Causal Inference: **Fundamentals of Partial Identification**

YES Workshop 2023
Eindhoven, 15th March 2023

Jakob Zeitler,
PhD Candidate in Foundational Artificial Intelligence
Centre for Doctoral Training in Foundational AI at UCL
mail@jakob-zeitler.de

The basis for this talk

The Causal Marginal Polytope for Bounding Treatment Effects
Jakob Zeitler, Ricardo Silva
https://arxiv.org/abs/2202.13851

Stochastic Causal Programming for Bounding Treatment Effects
Kirtan Padh, Jakob Zeitler, David Watson, Matt Kusner, Ricardo Silva, Niki Kilbertus
https://arxiv.org/abs/2202.10806

Learn more @ CLeaR (Causal Learning and Reasoning) 2023!

Literature Review: https://tinyurl.com/partial-identification

You have seen this before this week

Mats Stensrud, Tuesday, 9:30am

Slide 30 onwards

Example on Vitamin A supplementation in northern Sumatra

	$Z=0$		$Z=1$	
	$Y=1$	$Y=0$	$Y=1$	$Y=0$
$A=0$	0.0064	0.9936	0.0028	0.1972
$A=1$	0.0000	0.0000	0.0010	0.7990

$$-0.0054 \leq \mathbb{E}(Y^{a=1} - Y^{a=0}) \leq 0.1946$$

Consider now the effect among those who intend to be treated:

$$\mathbb{E}(Y^{a=1} - Y^{a=0} \mid A=1) = -0.0032.$$

and among those who intend to be untreated

$$-0.007 \leq \mathbb{E}(Y^{a=1} - Y^{a=0} \mid A=0) \leq 0.331 \; .$$

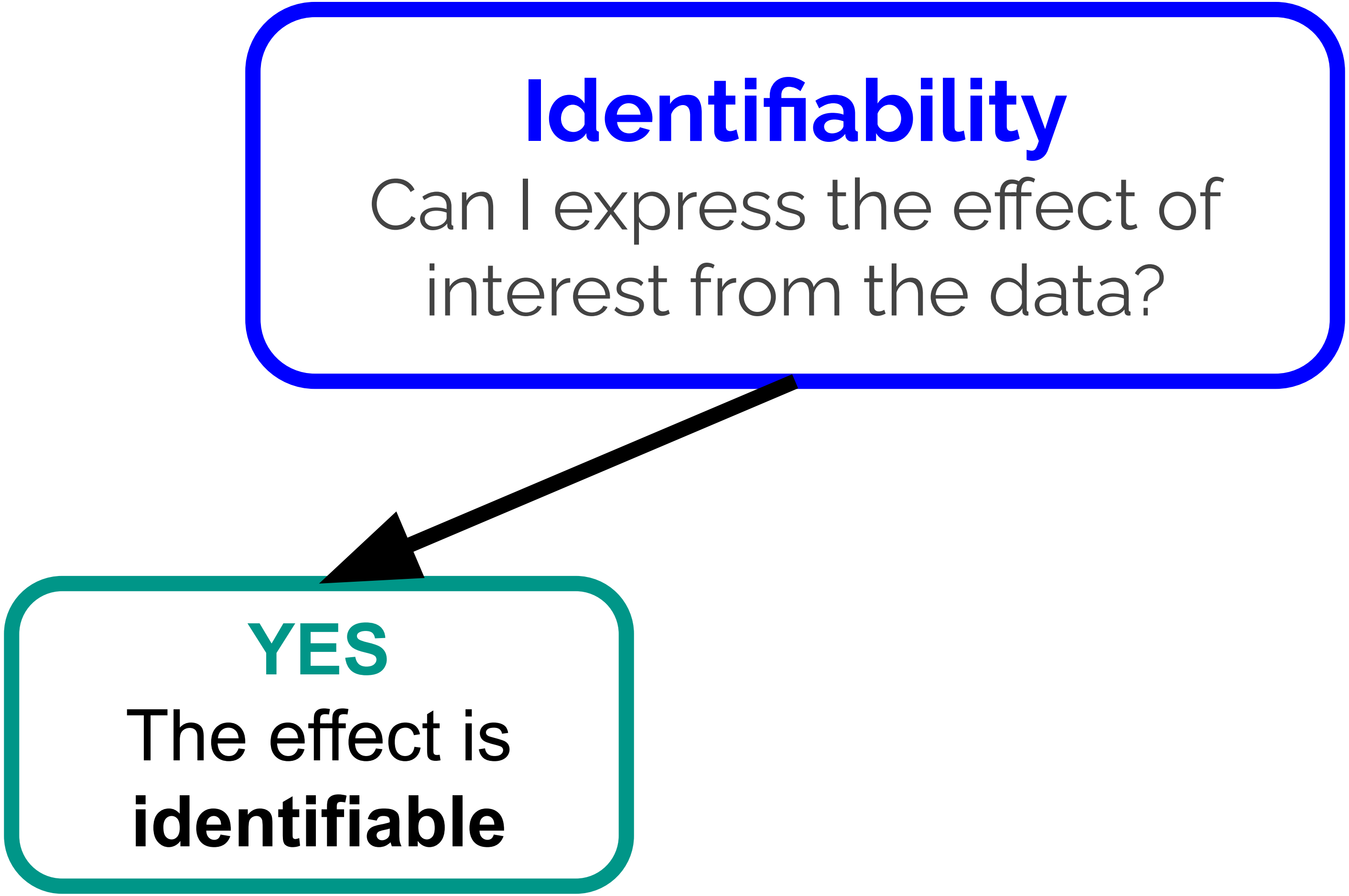
Program

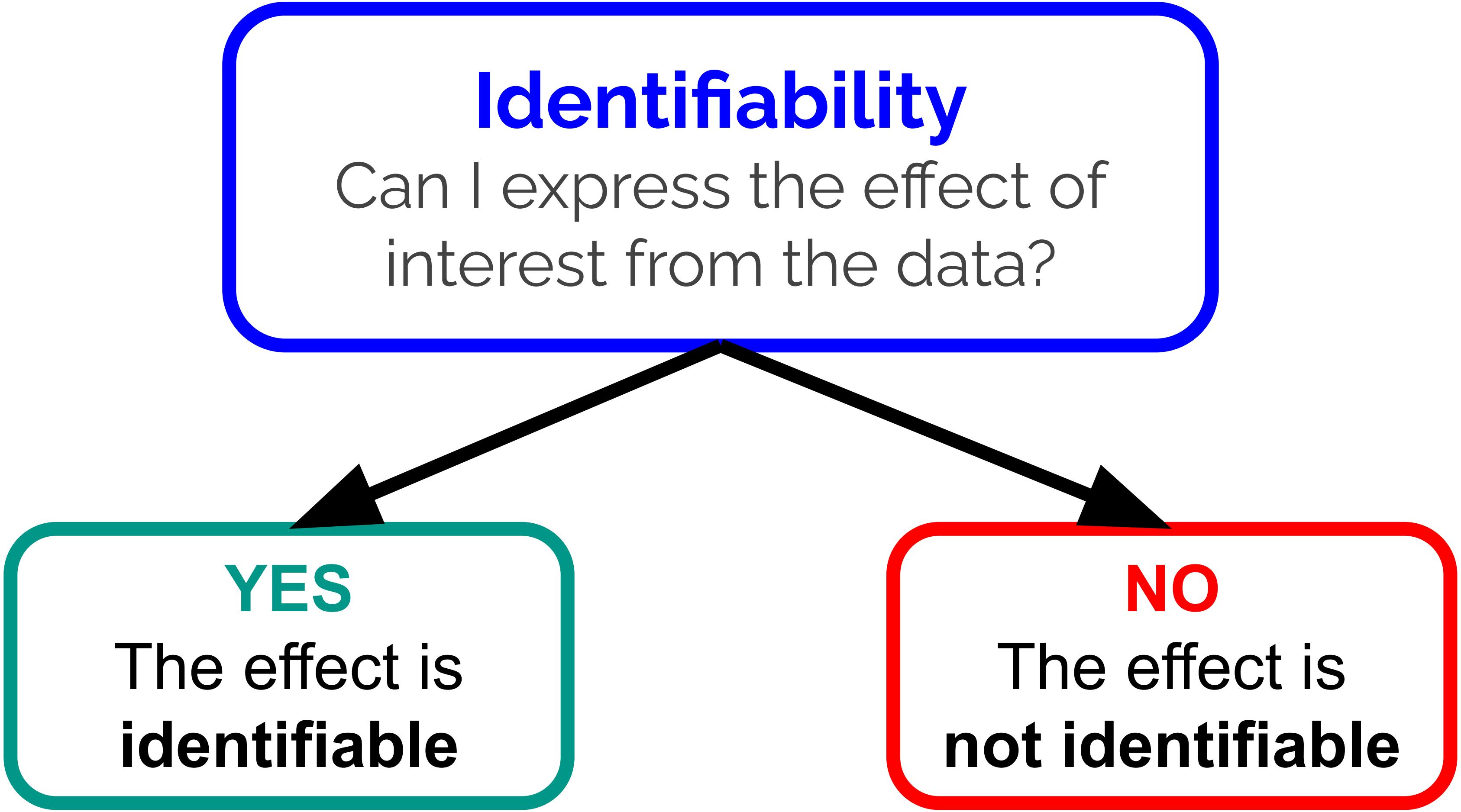
1. **Causal Inference**
2. **Partial Identification**
    a. Problem
    b. Challenges
    c. Solutions
3. **Q&A**

Causal Inference: **Identifiability**

Identifiability

Can I express the effect of interest from the data?


Identifiability
Can I express the effect of interest from the data?
YES
The effect is identifiable



Identifiability
Can I express the effect of interest from the data?
YES
The effect is identifiable
NO
The effect is not identifiable

We usually ask:

Is there confounding?

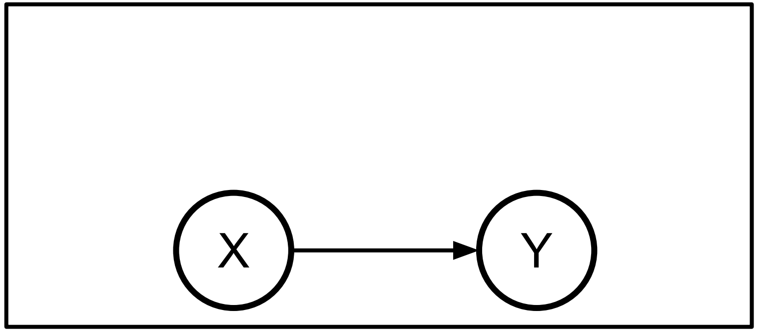
If yes, apply **“Back-Door Adjustment”**

Example: **Confounding backdoor adjustment**

If we want to adjust for confounding, we need to identify the confounders!

Scenario 1:

"No Confounding"

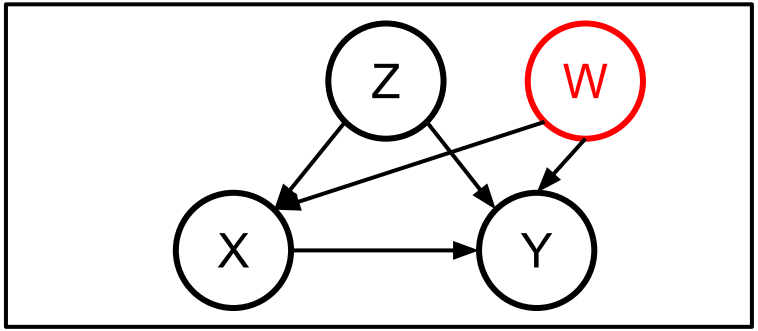


Average Treatment Effect (ATE):

$$P(Y = y | do(X = x))$$

Scenario 2:

"Confounding" Problem?



Average Treatment Effect (ATE):

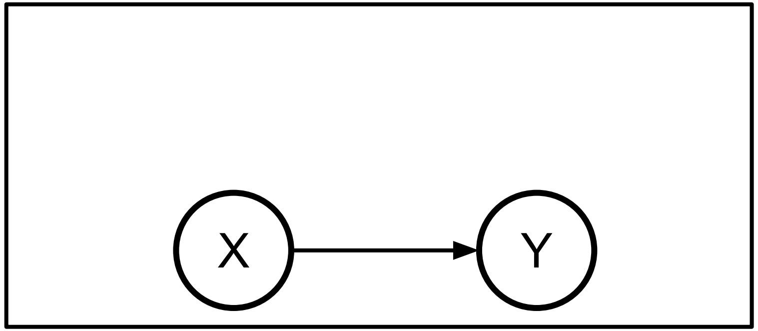
$$P(Y = y | do(X = x)) = \sum_z P(Y = y | X = x, Z = z) P(Z = z)$$

Example: **Confounding backdoor adjustment**

If we want to adjust for confounding, we need to identify the confounders!

Scenario 1:

“No Confounding”

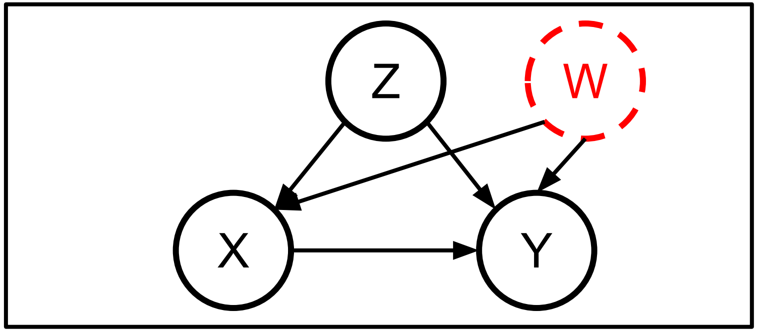


Average Treatment Effect (ATE):

$$P(Y = y|do(X = x))$$

Scenario 2:

“Confounding” Problem?



Average Treatment Effect (ATE):

$$P(Y = y|do(X = x)) = \sum_z P(Y = y|X = x, Z = z)P(Z = z)$$

Example: **Confounding backdoor adjustment**

If we w… …ed to identify the confounders!

Sc…

su…

Problem!

Assumption:
"No unmeasured confounding"

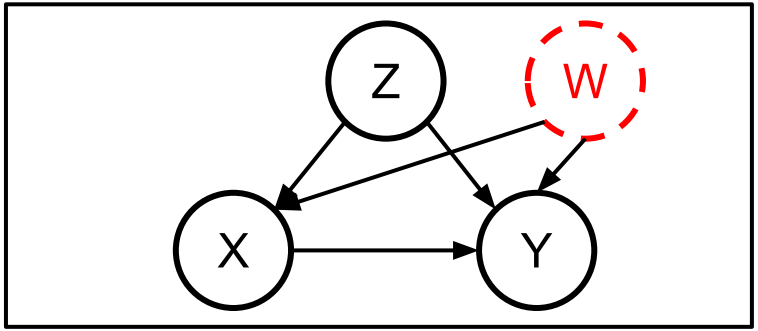
Violated?: Yes!
Because we don't observe **W, and cannot adjust for it.**

Consequence:
Biased effect estimate

Avera…

Scenario 2:

"Confounding" *Problem!*



Average Treatment Effect (ATE):

$$P(Y = y|do(X = x)) = \sum_z P(Y = y|X = x, Z = z)P(Z = z)$$

Assumptions are *essential* for Identification

Causal Inference is

a language for encoding your causal assumptions

It's all about the **assumptions** you make:

"No Causes in, No Causes out"

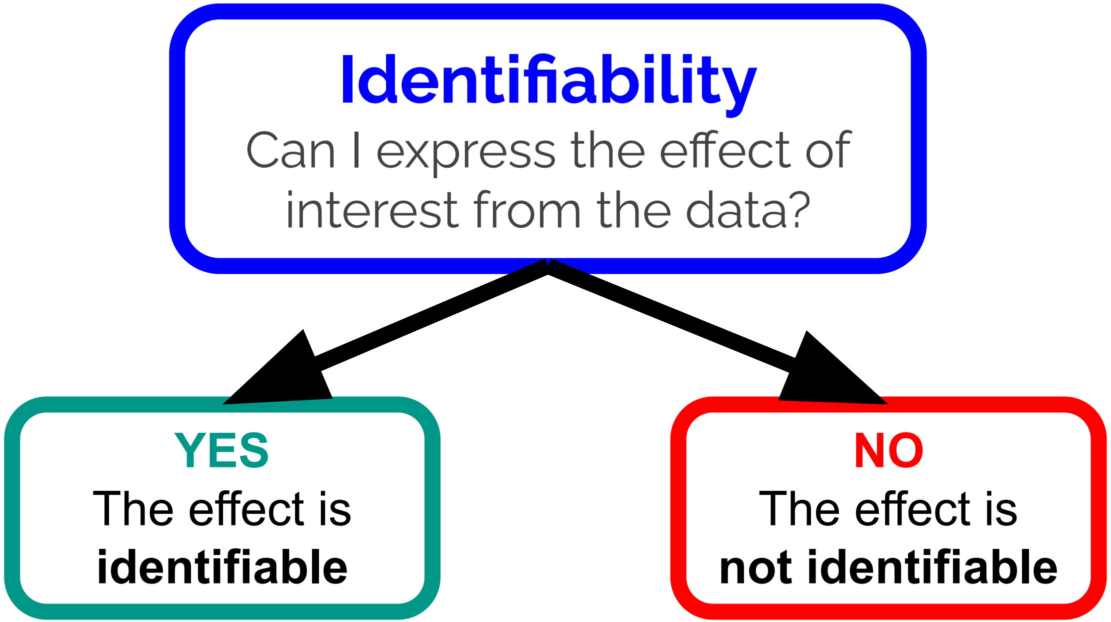
No assumptions, no identification

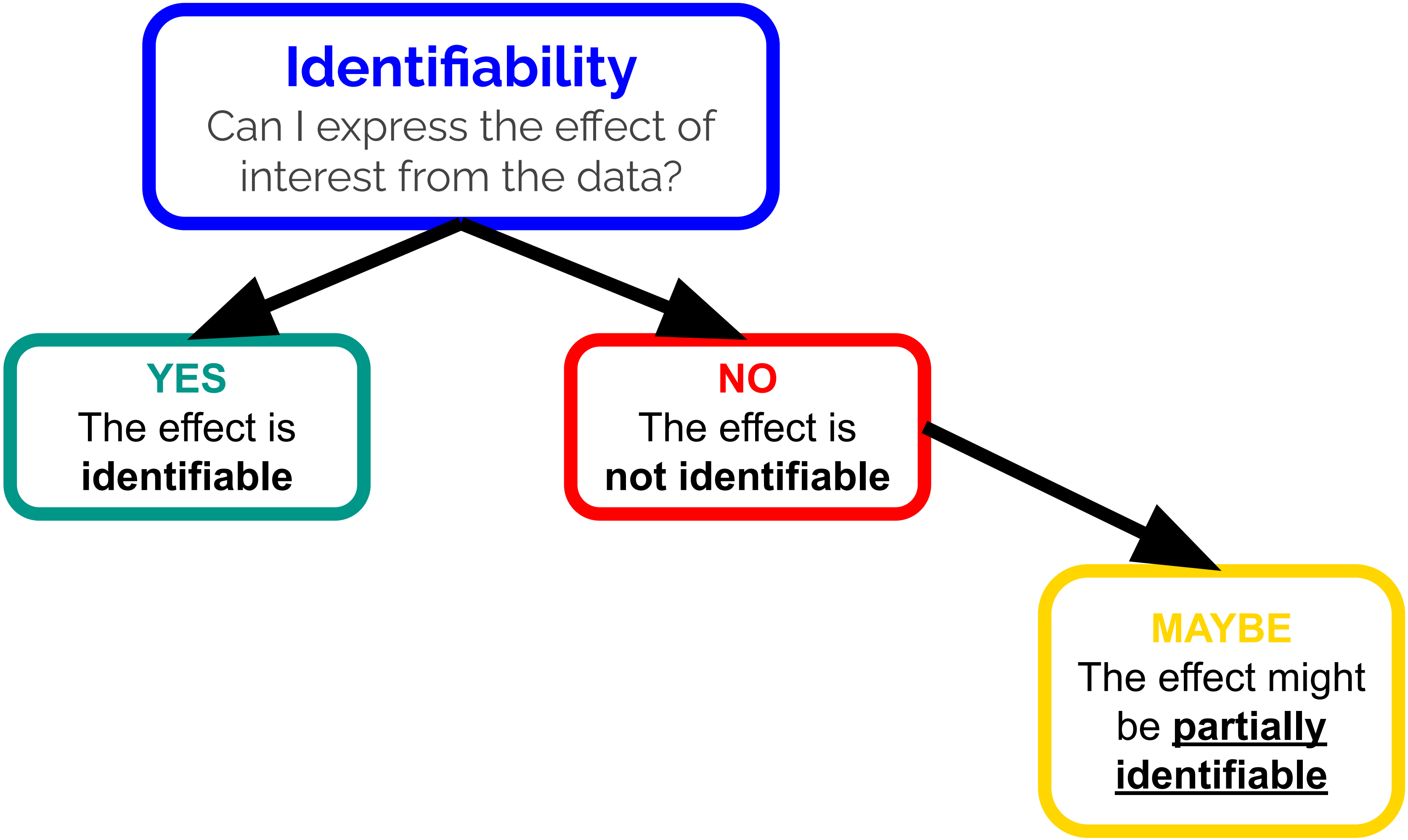
- Nancy Cartwright,

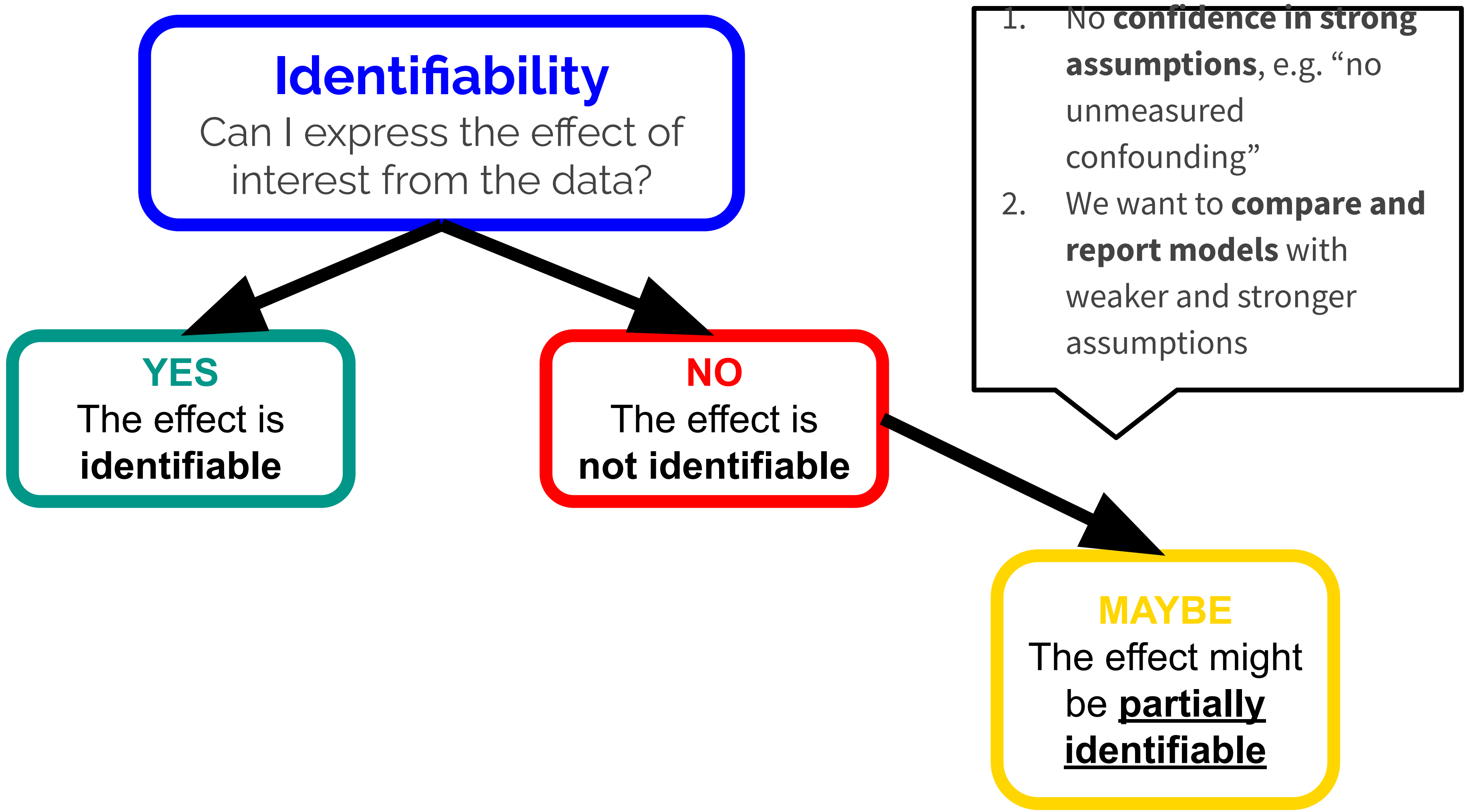
https://doi.org/10.1093/0198235070.003.0003

Partial Identification: **Why**

*Why is partial identification **relevant?***


Identifiability
Can I express the effect of interest from the data?
YES
The effect is identifiable
NO
The effect is not identifiable


Identifiability
Can I express the effect of interest from the data?
YES
The effect is identifiable
NO
The effect is not identifiable
MAYBE
The effect might be partially identifiable


Identifiability
Can I express the effect of interest from the data?
YES
The effect is identifiable
NO
The effect is not identifiable
MAYBE
The effect might be partially identifiable
1. No confidence in strong assumptions, e.g. "no unmeasured confounding"
2. We want to compare and report models with weaker and stronger assumptions

Bounds on effects of interest

Instead of doing our usual ***backdoor identification strategy*** of …

$$P(Y = y | do(X = x)) = \sum_z P(Y = y | X = x, Z = z) P(Z = z) \quad = 0.34$$

Bounds on effects of interest

Instead of doing our usual ***backdoor identification strategy*** of …

$$P(Y = y|do(X = x)) = \sum_z P(Y = y|X = x, Z = z)P(Z = z) \quad = 0.34$$

… we try to calculate **Bounds** on the effect of interest:

i.e. **Lower Bound < ATE(X->Y) < Upper Bound**

e.g $0.1 < P(Y = y|do(X = x)) < 0.4$

Let's review

So far, we have learned:

1. Causal Inference is all about **assumptions**
2. If we **assume a certain structure**, we can **apply the backdoor** adjustment for **full identification**
3. If we do not want to assume we know all confounders, we can calculate **bounds**, i.e. **partial identification**

Questions?

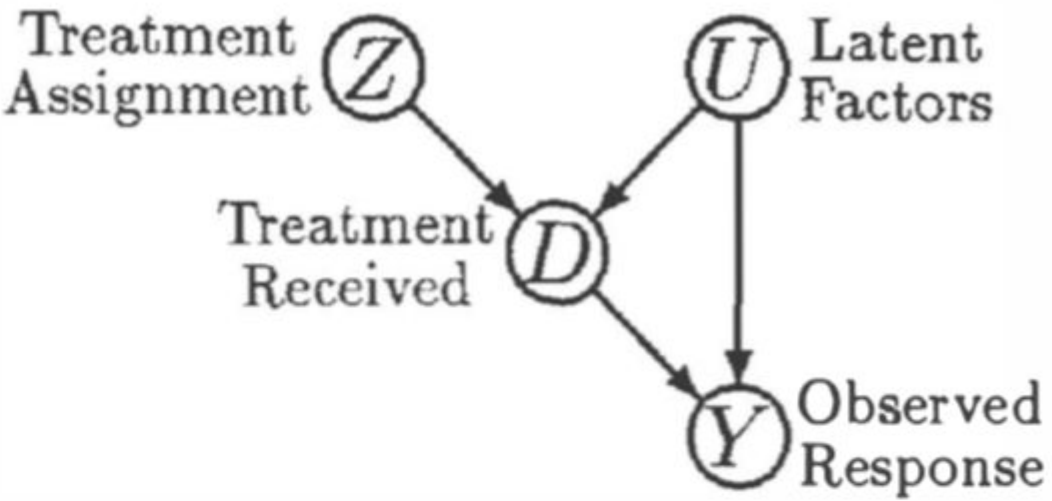
Explaining partial identification with the help of

Instrumental Variable Models

A classical workhorse method of causal inference, especially econometrics.

Instrumental Variable Model

Practically: “Imperfect Compliance“ Model



Assume:

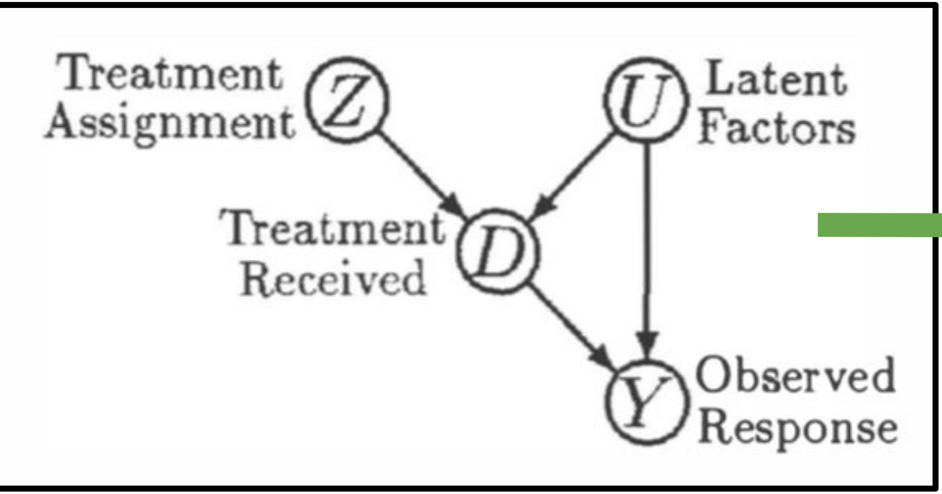
All binary, i.e. Z, D, Y = {0,1}

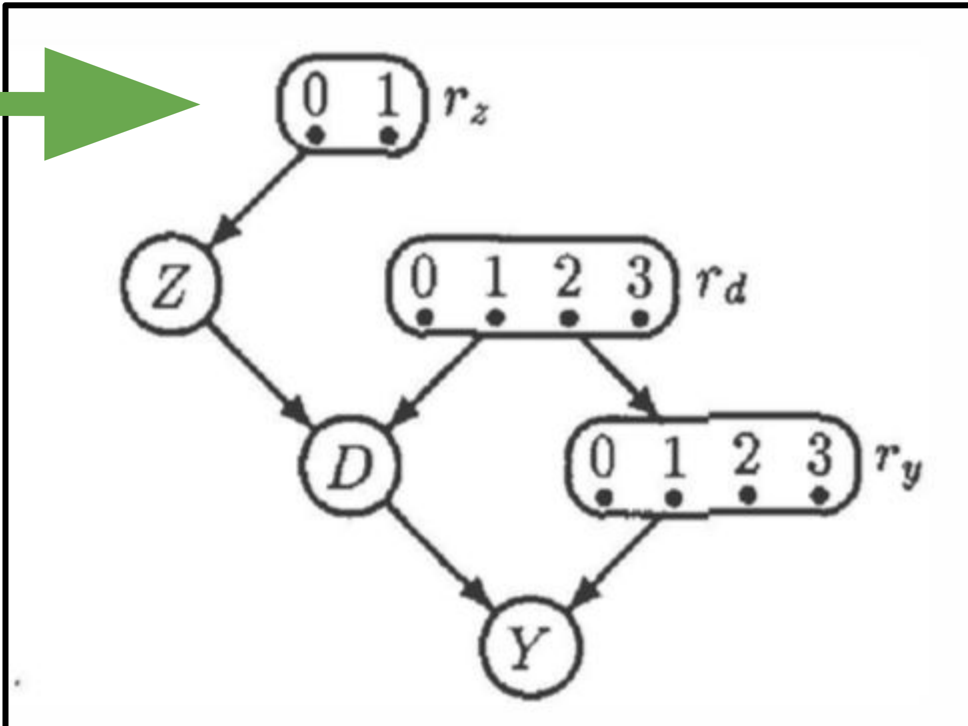
Introducing

Response Functions Variables

A very simple idea

Imperfect Compliance
with Response Functions





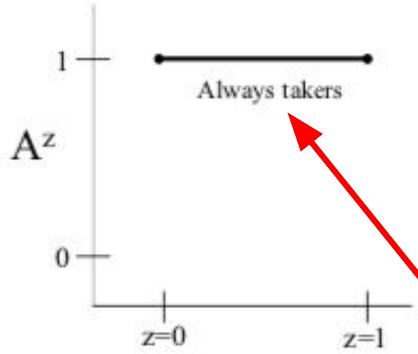
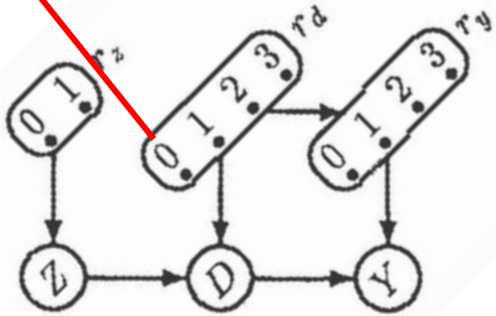

1
Always takers
A^z
0
z=0
z=1

Figure 16.4


r_z
r_d
r_y
0 1
0 1 2 3
0 1 2 3
Z
D
Y


GERICARE
NDC 57896-911-36
ADULT
LOW STRENGTH
CHEWABLE
ASPIRIN
Pain Reliever (NSAID)
Orange Flavored
PACKAGE NOT CHILD RESISTANT
36 Tablets

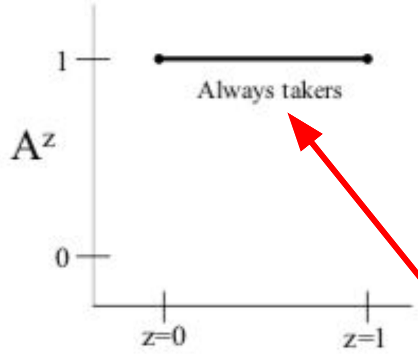


Figure 16.4

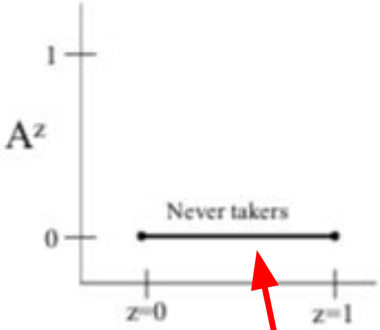
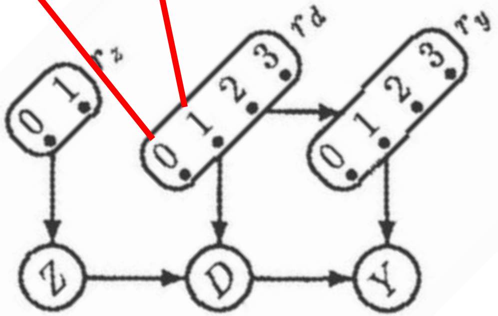


Figure 16.5



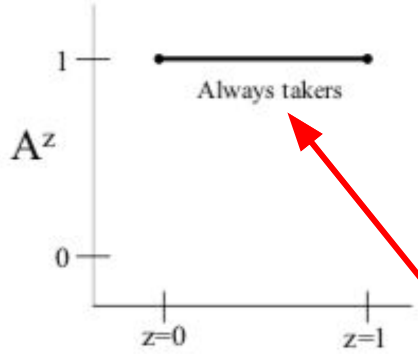


Figure 16.4

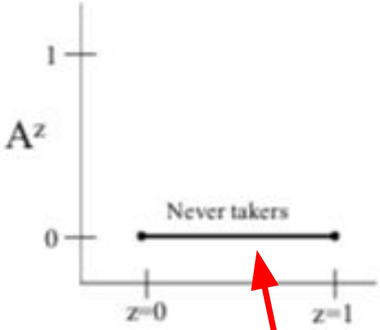


Figure 16.5

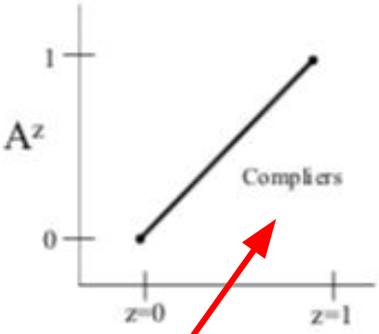
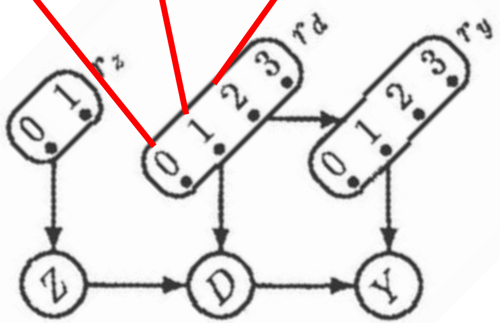


Figure 16.6



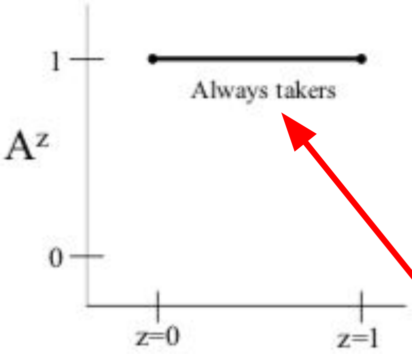


Figure 16.4

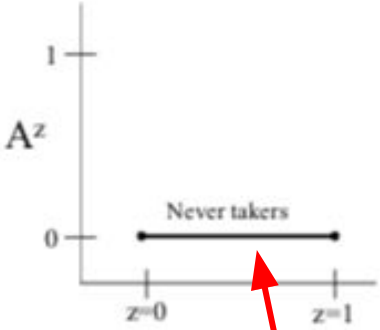


Figure 16.5

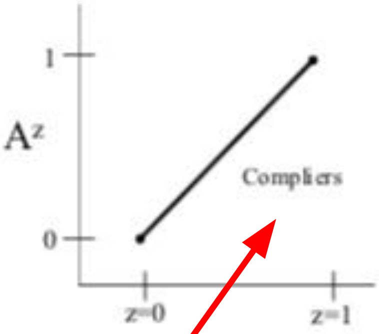


Figure 16.6

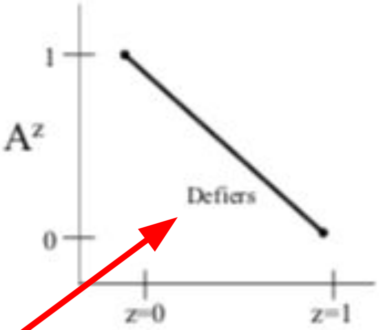
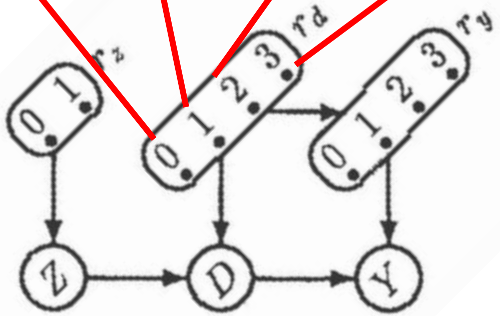


Figure 16.7



Expressing P(V={Z,D,Y}) **with** P(R={Rz,Rd,Ry})

$$P(\mathbf{v}) = \sum_{\mathbf{r}} P(\mathbf{r}) \prod_{V \in \mathbf{V}} \mathbb{I}(v; \mathsf{pa}_V, r_V),$$

Observed data P(v)

Latent model P(r)

Calculating Upper and Lower Bounds

Using Linear Programming

LPs are convex, i.e. solutions are exact

Linear Programming

$$\text{min/max} \quad \mathrm{TCE}(x_1, x_0) = \sum_{r_X, r_Y} P(r_X, r_Y)\mathbb{I}(y; x_1, r_Y) - \sum_{r_X, r_Y} P(r_X, r_Y)\mathbb{I}(y; x_0, r_Y),$$

Linear Programming

$$\begin{aligned}
\text{min/max} \quad & \mathrm{TCE}(x_1, x_0) = \sum_{r_X, r_Y} P(r_X, r_Y)\mathbb{I}(y; x_1, r_Y) - \sum_{r_X, r_Y} P(r_X, r_Y)\mathbb{I}(y; x_0, r_Y), \\
\text{s.t.} \quad & P(\mathbf{V}) = P(\mathcal{D}), \quad \sum_{\mathbf{r}} P(\mathbf{r}) = 1, \quad P(\mathbf{r}) \geq 0,
\end{aligned}$$

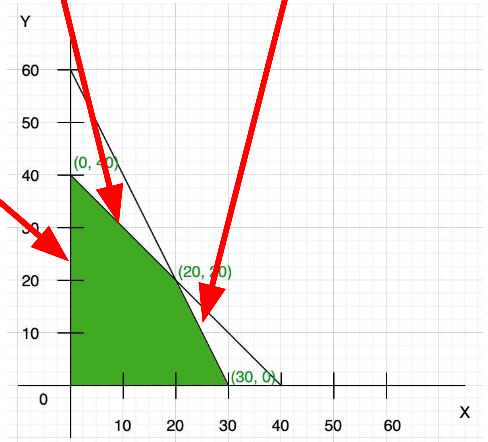
Linear Programming

$$\text{min/max} \quad \mathrm{TCE}(x_1, x_0) = \sum_{r_X, r_Y} P(r_X, r_Y)\mathbb{I}(y; x_1, r_Y) - \sum_{r_X, r_Y} P(r_X, r_Y)\mathbb{I}(y; x_0, r_Y),$$

$$\text{s.t.} \quad \boxed{P(\mathbf{V}) = P(\mathcal{D}),} \quad \boxed{\sum_{\mathbf{r}} P(\mathbf{r}) = 1, \quad P(\mathbf{r}) \geq 0,}$$

Matching observed data and latent model

Probability “Simplex”



Linear Programming

$$\text{min/max} \quad \text{TCE}(x_1, x_0) = \sum_{r_X, r_Y} P(r_X, r_Y)\mathbb{I}(y; x_1, r_Y) - \sum_{r_X, r_Y} P(r_X, r_Y)\mathbb{I}(y; x_0, r_Y),$$

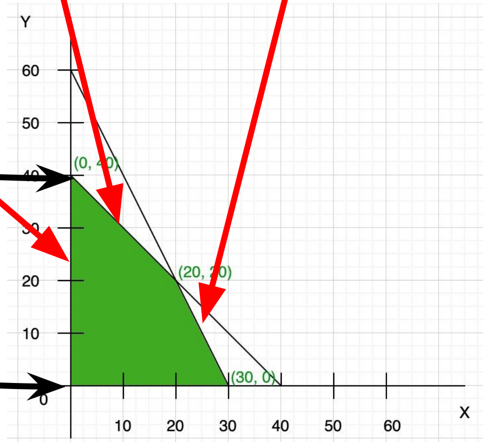
$$\text{s.t.} \quad \boxed{P(\mathbf{V}) = P(\mathcal{D}),} \quad \boxed{\sum_{\mathbf{r}} P(\mathbf{r}) = 1, \quad P(\mathbf{r}) \geq 0,}$$

Matching observed data and latent model

Probability “Simplex”

Upper Bound

Lower Bound



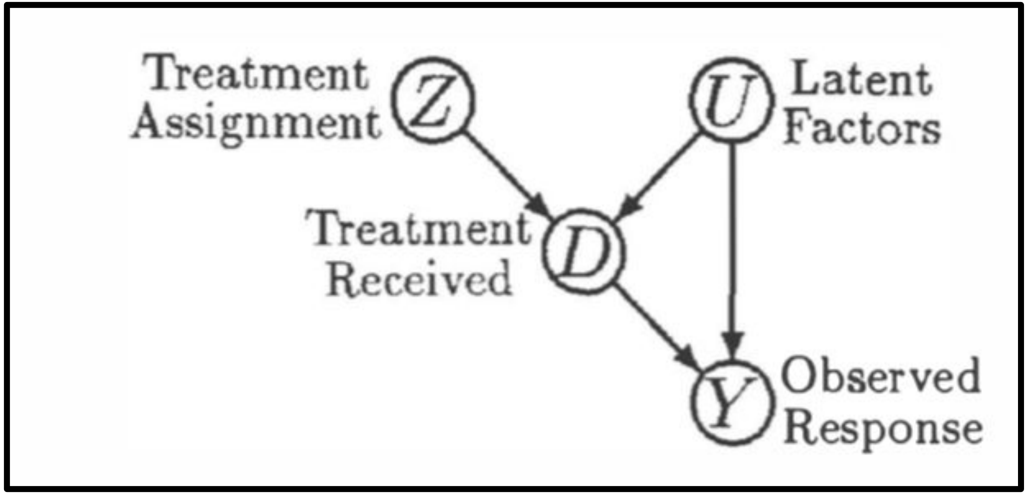
Partial Identification: **Challenges**

*Why is partial identification **hard?***

Beyond Instrumental Variable Models

How to generalise response functions variables

So far ...


Treatment
Assignment
Z
U
Latent
Factors
Treatment
Received
D
Y
Observed
Response

Response Functions Variables are *universal*

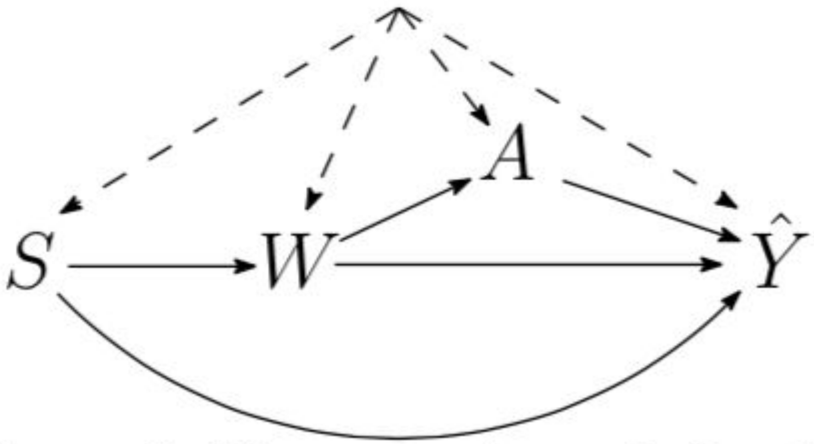
Response Function Variables:

1. Number of states “coming in” from parents
2. Numbers of “states” of node (the domain)

$$N_V = |V|^{|\mathrm{PA}_V|}$$

*How many Response Function states **per variable**?*

$$N_V = |V|^{|\mathsf{PA}_V|}$$



S =

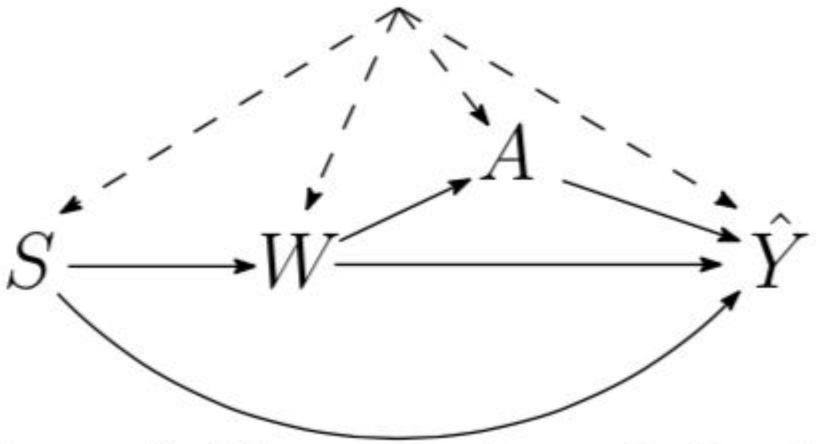
W =

A =

Y =

*How many Response Function states **per variable**?*

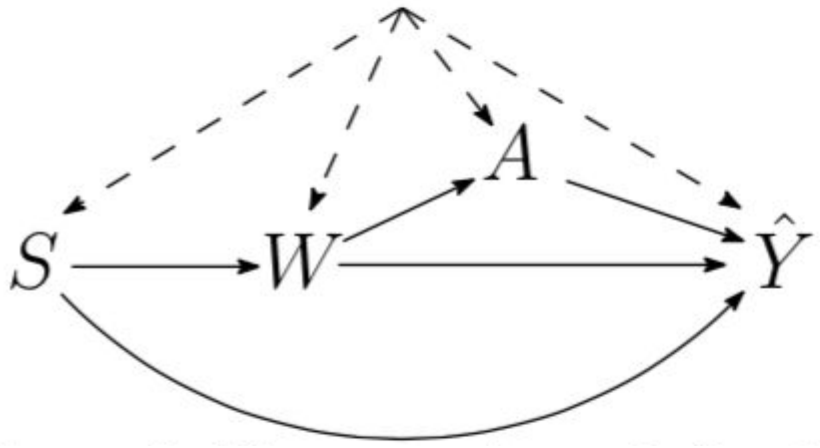
$$N_V = |V|^{|\mathsf{PA}_V|}$$



S = 2
W =
A =
Y =

*How many Response Function states **per variable**?*

$$N_V = |V|^{|\mathrm{PA}_V|}$$



S = 2

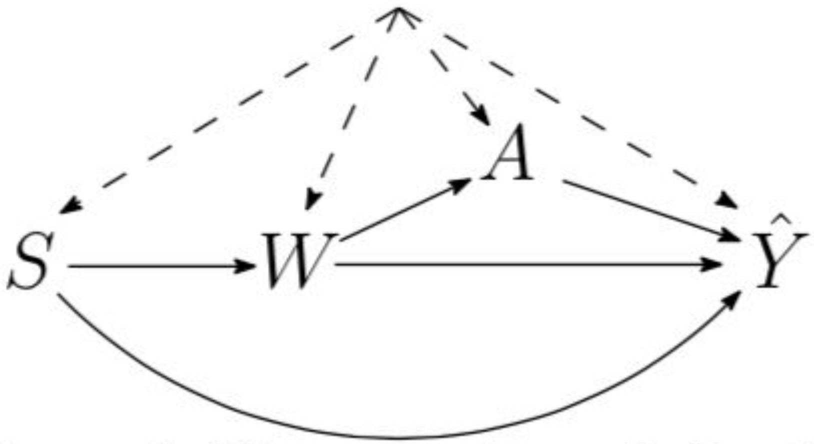
W = 2 ** 2

A =

Y =

How many Response Function states ***per variable****?*

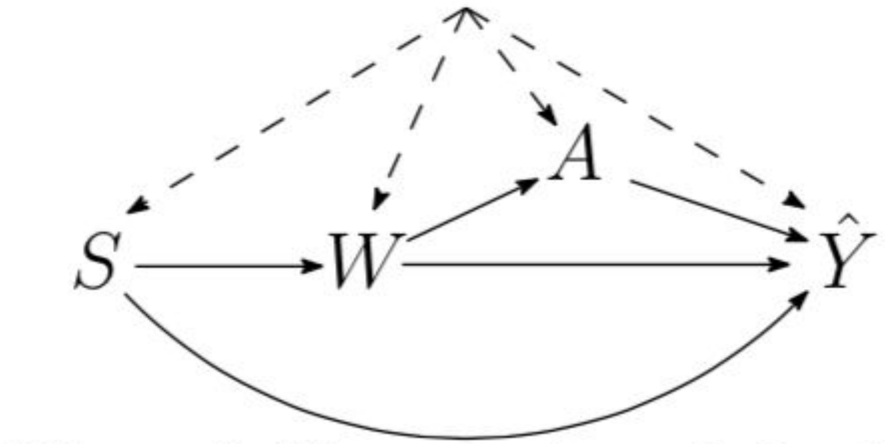
$$N_V = |V|^{|\mathsf{PA}_V|}$$



S = 2
W = 2 ** 2
A = 2 ** 2
Y =

How many Response Function states ***per variable****?*

$$N_V = |V|^{|\mathrm{PA}_V|}$$



S = 2

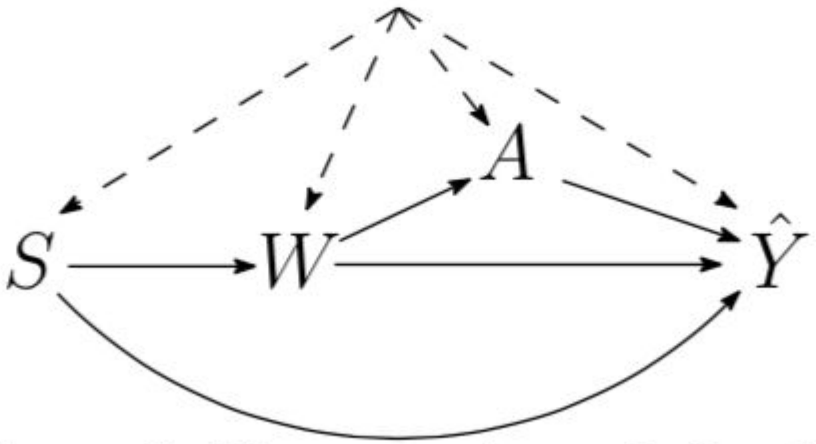
W = 2 ** 2

A = 2 ** 2

Y = 2 ** (2 * 2 * 2) = 256 (!!!)

How many Response Function states ***per variable****?*

$$N_V = |V|^{|\mathrm{PA}_V|}$$

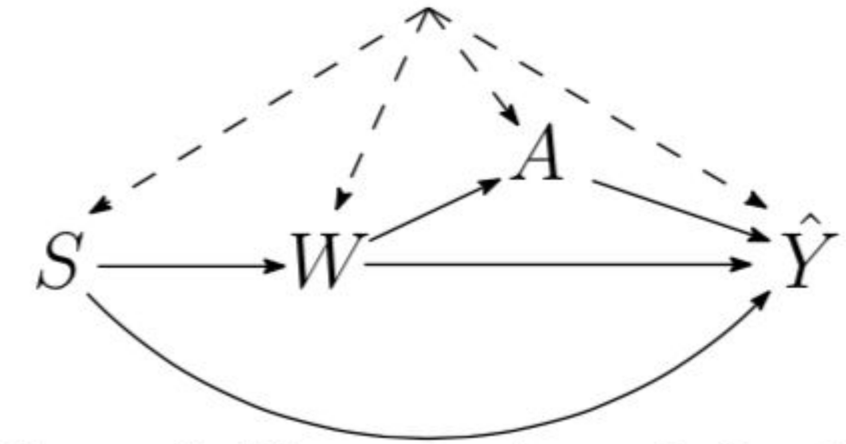


S = 2
W = 2 ** 2
A = 2 ** 2
Y = 2 ** (2 * 2 * 2) = 256 (!!!)

How many Response Function states ***for the whole model****?*

*How many Response Function states **per variable**?*

$$N_V = |V|^{|\mathrm{PA}_V|}$$



S = 2

W = 2 ** 2

A = 2 ** 2

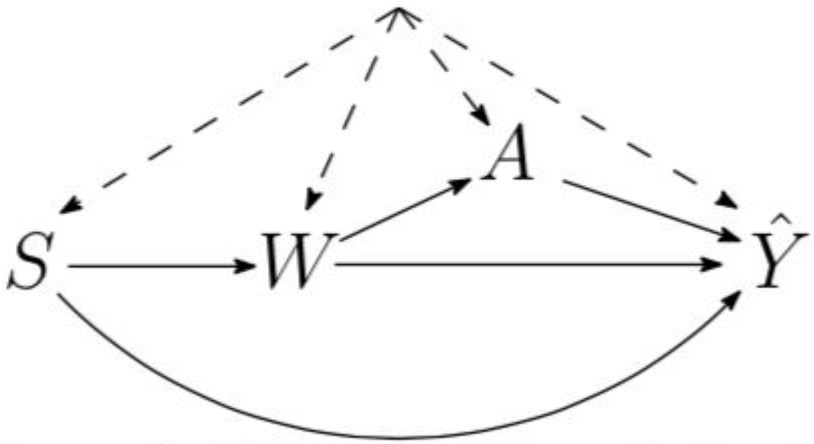
Y = 2 ** (2 * 2 * 2) = 256 (!!!)

*How many Response Function states **for the whole model**?*

2 * 4 * 4 * 256 = <u>8192</u>

How many Response Function states ***per variable****?*

$$N_V = |V|^{|\mathrm{PA}_V|}$$



S = 2
W = 2 ** 2
A = 2 ** 2
Y = 2 ** (2 * 2 * 2) = 256 (!!!)

How many Response Function states ***for the whole model****?*

2 * 4 * 4 * 256 = <u>8192</u>

Problem: RFV space increases *super-exponentially*.
How can we deal with that?

Partial Identification: **Solutions**

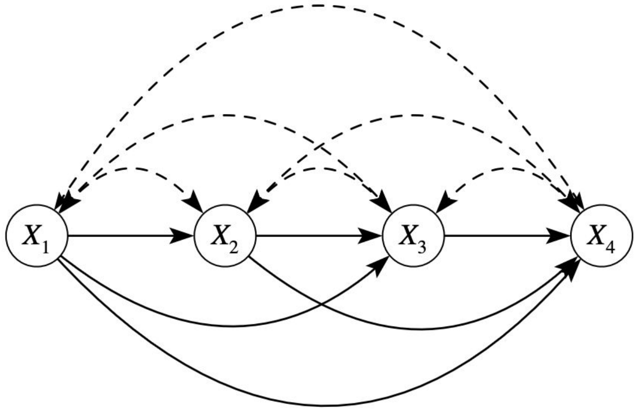
How is partial identification ***feasible?***

Solution: **Causal Marginal Polytope**

Idea: Consider smaller parts of the graph, aka. “marginals”

Example:

- 4 Variables
- Fully connected
- Fully confounded

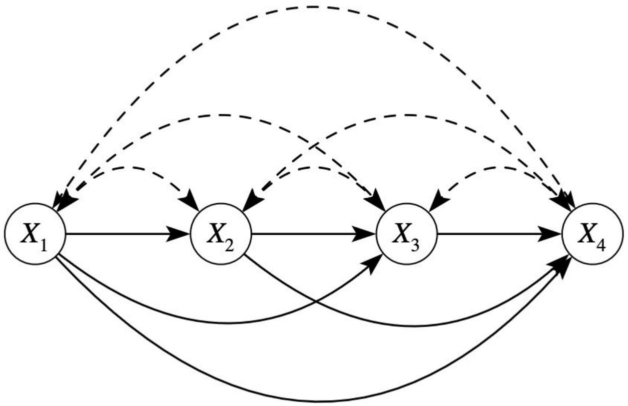


The Causal Marginal Polytope for Bounding Treatment Effects

Jakob Zeitler, Ricardo Silva

https://arxiv.org/abs/2202.13851

"Complete" model

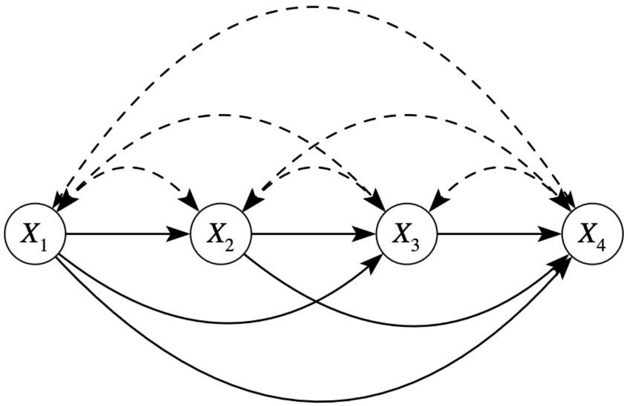


Marginals: 1, i.e. the complete 4 variable model

Number of Parameters: 32,768

Tightness of bounds: Sharp

“Complete” model

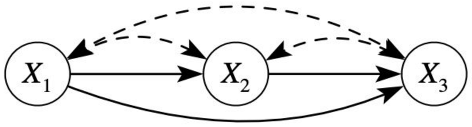


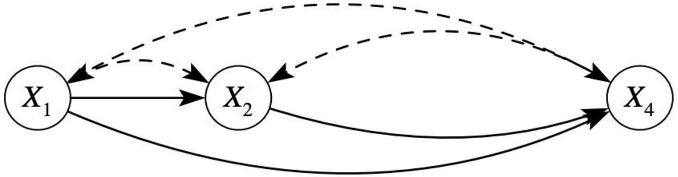
Marginals: 1, i.e. the complete 4 variable model

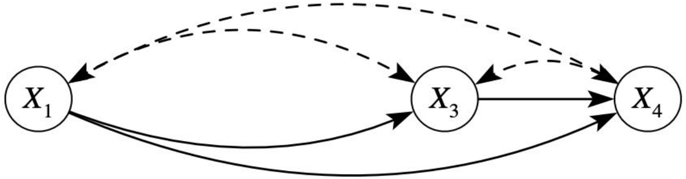
Number of Parameters: 32,768

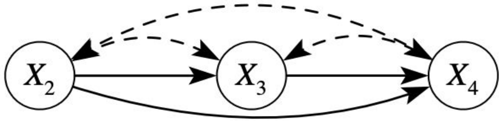
Tightness of bounds: Sharp

Causal Marginal Polytope







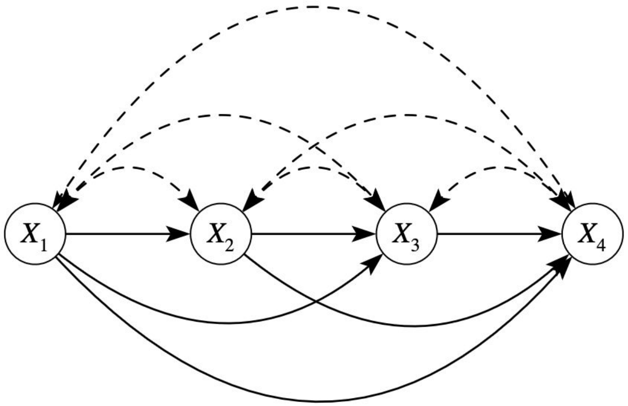


Marginals: 4, each with 3 variables

Number of Parameters:128*4=512

Tightness of bounds: More loose

“Complete” model

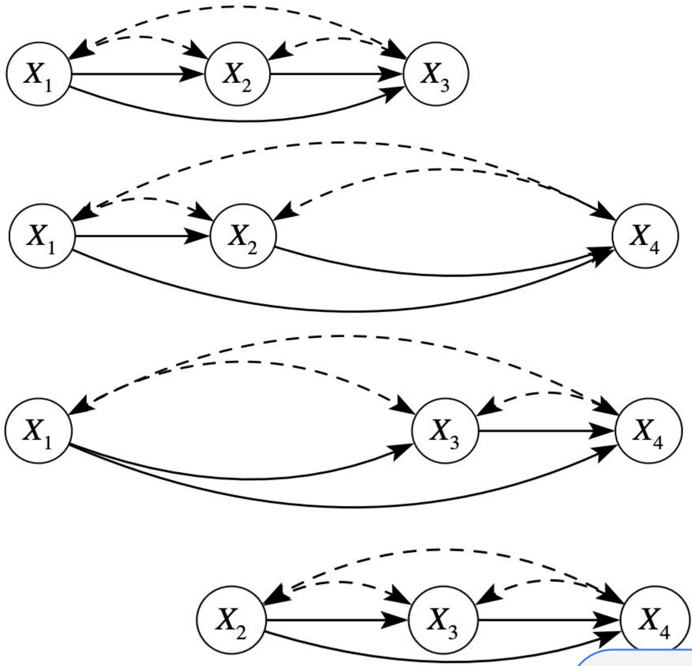


Marginals: 1, i.e. the complete 4 variable model

Number of Parameters: 32,768

Tightness of bounds: Sharp

Causal Marginal Polytope



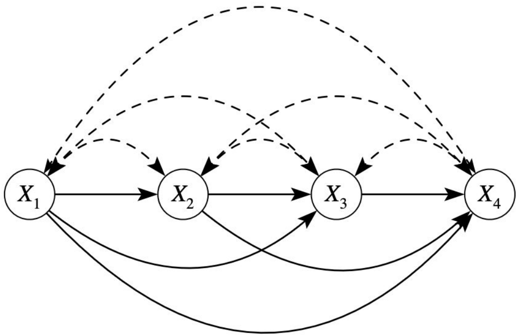
Marginals: 4, each with 3 variables

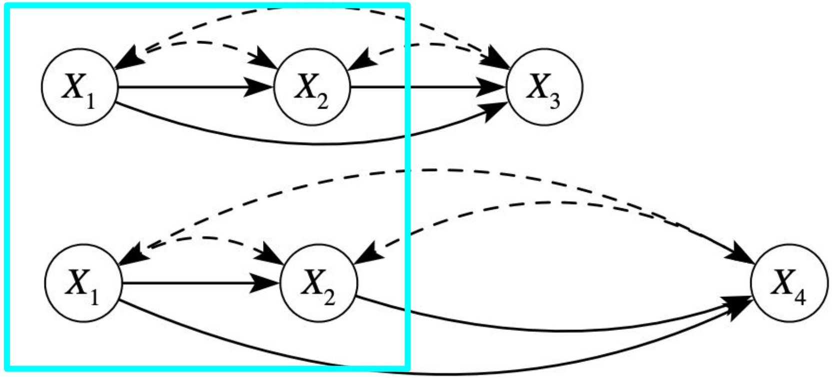
Number of Parameters:128*4=512

Tightness of bounds: More loose

Let’s add constraints to tighten the bounds

Constraint A: **Overlaps**

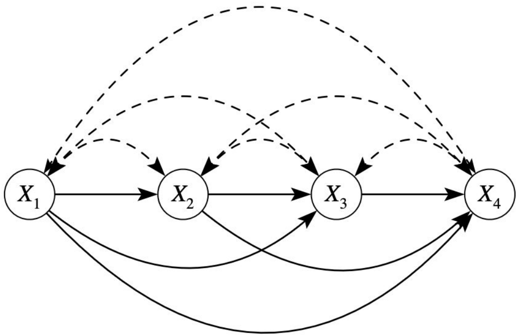


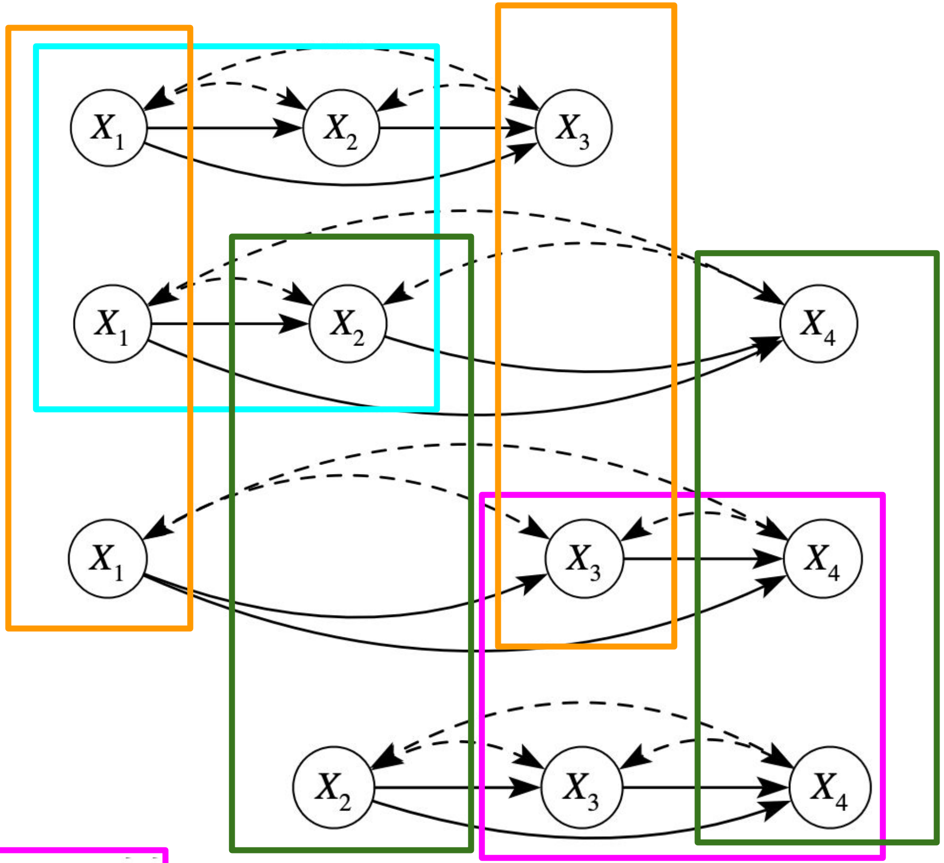


$$\boxed{\sum_{X_3} P(X_1, X_2, X_3) = \sum_{X_4} P(X_1, X_2, X_4)}$$

$\{\boxed{\{X_1, X_2\}},$

Constraint A: **Overlaps**





$$\{\{X_1, X_2\}, \{X_2, X_4\}, \{X_1, X_3\}, \{X_3, X_4\}\}.$$

$$
\begin{aligned}
&\min/\max \quad \boxed{\text{Objective: e.g. TCE}} \\
&\text{s.t.} \quad P(\mathbf{V}) = P(\mathcal{D}), \quad \sum_{\mathbf{r}} P(\mathbf{r}) = 1, \quad P(\mathbf{r}) \geq 0,
\end{aligned}
$$

Constraint A:

Overlaps

$$\text{min/max} \quad \text{Objective: e.g. TCE}$$

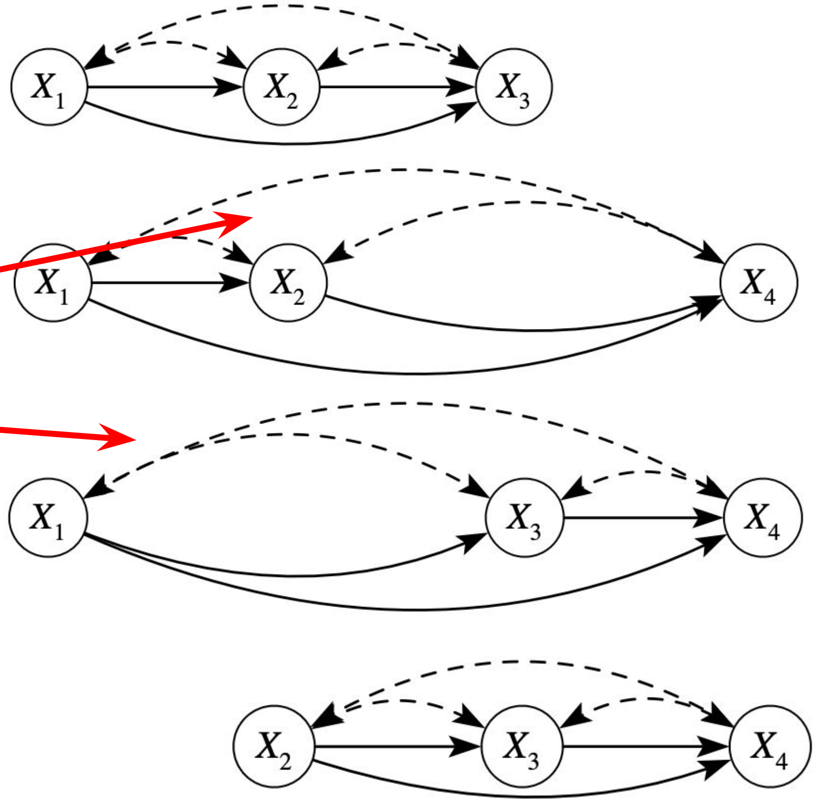
$$\text{s.t.} \quad P(\mathbf{V}) = P(\mathcal{D}), \quad \sum_{\mathbf{r}} P(\mathbf{r}) = 1, \quad P(\mathbf{r}) \geq 0,$$

Constraint A:
Overlaps

$$\sum_{X_3} P(X_1, X_2, X_3) = \sum_{X_4} P(X_1, X_2, X_4)$$

Constraint B:
Expert knowledge

1. Bidirected edges
2. Directed edges



$$\min/\max \quad \boxed{\text{Objective: e.g. TCE}}$$

$$\text{s.t.} \quad P(\mathbf{V}) = P(\mathcal{D}), \quad \sum_{\mathbf{r}} P(\mathbf{r}) = 1, \quad P(\mathbf{r}) \geq 0,$$

Constraint A:
Overlaps

$$\sum_{X_3} P(X_1, X_2, X_3) = \sum_{X_4} P(X_1, X_2, X_4)$$

Constraint B:
Expert knowledge

$$\text{min/max} \quad \text{Objective: e.g. TCE}$$

$$\text{s.t.} \quad P(\mathbf{V}) = P(\mathcal{D}), \quad \sum_{\mathbf{r}} P(\mathbf{r}) = 1, \quad P(\mathbf{r}) \geq 0,$$

Constraint A:
Overlaps

$$\sum_{X_3} P(X_1, X_2, X_3) = \sum_{X_4} P(X_1, X_2, X_4)$$

Constraint B:
Expert knowledge

Directed edges

$$|P_{\mathcal{M}}(V_i = 1 \mid do(v_{pa_i \setminus j}), do(V_j = 1), v_{\mathcal{M}'}) -$$
$$P_{\mathcal{M}}(V_i = 1 \mid do(v_{pa_i \setminus j}), do(V_j = 0), v_{\mathcal{M}'})| \leq \epsilon_{ij},$$

$$\text{min/max}$$

Objective: e.g. TCE

$$\text{s.t.} \quad P(\mathbf{V}) = P(\mathcal{D}), \quad \sum_{\mathbf{r}} P(\mathbf{r}) = 1, \quad P(\mathbf{r}) \geq 0,$$

Constraint A:
Overlaps

$$\sum_{X_3} P(X_1, X_2, X_3) = \sum_{X_4} P(X_1, X_2, X_4)$$

Constraint B:
Expert knowledge

Directed edges

$$|P_{\mathcal{M}}(V_i = 1 \mid do(v_{pa_i \setminus j}), do(V_j = 1), v_{\mathcal{M}'}) -$$

$$P_{\mathcal{M}}(V_i = 1 \mid do(v_{pa_i \setminus j}), do(V_j = 0), v_{\mathcal{M}'})| \leq \epsilon_{ij},$$

Bidirected edges

$$|P_{\mathcal{M}}(V_i = 1 \mid do(v_{pa_{ij}}), do(V_j = v_j), v_{\mathcal{M}'}) -$$

$$P_{\mathcal{M}}(V_i = 1 \mid do(v_{pa_{ij}}), V_j = v_j, v_{\mathcal{M}'})| \leq \epsilon_{ij}^{C},$$

Example: **Simulation**

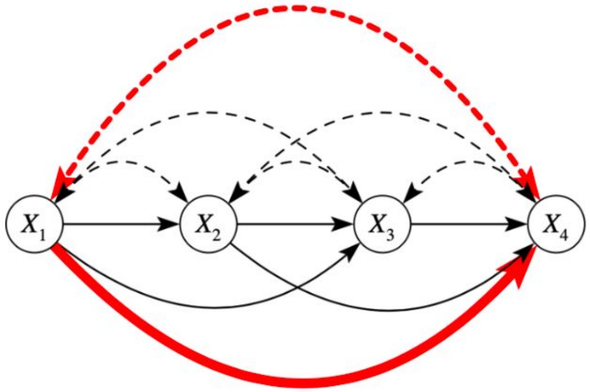
- 4 Variables
- We constrain the red directed and bidirected edges
- We supply the following data regimes

$do(\emptyset)$

$do(x_2), do(x_3)$

$do(x_2, x_3)$

$do(x_1 = 0, x_3 = 0), do(x_1 = 0, x_3 = 1)$



Example: **Simulation**

- 4 Variables
- We constrain the red directed and bidirected edges
- We supply the following data regimes

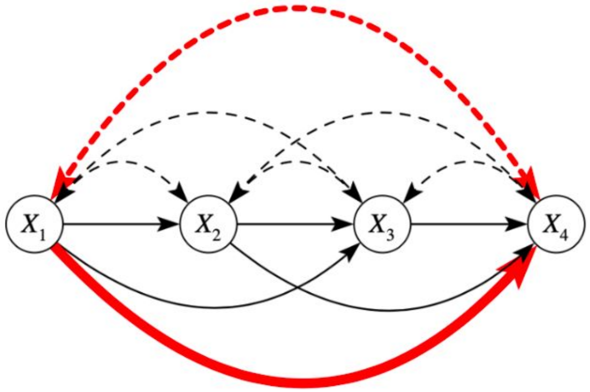
$do(\emptyset)$

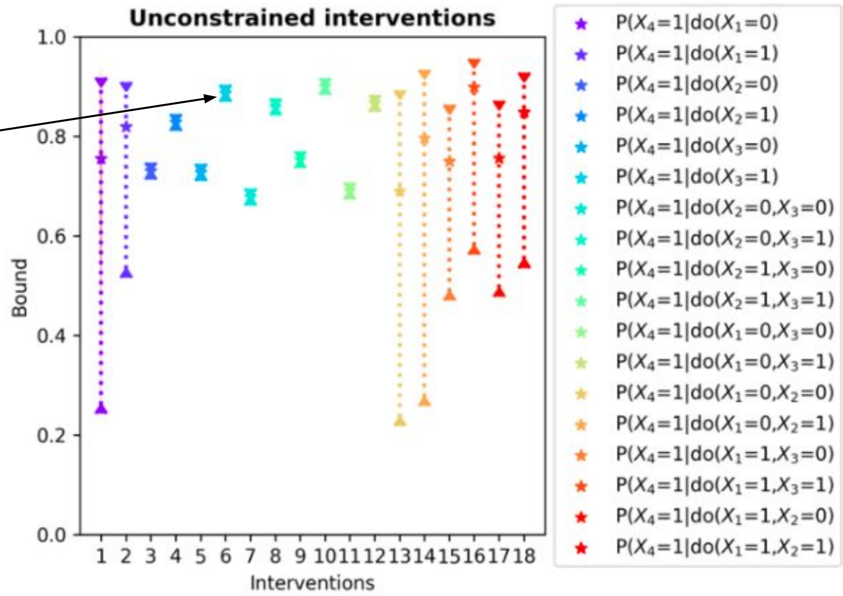
$do(x_2), do(x_3)$

$do(x_2, x_3)$

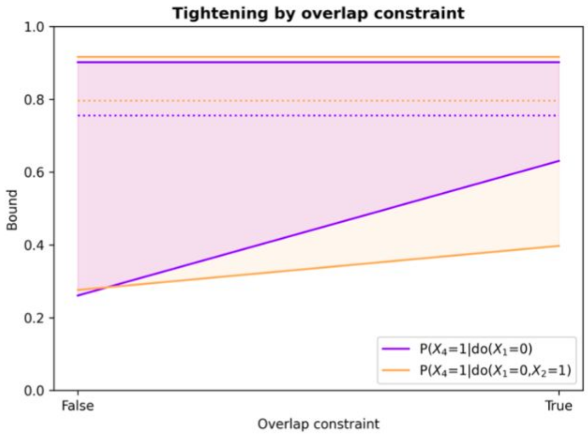
$do(x_1 = 0, x_3 = 0), do(x_1 = 0, x_3 = 1)$

Identification from data



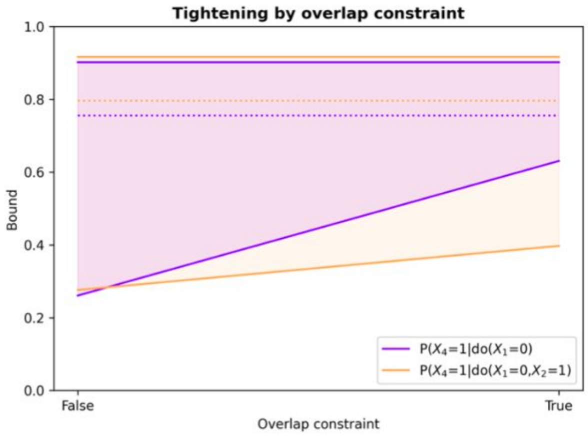


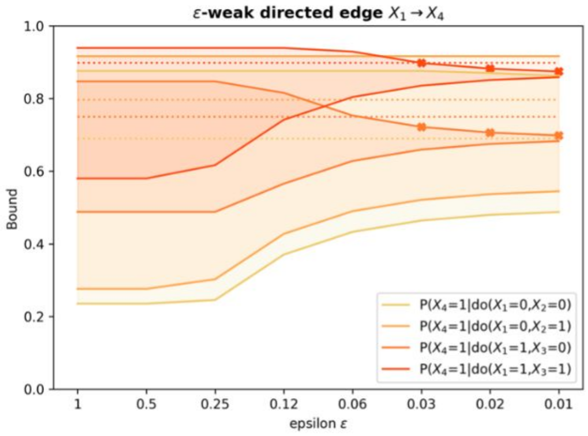
Results: Each constraint by itself



Tightening of bounds by constraining the **overlap of the margins**.

Results: Each constraint by itself

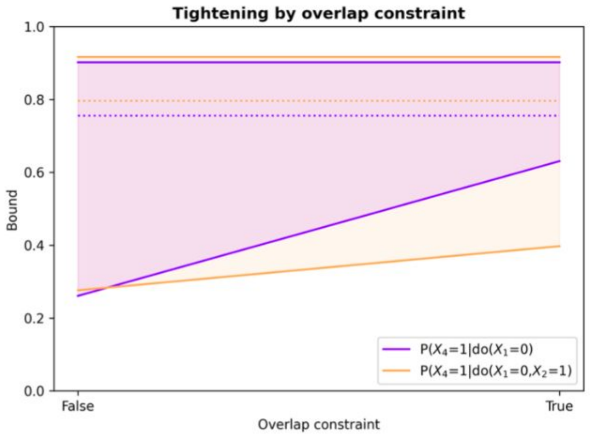


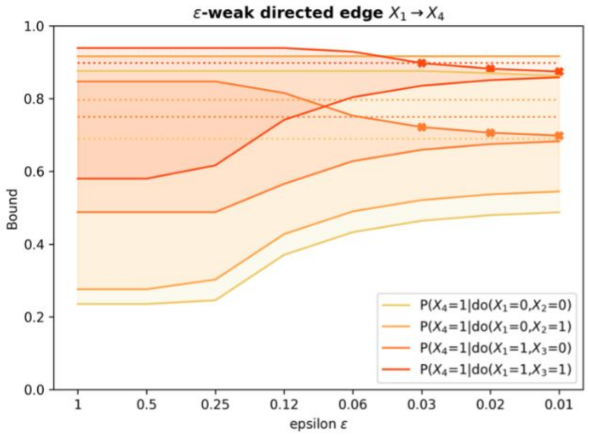


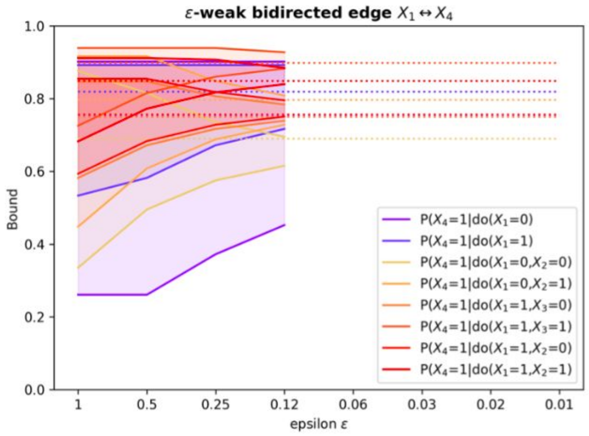
Tightening of bounds by constraining the **overlap of the margins**.

A **weak directed edge** tightens bounds. Invalid bounds are marked with a cross.

Results: Each constraint by itself







Tightening of bounds by constraining the **overlap of the margins**.

A **weak directed edge** tightens bounds. Invalid bounds are marked with a cross.

A **weak bidirected** edge tightens bounds. Lower epsilon values are falsified by infeasibility.

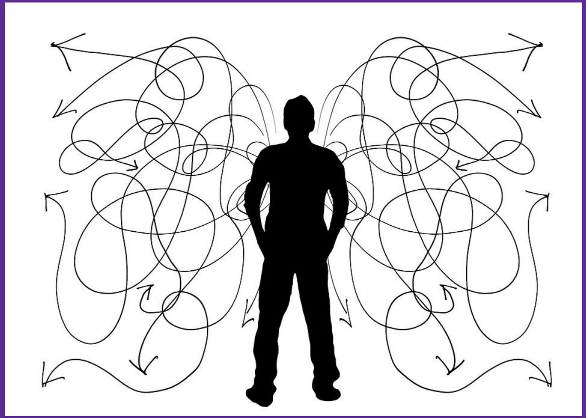
Let's review, again

We have learned:

1. If we do not want to assume we know all confounders, we can calculate **bounds**, i.e. **partial identification**
2. But *exact* partial identification is hard: **there is no free lunch**.

Partial Identification: **Alternatives**

*What **alternative solutions** are there?*



What's really behind partial identification

Take the IV model on the left, with

two *possible models for B.*

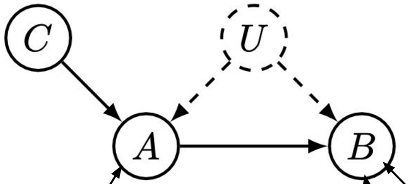


Figure 1: Example: A, B, C are observed variables and U is a hidden variable.

U	C	$A = f_A$
0	0	0
0	1	0
1	0	1
1	1	1

Table 1: Equation $f_A(c, u)$ for determining values of A.

U	A	$B = f_B^1$	$B = f_B^2$
0	0	0	0
0	1	1	0
1	0	0	0
1	1	1	1

Table 2: Equations $f_B^1(a, u)$ and $f_B^2(a, u)$ for determining values of B.

See Hu et al. **A Generative Adversarial Framework for Bounding Confounded Causal Effects**, 2021

What's really behind partial identification

Take the IV model on the left, with

***two** possible models for B.*

1. Setting P(U = 1) as p
2. The first model implies:
   a. $P(B = 1 | do(A = 1)) = 1$
   b. $P(B = 1 | do(A = 0)) = 0$
   c. Therefore, ACE= 1 - 0 = 1

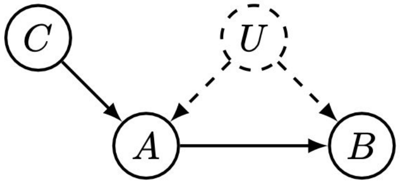


Figure 1: Example: A, B, C are observed variables and U is a hidden variable.

U	C	$A = f_A$
0	0	0
0	1	0
1	0	1
1	1	1

Table 1: Equation $f_A(c, u)$ for determining values of A.

U	A	$B = f_B^1$	$B = f_B^2$
0	0	0	0
0	1	1	0
1	0	0	0
1	1	1	1

Table 2: Equations $f_B^1(a, u)$ and $f_B^2(a, u)$ for determining values of B.

See Hu et al. **A Generative Adversarial Framework for Bounding Confounded Causal Effects**, 2021

What's really behind partial identification

Take the IV model on the left, with

***two** possible models for B.*

1. Setting P(U = 1) as p
2. The first model implies:
   a. P (B = 1|do(A = 1)) = 1
   b. P(B = 1|do(A = 0)) = 0
   c. Therefore, ACE= 1 - 0 = 1
3. The second model implies:
   a. P (B = 1|do(A = 1)) = **p**
   b. P (B = 1|do(A = 0)) = 0
   c. Therefore, ACE=p - 0 = p
4. We cannot reject either model
5. Therefore, the bounds are: [p,1]

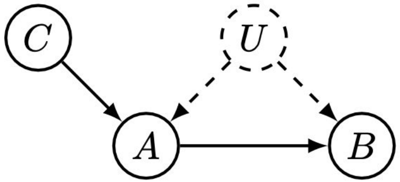


Figure 1: Example: A, B, C are observed variables and U is a hidden variable.

U	A	$B = f_B^1$	$B = f_B^2$
0	0	0	0
0	1	1	0
1	0	0	0
1	1	1	1

Table 2: Equations $f_B^1(a, u)$ and $f_B^2(a, u)$ for determining values of B.

See Hu et al. **A Generative Adversarial Framework for Bounding Confounded Causal Effects**, 2021

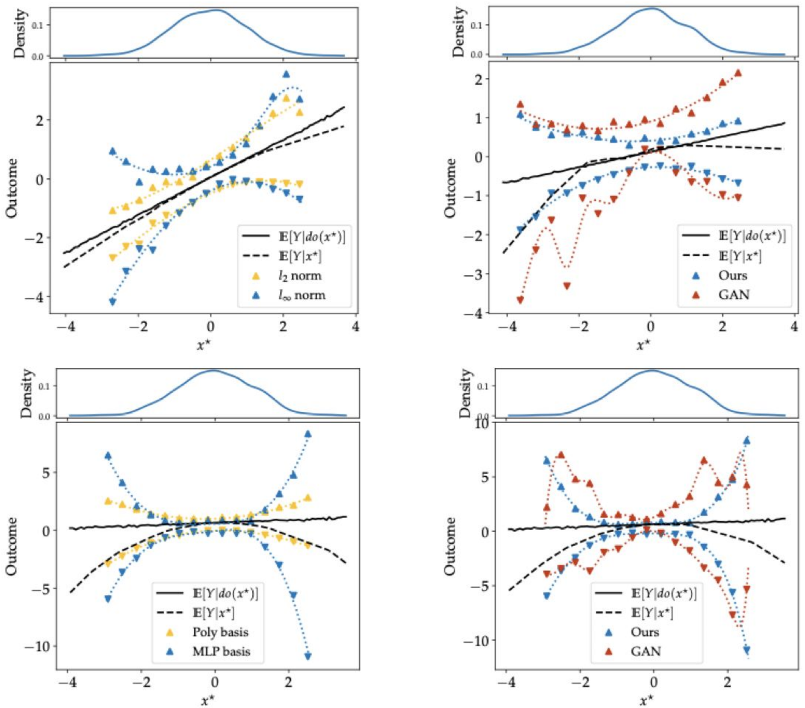
Approximate Partial Identification

1. Define a class of possible models
2. Reject models that disagree with that data
3. Calculate ATEs for all models left
4. The highest and lowest are your bounds

$$\min_{\eta \in \mathbb{R}^d} / \max \quad o_{x^\star}(\eta) = \psi(x^\star)^\top \, \mathbb{E}_N[\mu_{\eta_0}(N)]$$

$$\text{subject to} \quad \mathrm{dist}(p_\eta, \hat{p}) \leq \epsilon \,.$$

Stochastic Causal Programming for Bounding Treatment Effects
Kirtan Padh, Jakob Zeitler, David Watson, Matt Kusner, Ricardo Silva, Niki Kilbertus
https://arxiv.org/abs/2202.10806



Where we end today:

The goal of causal modelling is **not identification as such**, but to make the best out of the observables that we do have.

Don't say: ***"I can't provide a unique solution from your assumptions, go home."***

Do say: ***"These are all the solutions compatible with your assumptions".***

No academic talk without a ***meme***

how it started	how it's going
The **core question** of causal inference: **Identifiability**	**Make the best out of the observables** that we have
"I can't provide a unique solution from your assumptions, go home."	***"These are all the solutions compatible with your assumptions".***

Questions & Answers

Time for

- Questions
- Discussion
- Feedback
- 40+ slides in the Appendix

You can also talk to me about

1. **Synthetic Control**
2. **(Causal) Bayesian Optimisation**
3. **Topological Perspectives of Causality**

partial-identification.com

- Book (draft)
- 45 minute version of this talk

Contribute to Causal Genealogy: genealogy.causality.link

External Link to Spreadsheet for Editing: https://docs.google.com/spreadsheets/d/1rNrb7rvIVnUR0CPBqbvJNNlx76wfA5038rJ5zHrSZFk/edit?usp=sharing
How to use this Geneology (inspired by https://www.genealogy.math.ndsu.nodak.edu)
This spreadsheet is an overview of the researchers in causal inference.
For beginners, this is hopefully a useful resource.

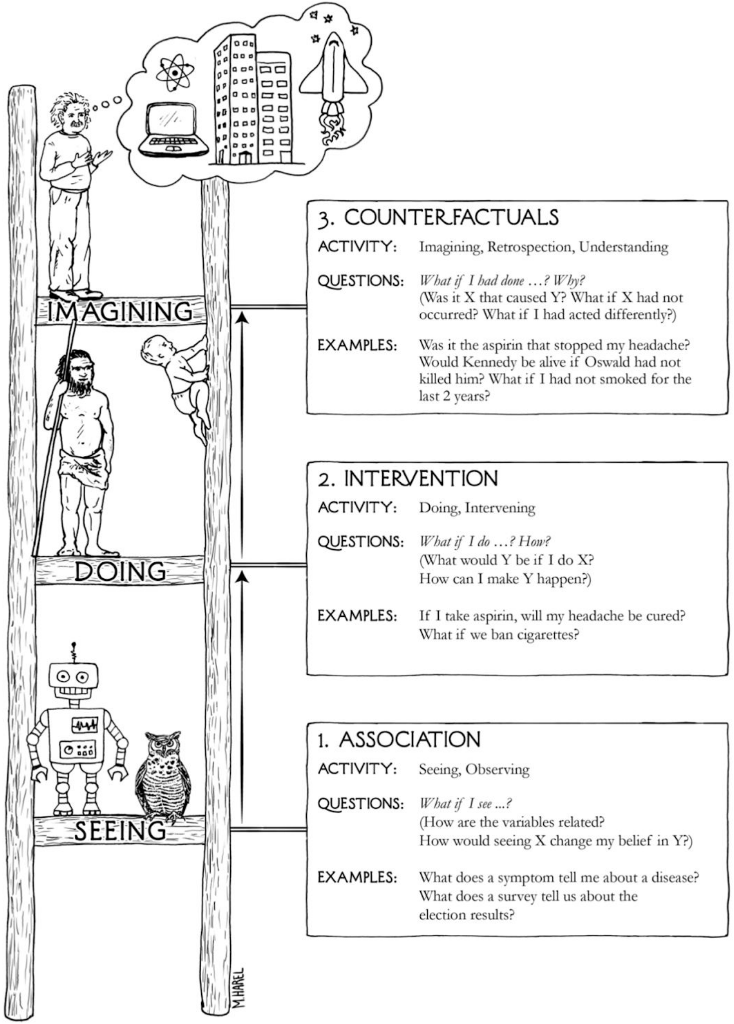
FAQ
How do I add more people to the sheet?
Simply use the Google Docs commenting function and we will turn your comment into an entry.

Name	Institution	Supervisor	Location	Previous Positions	Link
Academia					
UCLA					
Judea Pearl	UCLA	?	US	Rutgers, Technion, Newark College of Engineering	http://b
Wesley Salmon	UCLA	Hans Reichenbach	US	?	
Hans Reichenbach	UCLA	Paul Hensel, Max Noeth	US	Berlin, Istanbul, Erlangen	
John Hopkins					
Ilya Shpitser	John Hopkins		US	UCLA, Judea Pearl	https://
Oregon State University					
Karthika Mohan	Oregon State University	Judea Pearl	US		http://w
CMU					
Kun Zhang	CMU		Pittsburgh, US	MPI Tübingen	
Clark Glymour	CMU	Wesley Salmon	Pittsburgh, US		
Peter Spirtes	CMU		Pittsburgh, US		
ETH Zürich					
Peter Bühlmann	ETH		Zürich	?	
Marloes Maathuis	ETH		Zürich	?	https://
Nicolai Meinshausen	ETH				

Appendix

The Causal Ladder

(Pearl, 2018: "Book of Why")



Two schools of thought: Apples and Oranges?

	Potential Outcomes	Graphical Models (DAGs)
Areas	Epidemiology, Econometrics	Computer Science, Artificial Intelligence
Schools	Harvard, Berkeley and more	UCLA, CMU and more

Discussion see here:
https://www.jstor.org/stable/4616823?seq=1

Compare: Identifiability in RCTs and obs. studies

Randomized experiments

(Conditional) Exchangeability

True by design! (randomized)

Identifiability

Calculate effect via e.g. IPW.

Observational studies *(The focus of this talk)*

True?

Identifiability?

(Alternatively, if these conditions are not met, one can try to use a *instrumental variable model*)

Compare: Identifiability in RCTs and obs. studies

Randomized experiments

(Conditional) Exchangeability → *True by design! (randomized)* → **Identifiability**

Calculate effect via e.g. IPW.

Observational studies *(The focus of this talk)*

Consistency

(Conditional) Exchangeability

Positivity

→ *True?* →

Identifiability

Observational study acts *as if* it was a conditionally randomized study.

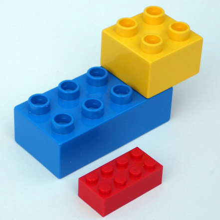
Calculate effect via e.g. IPW.

(Alternatively, if these conditions are not met, one can try to use a *instrumental variable model*)

What does the causal graph look like?

If we want to adjust for confounding, we need to ***identify the confounders!***

Define SCM: <**V**, **F**, **U**>
V: observed variables
F: structural equations
U: background variables, with P(**U**)



“The building blocks of causal DAGs (directed acyclic graphs)”

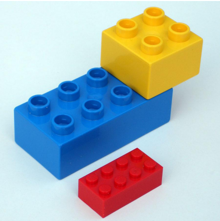
Pearl, Causality, 2009

What does the causal graph look like?

If we want to adjust for confounding, we need to ***identify the confounders!***

$$V = \{X, Y, Z\}, U = \{U_X, U_Y, U_Z\}, F = \{f_X, f_Y, f_Z\}$$

Define SCM: <**V**, **F**, **U**>
V: observed variables
F: structural equations
U: background variables, with P(**U**)



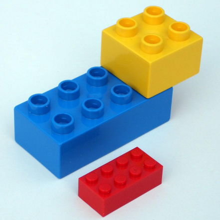
This SCM implies this DAG:

"The building blocks of causal DAGs (directed acyclic graphs)"

Pearl, Causality, 2009

What does the causal graph look like?

If we want to adjust for confounding, we need to ***identify the confounders!***

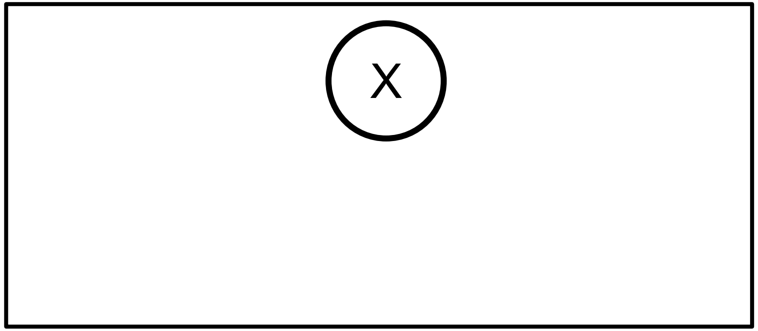


Define SCM: <**V**, **F**, **U**>
V: observed variables
F: structural equations
U: background variables, with P(**U**)

$$V = \{X, Y, Z\}, U = \{U_X, U_Y, U_Z\}, F = \{f_X, f_Y, f_Z\}$$

$$f_X : X = U_X$$

This SCM implies this DAG:



"The building blocks of causal DAGs (directed acyclic graphs)"

Pearl, Causality, 2009

What does the causal graph look like?

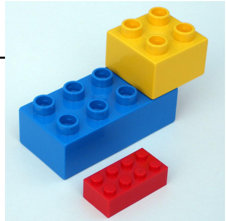
*If we want to adjust for confounding, we need to **identify the confounders!***

$$V = \{X, Y, Z\}, U = \{U_X, U_Y, U_Z\}, F = \{f_X, f_Y, f_Z\}$$

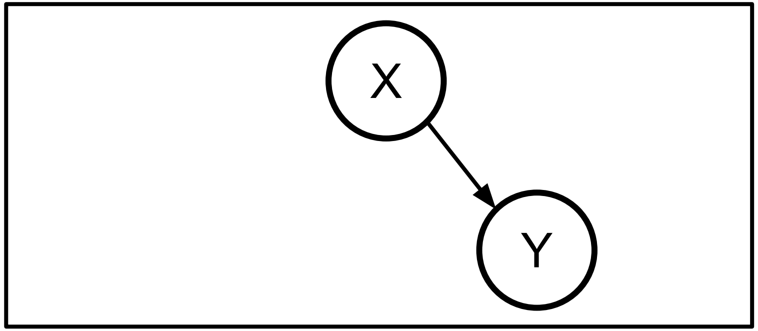
$$f_X : X = U_X$$

$$f_Y : Y = 4x + U_Y$$

Define SCM: <**V**, **F**, **U**>
V: observed variables
F: structural equations
U: background variables, with P(**U**)



This SCM implies this DAG:

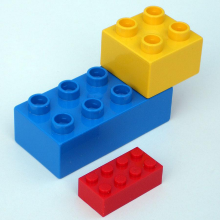


"The building blocks of causal DAGs (directed acyclic graphs)"

Pearl, Causality, 2009

What does the causal graph look like?

If we want to adjust for confounding, we need to ***identify the confounders!***



Define SCM: <**V**, **F**, **U**>
V: observed variables
F: structural equations
U: background variables, with P(**U**)

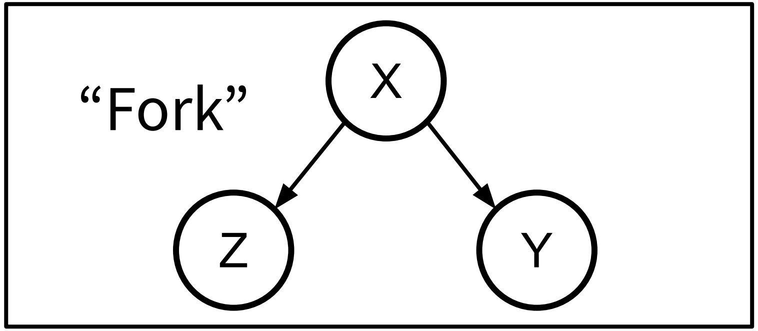
$$V = \{X, Y, Z\}, U = \{U_X, U_Y, U_Z\}, F = \{f_X, f_Y, f_Z\}$$

$$f_X : X = U_X$$

$$f_Y : Y = 4x + U_Y$$

$$f_Z : Z = \frac{x}{10} + U_Z$$

This SCM implies this DAG:



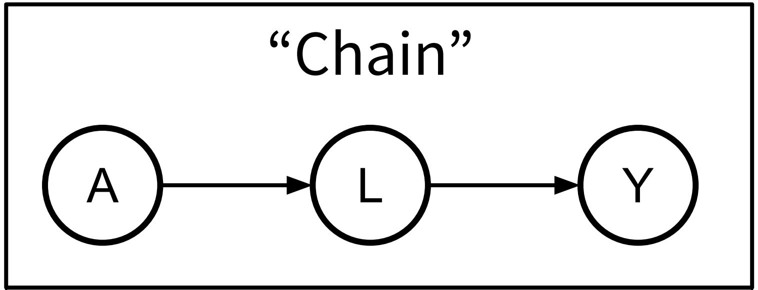
"The building blocks of causal DAGs (directed acyclic graphs)"

Pearl, Causality, 2009

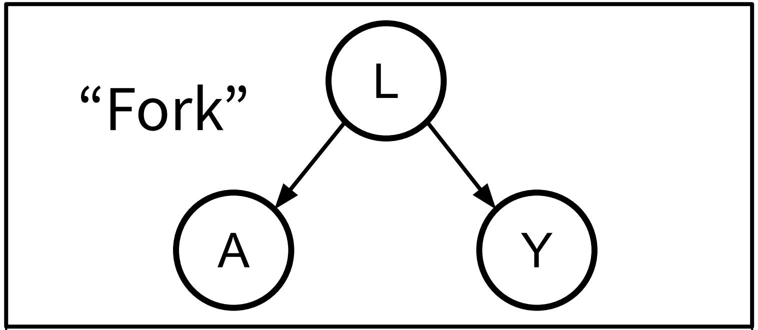
What does the causal graph look like?

If we want to adjust for confounding, we need to ***identify the confounders!***

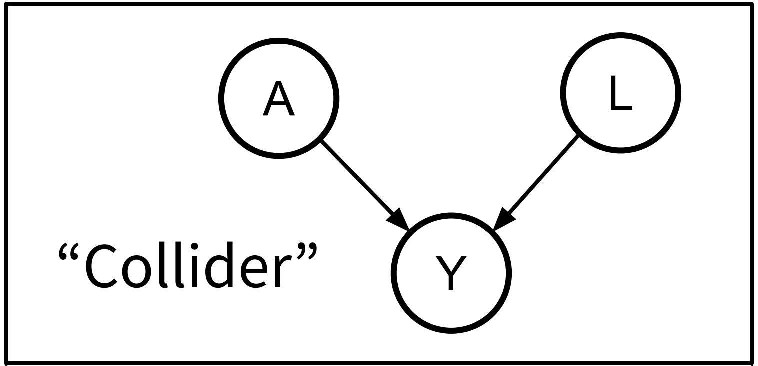
Define SCM: <**V**, **F**, **U**> (which implies a DAG)
V: observed variables
F: structural equations
U: background variables, with P(**U**)



V={A,L,Y}, **F**={f_A(U_A),f_L(A, U_L),f_Y(L, U_Y)}



V={A,L,Y}, **F**={f_A(L,U_A),f_L(U_L),f_Y(L, U_Y)}



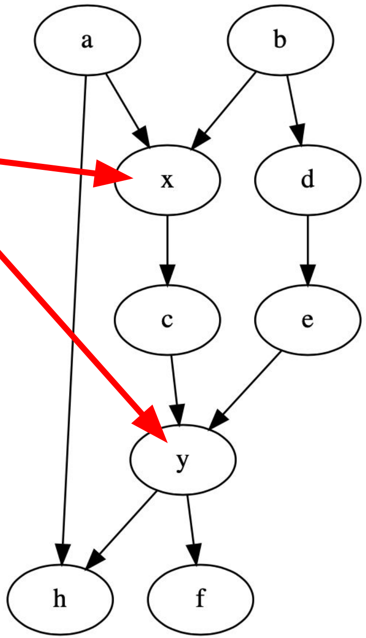
V={A,L,Y}, **F**={f_A(U_A),f_L(U_L),f_Y(L,A,U_Y)}

Pearl, Causality, 2009

Example: Choosing the right adjustment set

Effect of X on Y:

Choosing the correct **adjustment set Z** is critical for **unbiased** effect estimation.

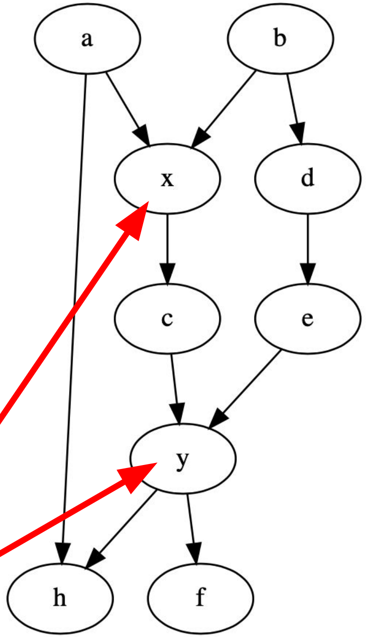


(Ian Barr, http://www.degeneratestate.org/posts/2018/Jul/10/causal-inference-with-python-part-2-causal-graphical-models/)

Example: Choosing the right adjustment set

Effect of X on Y:

Choosing the correct **adjustment set Z** is critical for unbiased effect estimation.

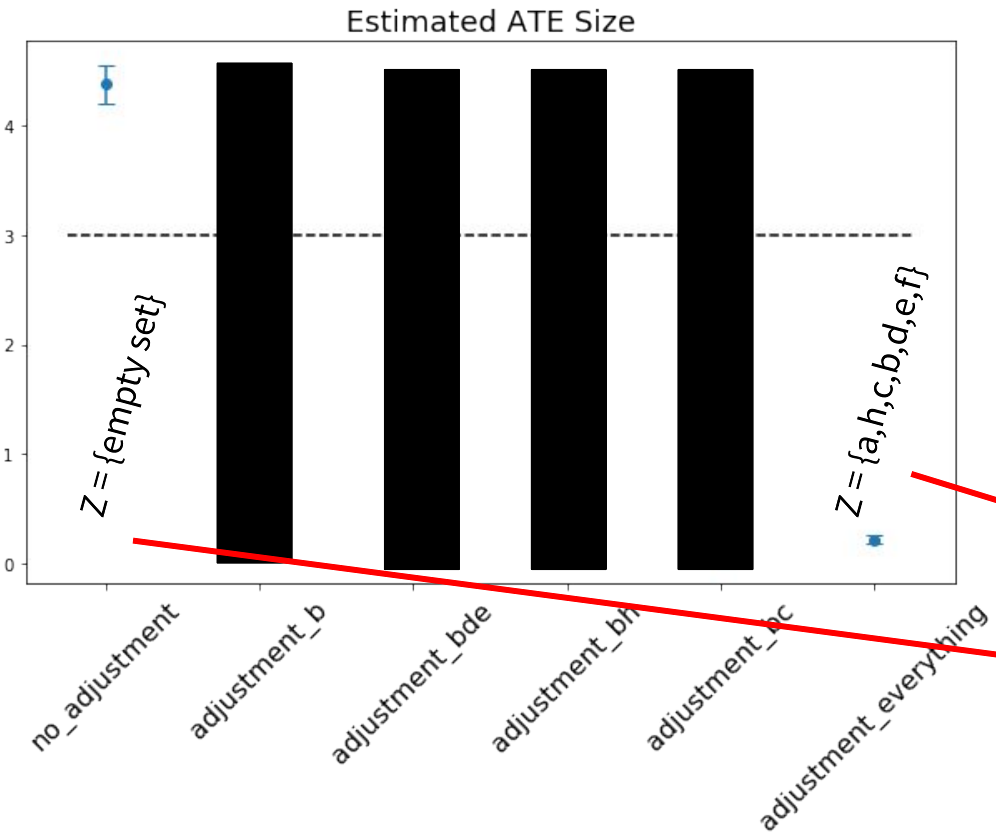


Backdoor adjusted ATE:

$$P(Y = y|do(X = x)) = \sum_z P(Y = y|X = x, \boxed{Z = z)P(Z = z)}$$

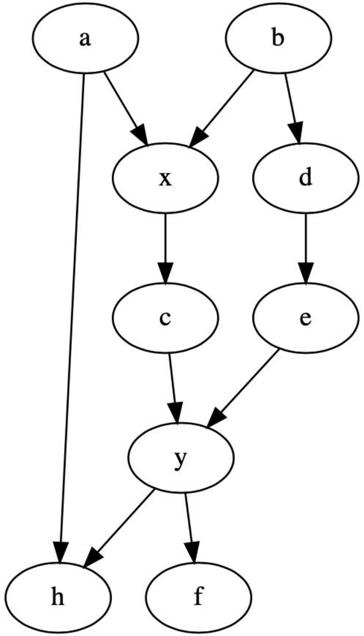
(Ian Barr, http://www.degeneratestate.org/posts/2018/Jul/10/causal-inference-with-python-part-2-causal-graphical-models/)

Example: Choosing the right adjustment set



Effect of X on Y:

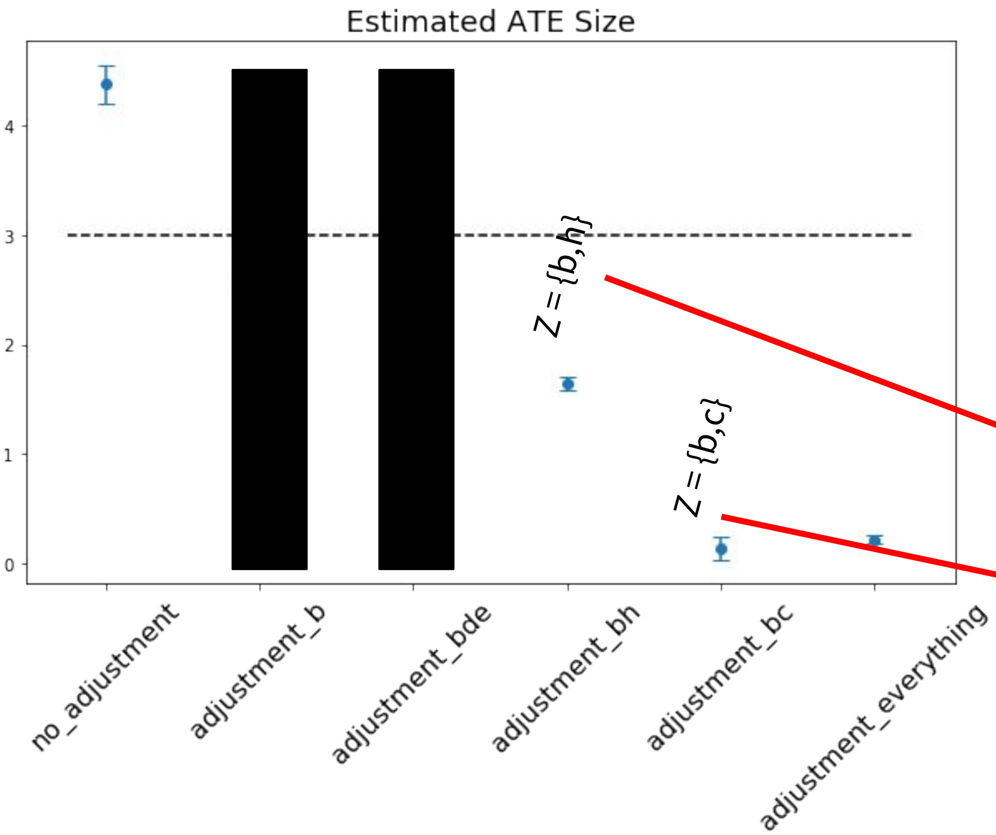
Choosing the correct **adjustment set Z** is critical for unbiased effect estimation.



Backdoor adjusted ATE:

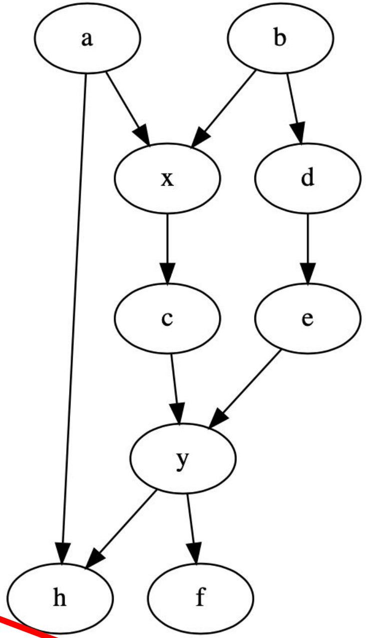
$$P(Y = y|do(X = x)) = \sum_z P(Y = y|X = x, Z = z)P(Z = z)$$

Example: Choosing the right adjustment set



Effect of X on Y:

Choosing the correct **adjustment set Z** is critical for unbiased effect estimation.

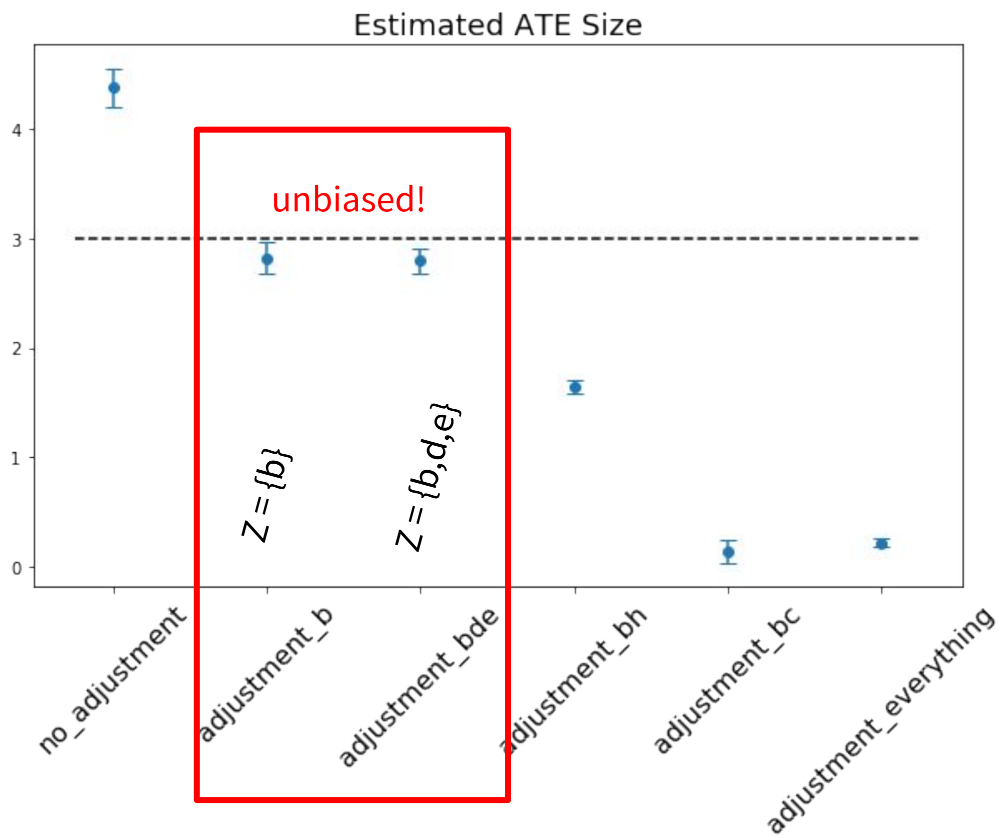


Backdoor adjusted ATE:

$$P(Y = y|do(X = x)) = \sum_z P(Y = y|X = x, Z = z)P(Z = z)$$

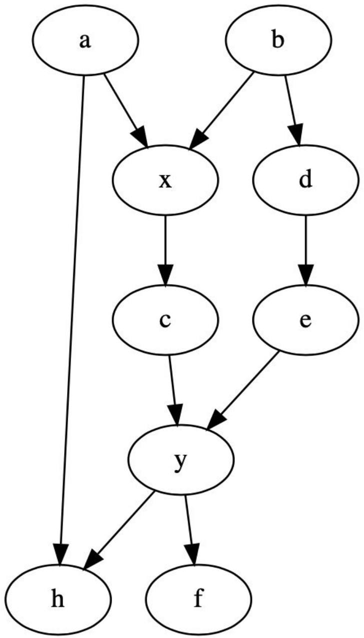
(Ian Barr, http://www.degeneratestate.org/posts/2018/Jul/10/causal-inference-with-python-part-2-causal-graphical-models/)

Example: Choosing the right adjustment set



Effect of X on Y:

Choosing the correct **adjustment set Z** is critical for unbiased effect estimation.

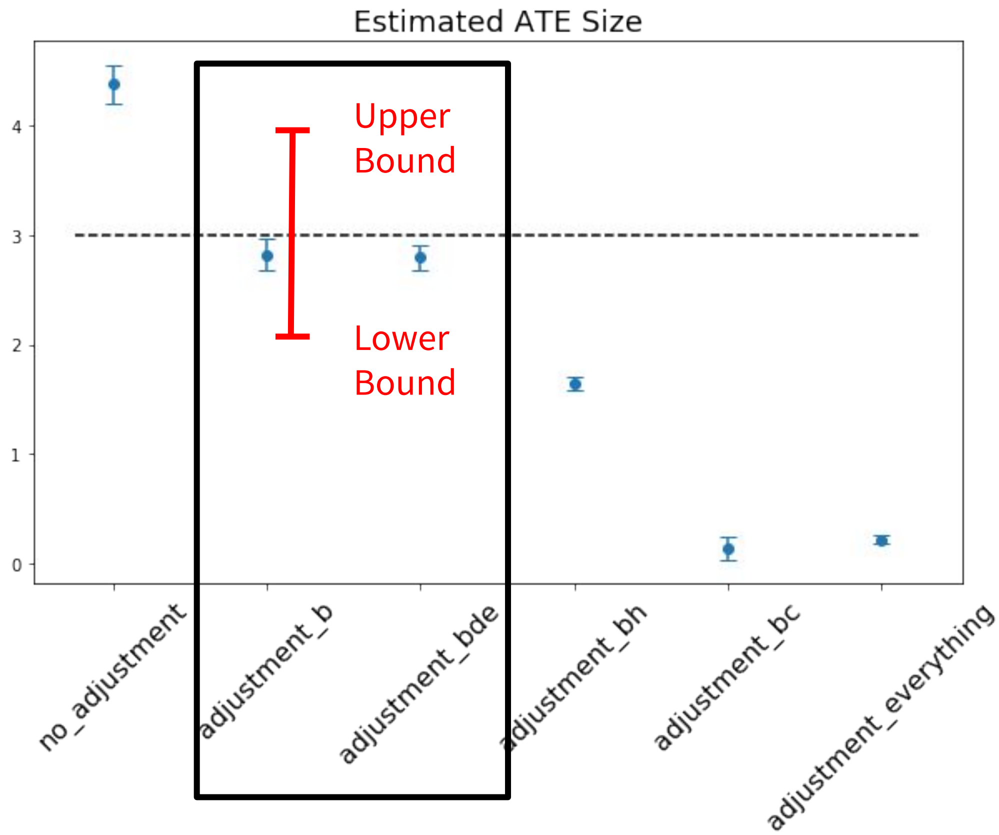


Backdoor adjusted ATE:

$$P(Y = y|do(X = x)) = \sum_z P(Y = y|X = x, Z = z)P(Z = z)$$

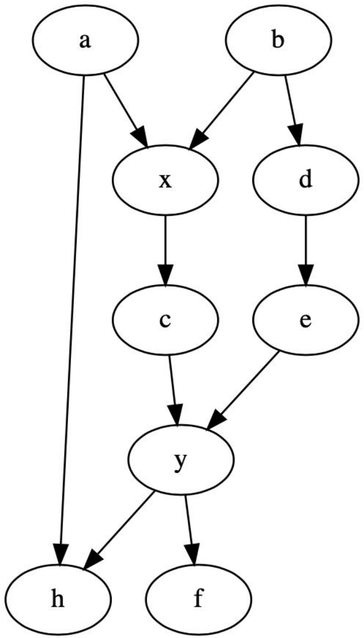
(Ian Barr, http://www.degeneratestate.org/posts/2018/Jul/10/causal-inference-with-python-part-2-causal-graphical-models/)

Example: Choosing the right adjustment set



Effect of X on Y:

Choosing the correct **adjustment set Z** is critical for unbiased effect estimation.



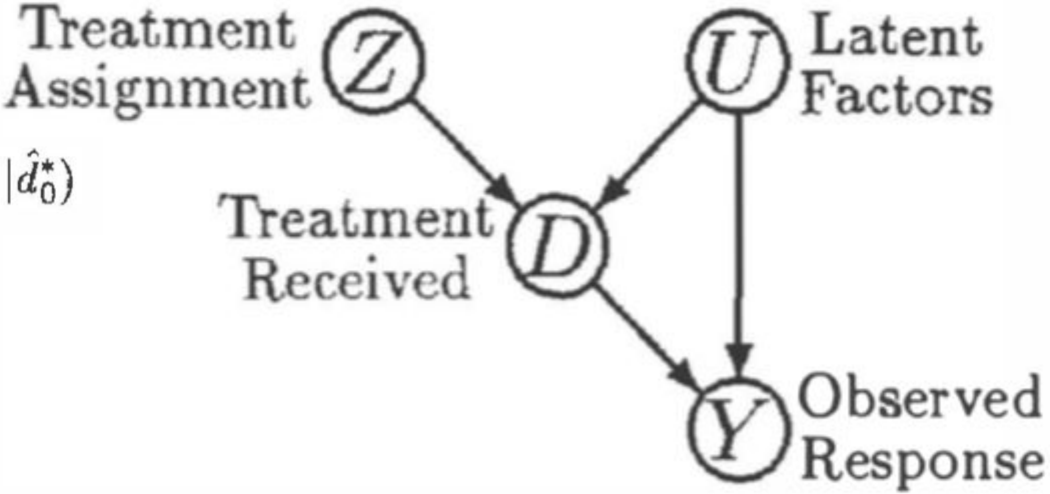
Backdoor adjusted ATE:

$$P(Y = y | do(X = x)) = \sum_z P(Y = y | X = x, Z = z) P(Z = z)$$

(Ian Barr, http://www.degeneratestate.org/posts/2018/Jul/10/causal-inference-with-python-part-2-causal-graphical-models/)

Effect Estimation

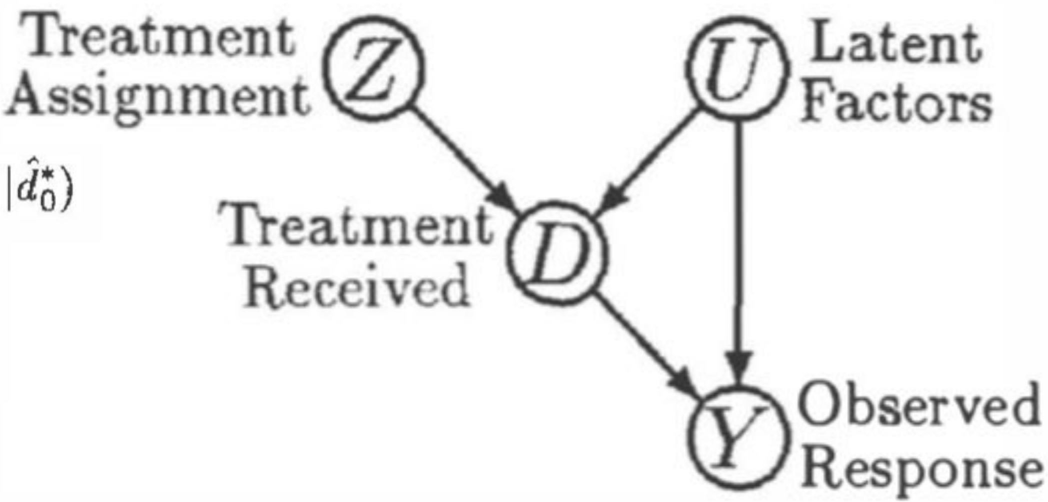
$$\mathrm{ACE}(D \rightarrow Y) = P(y_1^* | \hat{d}_1^*) - P(y_1^* | \hat{d}_0^*)$$



Effect Estimation

$$\mathrm{ACE}(D \rightarrow Y) = P(y_1^*|\hat{d}_1^*) - P(y_1^*|\hat{d}_0^*)$$

$$P(y^*|\hat{d}^*) = \sum_u P(y|d,u)P(u)$$

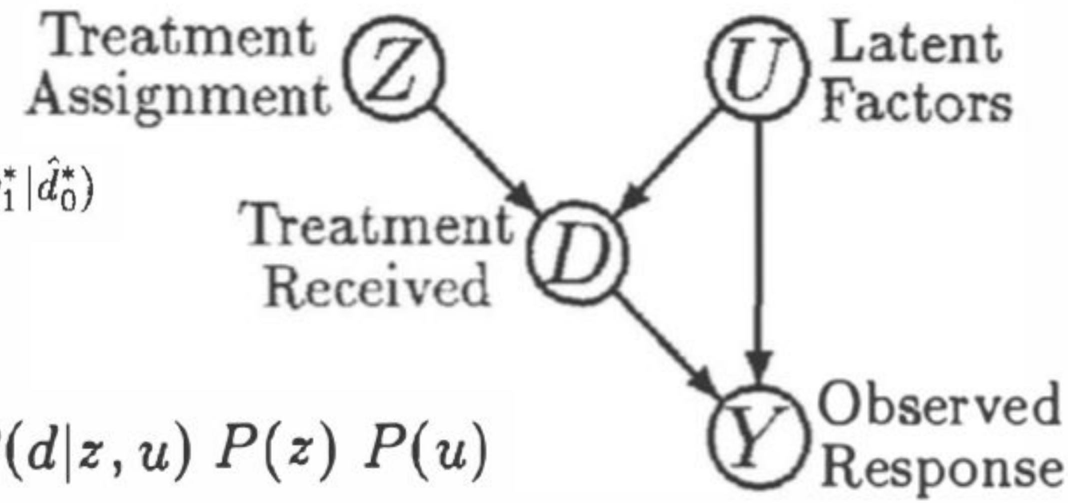


Effect Estimation

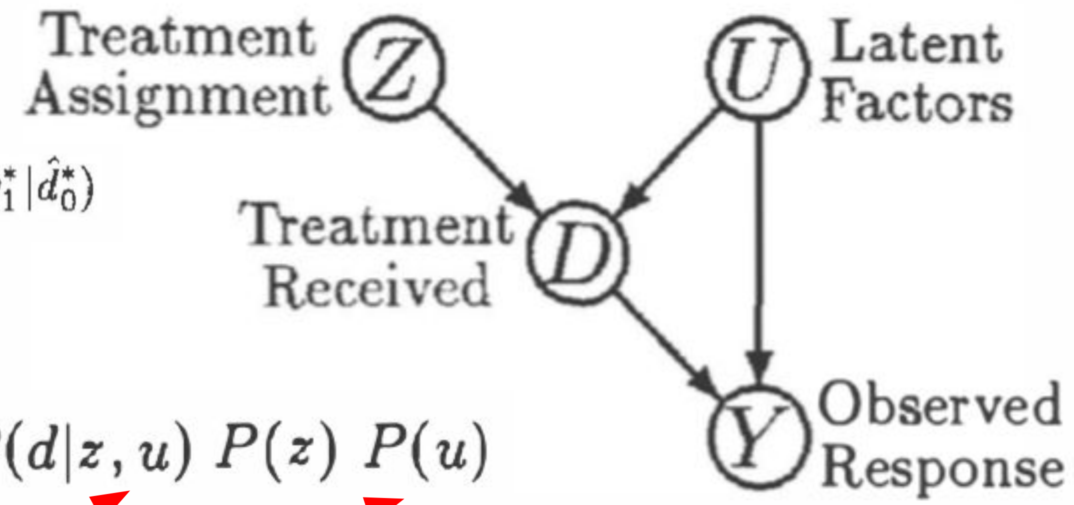
$$\mathrm{ACE}(D \rightarrow Y) = P(y_1^*|\hat{d}_1^*) - P(y_1^*|\hat{d}_0^*)$$

$$P(y^*|\hat{d}^*) = \sum_u P(y|d,u)P(u)$$

$$P(y,d,z,u) = P(y|d,u)\;P(d|z,u)\;P(z)\;P(u)$$



Effect Estimation



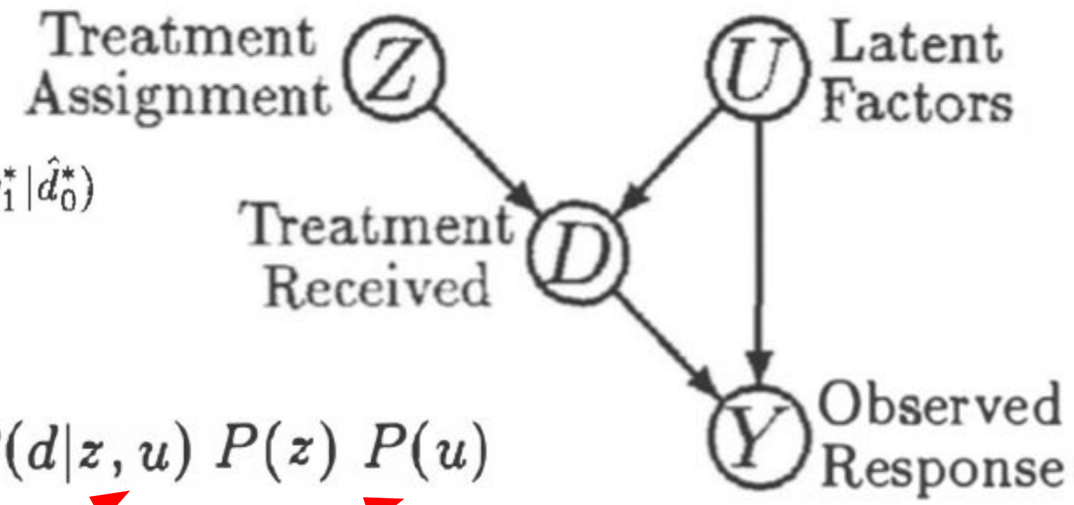
$$\mathrm{ACE}(D \rightarrow Y) = P(y_1^* | \hat{d}_1^*) - P(y_1^* | \hat{d}_0^*)$$

$$P(y^* | \hat{d}^*) = \sum_u P(y|d,u) P(u)$$

$$P(y, d, z, u) = P(y|d, u)\ P(d|z, u)\ P(z)\ P(u)$$

U: unobserved (latent)
Unidentifiable!

Effect Estimation



$$\mathrm{ACE}(D \rightarrow Y) = P(y_1^*|\hat{d}_1^*) - P(y_1^*|\hat{d}_0^*)$$

$$P(y^*|\hat{d}^*) = \sum_u P(y|d,u)P(u)$$

$$P(y,d,z,u) = P(y|d,u)\ P(d|z,u)\ P(z)\ P(u)$$

U: unobserved (latent)
Unidentifiable!

BUT: Observed marginal

$$P(y,d,z)$$

Standard Causal Model

Core Message Alert!!1111

$$v = f_V(\mathrm{pa}_V, u_V)$$

Problem: U_v can be anything: “any type with any domain”

Standard Causal Model

Core Message Alert!!1111

$$v = f_V(\mathrm{pa}_V, u_V)$$

Problem: U_v can be anything: “any type with any domain”

But: “For each u_v of U_v, the functional mapping from PA_v to V is ***particular deterministic response function***.”

Standard Causal Model

Core Message Alert!!1111

$$v = f_V(\mathrm{pa}_V, u_V)$$

Problem: U_v can be anything: “any type with any domain”

But: “For each u_v of U_v, the functional mapping from PA_v to V is ***particular deterministic response function***.”

Consequence: Can map each value of U_v to a deterministic response function.

Reason: Although the domain size of U_V is unknown , which might be very large or even infinite, the number of different deterministic response functions is **known and limited**, given the domain sizes of PA_v and V

Standard Causal Model

Core Message Alert!!1111

$$v = f_V(\mathrm{pa}_V, u_V)$$

Problem: U_v can be anything: “any type with any domain”

But: “For each u_v of U_v, the functional mapping from PA_v to V is ***particular deterministic response function***.”

Consequence: Can map each value of U_v to a deterministic response function.

Reason: Although the domain size of U_V is unknown , which might be very large or even infinite, the number of different deterministic response functions is **known and limited**, given the domain sizes of PA_v and V

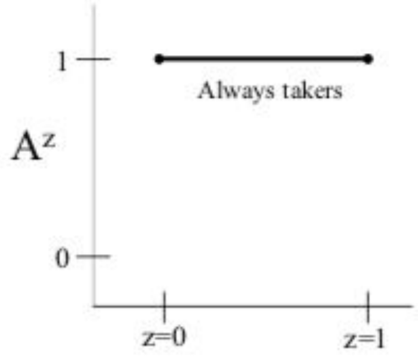


Figure 16.4

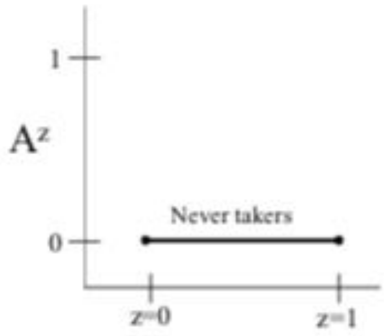


Figure 16.5

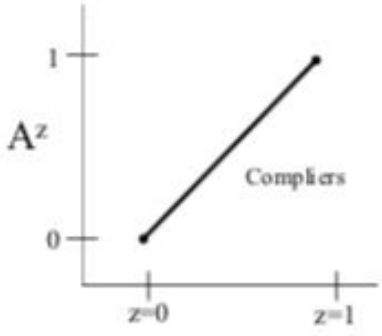


Figure 16.6

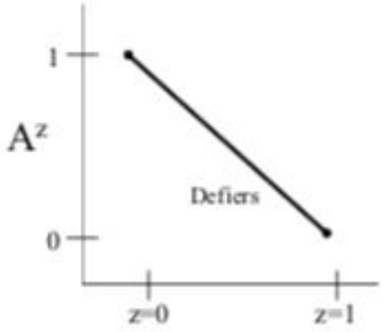
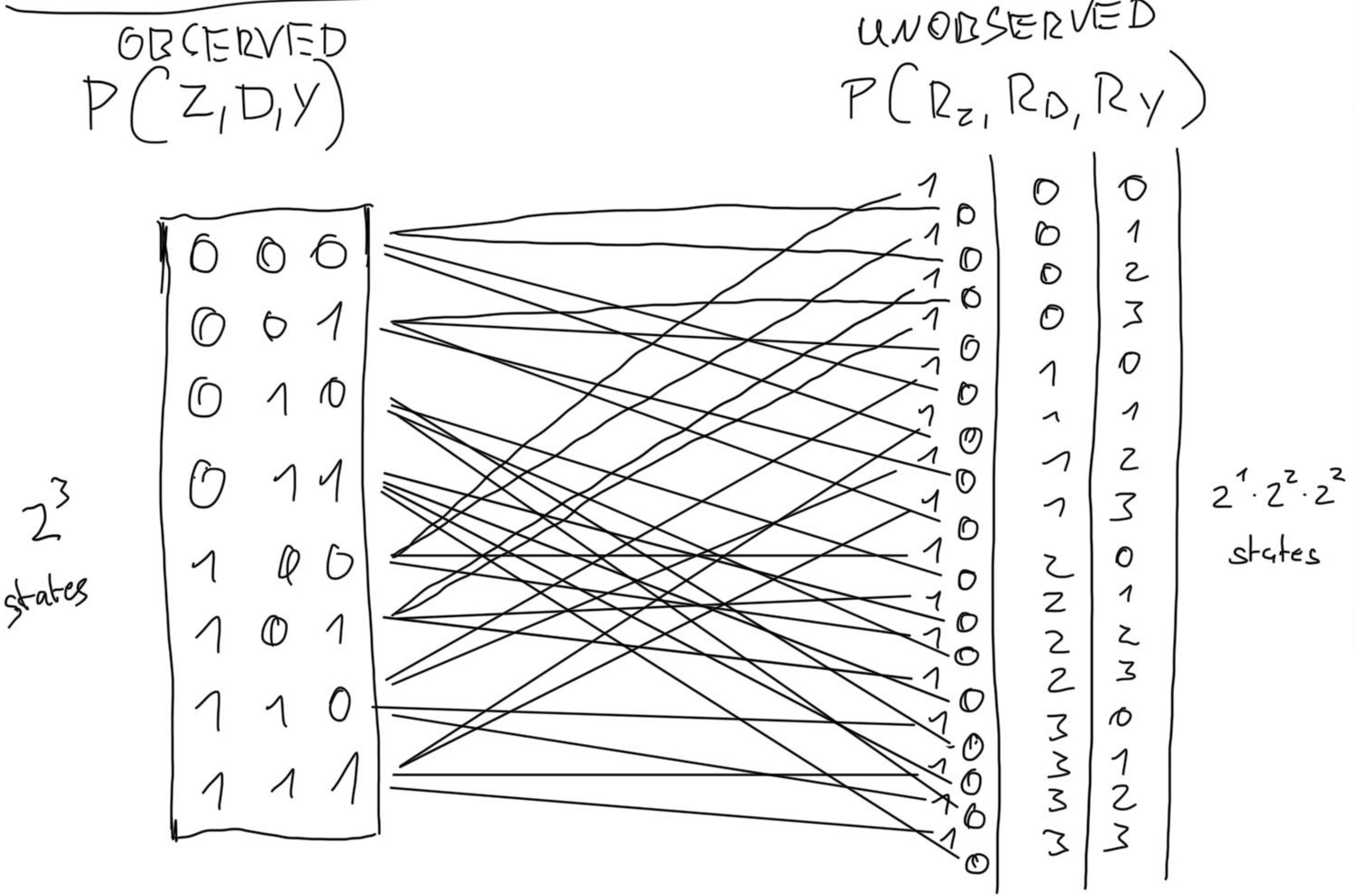


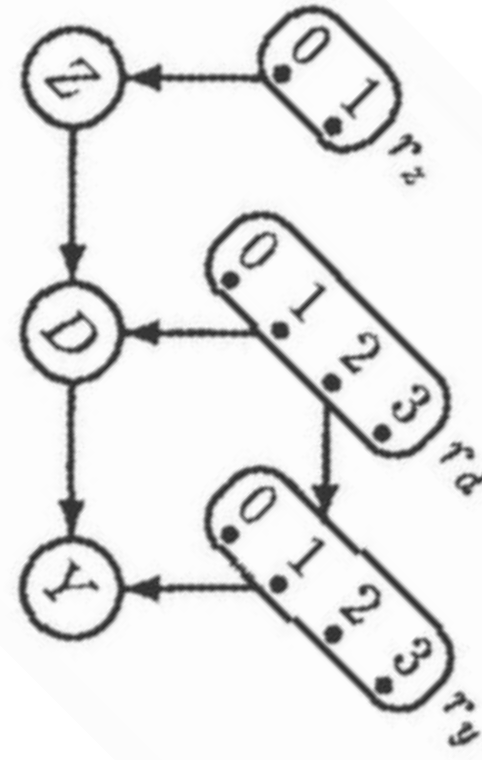
Figure 16.7

$$h_{d,0}(z) = d_0 \quad , \quad h_{d,1}(z) = \begin{cases} d_0 & \text{if } z = z_0 \\ d_1 & \text{if } z = z_1 \end{cases}$$

$$h_{d,3}(z) = d_1 \quad , \quad h_{d,2}(z) = \begin{cases} d_1 & \text{if } z = z_0 \\ d_0 & \text{if } z = z_1 \end{cases}$$

From P(V) to P(R): A visual representation





"Replacing the U with the R"

$$f_V(\mathsf{pa}_V, u_V) = f_V(\mathsf{pa}_V, \ell_V^{-1}(r_V)) = f_V \circ \ell_V^{-1}(\mathsf{pa}_V, r_V) = g_V(\mathsf{pa}_V, r_V),$$

where g_V is the composition of f_V and ℓ_V^{-1}

and denotes the response functions represented by r_V.

$$\mathbb{I}(v; \mathsf{pa}_V, r_V) = \begin{cases} 1 & \text{if } g_V(\mathsf{pa}_V, r_V) = v, \\ 0 & \text{otherwise,} \end{cases}$$

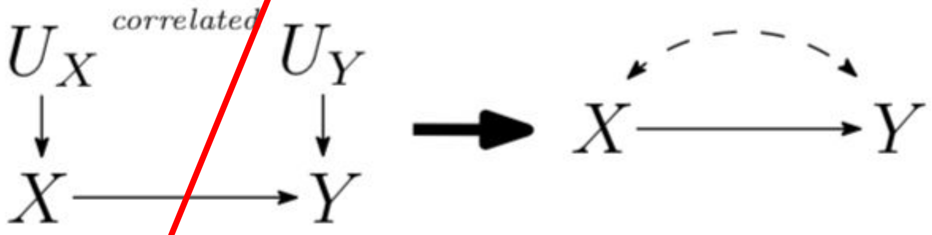
Different Formulations: *All the same stuff*

$$f_{W_i}(\mathbf{r}) = f_{W_i}(f_{W_{i1}}(\mathbf{r}), \ldots, f_{W_{ik_i}}(\mathbf{r}), r_{W_i}).$$

$$\begin{aligned} x_i &= f_{x_i}(\mathbf{r}) \\ &= f_{x_i}(f_{u_1}(\mathbf{r}), f_{u_2}(\mathbf{r}), \ldots, f_{u_k}(\mathbf{r}), r_{x_i}) \end{aligned}$$

$$f_V(\mathsf{pa}_V, u_V) = f_V(\mathsf{pa}_V, \ell_V^{-1}(r_V)) = f_V \circ \ell_V^{-1}(\mathsf{pa}_V, r_V) = g_V(\mathsf{pa}_V, r_V),$$

$$\begin{aligned} \text{min/max} \quad & \boxed{\text{Objective: e.g. TCE}} \\ \text{s.t.} \quad & P(\mathbf{V}) = P(\mathcal{D}), \quad \sum_{\mathbf{r}} P(\mathbf{r}) = 1, \quad P(\mathbf{r}) \geq 0, \end{aligned}$$



Example:

$$P(x, y) = \sum_{r_X, r_Y} P(r_X, r_Y)\mathbb{I}(x; r_X)\mathbb{I}(y; x, r_Y).$$

$$\text{TCE}(x_1, x_0) = \sum_{r_X, r_Y} P(r_X, r_Y)\mathbb{I}(y; x_1, r_Y) - \sum_{r_X, r_Y} P(r_X, r_Y)\mathbb{I}(y; x_0, r_Y),$$

Linear Programming

$$\begin{aligned} \text{min/max} \quad & P(\hat{y}_{s_1|\pi, s_0|\bar{\pi}} | \mathbf{o}) - P(\hat{y}_{s_0} | \mathbf{o}), \\ \text{s.t.} \quad & P(\mathbf{V}) = P(\mathcal{D}), \quad \sum_{\mathbf{r}} P(\mathbf{r}) = 1, \quad P(\mathbf{r}) \geq 0, \end{aligned}$$

Wu et al. (2019): The key idea is to

1. *parameterize the causal model using so-called response-function variables, whose distribution* ***captures all randomness encoded in the causal model****,*
2. *so that we can explicitly* ***traverse all possible causal models*** *to find the tightest possible bounds.*

Linear Programming

$$\begin{aligned} \text{min/max} \quad & P(\hat{y}_{s_1|\pi, s_0|\bar{\pi}}|\mathbf{o}) - P(\hat{y}_{s_0}|\mathbf{o}), \\ \text{s.t.} \quad & P(\mathbf{V}) = P(\mathcal{D}), \quad \sum_{\mathbf{r}} P(\mathbf{r}) = 1, \quad P(\mathbf{r}) \geq 0, \end{aligned}$$

Wu et al. (2019): The key idea is to

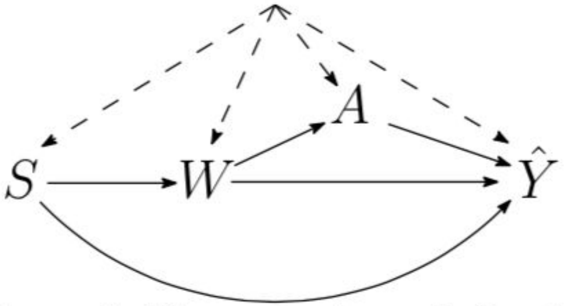
1. *parameterize the causal model using so-called response-function variables, whose distribution **captures all randomness encoded in the causal model**,*
2. *so that we can explicitly **traverse all possible causal models** to find the tightest possible bounds.*

Recap: Unidentifiable situation means that there exist two causal models which exactly agree with the same observational distribution.

Beyond Simple Effects

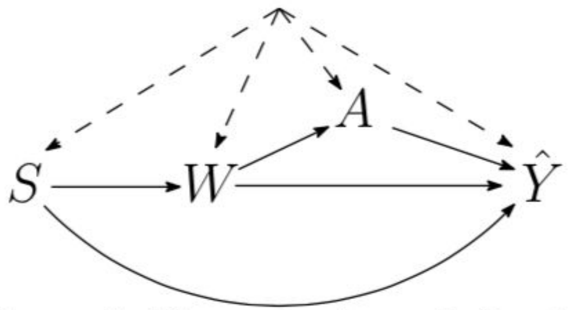
Everything you can imagine

$$\mathrm{TCE}(x_1, x_0) = P(y_{x_1}) - P(y_{x_0}).$$



$$\mathrm{TCE}(x_1, x_0) = P(y_{x_1}) - P(y_{x_0}).$$

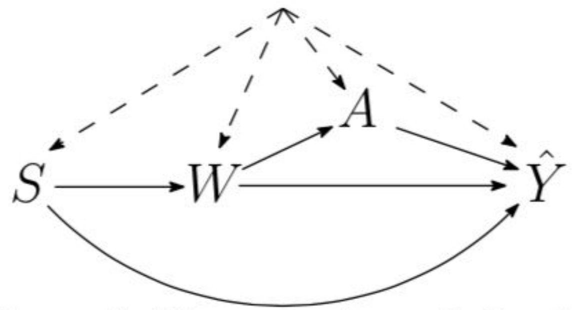
$$\mathrm{PE}_\pi(x_1, x_0) = P(y_{x_1|\pi, x_0|\bar{\pi}}) - P(y_{x_0}),$$



$$\mathrm{TCE}(x_1, x_0) = P(y_{x_1}) - P(y_{x_0}).$$

$$\mathrm{PE}_{\pi}(x_1, x_0) = P(y_{x_1|\pi, x_0|\bar{\pi}}) - P(y_{x_0}),$$

$$\mathrm{CE}(x_1, x_0|\mathbf{o}) = P(y_{x_1}|\mathbf{o}) - P(y_{x_0}|\mathbf{o}).$$

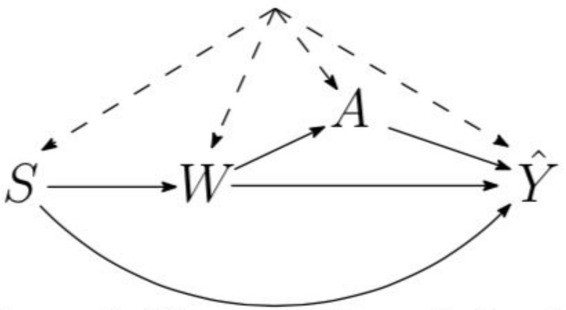


$$\mathrm{TCE}(x_1, x_0) = P(y_{x_1}) - P(y_{x_0}).$$

$$\mathrm{PE}_\pi(x_1, x_0) = P(y_{x_1|\pi, x_0|\bar{\pi}}) - P(y_{x_0}),$$

$$\mathrm{CE}(x_1, x_0|\mathbf{o}) = P(y_{x_1}|\mathbf{o}) - P(y_{x_0}|\mathbf{o}).$$

$$\mathrm{PCE}_\pi(x_1, x_0|\mathbf{o}) = P(y_{x_1|\pi, x_0|\bar{\pi}}|\mathbf{o}) - P(y_{x_0}|\mathbf{o}).$$



$$\pi = \{S \to W \to A \to \hat{Y}, \; S \to \hat{Y}\}$$

$$\mathrm{TCE}(x_1, x_0) = P(y_{x_1}) - P(y_{x_0}).$$

$$\mathrm{PE}_\pi(x_1, x_0) = P(y_{x_1|\pi, x_0|\bar{\pi}}) - P(y_{x_0}),$$

$$\mathrm{CE}(x_1, x_0|\mathbf{o}) = P(y_{x_1}|\mathbf{o}) - P(y_{x_0}|\mathbf{o}).$$

$$\mathrm{PCE}_\pi(x_1, x_0|\mathbf{o}) = P(y_{x_1|\pi, x_0|\bar{\pi}}|\mathbf{o}) - P(y_{x_0}|\mathbf{o}).$$

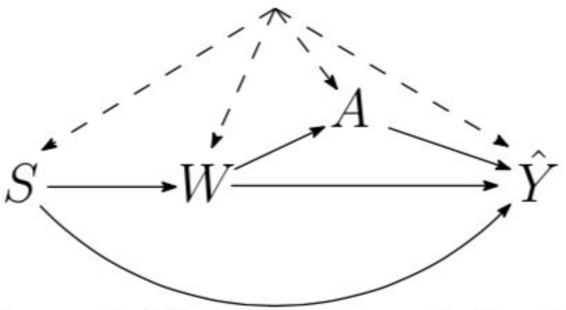


Table 1: Connection between previous fairness notions and PC fairness

NEW! →

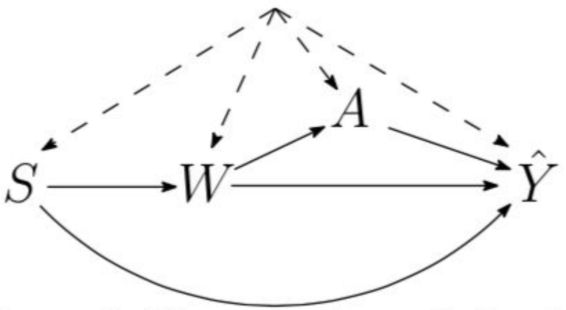
Description	References	Relating to PC fairness
Total effect	[19, 16]	$\mathbf{O} = \emptyset$ and $\pi = \Pi$
(System) Direct discrimination	[19, 7, 16]	$\mathbf{O} = \emptyset$ or $\{S\}$ and $\pi = \pi_d = \{S \to \hat{Y}\}$
(System) Indirect discrimination	[19, 7, 16]	$\mathbf{O} = \emptyset$ or $\{S\}$ and $\pi = \pi_i \subset \Pi$
Individual direct discrimination	[17]	$\mathbf{O} = \{S, \mathbf{X}\}$ and $\pi = \pi_d = \{S \to \hat{Y}\}$
Group direct discrimination	[18]	$\mathbf{O} = \mathbf{Q} = \mathsf{PA}_Y \backslash \{S\}$ and $\pi = \pi_d = \{S \to \hat{Y}\}$
Counterfactual fairness	[5, 9, 14]	$\mathbf{O} = \{S, \mathbf{X}\}$ and $\pi = \Pi$
Counterfactual error rate	[15]	$\mathbf{O} = \{S, Y\}$ and $\pi = \pi_d$ or π_i

$$\mathrm{TCE}(x_1, x_0) = P(y_{x_1}) - P(y_{x_0}).$$

$$\mathrm{PE}_\pi(x_1, x_0) = P(y_{x_1|\pi, x_0|\bar{\pi}}) - P(y_{x_0}),$$

$$\mathrm{CE}(x_1, x_0|\mathbf{o}) = P(y_{x_1}|\mathbf{o}) - P(y_{x_0}|\mathbf{o}).$$

$$\mathrm{PCE}_\pi(x_1, x_0|\mathbf{o}) = P(y_{x_1|\pi, x_0|\bar{\pi}}|\mathbf{o}) - P(y_{x_0}|\mathbf{o}).$$



Choice of O and pi determines effect! Just plug in into PCE_pi

Table 1: Connection between previous fairness notions and PC fairness

NEW! —>

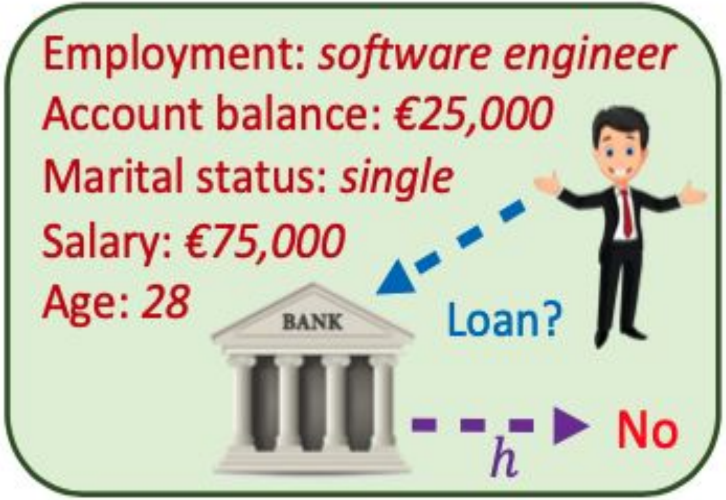
Description	References	Relating to PC fairness
Total effect	[19, 16]	$\mathbf{O} = \emptyset$ and $\pi = \Pi$
(System) Direct discrimination	[19, 7, 16]	$\mathbf{O} = \emptyset$ or $\{S\}$ and $\pi = \pi_d = \{S \to \hat{Y}\}$
(System) Indirect discrimination	[19, 7, 16]	$\mathbf{O} = \emptyset$ or $\{S\}$ and $\pi = \pi_i \subset \Pi$
Individual direct discrimination	[17]	$\mathbf{O} = \{S, \mathbf{X}\}$ and $\pi = \pi_d = \{S \to \hat{Y}\}$
Group direct discrimination	[18]	$\mathbf{O} = \mathbf{Q} = \mathrm{PA}_Y \backslash \{S\}$ and $\pi = \pi_d = \{S \to \hat{Y}\}$
Counterfactual fairness	[5, 9, 14]	$\mathbf{O} = \{S, \mathbf{X}\}$ and $\pi = \Pi$
Counterfactual error rate	[15]	$\mathbf{O} = \{S, Y\}$ and $\pi = \pi_d$ or π_i

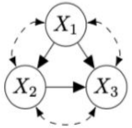
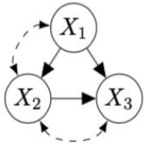
Partial Identification: **Applications**

How is partial identification ***useful?***

Algorithmic Recourse

Scenario: Joe wants to know what he needs to change to get his loan approved.



Graph	A 	B 
Confounding	Full	Partial
Recourse	Bounds (LB_{FC},UB_{FC})	Bounds (LB_{PC},UB_{PC})
Example	Joe's loan application will cross the acceptance threshold if his salary increases above €80k.	... if his salary increases above €78k, i.e. LB_{PB} might be tighter due to lack of confounding, 'flipping' earlier.