Causal Fairness Analysis

(Causal Inference II - Lecture 6)

Elias Bareinboim 

Drago Plecko 

Columbia University
Computer Science

Reference:

D. Plecko, E. Bareinboim.

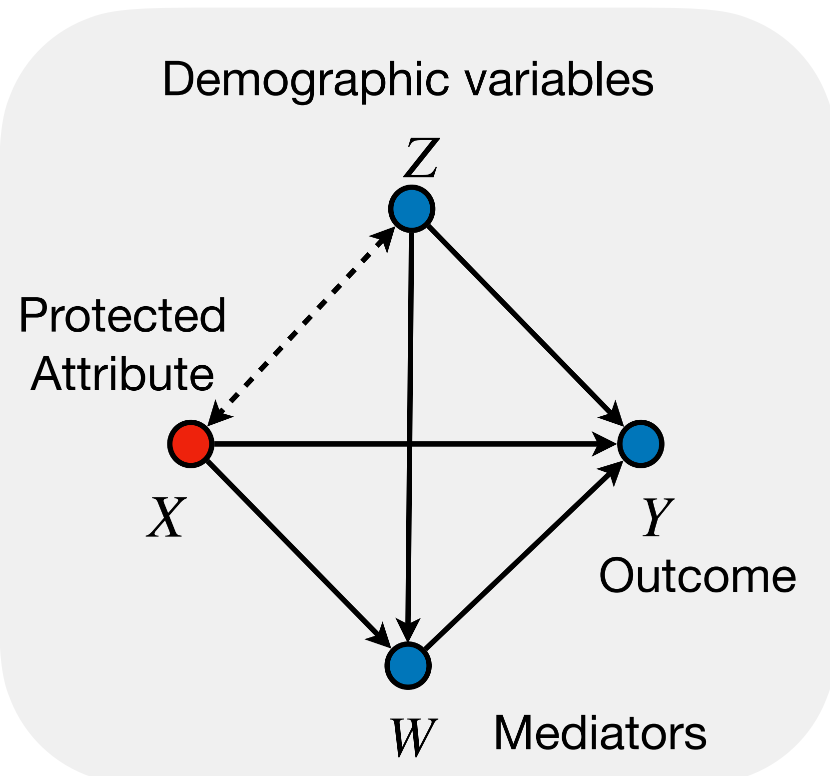
Causal Fairness Analysis.

TR R-90, CausalAI Lab, Columbia University.

https://causalai.net/r90.pdf

Moving beyond SFM

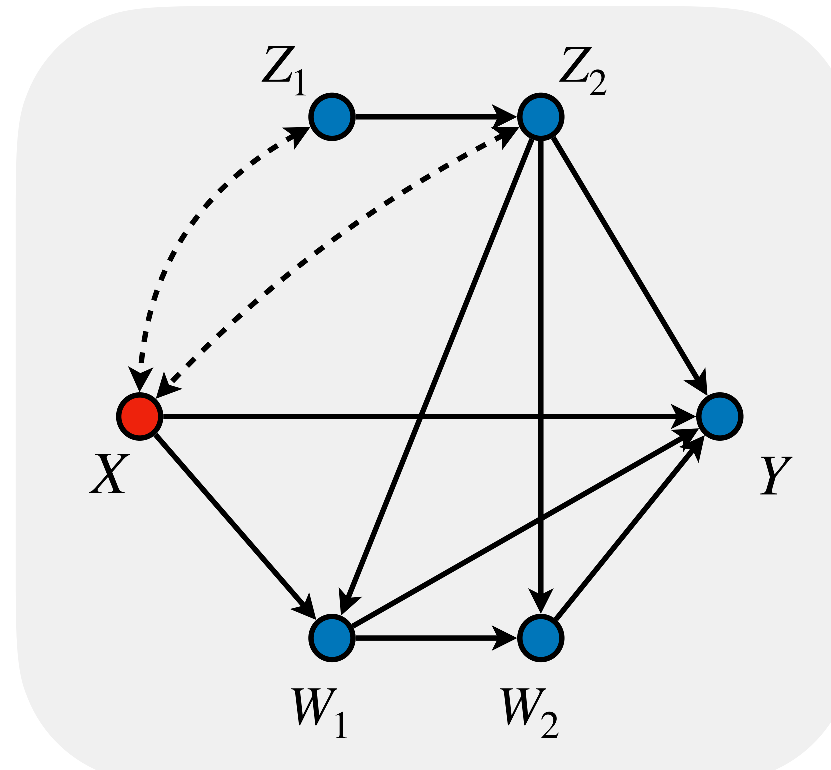
SFM



better resolution

Section 6

Diagram $\mathscr{G}$



Measures	direct, indirect, spurious
Business Necessity	$\{\{\varnothing\}, \{Z\}, \{W\}, \{Z, W\}\}$
Fair Prediction	Causal IF

Measures	variable specific
Business Necessity	any $V' \subseteq V$
Fair Prediction	fairadapt

Motivating Example

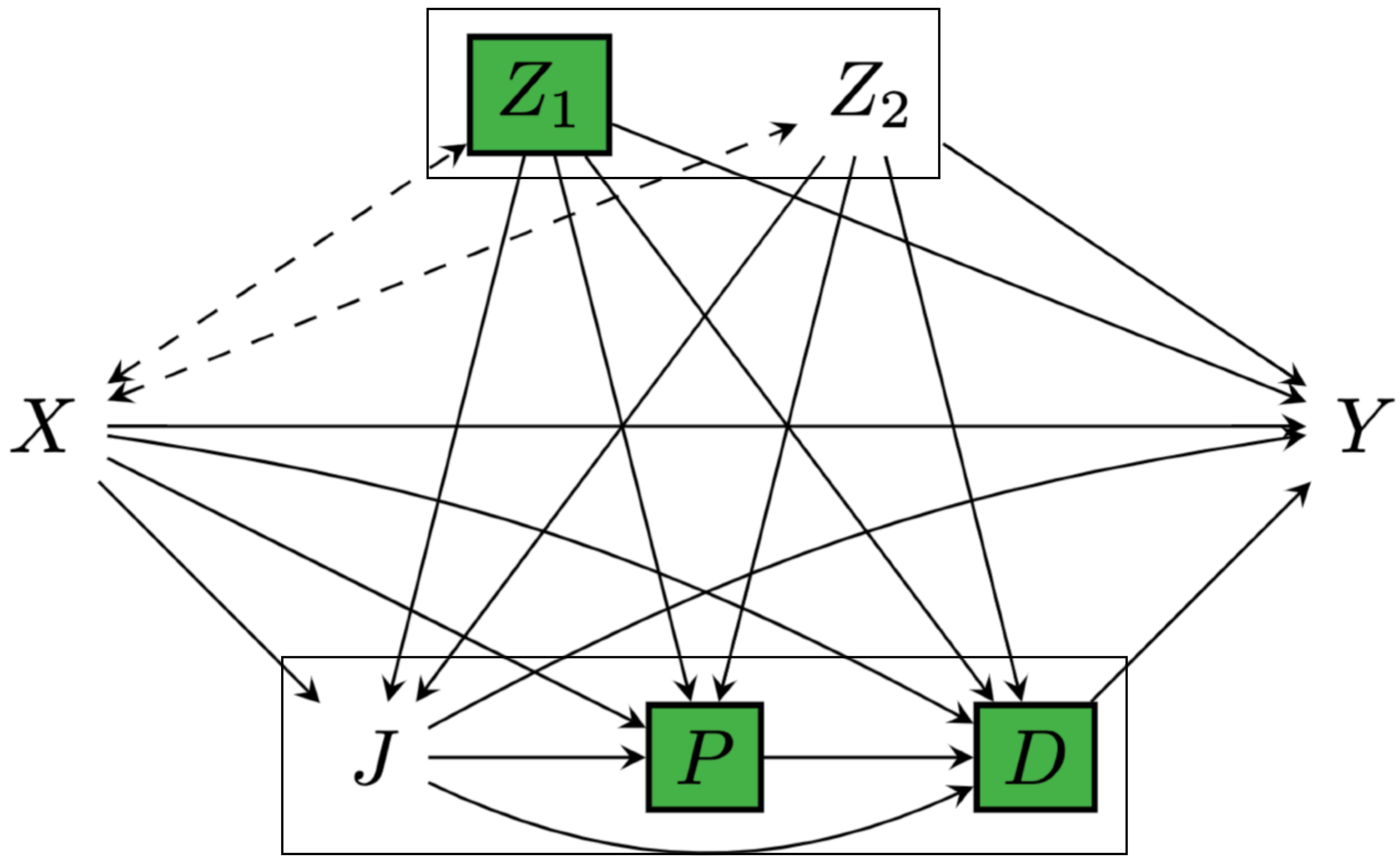
Example (COMPAS Business Necessity). Courts at Broward County, Florida, predict the risk of re-offending within 2 years, based on demographic information Z (Z_1 for gender, Z_2 for age), race X (x_0 denoting Majority, x_1 Minority), juvenile offense counts J, prior offense count P, and degree of charge D. A causal analysis using the Fairness Cookbook by ProPublica revealed that:

$$\text{Ctf-IE}_{x_1,x_0}(y \mid x_1) = -5.7\% \pm 0.5\%,$$
$$\text{Ctf-SE}_{x_1,x_0}(y) = -4.0\% \pm 0.9\%,$$

After the court hearing, the judge ruled that using the attributes age (Z_2), prior count (P), and charge degree (D) were not discriminatory, but using the attributes juvenile count (J) and gender (Z_1) was.

How can the ProPublica extend their findings based on this decision?

Motivating Example



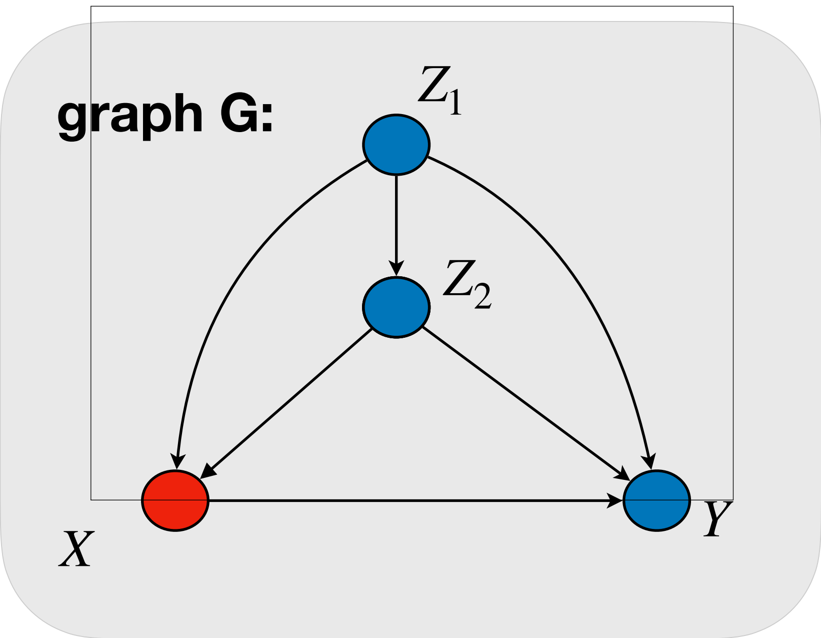
$$\text{Ctf-SE}_{x_1,x_0}(y) = \underbrace{\text{Ctf-SE}^{Z_1}_{x_1,x_0}(y)}_{\text{gender}} + \underbrace{\text{Ctf-SE}^{Z_2}_{x_1,x_0}(y)}_{\text{age}},$$

$$\text{Ctf-IE}_{x_1,x_0}(y \mid x_1) = \underbrace{\text{Ctf-IE}^{J}_{x_1,x_0}(y \mid x_1)}_{\text{juvenile count}} + \underbrace{\text{Ctf-IE}^{P}_{x_1,x_0}(y \mid x_1)}_{\text{prior count}} + \underbrace{\text{Ctf-IE}^{D}_{x_1,x_0}(y \mid x_1)}_{\text{charge degree}}.$$

Refining Spurious Effects

- We start by refining the spurious effect notion Exp-SE$_x(y)$

- What is our target in terms of Structural Fairness?



$$\text{Str-SE-BN}_X(Y) = \mathbb{1}(\text{an}^{\text{ex}}(Y) \cap \text{an}^{\text{ex}}(X) \cap U^C_{BN} = \emptyset)\,.$$

- How can we get a decomposition

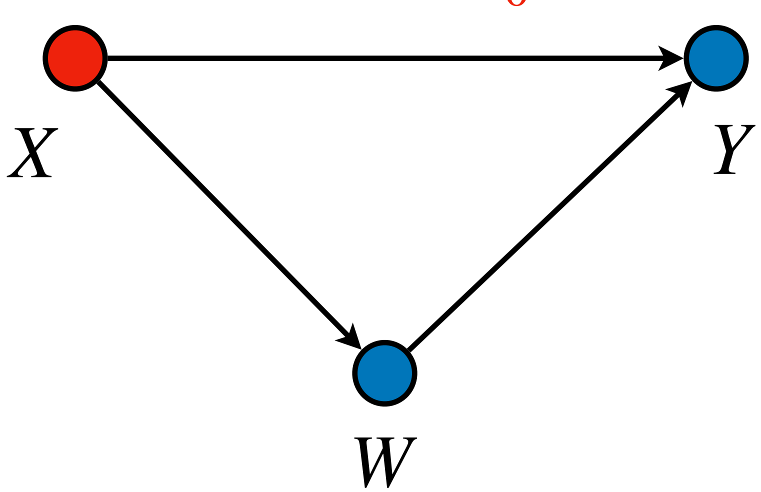
$$\text{Exp-SE}_x(y) = \text{Exp-SE}_x^{Z_1}(y) + \text{Exp-SE}_x^{Z_2}(y)$$?

New Primitive: Intuition

$$\mathscr{M}_{X=x}$$

"Behaves as" or
"Listens to" metaphor

$P(y_{x_1, W_{x_0}})$



Y listens to $X = x_1$,
W listens to $X = x_0$

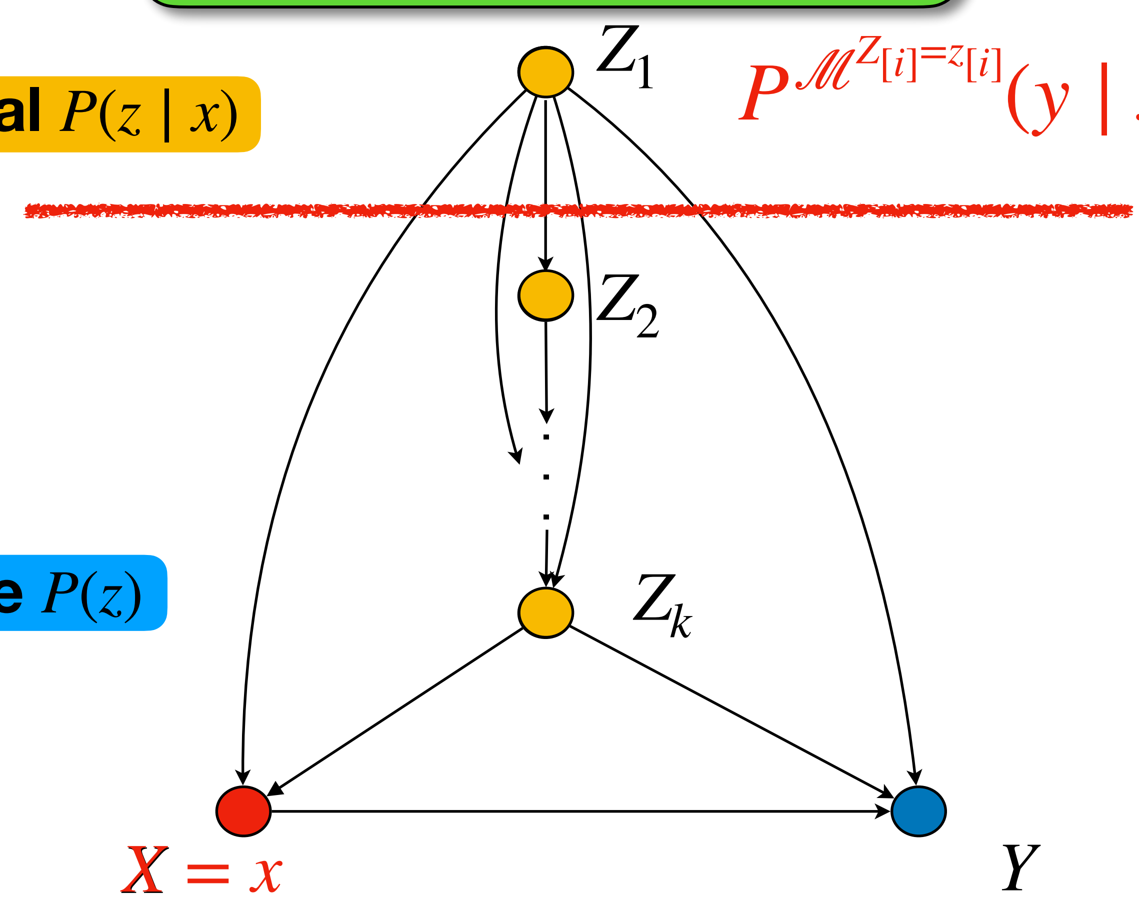
$$\mathscr{M}^{Z=z}$$

"Aware of" metaphor

Observational $P(z \mid x)$

$P^{\mathscr{M}^{Z[i]=z[i]}}(y \mid x)$

RCT style $P(z)$



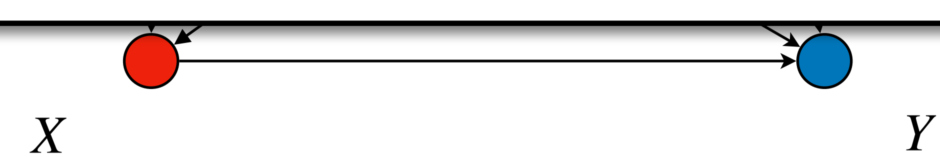
$Z_1, \ldots, Z_i$ unaware of the fact that $X = x$

Basic Idea: Integrated Submodel

Definition. Let $\mathscr{M}$ be an SCM. Let $Z' \subseteq Z$ be a subset of the exogenous variables. Define by $\mathscr{M}^{Z'}$ the following SCM

$$\mathscr{M}^{Z'=z'} = \sum_{z} P^{\mathscr{M}}(Z' = z')\mathscr{M}_{Z'=z'}.$$

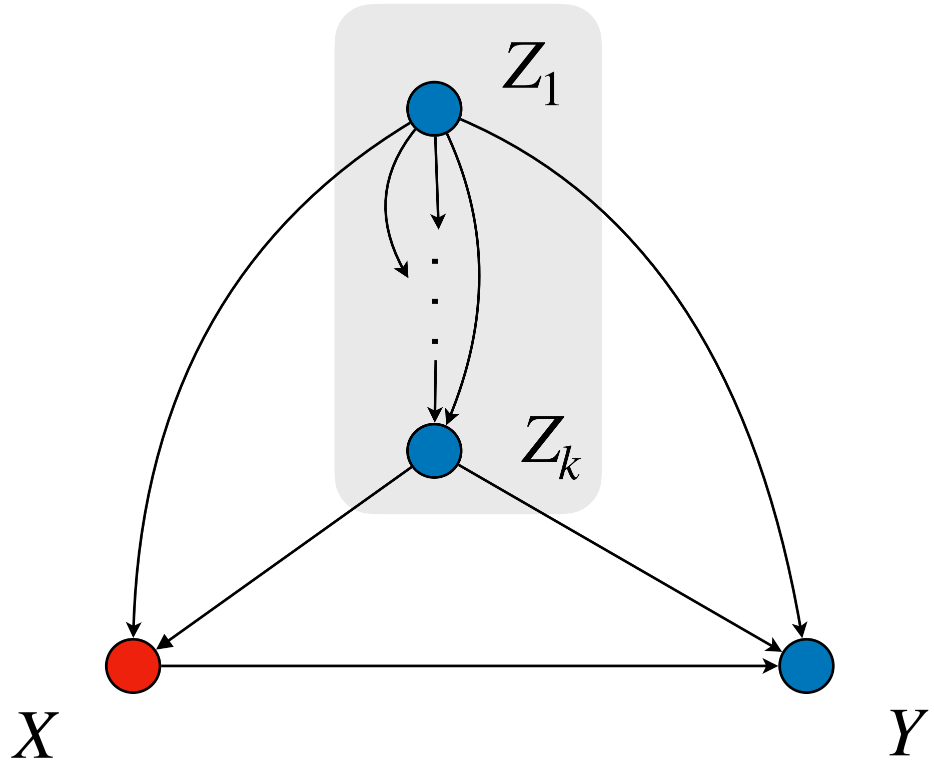
That is, in $\mathscr{M}^{Z'}$ the variables Z' are sampled from the observational distribution of the SCM, after which the submodel $\mathscr{M}_{Z'=z'}$ is used to obtain all other observables $V \backslash Z'$.



like in a randomized control trial

Basic Idea: Integrated Submodel

fixing Zs one-by-one



Z empty $\Longrightarrow X, Y$ associated as in the observational $P(V)$

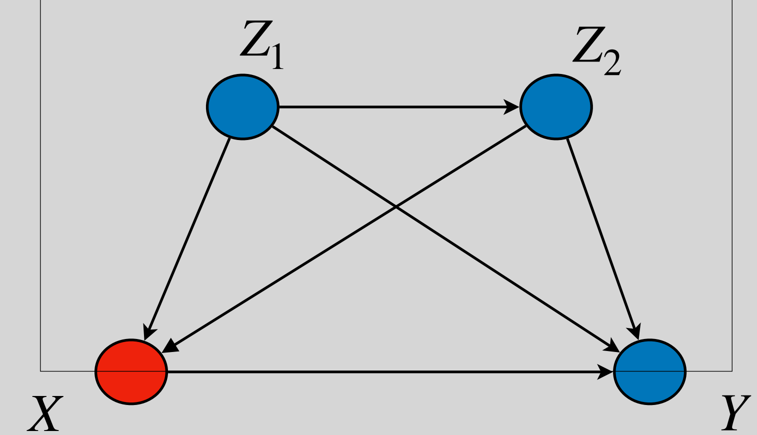
Z neither empty nor full $\Longrightarrow X, Y$ associated by some, but not all Us

all of $Z \Longrightarrow X, Y$ independent like in a randomized control trial

I-Submodel: Example

- How are conditional probabilities computed?

$P(y \mid x)$ in $\mathcal{M}^{Z_1}$ for



$$P^{\mathcal{M}^{Z_1}}(y \mid x) = \sum_{z_1} P^{\mathcal{M}}(z_1) P^{\mathcal{M}}(y \mid x, do(z_1))$$

$$= \sum_{z_1} P^{\mathcal{M}}(z_1) P^{\mathcal{M}}(y \mid x, z_1) \quad \text{(2nd rule of do-calculus)}$$

$$= \sum_{z_1, z_2} P^{\mathcal{M}}(z_1) P^{\mathcal{M}}(z_2 \mid z_1, x) P^{\mathcal{M}}(y \mid x, z_1, z_2).$$

ID!

Sampling-Evaluation Loop's Perspective

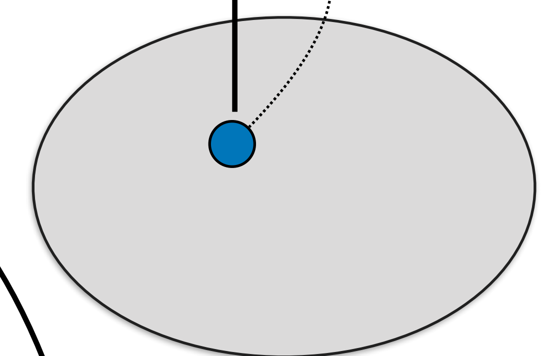
unit $u = (u_1, \ldots, u_k)$

$$Z_1 \leftarrow f_1(u_1)$$

$$Z_2 \leftarrow f_2(z_1, u_2)$$

$$\vdots$$

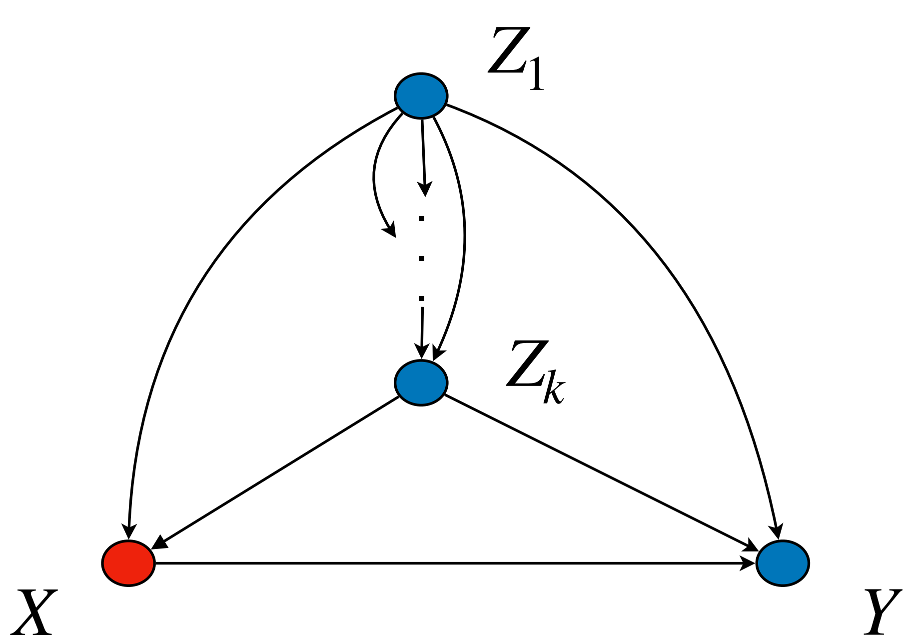
$$Y \leftarrow f_k(z_1, \ldots, z_{k_1}, x, u_k)$$



space of units $\mathscr{U}$

when you condition, information doesn't go past the red line!

How is this different than usual?



A. Sample a value of $Z_1 = z_1$

B. sample from the SCM $\mathscr{M}_{Z_1=z_1}$

C. take the mixture SCM $\sum_{z_1} P(z_1)\mathscr{M}_{Z_1=z_1}$

Spurious Decomposition

Theorem. Let $U_1, \ldots, U_k$ be the subset of exogenous variables that lie on top of a spurious trek between X and Y. Let $Z_{[i]}$ denote the variables $Z_1, \ldots, Z_i$ ($Z_{[0]}$ denotes the empty set $\emptyset$). Then, using the term

$$\textit{Exp-SE}_x^{A,B}(y) = P^{\mathcal{M}^A}(y \mid x) - P^{\mathcal{M}^B}(y \mid x),$$

we can decompose the experimental spurious effect as follows:

$$\begin{aligned}
\textit{Exp-SE}_x(y) &= P(y \mid x) - P(y_x) \\
&= \sum_{i=0}^{k-1} \textit{Exp-SE}_x^{Z_{[i]}, Z_{[i+1]}}(y) \\
&= \sum_{i=0}^{k-1} P^{\mathcal{M}^{Z_{[i]}}}(y \mid x) - P^{\mathcal{M}^{Z_{[i+1]}}}(y \mid x)\,.
\end{aligned}$$

Decomposing Exp-SE$_x(y)$

Target quantity to decompose: Exp-SE$_x(y) = P(y \mid x) - P(y_x)$

Using the definition of decomposition:

$$P(y \mid x) - P(y_x) = \sum_{z_1,z_2} P(y_x \mid x, z_1, z_2)[P(z_1, z_2 \mid x) - P(z_1)P(z_2 \mid z_1, x)]$$
$$+ \sum_{z_1,z_2} P(y_x \mid x, z_1, z_2)[P(z_1)P(z_2 \mid x, z_1) - P(z_1, z_2)]$$

Z1 contribution: Exp-SE$_x^{Z_1}(y)$

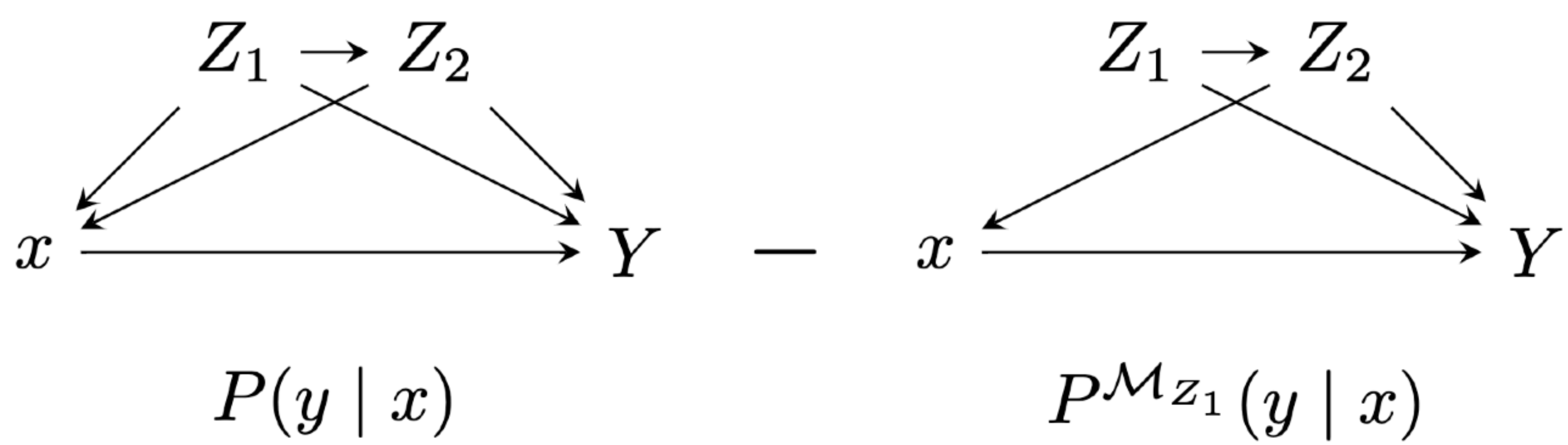
Z2 contribution: Exp-SE$_x^{Z_2}(y)$

$P(z_1, z_2 \mid x)$ -> assignment of X fully depends on Z (observational data)

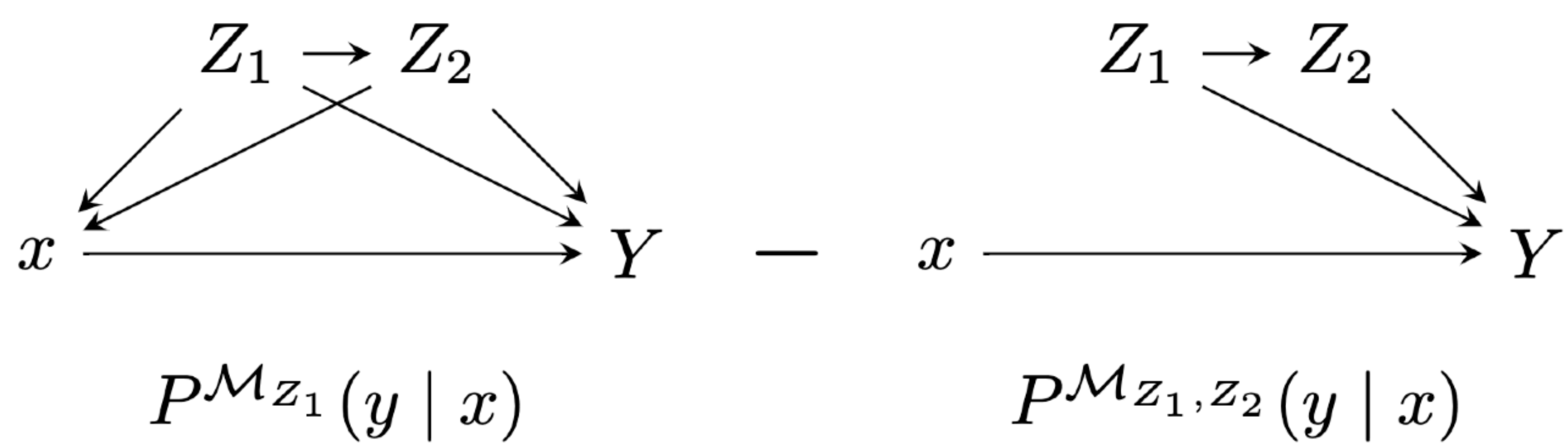
$P(z_1)P(z_2 \mid x, z_1)$ -> assignment of X fully depends on Z2, but the association of X and Z1 is the same as in an RCT (the between case)

$P(z_1, z_2)$ -> assignment of X does not depend on Z (classical RCT)

Decomposing Exp-SE$_x(y)$



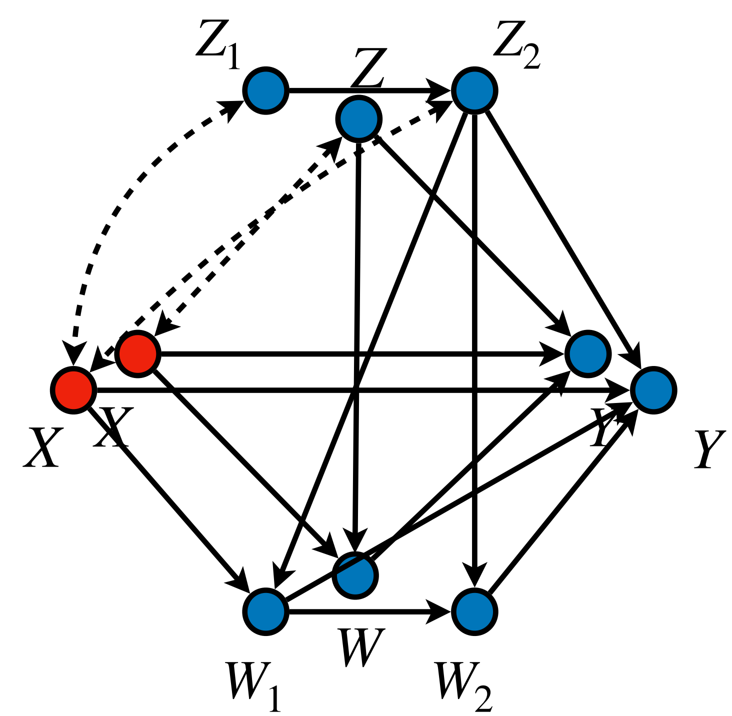
(a) $\text{Exp-SE}_x^{\emptyset, Z_1}(y)$.



(b) $\text{Exp-SE}_x^{Z_1, \{Z_1, Z_2\}}(y)$.

Towards latent decompositions

- We managed to decompose the spurious effect by attributing the variations to observable $Z_1, \ldots, Z_k$.
- When expanding the SFM, however, we might have bidirected confounding arrows - can we extend our approach?
- What is the best starting point?



Look at attribution of variations to $U_1, \ldots, U_k$ in the Markovian case

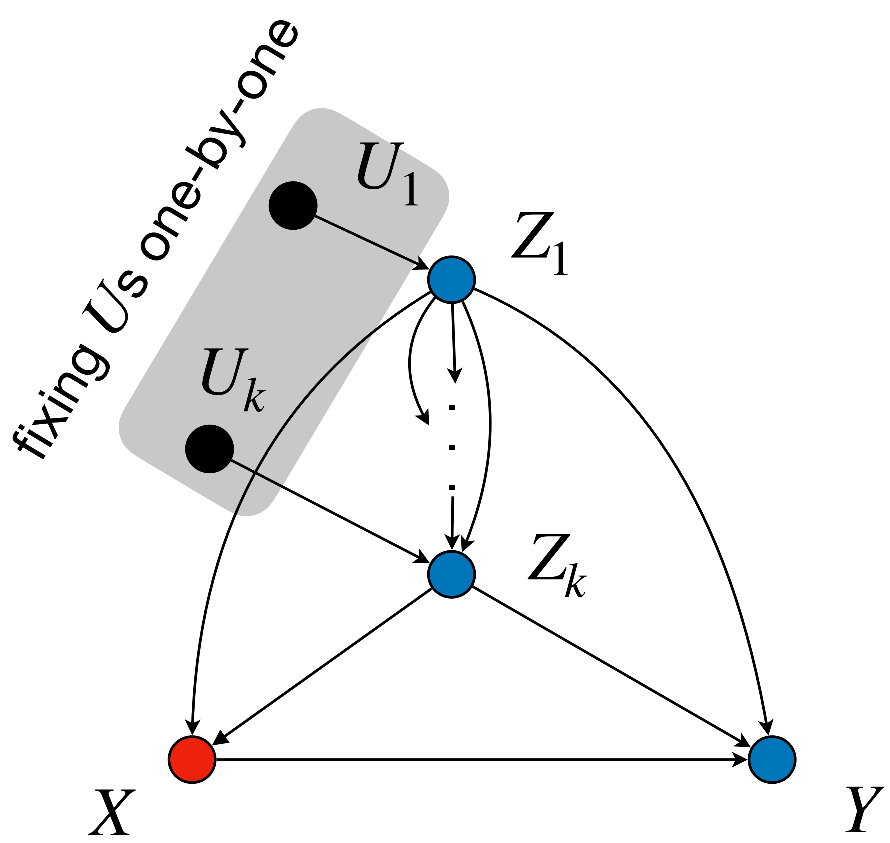
Exogenous Integrated Submodel

Definition. Let $\mathscr{M}$ be an SCM. Let $U_Z \subseteq U$ be a subset of the exogenous variables. Define by $\mathscr{M}^{U_Z}$ the following SCM

$$\mathscr{M}^{U_Z} = \sum_{u_Z} P^{\mathscr{M}}(U_Z = u_Z)\mathscr{M}_{U_Z=u_Z}.$$

That is, in $\mathscr{M}^{U_Z}$ the exogenous variables U_Z are determined from the distribution $P(U)$ of the SCM, after which the submodel $\mathscr{M}_{U_Z=u_Z}$ is used to obtain the all the observables V.

Exogenous Integrated Submodel



U_Z empty $\Longrightarrow X, Y$ associated as in the observational $P(V)$

U_Z neither empty nor full $\Longrightarrow$ X, Y associated by some, but

U_Z of all $Z \Longrightarrow X, Y$ independent like in a randomized control trial

Spurious Decomposition (Exogenous)

Theorem. Let $U_1, \dots, U_k$ be the subset of exogenous variables that lie on top of a spurious trek between X and Y. Let $U_{[i]}$ denote the variables $U_1, \dots, U_i$ ($U_{[0]}$ denotes the empty set $\emptyset$). Then, using the term

$$\textit{Exp-SE}_x^{A,B}(y) = P^{\mathcal{M}^A}(y \mid x) - P^{\mathcal{M}^B}(y \mid x),$$

we can decompose the experimental spurious effect as follows:

$$\begin{aligned}\textit{Exp-SE}_x(y) &= P(y \mid x) - P(y_x) \\ &= \sum_{i=0}^{k-1} \textit{Exp-SE}_x^{U_{[i]},U_{[i+1]}}(y) \\ &= \sum_{i=0}^{k-1} P^{\mathcal{M}^{U_{[i]}}}(y \mid x) - P^{\mathcal{M}^{U_{[i+1]}}}(y \mid x) \,.\end{aligned}$$

Spurious Decomposition Equivalence

Theorem. Let $Z_1, \ldots, Z_k$ be the confounders between variables X and Y, sorted in any valid topological ordering. Denote the exogenous variables corresponding to $Z_1, \ldots, Z_k$ as $U_1, \ldots, U_k$, respectively. Let $Z_{[i]} = \{Z_1, \ldots, Z_i\}$ and $U_{[i]} = \{U_1, \ldots, U_i\}$. It then holds that

$$P^{\mathcal{M}^{Z_{[i]}}}(V) = P^{\mathcal{M}^{U_{[i]}}}(V),$$

that is, the induced distributions over the observables V for the integrated submodel $\mathcal{M}^{Z_{[i]}}$ and the exogenous integrated submodel $\mathcal{M}^{U_{[i]}}$ are equal.

$\Longrightarrow$ we have an attribution with respect to latents that is equivalent (in Markovian, topological order case)

Spurious Decomposition Equivalence

Case 1: Fixing Z variables one by one

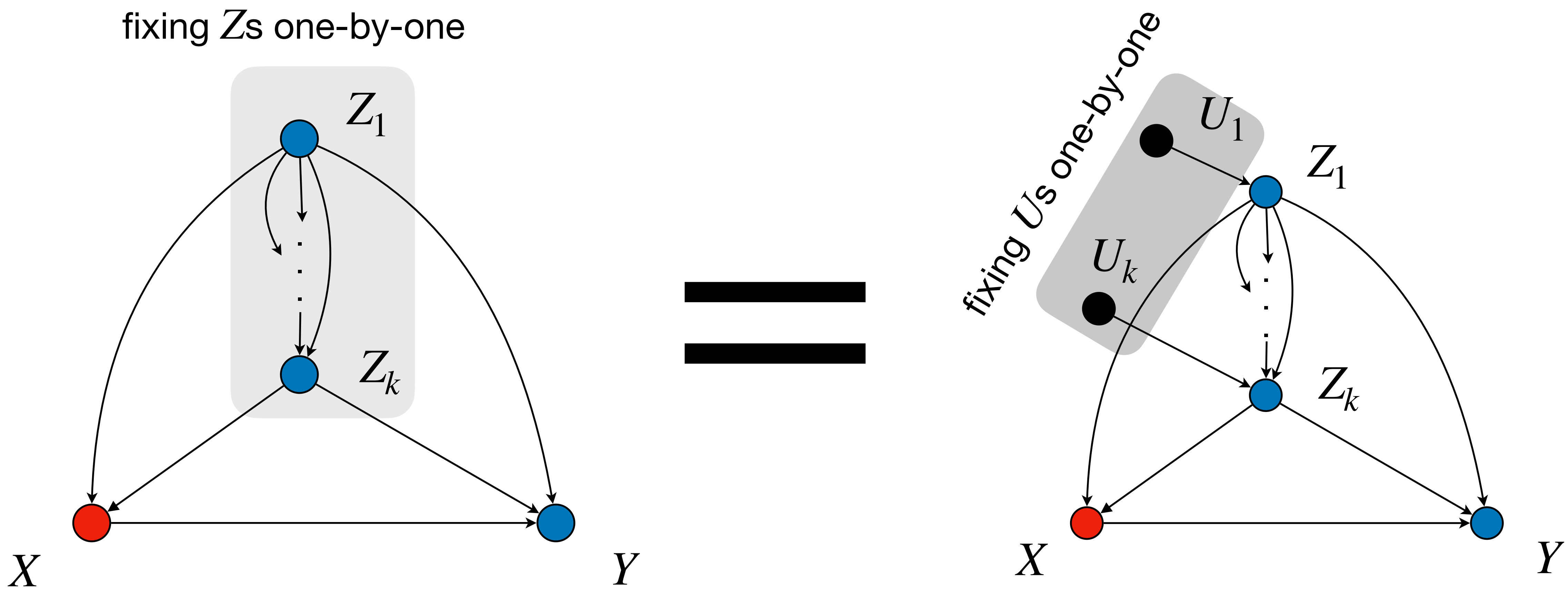
$$\textbf{Exp-SE}_x(y) = \underbrace{P^{\mathscr{M}^{Z[0]}}(y \mid x) - P^{\mathscr{M}^{Z[1]}}(y \mid x)}_{Z_1 \textbf{ contribution}} + \ldots + \underbrace{P^{\mathscr{M}^{Z[k-1]}}(y \mid x) - P^{\mathscr{M}^{Z[k]}}(y \mid x)}_{Z_k \textbf{ contribution}}.$$

Case 2: Fixing U variables one by one

same numbers!

$$\textbf{Exp-SE}_x(y) = \underbrace{P^{\mathscr{M}^{U[0]}}(y \mid x) - P^{\mathscr{M}^{U[1]}}(y \mid x)}_{U_1 \textbf{ contribution}} + \ldots + \underbrace{P^{\mathscr{M}^{U[k-1]}}(y \mid x) - P^{\mathscr{M}^{U[k]}}(y \mid x)}_{U_k \textbf{ contribution}}.$$

Spurious Decomposition Equivalence



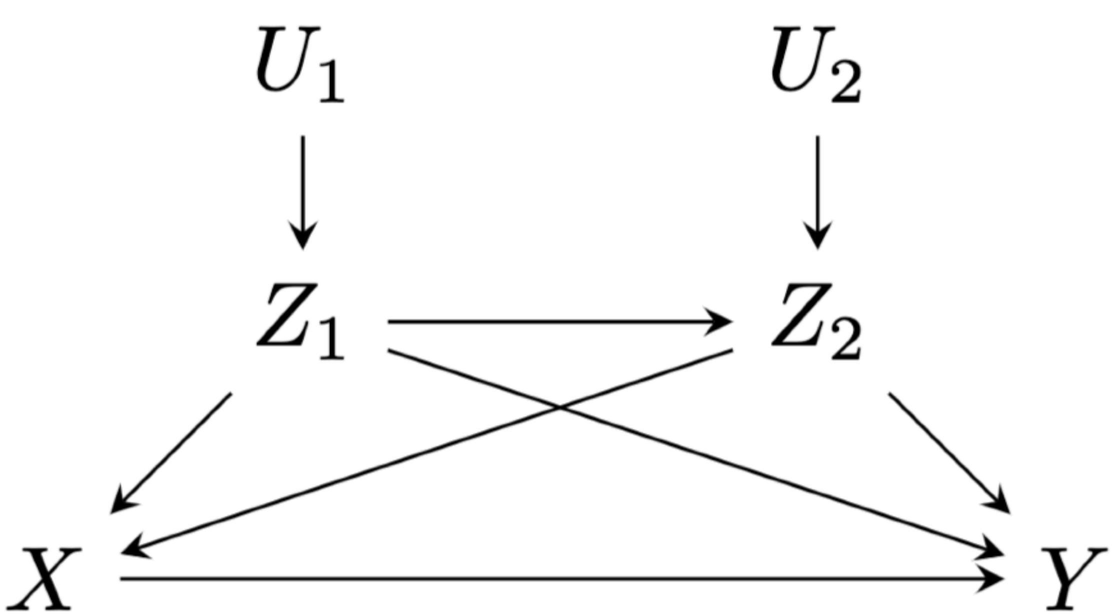
Can we use the same latent attribution approach to extend to Semi-Markovian models?

Note that we have a primitive that can attribute variations to the latent Us!

Semi-Markovian Models: Treks

Definition. Let $\mathcal{G}$ be the causal diagram of a Semi-Markovian model.
A trek τ from X to Y is an ordered pair of causal paths (g_l, g_r) with a common exogenous source $U_i \in U$. That is, g_l is a causal path $U_i \rightarrow \ldots \rightarrow X$ and g_r is a causal path $U_i \rightarrow \ldots \rightarrow Y$.

The common source U_i is called the top of the trek (ToT), denoted top(g_l, g_r). A trek is called spurious if g_r is a causal path from U_i to Y, i.e., not intercepted by X.



Spurious Treks:

$X \leftarrow Z_1 \leftarrow U_1 \rightarrow Z_1 \rightarrow Y$ **with top** U_1

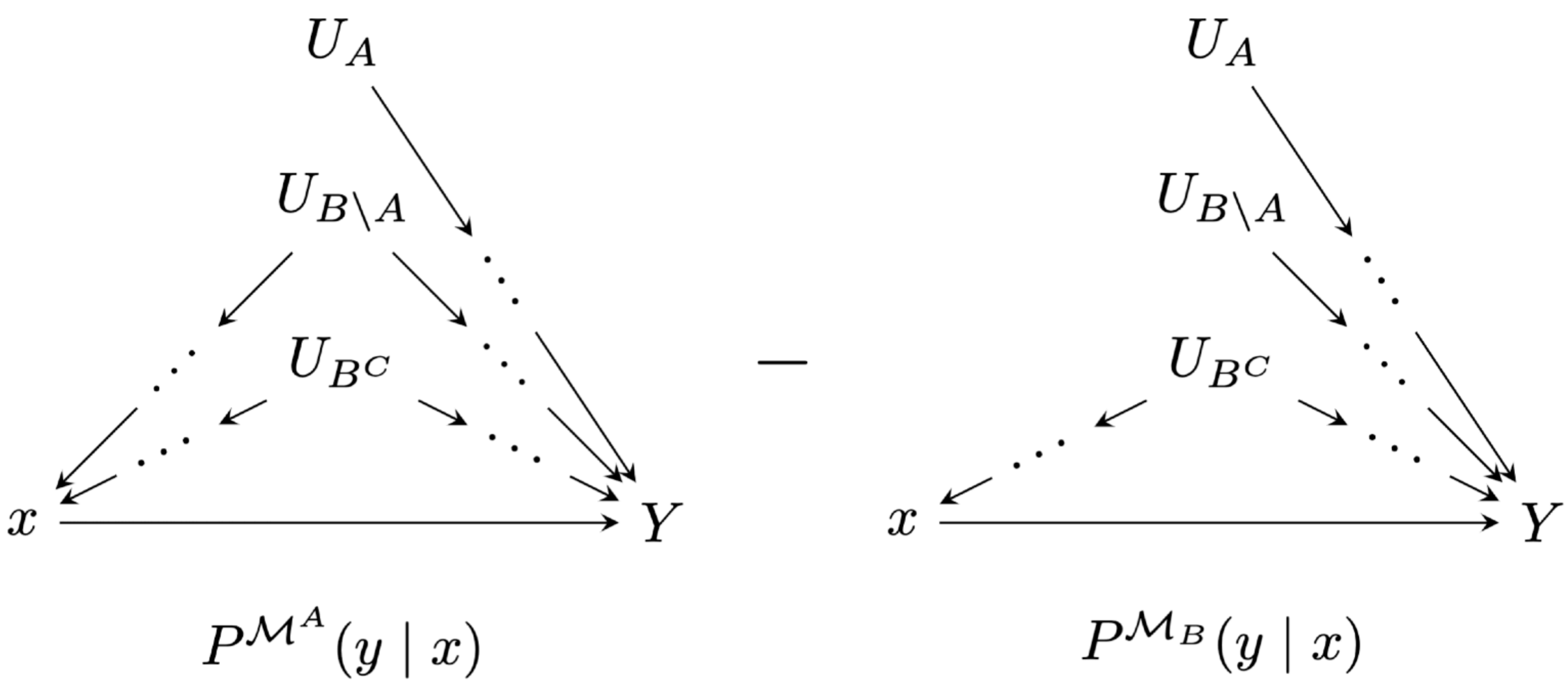
$X \leftarrow Z_2 \leftarrow U_2 \rightarrow Z_2 \rightarrow Y$ **with top** U_2

$X \leftarrow Z_1 \leftarrow U_1 \rightarrow Z_1 \rightarrow Z_2 \rightarrow Y$ **with top** U_1

Exogenous Set-Specific Effects

Definition. Let $U_{sToT} \subseteq U$ be the set of trek tops. Suppose $A \subseteq B \subseteq U_{sToT}$. The exogenous experimental spurious effect is defined as

$$\text{Exp-SE}_x^{A,B}(y) = P^{\mathcal{M}^A}(y \mid x) - P^{\mathcal{M}^B}(y \mid x)\,.$$



Admissibility with respect to Structural Fairness Measures

Lemma. Let $U_{BN} \subseteq U$ be a subset of the exogenous confounders of X, Y that fall under business necessity. Let U_{BN}^C denote the exogenous ancestors of X that do not fall under business necessity, that is $U_{BN}^C = \text{an}^{\text{ex}}(X) \backslash U_{BN}$. Then the measures Exp-SE$_x^{\emptyset, U_{BN}^C}(y)$, Exp-SE$_x^{U_{BN}, U}(y)$ are admissible with respect to the structural criterion Str-SE$(U_{BN})_X(Y)$, that is

$$(\textbf{Str-SE-BN}_X(Y) = 0) \implies (\textbf{Exp-SE}_x^{\emptyset, U_{BN}^C}(y) = 0)$$

$$(\textbf{Str-SE-BN}_X(Y) = 0) \implies (\textbf{Exp-SE}_x^{U_{BN}, U}(y) = 0)\,.$$

Since they are admissible, we will be able to add them to the Fairness Map (TBD)

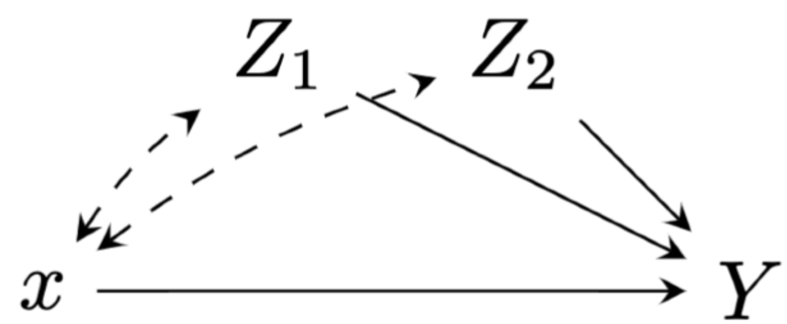
Semi-Markovian Spurious Decomposition

Theorem. Let $U_1, \ldots, U_k$ be the subset of exogenous variables that lie on top of a spurious trek between X and Y. Let $U_{[i]}$ denote the variables $U_1, \ldots, U_i$ ($U_{[0]}$ denotes the empty set $\emptyset$). The experimental spurious effect can be decomposed as follows:

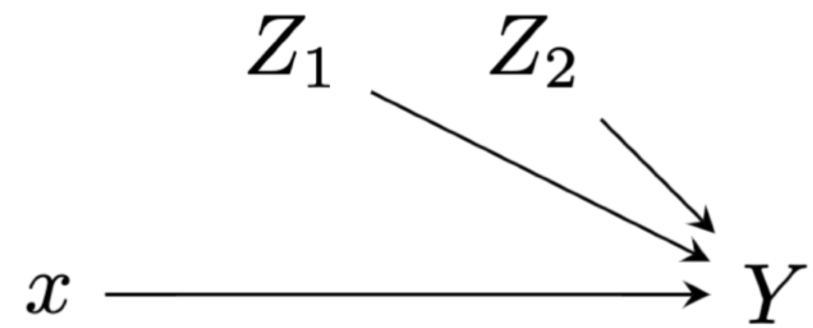
$$
\begin{aligned}
\textit{Exp-SE}_x(y) &= P(y \mid x) - P(y_x) \\
&= \sum_{i=0}^{k-1} \textit{Exp-SE}_x^{U_{[i]}, U_{[i+1]}}(y) \\
&= \sum_{i=0}^{k-1} P^{\mathscr{M}^{U_{[i]}}}(y \mid x) - P^{\mathscr{M}^{U_{[i+1]}}}(y \mid x) \,.
\end{aligned}
$$

Semi-Markovian Spurious Decomposition

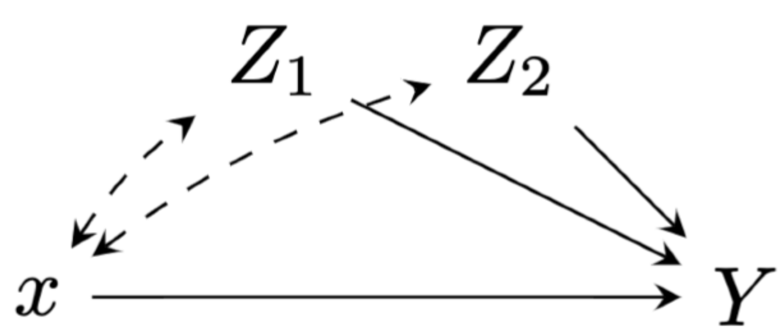
Exp-SE$_x(y) =$



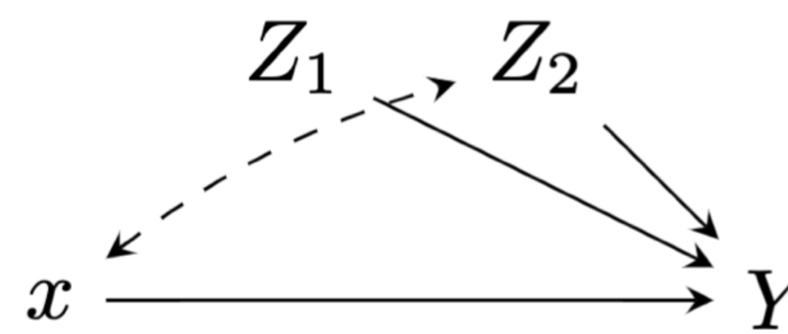
$-$



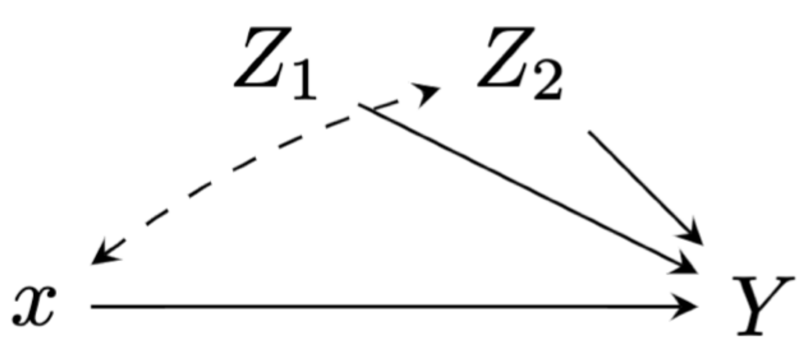
$=$



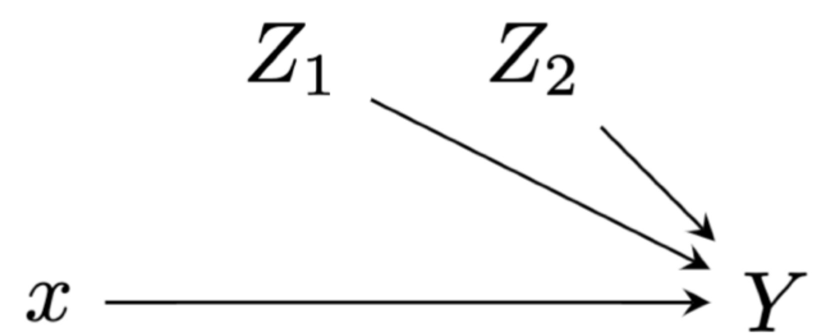
$-$



$+$



$-$



Identification of Spurious

Definition (Anchor Set). $\textbf{AS}(U_1, \ldots, U_l) = \bigcup_{i=1}^{l} \textbf{ch}(U_i) \setminus X\,.$ observables "touched" by U

Definition (Precedence Relation).

U_i topologically before U_j

$$U_i \overset{PR}{\leq} U_j \iff \textbf{AS}(U_j) \cap \{\textbf{AS}(U_i) \cup \textbf{an}(\textbf{AS}(U_i))\} \neq \emptyset\,.$$

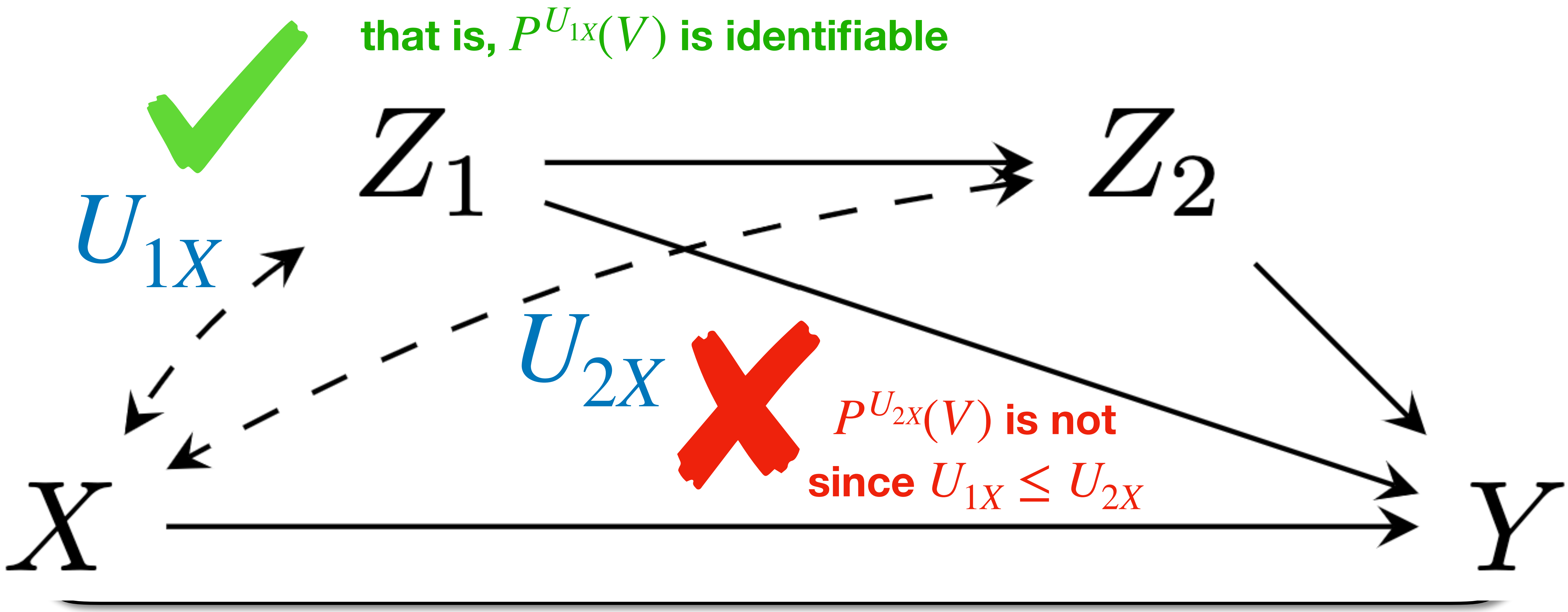
Theorem **(ID of Spurious Effects).** $P^{\mathcal{M}^A}(y \mid x)$ **is identifiable from observational data** $P(V)$ **if the following hold:**

(i) $Y \notin \textbf{AS}(A)$ Y not touched

(ii) $\textbf{AS}(A) \cap \textbf{AS}(U_{sToT} \setminus A) = \emptyset$ touched observables disjoint

(iii) there is no $U_j \in U_{sToT} \setminus A$ **such that** $\exists\, U_i \in A$ **for which** $U_j \overset{PR}{\leq} U_i$**.**

no precedence between set elements



Theorem **(ID of Spurious Effects).** $P^{\mathcal{M}^A}(y \mid x)$ **is identifiable from observational data** $P(V)$ **if the following hold:**

(i) $Y \notin$ **AS**(A) Y **not touched**

(ii) AS$(A) \cap$ **AS**$(U_{sToT} \backslash A) = \varnothing$ **touched observables disjoint**

(iii) there is no $U_j \in U_{sToT} \backslash A$ **such that** $\exists\, U_i \in A$ **for which** $U_j \overset{PR}{\leq} U_i$.

no precedence between set elements

x-specific spurious?

- Target: Ctf-SE$_{x_0,x_1}(y) = P(y_{x_0} \mid x_1) - P(y \mid x_0)$

Definition (Exogenous x-specific Integrated Submodel). Define by $\mathcal{M}_x^{U_Z}$ the following SCM:

$$\mathcal{M}_x^{U_Z} = \sum_{u_Z} P^{\mathcal{M}}(U_Z = u_Z \mid X = x)\mathcal{M}_{U_Z = u_Z}.$$

Definition (Exogenous x-specific spurious).

$$\text{Ctf-SE}^{A,B}_{x_0,x_1}(y) = P^{\mathcal{M}^A_{x_1}}(y \mid x_0) - P^{\mathcal{M}^B_{x_1}}(y \mid x_0).$$

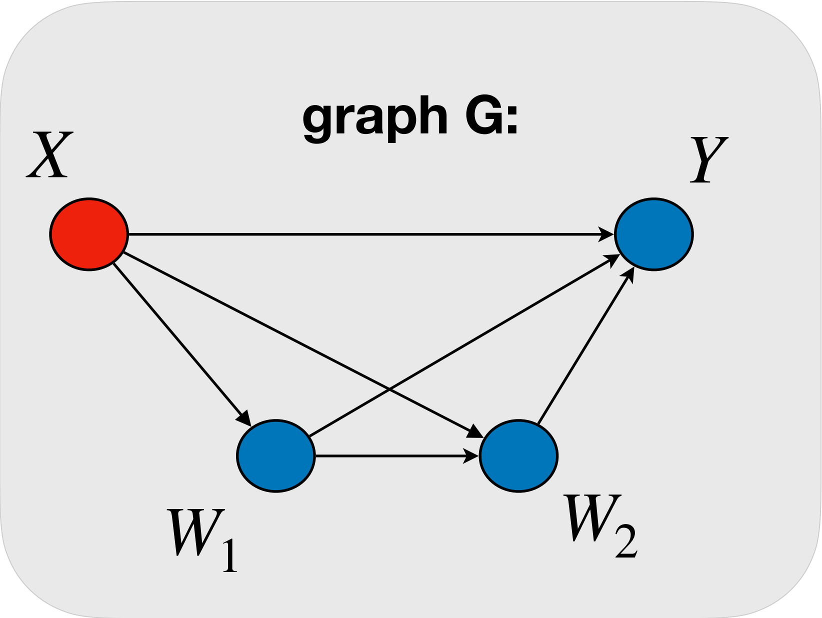
Theorem (x-specific exogenous spurious decomposition).

$$\text{Ctf-SE}_{x_0,x_1}(y) = \sum_{i=0}^{m-1} \text{Ctf-SE}^{U_{[i]},U_{[i+1]}}_{x_0,x_1}(y)$$

Refining Indirect Effects

- Target: refine the quantity NIE$_{x_0,x_1}(y)$
- What is our target in terms of Structural Fairness?

$$\text{Str-IE-BN}_X(Y) = \mathbb{1}(\text{an}(Y) \cap \text{ch}(X) \cap W^C_{BN} = \emptyset)\,.$$



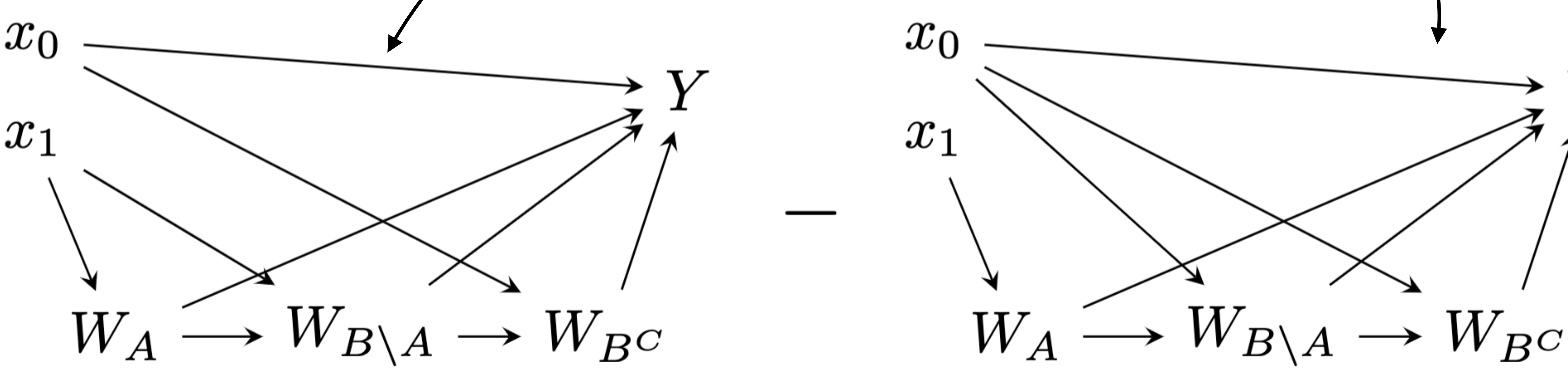
- How can we get a decomposition

$$\text{NIE}_{x_0,x_1}(y) = \text{NIE}^{W_1}_{x_0,x_1}(y) + \text{NIE}^{W_2}_{x_0,x_1}(y)\ ?$$

Set-specific indirect

Definition (Set-specific indirect effect). Let W_A, W_B be nested subsets of the mediators W, so that $W_A \subseteq W_B$. Let W_{A^C} and W_{B^C} denote the complements of W_A, W_B in W. We then define the E-specific indirect effect with respect to sets W_A, W_B as

$$E\text{-IE}_{x_0,x_1}^{W_A,W_B}(y) = P(y_{x_0,(W_B)_{x_1},(W_{B^C})_{x_0}} \mid E) - P(y_{x_0,(W_A)_{x_1},(W_{A^C})_{x_0}} \mid E) \,.$$



Admissibility with respect to Structural Measures

Lemma. Let $W_{BN} \subseteq W$ be a subset of the mediators that fall under business necessity. Then the measure $E\text{-IE}^{\varnothing, W_{BN}^C}_{x_0,x_1}(y)$ is admissible with respect to the structural criterion Str-IE$(W_{BN})_X(Y)$, that is

$$(\text{Str-IE-BN}_X(Y) = 0) \implies (E\text{-IE}^{\varnothing, W_{BN}^C}_{x_0,x_1}(y) = 0),$$

$$(\text{Str-IE-BN}_X(Y) = 0) \implies (E\text{-IE}^{W_{BN}, W}_{x_0,x_1}(y) = 0)\,.$$

Since they are admissible, we will be able to add them to the Fairness Map (TBC)

Decomposition of Indirect

Theorem. Let $W_1, \ldots, W_k$ denote the set of mediators, sorted in a topological order. Define $W_{[i]}$ as the set $\{W_1, \ldots, W_i\}$ and $W_{-[i]}$ as $\{W_{i+1}, \ldots, W_k\}$. The E-specific indirect effect can then be decomposed as

$$
\begin{aligned}
E\text{-}IE_{x_0,x_1}(y) &= P(y_{x_0,W_{x_1}} \mid E) - P(y_{x_0} \mid E) \\
&= \sum_{i=0}^{k-1} E\text{-}IE_{x_0,x_1}^{W_{[i]},W_{[i+1]}}(y) \\
&= \sum_{i=0}^{k-1} P(y_{x_0,(W_{[i+1]})_{x_1},(W_{-[i+1]})_{x_0}} \mid E) - P(y_{x_0,(W_{[i]})_{x_1},(W_{-[i]})_{x_0}} \mid E)\,.
\end{aligned}
$$

Lack of symmetry

- A lack of symmetry arises because we can consider either a $x_0 \to x_1$, or $x_1 \to x_0$ transition, and similarly for the BN transition. As a consequence, note that:

$$
\begin{aligned}
\text{Ctf-IE}_{x_0,x_1}(y \mid x) &= \underbrace{\text{Ctf-IE}^{\emptyset, W^C_{BN}}_{x_0,x_1}(y \mid x)}_{\text{discriminatory}} + \underbrace{\text{Ctf-IE}^{W^C_{BN}, W}_{x_0,x_1}(y \mid x)}_{\text{BN variations}} \\
&= \underbrace{\text{Ctf-IE}^{\emptyset, W_{BN}}_{x_0,x_1}(y \mid x)}_{\text{BN variations}} + \underbrace{\text{Ctf-IE}^{W_{BN}, W}_{x_0,x_1}(y \mid x)}_{\text{discriminatory}},
\end{aligned}
$$

and analogously for Ctf-IE$_{x_1,x_0}(y \mid x)$, and also for the spurious.

- How can we fix this problem?

- $\Longrightarrow$ Take an average over the transitions!

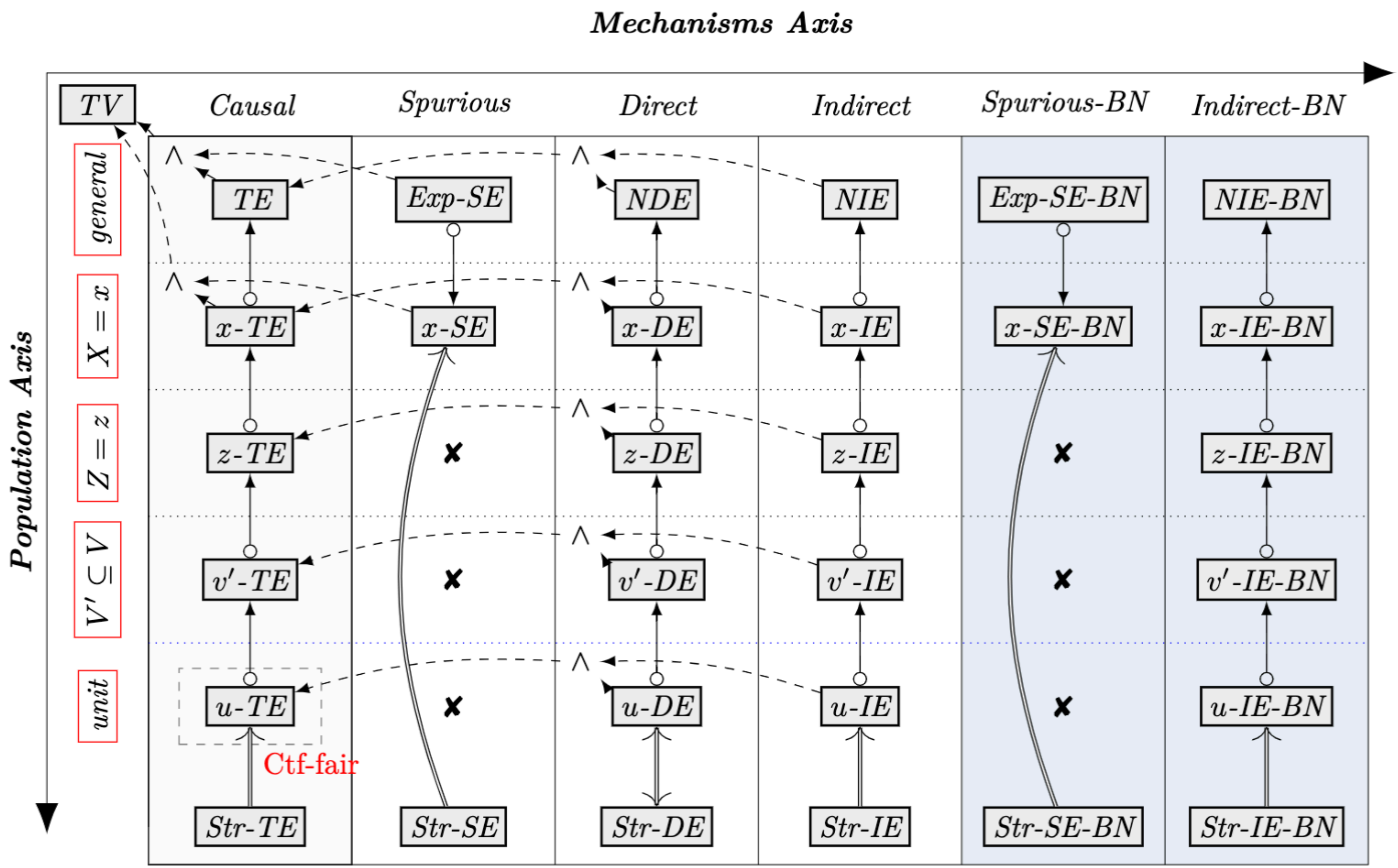
Lack of symmetry

Definition. Define the x-specific indirect and spurious measures under business necessity as

$$x\text{-IE}^{\text{sym-BN}}(y \mid x) = \frac{1}{4}\Big(\text{Ctf-IE}_{x_1,x_0}^{\emptyset, W_{BN}^C}(y \mid x) + \text{Ctf-IE}_{x_1,x_0}^{W_{BN}, W}(y \mid x) - \text{Ctf-IE}_{x_0,x_1}^{\emptyset, W_{BN}^C}(y \mid x) - \text{Ctf-IE}_{x_0,x_1}^{W_{BN}, W}(y \mid x)\Big)$$

$$x\text{-SE}^{\text{sym-BN}}(y) = \frac{1}{4}\Big(\text{Ctf-SE}_{x_1,x_0}^{\emptyset, U_{BN}^C}(y) + \text{Ctf-SE}_{x_1,x_0}^{U_{BN}, U}(y) - \text{Ctf-SE}_{x_0,x_1}^{\emptyset, U_{BN}^C}(y) - \text{Ctf-SE}_{x_0,x_1}^{U_{BN}, U}(y)\Big).$$

Extended Fairness Map



Task 1 (Extended)

Extended Fairness Cookbook

1) Obtain data on past decisions $\mathscr{D}$.

2) Determine the (possibly simplified) causal diagram $\mathscr{G}$ (w.r.t. underlying $\mathscr{M}^*$).

3) Determine the **Business Necessity** (BN) set (**now arbitrary!**).

4) Test the following two hypotheses:

$$H_0^{(\text{Ctf-SE}),\ \text{BN}} : \text{Ctf-SE}^{\text{sym-BN}}_{x_1,x_0}(y) = 0$$

rejected → evidence of disparate impact

not rejected → no evidence of disparate impact

$$H_0^{(\text{Ctf-IE}),\ \text{BN}} : \text{Ctf-IE}^{\text{sym-BN}}(y \mid x_0) = 0.$$

rejected → evidence of disparate impact

not rejected → no evidence of disparate impact

Task 2 (Extended)

Fairadapt: Sequential Optimal Transport

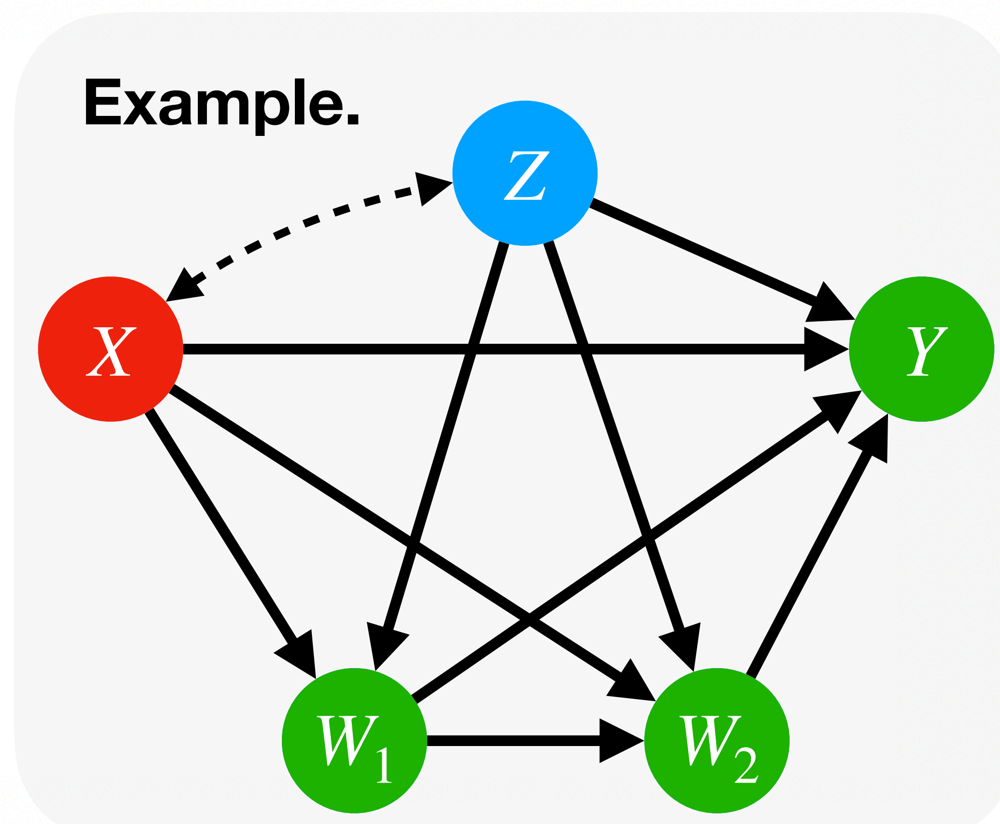
Plecko & Meinshausen, JMLR 2020

- joint optimal transport induces a dependency of W on Y, therefore *breaking the causal structure*
- instead, we perform the Optimal Transport sequentially

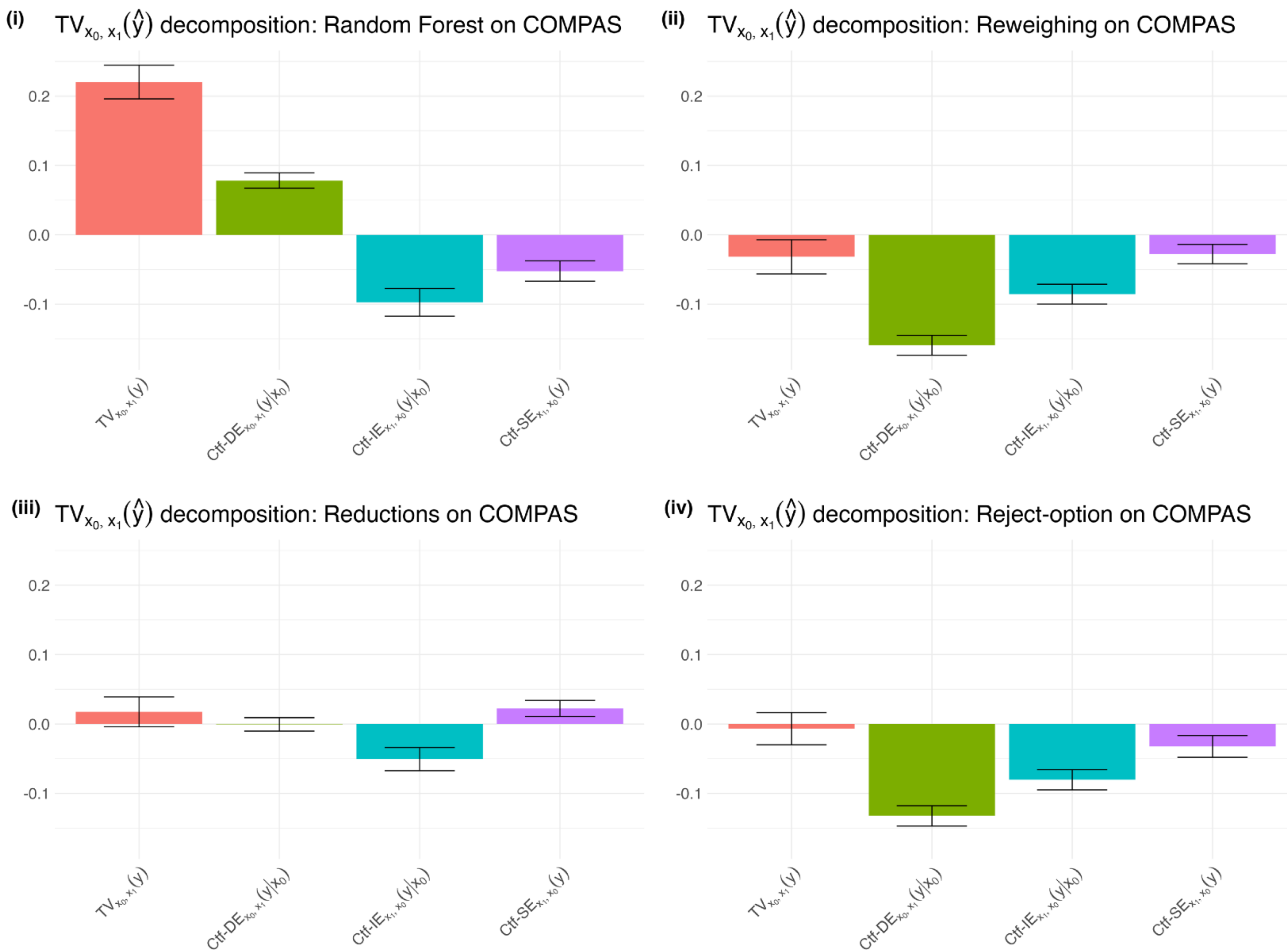
preserving relative achievement

Al
In
set

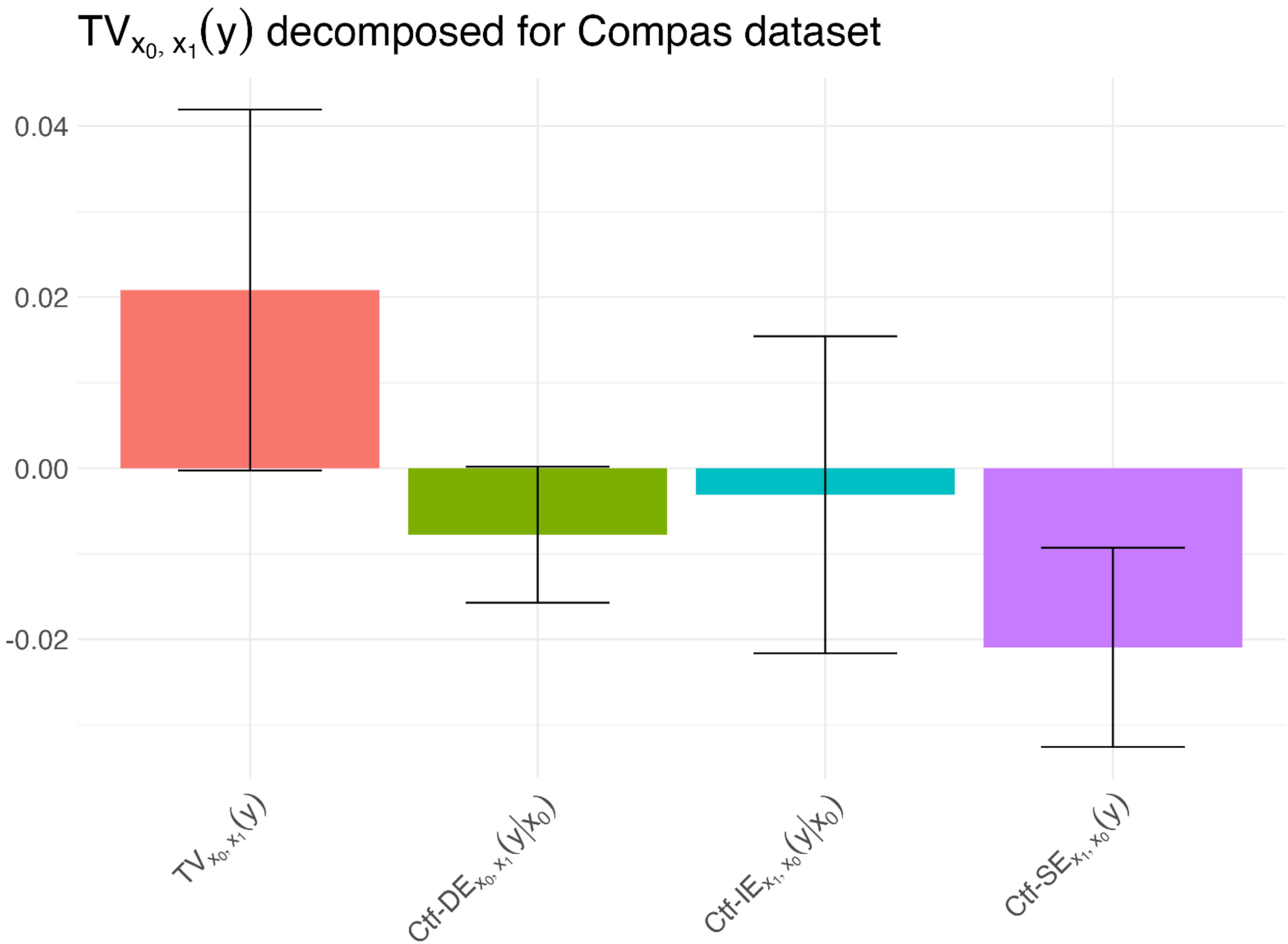
for $V_i \in \mathrm{de}(A)$ *in topological order* **do**
  learn function $V_i \leftarrow f_i(\mathrm{pa}(V_i), U_i)$
  infer quantiles U_i associated with the variable V_i
  transform values as $V_i^{(fp)} \leftarrow f_i(\mathrm{pa}(V_i)^{(fp)}, U_i)$
end
return $V^{(fp)}$



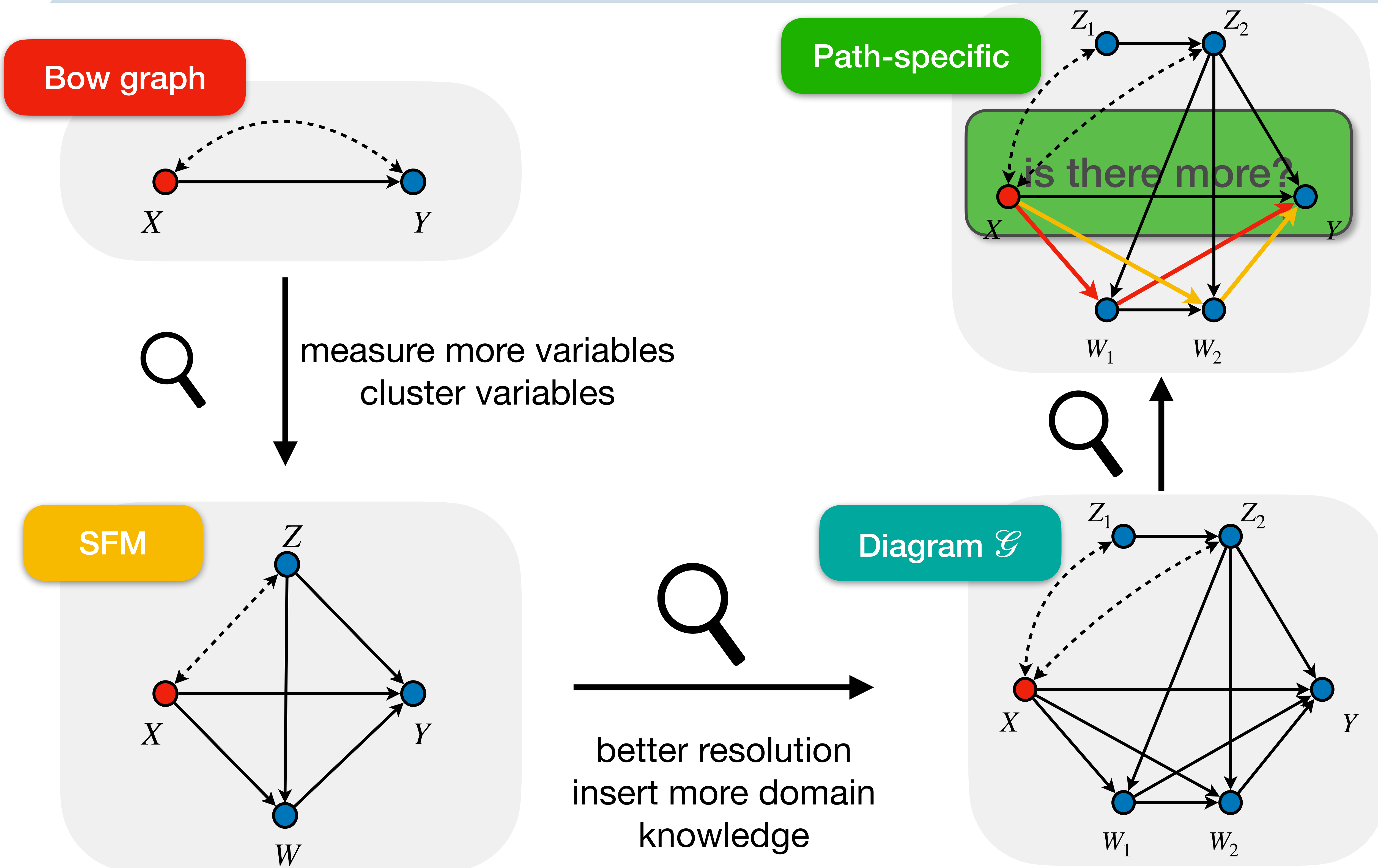
Recap: Fair Prediction Theorem on COMPAS



Fairadapt: Result on COMPAS



Complexity Cascade



Lectures' Recap - L1

- Foundations of Causal Inference
- Fairness Examples & the SFM
- FPCFA
- Legal Doctrines of Discrimination
- Structural Fairness Criteria / Doctrines
- Decomposing Variations
- Admissibility & Power
- Explainability Plane

Lectures' Recap - L2

- TV family of measures
- Power in practice
- Unit-level measures
- Towards x, z, y-specific
- TV family as contrasts
- Fairness Map
- Decomposability, Admissibility and Power in the Map

Lectures' Recap - L3

Corollaries of Fairness Map

Identification & Estimation

Understanding previous literature through the Map

Counterfactual Fairness
Individual Fairness
Predictive Parity

Lectures' Recap - L4 & 5

Task 1: Bias Quantification

Task 2: Fair Prediction

Fairness Cookbook

Biased Reality -> Biased Data -> Biased Future?

Quantification over time

Pre-, In-, Post- processing

Quantification with $Y, \widehat{Y}$

Fair Prediction Theorem

Lectures' Recap - L4 & 5

Task 3: Fair Decision-Making

Principal Fairness & Benefit Fairness

Chaining Predictions to Decisions Fails

Canonical Types

Different types of utility

Decomposing the Gap

Outcome Control Task

Task 3 fully blown version

Lectures’ Recap - L6

- Beyond the SFM
- Decomposing Indirect Effects
- Decomposing spurious effects: Integrated Submodels
- Admissibility with respect to Structural Fairness
- Integrated Submodels for Semi-Markovian models
- Extended Fairness Map
- Identifiability
- Fair Data Adaptation