An Introduction to Causal Inference

Conor Mayo-Wilson

Department of Philosophy
University of Washington

July 20th, 2015

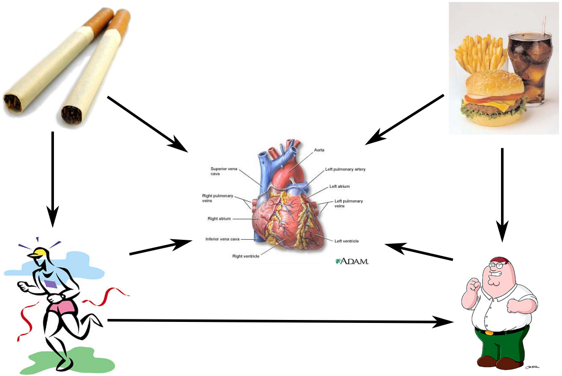
A Puzzle

- Search the Journal of the American Medical Association and you'll find over $270,000$ papers concerning causes of heart disease.

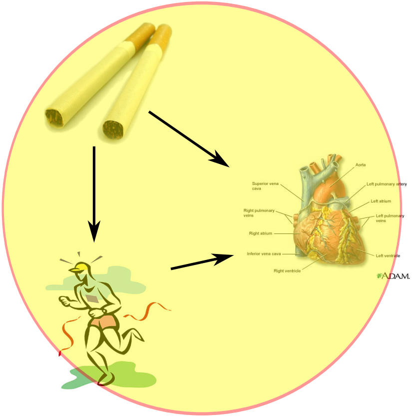
A Puzzle

- Search the Journal of the American Medical Association and you'll find over $270,000$ papers concerning causes of heart disease.
- The enormous number of papers is, in part, a result of the enormous number of factors that might be relevant to heart disease.

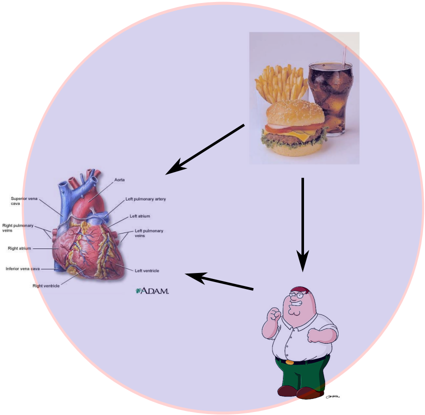
Integrating Scientific Evidence



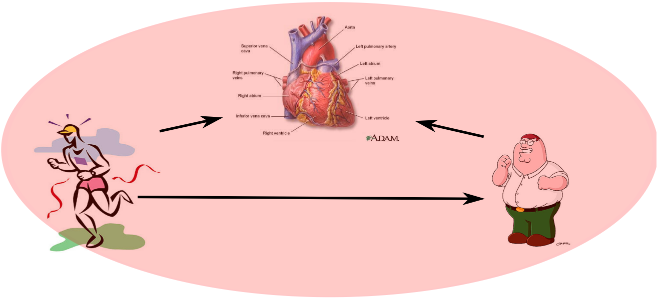
Integrating Scientific Evidence



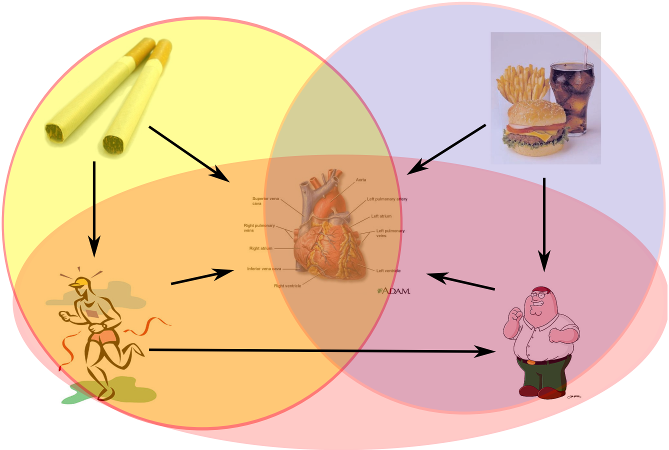
Integrating Scientific Evidence



Integrating Scientific Evidence

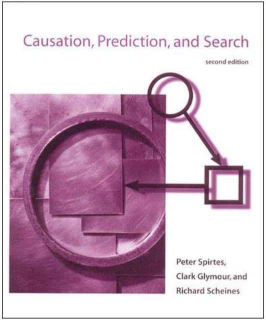


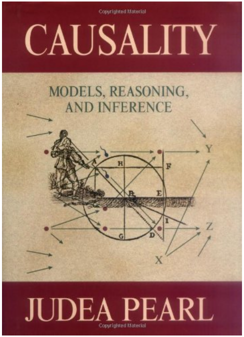
Integrating Scientific Evidence



Central Question: When, if ever, is it possible to integrate several causal theories, like those typical in medicine and the social sciences?

My Framework





Eliciting Intuitions

Please write down your answers to the following two questions.

Eliciting Intuitions

Please write down your answers to the following two questions.

Question 1: Mary has blonde hair. What is the probability that she has blue eyes?

Eliciting Intuitions

Please write down your answers to the following two questions.

Question 1: Mary has blonde hair. What is the probability that she has blue eyes?

Question 2: Suzy recently dyed her hair blonde. What is the probability that she has blue eyes?

Prediction: Observation vs. Intervention

The questions indicate the difference between predicting

- Based on **observational** data
- The effect of an intervention

Prediction: Observation vs. Intervention

Examples:

- High test scores indicate student learning.

Prediction: Observation vs. Intervention

Examples:

- High test scores indicate student learning.
  - If the teacher intervenes and corrects all of her student's mistakes on a standardized test before sending the exams to the central office, her actions will improve the test scores, but not student learning.

Prediction: Observation vs. Intervention

Examples:

- High test scores indicate student learning.
    - If the teacher intervenes and corrects all of her student's mistakes on a standardized test before sending the exams to the central office, her actions will improve the test scores, but not student learning.
- Whether someone's clothes smell like cigarettes allow us to predict more accurately whether or not she will acquire lung cancer later in life.

Prediction: Observation vs. Intervention

Examples:

- High test scores indicate student learning.
  - If the teacher intervenes and corrects all of her student's mistakes on a standardized test before sending the exams to the central office, her actions will improve the test scores, but not student learning.
- Whether someone's clothes smell like cigarettes allow us to predict more accurately whether or not she will acquire lung cancer later in life.
  - But if we intervene and Febreeze someone's wardrobe, we do not reduce her chances of lung cancer.

Predicting Intervention Effects

Predicting the effects of social policy and medical interventions requires two thing:

1. Discovering causal structure among the variables of interest,

Predicting Intervention Effects

Predicting the effects of social policy and medical interventions requires two thing:

1. Discovering causal structure among the variables of interest,
2. Rules for estimating the effect of an intervention from causal structure.

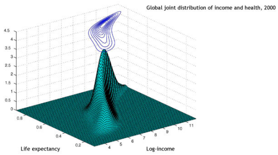
Predicting Intervention Effects

Predicting the effects of social policy and medical interventions requires two thing:

1. Discovering causal structure among the variables of interest,
2. Rules for estimating the effect of an intervention from causal structure.

We'll tackle these two issues in reverse order.

Probability Theory vs. Statistics



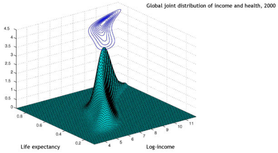
Income	Age at death
55K	77
87K	83
24K	63
⋮	⋮
⋮	⋮

Probability Distribution

Data

Probability Theory vs. Statistics

Probability Theory

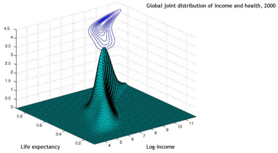


Income	Age at death
55K	77
87K	83
24K	63
⋮	⋮
⋮	⋮

Probability Distribution

Data

Probability Theory vs. Statistics



Probability Theory

Statistics

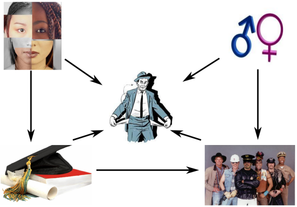
Income	Age at death
55K	77
87K	83
24K	63
⋮	⋮
⋮	⋮

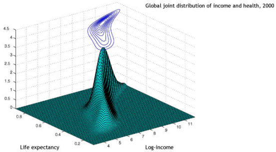
Probability Distribution

Data

Causal Bridge Principles

Bridge Principle





Causal Structure

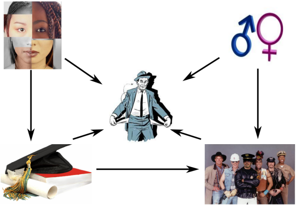
Probability Distribution

Causal Bridge Principles

Different bridge principles concern probabilistic facts of various levels of "coarseness."

Types of Bridge Principles

Bridge Principle



$\text{Gender} \perp\!\!\!\perp \text{Race}$

$\text{Education} \not\!\perp\!\!\!\perp \text{Income}$

$\text{Race} \perp\!\!\!\perp \text{Job} \mid \text{Education}$

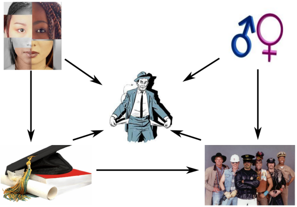
$\vdots$

Causal Structure

Conditional Independence

Types of Bridge Principles

Bridge Principle



$\text{Gender} \perp\!\!\!\perp \text{Race}$

$\text{Education} \not\perp\!\!\!\perp \text{Income}$

$\text{Race} \perp\!\!\!\perp \text{Job} \mid \text{Education}$

$\vdots$

Causal Structure

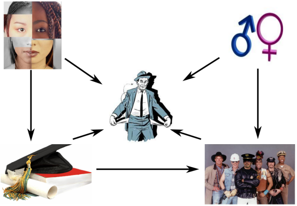
Conditional Independence

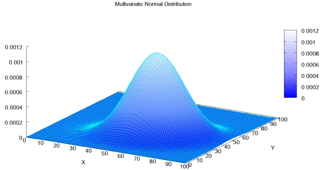
Bridge Principles:

- **Causal Markov Condition** (CMC): In causally sufficient systems, variables are conditionally independent of their non-effects given their direct causes.
- **Causal Faithfulness Condition** (CFC): No two variables are conditionally independent unless so entailed by the CMC.

Types of Bridge Principles

Bridge Principle



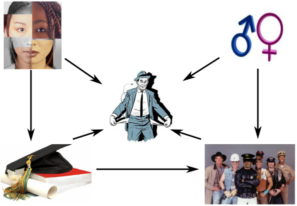


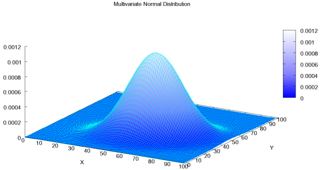
Causal Structure

Properties of Joint Distribution

Types of Bridge Principles

Bridge Principle





Causal Structure

Properties of Joint Distribution

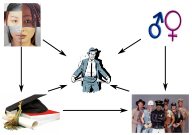
Bridge Principle: Each variable is a linear combination of its direct causes and an independent, normally distributed error term:

$\epsilon_i \sim N(\mu_i, \sigma_i^2)$

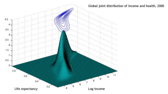
$X_i = \epsilon_i + \sum_{X_j \in \mathrm{PA}_G(X)} a_{i,j} X_j$

Utility of Bridge Principles

Bridge Principle



Causes



Probabilities

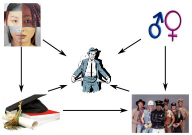
Probability Theory

Income	**Age**	**Gender**	...
55K	37	F	...
19K	28	M	...
123K	58	M	...
⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮

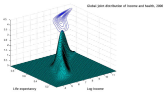
Data

Utility of Bridge Principles

Bridge Principle



Causes



Probabilities

Probability Theory

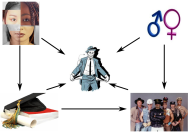
Prediction and Control

Income	Age	Gender	...
55K	37	F	...
19K	28	M	...
123K	58	M	...
⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮

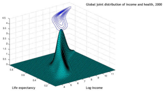
Data

Utility of Bridge Principles

Bridge Principle



Causes



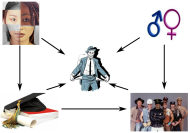
Probabilities

Income	Age	Gender	...
55K	37	F	...
19K	28	M	...
123K	58	M	...
⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮

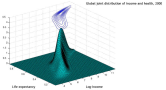
Data

Utility of Bridge Principles

Bridge Principle



Causes



Probabilities

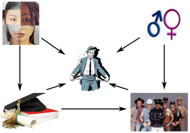
Statistics

Income	Age	Gender	...
55K	37	F	...
19K	28	M	...
123K	58	M	...
⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮

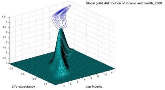
Data

Utility of Bridge Principles

Bridge Principle



Causes



Probabilities

Statistics

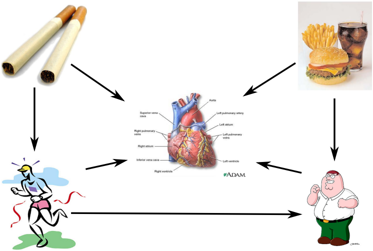
Causal Discovery

Income	Age	Gender	...
55K	37	F	...
19K	28	M	...
123K	58	M	...
⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮

Data

Outline

Causal Graphs



- We'll represent causal relationships using **directed graphs**.
- Nodes = Random Variables.
- In a causal graph $\mathcal{G}$, an edge $X \rightarrow Y$ represents that X is a direct cause of Y relative to the variables in $\mathcal{G}$.

Direct Causation

Question: What does it mean to say X is a **direct cause** of Y relative to a set of variables $\mathcal{V}$?

Answer: Roughly, it means that there's some way of changing X that will change Y, if variables in $\mathcal{V}$ other than Y are held fixed. Here's an example ...

Direct Causation

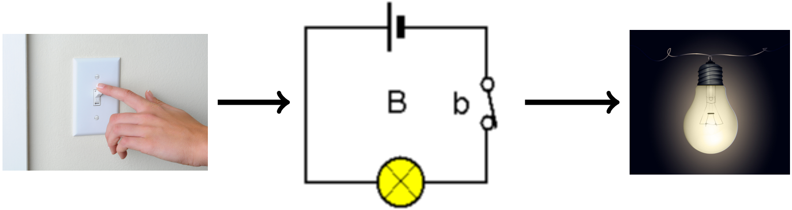
$\longrightarrow$

Let $\mathcal{V} = \{S, B\}$, where

- $S =$ "Switch", which can take the values "On" and "Off", and
- $B =$ "Bulb", which takes the the values "On" and "Off."

Flipping the switch can change the value of bulb. So S is a direct cause of B relative to $\mathcal{V}$.

Direct Causation

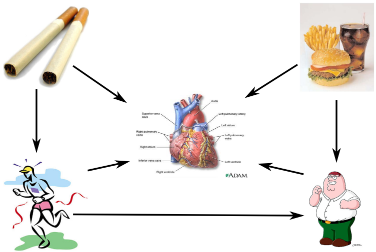


Let $\mathcal{V} = \{S, C, B\}$, where

- S and B are as before, and
- C = "Circuit", which takes two values: "Closed" and "Open."

If the circuit is held open (or closed), then changing flipping the light switch cannot change the value of bulb. So relative to $\mathcal{V}$, the variable S is **not** a direct cause of B.

Causal Graphs



We'll also assume graphs are **acyclic.**

- That is, there is no sequence $V_1 \to V_2 \to \ldots \to V_n \to V_1$.

Feedback loops

Question: Some real-world cases of causation appear cyclical. Can this framework represent those relationships?

- E.g., Infrequent exercise causes obesity, which in turn reduces one's desire to exercise.

Modeling causation over time

$$\text{Exercise}_t \longrightarrow \text{Weight}_{t+1} \longrightarrow \text{Exercise}_{t+2}$$

Response: Causation is not actually cyclical. Distinguish the variables at different **times**.

Modeling causation over time

Question: What bridge principles connect causal graphs to probability distributions?

Markov and Faithfulness

Causal Sufficiency

Definition: A collection of variables $\mathcal{V}$ is **causally sufficient** for any pair of variables $V_1, V_2 \in \mathcal{V}$, if U is a common cause of V_1 and V_2, then U is also in $\mathcal{V}$

.

Causal Sufficiency

Example: Consider the set $\mathcal{V} = \{X, Y, Z\}$, where

- X = Average number of cigarettes smoked in a day.
- Y = Yellow discoloration of fingernails (yes or no)
- Z = Patient has lung cancer (yes or no)

Causal Sufficiency

This system is likely causally sufficient.

- Are there any common causes of smoking (X) and having yellow fingernails (Y)? Probably not.

Causal Sufficiency

This system is likely causally sufficient.

- Are there any common causes of smoking (X) and having yellow fingernails (Y)? Probably not.
- Are there any common causes of smoking (X) and having lung cancer (Y)? Probably not (pace Fisher).

Causal Sufficiency

This system is likely causally sufficient.

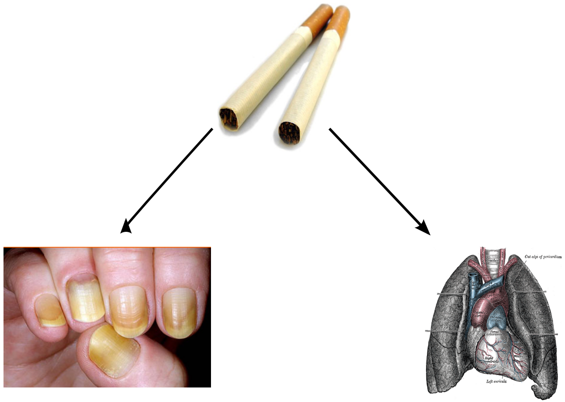
- Are there any common causes of smoking (X) and having yellow fingernails (Y)? Probably not.
- Are there any common causes of smoking (X) and having lung cancer (Y)? Probably not (pace Fisher).
- Are there any common causes of lung cancer (Z) and having yellow fingernails (Y)?
  - Sure. Smoking is a common cause, and it's in $\mathcal{V}$.

Causal Markov Condition (CMC)

Causal Markov Condition (CMC): In causally sufficient systems, any variable is conditionally independent of its non-effects given its direct causes.

Consequences of CMC

CMC captures intuitions about common causes.



Consequences of CMC

CMC captures intuitions about indirect causes.

Causal Faithfulness Condition (CFC)

Causal Faithfulness Condition (CFC): No two variables are conditionally independent unless so entailed by the CMC.

Consequences of CFC

Informally: Variables that are (conditionally) independent are not directly causally connected.

Consequences of CFC

Example:

- Suppose you wish to know whether drinking a glass of wine day reduces chance of heart disease.
- You find that, conditional on income, there is no correlation between drinking a glass of wine a day and incidence of heart disease.
- You conclude that drinking a glass of wine a day is not a cause of cardiovascular health.

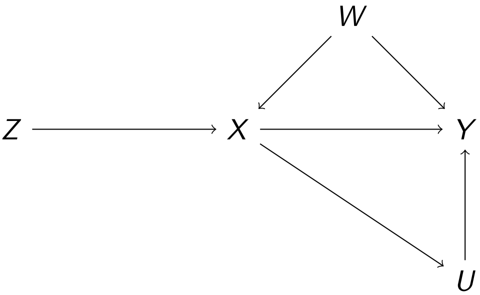
Consequences of CFC

Example:

- Suppose you wish to know whether drinking a glass of wine day reduces chance of heart disease.
- You find that, conditional on income, there is no correlation between drinking a glass of wine a day and incidence of heart disease.
- You conclude that drinking a glass of wine a day is not a cause of cardiovascular health.
- Your conclusion is valid only if and only if you assume faithfulness.
  - Wine-drinking and cardiovascular health are conditionally independent.
  - So by faithfulness, they are not directly causally connected.

Graphical Terminology

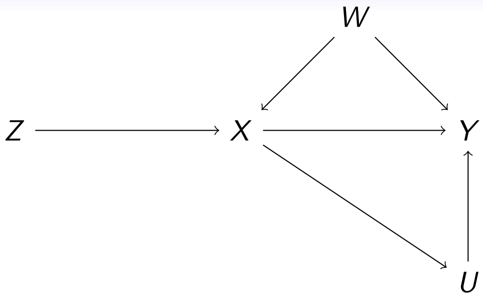
Parents and Children



Say V_1 is a **parent** of V_2 in $\mathcal{G}$ if there is an edge from V_1 to V_2. In this case, say V_2 is a **child** of V_1.

- Example: X is a parent of Y, whereas Z is **not** a parent of Y.

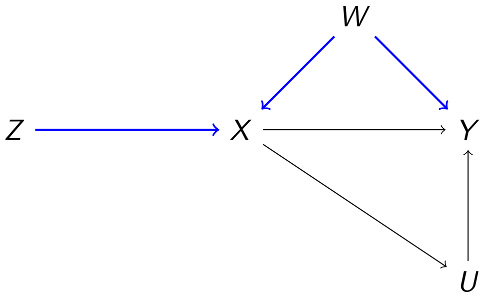
Ancestors and Descendants



Say X is a **ancestor** of Y in $\mathcal{G}$ if there is a sequence of edges $X \to V_1 \to V_2 \ldots \to Y$. In this case, we say Y is a **descendant** of X.

- Example: Z and X are both ancestors of Y.
- Note: Parents are ancestors, by definition. Children are descendants.

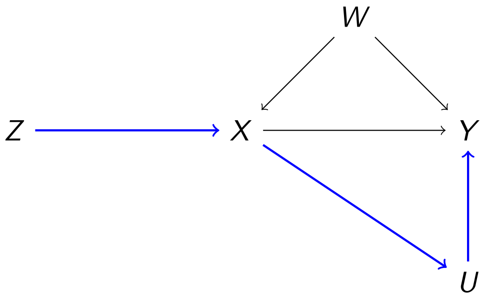
Paths



A **path** is a sequence of nodes where each variable in the sequence occurs precisely once.

- Example: $\langle Z, X, W, Y \rangle$ is a path.
- Example: $\langle X, U, Y, X \rangle$ is not a path.

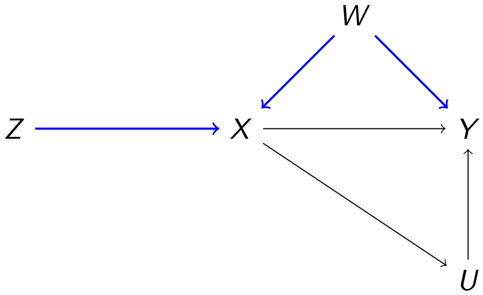
Directed Paths



A **directed path** is a path where each edge points in the same direction.

- Example: $\langle Z, X, U, Y \rangle$.
- Note there may be several directed paths between two variables.

Colliders



A variable V_k is a **collider** on a path $\langle V_1, V_2, \ldots, V_n \rangle$ if the edges on the path point "into" V_K, i.e., $V_{k-1} \rightarrow V_k \leftarrow V_{k+1}$ is in $\mathcal{G}$.

- Example: X is a collider on the path $\langle Z, X, W, Y \rangle$.

Blocking

A collection of variables $\mathcal{C}$ **blocks** a path from V_1 to V_2:

Blocking

A collection of variables $\mathcal{C}$ **blocks** a path from V_1 to V_2:

- The path contains a variable X in $\mathcal{C}$, and X is **not** a collider on the path.

Blocking

A collection of variables $\mathcal{C}$ **blocks** a path from V_1 to V_2:

- The path contains a variable X in $\mathcal{C}$, and X is **not** a collider on the path.
- The path contains a collider X that is **not** in $\mathcal{C}$, **and** none of X's descendants are in $\mathcal{C}$.

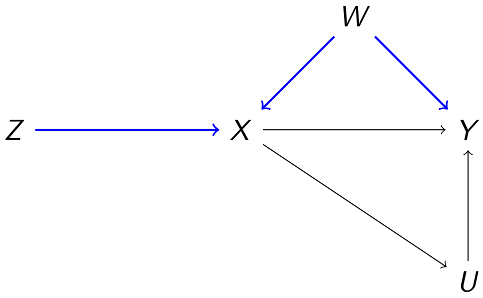
Blocking

A collection of variables $\mathcal{C}$ **blocks** a path from V_1 to V_2:

- The path contains a variable X in $\mathcal{C}$, and X is **not** a collider on the path.
- The path contains a collider X that is **not** in $\mathcal{C}$, **and** none of X's descendants are in $\mathcal{C}$.

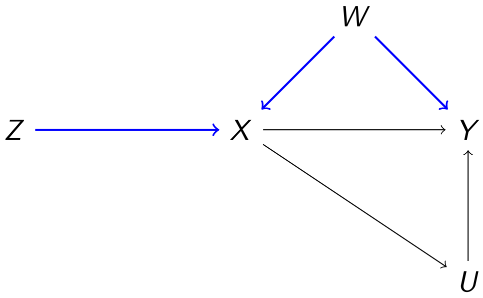
A path is **active** given $\mathcal{C}$ if and only if it is not blocked by $\mathcal{C}$.

Blocking



Example: Consider the path $\langle Z, X, W, Y \rangle$

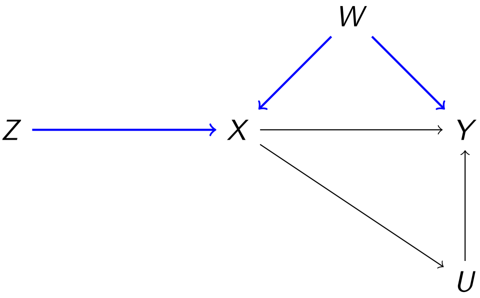
Blocking



Example: Consider the path $\langle Z, X, W, Y \rangle$

- The path is **not** blocked by $\mathcal{C} = \{X\}$, as X is a collider on the path.

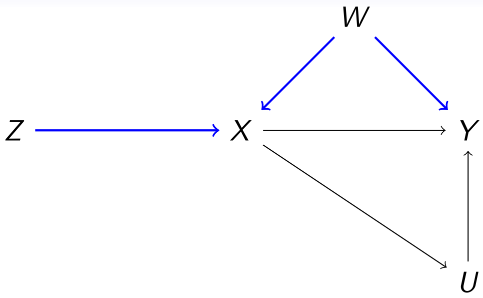
Blocking



Example: Consider the path $\langle Z, X, W, Y \rangle$

- The path is **not** blocked by $\mathcal{C} = \{X\}$, as X is a collider on the path.
- The path is blocked by $\mathcal{C} = \{X, W\}$, as $W in \mathcal{C}$ and is a non-collider on the path.

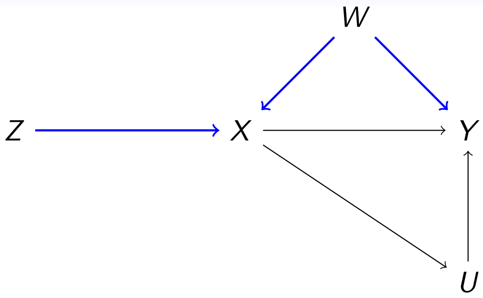
Blocking



Example: Consider the path $\langle Z, X, W, Y \rangle$

- The path is **not** blocked by $\mathcal{C} = \{U\}$, as U is a descendant of X, which is a collider on the path.

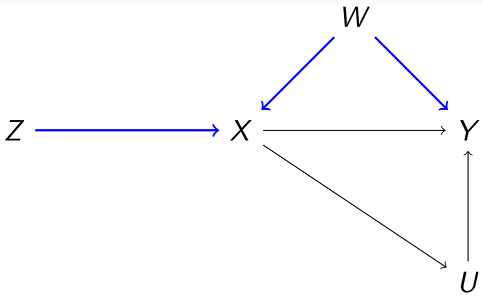
Blocking



Example: Consider the path $\langle Z, X, W, Y \rangle$

- The path is **not** blocked by $\mathcal{C} = \{U\}$, as U is a descendant of X, which is a collider on the path.
- The path is blocked by $\mathcal{C} = \emptyset$, as X is a collider and neither it nor any of its descendants are in $\mathcal{C}$.

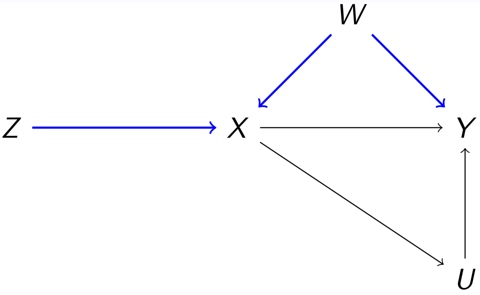
Blocking



Example: Consider the path $\langle Z, X, W, Y \rangle$

- The path is **not** blocked by $\mathcal{C} = \{U\}$, as U is a descendant of X, which is a collider on the path.

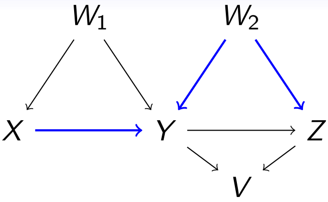
Blocking



Example: Consider the path $\langle Z, X, W, Y\rangle$

- The path is **not** blocked by $\mathcal{C} = \{U\}$, as U is a descendant of X, which is a collider on the path.
- The path is blocked by $\mathcal{C} = \emptyset$, as X is a collider and neither it nor any of its descendants are in $\mathcal{C}$.

Examples: Blocking



Question: Consider the path $\langle X, Y, W_2, Z \rangle$. Which of the following sets block the path?

- The empty set.
- $\{W_2\}$
- $\{W_2, V\}$
- $\{Y\}$
- $\{Y, W_2\}$

Examples: Blocking

$$\begin{tikzcd}
& W_1 \arrow[dl] \arrow[dr] & & W_2 \arrow[dl] \arrow[dr] & \\
X \arrow[rr] & & Y \arrow[rr] \arrow[dr] & & Z \arrow[dl] \\
& & & V &
\end{tikzcd}$$

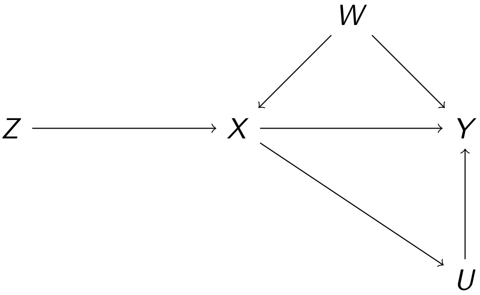
Question: Find all paths between X and Z that are active (i.e., not blocked) given

- The empty set.
- The set $\{Y\}$.

d-Separation

Definition: Two variables V_1 and V_2 are said to be d-separated given $\mathcal{C}$ if and only if every path between V_1 and V_2 is blocked by $\mathcal{C}$.

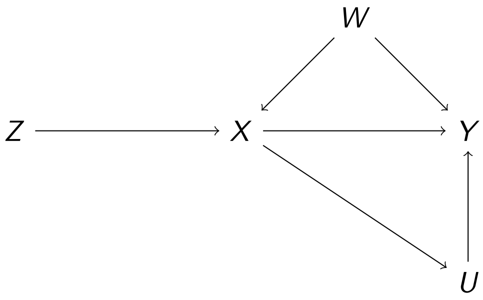
d-Separation



Example:

- Z and U are **not** d-separated given $\emptyset$.
  - Consider the path $\langle Z, X, U \rangle$.

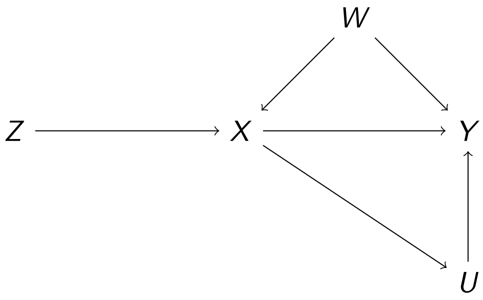
d-Separation



Example:

- Z and U are **not** d-separated given $\emptyset$.
  - Consider the path $\langle Z, X, U \rangle$.
- Z and U are *d*-separated given $\{X\}$.

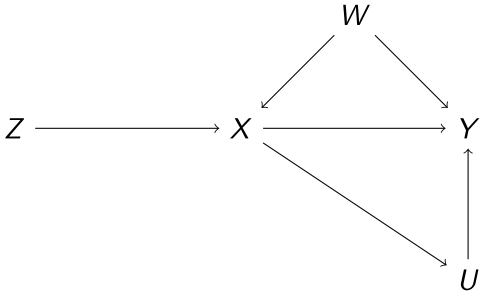
d-Separation



Example:

- Z and Y are **not** d-separated given $\{X\}$.
  - Consider the path $\langle Z, X, W, Y \rangle$.

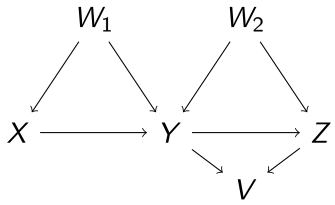
d-Separation



Example:

- Z and Y are **not** d-separated given $\{X\}$.
    - Consider the path $\langle Z, X, W, Y \rangle$.
- The are d-separated given $\{X, W\}$.

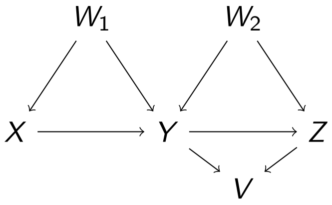
Exercise: d-Separation



Question: Are W_1 and W_2 d-separated given?

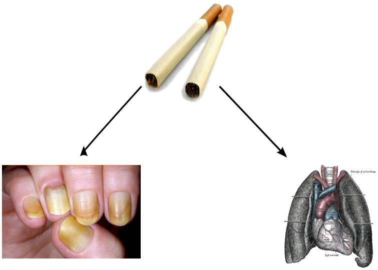
- The empty set.
- $\{Y\}$.
- $\{X, Z\}$.
- $\{V\}$.

Exercise: d-Separation



Question: Find all $\mathcal{C}$ such that X and Z are d-separated given $\mathcal{C}$.

d-Separation in Causal Graphs



- Yellow fingernails and lung cancer are **not** d-separated given the empty set.
- Yellow fingernails and lung cancer are d-separated given smoking.

d-Separation in Causal Graphs

- Smoking and lung cancer are *d*-separated given amount of tar.
- Smoking and lung cancer are **not** d-separated given the empty set.

d-Separation and Conditional Independence

- In the last two examples, the variables X and Y were d-separated given $\mathcal{C}$ precisely when X and Y were conditionally independent given the Markov condition.
- This was not a coincidence . . .

d-Separation and Conditional Independence

Theorem

If the CMC and CFC hold, then two variables V_1 and V_2 in a causal graph over a causally sufficient set of variables $\mathcal{V}$ are d-separated given $\mathcal{C}$ if and only if V_1 and V_2 are conditionally independent given $\mathcal{C}$.

d-Separation and Prediction

- Knowing which variables are conditionally independent is necessary for identifying what factors we need to control when estimating causal effect.

d-Separation and Prediction

- Knowing which variables are conditionally independent is necessary for identifying what factors we need to control when estimating causal effect.
- Given the causal graph, the notion of d-separation allows us to identify which variables are conditionally independent.

d-Separation and Prediction

- Knowing which variables are conditionally independent is necessary for identifying what factors we need to control when estimating causal effect.
- Given the causal graph, the notion of d-separation allows us to identify which variables are conditionally independent.
- Moral: The notion of d-separation will be useful in estimating causal effects.

Estimating Causal Effects

Markov condition and Factorization

Causal Markov Condition: Within a causally sufficient set of variables $\mathcal{V}$, any variable is conditionally independent of its non-effects given its direct causes.

Markov condition and Factorization

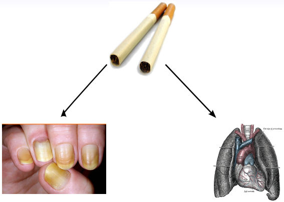
Causal Markov Condition: Within a causally sufficient set of variables $\mathcal{V}$, any variable is conditionally independent of its non-effects given its direct causes.

Equivalently: Let $P(V_1, V_2, \ldots, V_n)$ represent the joint distribution over a causally sufficient set of variables $\mathcal{V} = \{V_1, V_2, \ldots, V_n\}$. Then:

$$P(V_1, V_2, \ldots, V_n) = \Pi_{i \leq n} P(V_i | \mathrm{PA}_{\mathcal{G}}(V_i))$$

where $\mathrm{PA}_{\mathcal{G}}(V))$ is the set of parents of V in the causal graph $\mathcal{G}$ over $\mathcal{V}$.

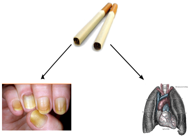
Markov condition and Factorization



Example: Let $\mathcal{V} = \{S, L, Y\}$ be three binary variables, where

- S represents "smoking".
- L represents "lung cancer"
- Y "yellow fingernails."
- $S(x) = 1$ represents that x smokes. $S(x) = 0$ represents x does not smoke.
- Similarly for L and Y.

Markov condition and Factorization

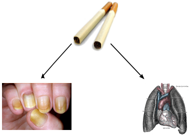


Example:

- $P(Smoker) = .178$
- $P(Yellow|Smoker) = .12$
- $P(Yellow|Non-Smoker) = .001$
- $P(Cancer|Smoker) = .076$
- $P(Cancer|Non-Smoker) = .02$

What's the probability that a randomly selected member of the population smokes, has lung cancer, and no yellow fingernails?

Markov condition and Factorization



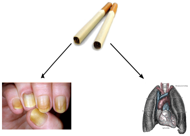
By the Markov condition:

$$P(V_1, V_2, \ldots, V_n) = \Pi_{i \leq n} P(V_i | \mathrm{PA}_{\mathcal{G}}(V_i))$$

In this case:

$$\begin{aligned} P(S=1, L=1, Y=0) &= P(S=1) \cdot P(L=1|S=1) \cdot P(Y=0|S=1) \\ &= 0.178 \cdot 0.076 \cdot 0.88 \\ &= 0.01190464 \end{aligned}$$

Markov condition and Factorization



Example: What's the probability that a randomly selected member of the population has lung cancer given she does not have yellow fingernails?

Markov condition and Factorization

Example: What's the probability that a randomly selected member of the population has lung cancer given she does not have yellow fingernails?

Answer: By the definition of conditional probability:

$$P(L = 1 | Y = 0) = \frac{P(L = 1, Y = 0)}{P(Y = 0)}$$

So we first compute $P(L = 1, Y = 0)$, and then we compute $P(Y = 0)$.

Markov condition and Factorization

$$P(L=1, Y=0) \quad = \quad P(S=1, L=1, Y=0) + P(S=0, L=1, Y=0)$$

Markov condition and Factorization

$$
\begin{aligned}
P(L=1, Y=0) &= P(S=1, L=1, Y=0) + P(S=0, L=1, Y=0) \\
&= [P(S=1) \cdot P(L=1|S=1) \cdot P(Y=0|S=1)] \\
&+ [P(S=0) \cdot P(L=1|S=0) \cdot P(Y=0|S=0)]
\end{aligned}
$$

by the Markov condition

Markov condition and Factorization

$$
\begin{aligned}
P(L=1, Y=0) &= P(S=1, L=1, Y=0) + P(S=0, L=1, Y=0) \\
&= [P(S=1) \cdot P(L=1|S=1) \cdot P(Y=0|S=1)] \\
&+ [P(S=0) \cdot P(L=1|S=0) \cdot P(Y=0|S=0)] \\
&\quad \text{by the Markov condition} \\
&= [.178 \cdot .076 \cdot .88] + [.822 \cdot .02 \cdot .999]
\end{aligned}
$$

Markov condition and Factorization

$$
\begin{aligned}
P(L=1, Y=0) &= P(S=1, L=1, Y=0) + P(S=0, L=1, Y=0) \\
&= [P(S=1)\cdot P(L=1|S=1)\cdot P(Y=0|S=1)] \\
&+ [P(S=0)\cdot P(L=1|S=0)\cdot P(Y=0|S=0)] \\
&\quad \text{by the Markov condition} \\
&= [.178\cdot .076\cdot .88] + [.822\cdot .02\cdot .999] \\
&= .028
\end{aligned}
$$

Markov condition and Factorization

$$\begin{aligned} P(Y=0) &= P(S=0,L=0,Y=0)+P(S=0,L=1,Y=0) \\ &+ P(S=1,L=0,Y=0)+P(S=1,L=1,Y=0) \end{aligned}$$

Markov condition and Factorization

$$
\begin{aligned}
P(Y=0) &= P(S=0, L=0, Y=0) + P(S=0, L=1, Y=0) \\
&+ P(S=1, L=0, Y=0) + P(S=1, L=1, Y=0) \\
&= [P(S=0) \cdot P(L=0|S=0) \cdot P(Y=0|S=0)] \\
&+ [P(S=0) \cdot P(L=1|S=0) \cdot P(Y=0|S=0)] \\
&+ [P(S=1) \cdot P(L=0|S=1) \cdot P(Y=0|S=1)] \\
&+ [P(S=1) \cdot P(L=1|S=1) \cdot P(Y=0|S=1)] \\
&\quad \text{by the Markov condition}
\end{aligned}
$$

Markov condition and Factorization

$$
\begin{aligned}
P(Y=0) &= P(S=0, L=0, Y=0) + P(S=0, L=1, Y=0) \\
&+ P(S=1, L=0, Y=0) + P(S=1, L=1, Y=0) \\
&= [P(S=0) \cdot P(L=0|S=0) \cdot P(Y=0|S=0)] \\
&+ [P(S=0) \cdot P(L=1|S=0) \cdot P(Y=0|S=0)] \\
&+ [P(S=1) \cdot P(L=0|S=1) \cdot P(Y=0|S=1)] \\
&+ [P(S=1) \cdot P(L=1|S=1) \cdot P(Y=0|S=1)] \\
&\quad \text{by the Markov condition} \\
&= [.822 \cdot .98 \cdot .999] + [.822 \cdot .02 \cdot .999] \\
&+ [.178 \cdot .924 \cdot .88)] + [.178 \cdot .076 \cdot .88] \\
&= 0.977818
\end{aligned}
$$

Markov condition and Factorization

Example: What's the probability that a randomly selected member of the population has lung cancer given she does not have yellow fingernails?

Answer:

$$
\begin{aligned}
P(L=1|Y=0) &= \frac{P(L=1, Y=0)}{P(Y=0)} \\
&= \frac{.028}{0.977818} \\
&= 0.02863518568
\end{aligned}
$$

Exercise with Markov condition

Exercise: Calculate the probability that a randomly selected member of the population has lung cancer given she has yellow fingernails.

Predictions after Interventions

Suppose I conduct an experiment in which I select 100 members of the population at random. I remove yellow stains from all their fingernails.

Question: What is the probability that a member of my population has lung cancer given her fingernails are not yellow?

Predicting the effect of a "surgical" intervention

Suppose we have a causal graph $\mathcal{G}$ and a probability distribution P over the variables of $\mathcal{G}$.

General Question: How should we compute the "new" causal graph $\mathcal{G}_{V=v}$ and the new probability distribution $P_{V=v}$ that results from intervening some variable V and setting its value to v?

Predicting the effect of a "surgical" intervention

Suppose we have a causal graph $\mathcal{G}$ and a probability distribution P over the variables of $\mathcal{G}$.

General Question: How should we compute the "new" causal graph $\mathcal{G}_{V=v}$ and the new probability distribution $P_{V=v}$ that results from intervening some variable V and setting its value to v?
Answer:

- $\mathcal{G}_{V=v}$ is identical to $\mathcal{G}$, except $\mathcal{G}_{V=v}$ contains no edges that point **into** V.

Predicting the effect of a "surgical" intervention

Suppose we have a causal graph $\mathcal{G}$ and a probability distribution P over the variables of $\mathcal{G}$.

General Question: How should we compute the "new" causal graph $\mathcal{G}_{V=v}$ and the new probability distribution $P_{V=v}$ that results from intervening some variable V and setting its value to v?
Answer:

- $\mathcal{G}_{V=v}$ is identical to $\mathcal{G}$, except $\mathcal{G}_{V=v}$ contains no edges that point **into** V.
- $P_{V=v}$ is the unique distribution satisfying the following constraints:
  - $P_{V=v}(V = v) = 1$ (i.e., your intervention was successful),

Predicting the effect of a "surgical" intervention

Suppose we have a causal graph $\mathcal{G}$ and a probability distribution P over the variables of $\mathcal{G}$.

General Question: How should we compute the "new" causal graph $\mathcal{G}_{V=v}$ and the new probability distribution $P_{V=v}$ that results from intervening some variable V and setting its value to v?
Answer:

- $\mathcal{G}_{V=v}$ is identical to $\mathcal{G}$, except $\mathcal{G}_{V=v}$ contains no edges that point **into** V.
- $P_{V=v}$ is the unique distribution satisfying the following constraints:
  - $P_{V=v}(V = v) = 1$ (i.e., your intervention was successful),
  - $P_{V=v}(X|\mathrm{PA}_{\mathcal{G}_{V=v}}(X)) = P(X|\mathrm{PA}_{\mathcal{G}}(X))$ for all $X \neq V$, and

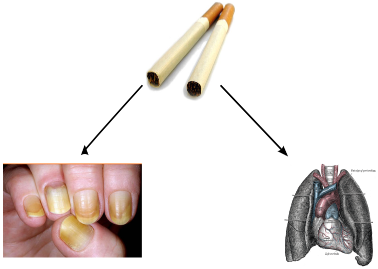
Predicting the effect of a "surgical" intervention

Suppose we have a causal graph $\mathcal{G}$ and a probability distribution P over the variables of $\mathcal{G}$.

General Question: How should we compute the "new" causal graph $\mathcal{G}_{V=v}$ and the new probability distribution $P_{V=v}$ that results from intervening some variable V and setting its value to v?
Answer:

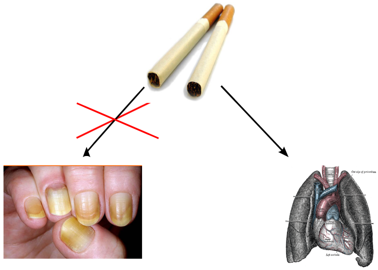
- $\mathcal{G}_{V=v}$ is identical to $\mathcal{G}$, except $\mathcal{G}_{V=v}$ contains no edges that point **into** V.
- $P_{V=v}$ is the unique distribution satisfying the following constraints:
  - $P_{V=v}(V = v) = 1$ (i.e., your intervention was successful),
  - $P_{V=v}(X|\mathrm{PA}_{\mathcal{G}_{V=v}}(X)) = P(X|\mathrm{PA}_{\mathcal{G}}(X))$ for all $X \neq V$, and
  - $P_{V=v}$ satisfies the CMC with respect to the causal graph $\mathcal{G}_{V=v}$.

Surgical Interventions: Graph



;

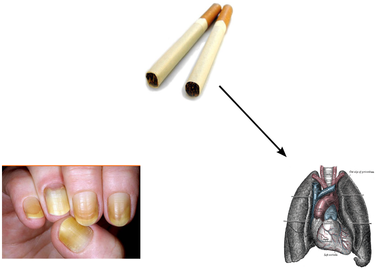
Surgical Interventions: Graph



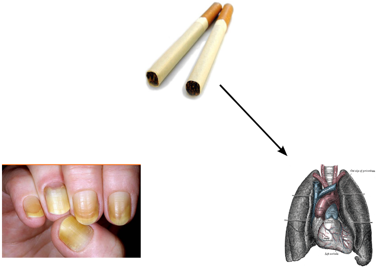
;

Surgical Interventions: Graph

;



Surgical Interventions: Probability



- $P_{Y=0}(Smoker) = .178$
- $P_{Y=0}(Cancer|Smoker) = .076$
- $P_{Y=0}(Cancer|NonSmoker) = .02$
- $P_{Y=0}(Yellow) = 0$

Post Intervention Probabilities

Question: What is the probability that a member of my population has lung cancer given her fingernails are not yellow?

Answer: By the definition of conditional probability:

$$P_{Y=0}(L=1|Y=0) = \frac{P_{Y=0}(L=1, Y=0)}{P_{Y=0}(Y=0)} = P_{Y=0}(L=1, Y=0)$$

as $P_{Y=0}(Y=0) = 1$ in my population.

To calculate $P_{Y=0}(L=1)$, we apply the Markov condition in the new graph.

Post Intervention Probabilities

$$
\begin{aligned}
P_{Y=0}(L=1) \quad &= \quad [P_{Y=0}(L=1, S=1, Y=0) + P_{Y=0}(L=1, S=0, Y=0)] \\
&+ \quad [P_{Y=0}(L=1, S=1, Y=1) + P_{Y=0}(L=1, S=0, Y=1)]
\end{aligned}
$$

Post Intervention Probabilities

$$\begin{aligned} P_{Y=0}(L=1) &= [P_{Y=0}(L=1,S=1,Y=0)+P_{Y=0}(L=1,S=0,Y=0)] \\ &+ [P_{Y=0}(L=1,S=1,Y=1)+P_{Y=0}(L=1,S=0,Y=1)] \\ \\ &= [P_{Y=0}(L=1,S=1,Y=0)+P_{Y=0}(L=1,S=0,Y=0)] \\ &\quad \text{as } P_{Y=0}(Y=1)=0 \end{aligned}$$

Post Intervention Probabilities

$$
\begin{aligned}
P_{Y=0}(L=1) &= [P_{Y=0}(L=1,S=1,Y=0)+P_{Y=0}(L=1,S=0,Y=0)] \\
&+ [P_{Y=0}(L=1,S=1,Y=1)+P_{Y=0}(L=1,S=0,Y=1)] \\
\\
&= [P_{Y=0}(L=1,S=1,Y=0)+P_{Y=0}(L=1,S=0,Y=0)] \\
&\quad \text{as } P_{Y=0}(Y=1)=0 \\
&= [P_{Y=0}(S=1)\cdot P_{Y=0}(L=1|S=1)\cdot P_{Y=0}(Y=0)] \\
&+ [P_{Y=0}(S=0)\cdot P_{Y=0}(L=1|S=0)\cdot P_{Y=0}(Y=0)] \\
&\quad \text{by the Markov condition} \\
&= [P(S=1)\cdot P(L=1|S=1)] \\
&+ [P(S=0)\cdot P(L=1|S=0)] \\
&\quad \text{by definition of } P_{Y=0}
\end{aligned}
$$

Post Intervention Probabilities

$$
\begin{aligned}
P_{Y=0}(L=1) &= [P_{Y=0}(L=1, S=1, Y=0) + P_{Y=0}(L=1, S=0, Y=0)] \\
&+ [P_{Y=0}(L=1, S=1, Y=1) + P_{Y=0}(L=1, S=0, Y=1)] \\
\\
&= [P_{Y=0}(L=1, S=1, Y=0) + P_{Y=0}(L=1, S=0, Y=0)] \\
&\quad \text{as } P_{Y=0}(Y=1) = 0 \\
&= [P_{Y=0}(S=1) \cdot P_{Y=0}(L=1|S=1) \cdot P_{Y=0}(Y=0)] \\
&+ [P_{Y=0}(S=0) \cdot P_{Y=0}(L=1|S=0) \cdot P_{Y=0}(Y=0)] \\
&\quad \text{by the Markov condition} \\
&= [P(S=1) \cdot P(L=1|S=1)] \\
&+ [P(S=0) \cdot P(L=1|S=0)] \\
&\quad \text{by definition of } P_{Y=0} \\
&= [.178 \cdot .076 \cdot 1] + [.822 \cdot .02 \cdot 1] \\
&= .0299 > 0.0286 = P(L=1|Y=0)
\end{aligned}
$$

Exercises

Exercise 1: Calculate the probability of lung cancer in an experimental group in which intervention forces everyone to smoke.

Exercise 2: Calculate the probability of lung cancer in an experimental group in which intervention forces everyone to smoke.

Four Problems

This procedure for calculating the effect of an intervention has at least four shortcomings:

- **Computational:** Is there a shorter formula that helps us calculate the probabilities?

Four Problems

This procedure for calculating the effect of an intervention has at least four shortcomings:

- **Computational:** Is there a shorter formula that helps us calculate the probabilities?
- **"Soft" Interventions:** Some interventions don't cut off all causal paths to the variable.
    - E.g., If we intervene on after-tax income by lowering taxes, we don't "set" individuals incomes to values, as we did above.

Four Problems

This procedure for calculating the effect of an intervention has at least four shortcomings:

- **Computational:** Is there a shorter formula that helps us calculate the probabilities?
- **"Soft" Interventions:** Some interventions don't cut off all causal paths to the variable.
  - E.g., If we intervene on after-tax income by lowering taxes, we don't "set" individuals incomes to values, as we did above.
- **No Latents:** Our procedure does not seem to work if there are latent variables we haven't measured.

Four Problems

This procedure for calculating the effect of an intervention has at least four shortcomings:

- **Computational:** Is there a shorter formula that helps us calculate the probabilities?
- **"Soft" Interventions:** Some interventions don't cut off all causal paths to the variable.
  - E.g., If we intervene on after-tax income by lowering taxes, we don't "set" individuals incomes to values, as we did above.
- **No Latents:** Our procedure does not seem to work if there are latent variables we haven't measured.
- **Known Causal Structure:** Our procedure assumes we have the graph.

Back-Door Criterion

Question 1: Is there a shorter formula that helps us calculate the probabilities?

Answer: There are several. Here's one due to Pearl [2000]

Back-Door Criterion

Given a pair of variables $\langle X, Y \rangle$, say a set $\mathcal{C}$ satisfies the **back door criterion** if

- $\mathcal{C}$ contains no descendants of X, and
- $\mathcal{C}$ blocks every path between X and Y that contains an edge **into** X.

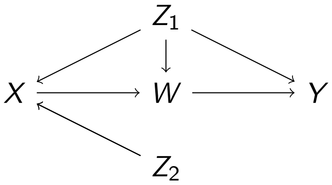
Example: Back-Door Criterion

$$
\begin{tikzcd}
 & Z_1 \arrow[dl] \arrow[d] \arrow[dr] & \\
X \arrow[r] & W \arrow[r] & Y \\
 & Z_2 \arrow[ul] &
\end{tikzcd}
$$

Example: Consider the pair $\langle X, Y \rangle$.

- $\{Z_1, Z_2\}$ satisfies the backdoor criterion

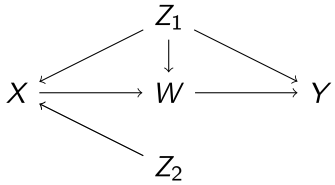
Example: Back-Door Criterion



Example: Consider the pair $\langle X, Y \rangle$.

- $\{Z_1, Z_2\}$ satisfies the backdoor criterion
- $\{Z_1\}$ does **not** because it fails to block the path $\langle X, Z_2, Y$, which contains an arrow into X.

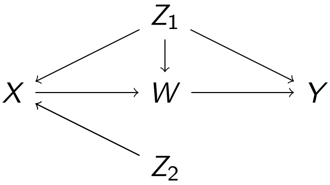
Example: Back-Door Criterion



Example: Consider the pair $\langle X, Y\rangle$.

- $\{Z_1, Z_2\}$ satisfies the backdoor criterion
- $\{Z_1\}$ does **not** because it fails to block the path $\langle X, Z_2, Y$, which contains an arrow into X.
- $\{Z_1, Z_2, W\}$ does **not** because it contains W, which is a descendant of X.

Exercise: Back-Door Criterion



Example: Which sets satisfy the backdoor criterion with respect to $\langle W, Y \rangle$?

Utility of the Backdoor Criterion

Theorem

[Pearl, 1993] If $\mathcal{C}$ satisfies the backdoor criterion with respect to $\langle X, Y \rangle$, then

$$P_{X=x}(Y = y) = \sum_{Z \in \mathcal{C}} \sum_{z \in ran(Z)} P(Y = y | X = x, Z = z) \cdot P(Z = z).$$

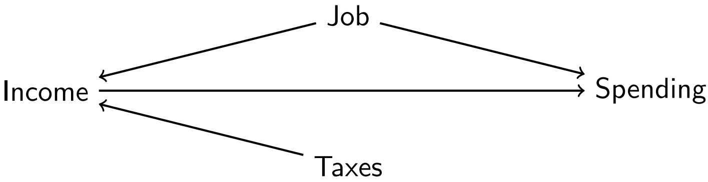
Moral: Computation of post-intervention effects is not so hard.

Soft Interventions

Question 2: Is there a way to calculate the effects of interventions that do not "surgically" remove edges from a graph?

Answer: The same approach as before works with small modifications . . .

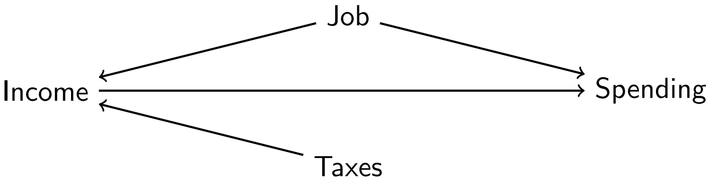
Soft Interventions



Example: Suppose I want to know the effect of increased income on consumer spending.

- I intervene on "Income" by giving 100 study participants $50.
- I haven't eliminated all of their other sources of income, i.e., I haven't "cut off" causes of their income.
  - So I haven't "set" anyone's income to a value.

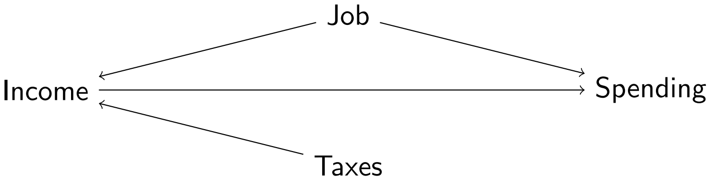
Soft Interventions



Example: Suppose I want to know the effect of increased income on consumer spending.

- I intervene on "Income" by giving 100 study participants $50.
- I haven't eliminated all of their other sources of income, i.e., I haven't "cut off" causes of their income.
  - So I haven't "set" anyone's income to a value.
- This type of intervention is sometimes called a **soft intervention**.

Modeling Soft Interventions



Key Move: I do know something about the relationship between the post-intervention income distribution P_I and the pre-intervention distribution P, namely:

$$P_I(Income = x | \mathrm{PA}_{\mathcal{G}}(Income)) = P(Income = x - 50 | \mathrm{PA}_{\mathcal{G}}(Income))$$

Predicting the effect of a soft intervention

General Question: How should we compute the effect of an intervention that changes the probability distribution of a particular variable V to some specific distribution Q?

Predicting the effect of a soft intervention

General Question: How should we compute the effect of an intervention that changes the probability distribution of a particular variable V to some specific distribution Q? Answer:

- $\mathcal{G}_{P(V)=Q}$ is identical to $\mathcal{G}$.
- $P_{P(V)=Q}$ is the unique distribution satisfying the following constraints:
    - $P_{P(V)=Q}(V|\mathrm{PA}_{\mathcal{G}}(V)) = Q$ (i.e., your intervention was successful),
    - $P_{P(V)=Q}(X|\mathrm{PA}_{\mathcal{G}}(X)) = P(X|\mathrm{PA}_{\mathcal{G}}(X))$ for all $X \neq V$, and
    - $P_{P(V)=Q}$ satisfies the CMC with respect to the causal graph $\mathcal{G}$.

Predicting the effect of a soft intervention

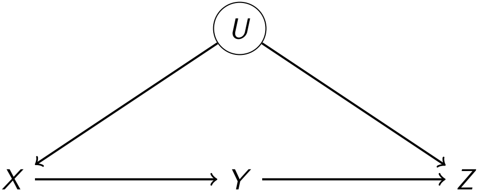
Moral: We can then use $P_{P(V)=Q}$ and $\mathcal{G}$ in the same way as we did before to predict the effect of an intervention.

Back-Door Criterion

Question 3: The formulas for calculating causal effects seem inapplicable if there are latent common causes (i.e., common causes that we have not measured). Is that true and can the assumption be dropped?

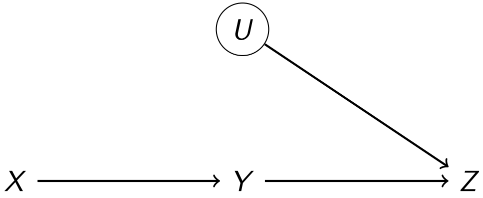
Answer: Yes and yes.

Front-Door Criterion



- Suppose you observe X, Y, and Z, but you have not observed U.
- What is the effect on Z of setting X to the value x (i.e., a "hard" intervention)?

Front-Door Criterion



Step 1: Cut the edge into X.

Front-Door Criterion

$$\begin{tikzcd} & & U \arrow[dr] & \\ X \arrow[r] & Y \arrow[rr] & & Z \end{tikzcd}$$

Step 2: Compute new probabilities

$$P(Z=z) = \sum_u \sum_y P(Z=z|Y=y, U=u)$$

Problem: But we don't know the distribution over U!

Front-Door Criterion

- **Luckily:** In this case you can rewrite the equation so that it does not contain U!
- Pearl provides a general set of conditions under which you can do this called the front door criterion.
- See [?]

Four Problems

- **Computational:** Is there a shorter formula that helps us calculate the probabilities?
- **"Soft" Interventions:** Some interventions don't cut off all causal paths to the variable.
- **No Latents:** Our procedure does not seem to work if there are latent variables we haven't measured.
- **Known Causal Structure:** Our procedure assumes we know the causal graph.

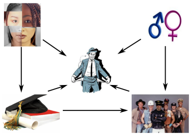
Taking Stock: We've discussed the first three questions. Let's attack the fourth.

Outline

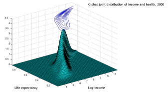
1. Causal Graphs
2. Markov and Faithfulness
3. Graphical Terminology
4. Estimating Causal Effects
5. Discovery
6. My Research

Utility of Bridge Principles

Bridge Principle



Causes



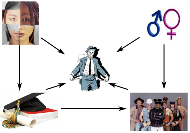
Probabilities

Income	Age	Gender	...
55K	37	F	...
19K	28	M	...
123K	58	M	...
⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮

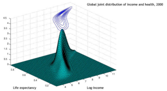
Data

Utility of Bridge Principles

Bridge Principle



Causes



Probabilities

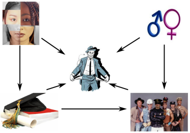
Statistics

Income	Age	Gender	...
55K	37	F	...
19K	28	M	...
123K	58	M	...
⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮

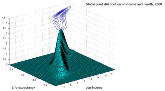
Data

Utility of Bridge Principles

Bridge Principle



Causes



Probabilities

Statistics

Causal Discovery

Income	Age	Gender	...
55K	37	F	...
19K	28	M	...
123K	58	M	...
⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮

Data

Statistical Tests for Independence

Income	Age	Gender	...
55K	42	M	...
87K	65	F	...
24K	63	M	...
⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮

Data

Probability

[?]

Statistical Tests for Independence

Income	Age	Gender	...
55K	42	M	...
87K	65	F	...
24K	63	M	...
⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮

Data

Statistics $\Longrightarrow$

Income $\not\!\perp\!\!\!\perp$ Gender
Income $\perp\!\!\!\perp$ Gender | Age
Income $\not\!\perp\!\!\!\perp$ Age
⋮

Probability

Assumption: We'll assume we can test for **conditional independence** among different variables. [?]

From Independence Tests to Causation

Question: How can knowing which variables are independent of others tell us about the underlying causal structure?

Recall our definition of blocking . . .

Blocking

A collection of variables $\mathcal{C}$ **blocks** a path from V_1 to V_2:

Blocking

A collection of variables $\mathcal{C}$ **blocks** a path from V_1 to V_2:

- The path contains a variable X in $\mathcal{C}$, and X is **not** a collider on the path.

Blocking

A collection of variables $\mathcal{C}$ **blocks** a path from V_1 to V_2:

- The path contains a variable X in $\mathcal{C}$, and X is **not** a collider on the path.
- The path contains a collider X that is **not** in $\mathcal{C}$, **and** none of X's descendants are in $\mathcal{C}$.

Blocking

A collection of variables $\mathcal{C}$ **blocks** a path from V_1 to V_2:

- The path contains a variable X in $\mathcal{C}$, and X is **not** a collider on the path.
- The path contains a collider X that is **not** in $\mathcal{C}$, **and** none of X's descendants are in $\mathcal{C}$.

A path is **active** given $\mathcal{C}$ if and only if it is not blocked by $\mathcal{C}$.

d-Separation

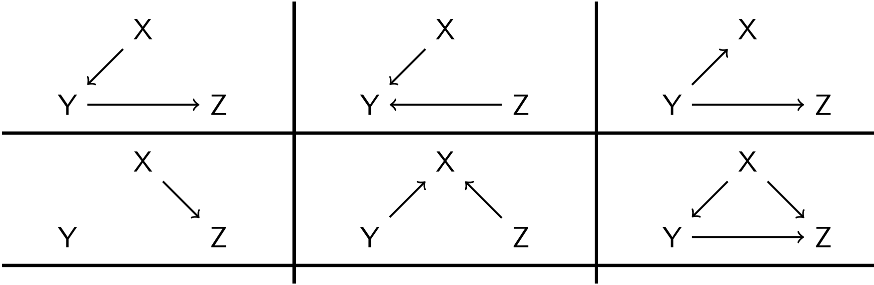
Definition: Two variables V_1 and V_2 are said to be d-separated given $\mathcal{C}$ if and only if every path between V_1 and V_2 is blocked by $\mathcal{C}$.

d-Separation and Conditional Independence

Theorem

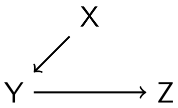
If the CMC and CFC hold, then two variables V_1 and V_2 in a causal graph over a causally sufficient set of variables $\mathcal{V}$ are d-separated given $\mathcal{C}$ if and only if V_1 and V_2 are conditionally independent given $\mathcal{C}$.

Exercise



Exercise: For each graph above, write down all of the conditional independence entailed by the Markov condition. Let's do one graph together.

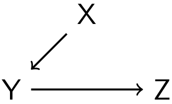
Worked Exercise: Reading Independences off a graph



Consider the pair $\langle X, Y \rangle$ first.

- There is a path $X \to Y$ between the two (an edge is a path!).

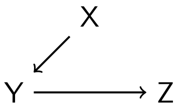
Worked Exercise: Reading Independences off a graph



Consider the pair $\langle X, Y \rangle$ first.

- There is a path $X \to Y$ between the two (an edge is a path!).
- Since there are no variables other than X and Y on the path, it is never blocked!

Worked Exercise: Reading Independences off a graph



Consider the pair $\langle X, Y \rangle$ first.

- There is a path $X \to Y$ between the two (an edge is a path!).
- Since there are no variables other than X and Y on the path, it is never blocked!
- So X and Y are not d-separated given any subset of $\{Z\}$.

Worked Exercise: Reading Independences off a graph

$$\begin{tikzcd}
 & X \arrow[dl] & \\
Y \arrow[rr] & & Z
\end{tikzcd}$$

Consider the pair $\langle X, Y \rangle$ first.

- There is a path $X \rightarrow Y$ between the two (an edge is a path!).
- Since there are no variables other than X and Y on the path, it is never blocked!
- So X and Y are not d-separated given any subset of $\{Z\}$.
- By the above theorem, X and Y are not conditionally independent given any subset of $\{Z\}$.

Worked Exercise: Reading Independences off a graph

$$\begin{tikzcd}
 & X \arrow[dl] & \\
Y \arrow[rr] & & Z
\end{tikzcd}$$

Consider the pair $\langle Y, Z \rangle$ next.

By the same reasoning, there are no conditional independences we should expect among these two variables.

Worked Exercise: Reading Independences off a graph

$$\begin{tikzcd} & X \arrow[dl] & \\ Y \arrow[rr] & & Z \end{tikzcd}$$

Finally, consider the pair $\langle X, Z \rangle$.

There is one path, $X \to Y \to Z$ between the two variables.

- The path is **not** blocked by the empty-set. By the theorem, X and Z are dependent given the empty set.

Worked Exercise: Reading Independences off a graph

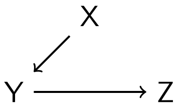
$$\begin{tikzcd} & X \arrow[dl] & \\ Y \arrow[rr] & & Z \end{tikzcd}$$

Finally, consider the pair $\langle X, Z \rangle$.

There is one path, $X \rightarrow Y \rightarrow Z$ between the two variables.

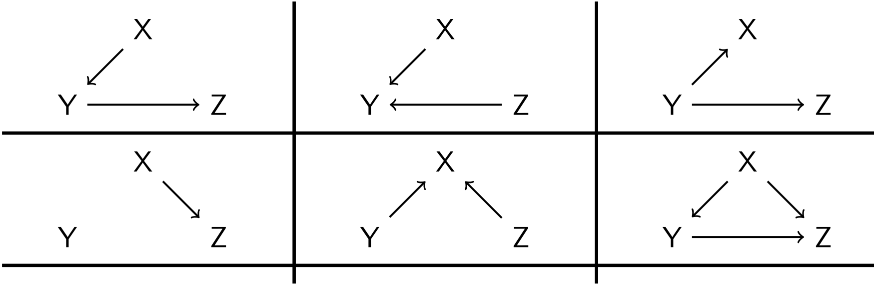
- The path is **not** blocked by the empty-set. By the theorem, X and Z are dependent given the empty set.
- It is blocked by $\{Y\}$. By the theorem, X and Z are conditionally independent given $\{Y\}$.

Worked Exercise: Reading Independences off a graph



So the only conditional independence that holds in the graph is $X \amalg Z | \{Y\}$.

Exercise



Exercise: For each graph above, write down all of the conditional independence entailed by the Markov condition.

Exercise Answers

$$\begin{tikzcd} & X \arrow[dl] & \\ Y \arrow[rr] & & Z \end{tikzcd}$$

$X \amalg Y | \{Z\}$

$$\begin{tikzcd} & X \arrow[dl] & \\ Y & & Z \arrow[ll] \end{tikzcd}$$

$X \amalg Z$

$$\begin{tikzcd} & X & \\ Y \arrow[ur] \arrow[rr] & & Z \end{tikzcd}$$

$X \amalg Z | \{Y\}$

$$\begin{tikzcd} & X \arrow[dr] & \\ Y & & Z \end{tikzcd}$$

$X \amalg Y | \emptyset, \{Z\}$
$Y \amalg Z | \emptyset, \{X\}$

$$\begin{tikzcd} & X & \\ Y \arrow[ur] & & Z \arrow[ul] \end{tikzcd}$$

$Y \amalg Z$

$$\begin{tikzcd} & X \arrow[dl] \arrow[dr] & \\ Y \arrow[rr] & & Z \end{tikzcd}$$

No independences

Exercise: Independence and Causal Discovery

I gather data concerning three variables among applicants to MIT: gender (G), admission to MIT (A), and SAT score (S). Here are my findings:

- Women are admitted to MIT at a higher rate then are men.
- But gender and SAT score are independent of one another.
- SAT score is positively correlated with admission to MIT.

Exercise: Assume there are no latent common causes of A, G, and S. What causal graphs are best supported by my findings?

Independence and Causal Discovery

$$
\begin{tikzcd}
\text{Gender} \arrow[dr] & & \text{SAT Score} \arrow[dl] \\
& \text{Admission to MIT} &
\end{tikzcd}
$$

Dependences and Independences
Gender $\not\perp\!\!\!\perp$ Admission
SAT $\not\perp\!\!\!\perp$ Admission
Gender $\perp\!\!\!\perp$ SAT Score
$\vdots$

Answer: Only the above graph. More on this example later, however . . .

Probability and Causation

Moral: If you believe your set of variables is causally sufficient,

- The CMC and CFC are useful because they allow one to infer **causal** conclusions from **probabilistic** relations among the variables.

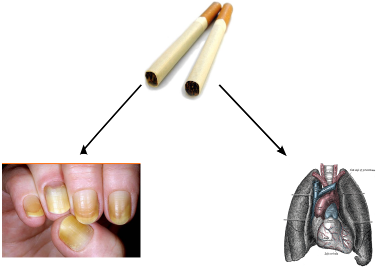
Probability and Causation

Moral: If you believe your set of variables is causally sufficient,

- The CMC and CFC are useful because they allow one to infer **causal** conclusions from **probabilistic** relations among the variables.
- The CMC and CFC associate every causal graph with a set of **sentences** concerning which variables are independent of others.

(We'll return to the question of causal inference with latent variables in a bit.)

Graphs and Independence Constraints



$$\Downarrow$$

"Lung cancer is independent of yellow fingernails given smoking."

$I(L, Y, S)$

Probability and Causation

- Since the CMC and CFC associate every causal graph with conditional independence statements,

Probability and Causation

- Since the CMC and CFC associate every causal graph with conditional independence statements,
- One can infer "backwards" from those statements (inferred from statistical tests) to causal structure.

Indistinguishable Causal Theories

- However, there will be multiple causal graphs satisfying the same probabilistic relations.

Indistinguishable Causal Theories

- However, there will be multiple causal graphs satisfying the same probabilistic relations.
- So several causal graphs might be indistinguishable assuming only the CMC and CFC.

Indistinguishable Causal Theories

$$Y \longleftarrow X \longrightarrow Z$$

Indistinguishable Causal Theories

$$\mathsf{Y} \longleftarrow \mathsf{X} \longrightarrow \mathsf{Z} \quad \xRightarrow{\text{CMC and CFC}} \quad Y \amalg Z | X$$

Indistinguishable Causal Theories

$\mathsf{Y} \longrightarrow \mathsf{X} \longrightarrow \mathsf{Z}$

$$\begin{tikzcd} \mathsf{Y} \longleftarrow \mathsf{X} \longrightarrow \mathsf{Z} \arrow[r, Rightarrow, bend left, "\text{CMC and CFC}"] & Y \perp\!\!\!\perp Z | X \arrow[l, Rightarrow, bend left, "\text{Discovery}"] \end{tikzcd}$$

$\mathsf{Y} \longleftarrow \mathsf{X} \longleftarrow \mathsf{Z}$

Indistinguishable Causal Theories

Nonetheless, indistinguishable causal graphs often share structure . . .

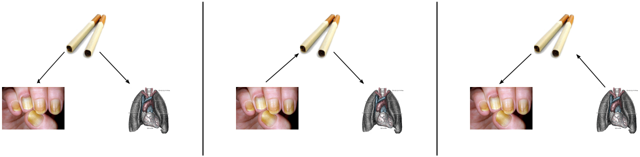
Markov Equivalence

"Lung cancer is independent of yellow fingernails given smoking."

Markov Equivalence

"Lung cancer is independent of yellow fingernails given smoking."

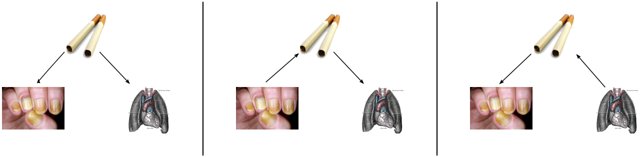
$$\Downarrow$$



Markov Equivalence

"Lung cancer is independent of yellow fingernails given smoking."

$$\Downarrow$$



$$\Downarrow$$

Developing yellow fingernails does not directly cause lung cancer, nor vice versa.

Background Information and Discovery

We will discuss what we can learn from statistical facts **alone**, but of course, some causal graphs will be ruled out by background information, such as

- Temporal information (e.g., People rarely start smoking after they acquire lung cancer),
- Plausible mechanisms (e.g., We have no idea how developing yellow fingernails could produce a desire to smoke),
- And more.

Conditional Independence Constraints

If G is a graph over the variables V, let I_G^V be the set of Independence statements associated with G, which describe which variables are conditionally independent of each other.

Markov Equivalence

Definition

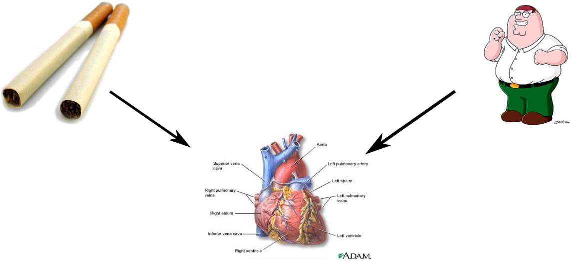
Two graphs are Markov Equivalent if they get mapped to the same set of independence sentences. In other words, G and H are Markov equivalent if $\mathrm{I}_G^V = \mathrm{I}_H^V$. Let $[G]$ be the Markov equivalence class of G.

Good News

As I said, Markov equivalent graphs often share quite a bit of structure.

Colliders

Say $X \to Z \leftarrow Y$ is an **unshielded collider** if neither X nor Y is a parent of one another.

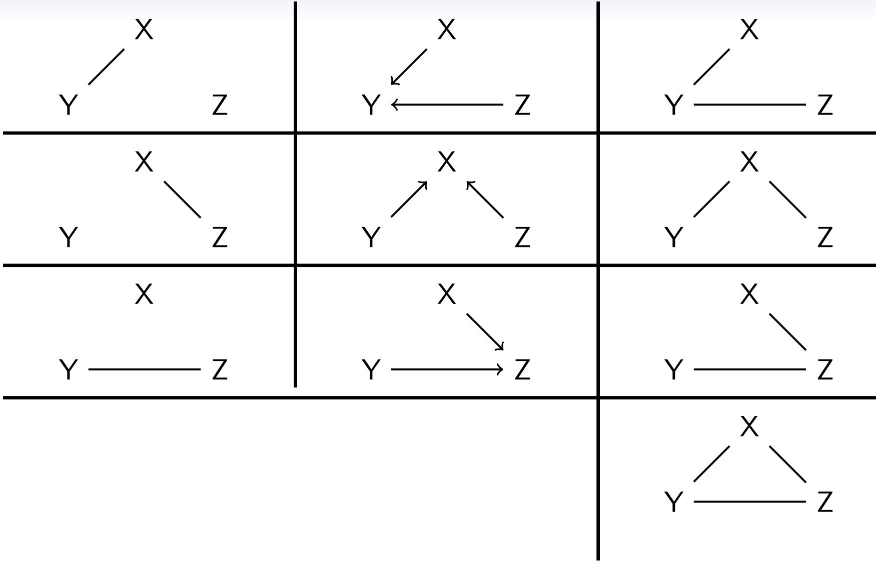


A Striking Theorem

Theorem ([Pearl and Verma, 1991])

Two causal graphs are Markov Equivalent if and only if the have the same edges and unshielded colliders.

Markov Equivalence Classes



Exercise: Markov Equivalence

$$X \longleftarrow Y \longrightarrow W \longrightarrow Z$$

$$X \longleftarrow Y \longleftarrow W \longrightarrow Z$$

$$X \longrightarrow Y \longleftarrow W \longrightarrow Z$$

$$X \longrightarrow W \longleftarrow Y \longrightarrow Z$$

Exercise: Which of above graphs are Markov equivalent?

Causal Discovery Algorithms

Moral: If you know which variables are conditionally independent of which others, you can

- Find one causal graph $\mathcal{G}$ that satisfies those independences, and
- Find all members of the Markov equivalence class of $\mathcal{G}$

The resulting class is the set of causal graphs that best explains your data $\mathcal{G}$.

Causal Discovery Algorithms

Question: How can we find a graph satisfying all our observed independences?

Answer: There are various algorithms you can use [Spirtes et al., 2000].

SGS Algorithm

Step 1: Draw a graph in which all variables are connected by **undirected** edges.

Step 2: If $X \amalg Y | \mathcal{C}$ for some set $\mathcal{C}$, remove the edge $X - Y$ from your graph.

Step 3: For each triple of variables X, Y, Z, if $X - Y$ and $Y - Z$ are in the graph, but $X - Z$ is not, orient the edges $X \rightarrow Y \leftarrow Z$ if $X \not\amalg Z | \mathcal{C}$ for all $\mathcal{C}$ containing Y.

Step 4: Repeat the following procedure until you can no longer orient any more edges:

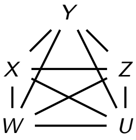
- If $X \rightarrow Y$ and $Y - Z$ appear in the graph, then orient $Y \rightarrow Z$.
- If there is a directed path from X to Y and the edge $X - Y$ is in the graph, then orient the edge $X \rightarrow Y$.

SGS Algorithm: Example

Among five variables $\{U, W, X, Y, Z\}$, you find only the following independences:

$$
\begin{aligned}
&U \amalg X | \{W\}, \{W, Z\}, \{W, Y, Z\} \\
&U \amalg Y | \{X\}, \{W, Z\}, \{W, X, Z\} \\
&U \amalg Z \\
&W \amalg Y | \{X, Z\}, \{X, Z, U\} \\
&X \amalg Z | \{W\}, \{W, U\}
\end{aligned}
$$

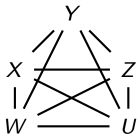
SGS Algorithm: Example



$U \perp\!\!\!\perp X | \{W\}, \{W, Z\}, \{W, Y, Z\}$
$U \perp\!\!\!\perp Y | \{X\}, \{W, Z\}, \{W, X, Z\}$
$U \perp\!\!\!\perp Z$
$W \perp\!\!\!\perp Y | \{X, Z\}, \{X, Z, U\}$
$X \perp\!\!\!\perp Z | \{W\}, \{W, U\}$

Step 1: Draw a graph in which all variables are connected by undirected edges.

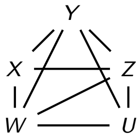
SGS Algorithm: Example



$U \amalg X | \{W\}, \{W, Z\}, \{W, Y, Z\}$
$U \amalg Y | \{X\}, \{W, Z\}, \{W, X, Z\}$
$U \amalg Z$
$W \amalg Y | \{X, Z\}, \{X, Z, U\}$
$X \amalg Z | \{W\}, \{W, U\}$

Step 2: If $X \amalg Y | \mathcal{C}$ for some set $\mathcal{C}$, remove the edge $X - Y$ from your graph.

SGS Algorithm: Example

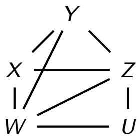


$U \perp\!\!\!\perp X | \{W\}, \{W, Z\}, \{W, Y, Z\}$
$U \perp\!\!\!\perp Y | \{X\}, \{W, Z\}, \{W, X, Z\}$
$U \perp\!\!\!\perp Z$
$W \perp\!\!\!\perp Y | \{X, Z\}, \{X, Z, U\}$
$X \perp\!\!\!\perp Z | \{W\}, \{W, U\}$

Step 2: If $X \perp\!\!\!\perp Y | \mathcal{C}$ for some set $\mathcal{C}$, remove the edge $X - Y$ from your graph.

- Remove $U - X$.

SGS Algorithm: Example

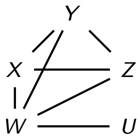


$U \amalg X | \{W\}, \{W, Z\}, \{W, Y, Z\}$
$U \amalg Y | \{X\}, \{W, Z\}, \{W, X, Z\}$
$U \amalg Z$
$W \amalg Y | \{X, Z\}, \{X, Z, U\}$
$X \amalg Z | \{W\}, \{W, U\}$

Step 2: If $X \amalg Y | \mathcal{C}$ for some set $\mathcal{C}$, remove the edge $X - Y$ from your graph.

- Remove $U - X$.
- Remove $U - Y$.

SGS Algorithm: Example

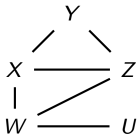


$U \amalg X | \{W\}, \{W, Z\}, \{W, Y, Z\}$
$U \amalg Y | \{X\}, \{W, Z\}, \{W, X, Z\}$
$U \amalg Z$
$W \amalg Y | \{X, Z\}, \{X, Z, U\}$
$X \amalg Z | \{W\}, \{W, U\}$

Step 2: If $X \amalg Y | \mathcal{C}$ for some set $\mathcal{C}$, remove the edge $X - Y$ from your graph.

- Remove $U - X$.
- Remove $U - Y$.
- Remove $U - Z$.

SGS Algorithm: Example

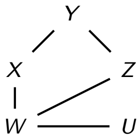


$U \perp\!\!\!\perp X | \{W\}, \{W, Z\}, \{W, Y, Z\}$
$U \perp\!\!\!\perp Y | \{X\}, \{W, Z\}, \{W, X, Z\}$
$U \perp\!\!\!\perp Z$
$W \perp\!\!\!\perp Y | \{X, Z\}, \{X, Z, U\}$
$X \perp\!\!\!\perp Z | \{W\}, \{W, U\}$

Step 2: If $X \perp\!\!\!\perp Y | \mathcal{C}$ for some set $\mathcal{C}$, remove the edge $X - Y$ from your graph.

- Remove $U - X$.
- Remove $U - Y$.
- Remove $U - Z$.
- Remove $W - Y$.

SGS Algorithm: Example

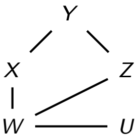


$U \perp\!\!\!\perp X | \{W\}, \{W, Z\}, \{W, Y, Z\}$
$U \perp\!\!\!\perp Y | \{X\}, \{W, Z\}, \{W, X, Z\}$
$U \perp\!\!\!\perp Z$
$W \perp\!\!\!\perp Y | \{X, Z\}, \{X, Z, U\}$
$X \perp\!\!\!\perp Z | \{W\}, \{W, U\}$

Step 2: If $X \perp\!\!\!\perp Y | \mathcal{C}$ for some set $\mathcal{C}$, remove the edge $X - Y$ from your graph.

- Remove $U - X$.
- Remove $U - Y$.
- Remove $U - Z$.
- Remove $W - Y$.
- Remove $X - Z$.

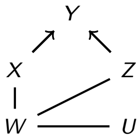
SGS Algorithm: Example



$U \perp\!\!\!\perp X | \{W\}, \{W, Z\}, \{W, Y, Z\}$
$U \perp\!\!\!\perp Y | \{X\}, \{W, Z\}, \{W, X, Z\}$
$U \perp\!\!\!\perp Z$
$W \perp\!\!\!\perp Y | \{X, Z\}, \{X, Z, U\}$
$X \perp\!\!\!\perp Z | \{W\}, \{W, U\}$

Step 3: For each triple of variables V_1, V_2, V_3, if $V_1 - V_2$ and $V_2 - V_3$ are in the graph, but $V_1 - V_3$ is not, orient the edges $V_1 \rightarrow V_2 \leftarrow V_3$ if $V_1 \not\perp\!\!\!\perp V_3 | \mathcal{C}$ for all $\mathcal{C}$ containing V_2.

SGS Algorithm: Example

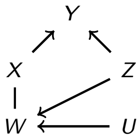


$U \perp\!\!\!\perp X | \{W\}, \{W, Z\}, \{W, Y, Z\}$
$U \perp\!\!\!\perp Y | \{X\}, \{W, Z\}, \{W, X, Z\}$
$U \perp\!\!\!\perp Z$
$W \perp\!\!\!\perp Y | \{X, Z\}, \{X, Z, U\}$
$X \perp\!\!\!\perp Z | \{W\}, \{W, U\}$

Step 3: For each triple of variables V_1, V_2, V_3, if $V_1 - V_2$ and $V_2 - V_3$ are in the graph, but $V_1 - V_3$ is not, orient the edges $V_1 \rightarrow V_2 \leftarrow V_3$ if $V_1 \not\!\perp\!\!\!\perp V_3 | \mathcal{C}$ for all $\mathcal{C}$ containing V_2.

- Orient $X \rightarrow Y \leftarrow Z$ as $X \not\!\perp\!\!\!\perp Z | \mathcal{C}$ for all $\mathcal{C}$ containing Y.

SGS Algorithm: Example

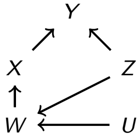


$U \perp\!\!\!\perp X | \{W\}, \{W, Z\}, \{W, Y, Z\}$
$U \perp\!\!\!\perp Y | \{X\}, \{W, Z\}, \{W, X, Z\}$
$U \perp\!\!\!\perp Z$
$W \perp\!\!\!\perp Y | \{X, Z\}, \{X, Z, U\}$
$X \perp\!\!\!\perp Z | \{W\}, \{W, U\}$

Step 3: For each triple of variables V_1, V_2, V_3, if $V_1 - V_2$ and $V_2 - V_3$ are in the graph, but $V_1 - V_3$ is not, orient the edges $V_1 \rightarrow V_2 \leftarrow V_3$ if $V_1 \not\perp\!\!\!\perp V_3 | \mathcal{C}$ for all $\mathcal{C}$ containing V_2.

- Orient $X \rightarrow Y \leftarrow Z$ as $X \not\perp\!\!\!\perp Z | \mathcal{C}$ for all $\mathcal{C}$ containing Y.
- Orient $Z \rightarrow W \leftarrow U$ as $U \not\perp\!\!\!\perp Z | \mathcal{C}$ for all $\mathcal{C}$ containing W.

SGS Algorithm: Example



$U \perp\!\!\!\perp X | \{W\}, \{W, Z\}, \{W, Y, Z\}$
$U \perp\!\!\!\perp Y | \{X\}, \{W, Z\}, \{W, X, Z\}$
$U \perp\!\!\!\perp Z$
$W \perp\!\!\!\perp Y | \{X, Z\}, \{X, Z, U\}$
$X \perp\!\!\!\perp Z | \{W\}, \{W, U\}$

Step 4: Repeat the following procedure until you can no longer orient any more edges:

- If $X \rightarrow Y$ and $Y - Z$ appear in the graph, then orient $Y \rightarrow Z$.
- If there is a directed path from X to Y and the edge $X - Y$ is in the graph, then orient the edge $X \rightarrow Y$.

Causal Discovery Algorithms

Question: Are there computer programs that will (i) conduct conditional independence tests for me, and then (b) output all causal graphs that are compatible with my data?

Answer: Yes. Download `http://www.phil.cmu.edu/tetrad/current.html`.

Latent Variables

Problem:

- The CMC holds only when the set of variables are causally sufficient (i.e., there are no unmeasured common causes of any variables in my set).
- But that's never true in practice.
- So can any of this help me discover causal structure in the real world?

Here's an example to motivate the problem ...

Inference without sufficiency

Example: As you look at applicants for a job in your research group, you notice there is a strong correlation between grey hair and CV length.

If these variables were causally sufficient, the CMC and CFC would entail that one is a cause of the other.

- So you would get a longer CV by dyeing your hair grey, or
- Padding your CV would make your hair grey.

Causal discovery with latent variables

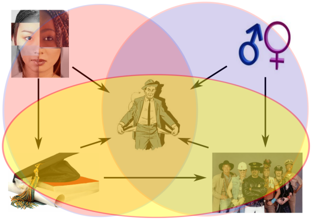
Solution: To correct this problem, we need to

- Introduce a different type of graph, which can represent the possibility that two variables share an unmeasured common cause, and
- Explain how to make inferences from statistical data to these other graphs.

Both of these tasks are possible. See the discussion of the FCI algorithm in [Spirtes et al., 2000]

My Research

Combining Studies



Causal Structure

Bridge Principle

Income	Gender	Occupation
55K	F	Programmer
65K	M	Accountant
32K	M	Teacher
⋮	⋮	⋮

Income	Race	Education
55K	Black	College
65K	White	College
32K	Hispanic	College
⋮	⋮	⋮

Income	Education	Occupation
55K	College	Programmer
65K	White	Accountant
32K	Hispanic	Teacher
⋮	⋮	⋮

Multiple Data Sets

Limitations of Verma and Pearl's Theorem

- Recall, Verma and Pearl's theorem showed that if we can list all the conditional independences among variables $\mathcal{V}$, then assuming the CFC, CMC, and causal sufficiency, we can determine the unshielded colliders and direct causal connections in the true graph.

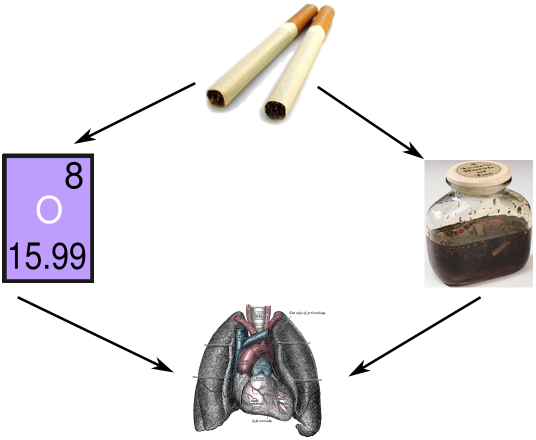
Limitations of Verma and Pearl's Theorem

- Recall, Verma and Pearl's theorem showed that if we can list all the conditional independences among variables $\mathcal{V}$, then assuming the CFC, CMC, and causal sufficiency, we can determine the unshielded colliders and direct causal connections in the true graph.
- Applying Verma and Pearl's theorem, however, requires measuring all variables simultaneously.

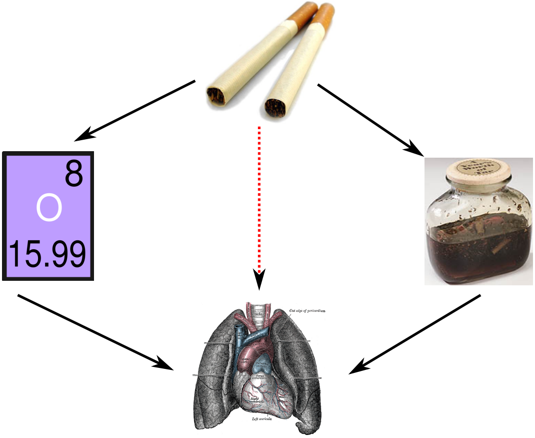
Limitations of Verma and Pearl's Theorem

- Recall, Verma and Pearl's theorem showed that if we can list all the conditional independences among variables $\mathcal{V}$, then assuming the CFC, CMC, and causal sufficiency, we can determine the unshielded colliders and direct causal connections in the true graph.
- Applying Verma and Pearl's theorem, however, requires measuring all variables simultaneously.
- Why? For example, one graph G might imply v is independent of v' given some large number of variables S, whereas H might not. But to detect whether such an independence holds, one generally must measure v, v' and S simultaneously.

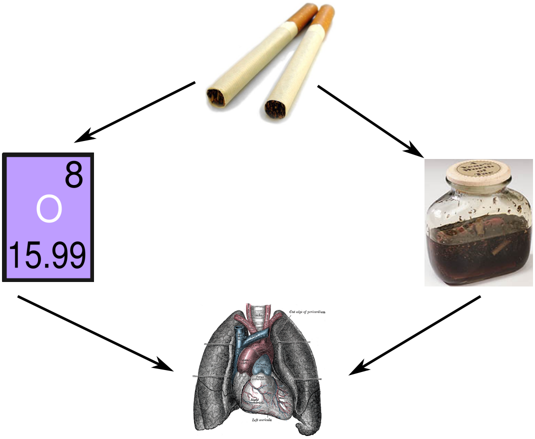
A Fictitious Example



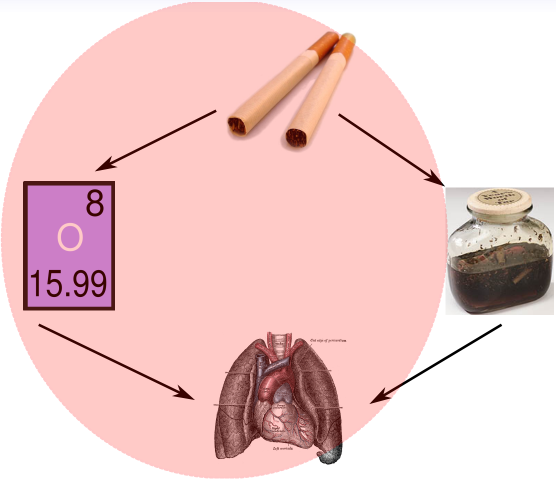
A Fictitious Example



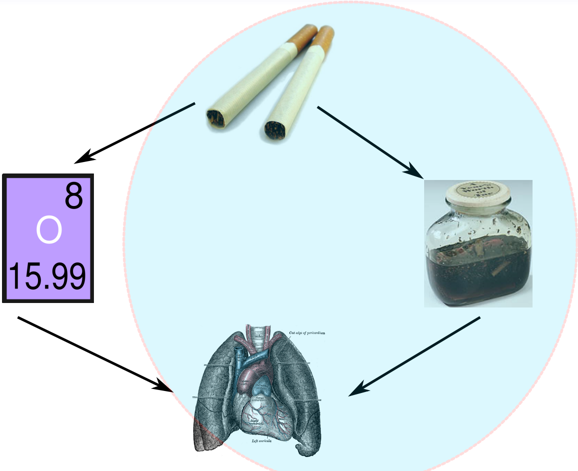
A Fictitious Example

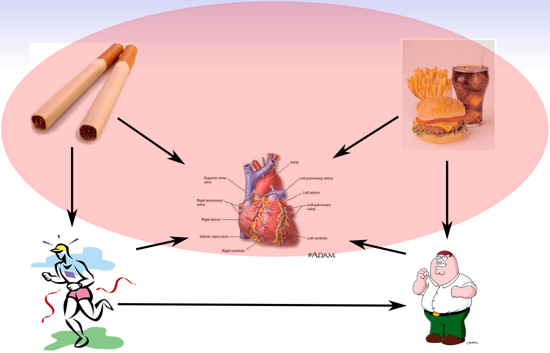


A Fictitious Example



A Fictitious Example





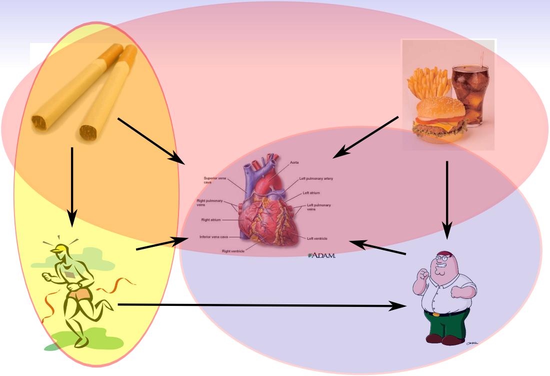
Study 1	Independence Facts
Heart Disease	$I(S, D, \emptyset)$
Smoking	
Diet	

Let V be the variables under investigation.

- Think of a subset $U \subset V$ of as a single observational study.

Let V be the variables under investigation.

- Think of a subset $U \subset V$ of as a single observational study.
- If one conducted the study U, one would learn I_G^U, the set of conditional independence constraints associated with the graph G that mention only the variables in U.



Study 1	Study 2	Study 3
Heart Disease	Heart Disease	Smoking
Smoking	Weight	Exercise
Diet		

- Think of a collection of observational studies is a subset $\mathcal{U} \subseteq 2^V$ of the power set of the variables under investigation.

- Think of a collection of observational studies is a subset $\mathcal{U} \subseteq 2^V$ of the power set of the variables under investigation.
- If one conducts the studies $\mathcal{U}$, one would learn $\mathrm{I}_G^{\mathcal{U}} := \bigcup_{U \in \mathcal{U}} \mathrm{I}_G^U$.

$\mathcal{U}$-Equivalence

Let V be any set, and $\mathcal{U} \subseteq 2^V$.

Definition

Two graphs G and H are $\mathcal{U}$-Equivalent if $\mathrm{I}_G^{\mathcal{U}} = \mathrm{I}_H^{\mathcal{U}}$. Write $G \equiv_{\mathcal{U}} H$ when this is the case, and let $[G]_{\mathcal{U}}$ be the $\mathcal{U}$-equivalence class of G.

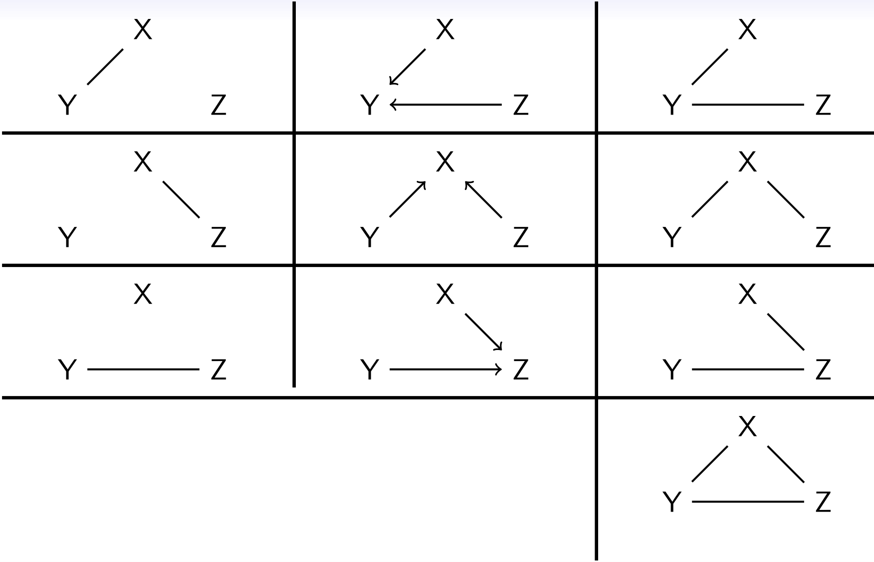
k-Equivalence

Definition

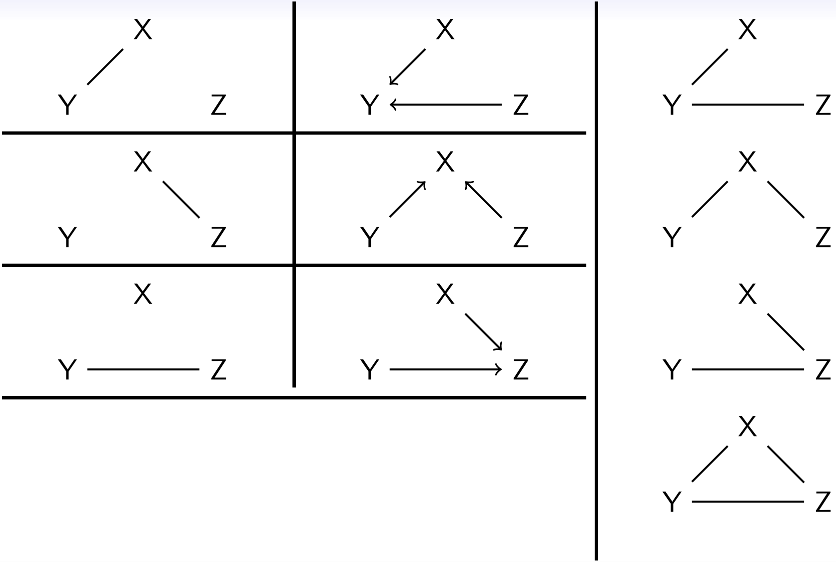
Let $k < |V|$ and $\mathcal{U}_k$ be all subsets of V of size k. Say two graphs G and H are k-Equivalent if $G \equiv_{\mathcal{U}_k} H$.

How do indistinguishability relations for various bridge principles change (if at all) if one must synthesize causal theories inferred from multiple data sets?

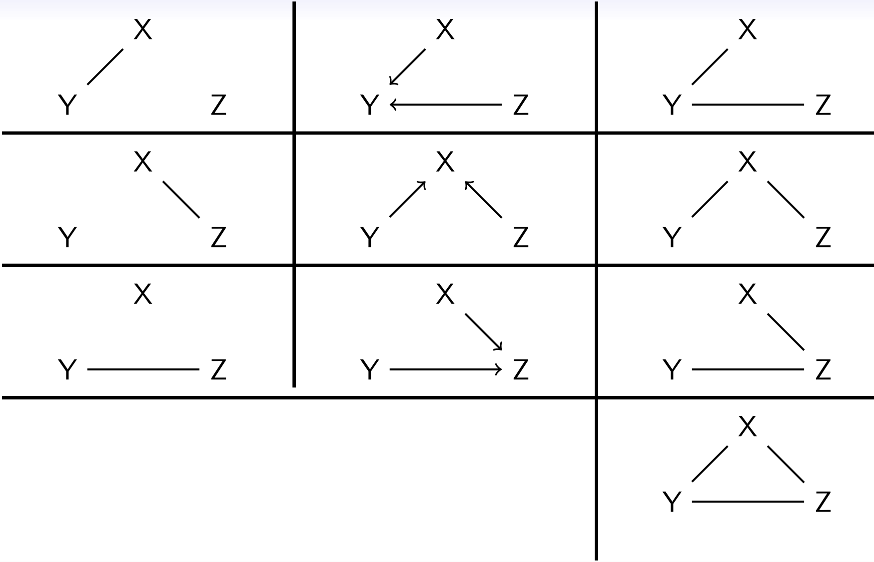
Markov Equivalence Classes



2-Equivalence Classes



2-Equivalence, Binary variables, Positive Joint



Existing research

- **My past research:** Characterizing these indistinguishability relations when studies are combined.
- **Other research:** Developing **algorithms** for inferring causal structure from overlapping studies [Danks, 2002, Tillman et al., 2008, Tillman and Spirtes, 2011]

Questions?

Danks, D. (2002). Learning the causal structure of overlapping variable sets. In *Discovery Science*, pages 178–191.

Pearl, J. (1993). [Bayesian Analysis in Expert Systems]: Comment: Graphical Models, Causality and Intervention. *Statistical Science*, pages 266–269.

Pearl, J. (2000). *Causality: models, reasoning, and inference*, volume 47. Cambridge Univ Press.

Pearl, J. and Verma, T. (1991). A theory of inferred causation.

Spirtes, P., Glymour, C. N., and Scheines, R. (2000). *Causation, prediction, and search*. The MIT Press.

Tillman, R., Danks, D., and Glymour, C. (2008). Integrating locally learned causal structures with overlapping variables. *Advances in Neural Information Processing Systems*, 21:1665–1672.

Tillman, R. and Spirtes, P. (2011). Learning equivalence classes of acyclic models with latent and selection variables from multiple datasets with overlapping variables. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS 2011)*.