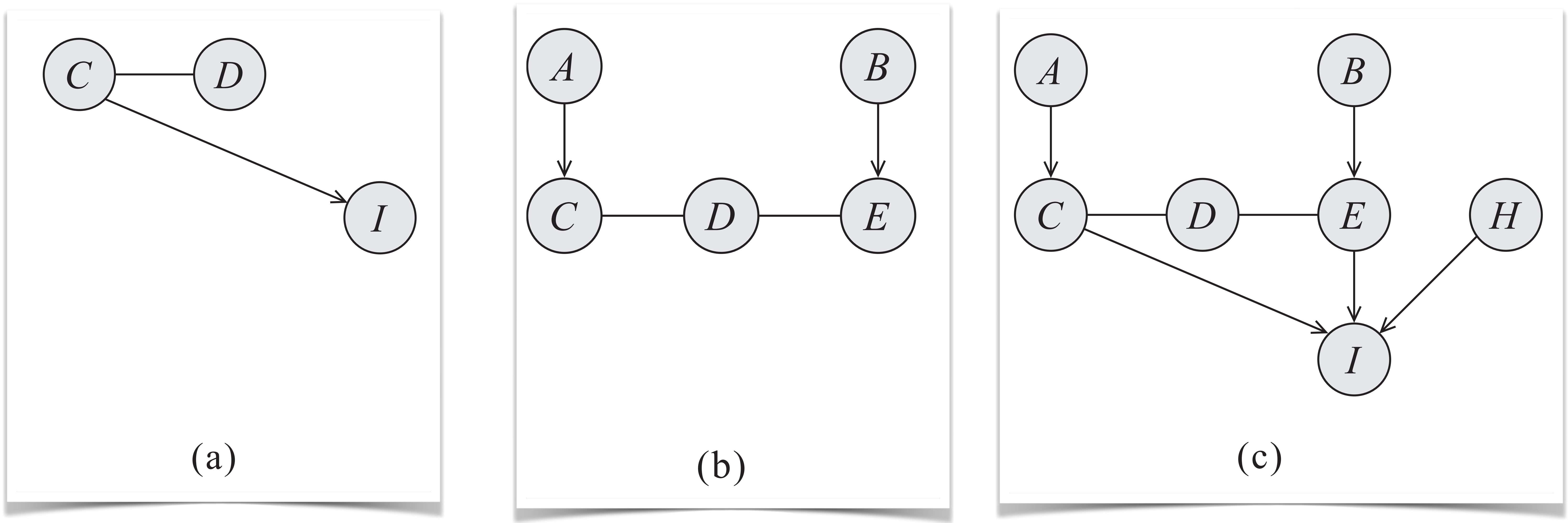
$X_i - X_j \in E$: X_i, X_j are *neighbours*

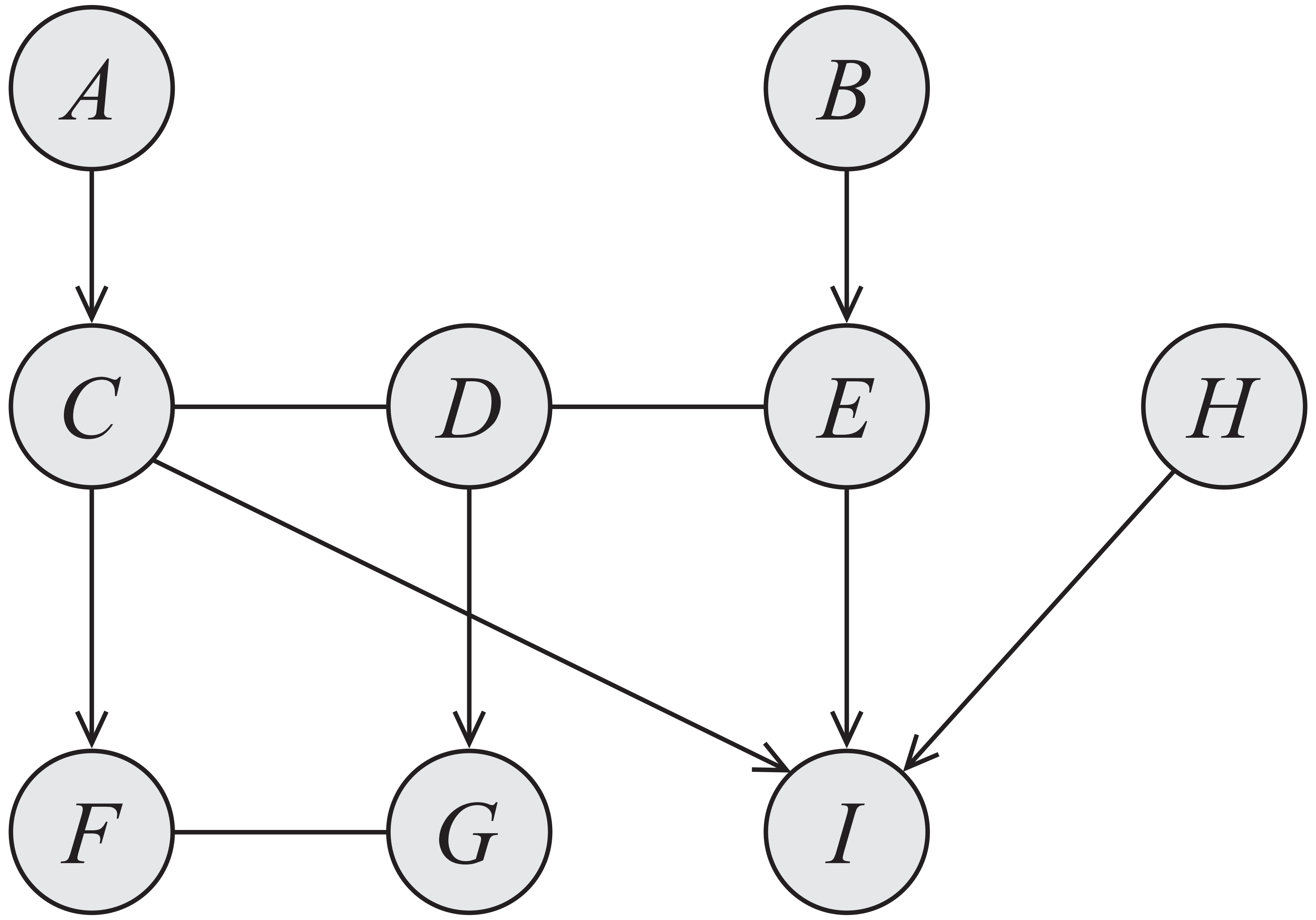
Notation: Nb_X : neighbours of X

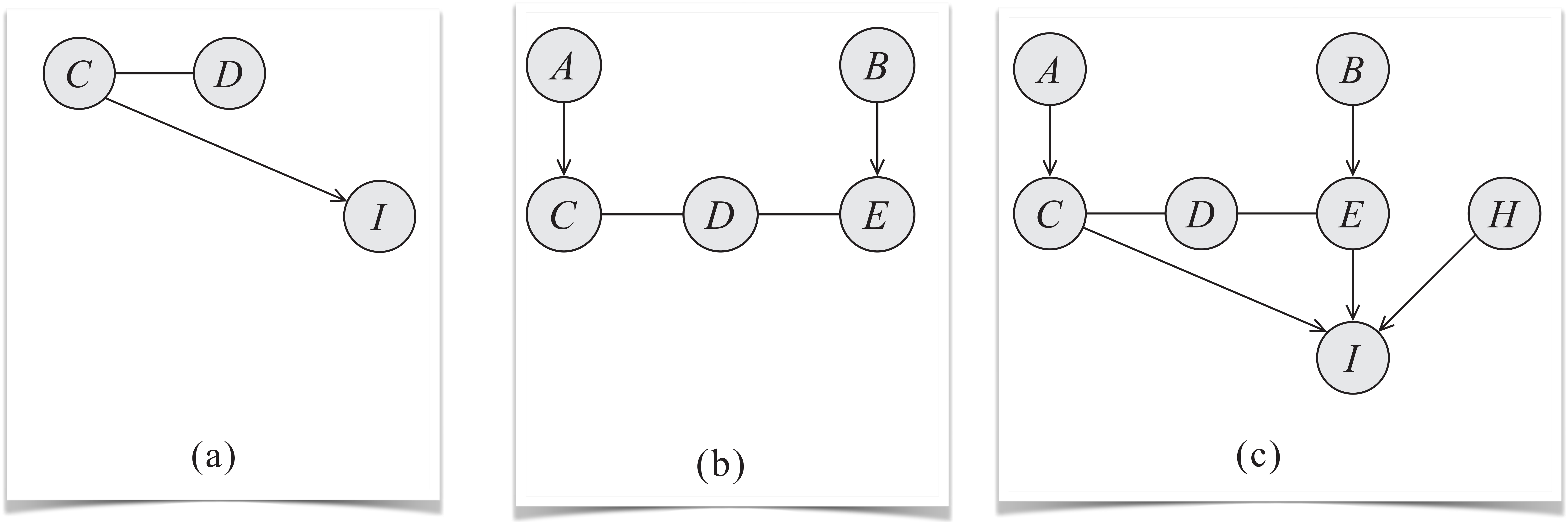
The set of all parents and neighbours of X combined, $\mathrm{Pa}(X) \cup \mathrm{Nb}(X)$, is called the *boundary* of X.

Notation: $\mathrm{Boundary}_X$


A
B
C
D
E
H
F
G
I


C
D
I
(a)
A
B
C
D
E
(b)
A
B
C
D
E
H
I
(c)





(a) $K = (V, E)$; let $\boldsymbol{X} \in V$. The induced subgraph $K[\boldsymbol{X}]$ is the graph $(\boldsymbol{X}, E')$, where E' are all the edges $X \leftrightharpoons Y \in E$ such that $X, Y \in \boldsymbol{X}$.
(example: $K[C, D, I]$)

- A subgraph over $\boldsymbol{X}$ is complete if every two nodes in $\boldsymbol{X}$ are connected by some edge.The set $\boldsymbol{X}$ is often called a clique; we say a clique $\boldsymbol{X}$ is maximal if for any superset of nodes $\boldsymbol{Y} \supset \boldsymbol{X}$, $\boldsymbol{Y}$ is not a clique.

(a) $K = (V, E)$; let $\boldsymbol{X} \in V$. The induced subgraph $K[\boldsymbol{X}]$ is the graph $(\boldsymbol{X}, E')$, where E' are all the edges $X \leftrightharpoons Y \in E$ such that $X, Y \in \boldsymbol{X}$.
(example: $K[C, D, I]$)

- A subgraph over $\boldsymbol{X}$ is complete if every two nodes in $\boldsymbol{X}$ are connected by some edge.The set $\boldsymbol{X}$ is often called a clique; we say a clique $\boldsymbol{X}$ is maximal if for any superset of nodes $\boldsymbol{Y} \supset \boldsymbol{X}$, $\boldsymbol{Y}$ is not a clique.

(b) and (c) A subset of nodes $\boldsymbol{X} \in V$ is upwardly closed in K if, for any $X \in \boldsymbol{X}$, we have that $\text{Boundary}_X \subset \boldsymbol{X}$. The upward closure of $\boldsymbol{X}$ is the minimal upwardly closed subset $\boldsymbol{Y}$ that contains $\boldsymbol{X}$. The upwardly closed subgraph of $\boldsymbol{X}$, denoted $K^+[\boldsymbol{X}]$, is the induced subgraph over $\boldsymbol{Y}$, $K[\boldsymbol{Y}]$.

Bayesian networks

PGM chapter 3

A Bayesian network is specified by a directed acyclic graph $G = (V, E)$ with:

1. One node $i \in V$ for each random variable X_i
2. One conditional probability distribution (CDP) per node, $p(x_i \mid \boldsymbol{x}_{\mathrm{Pa}(i)})$, specifying the variable's probability conditioned on its parents' values.

Bayesian networks

PGM chapter 3

A Bayesian network is specified by a directed acyclic graph $G = (V, E)$ with:

1. One node $i \in V$ for each random variable X_i
2. One conditional probability distribution (CDP) per node, $p(x_i \,|\, \boldsymbol{x}_{\mathrm{Pa}(i)})$, specifying the variable's probability conditioned on its parents' values.

Corresponds 1-1 with a particular factorisation of the joint distribution:

Bayesian networks

PGM chapter 3

A Bayesian network is specified by a directed acyclic graph $G = (V, E)$ with:

1. One node $i \in V$ for each random variable X_i
2. One conditional probability distribution (CDP) per node, $p(x_i \,|\, \boldsymbol{x}_{\mathrm{Pa}(i)})$, specifying the variable's probability conditioned on its parents' values.

Corresponds 1-1 with a particular factorisation of the joint distribution:

$$p(x_1, \ldots, x_n) = \prod_{i \in V} p(x_i \,|\, \boldsymbol{x}_{\mathrm{Pa}(i)})$$

Bayesian networks

PGM chapter 3

A Bayesian network is specified by a directed acyclic graph $G = (V, E)$ with:

1. One node $i \in V$ for each random variable X_i
2. One conditional probability distribution (CDP) per node, $p(x_i \,|\, \boldsymbol{x}_{\mathrm{Pa}(i)})$, specifying the variable's probability conditioned on its parents' values.

Corresponds 1-1 with a particular factorisation of the joint distribution:

$$p(x_1, \ldots, x_n) = \prod_{i \in V} p(x_i \,|\, \boldsymbol{x}_{\mathrm{Pa}(i)})$$

Powerful framework for designing algorithms to perform probability computations.

Semantics and Factorisation

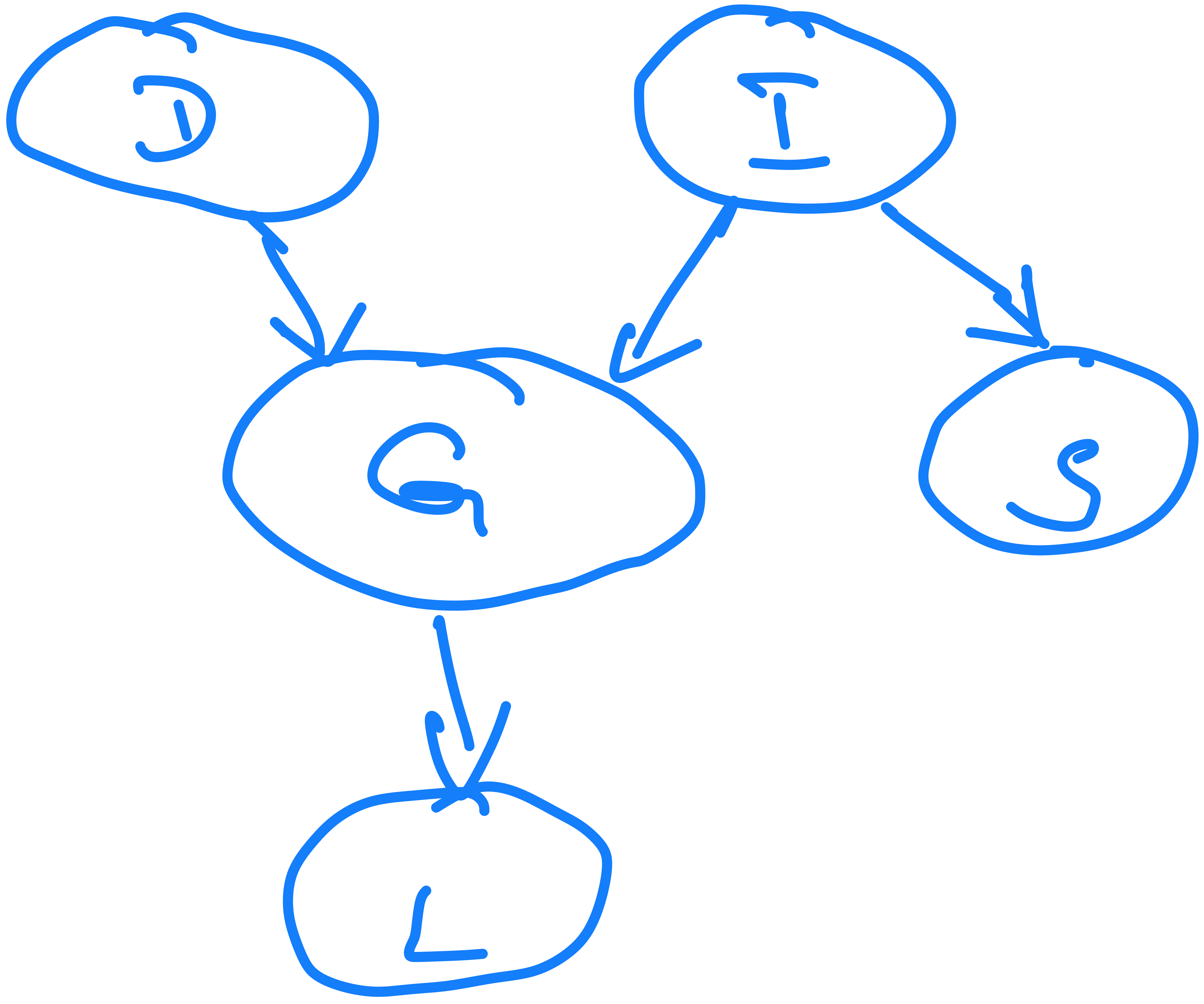
Bayesian network example

- Grade
- Course Difficulty
- Student Intelligence
- Student SAT
- Reference Letter

For this example, everything is binary, except grade (3 values)

Bayesian network example

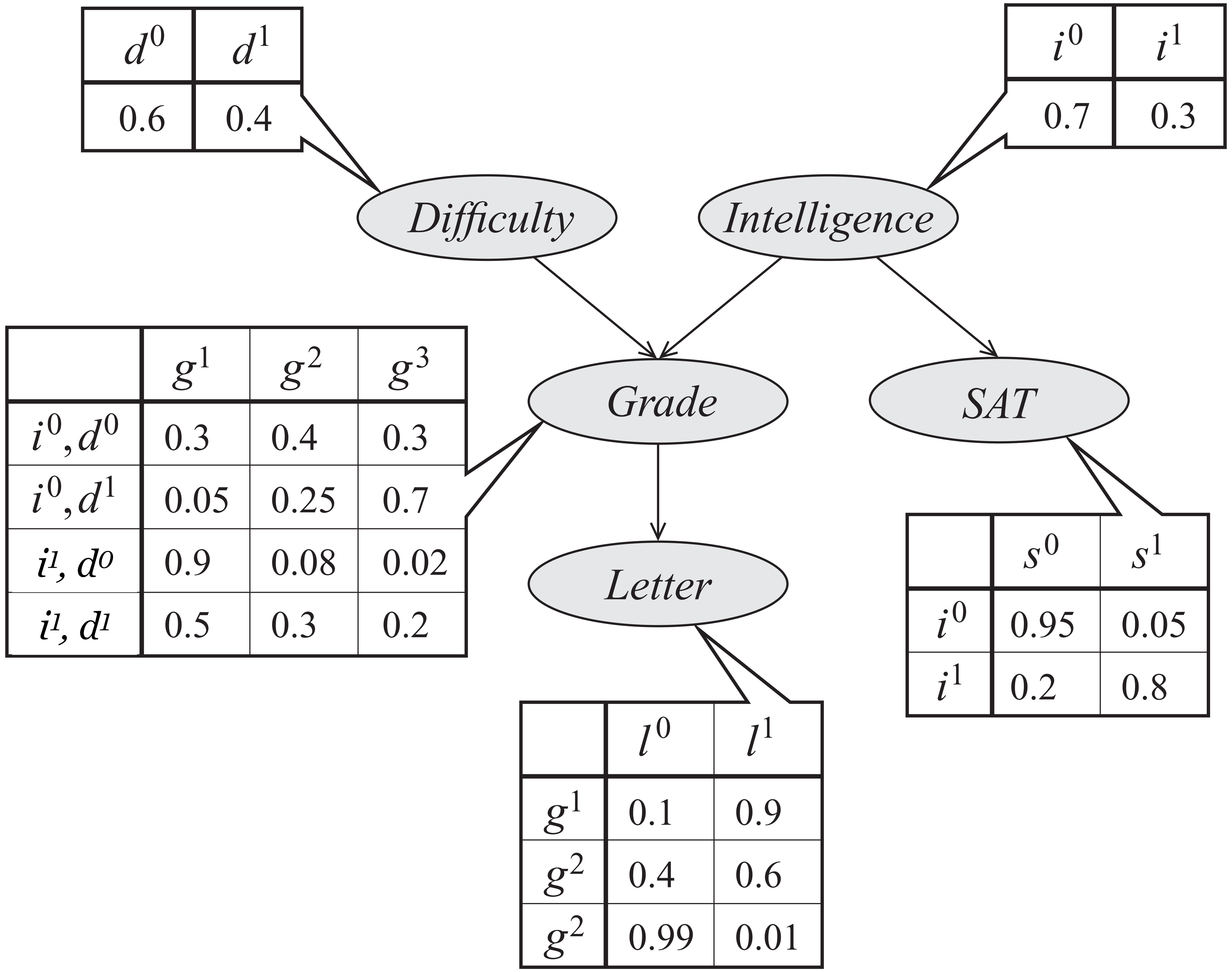
- Grade
- Course Difficulty
- Student Intelligence
- Student SAT
- Reference Letter



For this example, everything is binary, except grade (3 values)

Example

- Consider the Bayesian network here:



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

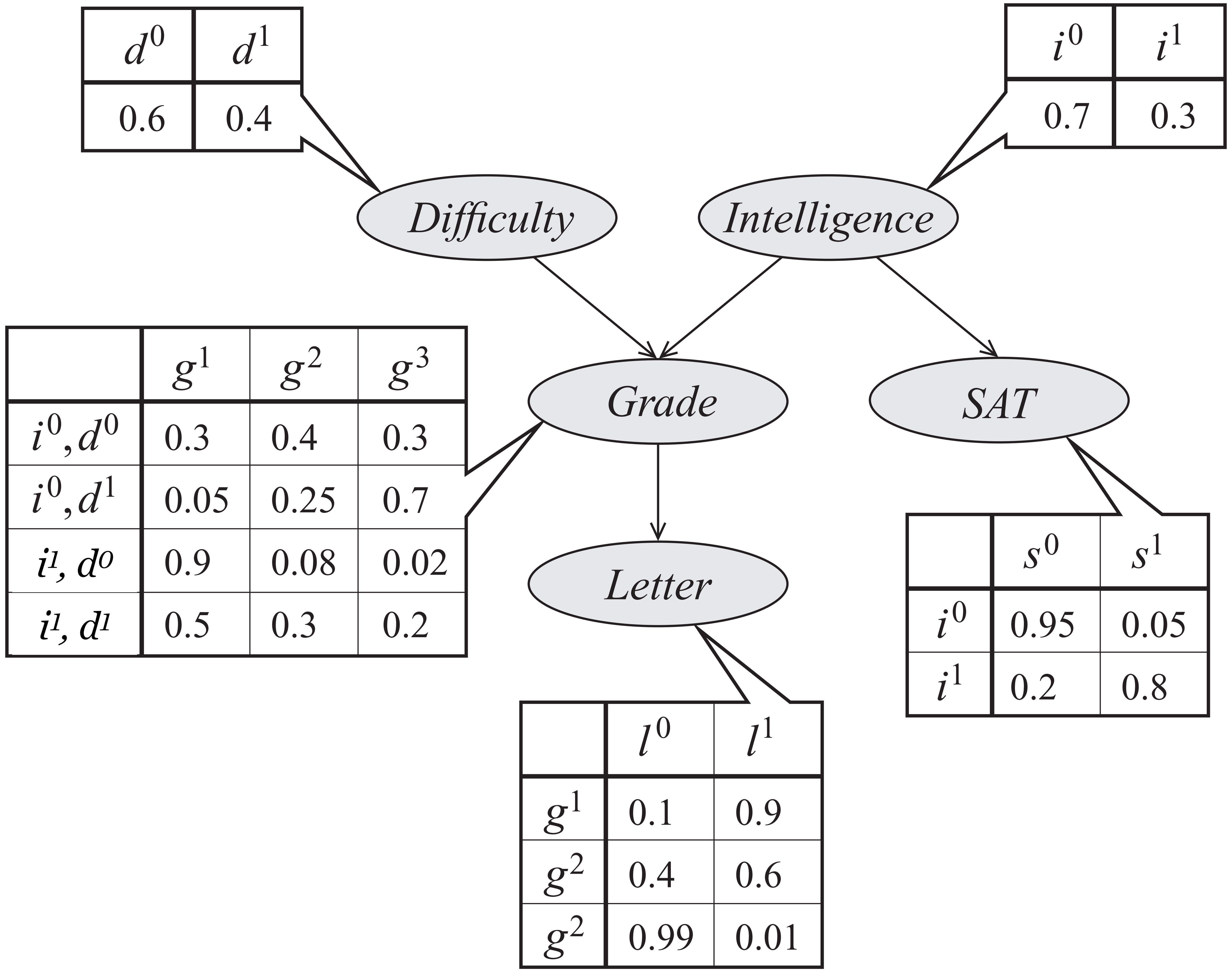
	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^2	0.99	0.01

Example

- Consider the Bayesian network here:

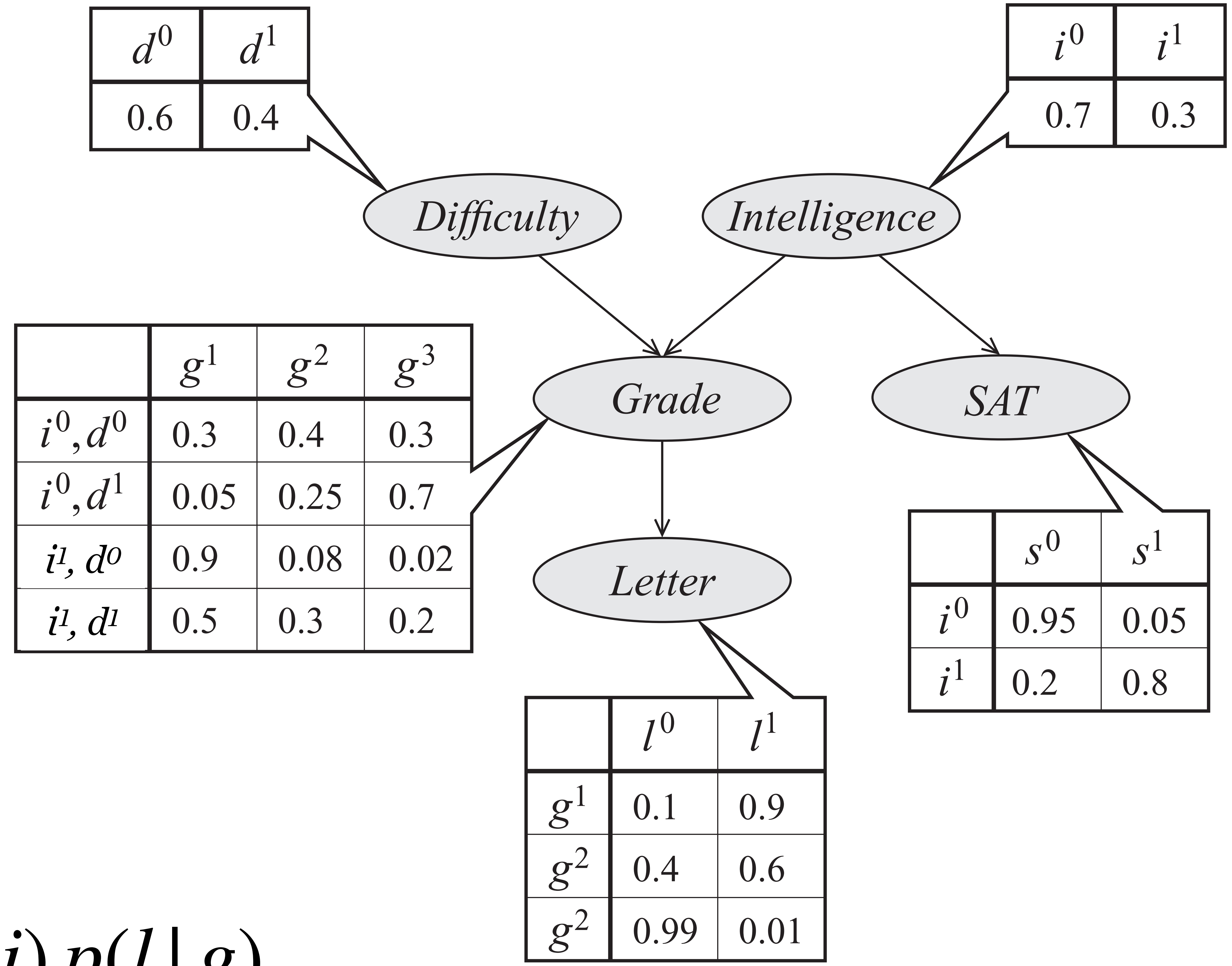
- What is its joint distribution?

$$p(x_1, \ldots, x_n) = \prod_{i \in V} p(x_i \,|\, \boldsymbol{x}_{\mathsf{Pa}(i)})$$



Example

- Consider the Bayesian network here:



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^2	0.99	0.01

- What is its joint distribution?

$$p(x_1, \ldots, x_n) = \prod_{i \in V} p(x_i \,|\, \boldsymbol{x}_{\mathsf{Pa}(i)})$$

$$p(d, i, g, s, l) = p(d)\, p(i)\, p(g \,|\, i, d)\, p(s \,|\, i)\, p(l \,|\, g)$$

Bayesian networks

- A Bayesian network is specified by a directed acyclic graph $G = (V, E)$ with:
  1. One node $i \in V$ for each random variable X_i
  2. One conditional probability distribution (CDP) per node, $p(x_i \,|\, \boldsymbol{x}_{\mathsf{Pa}(i)})$, specifying the variable's probability conditioned on its parents' values.
- Corresponds 1-1 with a particular factorisation of the joint distribution:

$$p(x_1, \ldots, x_n) = \prod_{i \in V} p(x_i \,|\, \boldsymbol{x}_{\mathsf{Pa}(i)})$$

BN is a legal distribution: $\sum P = 1$

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

BN is a legal distribution: $\sum P = 1$

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

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

BN is a legal distribution: $\sum P = 1$

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

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

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

BN is a legal distribution: $\sum P = 1$

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

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

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

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

BN is a legal distribution: $\sum P = 1$

$$
\begin{aligned}
\sum_{D,I,G,S,L} P(D,I,G,S,L) &= \sum_{D,I,G,S,L} P(D)\,P(I)\,P(G\,|\,I,D)\,P(S\,|\,I)\,P(L\,|\,G) \\
&= \sum_{D,I,G,S} P(D)\,P(I)\,P(G\,|\,I,D)\,P(S\,|\,I) \sum_{L} P(L\,|\,G) \\
&= \sum_{D,I,G,S} P(D)\,P(I)\,P(G\,|\,I,D)\,P(S\,|\,I) \\
&= \sum_{D,I,G} P(D)\,P(I)\,P(G\,|\,I,D) \sum_{S} P(S\,|\,I) \\
&= \sum_{D,I} P(D)\,P(I) \sum_{G} P(G\,|\,I,D)
\end{aligned}
$$

P factorises over G

- Let G be a graph over $X_1, \ldots, X_n$.
- P factorises over G if

$$P(X_1, \ldots, X_n) = \prod_i P(X_i \mid \mathrm{Pa}_G(X_i))$$

P factorises over G

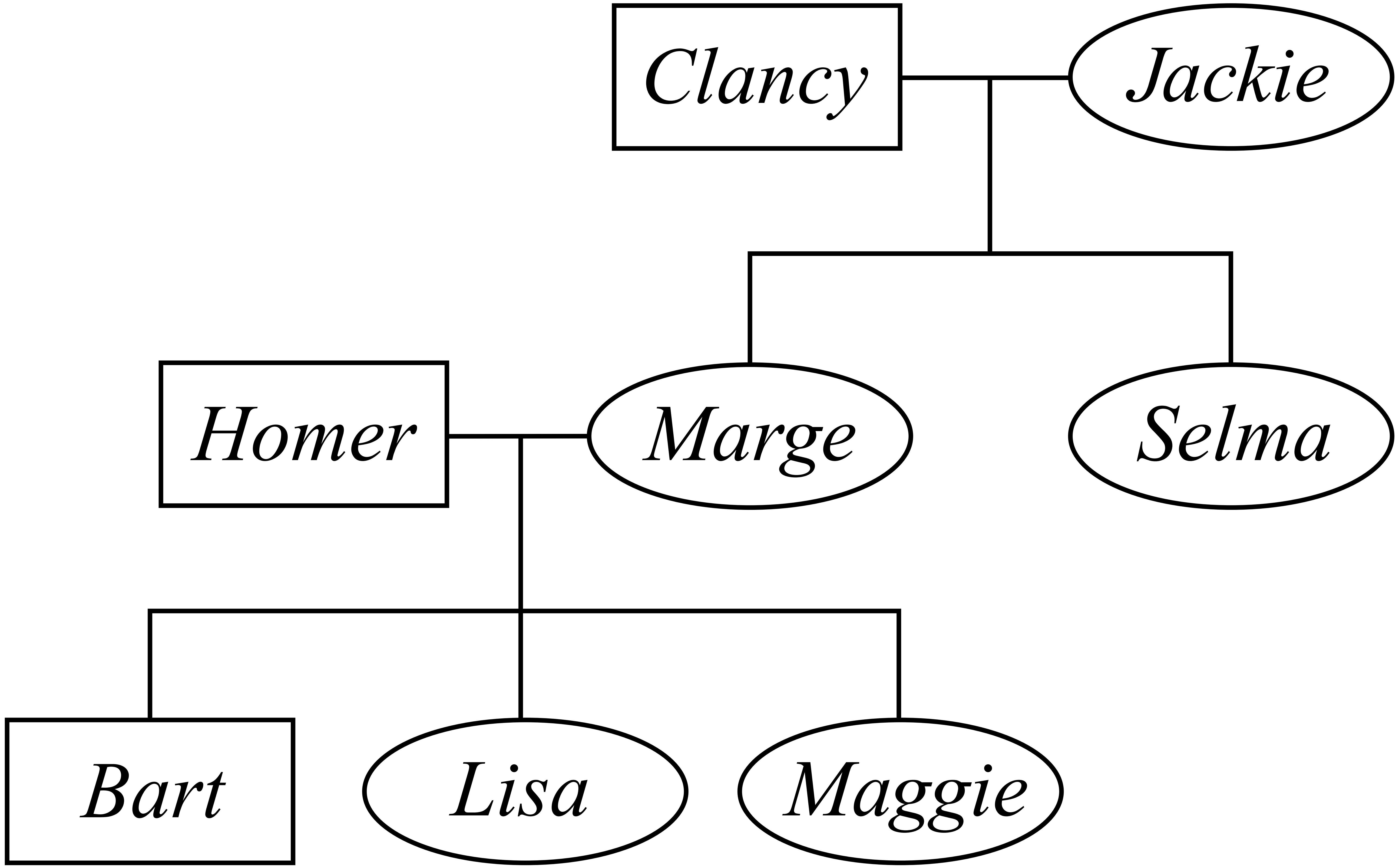
- Let G be a graph over $X_1, \ldots, X_n$.
- P factorises over G if

$$P(X_1, \ldots, X_n) = \prod_i P(X_i \mid \mathrm{Pa}_G(X_i))$$

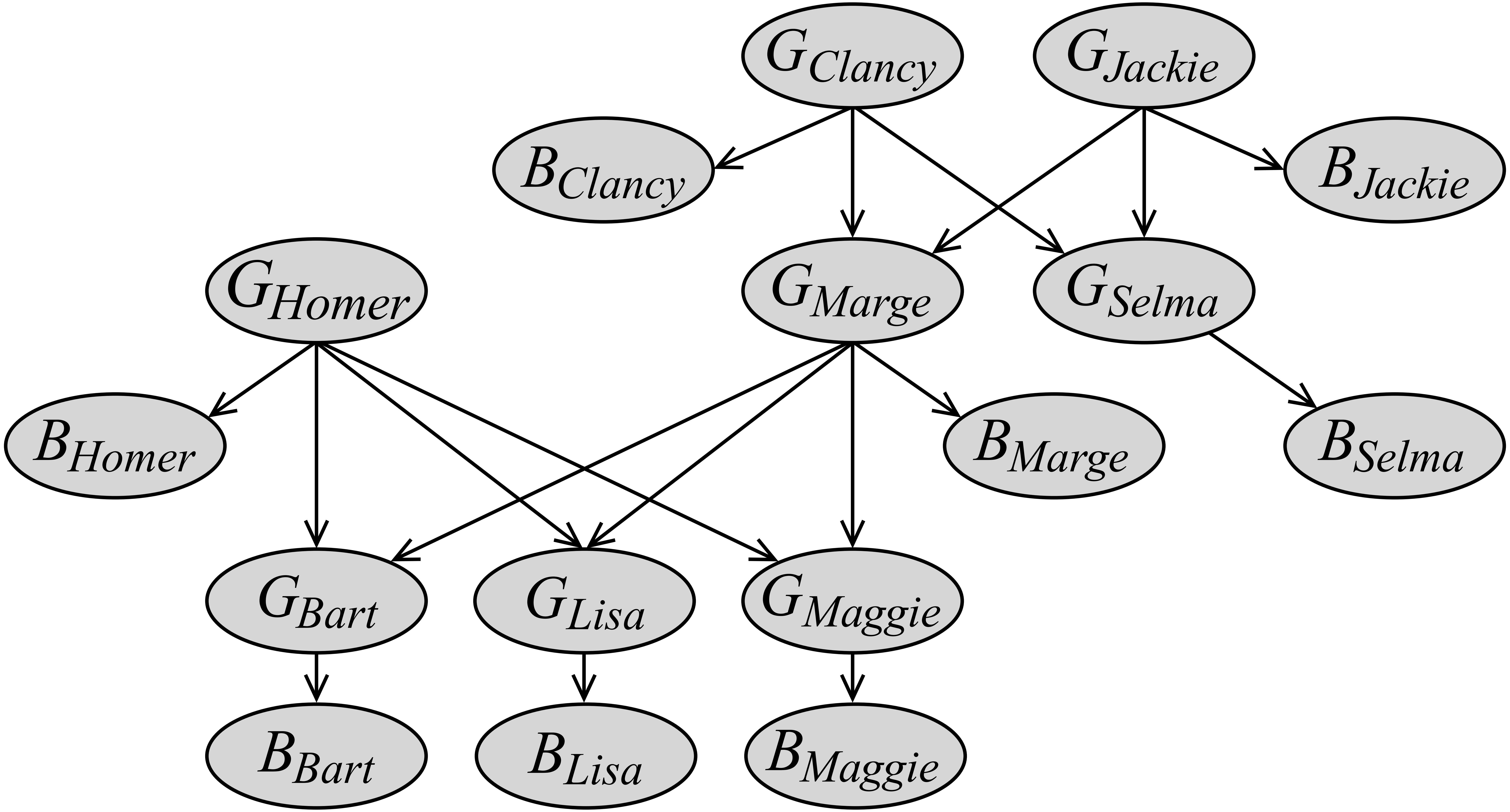
Genetic Inheritance

Genotype
AA, AB, AO, BO, BB, OO

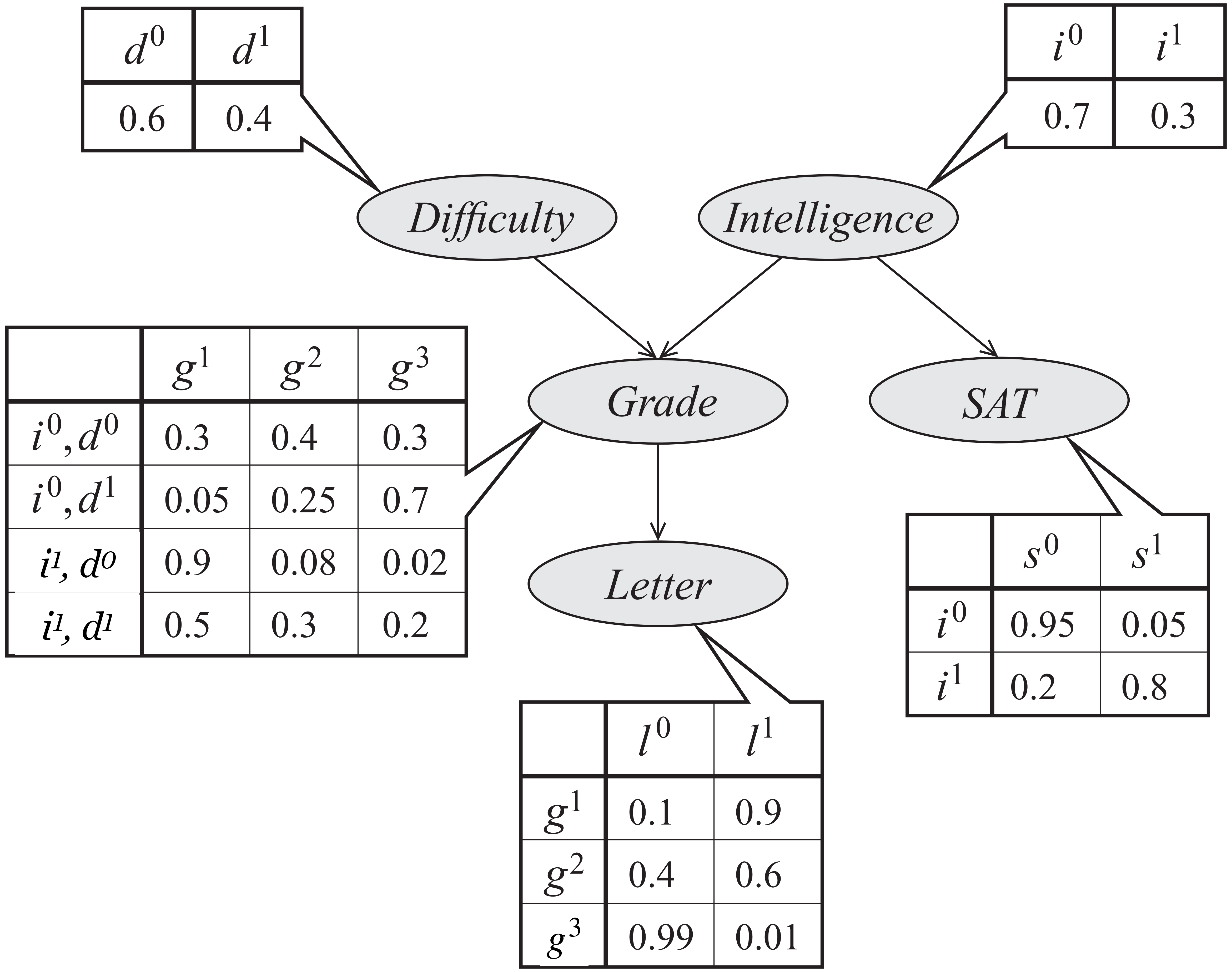
Phenotype
A, B, AB, O



BNs for Genetic Inheritance

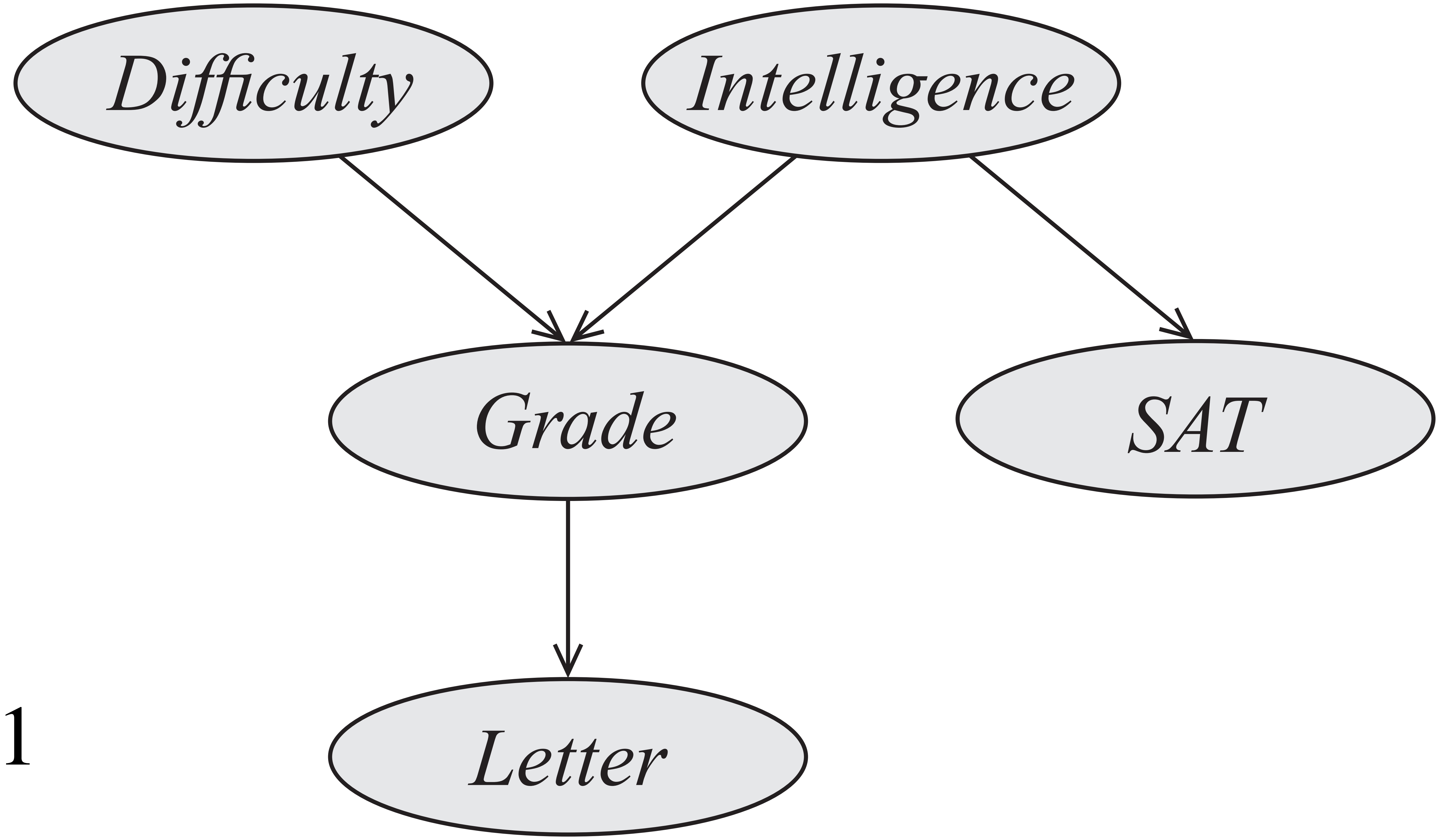


Reasoning Patterns



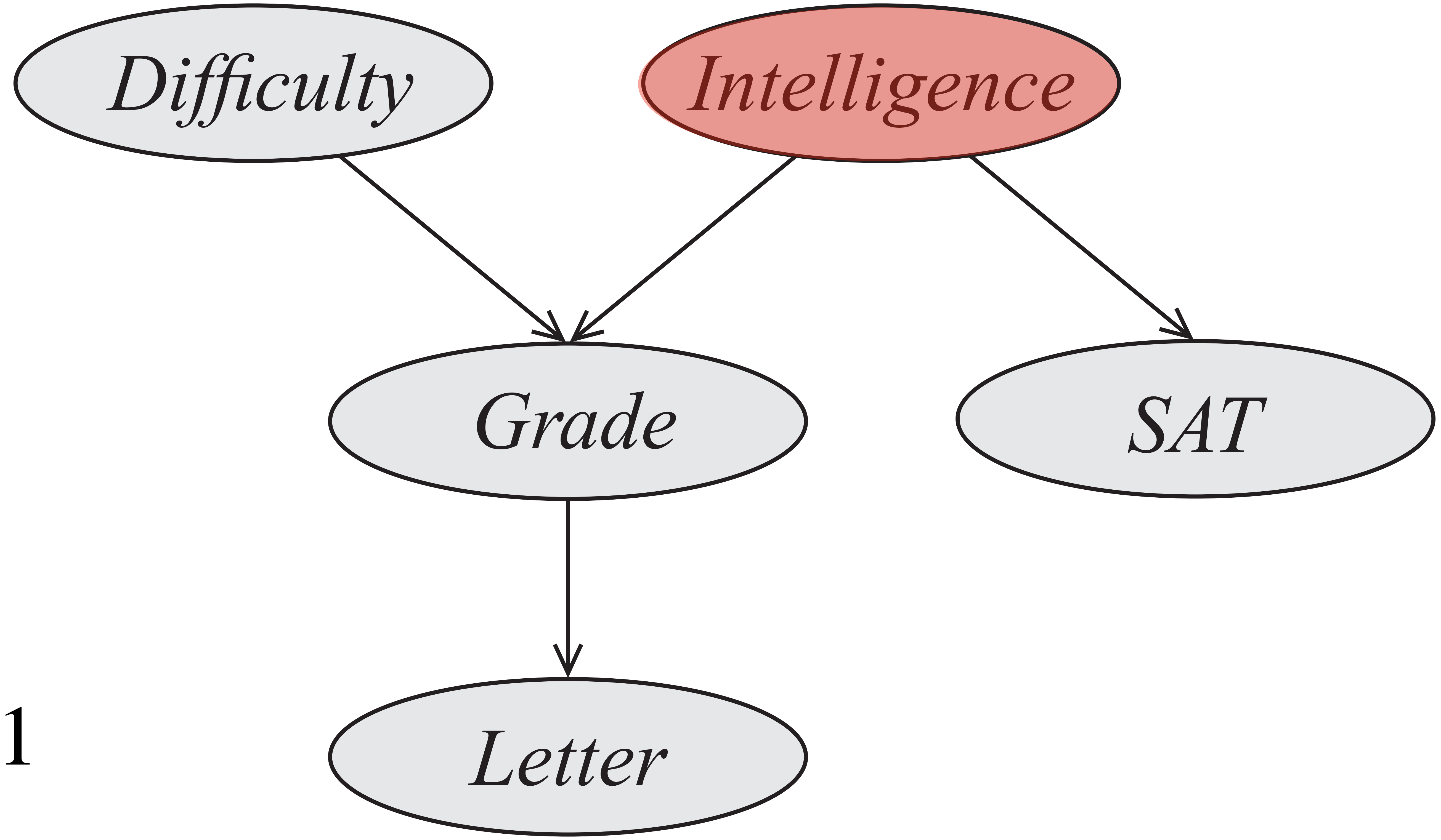
Difficulty
Intelligence
Grade
SAT
Letter
d^0 d^1
0.6 0.4
i^0 i^1
0.7 0.3
g^1 g^2 g^3
i^0, d^0 0.3 0.4 0.3
i^0, d^1 0.05 0.25 0.7
i^1, d^0 0.9 0.08 0.02
i^1, d^1 0.5 0.3 0.2
s^0 s^1
i^0 0.95 0.05
i^1 0.2 0.8
l^0 l^1
g^1 0.1 0.9
g^2 0.4 0.6
g^3 0.99 0.01

Causal reasoning



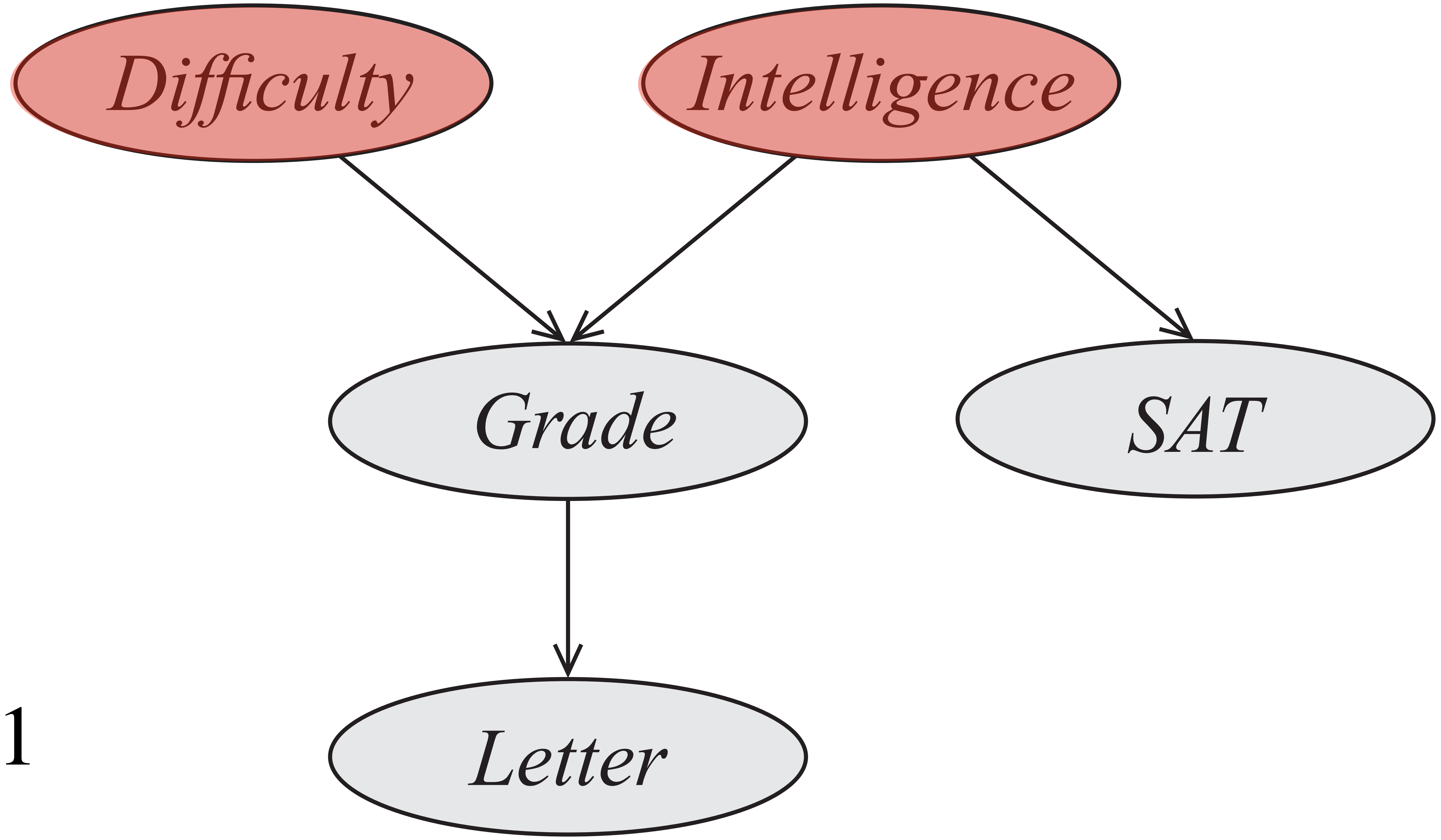
- $P(l^1) \approx 0.5$
- $P(l^1 \mid i^0) \approx 0.39$
- $P(l^1 \mid i^0, d^0) \approx 0.51$

Causal reasoning



- $P(l^1) \approx 0.5$
- $P(l^1 \mid i^0) \approx 0.39$
- $P(l^1 \mid i^0, d^0) \approx 0.51$

Causal reasoning

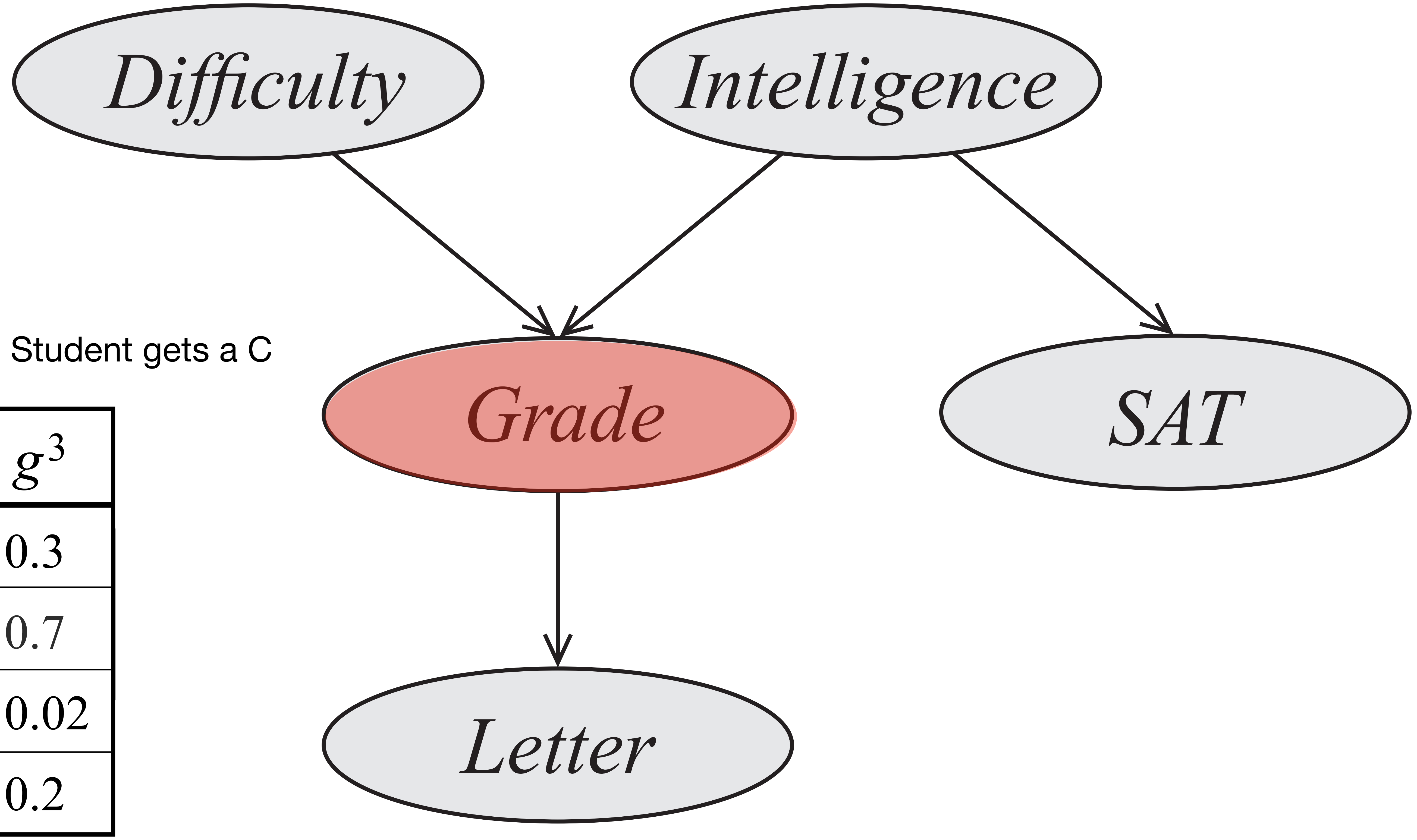


- $P(l^1) \approx 0.5$
- $P(l^1 \mid i^0) \approx 0.39$
- $P(l^1 \mid i^0, d^0) \approx 0.51$

Evidential reasoning

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

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

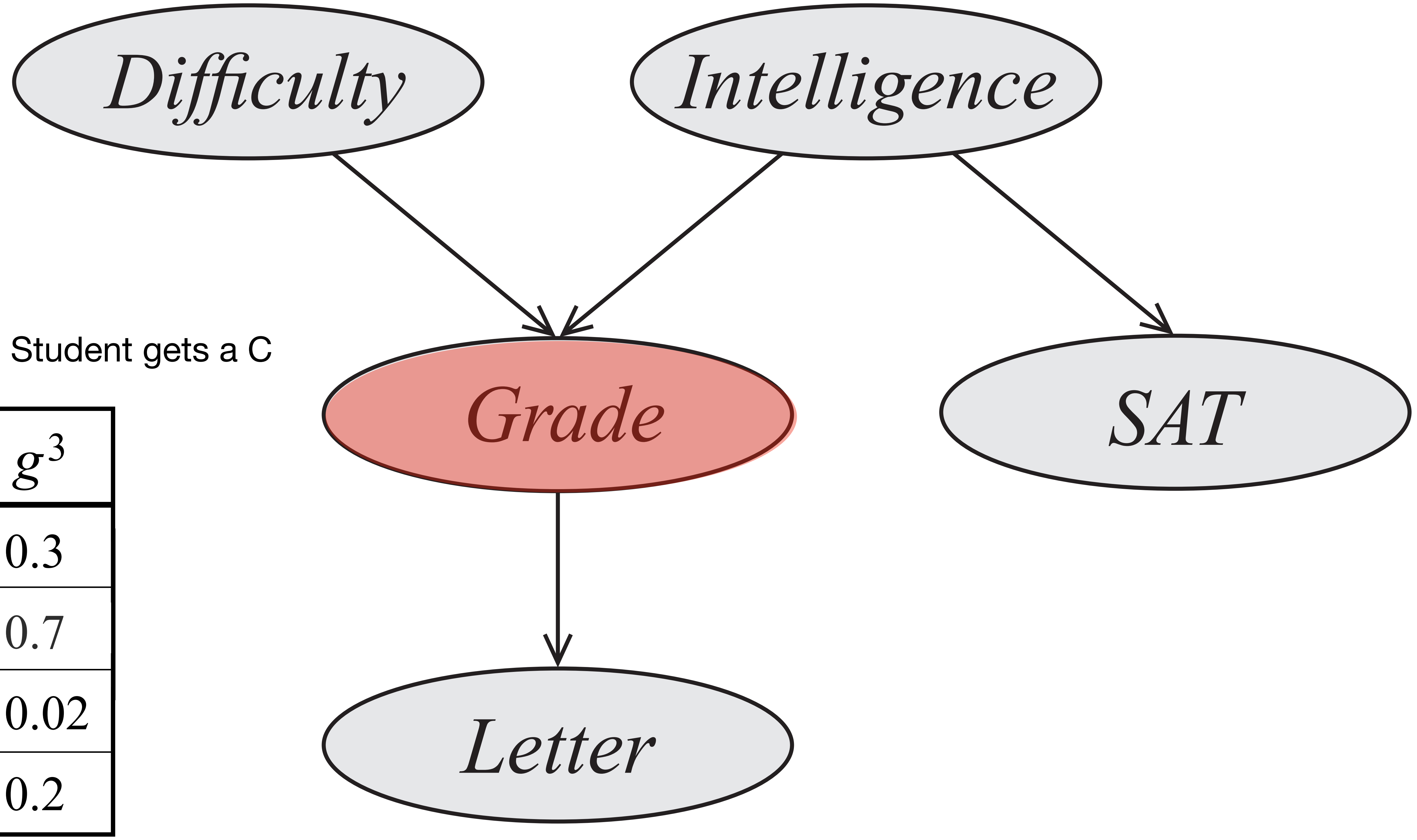


	g^1	g^2	g^3
i^0, d^0	0.3	0.4	0.3
i^0, d^1	0.05	0.25	0.7
i^1, d^0	0.9	0.08	0.02
i^1, d^1	0.5	0.3	0.2

Evidential reasoning

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

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



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

Intercausal reasoning

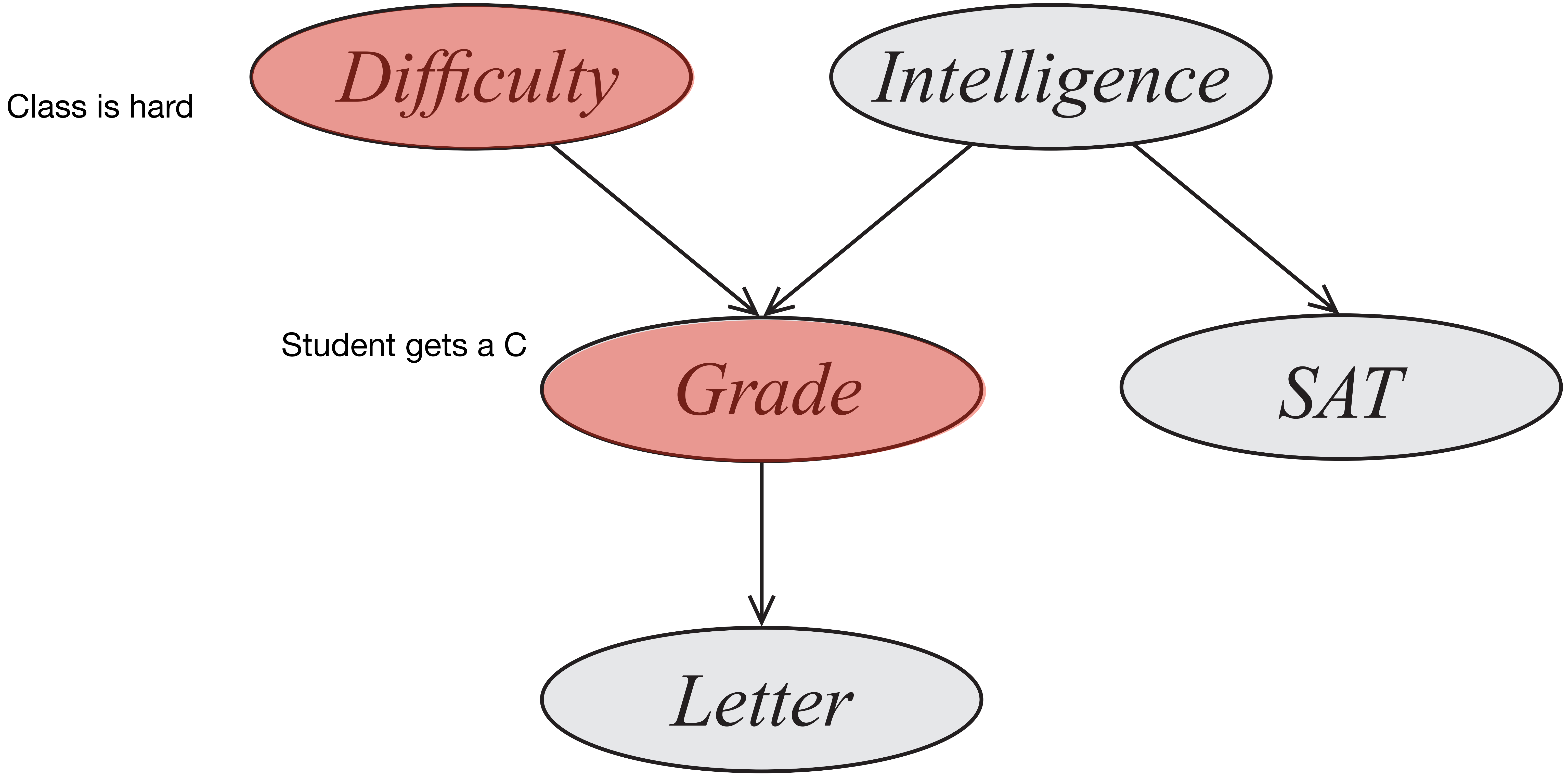
$$P(d^1) = 0.4$$

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

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

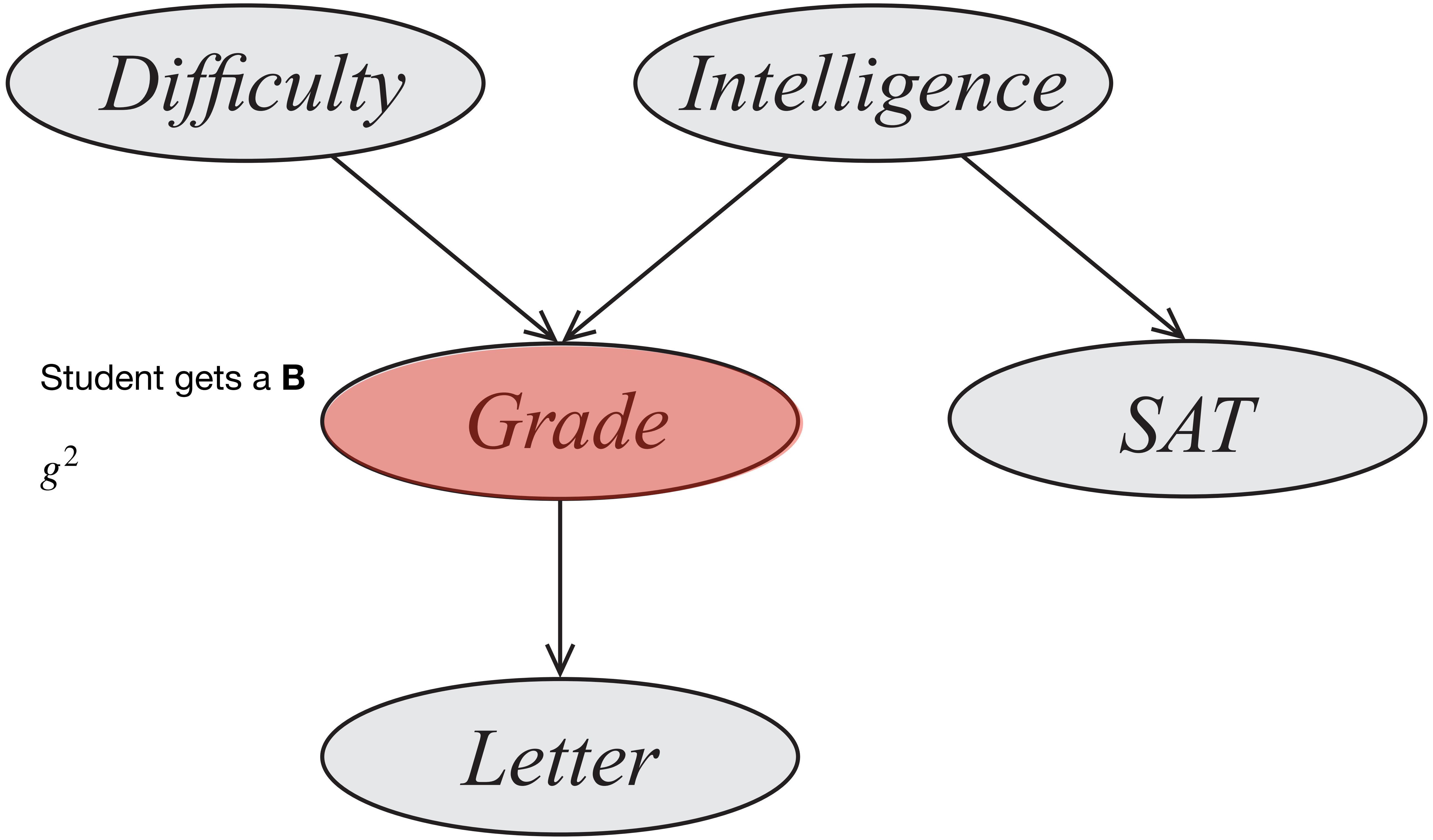
$$P(i^1 \mid g^3) \approx 0.08$$

$$P(i^1 \mid g^3, d^1) \approx 0.11$$



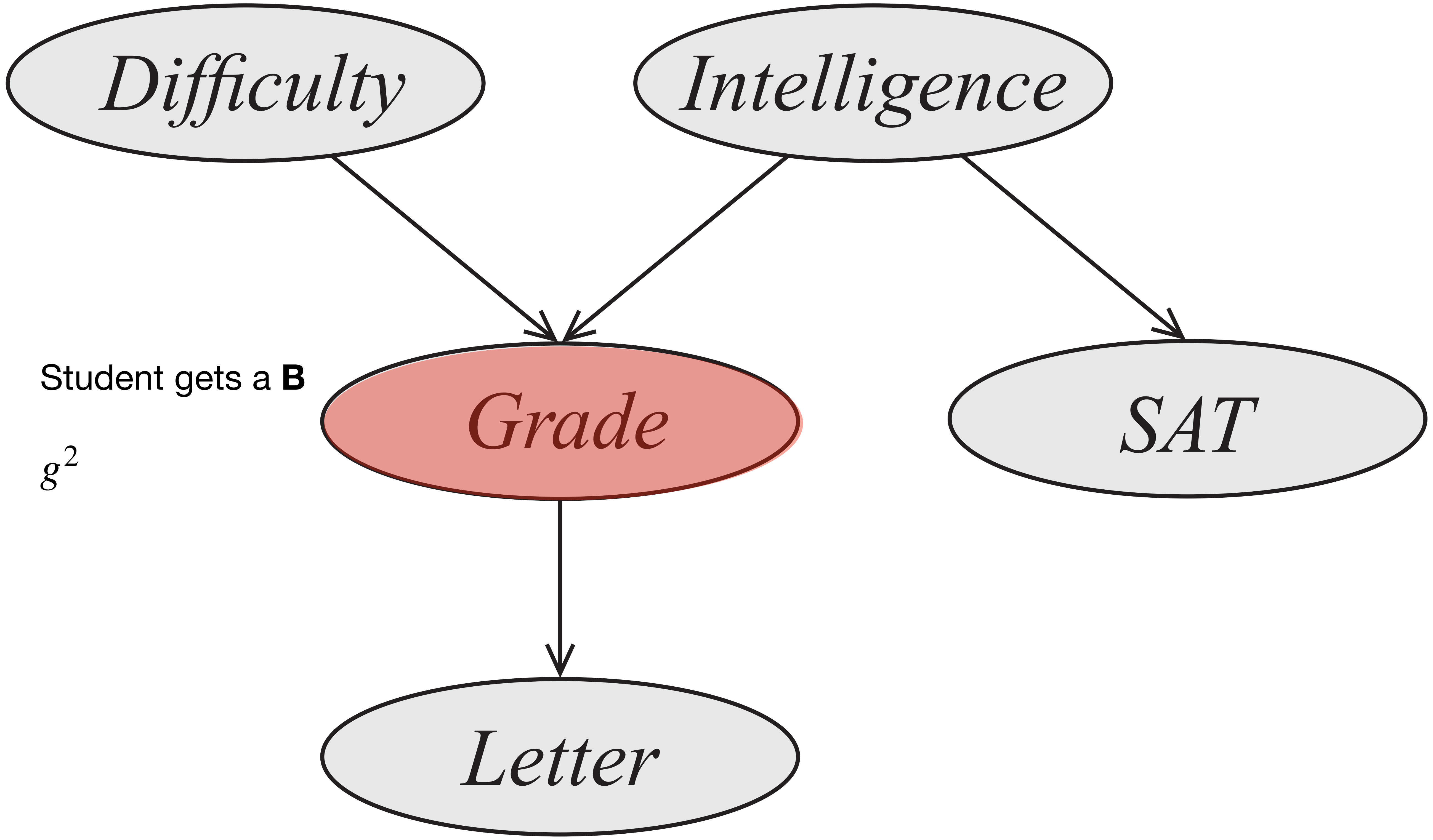
Intercausal reasoning

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



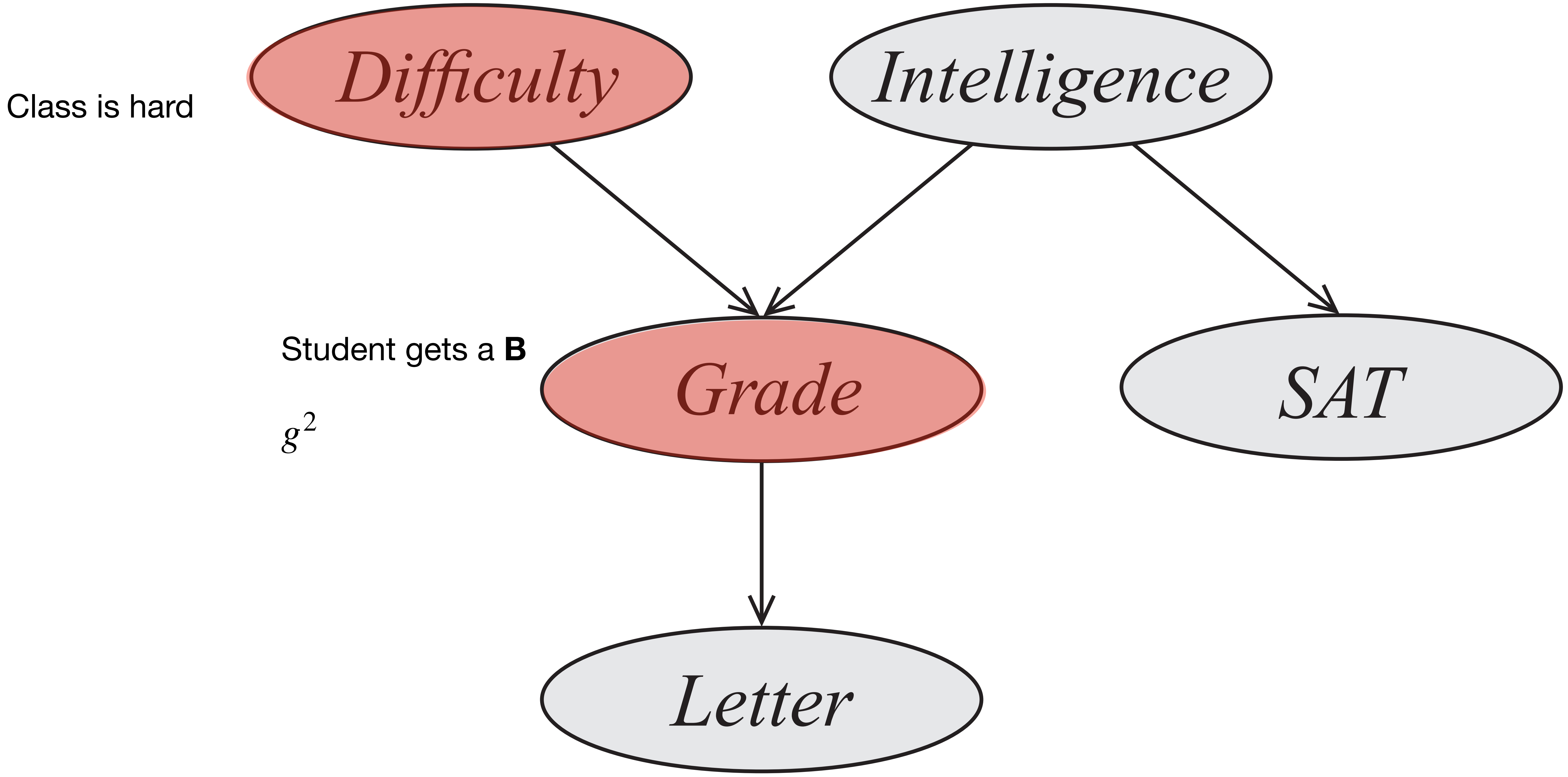
Intercausal reasoning

$$P(i^1) = 0.3$$
$$P(i^1 \mid g^2) \approx 0.175$$



Intercausal reasoning

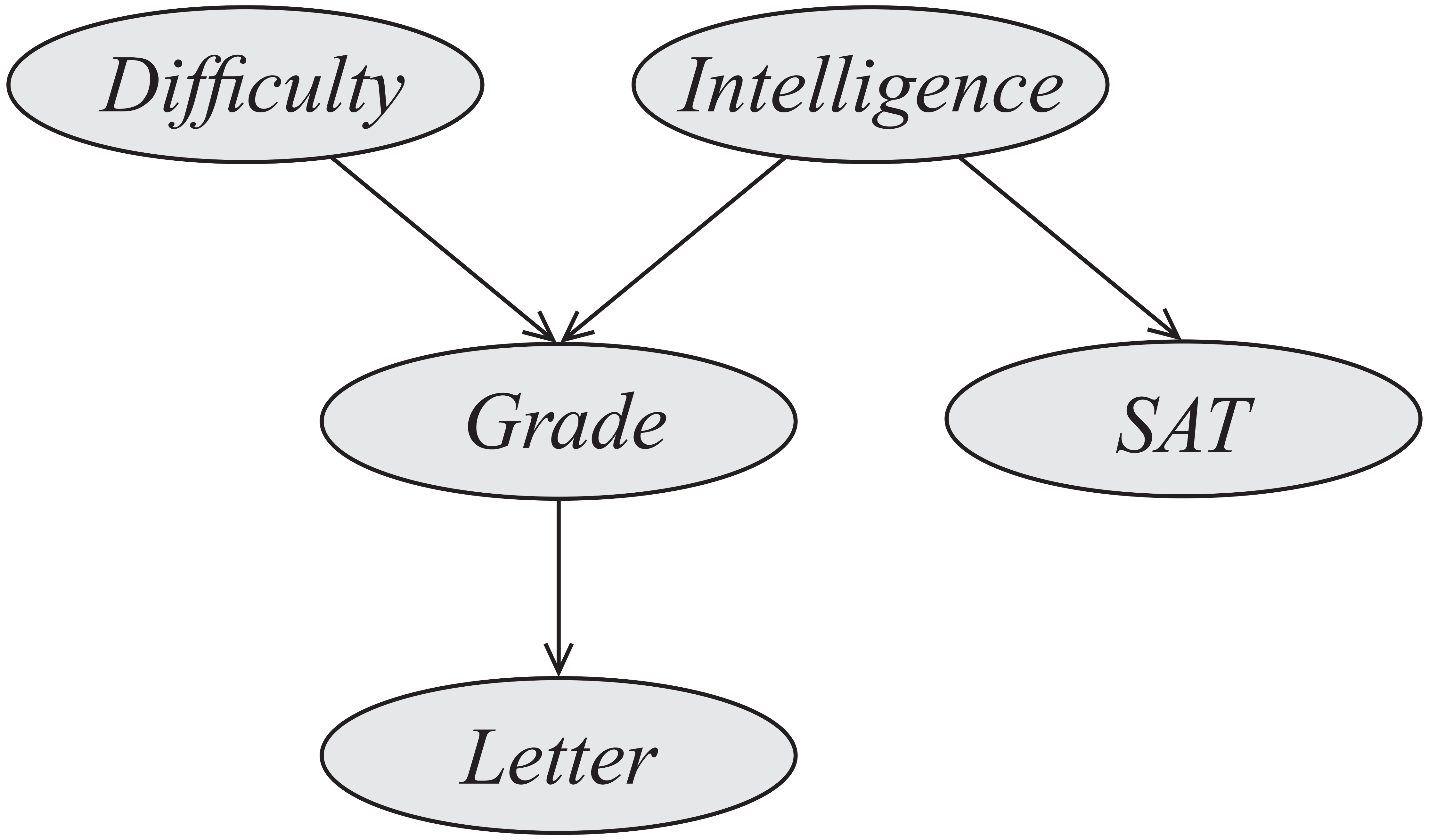
$$P(i^1) = 0.3$$
$$P(i^1 \mid g^2) \approx 0.175$$
$$P(i^1 \mid g^2, d^1) \approx 0.34$$



Flow of Probability Influence

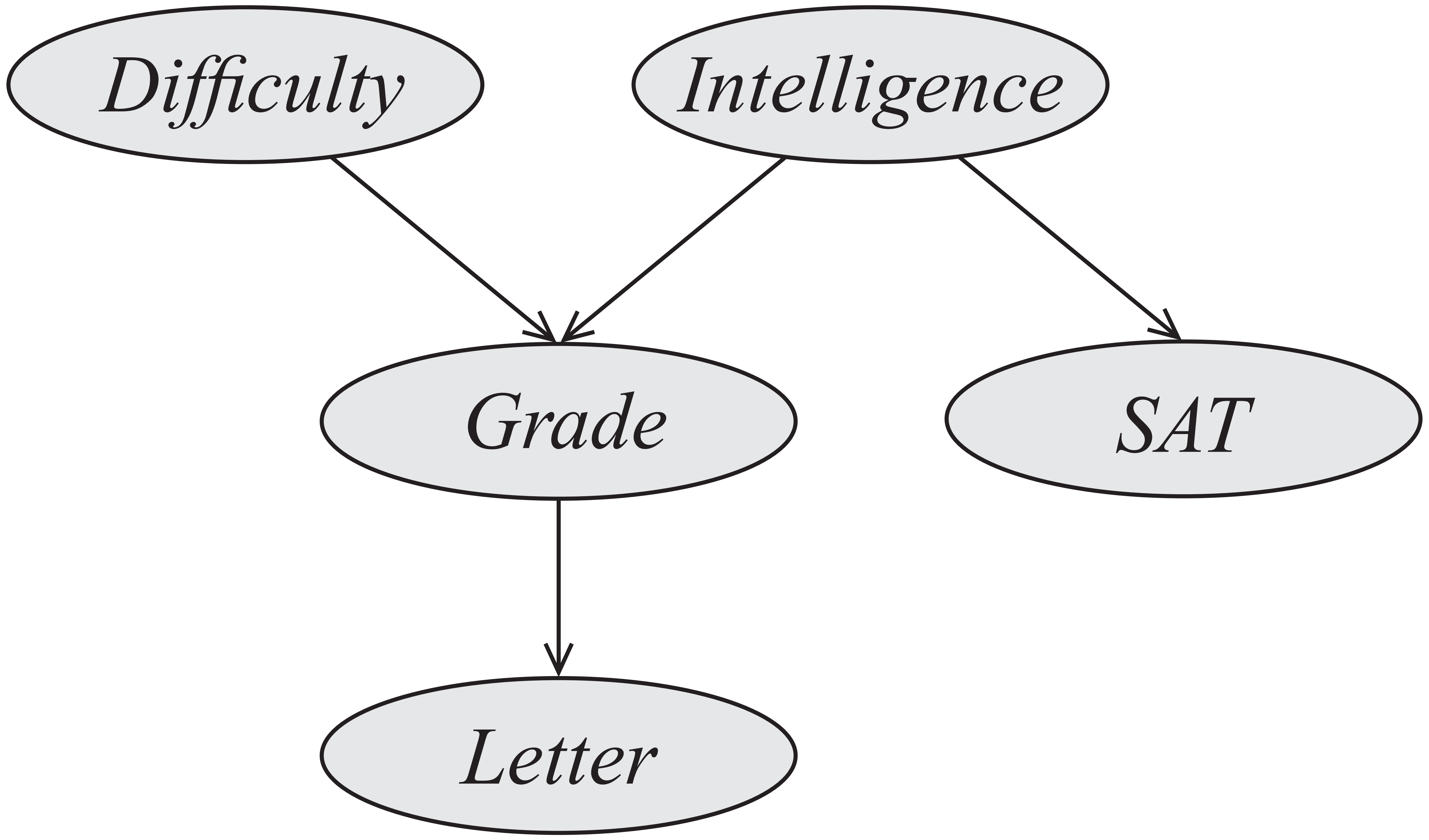
When can X influence Y

- $X \to Y$



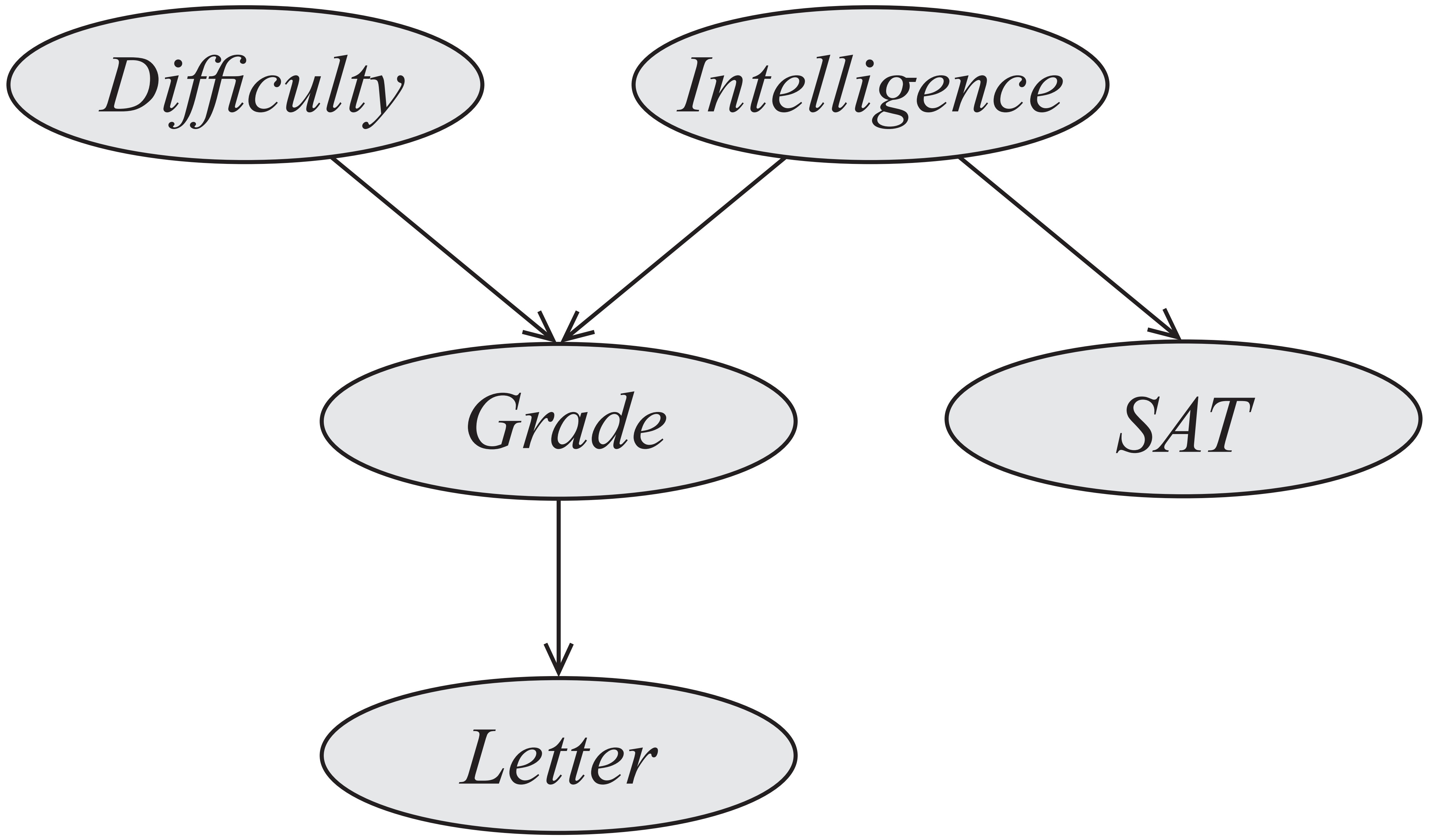
When can X influence Y

- $X \rightarrow Y$ ✓



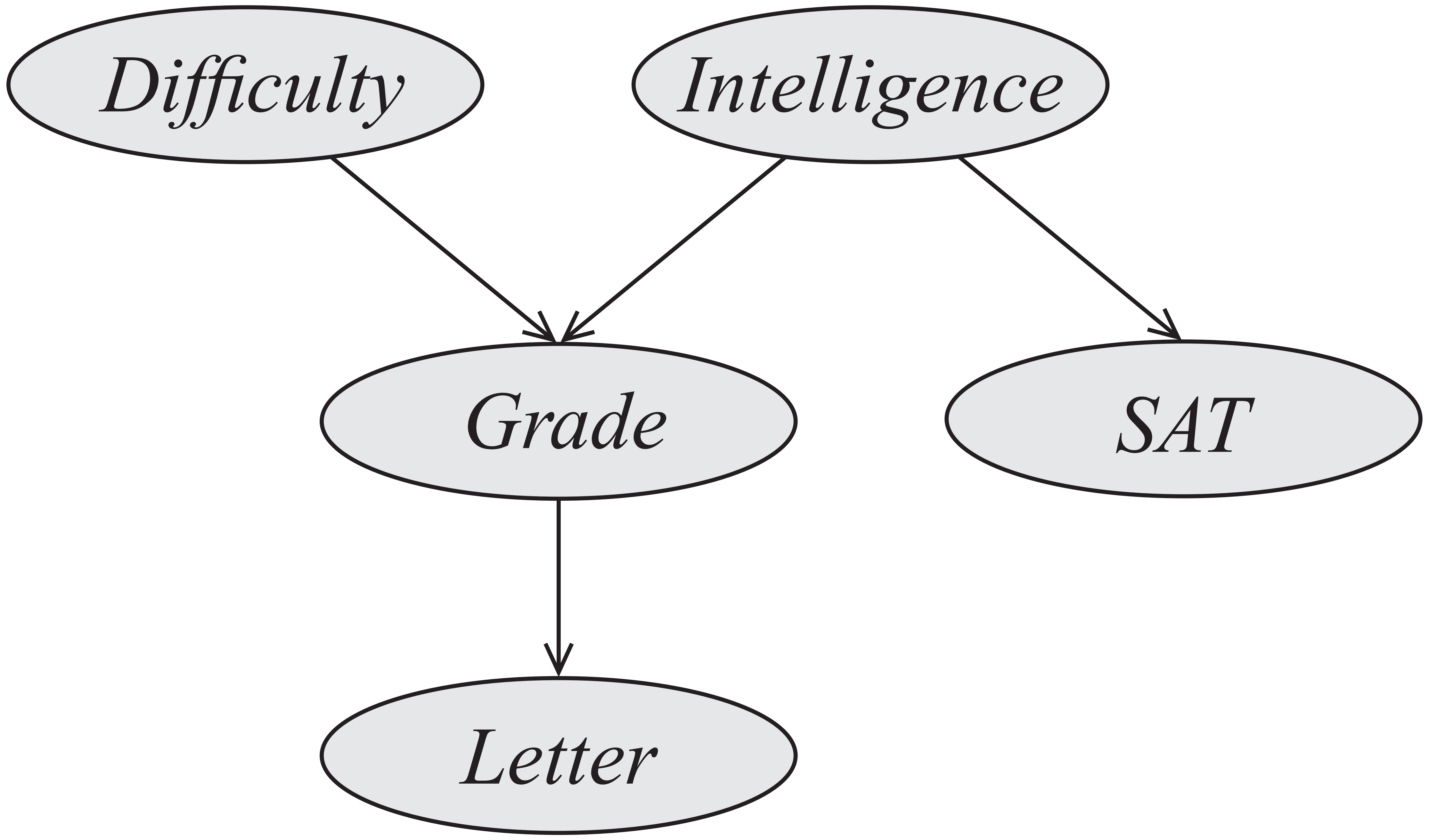
When can X influence Y

- $X \rightarrow Y$ ✓
- $X \leftarrow Y$



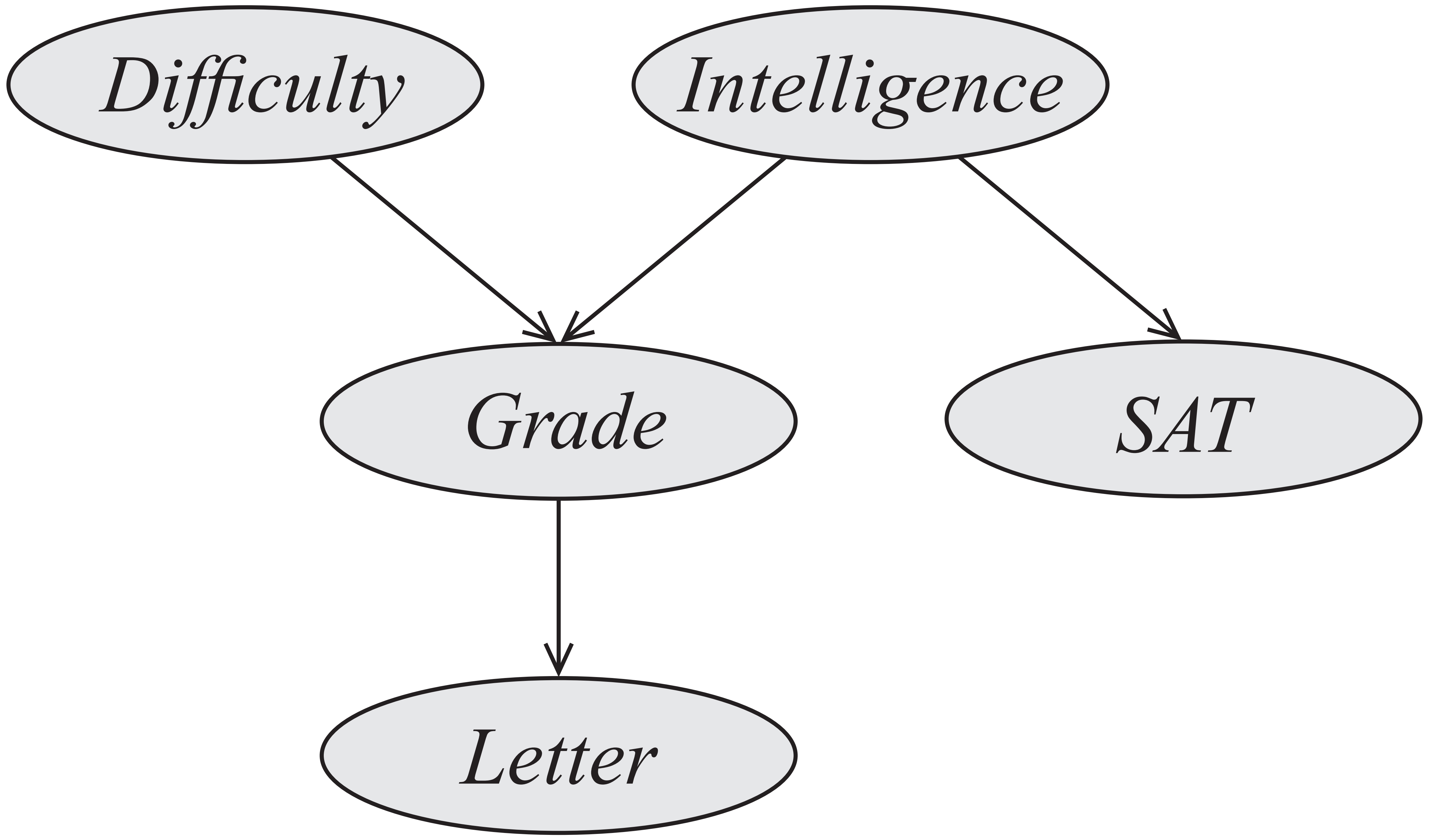
When can X influence Y

- $X \rightarrow Y$ ✓
- $X \leftarrow Y$ ✓



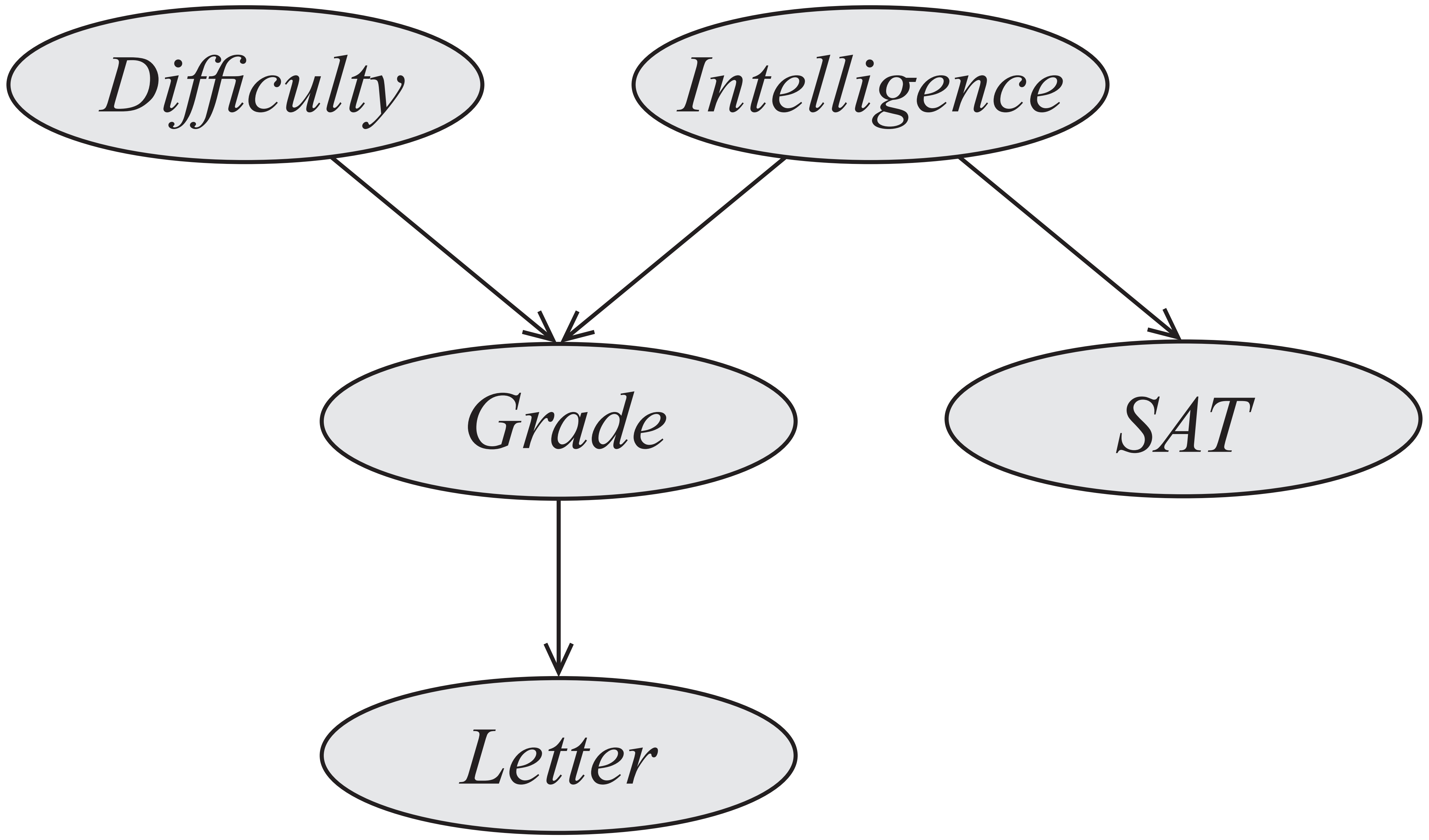
When can X influence Y

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



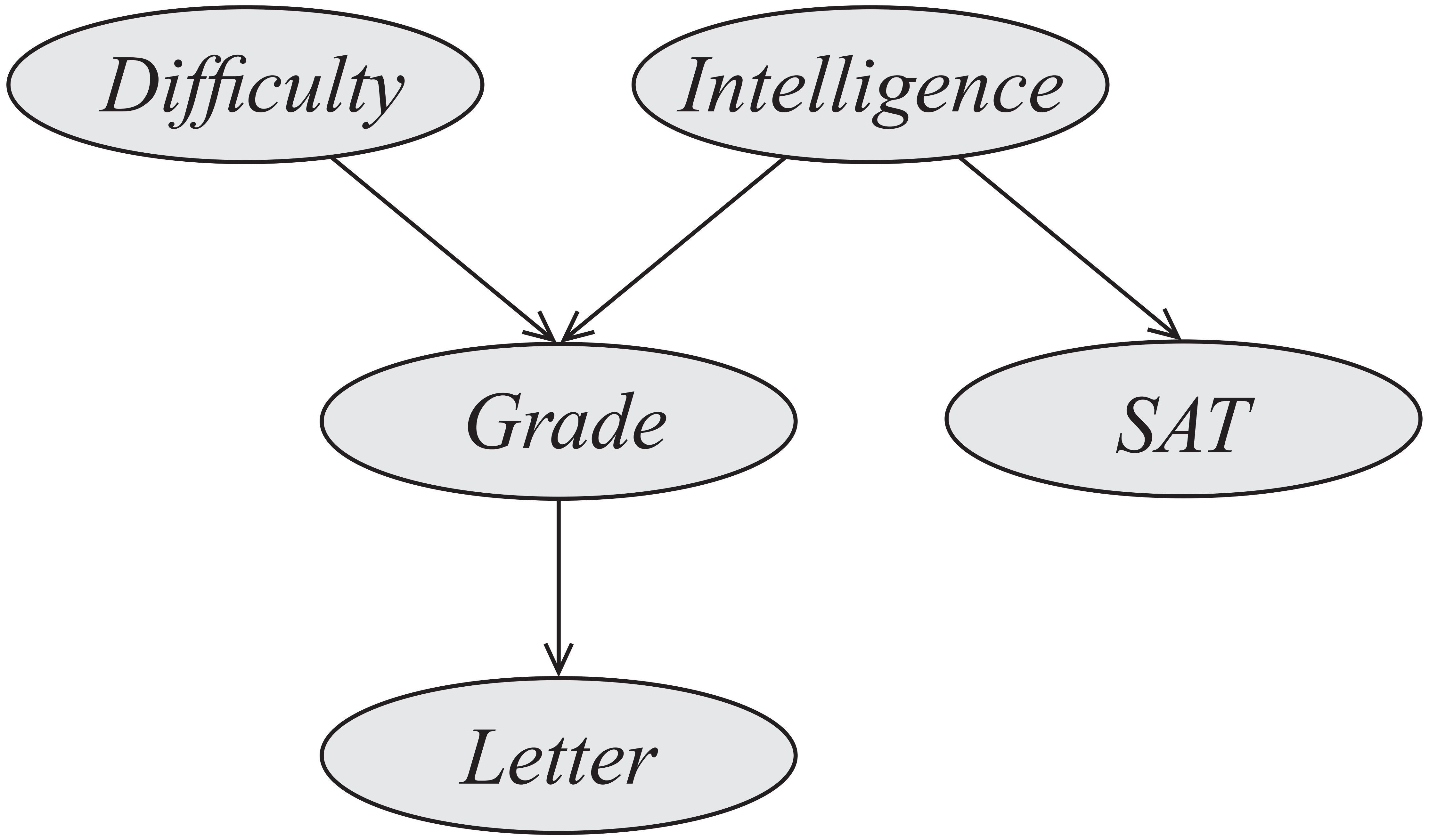
When can X influence Y

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



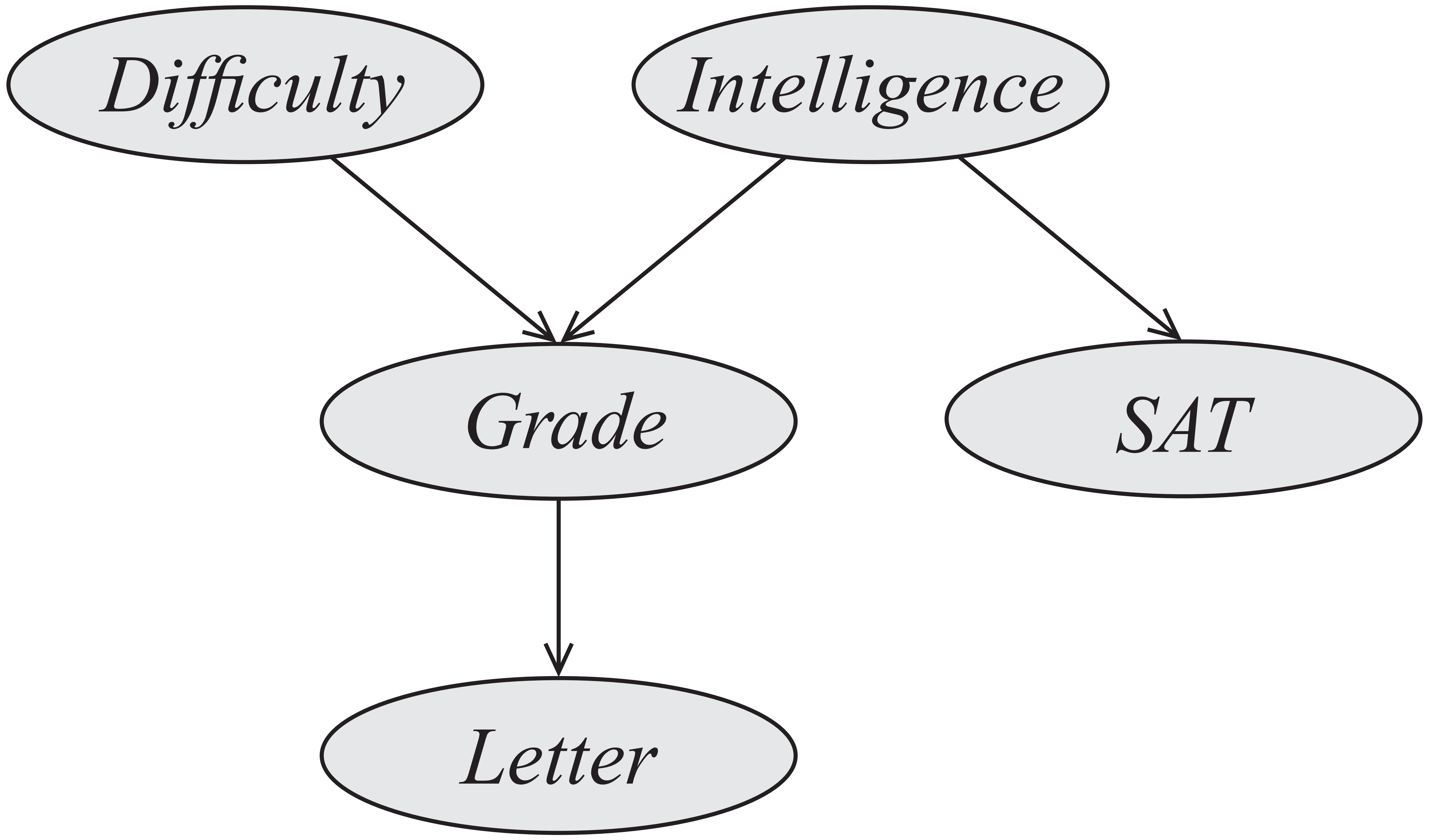
When can X influence Y

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



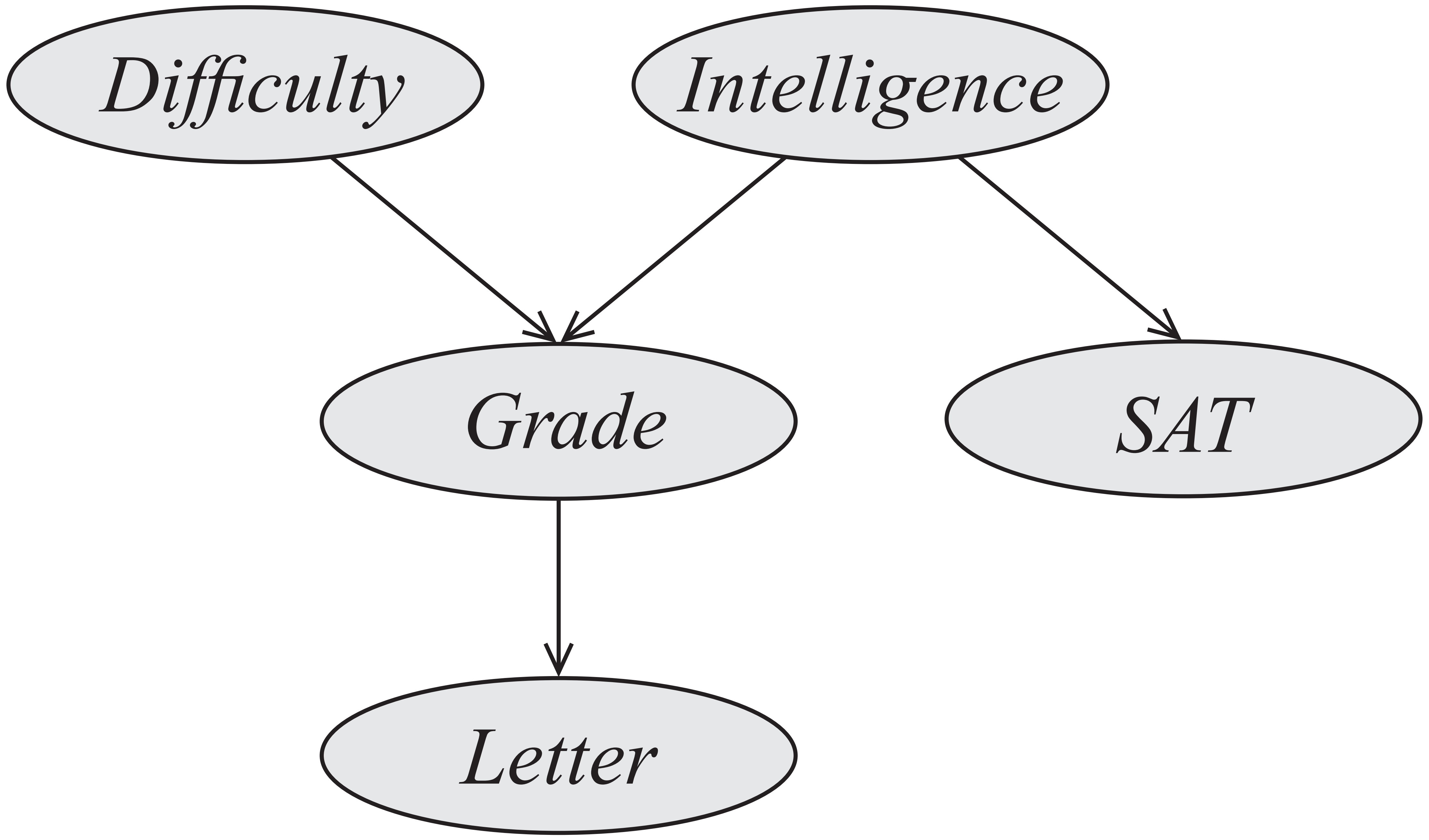
When can X influence Y

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



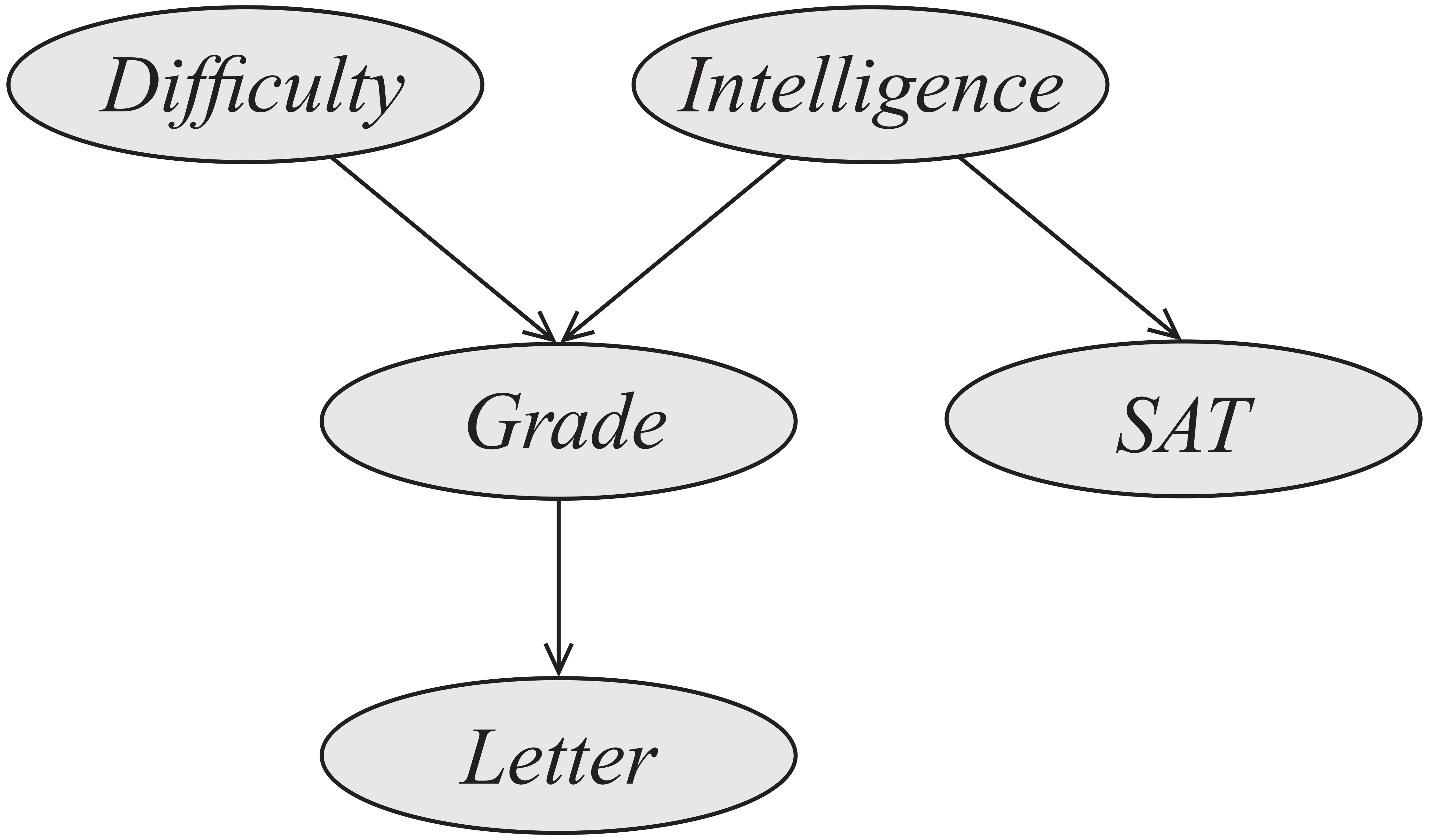
When can X influence Y

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



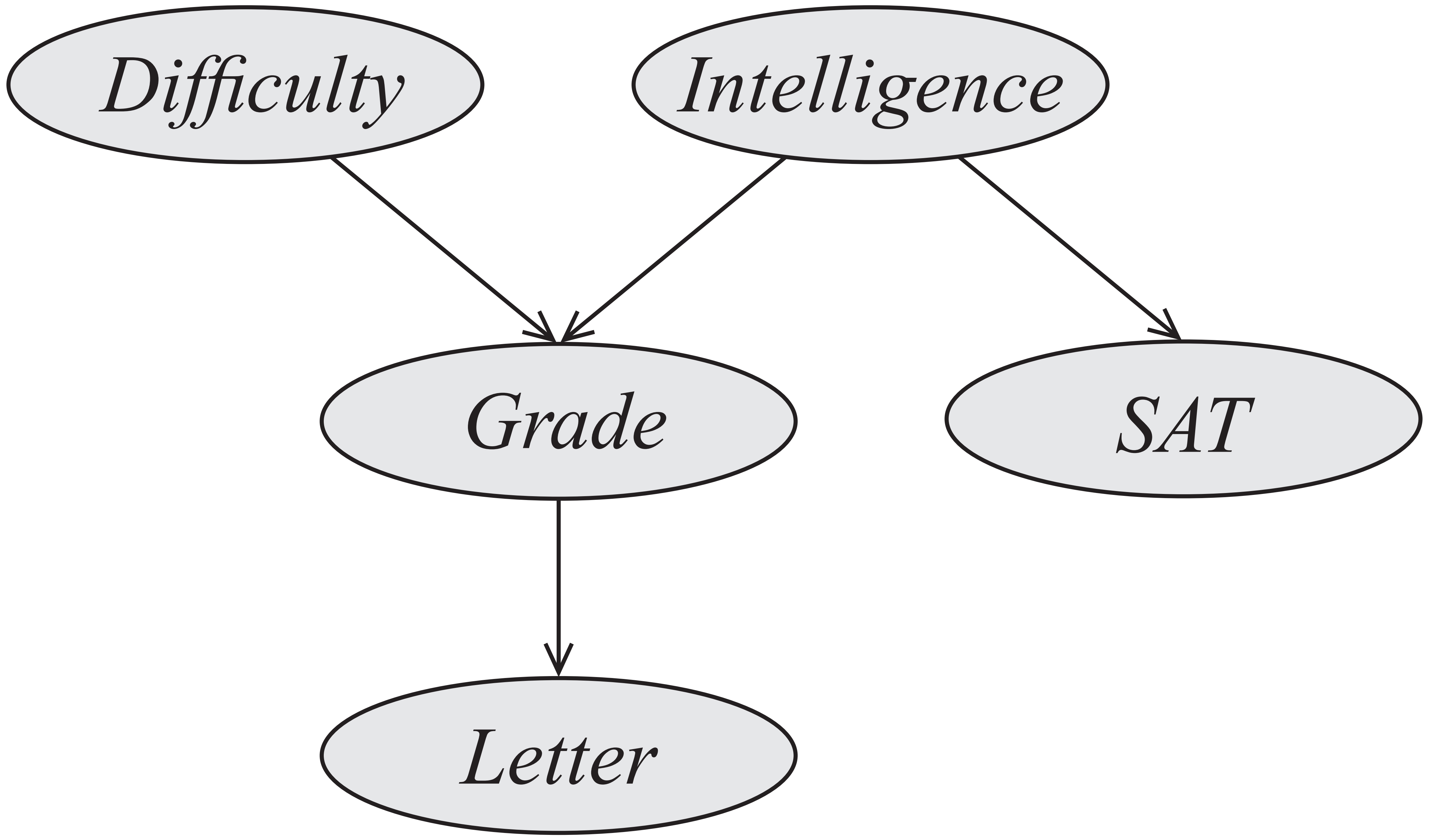
When can X influence Y

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



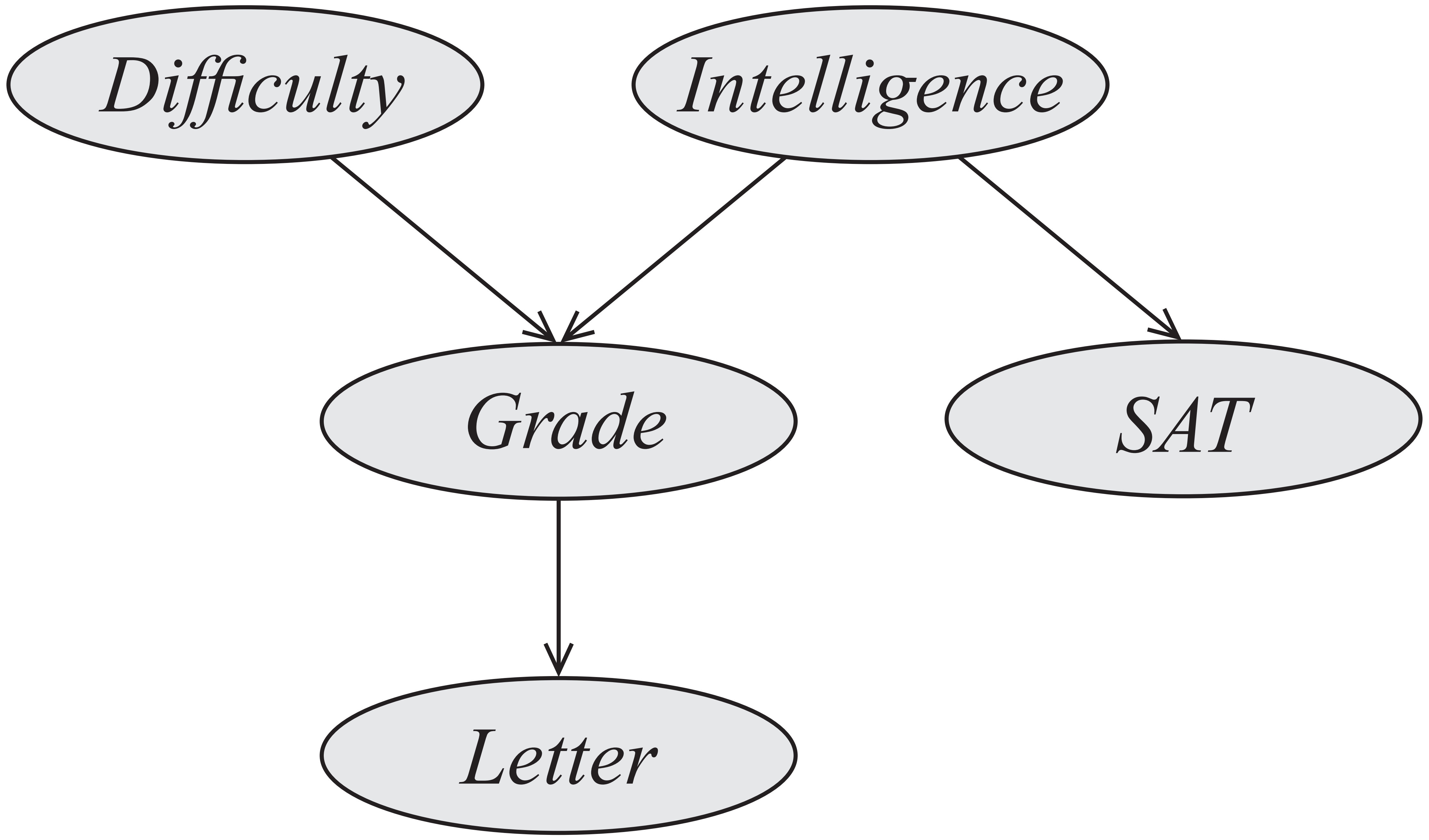
When can X influence Y

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



When can X influence Y

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



Active trails

A *trail* through a graph is a sequence of vertices connected by a sequence of edges, without using the same edge twice.

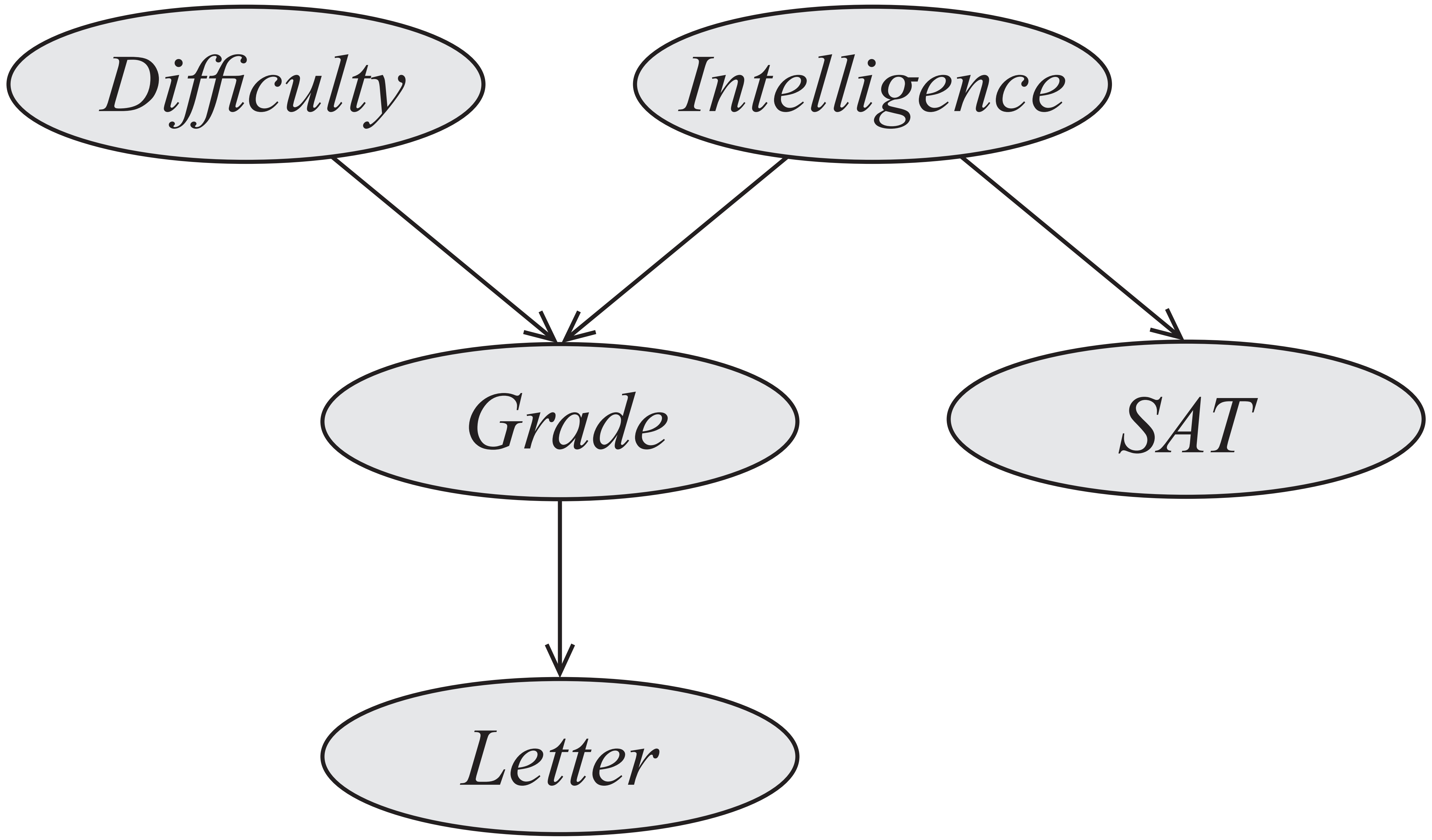
A trail $X_1 - \ldots - X_k$ is *active* if:

it has no v-structures $X_{i-1} \rightarrow X_i \leftarrow X_{i+1}$

When can X influence Y

given evidence about Z

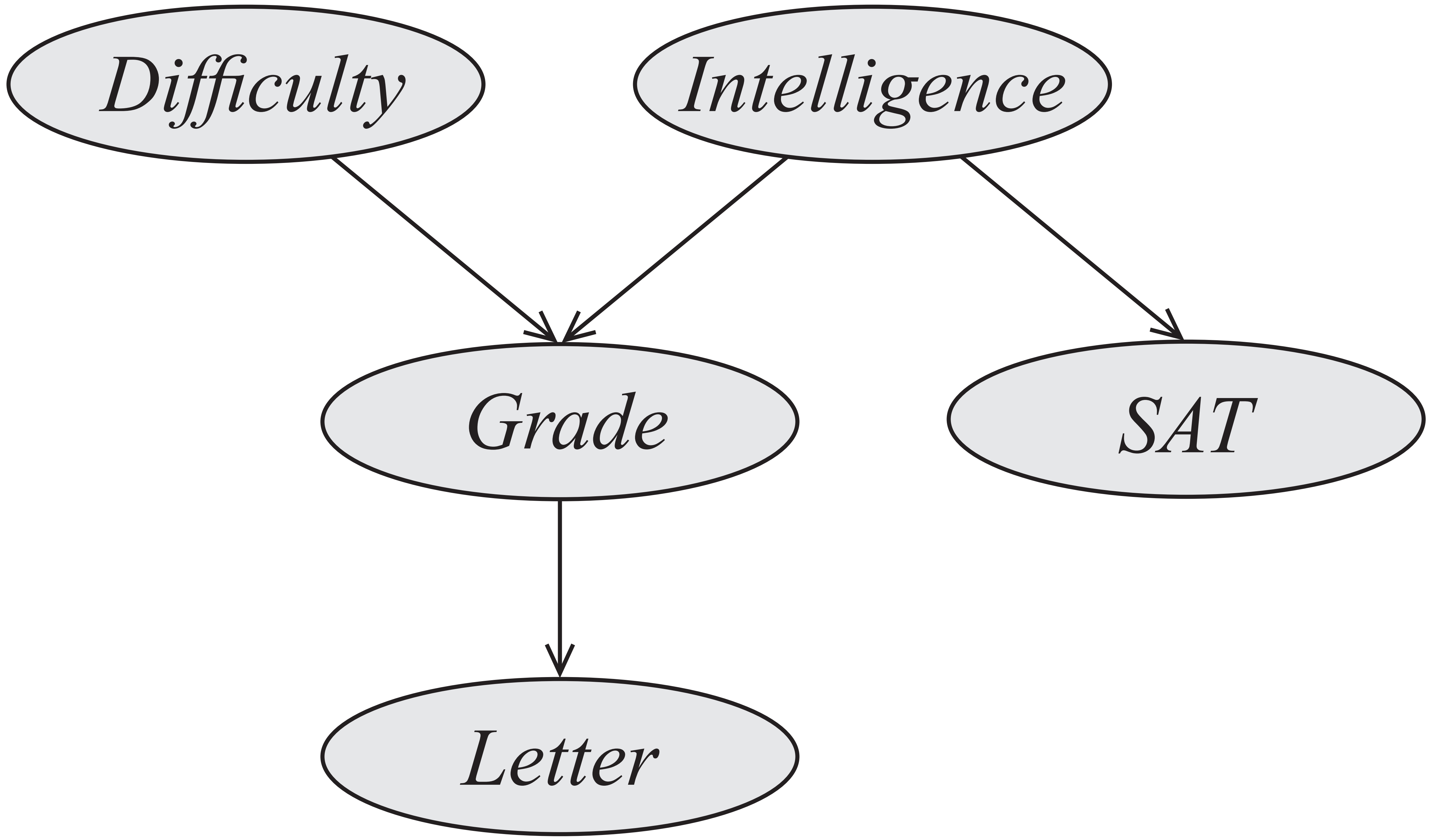
- $X \rightarrow Y$
- $X \leftarrow Y$



When can X influence Y

given evidence about Z

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

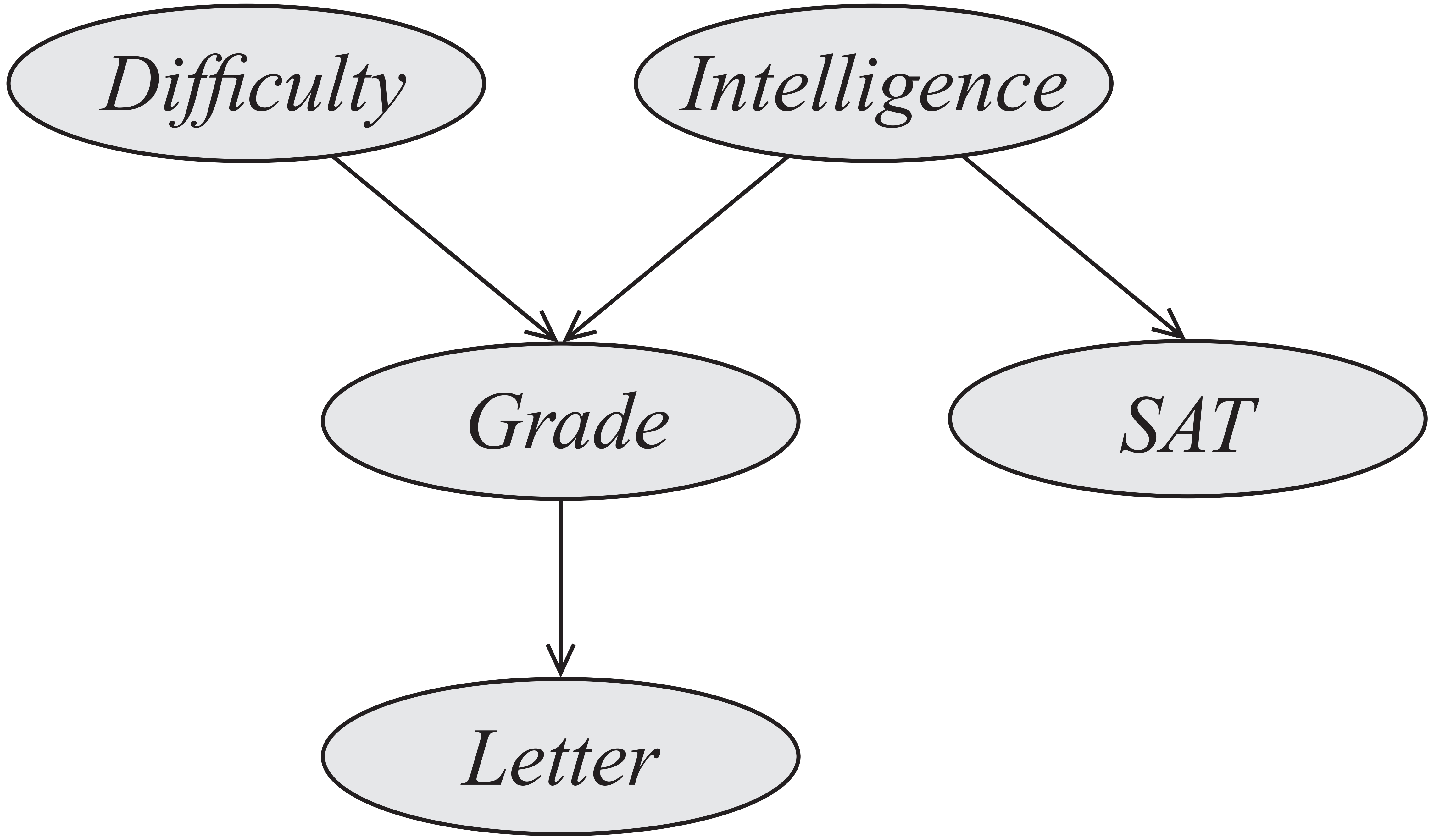


When can X influence Y

given evidence about Z

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

$W \notin \mathbf{Z}$	$W \in \mathbf{Z}$

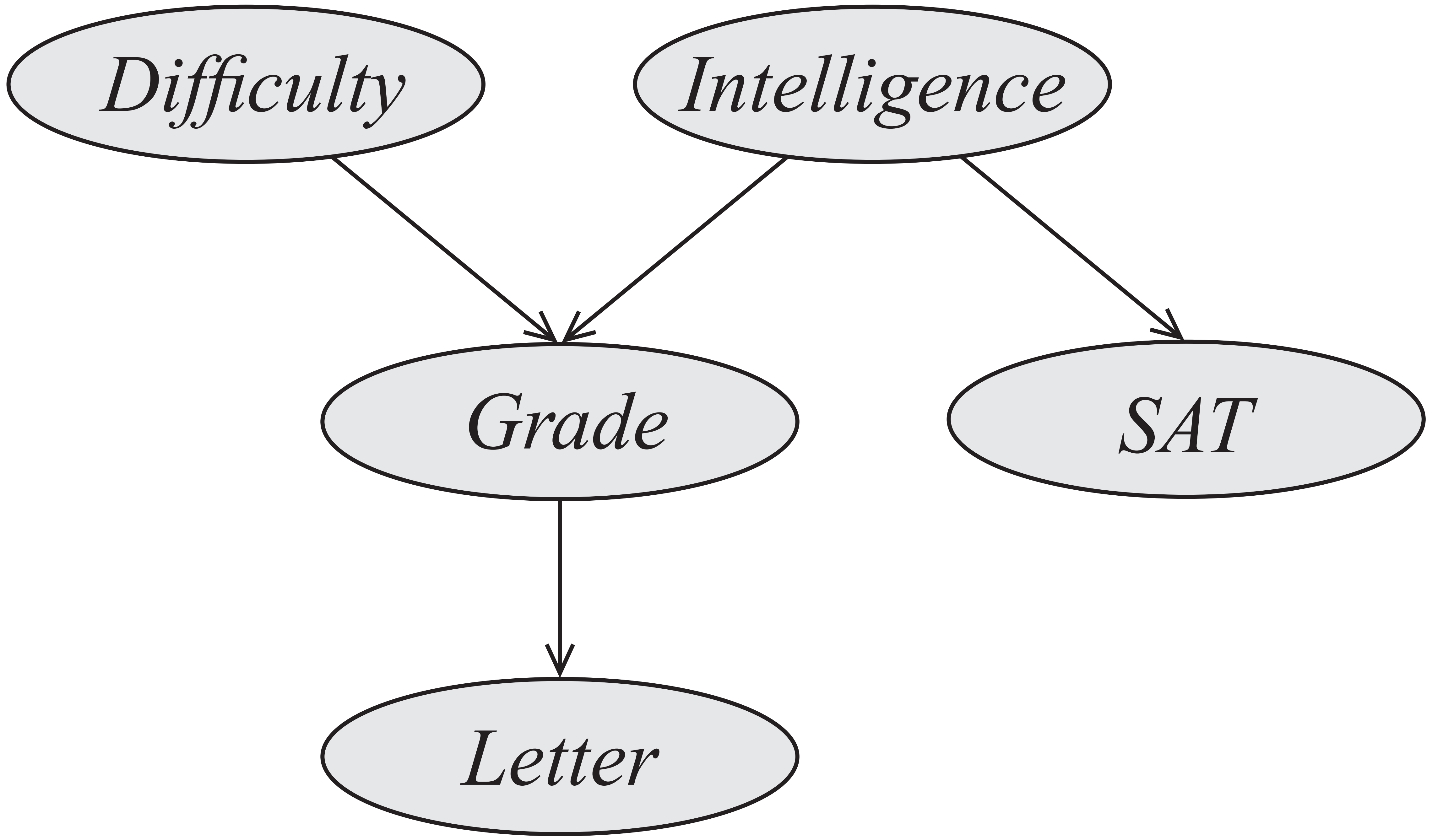


When can X influence Y

given evidence about Z

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

	$W \notin Z$	$W \in Z$
- $X \rightarrow W \rightarrow Y$	✓	
- $X \leftarrow W \leftarrow Y$		
- $X \leftarrow W \rightarrow Y$		
- $X \rightarrow W \leftarrow Y$		

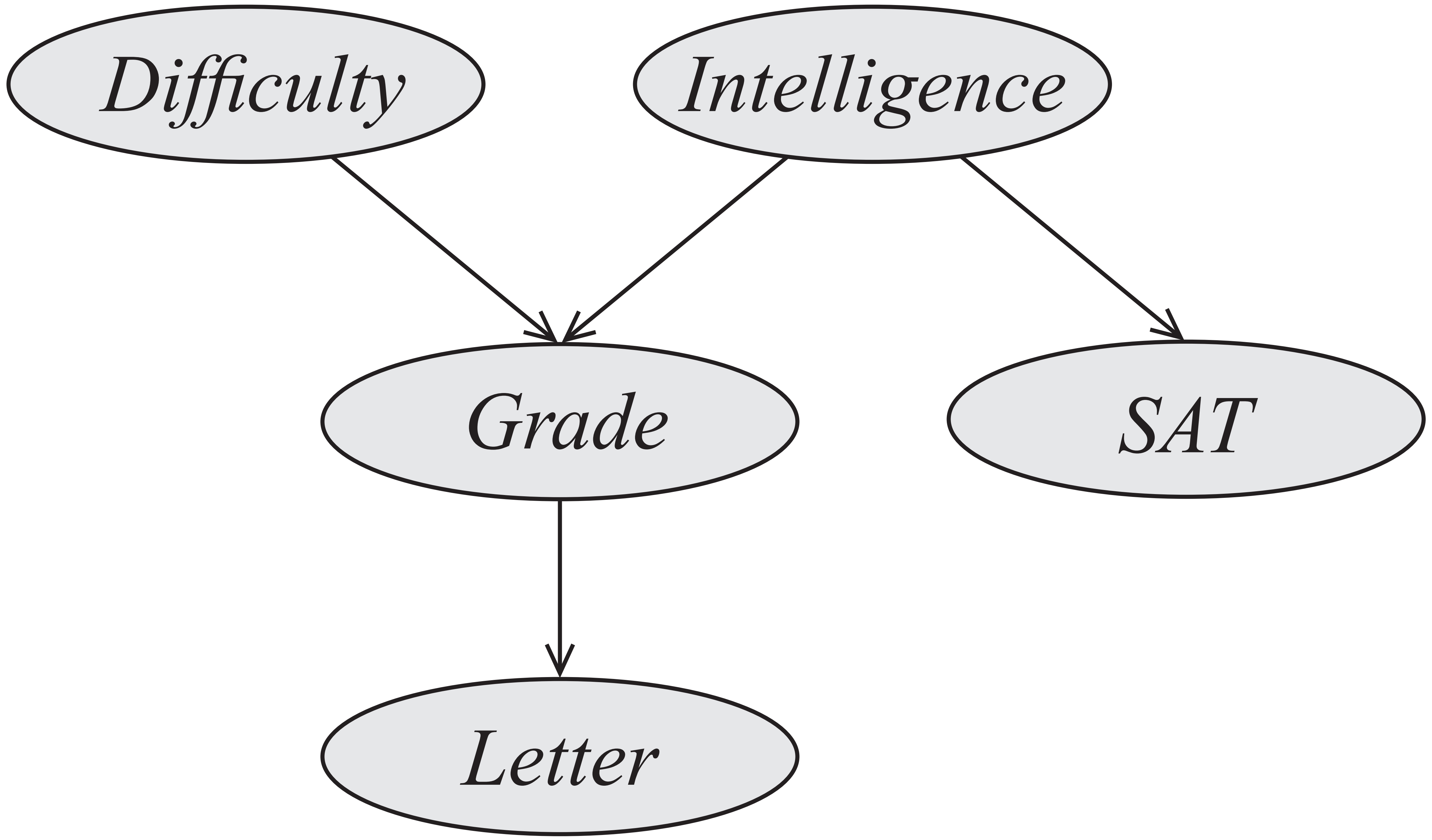


When can X influence Y

given evidence about Z

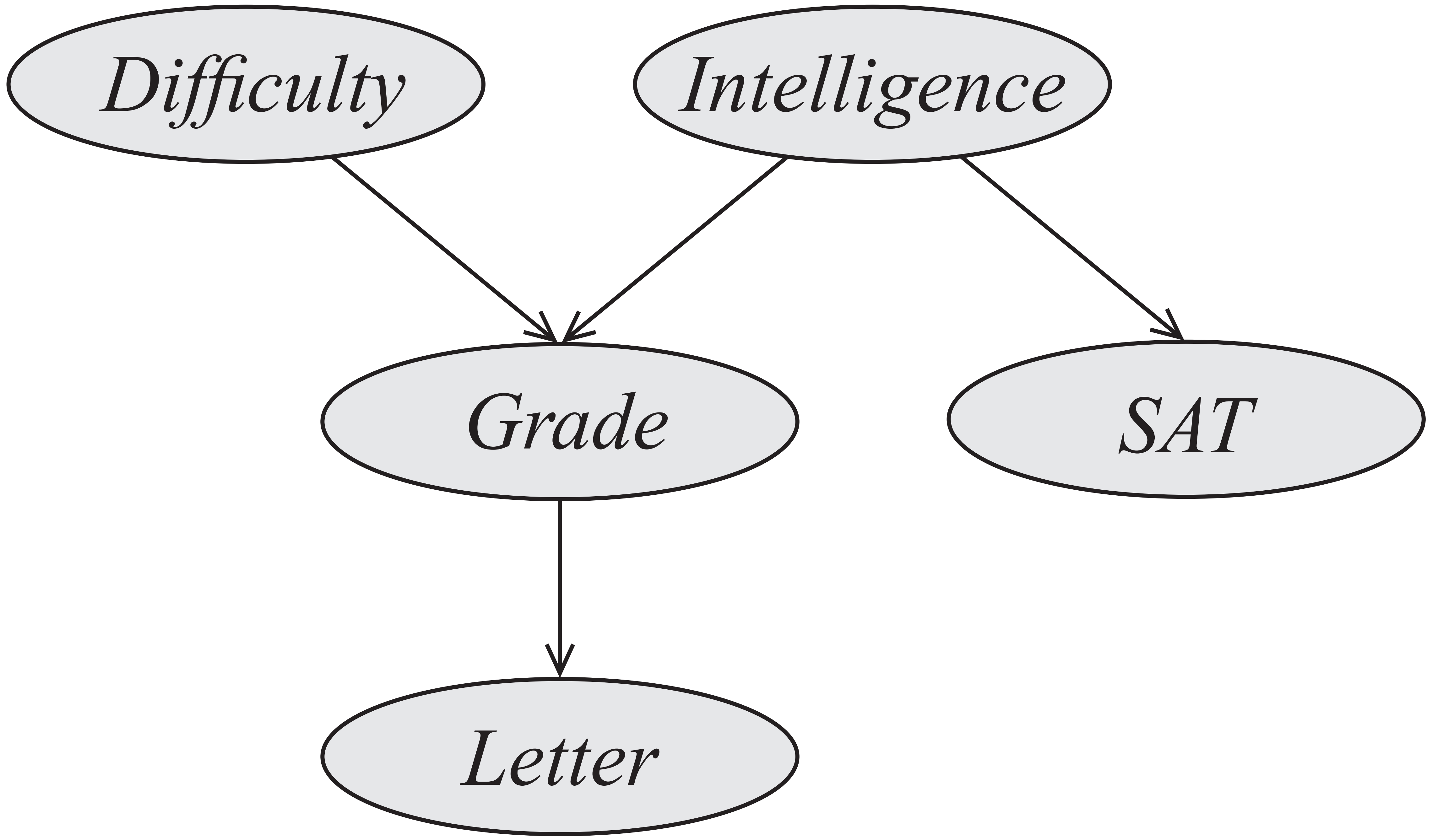
- $X \rightarrow Y$
- $X \leftarrow Y$
- $X \rightarrow W \rightarrow Y$
- $X \leftarrow W \leftarrow Y$
- $X \leftarrow W \rightarrow Y$
- $X \rightarrow W \leftarrow Y$

$W \notin Z$	$W \in Z$
✓	✗



When can X influence Y
given evidence about Z

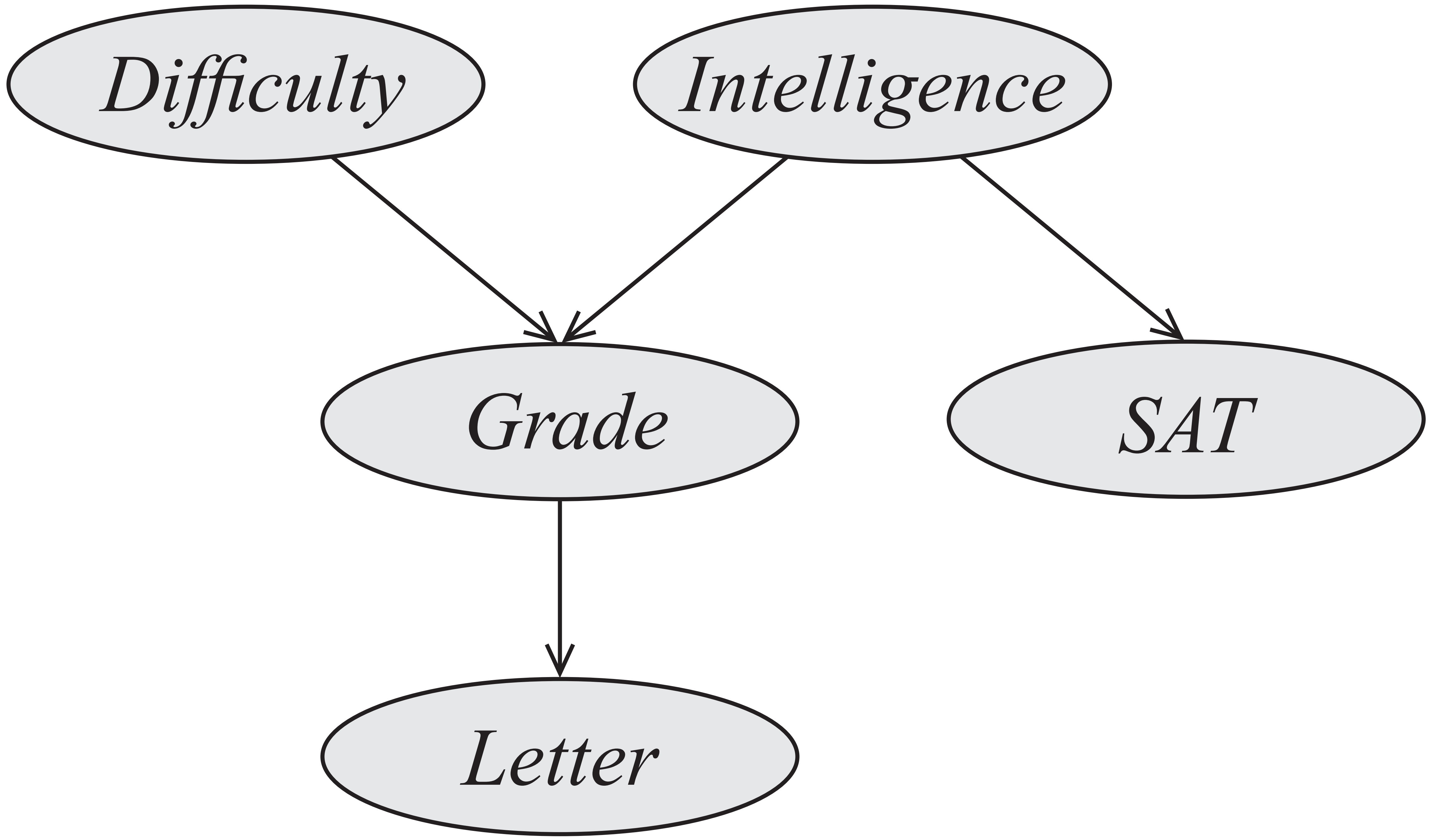
	$W \notin Z$	$W \in Z$
- $X \to Y$		
- $X \leftarrow Y$		
- $X \to W \to Y$	✓	✗
- $X \leftarrow W \leftarrow Y$	✓	
- $X \leftarrow W \to Y$		
- $X \to W \leftarrow Y$		



When can X influence Y

given evidence about Z

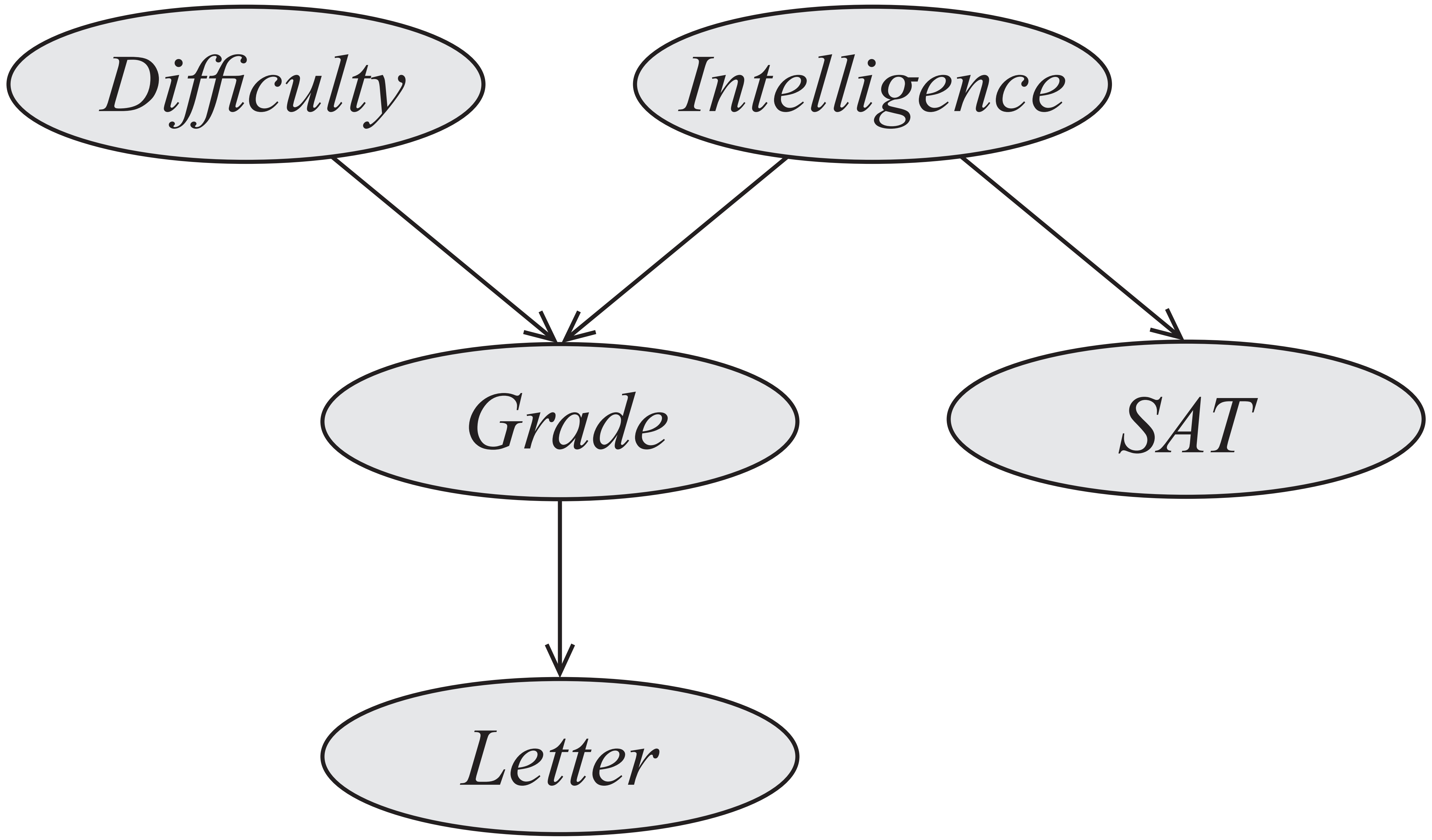
	$W \notin Z$	$W \in Z$
- $X \to Y$		
- $X \leftarrow Y$		
- $X \to W \to Y$	✓	✗
- $X \leftarrow W \leftarrow Y$	✓	✗
- $X \leftarrow W \to Y$		
- $X \to W \leftarrow Y$		



When can X influence Y

given evidence about Z

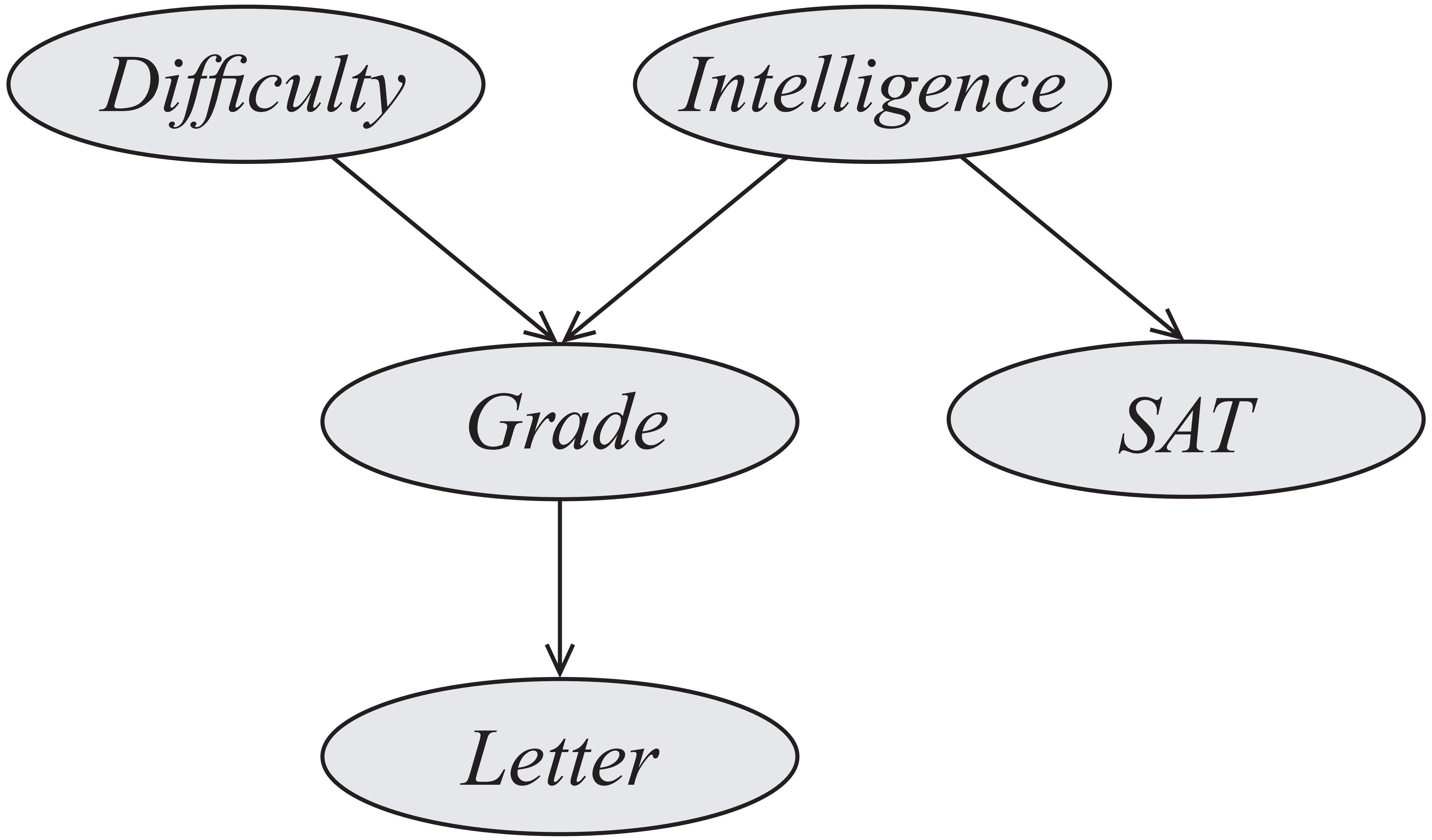
	$W \notin Z$	$W \in Z$
- $X \rightarrow Y$		
- $X \leftarrow Y$		
- $X \rightarrow W \rightarrow Y$	✓	✗
- $X \leftarrow W \leftarrow Y$	✓	✗
- $X \leftarrow W \rightarrow Y$	✓	
- $X \rightarrow W \leftarrow Y$		



When can X influence Y

given evidence about Z

	$W \notin Z$	$W \in Z$
- $X \rightarrow Y$		
- $X \leftarrow Y$		
- $X \rightarrow W \rightarrow Y$	✓	✗
- $X \leftarrow W \leftarrow Y$	✓	✗
- $X \leftarrow W \rightarrow Y$	✓	✗
- $X \rightarrow W \leftarrow Y$		

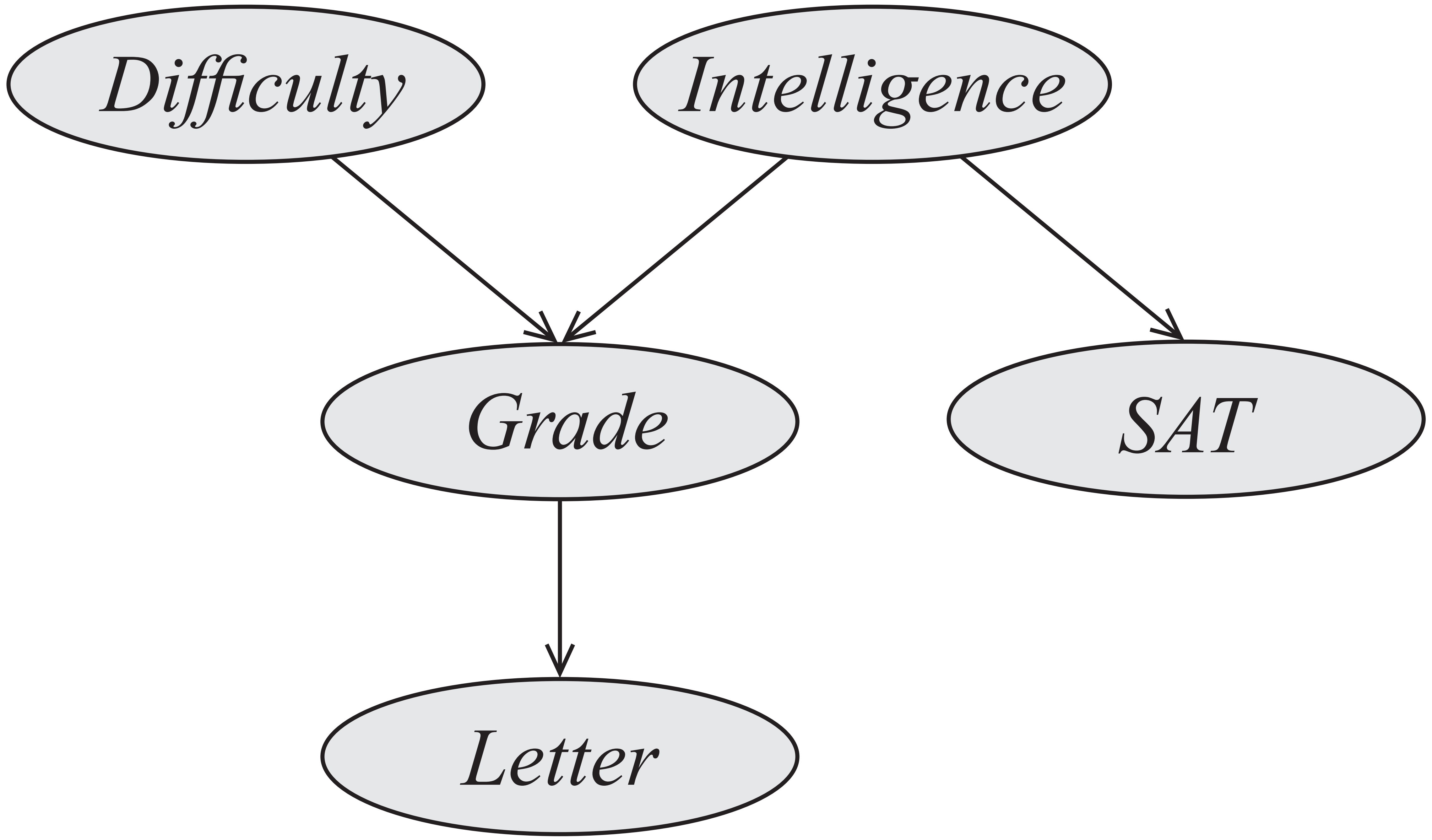


When can X influence Y

given evidence about Z

	$W \notin Z$	$W \in Z$
- $X \rightarrow Y$		
- $X \leftarrow Y$		
- $X \rightarrow W \rightarrow Y$	✓	✗
- $X \leftarrow W \leftarrow Y$	✓	✗
- $X \leftarrow W \rightarrow Y$	✓	✗
- $X \rightarrow W \leftarrow Y$	✗	

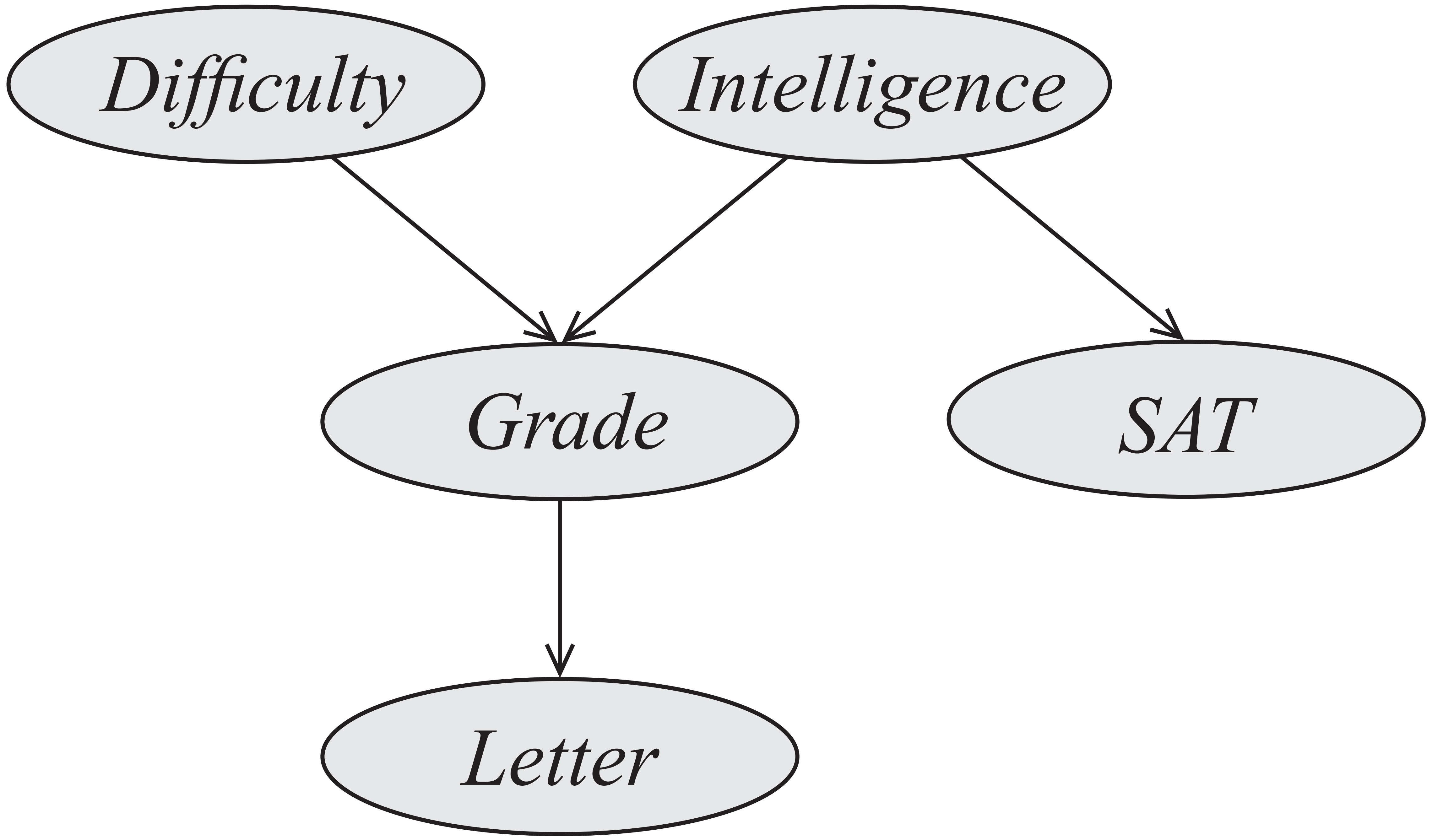
if W and all of its descendants not in Z



When can X influence Y

given evidence about Z

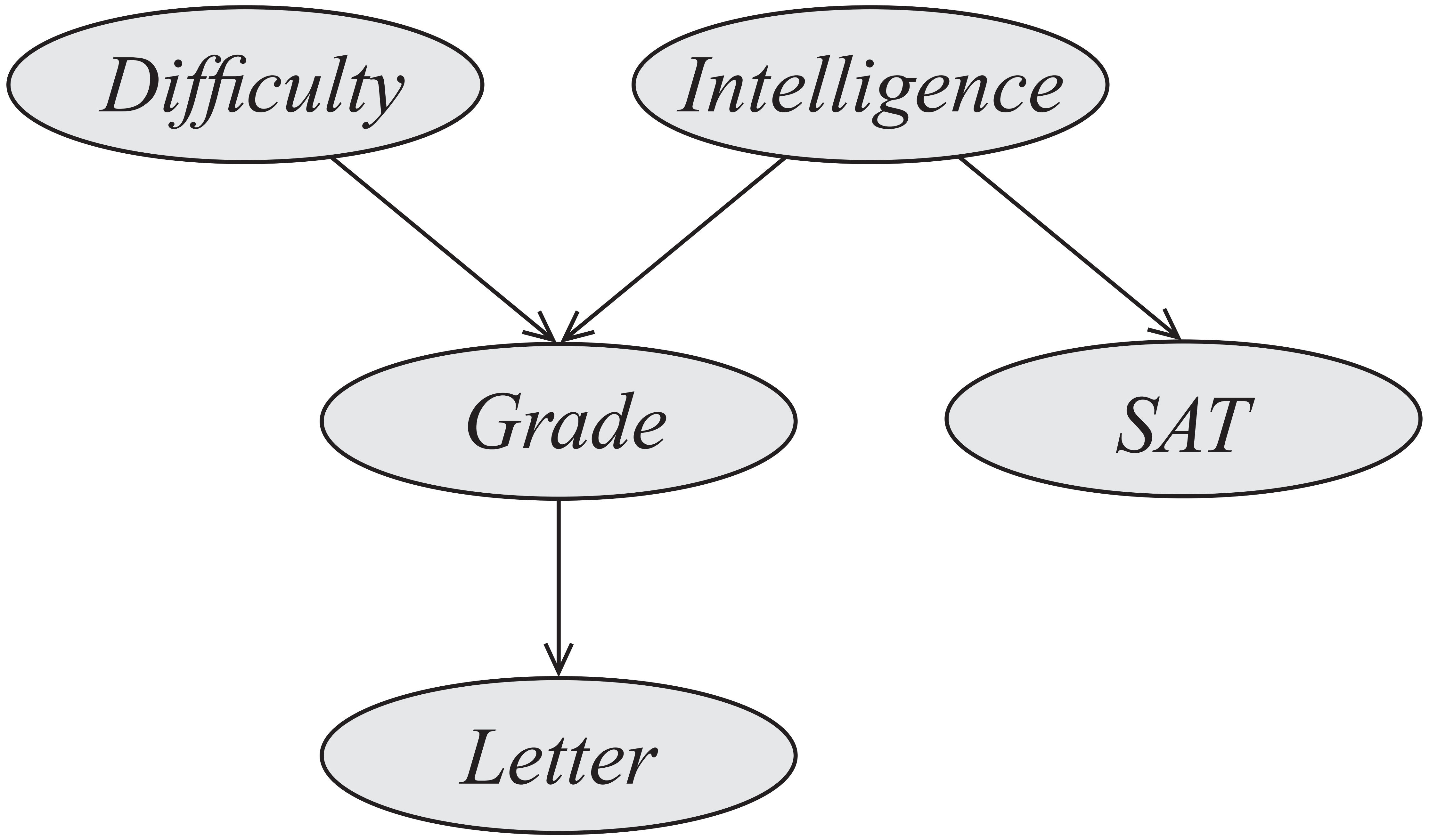
	$W \notin Z$	$W \in Z$
- $X \rightarrow Y$		
- $X \leftarrow Y$		
- $X \rightarrow W \rightarrow Y$	✓	✗
- $X \leftarrow W \leftarrow Y$	✓	✗
- $X \leftarrow W \rightarrow Y$	✓	✗
- $X \rightarrow W \leftarrow Y$	✗	✓
	if W and all of its descendants not in Z	If W or one of its descendants is observed



When can X influence Y

Given evidence about Z

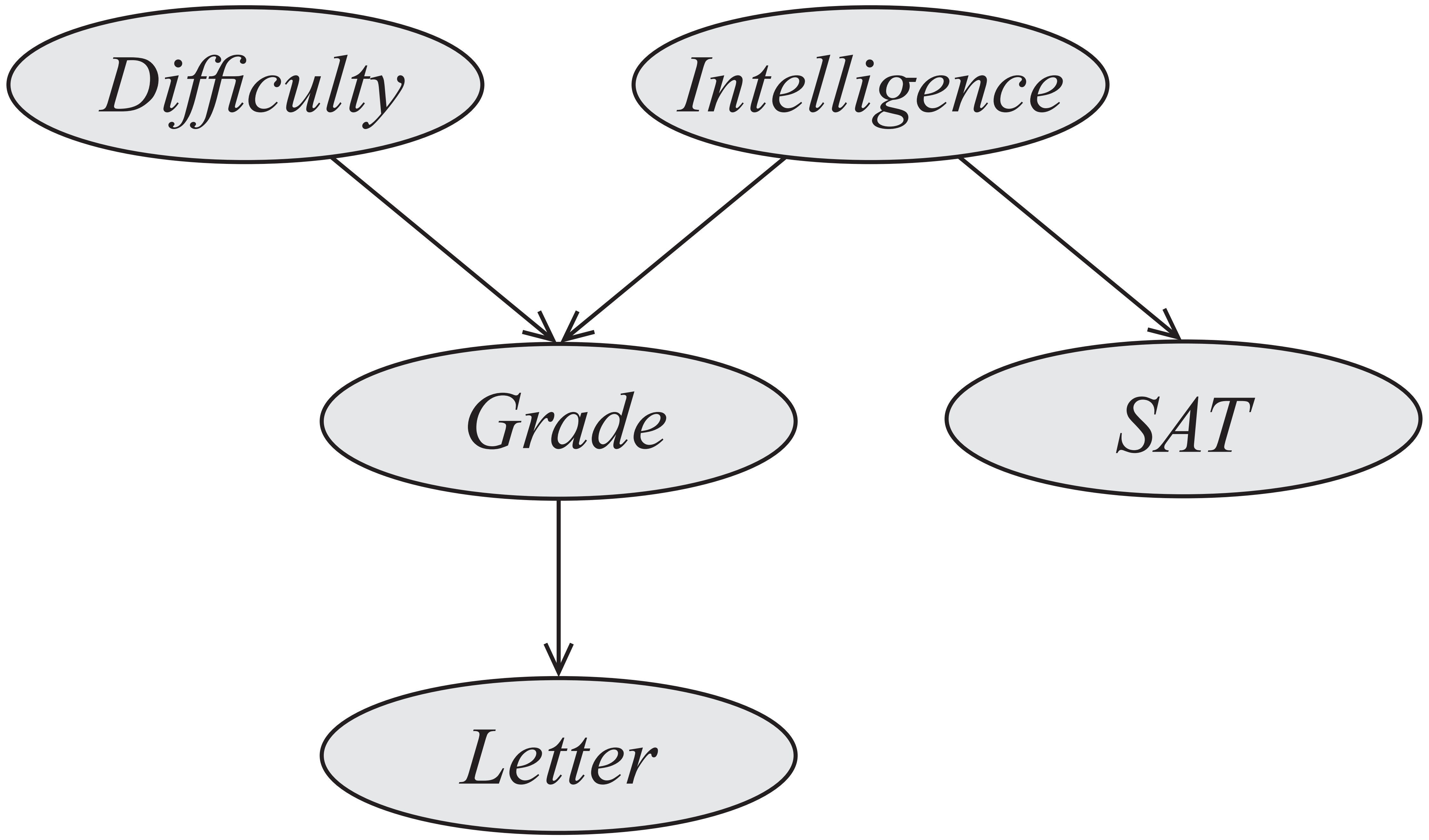
- $S - I - G - D$ allows influence to flow when?



When can X influence Y

Given evidence about Z

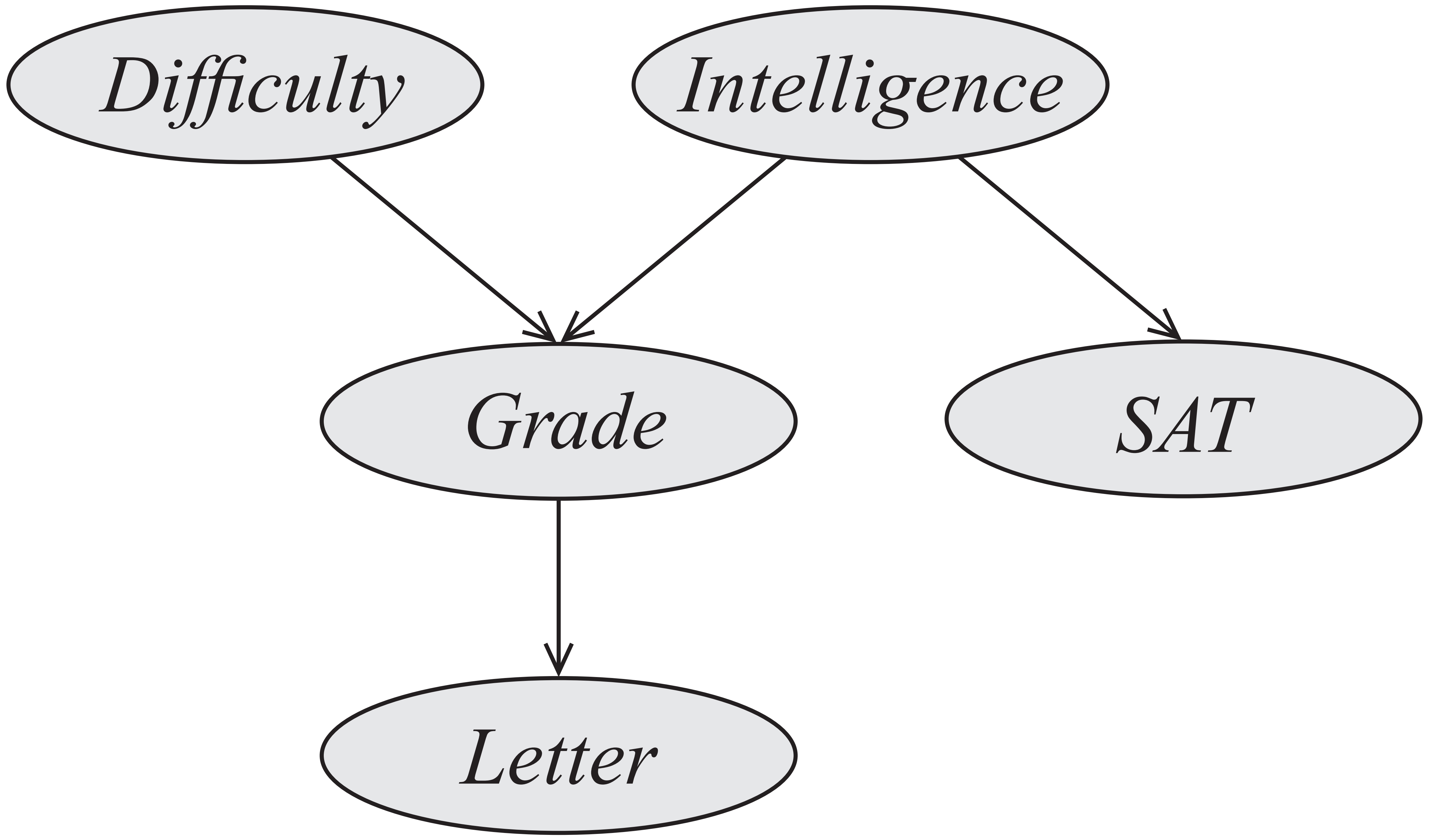
- $S - I - G - D$ allows influence to flow when?
  - I is observed



When can X influence Y

Given evidence about Z

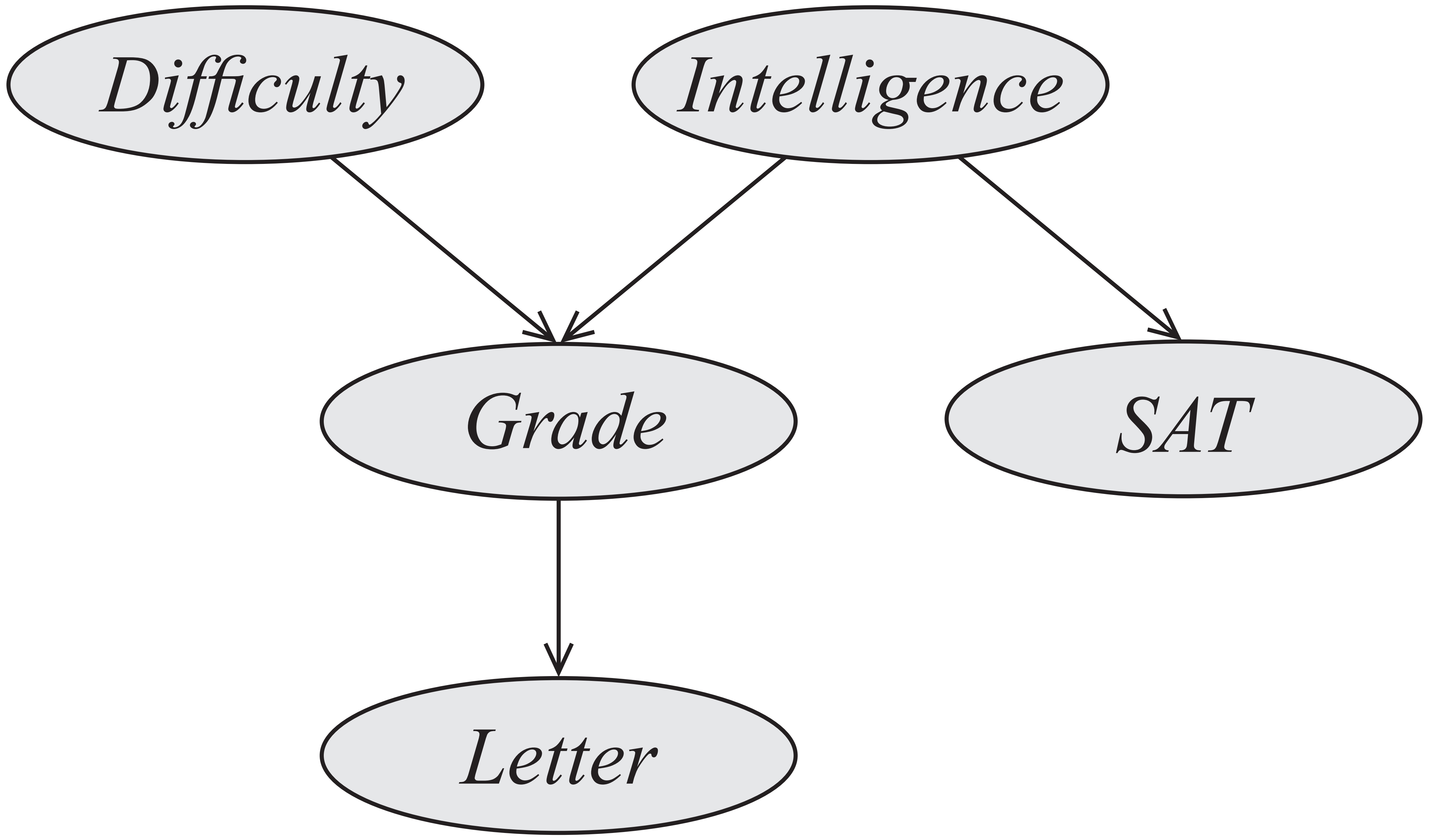
- $S - I - G - D$ allows influence to flow when?
  - I is observed ×



When can X influence Y

Given evidence about Z

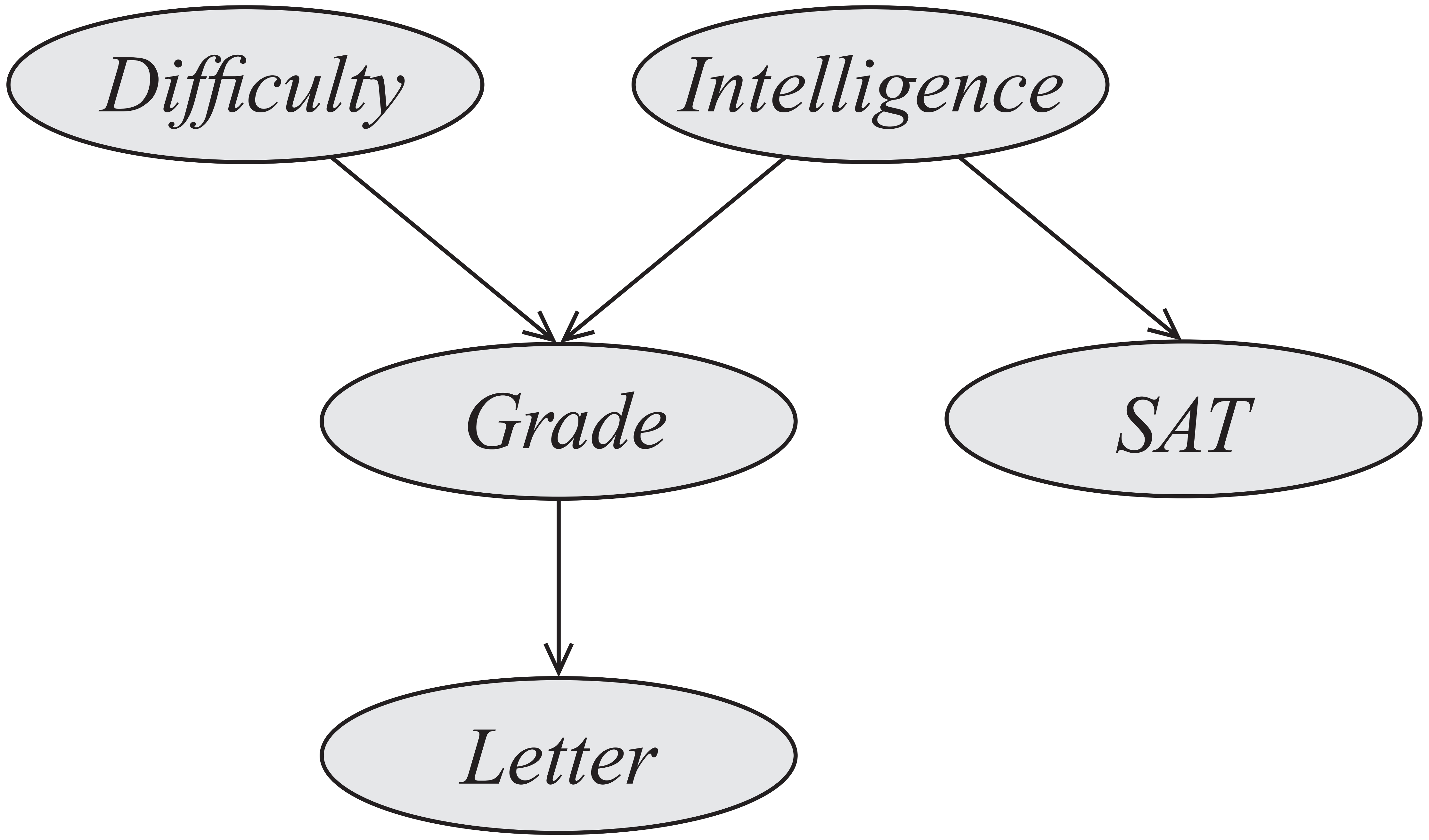
- $S - I - G - D$ allows influence to flow when?
  - I is observed ×
  - I is not observed, nothing else



When can X influence Y

Given evidence about Z

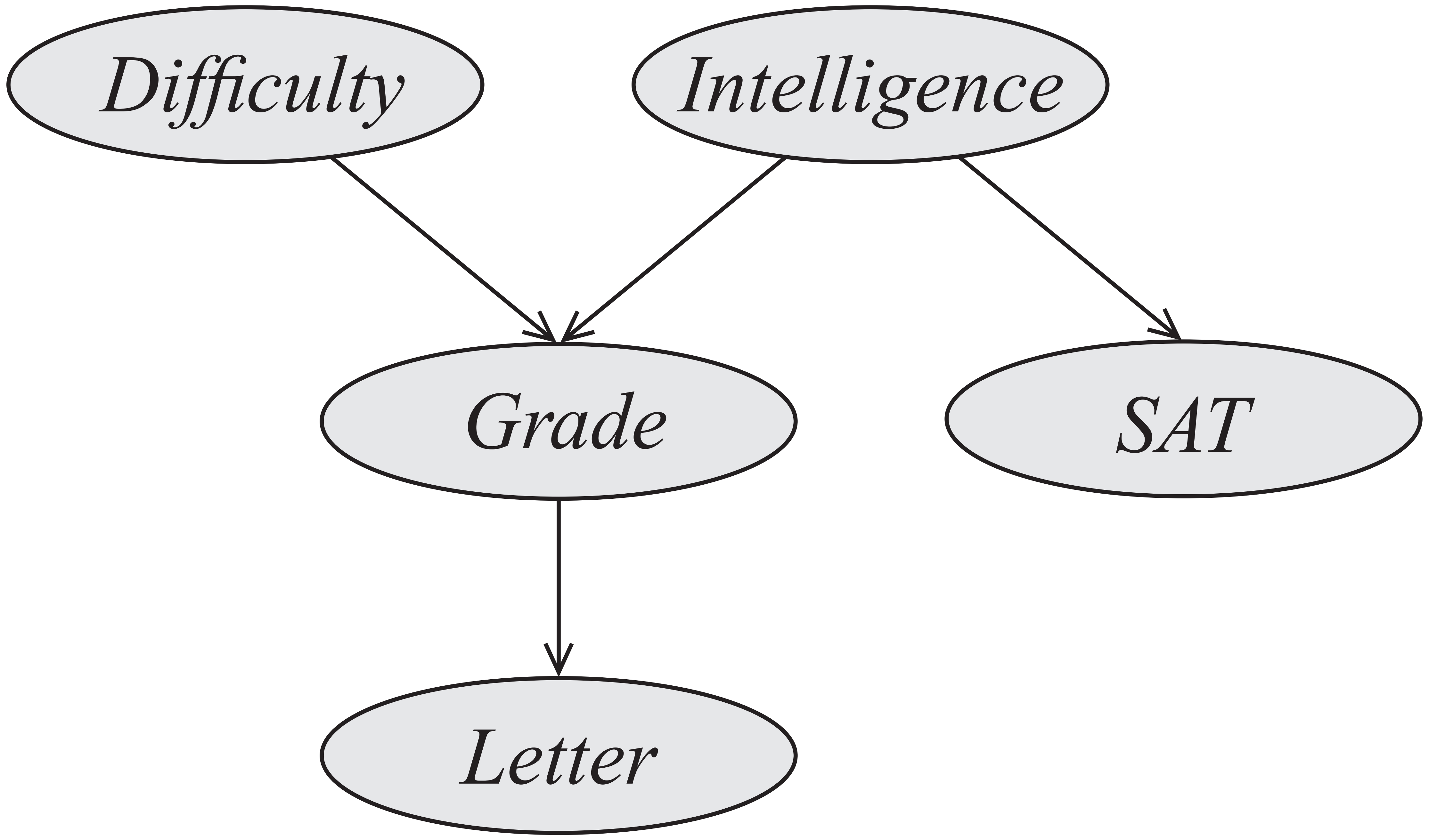
- $S - I - G - D$ allows influence to flow when?
  - I is observed ✗
  - I is not observed, nothing else ✗



When can X influence Y

Given evidence about Z

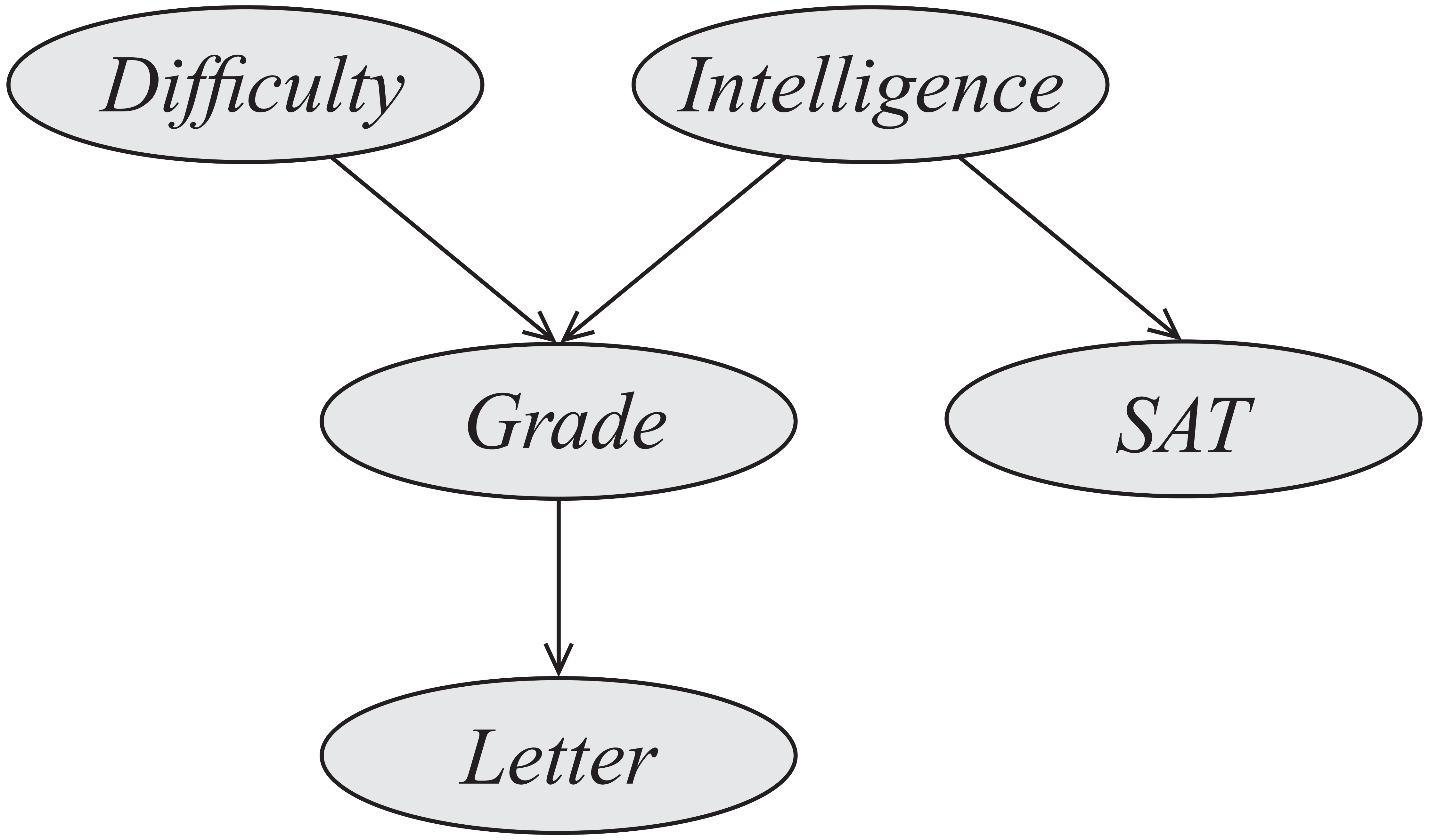
- $S - I - G - D$ allows influence to flow when?
  - I is observed ×
  - I is not observed, nothing else ×
  - I not observed, G is observed



When can X influence Y

Given evidence about Z

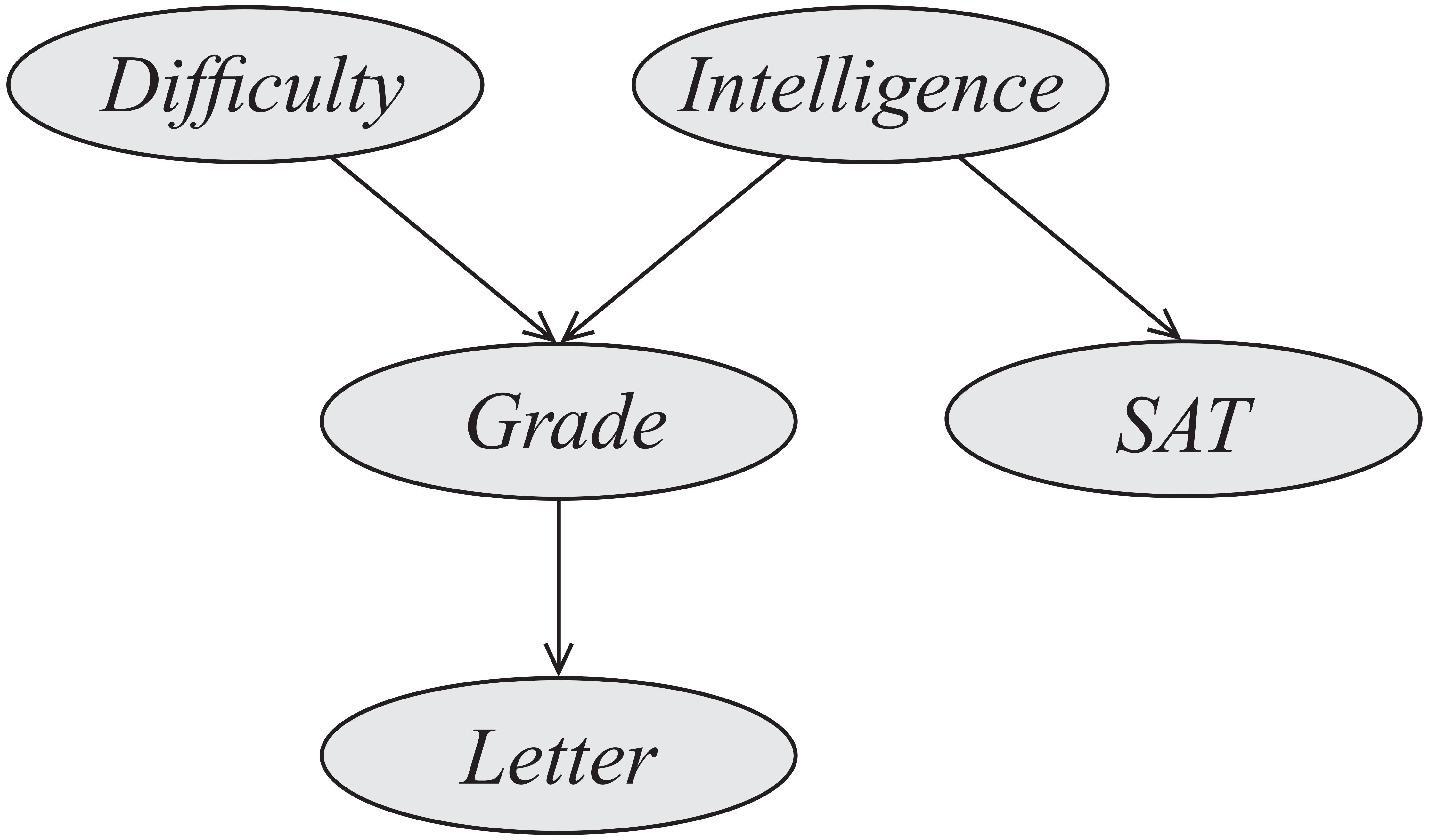
- $S - I - G - D$ allows influence to flow when?
  - I is observed ✗
  - I is not observed, nothing else ✗
  - I not observed, G is observed ✓



When can X influence Y

Given evidence about Z

- $S - I - G - D$ allows influence to flow when?
  - I is observed ✗
  - I is not observed, nothing else ✗
  - I not observed, G is observed ✓
  - ..



Active trails

- A trail $X_1 - \ldots - X_k$ is active given $\boldsymbol{Z}$ 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 $\in \boldsymbol{Z}$
  - No other X_i is in $\boldsymbol{Z}$

Independencies in BN

Independence

- For events α, β, $P \models \alpha \perp \beta$ if:

Independence

- For events α, β, $P \models \alpha \perp \beta$ if:
  - $P(\alpha, \beta) = P(\alpha)\, P(\beta)$

Independence

- For events $\alpha, \beta, P \models \alpha \perp \beta$ if:
  - $P(\alpha, \beta) = P(\alpha)\,P(\beta)$
  - $P(\alpha \,|\, \beta) = P(\alpha)$

Independence

- For events α, β, $P \models \alpha \perp \beta$ if:
  - $P(\alpha, \beta) = P(\alpha)\, P(\beta)$
  - $P(\alpha \,|\, \beta) = P(\alpha)$
  - $P(\beta \,|\, \alpha) = P(\beta)$

Independence

- For events $\alpha, \beta, P \models \alpha \perp \beta$ if:
  - $P(\alpha, \beta) = P(\alpha)\, P(\beta)$
  - $P(\alpha \,|\, \beta) = P(\alpha)$
  - $P(\beta \,|\, \alpha) = P(\beta)$
- For random variables $X, Y, P \models X \perp Y$ if:

Independence

- For events α, β, $P \models \alpha \perp \beta$ if:
  - $P(\alpha, \beta) = P(\alpha)\, P(\beta)$
  - $P(\alpha \mid \beta) = P(\alpha)$
  - $P(\beta \mid \alpha) = P(\beta)$
- For random variables X, Y, $P \models X \perp Y$ if:
  - $P(X, Y) = P(X)\, P(Y)$

Independence

- For events α, β, $P \models \alpha \perp \beta$ if:
  - $P(\alpha, \beta) = P(\alpha)\, P(\beta)$
  - $P(\alpha \mid \beta) = P(\alpha)$
  - $P(\beta \mid \alpha) = P(\beta)$
- For random variables X, Y, $P \models X \perp Y$ if:
  - $P(X, Y) = P(X)\, P(Y)$
  - $P(X \mid Y) = P(X)$

Independence

- For events α, β, $P \models \alpha \perp \beta$ if:
  - $P(\alpha, \beta) = P(\alpha)\, P(\beta)$
  - $P(\alpha \,|\, \beta) = P(\alpha)$
  - $P(\beta \,|\, \alpha) = P(\beta)$
- For random variables X, Y, $P \models X \perp Y$ if:
  - $P(X, Y) = P(X)\, P(Y)$
  - $P(X \,|\, Y) = P(X)$
  - $P(Y \,|\, X) = P(Y)$

Independence

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

Independence

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

I	D	Prob
i^0	d^0	0.42
i^0	d^1	0.18
i^1	d^0	0.28
i^1	d^1	0.12

Independence

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

I	D	Prob
i^0	d^0	0.42
i^0	d^1	0.18
i^1	d^0	0.28
i^1	d^1	0.12

I	Prob
i^0	0.6
i^1	0.4

D	Prob
d^0	0.7
d^1	0.3

Conditional Independence

- For random variables X, Y, Z,

  $P \models X \perp Y \mid Z$ if: (P satisfies conditional independence)

Conditional Independence

- For random variables X, Y, Z,

  $P \models X \perp Y \mid Z$ if: (P satisfies conditional independence)

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

Conditional Independence

- For random variables X, Y, Z,

$P \models X \perp Y | Z$ if: (P satisfies conditional independence)

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

Conditional Independence

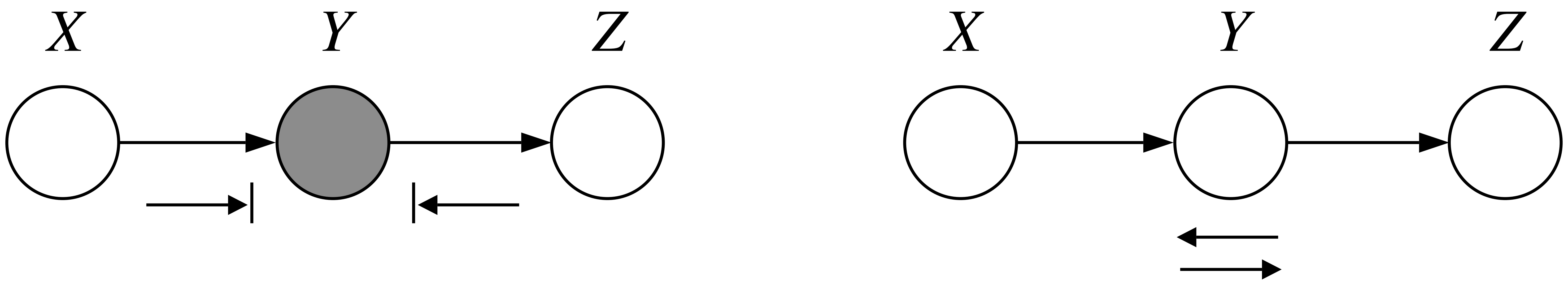
- For random variables X, Y, Z,

  $P \models X \perp Y | Z$ if: (P satisfies conditional independence)

  - $P(X, Y | Z) = P(X | Z) P(Y | Z)$
  - $P(X | Y, Z) = P(X | Z)$
  - $P(Y | X, Z) = P(Y, Z)$

D-separation ("directed separated") in Bayesian networks

- Algorithm to calculate whether $X \perp Z \mid \boldsymbol{Y}$ by looking at graph separation
- Look to see if there is an **active path** between X and Z when variables $\boldsymbol{Y}$ are observed:



(a) (b)

- If no such path, then X and Z are d-separated with respect to $\boldsymbol{Y}$.
- d-separation reduces statistical independencies (hard) to connectivity graphs (easy)
- Important because it allows us to quickly prune the Bayesian network, finding just the relevant variables for answering a query

Algorithm 3.1 Algorithm for finding nodes reachable from X given Z via active trails

```
Procedure Reachable (
  𝒢,   // Bayesian network graph
  X,   // Source variable
  Z    // Observations
)
    // Phase I: Insert all ancestors of Z into A
  L ← Z     // Nodes to be visited
  A ← ∅     // Ancestors of Z
  while L ≠ ∅
    Select some Y from L
    L ← L − {Y}
    if Y ∉ A then
      L ← L ∪ Pa_Y     // Y's parents need to be visited
    A ← A ∪ {Y}     // Y is ancestor of evidence

    // Phase II: traverse active trails starting from X
  L ← {(X, ↑)}     // (Node,direction) to be visited
  V ← ∅     // (Node,direction) marked as visited
  R ← ∅     // Nodes reachable via active trail
  while L ≠ ∅
    Select some (Y, d) from L
    L ← L − {(Y, d)}
    if (Y, d) ∉ V then
      if Y ∉ Z then
        R ← R ∪ {Y}     // Y is reachable
      V ← V ∪ {(Y, d)}     // Mark (Y, d) as visited
      if d = ↑ and Y ∉ Z then     // Trail up through Y active if Y not in Z
        for each Z ∈ Pa_Y
          L ← L ∪ {(Z, ↑)}     // Y's parents to be visited from bottom
        for each Z ∈ Ch_Y
          L ← L ∪ {(Z, ↓)}     // Y's children to be visited from top
      else if d = ↓ then     // Trails down through Y
        if Y ∉ Z then
            // Downward trails to Y's children are active
          for each Z ∈ Ch_Y
            L ← L ∪ {(Z, ↓)}     // Y's children to be visited from top
        if Y ∈ A then     // v-structure trails are active
          for each Z ∈ Pa_Y
            L ← L ∪ {(Z, ↑)}     // Y's parents to be visited from bottom
return R
```

Algorithm 3.1 Algorithm for finding nodes reachable from X given Z via active trails

Procedure Reachable (
  $\mathcal{G}$, // Bayesian network graph
  X, // Source variable
  $\boldsymbol{Z}$ // Observations
)
  // Phase I: Insert all ancestors of $\boldsymbol{Z}$ into $\boldsymbol{A}$
  $\boldsymbol{L} \leftarrow \boldsymbol{Z}$ // Nodes to be visited
  $\boldsymbol{A} \leftarrow \emptyset$ // Ancestors of $\boldsymbol{Z}$
  while $\boldsymbol{L} \neq \emptyset$
    Select some Y from $\boldsymbol{L}$
    $\boldsymbol{L} \leftarrow \boldsymbol{L} - \{Y\}$
    if $Y \notin \boldsymbol{A}$ **then**
      $\boldsymbol{L} \leftarrow \boldsymbol{L} \cup \mathrm{Pa}_Y$ // Y's parents need to be visited
    $\boldsymbol{A} \leftarrow \boldsymbol{A} \cup \{Y\}$ // Y is ancestor of evidence

  // Phase II: traverse active trails starting from X
  $\boldsymbol{L} \leftarrow \{(X, \uparrow)\}$ // (Node,direction) to be visited
  $\boldsymbol{V} \leftarrow \emptyset$ // (Node,direction) marked as visited
  $R \leftarrow \emptyset$ // Nodes reachable via active trail
  while $\boldsymbol{L} \neq \emptyset$
    Select some (Y, d) from $\boldsymbol{L}$
    $\boldsymbol{L} \leftarrow \boldsymbol{L} - \{(Y, d)\}$
    if $(Y, d) \notin \boldsymbol{V}$ **then**
      if $Y \notin \boldsymbol{Z}$ **then**
        $R \leftarrow R \cup \{Y\}$ // Y is reachable
      $\boldsymbol{V} \leftarrow \boldsymbol{V} \cup \{(Y, d)\}$ // Mark (Y, d) as visited
      if $d = \uparrow$ **and** $Y \notin \boldsymbol{Z}$ **then** // Trail up through Y active if Y not in $\boldsymbol{Z}$
        for each $Z \in \mathrm{Pa}_Y$
          $\boldsymbol{L} \leftarrow \boldsymbol{L} \cup \{(Z, \uparrow)\}$ // Y's parents to be visited from bottom
        for each $Z \in \mathrm{Ch}_Y$
          $\boldsymbol{L} \leftarrow \boldsymbol{L} \cup \{(Z, \downarrow)\}$ // Y's children to be visited from top
      else if $d = \downarrow$ **then** // Trails down through Y
        if $Y \notin \boldsymbol{Z}$ **then**
          // Downward trails to Y's children are active
          for each $Z \in \mathrm{Ch}_Y$
            $\boldsymbol{L} \leftarrow \boldsymbol{L} \cup \{(Z, \downarrow)\}$ // Y's children to be visited from top
        if $Y \in \boldsymbol{A}$ **then** // v-structure trails are active
          for each $Z \in \mathrm{Pa}_Y$
            $\boldsymbol{L} \leftarrow \boldsymbol{L} \cup \{(Z, \uparrow)\}$ // Y's parents to be visited from bottom
  return R

Procedure Reachable (
  $\mathcal{G}$, // Bayesian network graph
  X, // Source variable
  $\boldsymbol{Z}$ // Observations
)
  // Phase I: Insert all ancestors of $\boldsymbol{Z}$ into $\boldsymbol{A}$
  $\boldsymbol{L} \leftarrow \boldsymbol{Z}$ // Nodes to be visited
  $\boldsymbol{A} \leftarrow \emptyset$ // Ancestors of $\boldsymbol{Z}$
  while $\boldsymbol{L} \neq \emptyset$
    Select some Y from $\boldsymbol{L}$
    $\boldsymbol{L} \leftarrow \boldsymbol{L} - \{Y\}$
    if $Y \notin \boldsymbol{A}$ **then**
      $\boldsymbol{L} \leftarrow \boldsymbol{L} \cup \mathrm{Pa}_Y$ // Y's parents need to be visited
    $\boldsymbol{A} \leftarrow \boldsymbol{A} \cup \{Y\}$ // Y is ancestor of evidence

Algorithm 3.1 Algorithm for finding nodes reachable from X given Z via active trails

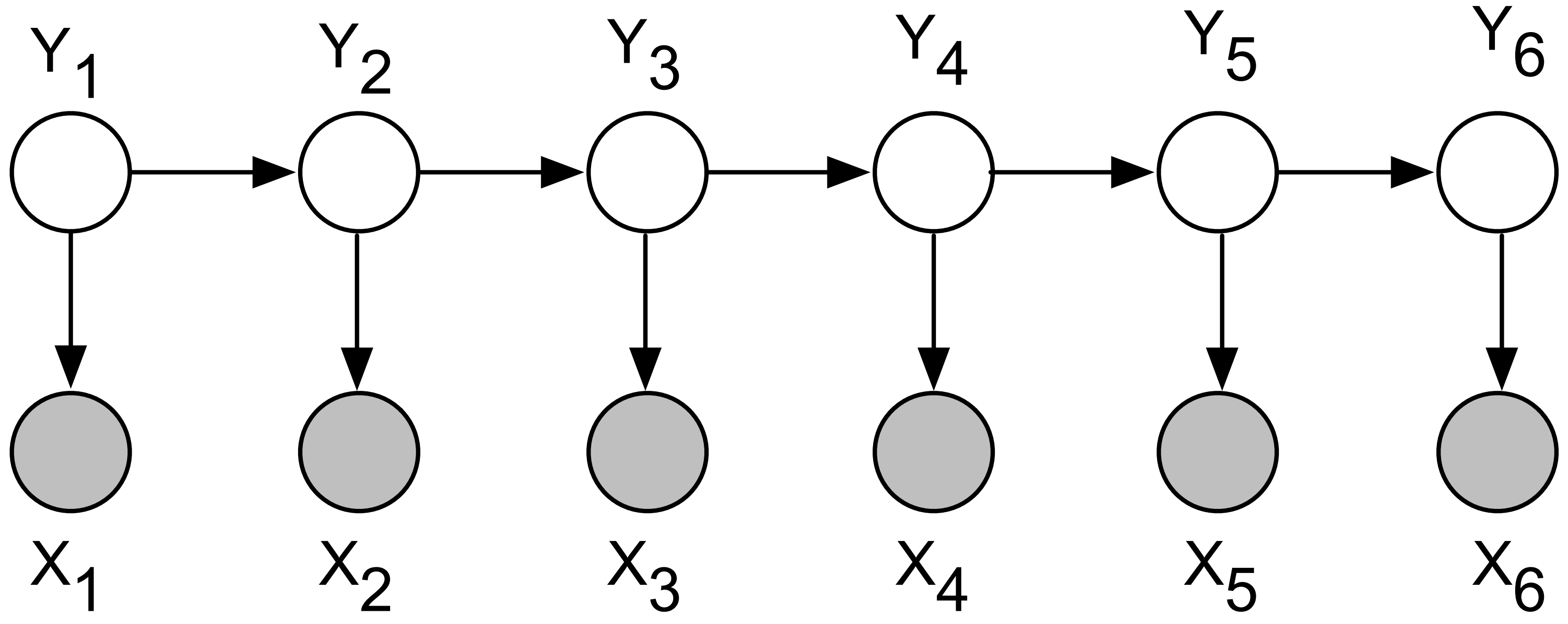
```
Procedure Reachable (
   𝒢,   // Bayesian network graph
   X,   // Source variable
   Z    // Observations
)
     // Phase I: Insert all ancestors of Z into A
   L ← Z     // Nodes to be visited
   A ← ∅     // Ancestors of Z
   while L ≠ ∅
     Select some Y from L
     L ← L − {Y}
     if Y ∉ A then
       L ← L ∪ Pa_Y     // Y's parents need to be visited
     A ← A ∪ {Y}     // Y is ancestor of evidence

     // Phase II: traverse active trails starting from X
   L ← {(X, ↑)}     // (Node,direction) to be visited
   V ← ∅     // (Node,direction) marked as visited
   R ← ∅     // Nodes reachable via active trail
   while L ≠ ∅
     Select some (Y, d) from L
     L ← L − {(Y, d)}
     if (Y, d) ∉ V then
       if Y ∉ Z then
         R ← R ∪ {Y}     // Y is reachable
       V ← V ∪ {(Y, d)}     // Mark (Y, d) as visited
       if d =↑ and Y ∉ Z then     // Trail up through Y active if Y not in Z
         for each Z ∈ Pa_Y
           L ← L ∪ {(Z, ↑)}     // Y's parents to be visited from bottom
         for each Z ∈ Ch_Y
           L ← L ∪ {(Z, ↓)}     // Y's children to be visited from top
       else if d =↓ then     // Trails down through Y
         if Y ∉ Z then
             // Downward trails to Y's children are active
           for each Z ∈ Ch_Y
             L ← L∪{(Z, ↓)}     // Y's children to be visited from top
         if Y ∈ A then     // v-structure trails are active
           for each Z ∈ Pa_Y
             L ← L∪{(Z, ↑)}     // Y's parents to be visited from bottom
 return R
```


```
     // Phase II: traverse active trails starting from X
   L ← {(X, ↑)}     // (Node,direction) to be visited
   V ← ∅     // (Node,direction) marked as visited
   R ← ∅     // Nodes reachable via active trail
   while L ≠ ∅
     Select some (Y, d) from L
     L ← L − {(Y, d)}
     if (Y, d) ∉ V then
       if Y ∉ Z then
         R ← R ∪ {Y}     // Y is reachable
       V ← V ∪ {(Y, d)}     // Mark (Y, d) as visited
       if d =↑ and Y ∉ Z then     // Trail up through Y active if Y not in Z
         for each Z ∈ Pa_Y
           L ← L ∪ {(Z, ↑)}     // Y's parents to be visited from bottom
         for each Z ∈ Ch_Y
           L ← L ∪ {(Z, ↓)}     // Y's children to be visited from top
       else if d =↓ then     // Trails down through Y
         if Y ∉ Z then
             // Downward trails to Y's children are active
           for each Z ∈ Ch_Y
             L ← L∪{(Z, ↓)}     // Y's children to be visited from top
         if Y ∈ A then     // v-structure trails are active
           for each Z ∈ Pa_Y
             L ← L∪{(Z, ↑)}     // Y's parents to be visited from bottom
 return R
```

What are some frequently used graphical models?

Hidden Markov models



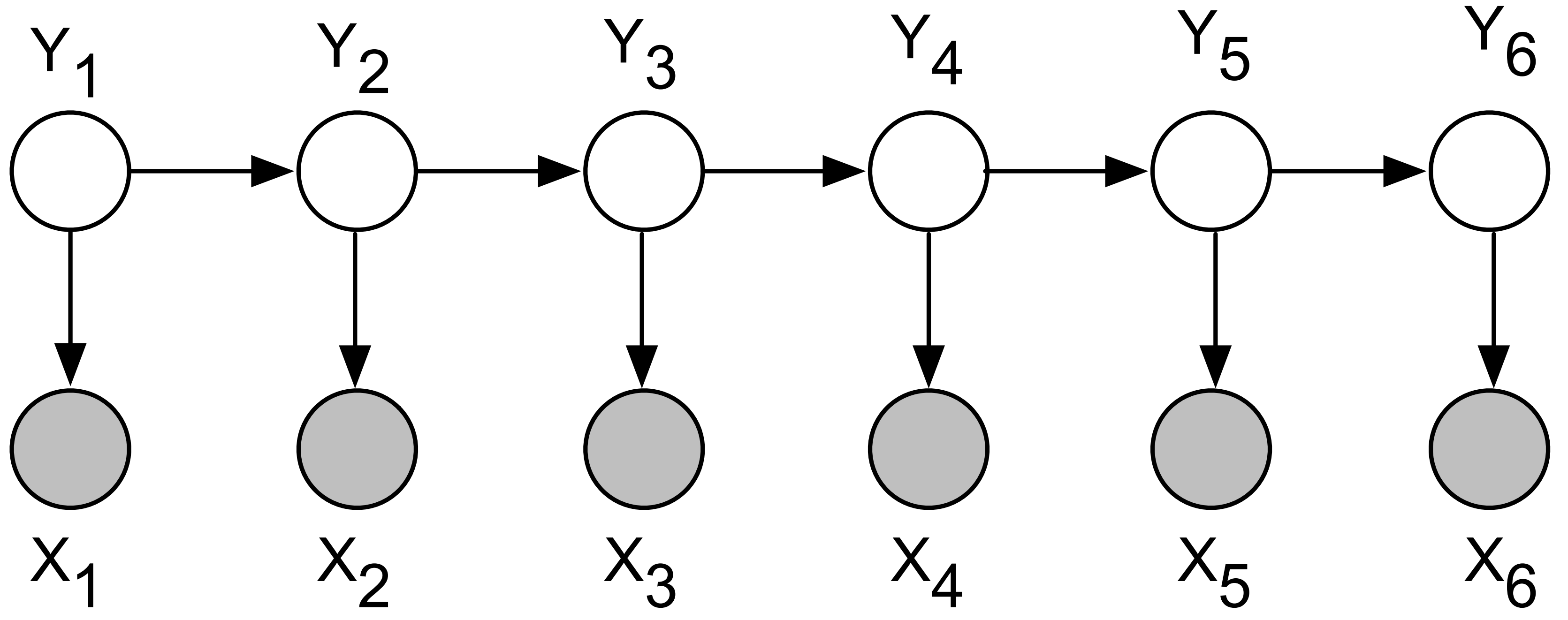
- Frequently used for speech recognition and part-of-speech tagging
- Joint distribution factors as

$$p(\boldsymbol{y}, \boldsymbol{x}) = p(y_1)\, p(x_1 \,|\, y_1) \prod_{t=2}^{T} p(y_t \,|\, y_{t-1})\, p(x_t \,|\, y_t)$$

  - $p(y_1)$ is the distribution for the starting state
  - $p(y_t \,|\, y_{t-1})$ is the transition probability between any two states
  - $p(x_t \,|\, y_t)$ is the emission probability
- What are the conditional independencies here?

  For example, $Y_1 \perp \{Y_3, \ldots, Y_6\} \,|\, Y_2$

Hidden Markov models



- Joint distribution factors as

$$p(\boldsymbol{y}, \boldsymbol{x}) = p(y_1)\, p(x_1 \,|\, y_1) \prod_{t=2}^{T} p(y_t \,|\, y_{t-1})\, p(x_t \,|\, y_t)$$

- A **homogeneous** HMM uses the same parameters (β and α below) for each transition and emission distribution (**parameter sharing**):

$$p(\boldsymbol{y}, \boldsymbol{x}) = p(y_1)\, \alpha_{x_1, y_1} \prod_{t=2}^{T} \beta_{y_t, y_{t-1}}\, \alpha_{x_t, y_t}$$

How many parameters need to be learned?

Mixture of Gaussians

- The N-dim multivariate normal distribution, $\mathcal{N}(\mu, \Sigma)$, has density:

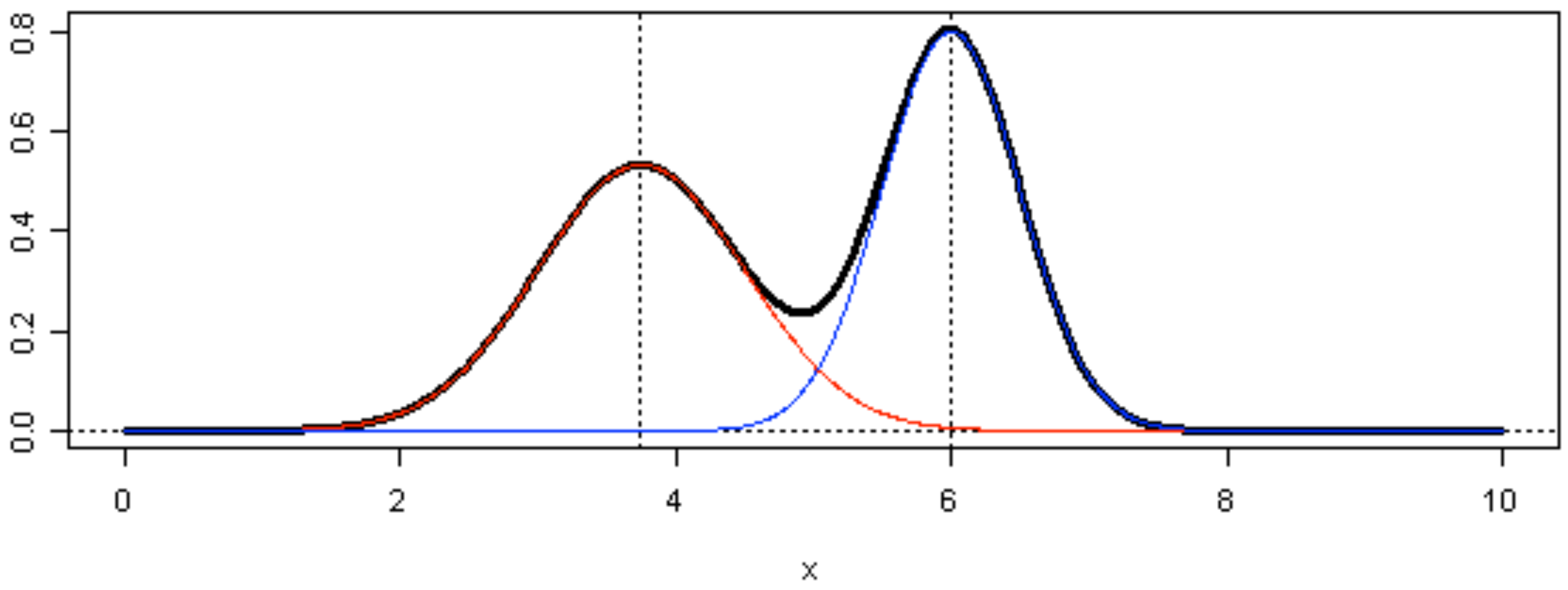
$$p(\boldsymbol{x}) = \frac{1}{(2\pi)^{N/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \mu)^\top \Sigma^{-1} (\boldsymbol{x} - \mu)\right)$$

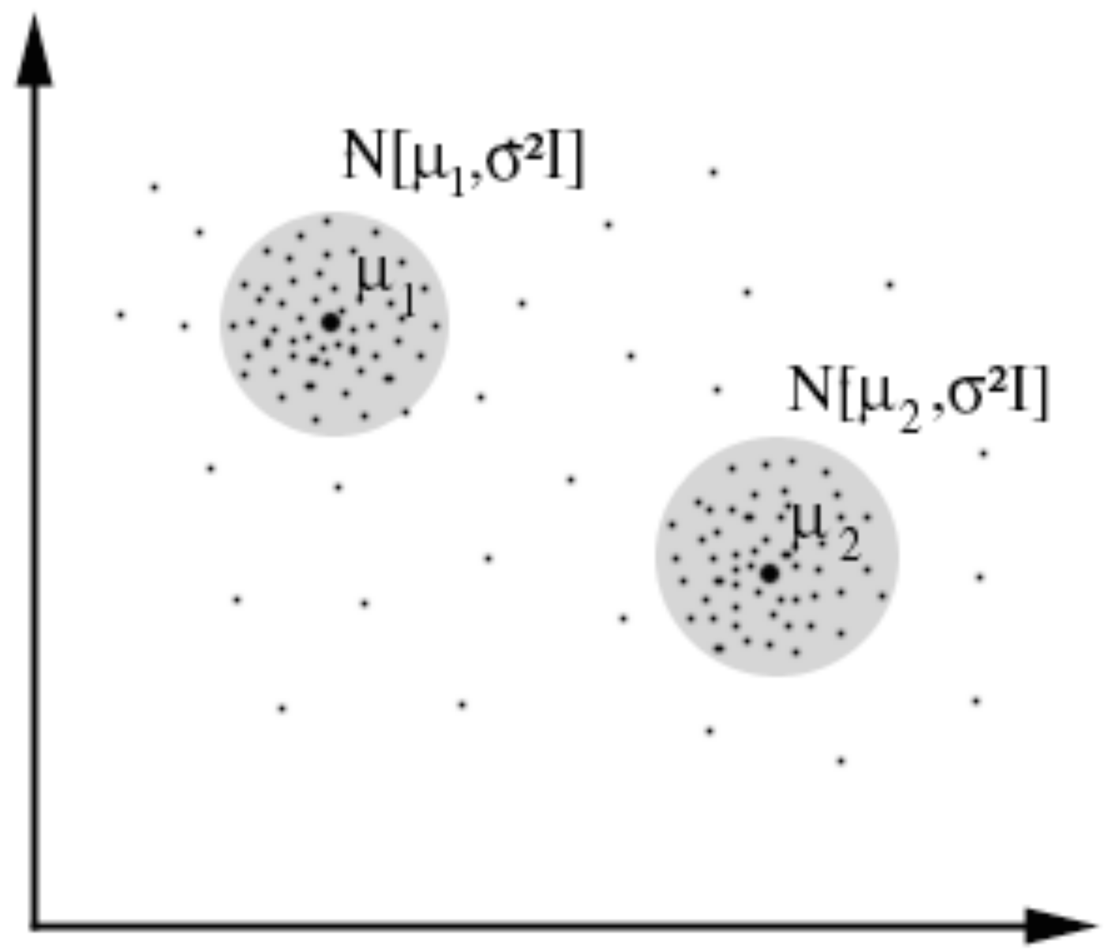
- Suppose we have k Gaussians given by μ_k and Σ_k, and a distribution θ over the numbers $1,\ldots,k$

- Mixture of Gaussians distribution $p(y, \boldsymbol{x})$ given by
  1. Sample $y \sim \theta$ (specifies which Gaussian to use)
  2. Sample $x \sim \mathcal{N}(\mu_y, \Sigma_y)$

Mixture of Gaussians

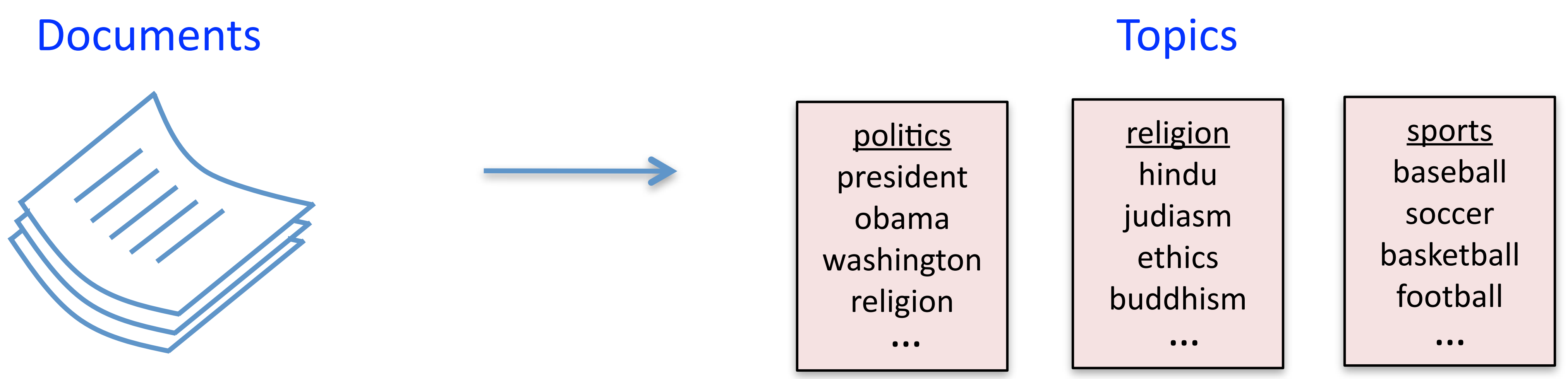
- The marginal distribution over $\boldsymbol{x}$ looks like:



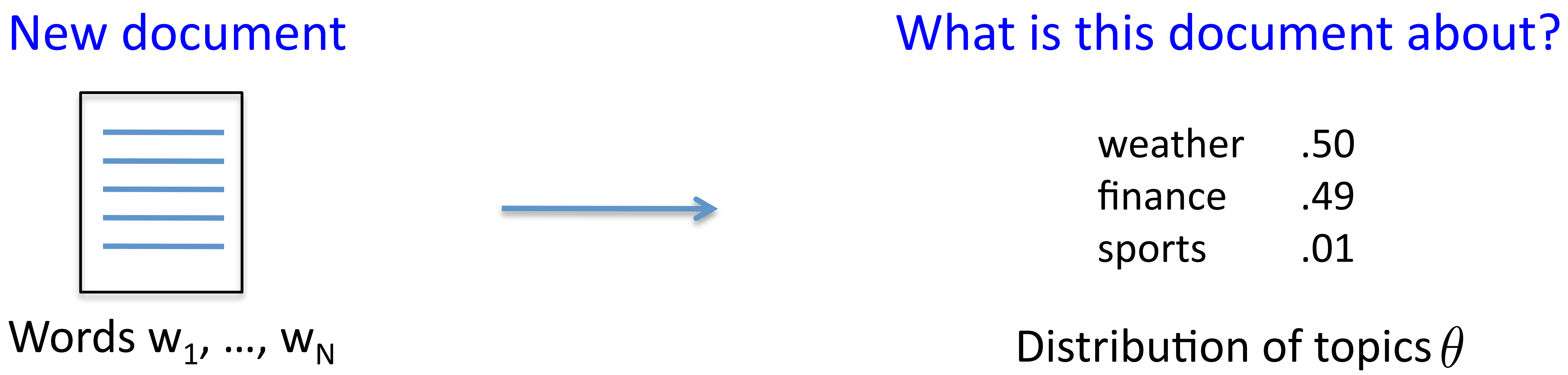


Latent Dirichlet allocation

- Topic models are powerful tools for exploring large data sets and for making inferences about the content of documents



- Many applications in information retrieval, document summarisation, and classification



- LDA is one of the simplest and most widely used topic models

Generative model for a document in LDA

1. Sample the document's topic distribution θ (aka topic vector)

$$\theta \sim \text{Dirichlet}(\alpha_{1:T})$$

   where the $\{\alpha_t\}_{t=1}^T$ are fixed hyperparameters. Thus θ is a distribution over T topics with mean $\theta_t = \alpha_t / \sum_{t'} \alpha_{t'}$

2. For $i = 1 : N$, sample the **topic** z_i of the *i*'th word

$$z_i | \theta \sim \theta$$

3. …and then sample the actual **word** w_i from the z_i'th topic

$$w_i | z_i \sim \beta_{z_i}$$

   where $\{\beta_t\}_{t=1}^T$ are the *topics* (a fixed connection of distributions on words).

Summary so far

- **Bayesian networks** given by (G, P) where P is specified as a set of local **conditional probability distributions** associated with G's nodes
- One interpretation of a BN is as a **generative model**, where variables are sampled in topological order
- Local and global independence properties identifiable via **d-separation** criteria
- Computing the probability of any assignment is obtained by multiplying CDPs
  - **Bayes' rule** is used to compute conditional probabilities
  - Marginalisation or **inference** is often computationally difficult
- Examples: **naive Bayes, hidden Markov models, latent Dirichlet allocation**

Factors

P factorises over G

- Let G be a graph over $X_1, \ldots, X_n$.
- P factorises over G if

$$P(X_1, \ldots, X_n) = \prod_i P(X_i \,|\, \mathrm{Pa}_G(X_i))$$

Factors

- A factor $\phi(X_1, \ldots, X_k)$

  $\phi : \text{Val}(X_1, \ldots, X_k) \to \mathbb{R}$

- Scope $= \{X_1, \ldots, X_k\}$

Factors

a factor is a function

- A factor $\phi(X_1, \ldots, X_k)$

  $\phi : \text{Val}(X_1, \ldots, X_k) \rightarrow \mathbb{R}$

- Scope $= \{X_1, \ldots, X_k\}$

Factors

a factor is a function

- A factor $\phi(X_1, \ldots, X_k)$

  $\phi : \mathrm{Val}(X_1, \ldots, X_k) \to \mathbb{R}$

the variable arguments
of that function

- Scope $= \{X_1, \ldots, X_k\}$

A joint distribution is a factor

$P(I, D, G)$

I	D	G	Probability
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

A joint distribution is a factor

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

I	D	G	Probability
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

Here, they sum to 1.0, but for general factors, they don't have to...

Unnormalised measure $P(I, D, g^1)$

$P(I, D, g^1)$

I	D	G	Probability
i^0	d^0	g^1	0.126
i^0	d^1	g^1	0.009
i^1	d^0	g^1	0.252
i^1	d^1	g^1	0.06

Scope $= \{I, D\}$

Unnormalised measure $P(I, D, g^1)$

$$\underbrace{P(I, D}_{\text{Scope}}, g^1)$$

I	D	G	Probability
i^0	d^0	g^1	0.126
i^0	d^1	g^1	0.009
i^1	d^0	g^1	0.252
i^1	d^1	g^1	0.06

Scope $= \{I, D\}$

Conditional Probability Distribution (CPD)

$$P(G \mid I, D)$$

	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

General factors

A	B	ϕ
a^0	b^0	30
a^0	b^1	5
a^1	b^0	1
a^1	b^1	10

General factors

don't have to be probabilities

A	B	ϕ
a^0	b^0	30
a^0	b^1	5
a^1	b^0	1
a^1	b^1	10

Factor product

$\phi_1(A, B)$

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

$\phi_2(B, C)$

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

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

Factor product

$\phi_1(A, B)$

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

$\phi_2(B, C)$

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

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

Factor product

$\phi_1(A, B)$

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

$\phi_2(B, C)$

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

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

Factor product

$\phi_1(A, B)$

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

$\phi_2(B, C)$

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

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

Factor Marginalisation

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

Factor Marginalisation

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

Factor Reduction

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

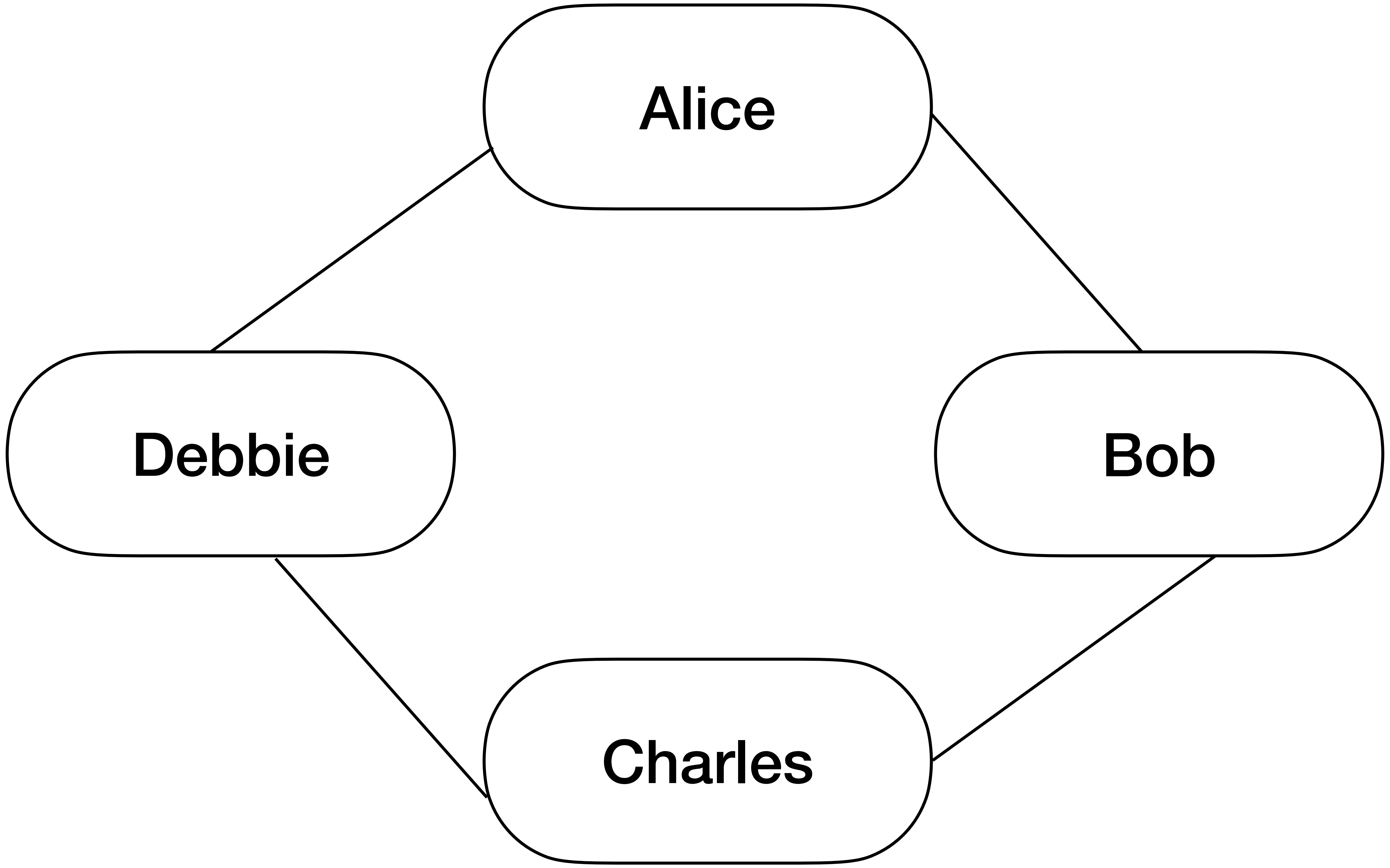
a^1	b^1	c^1	*0.25*
a^1	b^2	c^1	0.08
a^2	b^1	c^1	0.05
a^2	b^2	c^1	0
a^3	b^1	c^1	0.15
a^3	b^2	c^1	0.09

Factors

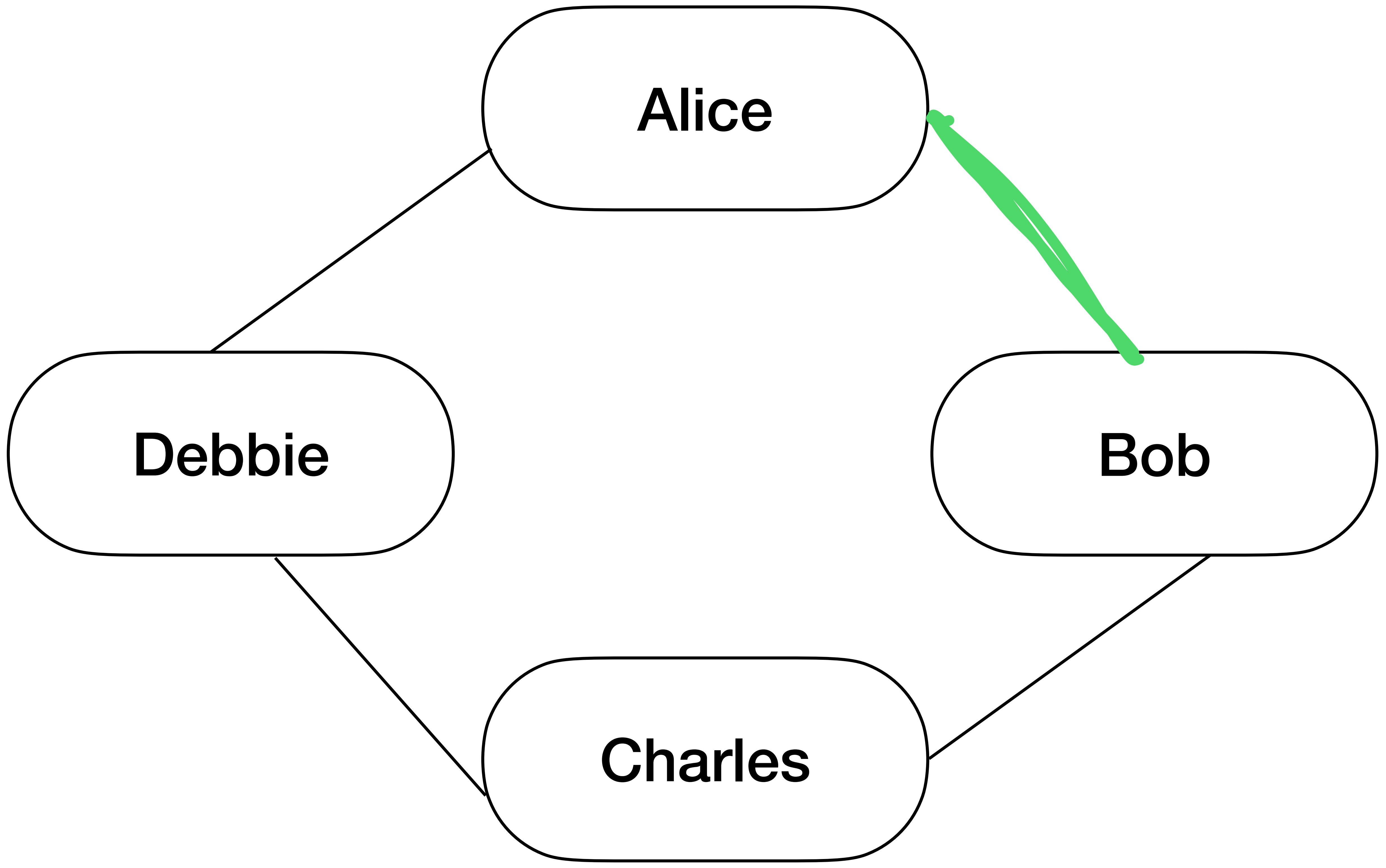
- Factors are a building block for defining distributions in high-dimensional spaces
- We have seen a set of basic operations on factors that are useful for manipulating these probability distributions

Markov Networks

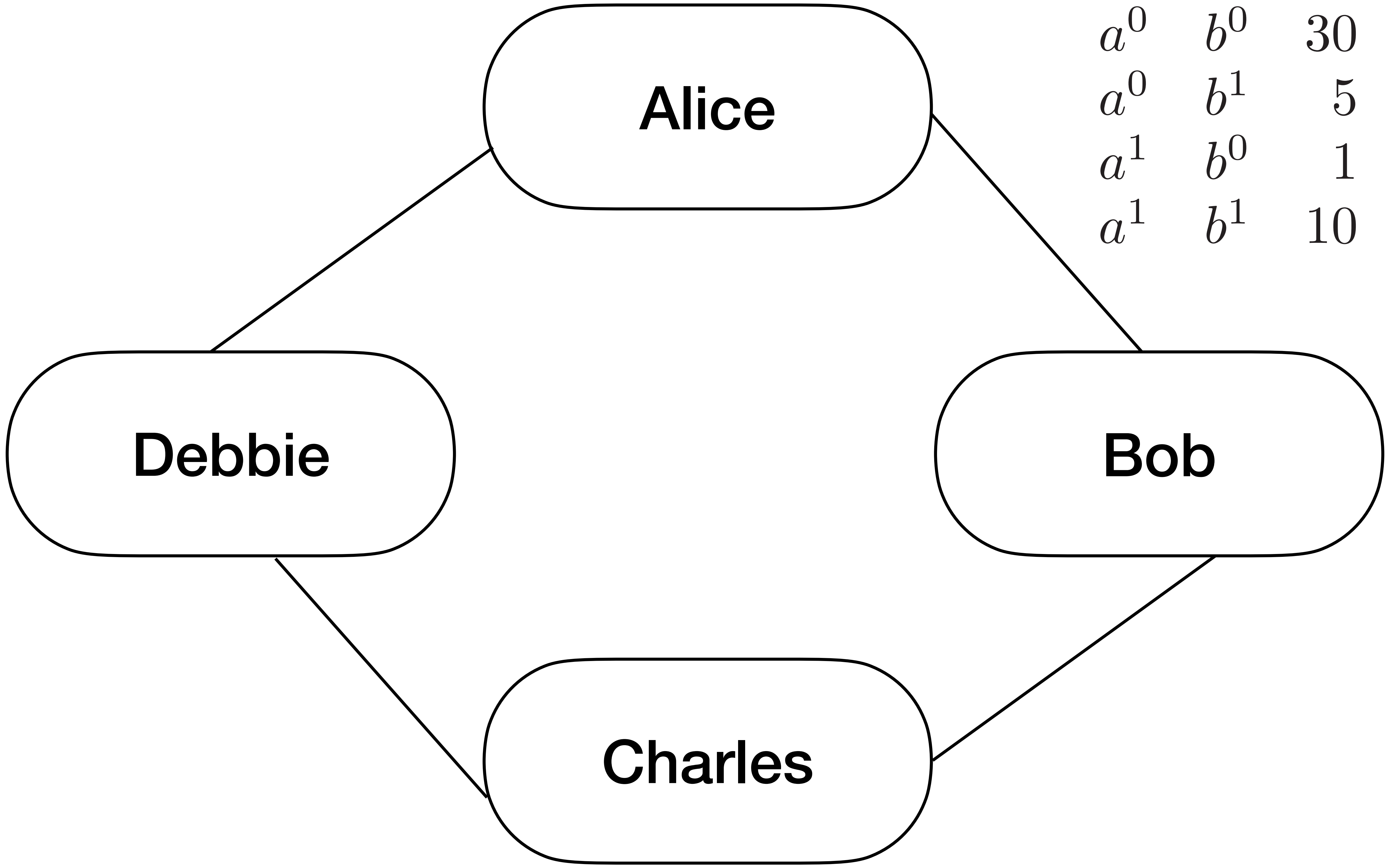
Pairwise Markov Networks



Pairwise Markov Networks



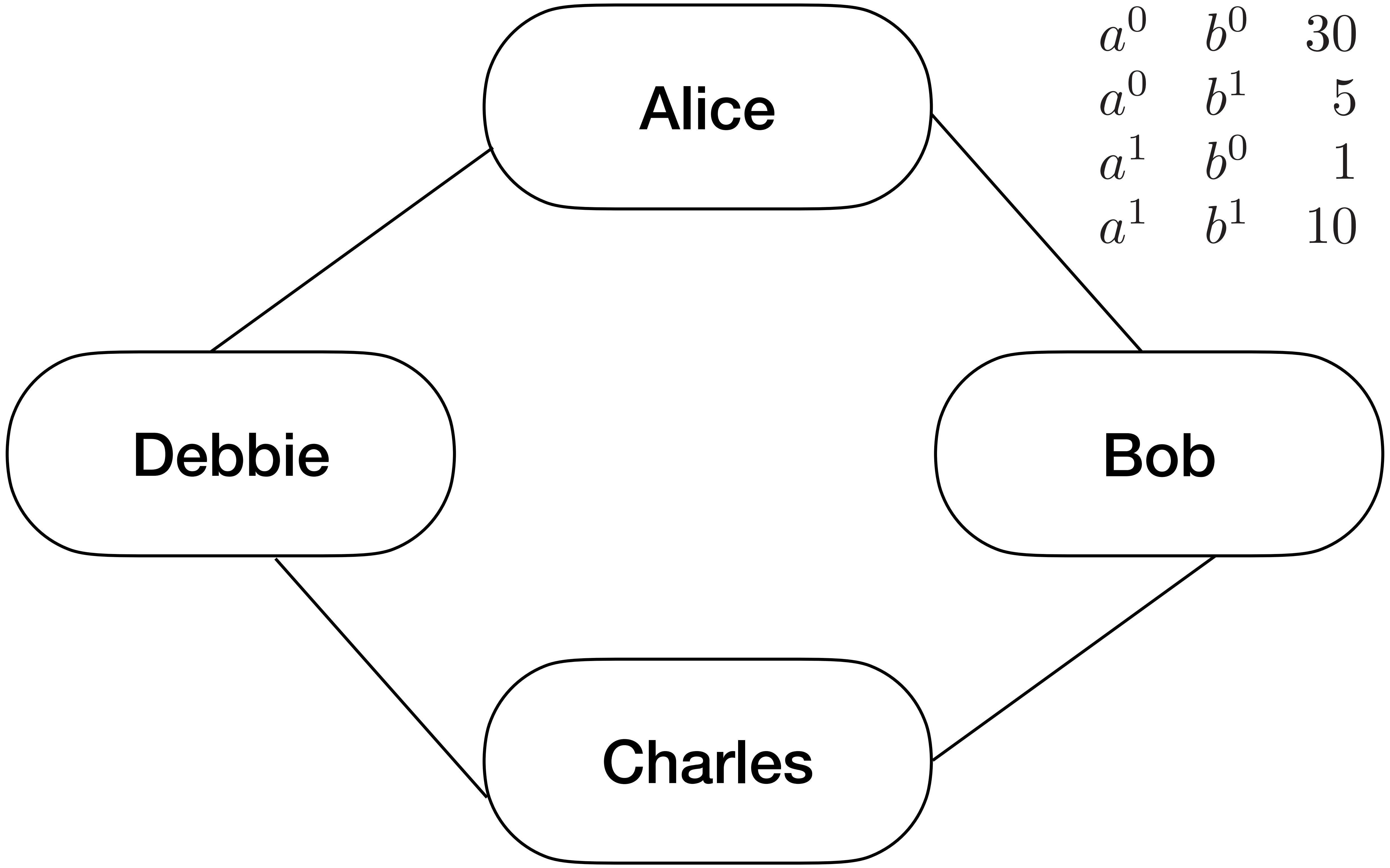
Pairwise Markov Networks



$$\phi_1(A, B)$$

a^0	b^0	30
a^0	b^1	5
a^1	b^0	1
a^1	b^1	10

Pairwise Markov Networks

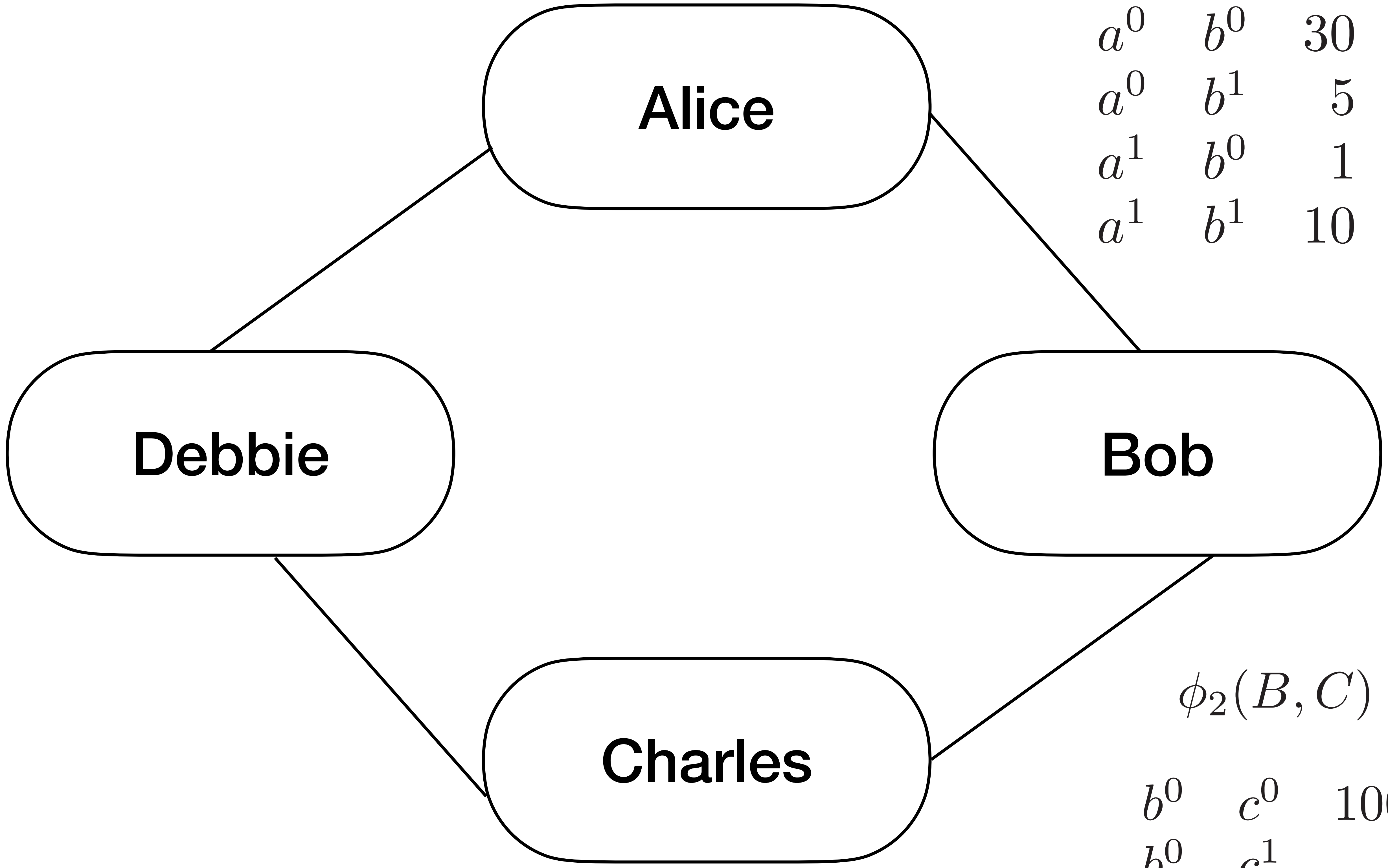


$$\phi_1(A, B)$$

a^0	b^0	30
a^0	b^1	5
a^1	b^0	1
a^1	b^1	10

- affinity
- compatibility
- soft constraints

Pairwise Markov Networks



$$\phi_1(A, B)$$

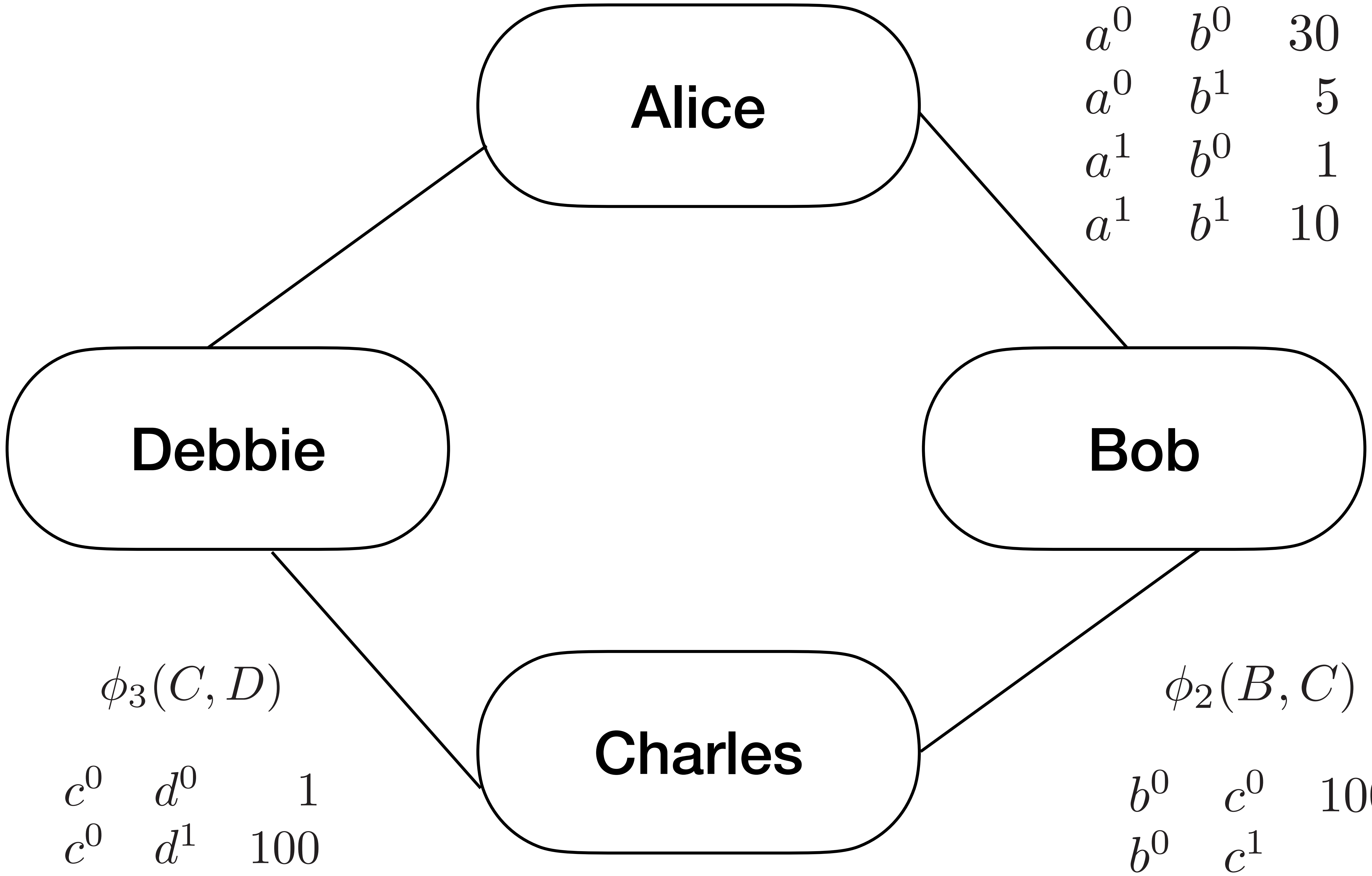
$$\begin{array}{llr} a^0 & b^0 & 30 \\ a^0 & b^1 & 5 \\ a^1 & b^0 & 1 \\ a^1 & b^1 & 10 \end{array}$$

$$\phi_2(B, C)$$

$$\begin{array}{llr} b^0 & c^0 & 100 \\ b^0 & c^1 & 1 \\ b^1 & c^0 & 1 \\ b^1 & c^1 & 100 \end{array}$$

- affinity
- compatibility
- soft constraints

Pairwise Markov Networks



$\phi_1(A, B)$

a^0	b^0	30
a^0	b^1	5
a^1	b^0	1
a^1	b^1	10

$\phi_2(B, C)$

b^0	c^0	100
b^0	c^1	1
b^1	c^0	1
b^1	c^1	100

$\phi_3(C, D)$

c^0	d^0	1
c^0	d^1	100
c^1	d^0	100
c^1	d^1	1

- affinity
- compatibility
- soft constraints

Pairwise Markov Networks

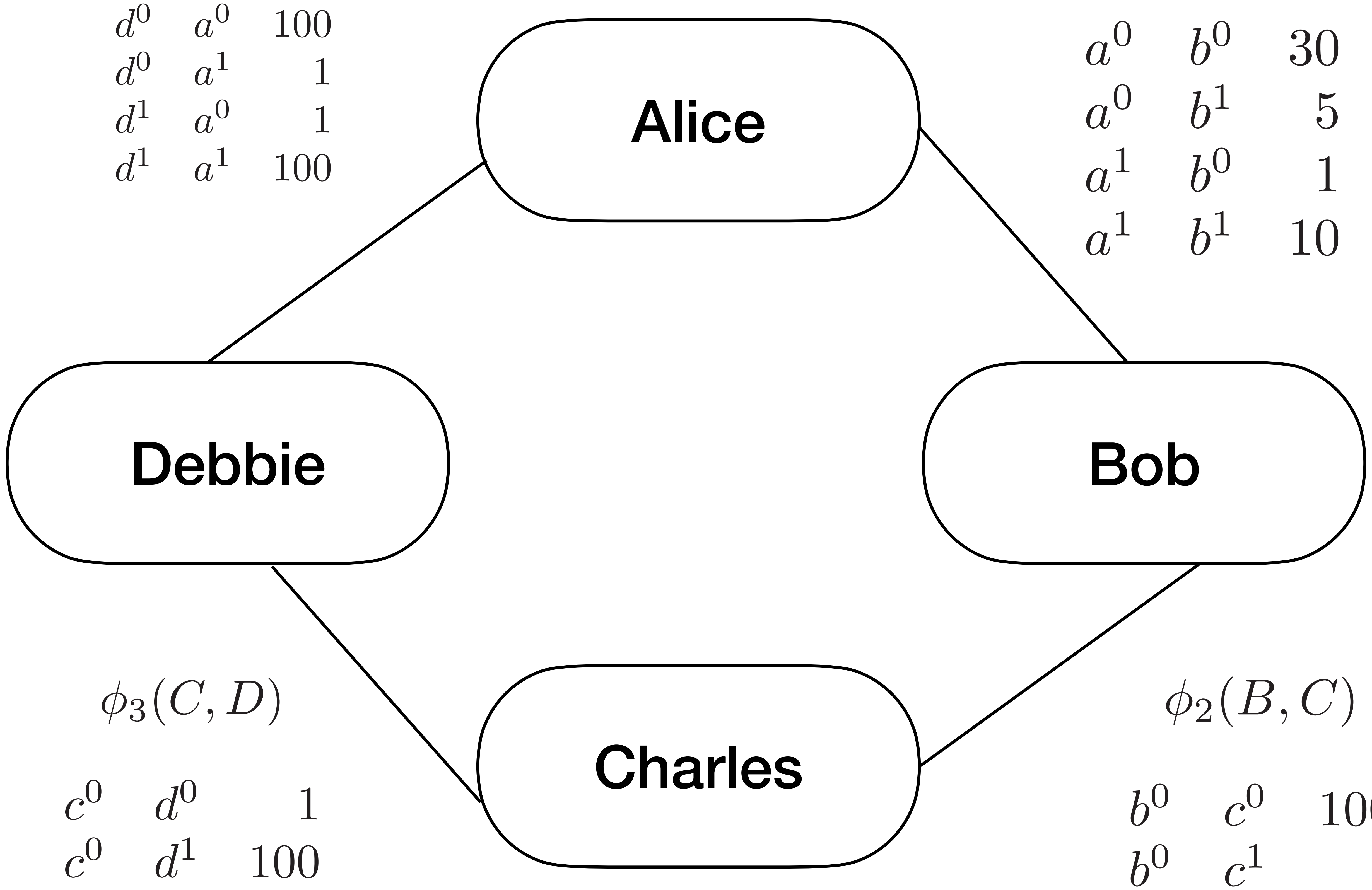
- affinity
- compatibility
- soft constraints

$\phi_4(D, A)$

d^0	a^0	100
d^0	a^1	1
d^1	a^0	1
d^1	a^1	100

$\phi_1(A, B)$

a^0	b^0	30
a^0	b^1	5
a^1	b^0	1
a^1	b^1	10



$\phi_3(C, D)$

c^0	d^0	1
c^0	d^1	100
c^1	d^0	100
c^1	d^1	1

$\phi_2(B, C)$

b^0	c^0	100
b^0	c^1	1
b^1	c^0	1
b^1	c^1	100

$$\tilde{P}(A,B,C,D) = \phi_1(A,B) \times \phi_2(B,C) \times \phi_3(C,D) \times \phi_4(D,A)$$

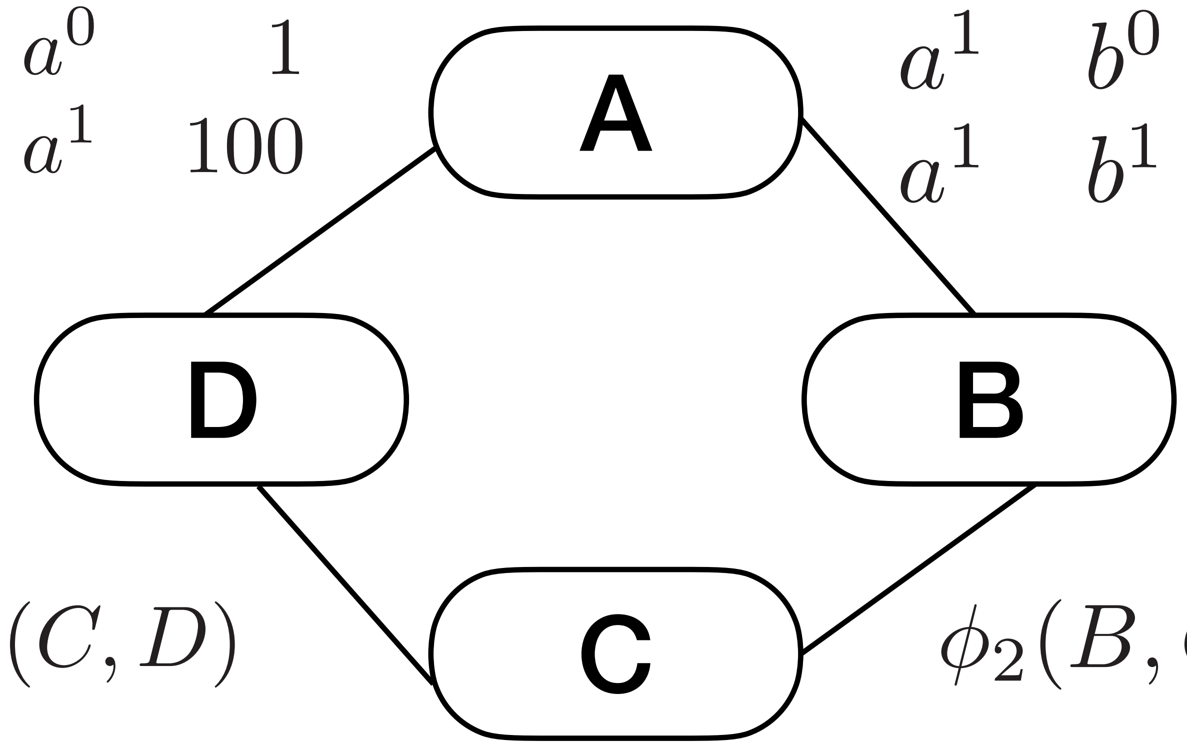
Unnormalised measure

$\phi_4(D,A)$

d^0	a^0	100
d^0	a^1	1
d^1	a^0	1
d^1	a^1	100

$\phi_1(A,B)$

a^0	b^0	30
a^0	b^1	5
a^1	b^0	1
a^1	b^1	10



$\phi_3(C,D)$

c^0	d^0	1
c^0	d^1	100
c^1	d^0	100
c^1	d^1	1

$\phi_2(B,C)$

b^0	c^0	100
b^0	c^1	1
b^1	c^0	1
b^1	c^1	100

$$\tilde{P}(A, B, C, D) = \phi_1(A, B) \times \phi_2(B, C) \times \phi_3(C, D) \times \phi_4(D, A)$$

Unnormalised measure

$$P(A, B, C, D) = \frac{1}{Z}\tilde{P}(A, B, C, D)$$

Partition function

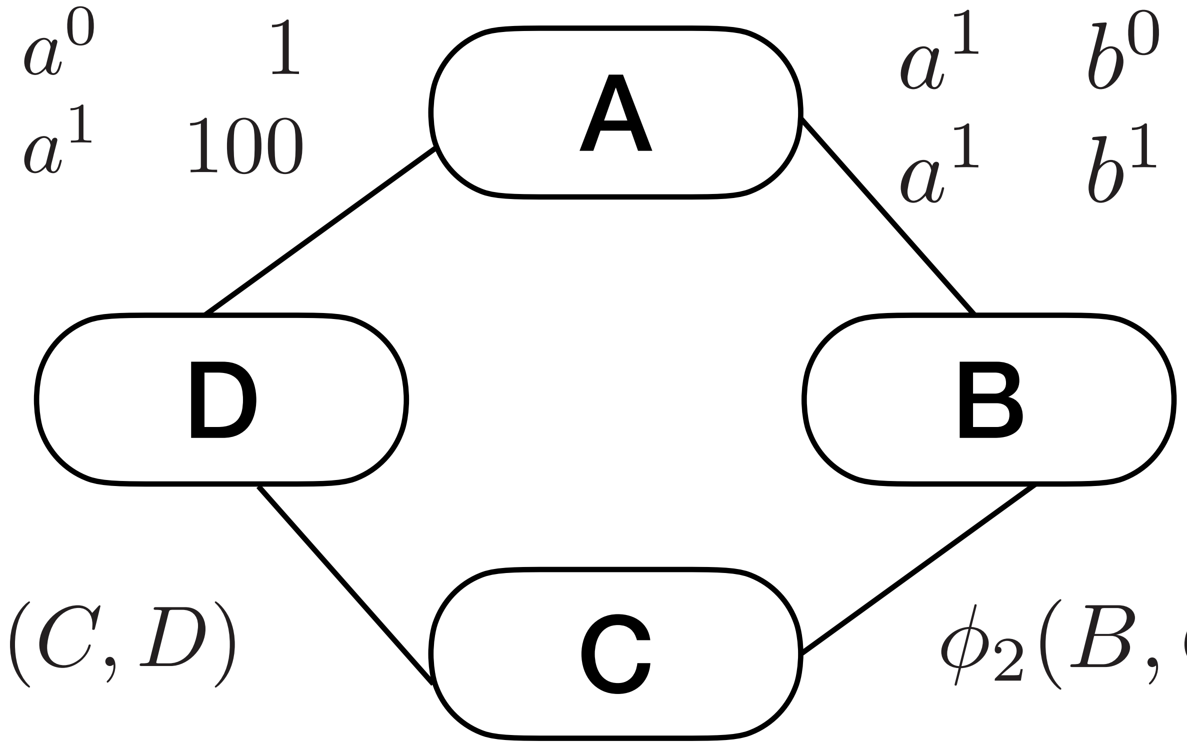
Assignment				*Unnormalized*
a^0	b^0	c^0	d^0	$300,000$
a^0	b^0	c^0	d^1	$300,000$
a^0	b^0	c^1	d^0	$300,000$
a^0	b^0	c^1	d^1	30
a^0	b^1	c^0	d^0	500
a^0	b^1	c^0	d^1	500
a^0	b^1	c^1	d^0	$5,000,000$
a^0	b^1	c^1	d^1	500
a^1	b^0	c^0	d^0	100
a^1	b^0	c^0	d^1	$1,000,000$
a^1	b^0	c^1	d^0	100
a^1	b^0	c^1	d^1	100
a^1	b^1	c^0	d^0	10
a^1	b^1	c^0	d^1	$100,000$
a^1	b^1	c^1	d^0	$100,000$
a^1	b^1	c^1	d^1	$100,000$

$\phi_4(D, A)$

d^0	a^0	100
d^0	a^1	1
d^1	a^0	1
d^1	a^1	100

$\phi_1(A, B)$

a^0	b^0	30
a^0	b^1	5
a^1	b^0	1
a^1	b^1	10



$\phi_3(C, D)$

c^0	d^0	1
c^0	d^1	100
c^1	d^0	100
c^1	d^1	1

$\phi_2(B, C)$

b^0	c^0	100
b^0	c^1	1
b^1	c^0	1
b^1	c^1	100

$$\tilde{P}(A,B,C,D) = \phi_1(A,B) \times \phi_2(B,C) \times \phi_3(C,D) \times \phi_4(D,A)$$

Unnormalised measure

$$P(A,B,C,D) = \frac{1}{Z}\tilde{P}(A,B,C,D)$$

Partition function

Assignment				*Unnormalized*	*Normalized*
a^0	b^0	c^0	d^0	$300,000$	0.04
a^0	b^0	c^0	d^1	$300,000$	0.04
a^0	b^0	c^1	d^0	$300,000$	0.04
a^0	b^0	c^1	d^1	30	$4.1 \cdot 10^{-6}$
a^0	b^1	c^0	d^0	500	$6.9 \cdot 10^{-5}$
a^0	b^1	c^0	d^1	500	$6.9 \cdot 10^{-5}$
a^0	b^1	c^1	d^0	$5,000,000$	0.69
a^0	b^1	c^1	d^1	500	$6.9 \cdot 10^{-5}$
a^1	b^0	c^0	d^0	100	$1.4 \cdot 10^{-5}$
a^1	b^0	c^0	d^1	$1,000,000$	0.14
a^1	b^0	c^1	d^0	100	$1.4 \cdot 10^{-5}$
a^1	b^0	c^1	d^1	100	$1.4 \cdot 10^{-5}$
a^1	b^1	c^0	d^0	10	$1.4 \cdot 10^{-6}$
a^1	b^1	c^0	d^1	$100,000$	0.014
a^1	b^1	c^1	d^0	$100,000$	0.014
a^1	b^1	c^1	d^1	$100,000$	0.014

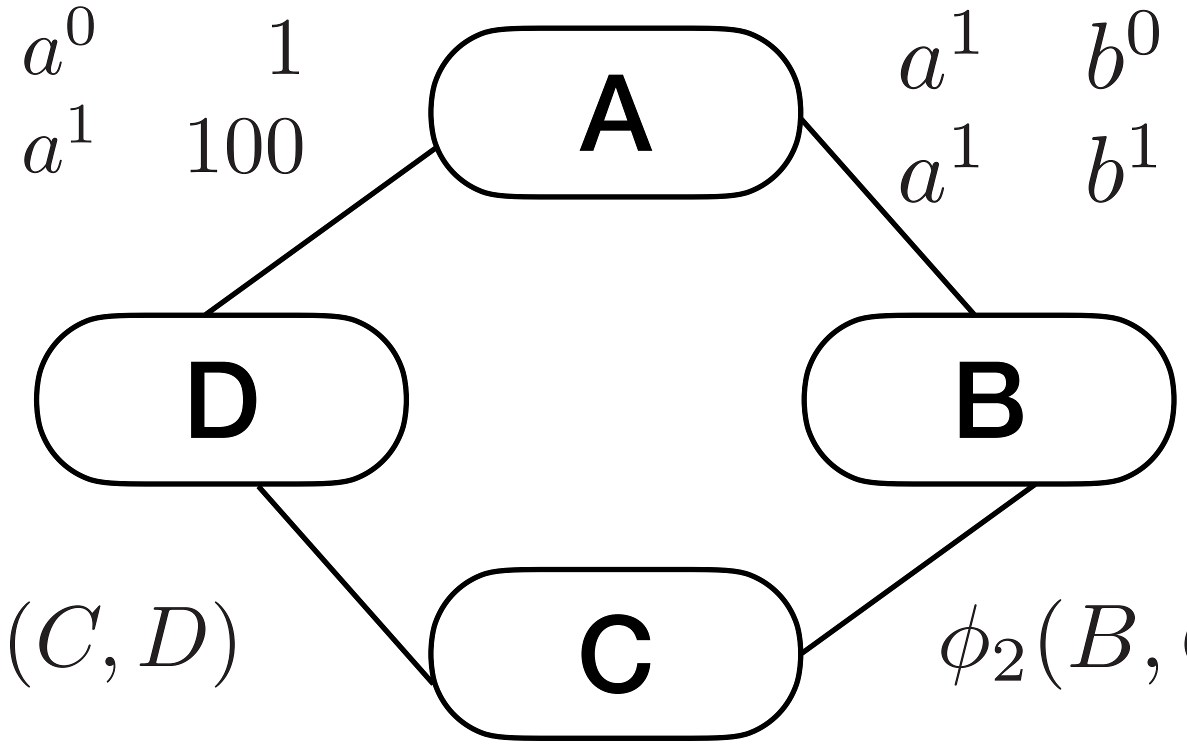
Z

$\phi_4(D,A)$

d^0	a^0	100
d^0	a^1	1
d^1	a^0	1
d^1	a^1	100

$\phi_1(A,B)$

a^0	b^0	30
a^0	b^1	5
a^1	b^0	1
a^1	b^1	10



$\phi_3(C,D)$

c^0	d^0	1
c^0	d^1	100
c^1	d^0	100
c^1	d^1	1

$\phi_2(B,C)$

b^0	c^0	100
b^0	c^1	1
b^1	c^0	1
b^1	c^1	100

What does the pairwise factor $\phi_1(A, B)$ mean?

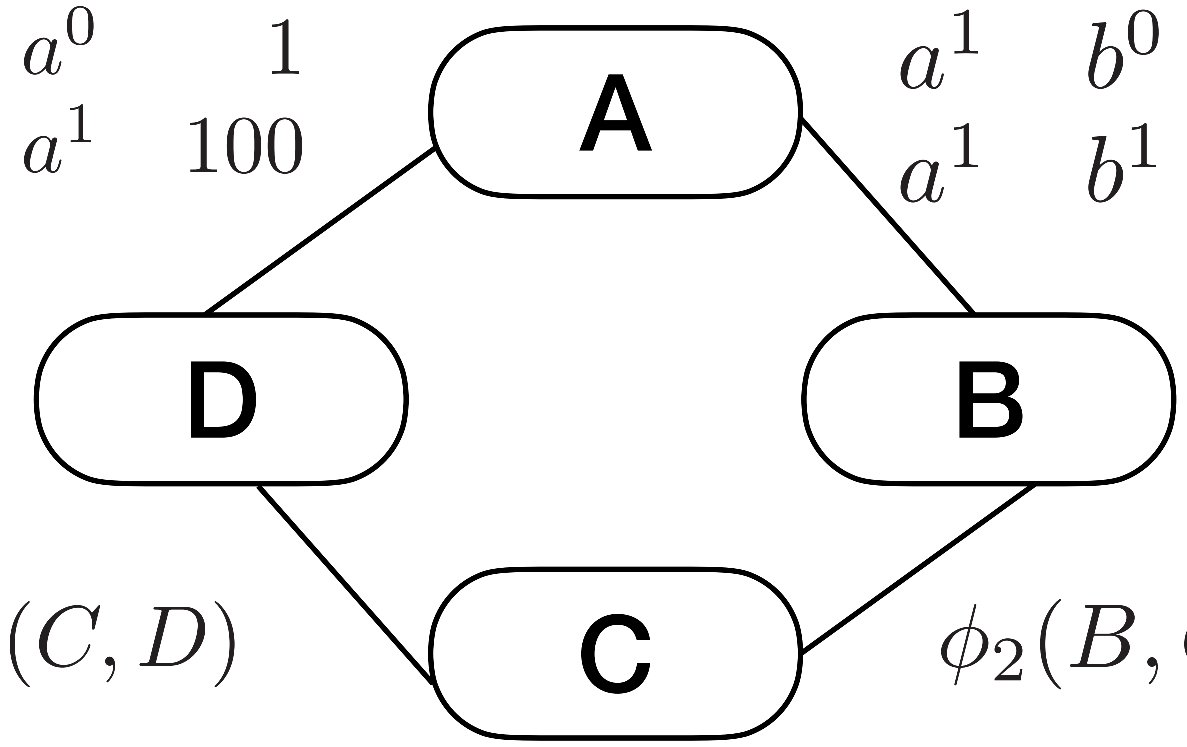
It is the local “happiness” between A and B. How does it relate to a probability distribution?

$\phi_4(D, A)$

d^0	a^0	100
d^0	a^1	1
d^1	a^0	1
d^1	a^1	100

$\phi_1(A, B)$

a^0	b^0	30
a^0	b^1	5
a^1	b^0	1
a^1	b^1	10



$\phi_3(C, D)$

c^0	d^0	1
c^0	d^1	100
c^1	d^0	100
c^1	d^1	1

$\phi_2(B, C)$

b^0	c^0	100
b^0	c^1	1
b^1	c^0	1
b^1	c^1	100

What does the pairwise factor $\phi_1(A, B)$ mean?

It is the local "happiness" between A and B. How does it relate to a probability distribution?

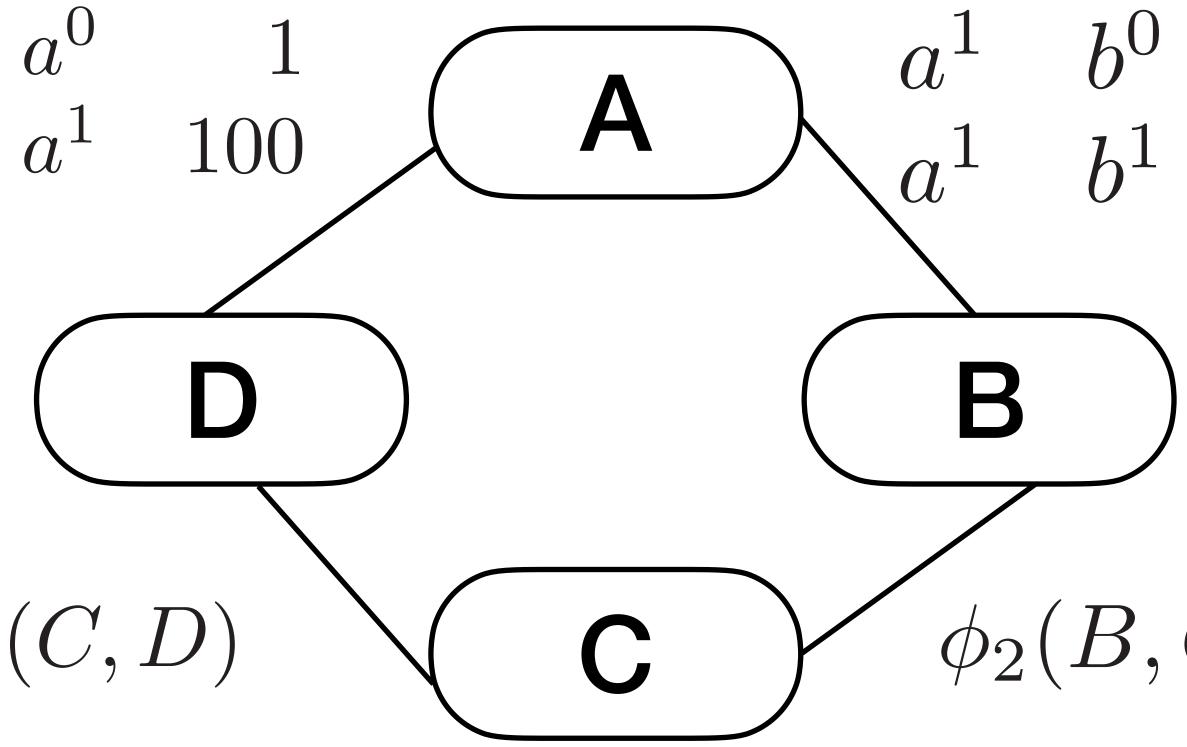
- $P(A, B)$
- $P(A \mid B)$
- $P(A, B \mid C, D)$
- ?

$\phi_4(D, A)$

d^0	a^0	100
d^0	a^1	1
d^1	a^0	1
d^1	a^1	100

$\phi_1(A, B)$

a^0	b^0	30
a^0	b^1	5
a^1	b^0	1
a^1	b^1	10



$\phi_3(C, D)$

c^0	d^0	1
c^0	d^1	100
c^1	d^0	100
c^1	d^1	1

$\phi_2(B, C)$

b^0	c^0	100
b^0	c^1	1
b^1	c^0	1
b^1	c^1	100

What does the pairwise factor $\phi_1(A, B)$ mean?

$P_\phi(A, B)$

a^0	b^0	0.13
a^0	b^1	0.69
a^1	b^0	0.14
a^1	b^1	0.04

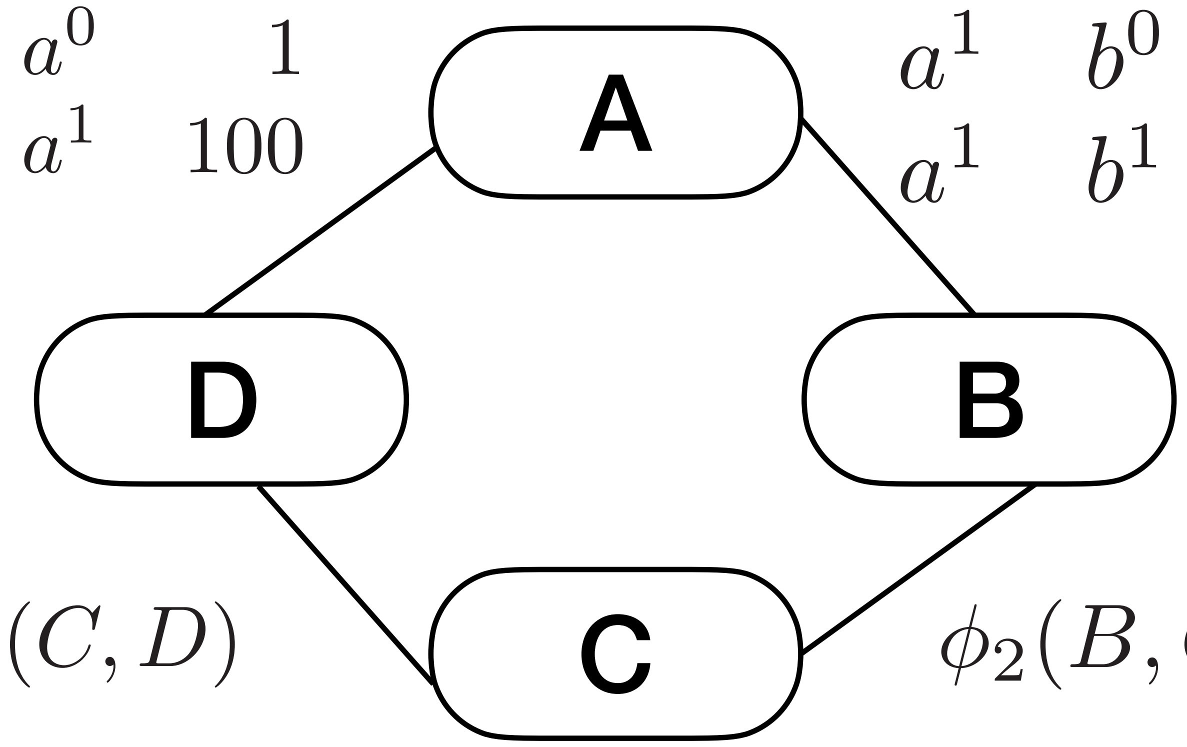
$$\phi = \{\phi_1, \phi_2, \phi_3, \phi_4\}$$

$\phi_4(D, A)$

d^0	a^0	100
d^0	a^1	1
d^1	a^0	1
d^1	a^1	100

$\phi_1(A, B)$

a^0	b^0	30
a^0	b^1	5
a^1	b^0	1
a^1	b^1	10



$\phi_3(C, D)$

c^0	d^0	1
c^0	d^1	100
c^1	d^0	100
c^1	d^1	1

$\phi_2(B, C)$

b^0	c^0	100
b^0	c^1	1
b^1	c^0	1
b^1	c^1	100

What does the pairwise factor $\phi_1(A, B)$ mean?

$P_\phi(A, B)$

a^0	b^0	0.13
a^0	b^1	0.69
a^1	b^0	0.14
a^1	b^1	0.04

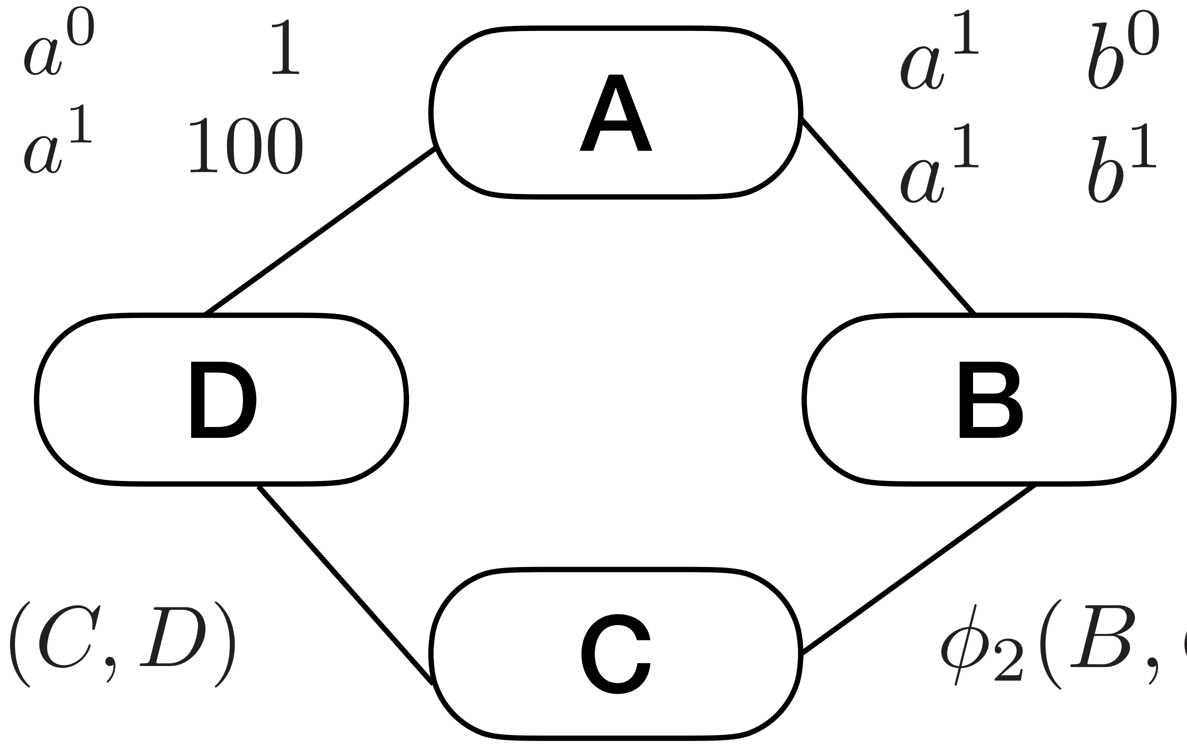
$\phi = \{\phi_1, \phi_2, \phi_3, \phi_4\}$

$\phi_4(D, A)$

d^0	a^0	100
d^0	a^1	1
d^1	a^0	1
d^1	a^1	100

$\phi_1(A, B)$

a^0	b^0	30
a^0	b^1	5
a^1	b^0	1
a^1	b^1	10



$\phi_3(C, D)$

c^0	d^0	1
c^0	d^1	100
c^1	d^0	100
c^1	d^1	1

$\phi_2(B, C)$

b^0	c^0	100
b^0	c^1	1
b^1	c^0	1
b^1	c^1	100

What does the pairwise factor $\phi_1(A, B)$ mean?

$P_\phi(A, B)$

a^0	b^0	0.13
a^0	b^1	0.69
a^1	b^0	0.14
a^1	b^1	0.04

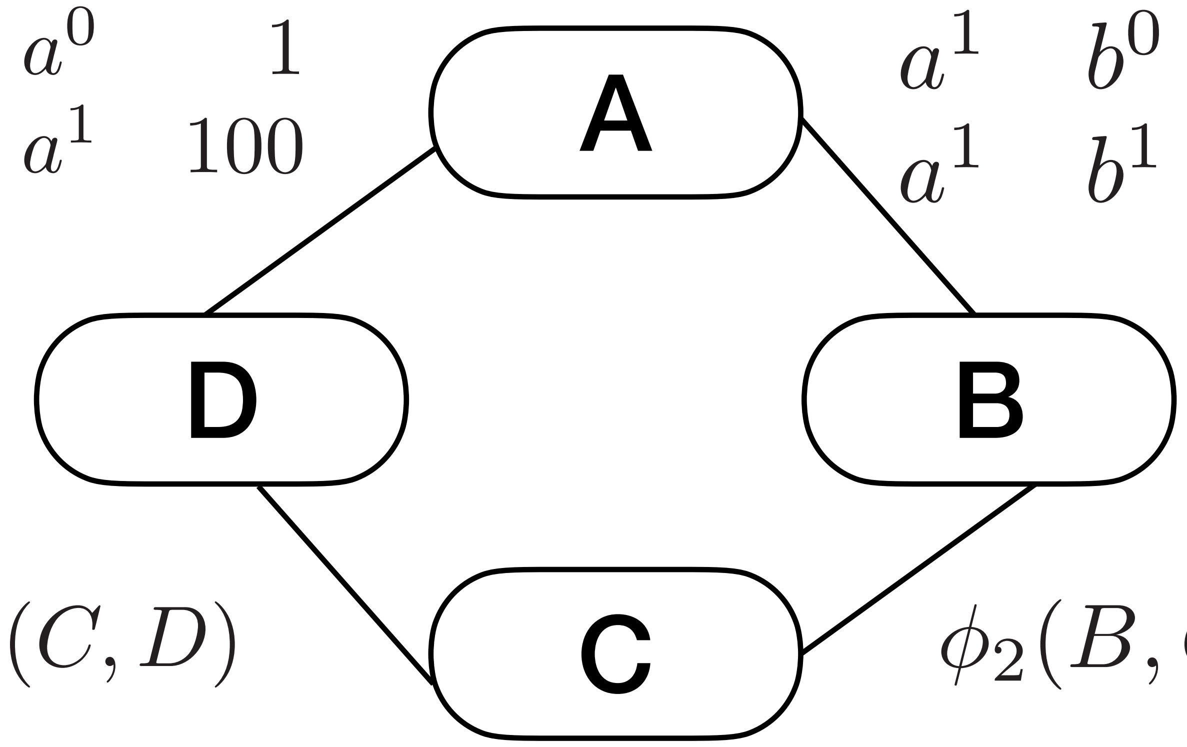
$\phi = \{\phi_1, \phi_2, \phi_3, \phi_4\}$

$\phi_4(D, A)$

d^0	a^0	100
d^0	a^1	1
d^1	a^0	1
d^1	a^1	100

$\phi_1(A, B)$

a^0	b^0	30
a^0	b^1	5
a^1	b^0	1
a^1	b^1	10



$\phi_3(C, D)$

c^0	d^0	1
c^0	d^1	100
c^1	d^0	100
c^1	d^1	1

$\phi_2(B, C)$

b^0	c^0	100
b^0	c^1	1
b^1	c^0	1
b^1	c^1	100

What does the pairwise factor $\phi_1(A, B)$ mean?

$P_\phi(A, B)$

a^0	b^0	0.13
a^0	b^1	0.69
a^1	b^0	0.14
a^1	b^1	0.04

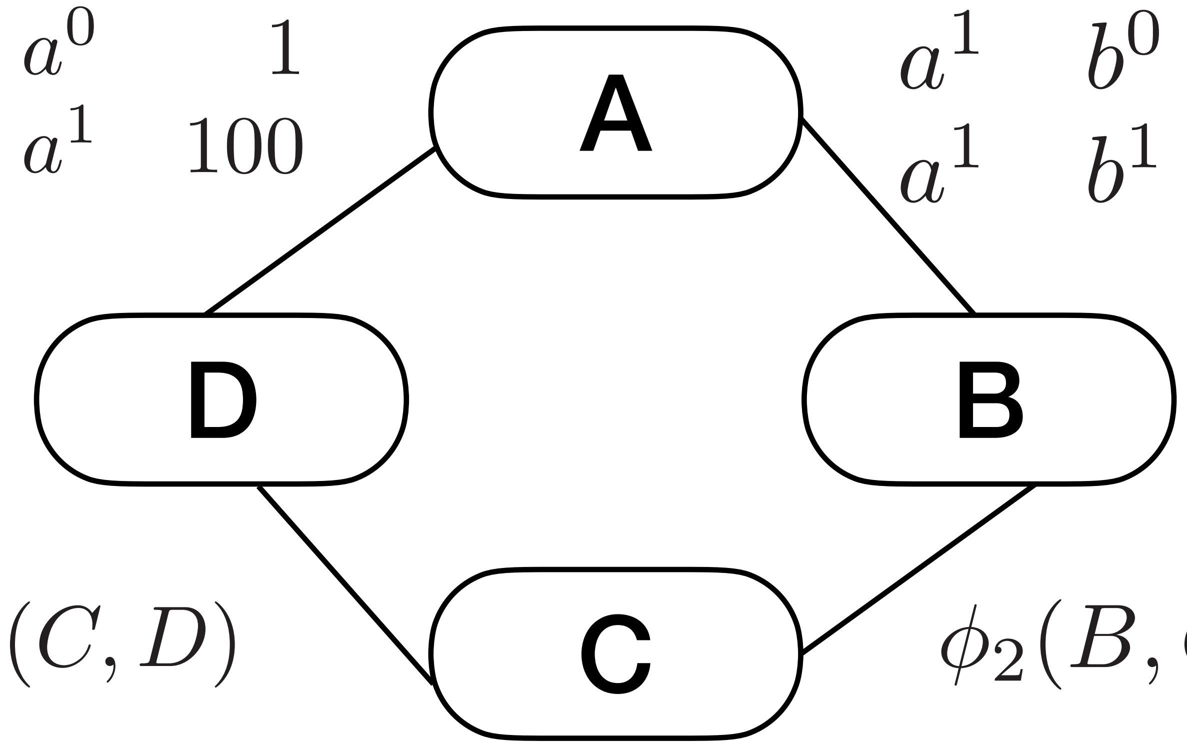
$$\phi = \{\phi_1, \phi_2, \phi_3, \phi_4\}$$

$\phi_4(D, A)$

d^0	a^0	100
d^0	a^1	1
d^1	a^0	1
d^1	a^1	100

$\phi_1(A, B)$

a^0	b^0	30
a^0	b^1	5
a^1	b^0	1
a^1	b^1	10



$\phi_3(C, D)$

c^0	d^0	1
c^0	d^1	100
c^1	d^0	100
c^1	d^1	1

$\phi_2(B, C)$

b^0	c^0	100
b^0	c^1	1
b^1	c^0	1
b^1	c^1	100

What does the pairwise factor $\phi_1(A, B)$ mean?

$P_\phi(A, B)$

a^0	b^0	0.13
a^0	b^1	0.69
a^1	b^0	0.14
a^1	b^1	0.04

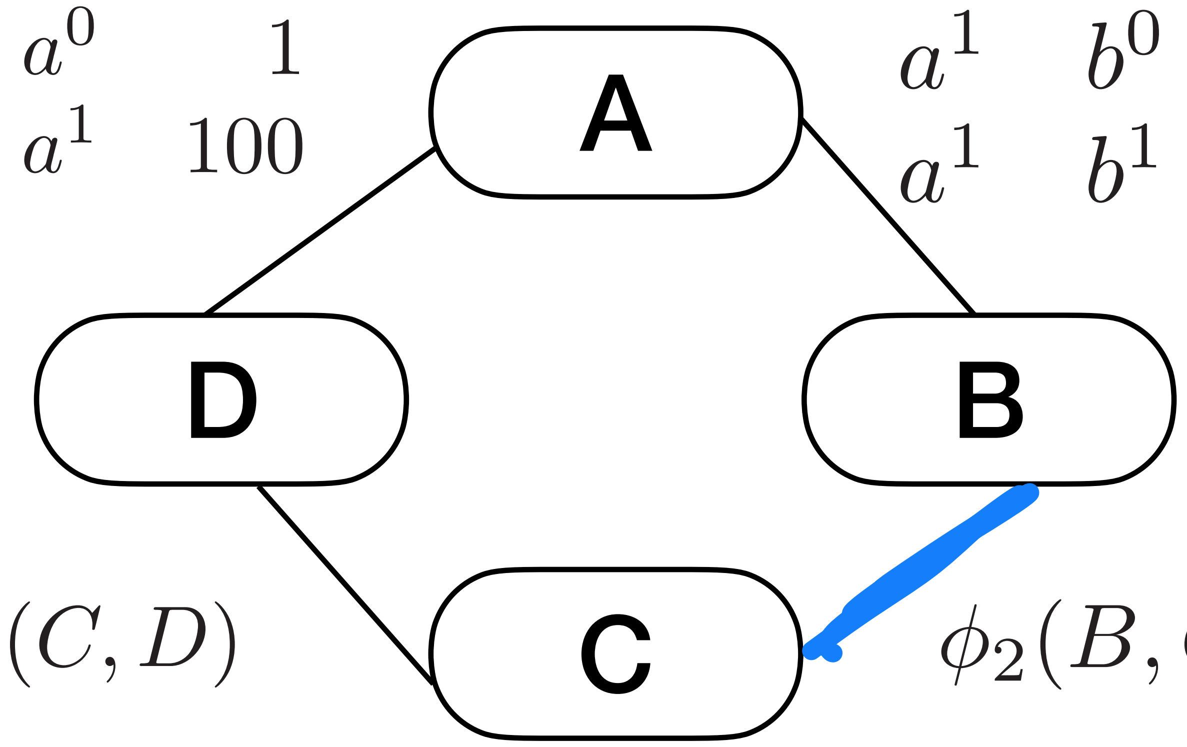
$$\phi = \{\phi_1, \phi_2, \phi_3, \phi_4\}$$

$\phi_4(D, A)$

d^0	a^0	100
d^0	a^1	1
d^1	a^0	1
d^1	a^1	100

$\phi_1(A, B)$

a^0	b^0	30
a^0	b^1	5
a^1	b^0	1
a^1	b^1	10



$\phi_3(C, D)$

c^0	d^0	1
c^0	d^1	100
c^1	d^0	100
c^1	d^1	1

$\phi_2(B, C)$

b^0	c^0	100
b^0	c^1	1
b^1	c^0	1
b^1	c^1	100

What does the pairwise factor $\phi_1(A, B)$ mean?

$$P_\phi(A, B)$$

a^0	b^0	0.13
a^0	b^1	0.69
a^1	b^0	0.14
a^1	b^1	0.04

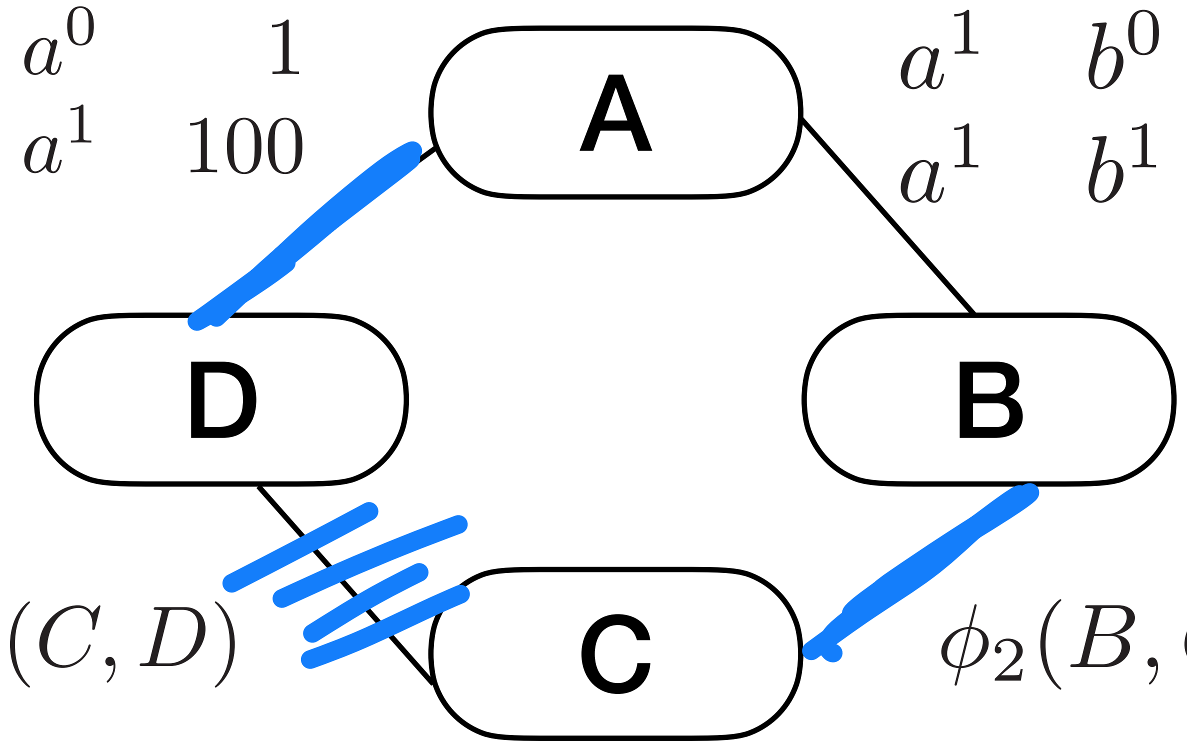
$$\phi = \{\phi_1, \phi_2, \phi_3, \phi_4\}$$

$\phi_4(D, A)$

d^0	a^0	100
d^0	a^1	1
d^1	a^0	1
d^1	a^1	100

$\phi_1(A, B)$

a^0	b^0	30
a^0	b^1	5
a^1	b^0	1
a^1	b^1	10



$\phi_3(C, D)$

c^0	d^0	1
c^0	d^1	100
c^1	d^0	100
c^1	d^1	1

$\phi_2(B, C)$

b^0	c^0	100
b^0	c^1	1
b^1	c^0	1
b^1	c^1	100

Pairwise Markov Networks

- A pairwise Markov network is an undirected graph whose nodes are random variables $X_1, \ldots, X_n$

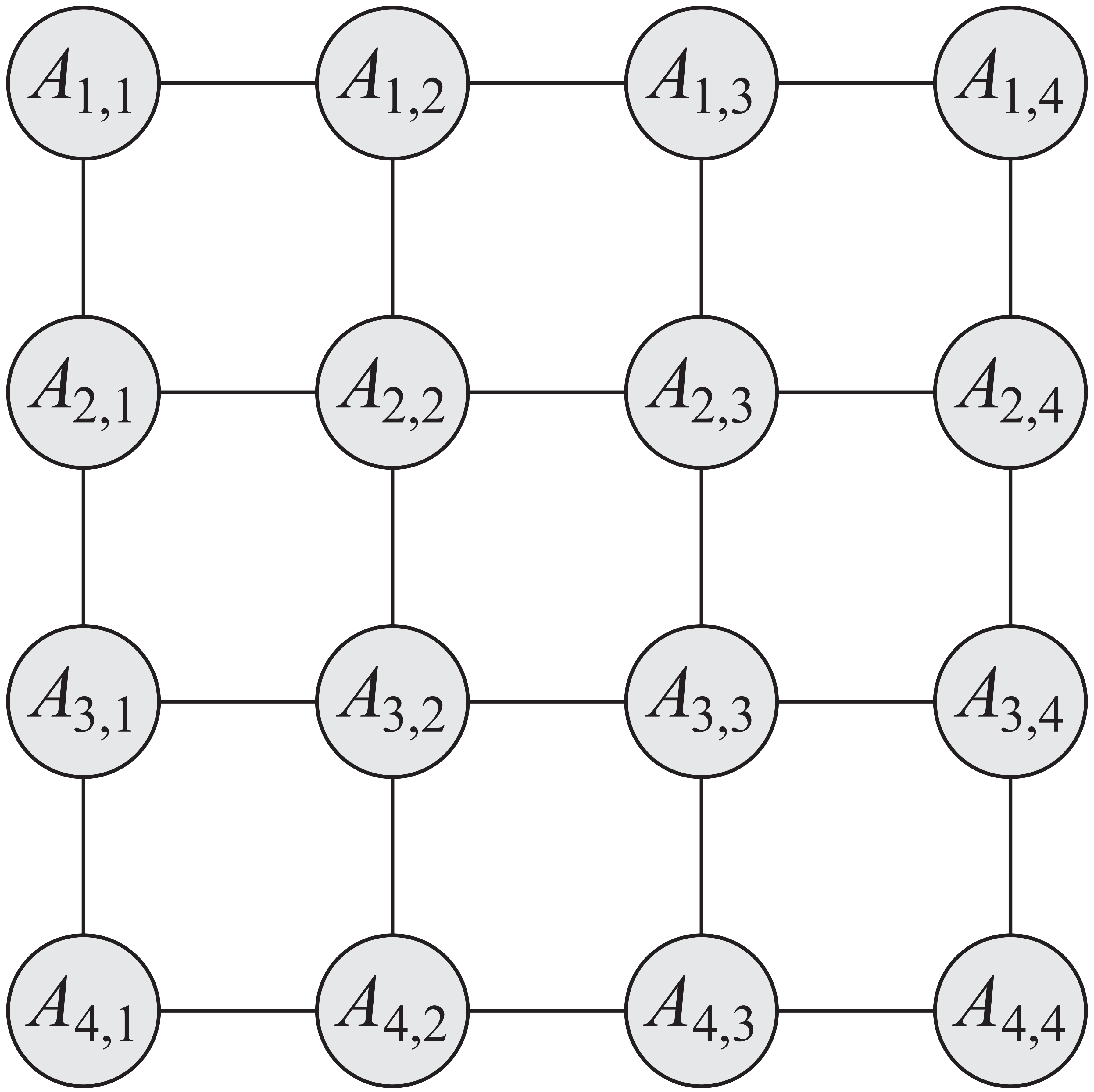
  and each edge $X_i - X_j$ is associated with a factor (aka potential) $\phi_{ij}(X_i, X_j)$

Pairwise Markov Networks

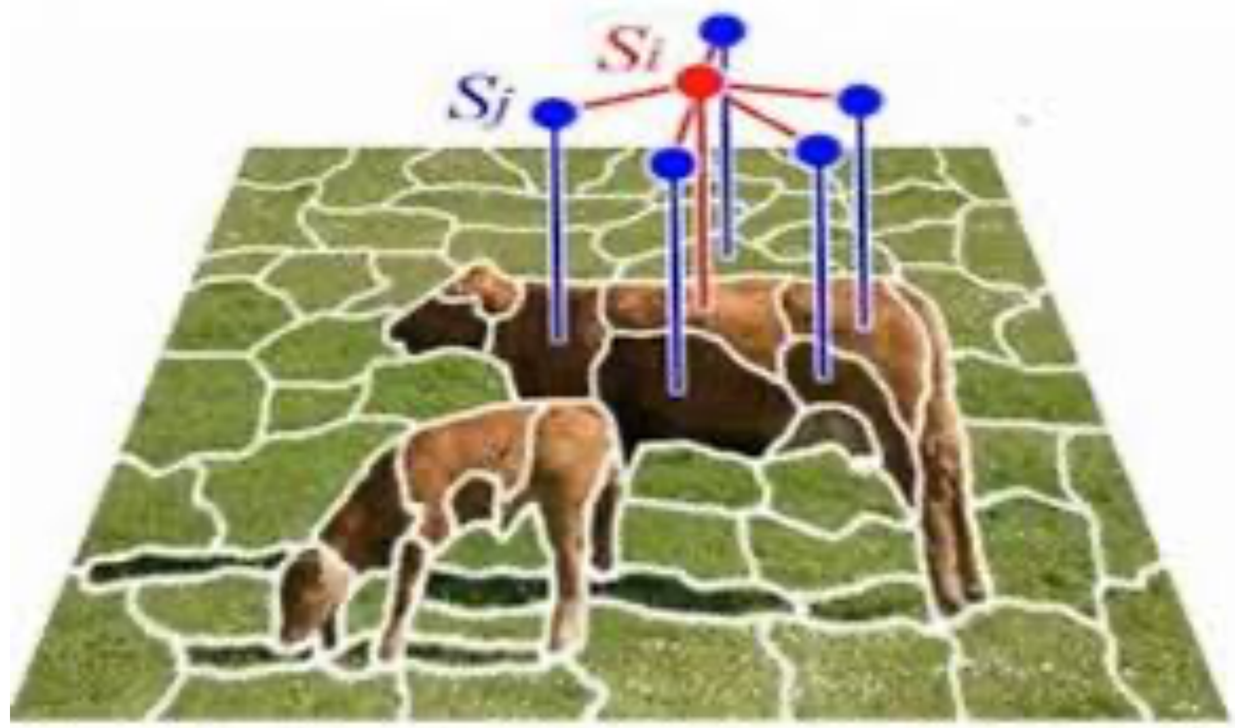
- A pairwise Markov network is an undirected graph whose nodes are random variables $X_1, \ldots, X_n$

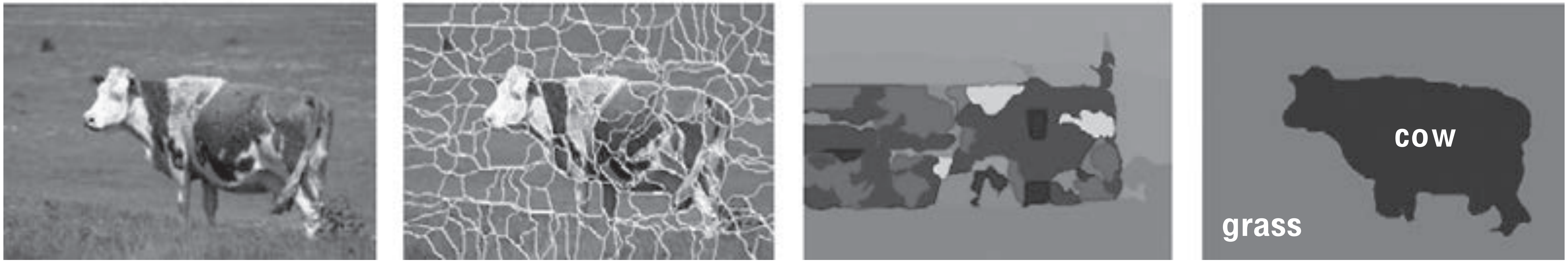
  and each edge $X_i - X_j$ is associated with a factor (aka potential) $\phi_{ij}(X_i, X_j)$

Pairwise Markov Networks



Pairwise Markov Networks





examples of image segmentation results (a) The original image. (b) An oversegmentation known as superpixels; each superpixel is associated with a random variable that designates its segment assignment. The use of superpixels reduces the size of the problems. (c) Result of segmentation using node potentials alone, so that each superpixel is classified independently. (d) Result of segmentation using a pairwise Markov network encoding interactions between adjacent superpixels.

Summary

- **Bayesian networks**
- Factors
- Markov networks