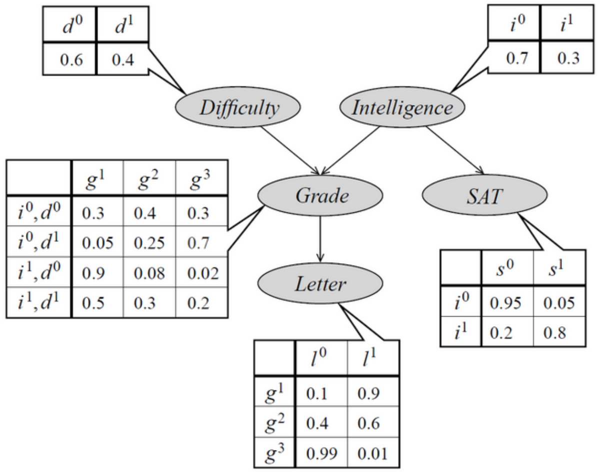
D-Separation

Dr. Xiaowei Huang

https://cgi.csc.liv.ac.uk/~xiaowei/

Up to now,

- Overview of Machine Learning
- Traditional Machine Learning Algorithms
- Deep learning
- Probabilistic Graphical Models
  - Introduction
  - I-Map, Perfect Map
  - Reasoning Patterns (Causal Reasoning, Evidential Reasoning, Intercausal Reasoning)



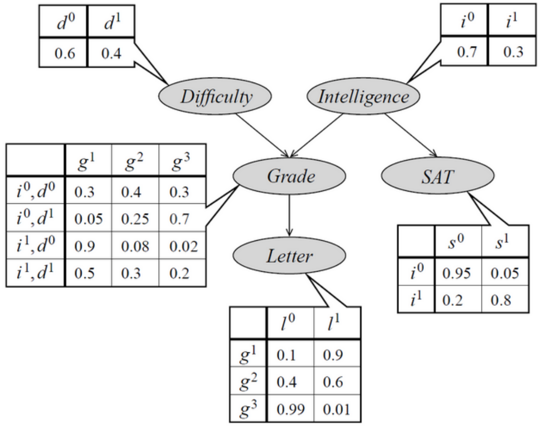
Recap: Local Independencies in a BN

- A BN G is a directed acyclic graph whose nodes represent random variables $X_i,..,X_n$.
- Let $Pa(X_i)$ denote parents of X_i in G
- Let $Non\text{-}Desc(X_i)$ denote variables in G that are not descendants of X_i
- Then G encodes the following set of *conditional independence* assumptions denoted Il(G)
  - For each X_i: $(X_i \perp Non\text{-}Desc(X_i) \mid Pa(X_i))$
- Also known as *Local Markov Independencies*

Recap: Local Independencies

- Graph G with CPDs is equivalent to a set of independence assertions

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



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^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

- Local Conditional Independence Assertions (starting from leaf nodes):

$I(G) = \{(L \perp I,D,S \mid G),$ — L is conditionally independent of all other nodes given parent G

$(S \perp D,G,L \mid I),$ — S is conditionally independent of all other nodes given parent I

$(G \perp S \mid D,I),$ — Even given parents, G is NOT independent of descendant L

$(\mathrm{I} \perp D \mid \phi),$ — Nodes with no parents are marginally independent

$(D \perp I,S \mid \phi)\}$ — D is independent of non-descendants I and S

- Parents of a variable shield it from probabilistic influence
  - Once value of parents known, no influence of ancestors
- Information about descendants can change beliefs about a node

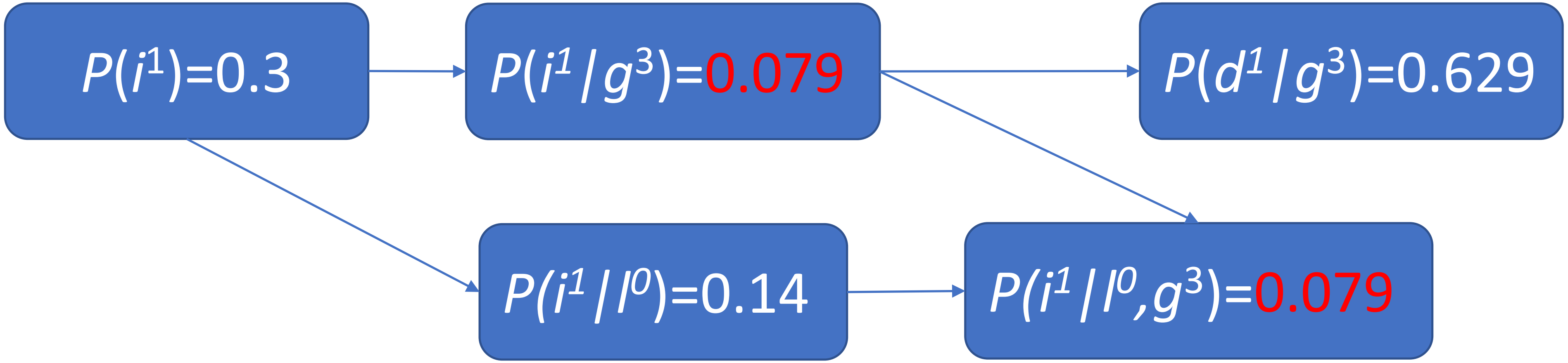
Can we have the following conditional independence?

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

Recap:

low grade drastically decreases the probability of high intelligence

low grade justifies the difficulty



A weak letter drastically decreases the probability of high intelligence

After knowing low grade, a weak letter won't make the probability of high intelligence lower.

$$I \perp L | G$$

Independencies in Graphs

- A graph structure G encodes a set of conditional independence assumptions *I*(G)
- Are there other independencies that we can read-off?
  - i.e., are there other independencies that hold for every distribution that factorizes over G?
- D-separation holds the key

Topics

- Why D-separation?
- What is D-separation?
- Algorithm for D-separation (extended materials)

Why D-separation?

Dependencies and Independencies

- Crucial for understanding network behaviour
- Independence properties are important for answering queries
  - Exploited to reduce computation of inference
  - A distribution P that factorizes over G satisfies $I(G)$

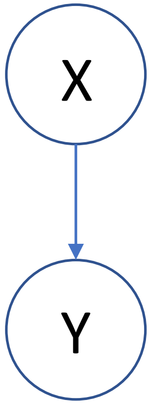
What is D-separation?

D-separation

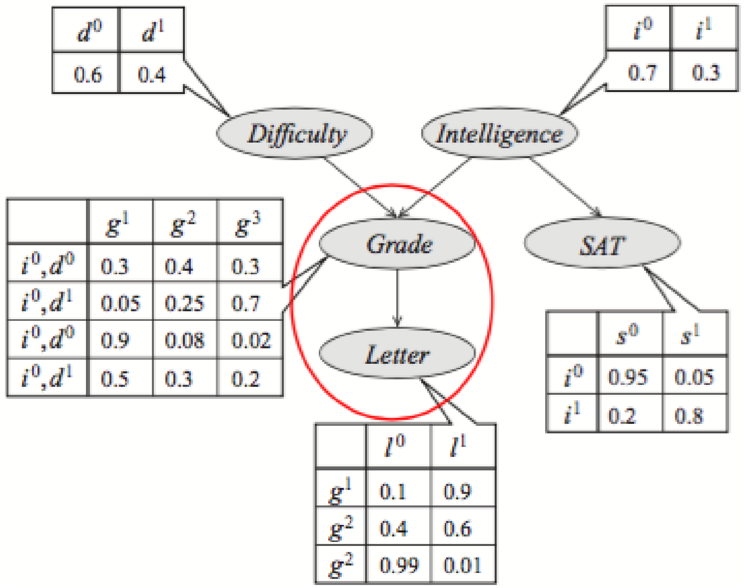
- Study independence properties for subgraphs (connected triples)
- Analyze complex cases in terms of triples along paths between variables
- ***D-separation:*** a condition / algorithm for answering such queries

- Definition: A procedure $\text{d-sep}_G(X \perp Y | Z)$ that given a DAG *G*, and sets *X, Y,* and *Z* returns either *yes* or *no, where* $\text{d-sep}_G(X \perp Y | Z) = \text{Yes}$ iff $(X \perp Y | Z)$ follows from *I(G)*

Direct Connection between X and Y



- X and Y are correlated regardless of any evidence about any other variables
  - E.g., Feature Y and character X are correlated
  - Grade G and Letter L are correlated
- If X and Y are directly connected we can get examples where they influence each other regardless of Z



d^0	d^1
0.6	0.4

i^0	i^1
0.7	0.3

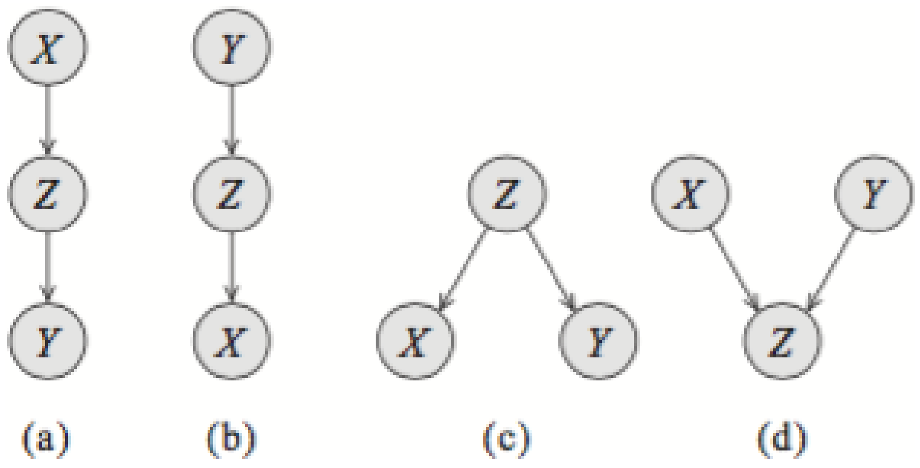
	g^1	g^2	g^3
i^0,d^0	0.3	0.4	0.3
i^0,d^1	0.05	0.25	0.7
i^0,d^0	0.9	0.08	0.02
i^0,d^1	0.5	0.3	0.2

	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

Indirect Connection between X and Y

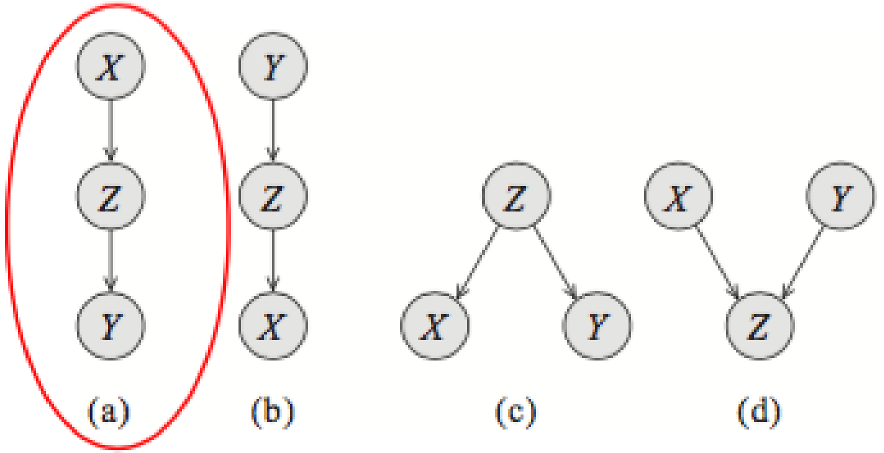
- Four cases where X and Y are connected via Z
- (a). Indirect causal effect
- (b). Indirect evidential effect
- (c). Common cause
- (d). Common effect



- We will see that first three cases are similar while fourth case (*V*-structure) is different

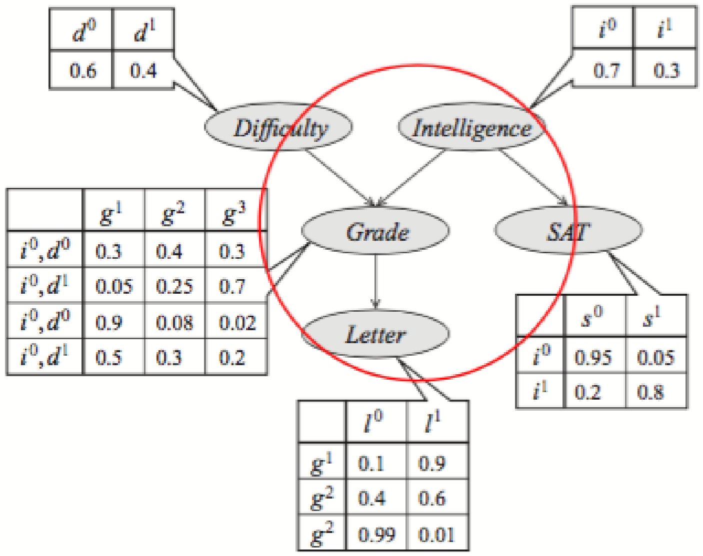
1. Indirect Causal Effect: *X->Z->Y*

- Cause *X* cannot influence effect *Y* if *Z* observed
  - Observed *Z* blocks influence
- If *Grade* observed then *I* does not influence *L*
  - *Intelligence* influences *Letter* if *Grade* is unobserved



Z = *Grade*

$$I \perp L \mid G$$



Recap:

low grade drastically decreases the probability of high intelligence

low grade justifies the difficulty

$P(i^1)=0.3$

$P(i^1 \mid g^3)=0.079$

$P(d^1 \mid g^3)=0.629$

$P(i^1 \mid l^0)=0.14$

$P(i^1 \mid l^0, g^3)=0.079$

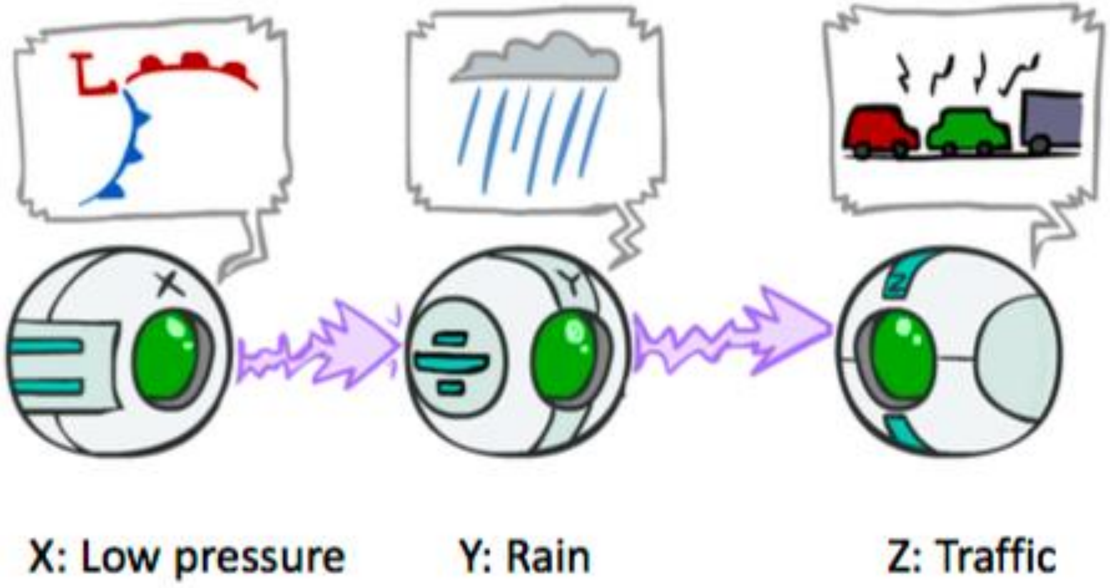
A weak letter drastically decreases the probability of high intelligence

After knowing low grade, a weak letter won't make the probability of high intelligence lower.

$$I \perp L \mid G$$

Causal Chains

- This configuration is a "causal chain"



$$P(x, y, z) = P(x)P(y|x)P(z|y)$$

Guaranteed X independent of Z given Y?

$$P(z|x, y) = \frac{P(x, y, z)}{P(x, y)}$$

$$= \frac{\cancel{P(x)}\cancel{P(y|x)}P(z|y)}{\cancel{P(x)}\cancel{P(y|x)}}$$

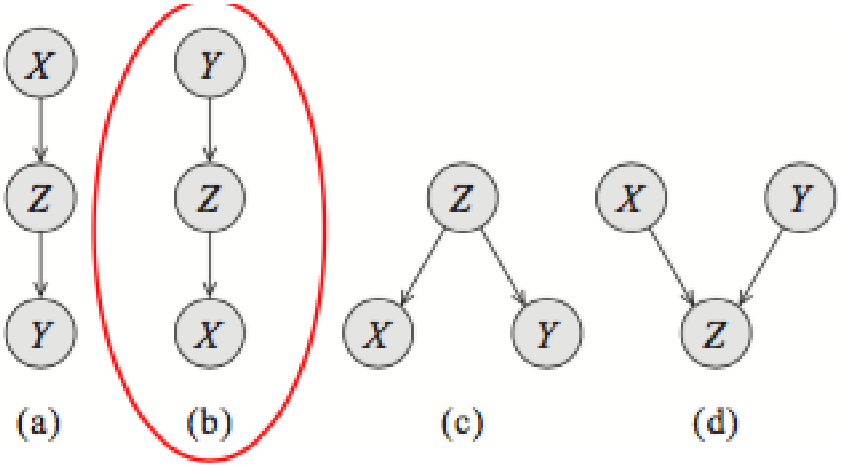
$$= P(z|y)$$

Yes!

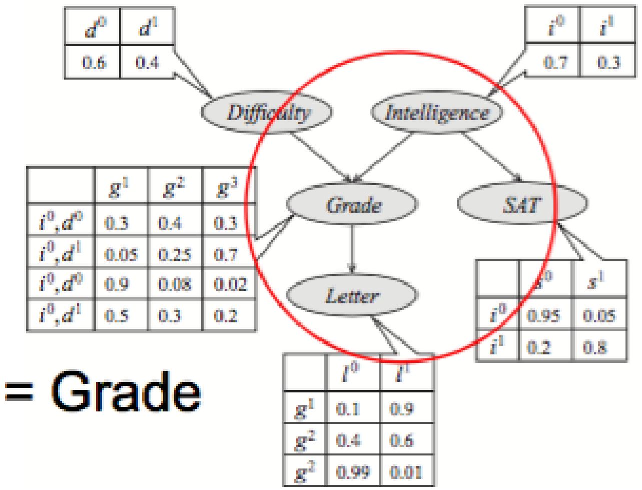
Evidence along the chain "blocks" the influence (makes "inactive")

2. Indirect Evidential Effect: *Y->Z->X*

- Evidence *X* can influence *Y* via *Z* only if *Z* is unobserved
  - Observed *Z* blocks influence
- If Grade unobserved, Letter influences assessment of Intelligence
- Dependency is a symmetric notion
  - $X \perp Y$ does not hold then $Y \perp X$ does not hold either

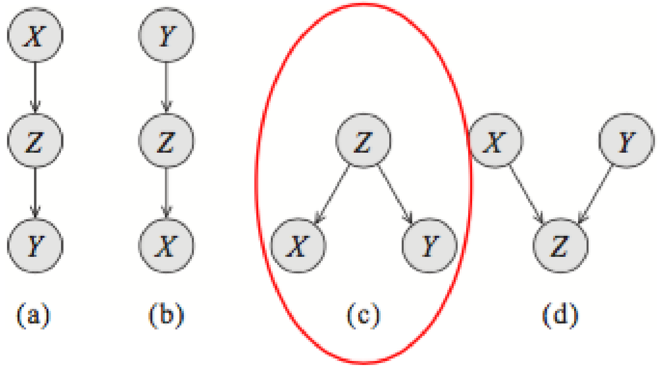


Z = Grade

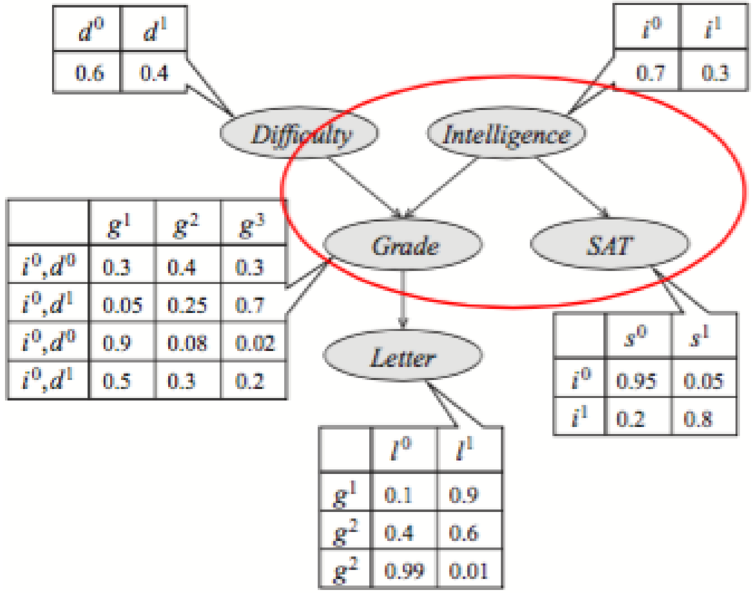


3. Common Cause: *X<-Z->Y*

- *X* can influence *Y* if and only if *Z* is not observed
  - Observed *Z* blocks
- Grade is correlated with SAT score
- But if Intelligence is observed then SAT provides no additional information



$$(S \perp G | I)$$



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

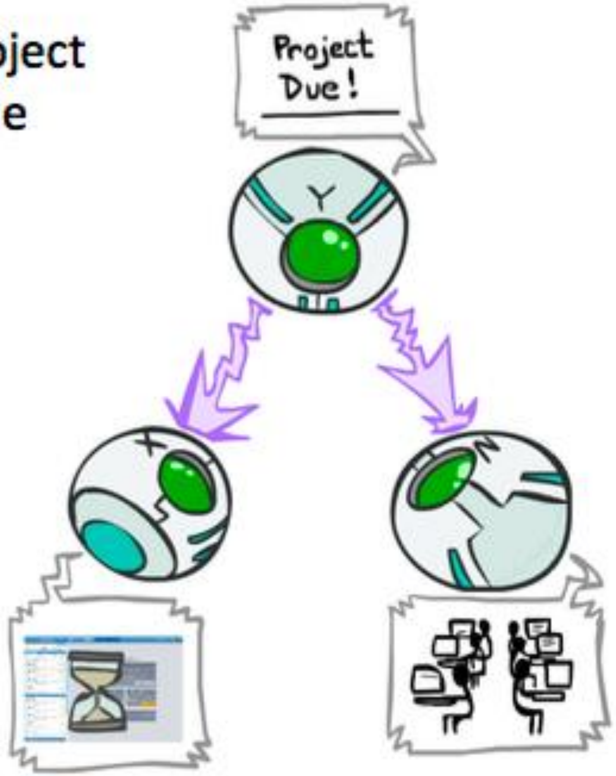
	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

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

Common Cause

- This configuration is a "common cause"

Y: Project due



X: Forums busy

Z: Lab full

$$P(x, y, z) = P(y)P(x|y)P(z|y)$$

- Guaranteed X and Z independent given Y?

$$P(z|x, y) = \frac{P(x, y, z)}{P(x, y)}$$

$$= \frac{\cancel{P(y)}\cancel{P(x|y)}P(z|y)}{\cancel{P(y)}\cancel{P(x|y)}}$$

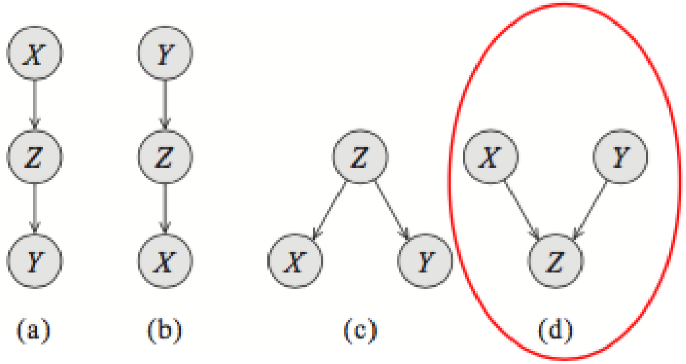
$$= P(z|y)$$

Yes!

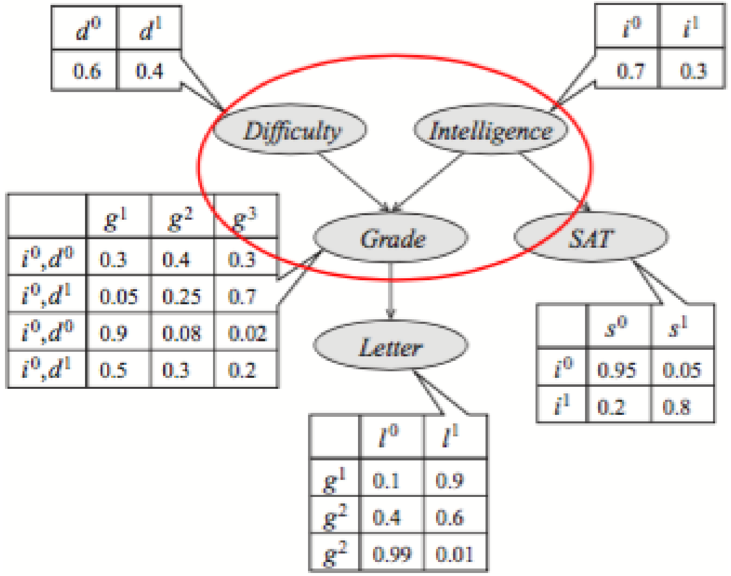
Observing the cause blocks influence between effects. (makes inactive)

4. Common Effect (V-structure) *X->Z<-Y*

- Influence cannot flow on trail *X->Z<-Y* if *Z* is not observed
  - Observed *Z* enables
  - Opposite to previous 3 cases (Observed *Z* blocks)
- When *G* not observed *I* and *D* are independent
- When *G* is observed, *I* and *D* are correlated

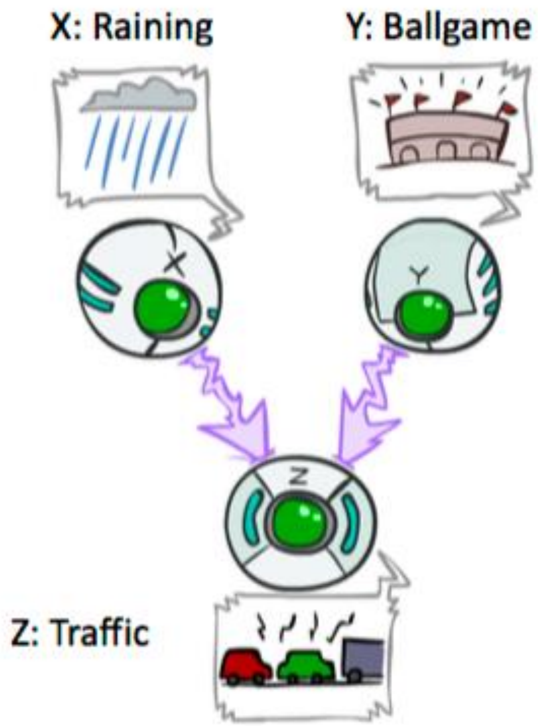


$$I \perp D | \sim G$$



Common Effect

- Last configuration: two causes of one effect (v-structures)

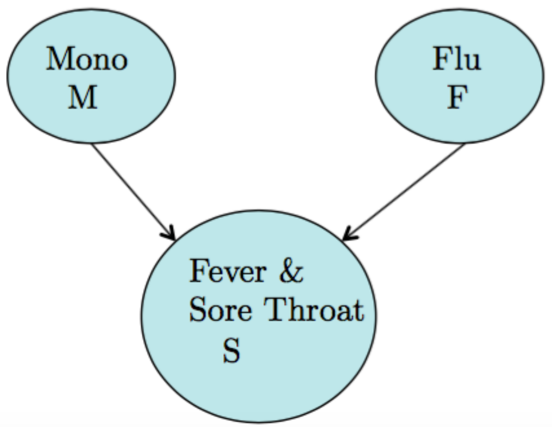


- Are X and Y independent?
  - *Yes*: the ballgame and the rain cause traffic, but they are not correlated
  - Still need to prove they must be (try it!)
- Are X and Y independent given Z?
  - *No*: seeing traffic puts the rain and the ballgame in competition as explanation.
- This is ***backwards*** from the other cases
  - Observing an effect activates influence between possible causes. (makes active!)

Recall: Common in Human Reasoning

- Binary Variables
- Fever & Sore Throat can be caused by mono and flu
- When flu is diagnosed probability of mono is reduced (although mono could still be present)
- It provides an alternative explanation of symptoms

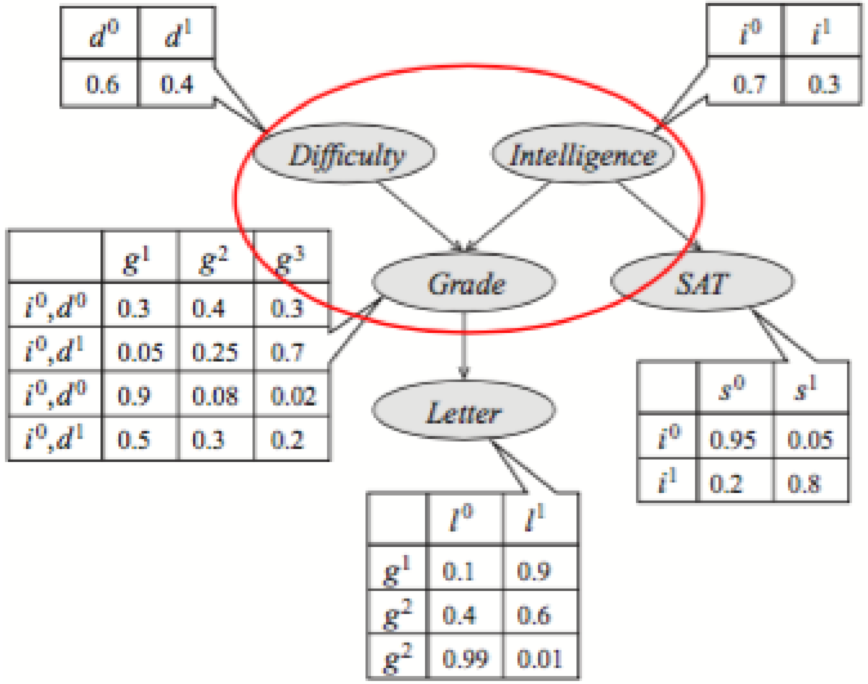
$$P(m^1 | s^1) > P(m^1 | s^1, f^1)$$



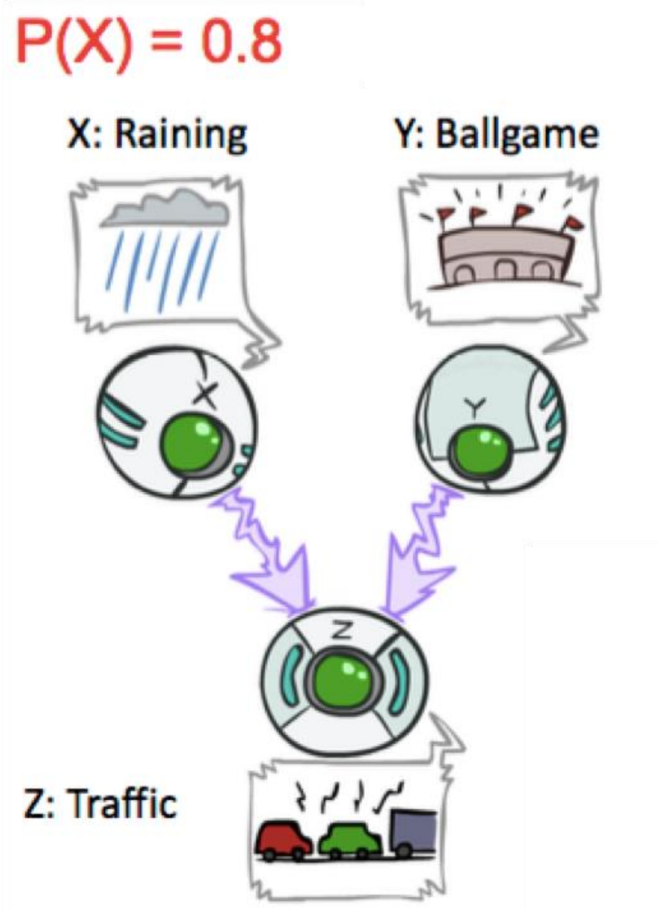
4. Common Effect (V-structure) *X->Z<-Y*

- *Grade* is not observed
- Observe weak letter *L*
  - Which indicates low *Grade*
  - Suffices to correlate *D* and *I*

Not child, but
descendants

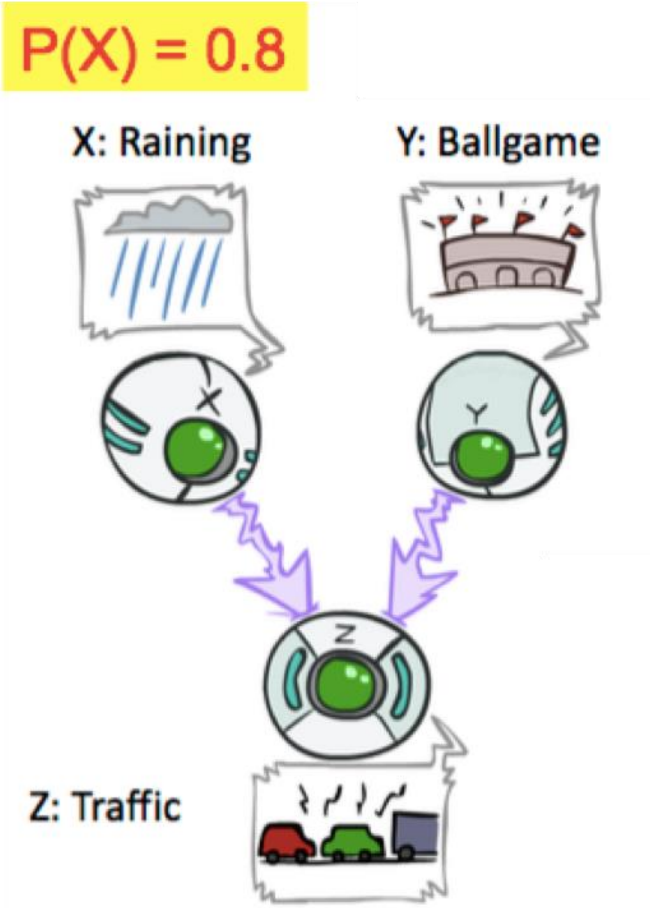


Example: Common Effect



X	Y	Z	P
T	T	T	0.076
T	T	F	0.004
T	F	T	0.576
T	F	F	0.144
F	T	T	0.162
F	T	F	0.002
F	F	T	0.090
F	F	F	0.009

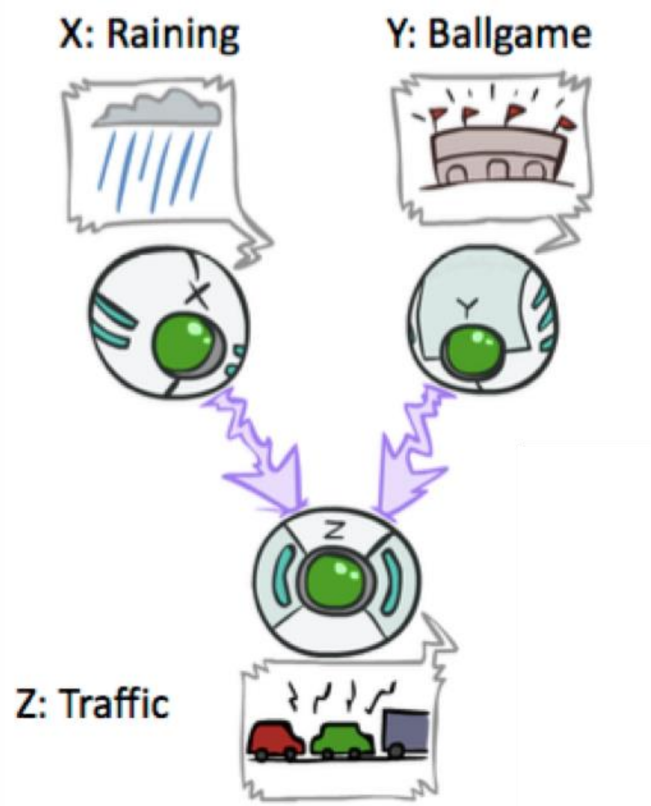
Example: Common Effect



X	Y	Z	P
T	T	T	0.076
T	T	F	0.004
T	F	T	0.576
T	F	F	0.144
F	T	T	0.162
F	T	F	0.002
F	F	T	0.090
F	F	F	0.009

Example: Common Effect

P(X) = 0.8



X	Y	Z	P
T	T	T	0.076
T	T	F	0.004
T	F	T	0.576
T	F	F	0.144
F	T	T	0.018
F	T	F	0.002
F	F	T	0.090
F	F	F	0.009

$$P(X|Y) = \frac{0.076+0.004}{0.076+0.004+0.018+0.002}$$

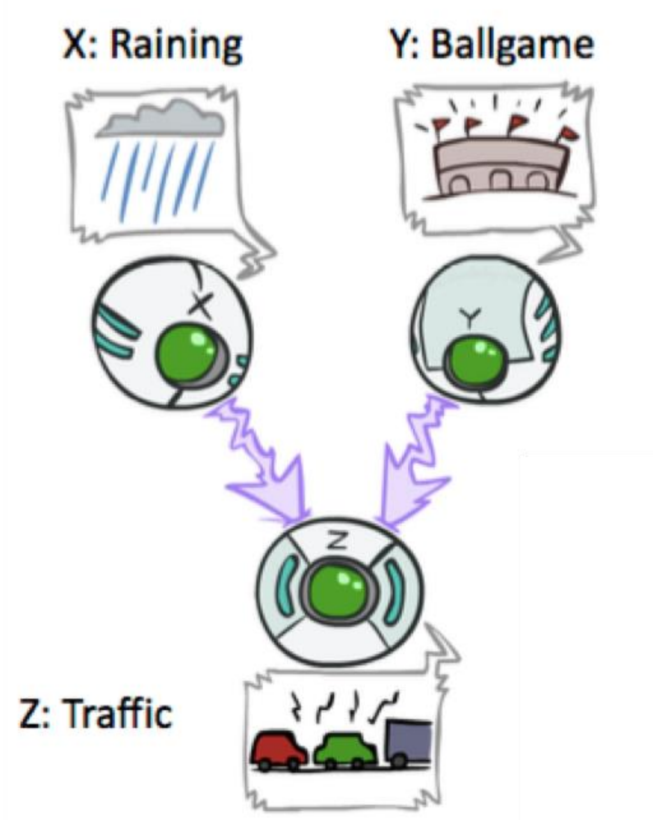
$$= 0.08 / 0.1$$

$$= 0.8$$

X and Y are independent!

But Suppose Also Know Z=T

$P(X) = 0.8$



$$P(X|Z) = \frac{.076+.576}{.076+.576+.018+.090}$$
$$= 0.652/0.76$$
$$= 0.858$$

$$P(X|Y,Z) = \frac{0.076}{0.076+0.018}$$
$$= 0.8085$$

X	Y	Z	P
T	T	T	0.076
T	T	F	0.004
T	F	T	0.576
T	F	F	0.144
F	T	T	0.018
F	T	F	0.002
F	F	T	0.090
F	F	F	0.009

X and Y are not independent given Z!

Summary of Indirect Connection

- Causal trail: *X->Z->Y*: active iff *Z* not observed
- Evidential Trail: *X<-Z<-Y*: active iff *Z* is not observed
- Common Cause: *X<-Z->Y*: active iff *Z* is not observed
- Common Effect: *X->Z<-Y*: active iff either *Z* or one of its descendants is observed

What is the general case?

The General Case

- General question: in a given BN, are two variables independent (given evidence)?

- Solution: analyze the graph

- Any complex example can be broken into repetitions of the three canonical cases

D-Separation

- Query: $X_i \perp X_j | \{X_{k_1}, ..., X_{k_n}\}$?
- Check all (undirected) paths between X_i and X_j
  - If one or more paths is active, then independence not guaranteed

$$X_i \not\perp X_j | \{X_{k_1}, ..., X_{k_n}\}$$

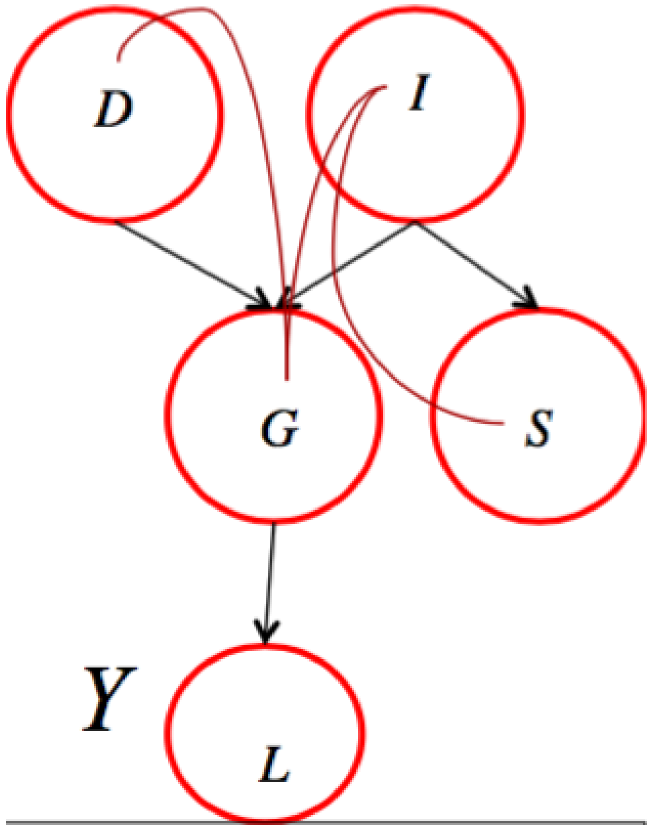
  - Otherwise (i.e. if ***all paths*** are inactive),
    then "D-separated" = independence ***is*** guaranteed

$$X_i \perp X_j | \{X_{k_1}, ..., X_{k_n}\}$$

Active Trail

- When influence can flow from *X* to *Y* via *Z* then trail *X—Z—Y* is active

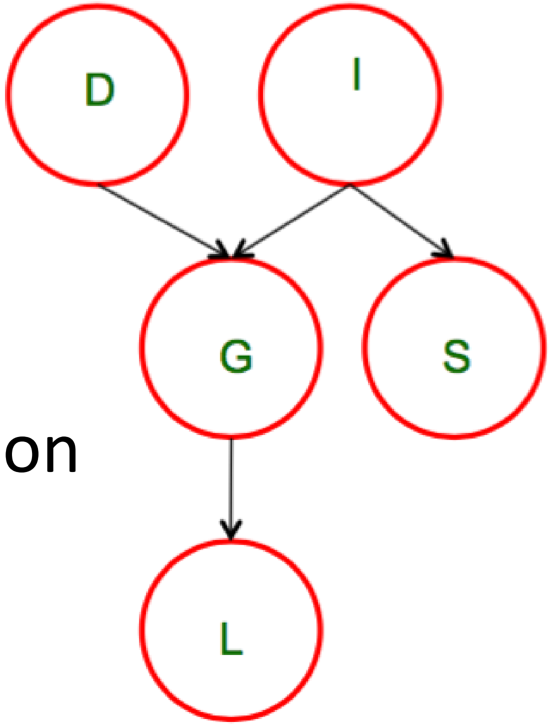
- Example: Consider Trail *D->G<-I->S*
- When observed
  - $Z = \{\emptyset\}$ *:* trail is *inactive* because v-structure *D->G<-I* is inactive
  - $Z = \{L\}$ *: active* (*D->G<-I* active) since L is descendant of G
  - $Z = \{L, I\}$ *: inactive* because observing I blocks *G<-I->S*



D-separation definition

- Let ***X,Y*** and ***Z*** be three sets of nodes in G.
- ***X*** and ***Y*** are d-separated given ***Z*** denoted $\text{d-sep}_G(X \perp Y | Z)$ if there is no active trail between any node *X*∈***X*** and *Y*∈***Y*** given ***Z***
- That is, nodes in ***X*** cannot influence nodes in ***Y***
- Provides notion of separation between nodes in a directed graph ("directed" separation)

Independencies from D-separation



- Definition

$$I(G) = \{(X \perp Y | Z) : \text{d-sep}_G(X \perp Y | Z)\}$$

  - Also called *Global Markov independencies*
  - Note: Derived purely from graph (using trails)
- Example: *Global* independence using D-separation
  - $(L \perp I, D, S | G) \in I(G)$
- Compare with *local* independencies
  - For each X_i: ($X_i \perp$ Non-Desc(X_i) | $Pa(X_i)$)

$$\{(L \perp I, D, S | G), (S \perp D, G, L | I), (G, S | D, I), (D \perp I, S | \emptyset)\} \subseteq I(G)$$

Algorithm for D-Separation

Algorithm for D-Separation

- Enumerate all trails between *X* & *Y* is inefficient
  - No. of trails is *exponential* with graph size
- Linear time algorithm has two phases
  - Algorithm *Reachable*(G,*X*,***Z***) returns nodes reachable from *X* given ***Z***
  - Phase 1 (simple)
    - Traverse bottom-up from leaves marking all nodes in ***Z*** or descendants in ***Z***; to enable v-structures
  - Phase 2 (subtle)
    - Traverse top-down from *X* to *Y,* stopping when blocked by a node

Structure Implications

- Given a Bayes net structure, can run d-separation algorithm to build a complete list of conditional independences that are necessarily true of the form

$$X_i \perp X_j | \{X_{k_1}, ..., X_{k_n}\}$$

- This list determines the set of probability distributions that can be represented

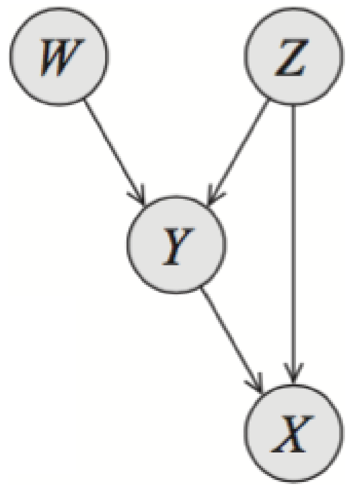
Pseudocode

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 V
    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
```

Example for D-separation algorithm



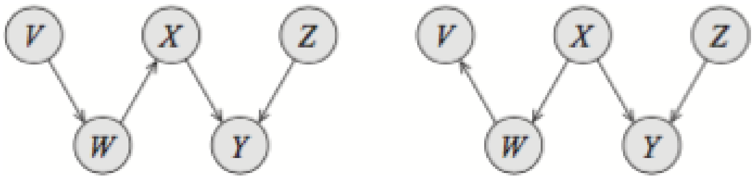
- Task: Find all nodes reachable from *X*
- Assume that *Y* is observed, i.e., $Y \in \boldsymbol{Z}$
- Assume algorithm first encounters *Y* via edge *Y* -> *X*
- Any extension of this trail is blocked by *Y*
- Trail *X*<-*Z*->*Y*<-*W* is not blocked by *Y*
- Thus when we encounter *Y* for the second time via the edge *Z*->*Y* we should not ignore it
- Therefore after the first visit to *Y* we can mark it as visited

  For trails coming from children of *Y*
  Not for purpose of trails coming from parents of *Y*

I-Equivalence

I-Equivalence

- Conditional independence assertion statements can be the same with different structures
- Two graphs G_1 and G_2 are *I- equivalent* if $I(G_1)=I(G_2)$
- Skeleton of a BN graph G is an undirected graph with an edge for every edge in G
- If two BN graphs have the same set of skeletons and v-structures then they are *I-equivalent*



Same skeleton
Same v-structure $X \rightarrow Y \leftarrow Z$

Bayes Nets Representation Summary

- Bayes nets compactly encode joint distributions
- Guaranteed independencies of distributions can be deduced from BN graph structure
- D-separation gives precise conditional independence guarantees from graph alone
- A Bayes’ net’s joint distribution may have further (conditional) independence that is not detectable until you inspect its specific distribution