# A Crash Course in Good and Bad Controls

C. Cinelli, A. Forney, J. Pearl

Federica Adobbati and Martina Saletta

*Final exam for the PhD course "Causal Networks: Learning and Inference"*
A.A. 2020-2021

# Introduction

**Problem:** decide whether adding a variable to a regression equation helps in the estimation of the *average causal effect (ACE)* of a treatment on an outcome

**Good controls**

Variables we should add

**Bad controls**

Variables we should not add

**Neutral controls**

Variables that do not contribute in deciding the causal effect, or contribute only under particular conditions

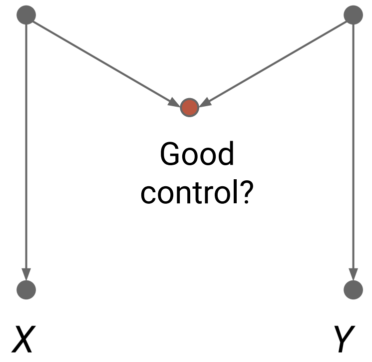
**Proposed solution:** decision based on the graph of the causal model

# Good and Bad Controls

Classical approaches in econometrics do not consider bad controls, discussing only about variables that should not be omitted

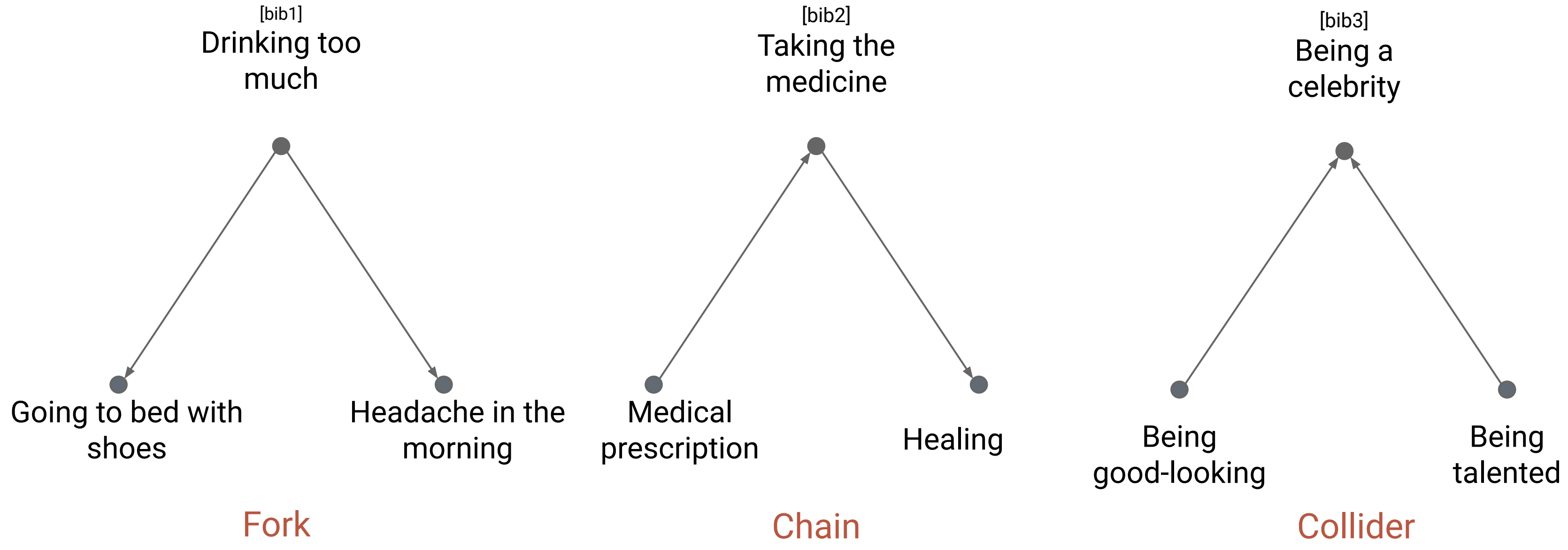
The few works discussing bad controls define them as variables that could be affected by the treatment

Graphical criteria provide us with necessary and sufficient conditions to decide good and bad controls



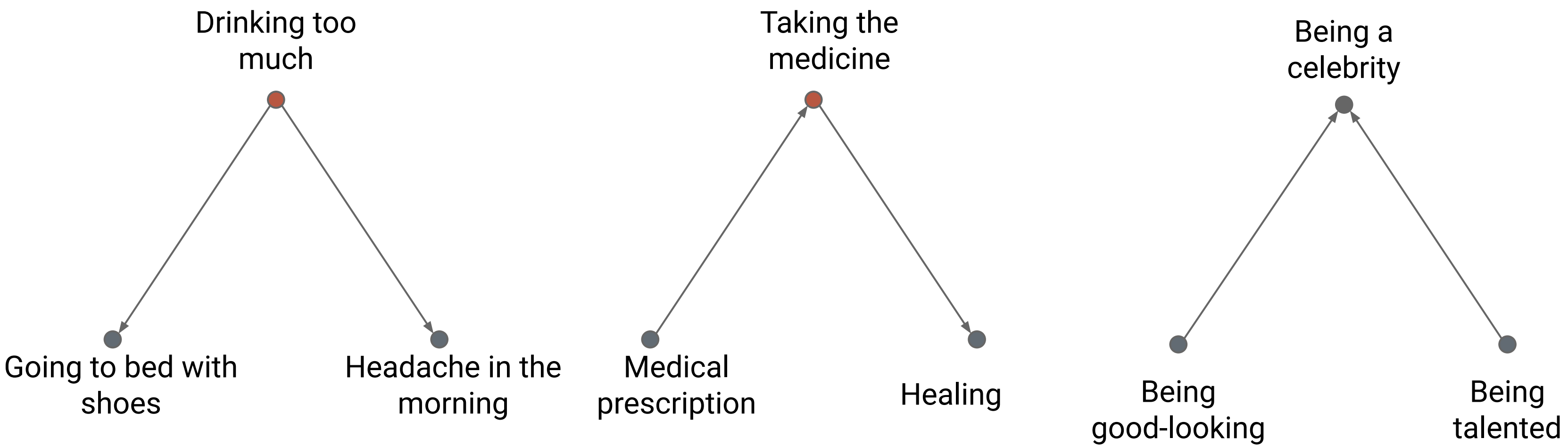
# Flow of association vs flow of causation

Causation flows along directed paths, association along any path without colliders



# Flow of association vs flow of causation

We can block the flow of association by controlling some variables



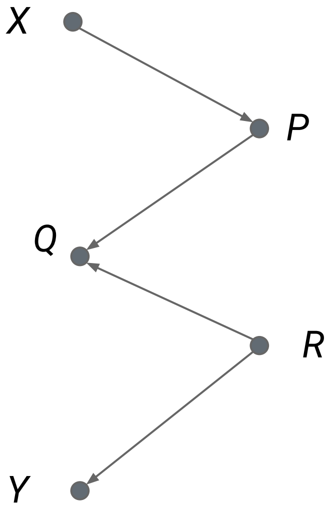

Fork Chain Collider

When we want to estimate the causal effect of a variable on another we need to block the non-causal paths connecting them!

# D-separation

A path is blocked by a set of nodes $\boldsymbol{S}$ if, and only if:

1. it contains a chain of nodes $A \rightarrow B \rightarrow C$ or a fork $A$ $B \leftarrow C$ and $B \in \boldsymbol{S}$;
2. it contains a collider $A \rightarrow B \leftarrow C$ and $B \notin \boldsymbol{S}$

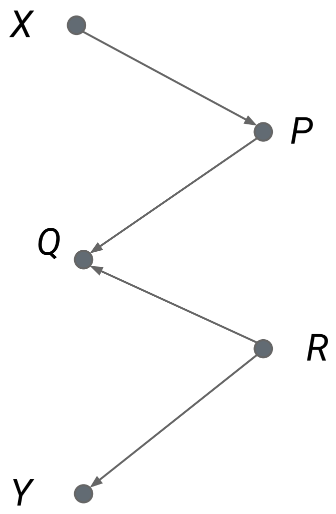


# D-separation

A path is blocked by a set of nodes ***S*** if, and only if:

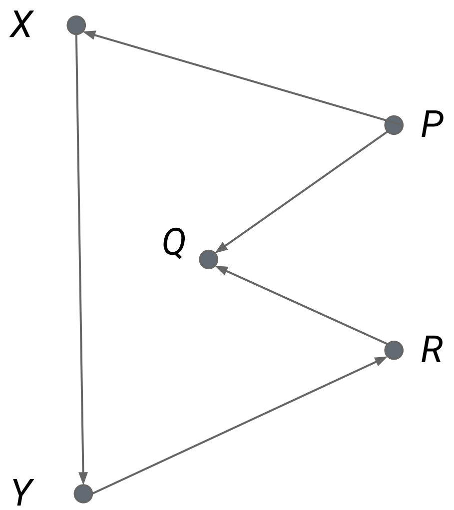
1. it contains a chain of nodes $A \rightarrow B \rightarrow C$ or a fork $A$ $B \leftarrow C$ and $B \in \boldsymbol{S}$;
2. it contains a collider $A \rightarrow B \leftarrow C$ and $B \notin \boldsymbol{S}$

Some sets blocking $X \rightarrow P \rightarrow Q \leftarrow R \rightarrow Y$:
***S*** = ∅; ***S*** = {Q, P}; ***S*** = {Q, R}

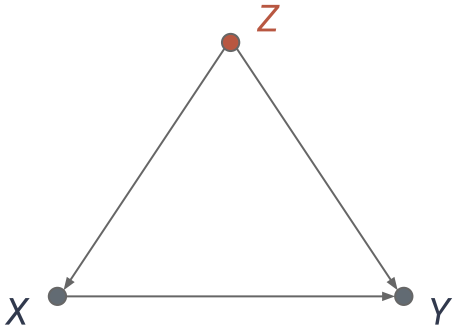


# Backdoor criterion

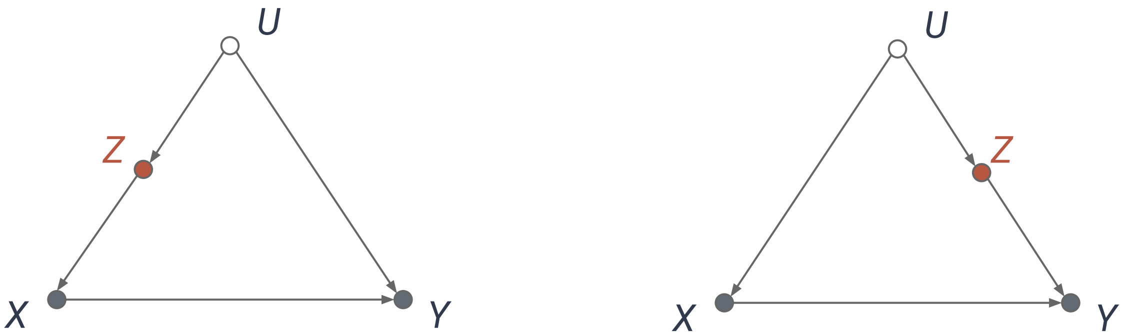
Given an ordered pair of variables *X* and *Y*, a set of variables ***S*** satisfies the backdoor criterion relative to *X* and *Y* if no node in ***S*** is a descendant of *X*, and ***S*** blocks every path between *X* and *Y* that contains an arrow into *X*




Some sets satisfying the backdoor criterion:
***S*** = ∅; ***S*** = {Q, P}; ***S*** = {Q, R}

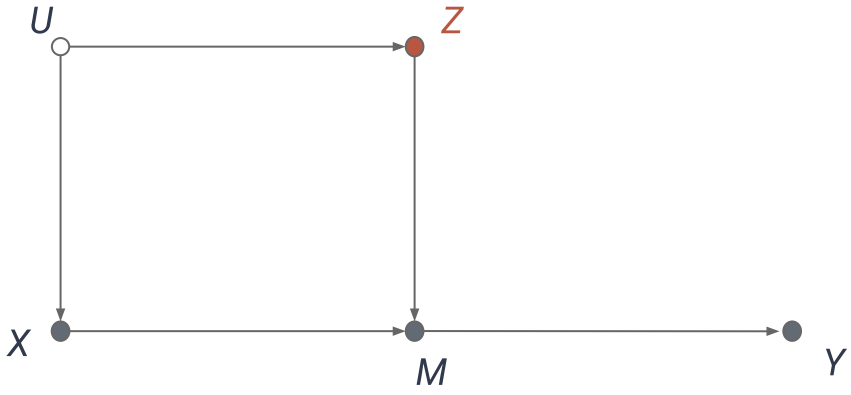


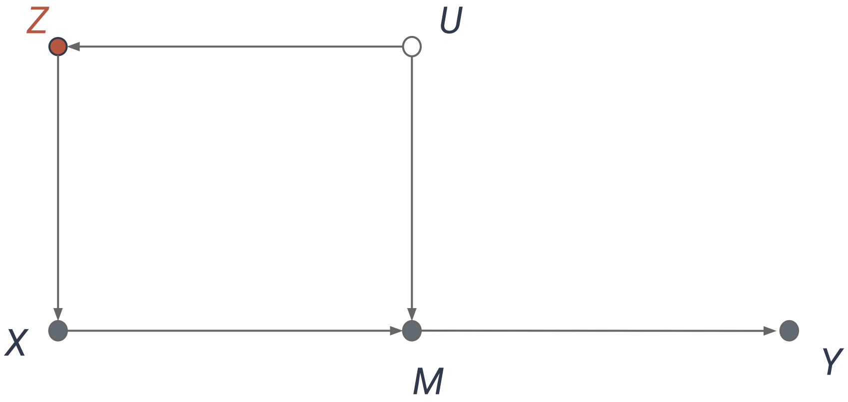

Controlling the common cause

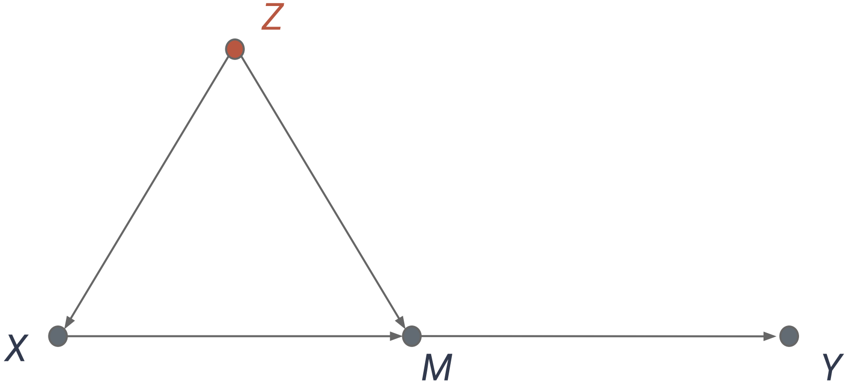



Blocking the backdoor path by indirectly controlling the confounder U

# Good Controls - Blocking the path through the (unobserved) confounder

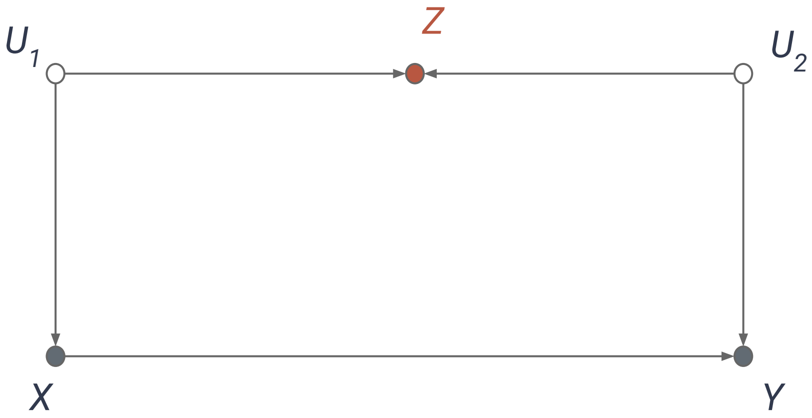



Blocking the backdoor in presence of forks and chains - case with a mediator

# Good Controls – Common causes and mediators

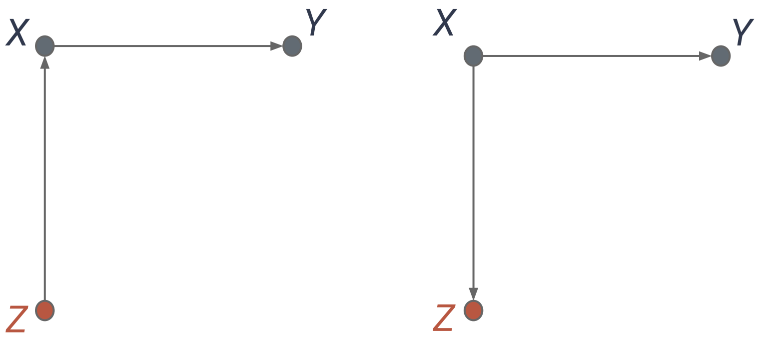


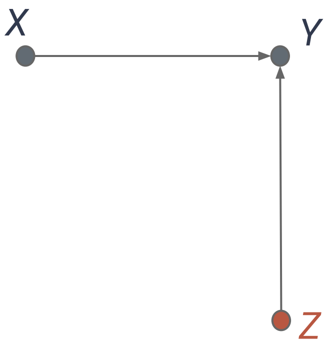

Opening a backdoor path by controlling a collider

Example taken from [bib4]

# Bad Control – Opening a backdoor path

*Z* is not a confounder, thus it does not block nor open any backdoor path.

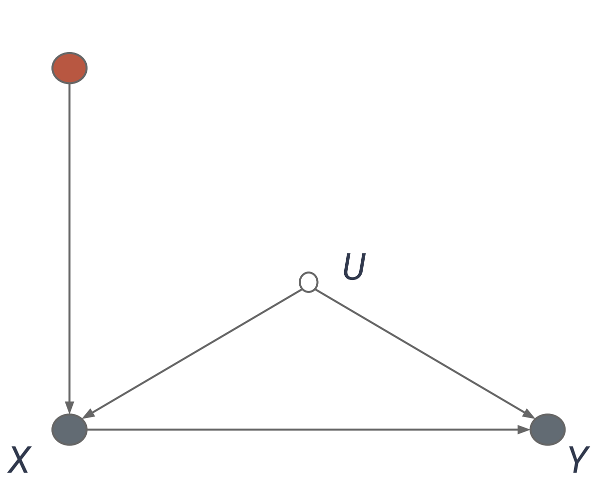



In finite samples *Z* hurts the precision of the ACE estimate, since it reduces the variation of *X*

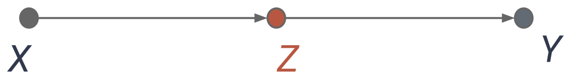
In finite samples *Z* helps improving the ACE estimate, since it reduces the variation of *Y*

Neutral Controls – Reducing the variation of the outcome or the treatment

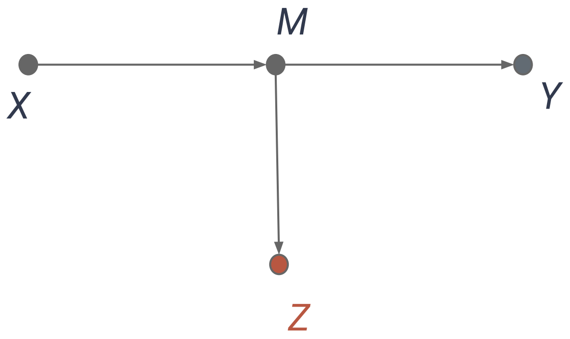



Pre-treatment bad control, since *Z* amplifies the existing bias

All the paths from the treatment to the outcome should be untouched



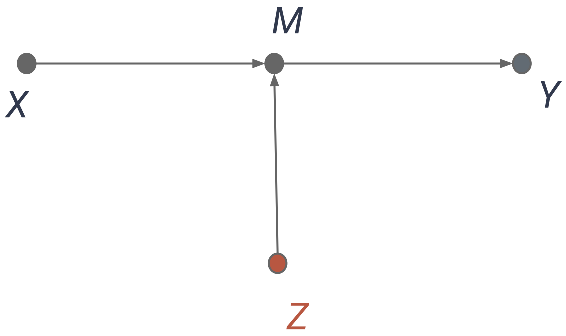

*Z* blocks the effect *Y* we want to estimate




Controlling *Z* is equivalent to partially control the mediator *M*

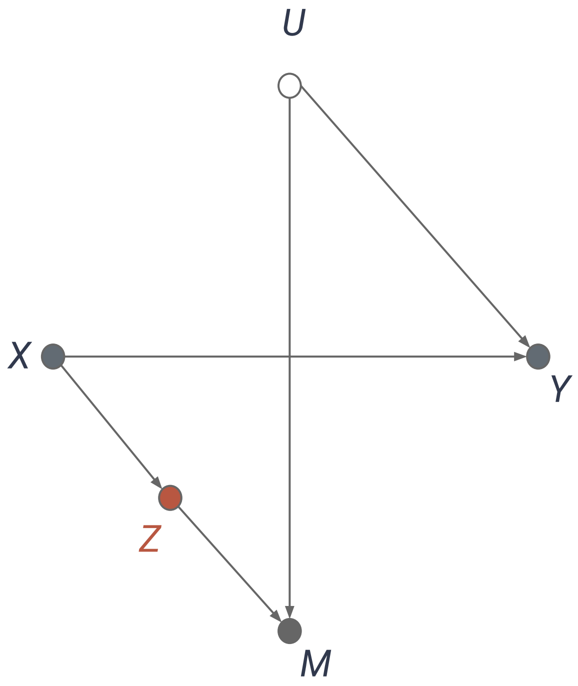
Both the models violate the backdoor criterion

## Bad Controls – (Partially) controlling the mediator




*Z* is a cause of the mediator *M*, and thus of *Y*.
In finite samples, controlling *Z* may increase precision of the ACE estimate

## Neutral Control - Controlling an external cause

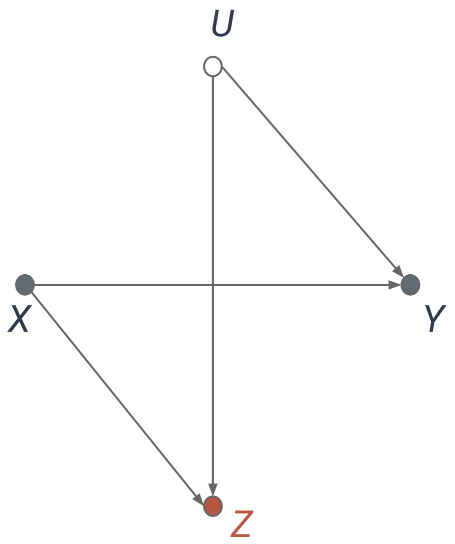



Controlling *Z* does not open any path, thus *Z* is neutral.

May hurt the precision of the ACE estimate in finite samples, since reduces the variation of the treatment *X*.

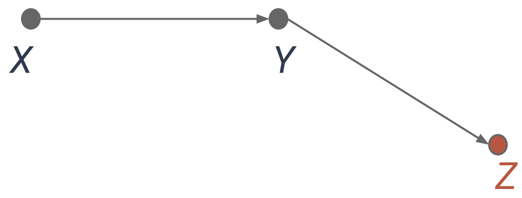
But in case of selection bias (e.g. *M=1*) controlling *Z* could help in obtaining the *M*-specific effect of *X* on *Y.*

## Neutral Control – Possibly helpful in case of selection bias




Controlling $Z$ opens the backdoor path $X \rightarrow Z \leftarrow U \rightarrow Y$

## Bad Control – Opening an undirected path




Controlling for *Z* induces bias in the ACE estimate.

## Bad Control – Selection bias

# Conclusions

Graphical criteria can be used for deciding when a variable should be included in a regression equation.

Some traditional econometrics practices are inaccurate.

Structural knowledge is crucial for stating whether a variable is a good or a bad control.

# References


[bib1] B. Neal. *Introduction to causal inference from a machine learning perspective.* Course lecture notes. 2020.

[bib2] J. Pearl and D. Mackenzie. *The book of why.* Basic books ed. 2018.

[bib3] “Berkson’s paradox” wikipedia page, version of 1st December 2021.

[bib4] F. Elwert and C. Winship. *Endogenous selection bias: the problem of conditioning on a collider variable.* Annual review of sociology. 2014.

*Thank you!*