Causal Inference With Contagion and Latent Homophily Under Full Interference

Yufeng Wu
Advised by Prof. Rohit Bhattacharya

May 13th, 2024

Williams College

Focus of This Thesis

Create new methods that can estimate causal effects from (social) network data.

Focus of This Thesis

Create new methods that can estimate causal effects from (social) network data.

"**Interference**": considers how different rows of data depend on each other in a given dataset.

Motivation

Many causal questions we ask are rooted in social interactions.

Motivation

Many causal questions we ask are rooted in social interactions.

- How effective can flu vaccine protect ourselves and people around us?

Motivation

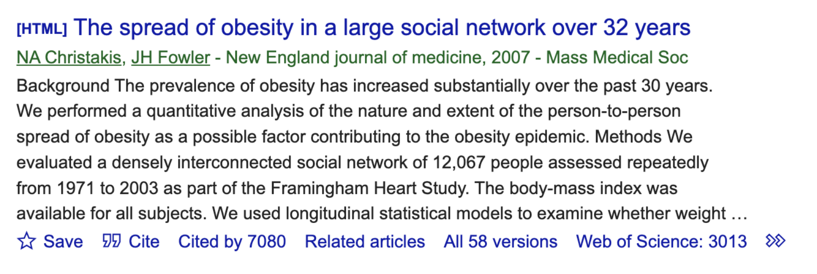
Many causal questions we ask are rooted in social interactions.

- How effective can flu vaccine protect ourselves and people around us?
- If I gain some weight, will it cause my friends to gain weight too?

Motivation

Many causal questions we ask are rooted in social interactions.

- How effective can flu vaccine protect ourselves and people around us?
- If I gain some weight, will it cause my friends to gain weight too?



[HTML] The spread of obesity in a large social network over 32 years

NA Christakis, JH Fowler - New England journal of medicine, 2007 - Mass Medical Soc

Background The prevalence of obesity has increased substantially over the past 30 years. We performed a quantitative analysis of the nature and extent of the person-to-person spread of obesity as a possible factor contributing to the obesity epidemic. Methods We evaluated a densely interconnected social network of 12,067 people assessed repeatedly from 1971 to 2003 as part of the Framingham Heart Study. The body-mass index was available for all subjects. We used longitudinal statistical models to examine whether weight …

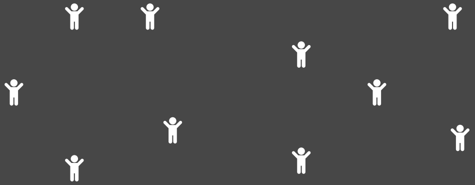
Why Network Data Requires Special Attention?

Why Network Data Requires Special Attention?

i.i.d. = independent and identically distributed

Why Network Data Requires Special Attention?

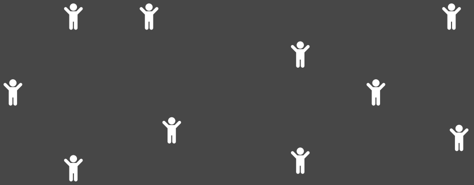
i.i.d. = independent and identically distributed



- People are similar.
- Information from one person cannot predict information of others.

Why Network Data Requires Special Attention?

i.i.d. $=$ independent and identically distributed



- People are similar.
- Information from one person cannot predict information of others. (**almost never true in social networks!**)

Dependent Data Complicates Causal Inference

Dependent Data Complicates Causal Inference

- High variance: estimations are less accurate, but still correct on average (not always a problem.)

Dependent Data Complicates Causal Inference

- High variance: estimations are less accurate, but still correct on average (not always a problem.)
- Bias: incorrect estimation, even with infinite data. (**always a problem!**)

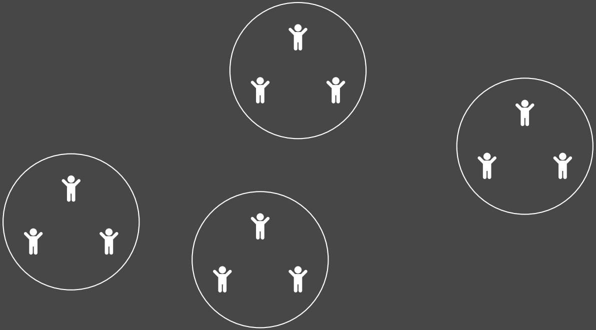
A Convenient Assumption: Partial Interference [1]

Assume i.i.d. “chunks” of data.

[1] Bhattacharya, Malinsky, and Shpitser 2020, Kang and Imbens 2016, Tchetgen and VanderWeele 2012, Hudgens and Halloran 2008

A Convenient Assumption: Partial Interference [1]

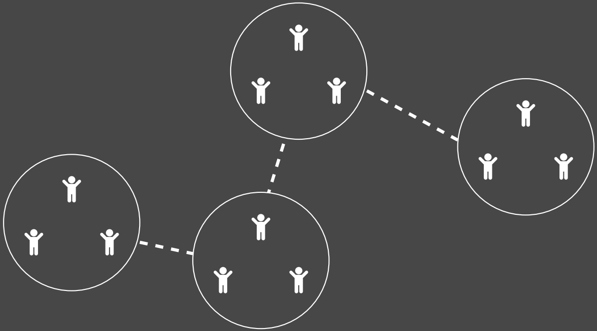
Assume i.i.d. “chunks” of data.



[1] Bhattacharya, Malinsky, and Shpitser 2020, Kang and Imbens 2016, Tchetgen and VanderWeele 2012, Hudgens and Halloran 2008

A Convenient Assumption: Partial Interference [2]

This assumption does not hold in general!



[2] Bhattacharya, Malinsky, and Shpitser 2020, Kang and Imbens 2016, Tchetgen and VanderWeele 2012, Hudgens and Halloran 2008

The More General Setting: Full Interference [3]

Everyone may interfere with anyone else in the network.

[3] Tchetgen Tchetgen, Fulcher, and Shpitser 2021, Tchetgen and VanderWeele 2012

The More General Setting: Full Interference [3]

Everyone may interfere with anyone else in the network.

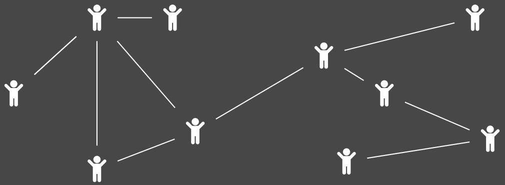
No restrictions on the structure of the network.

[3] Tchetgen Tchetgen, Fulcher, and Shpitser 2021, Tchetgen and VanderWeele 2012

The More General Setting: Full Interference [3]

Everyone may interfere with anyone else in the network.

No restrictions on the structure of the network.



[3] Tchetgen Tchetgen, Fulcher, and Shpitser 2021, Tchetgen and VanderWeele 2012

Why Might Data Be Dependent? [4]

[4] Shalizi and Thomas 2011, Ogburn and VanderWeele 2014, Lauritzen and Richardson 2002, Shpitser 2015

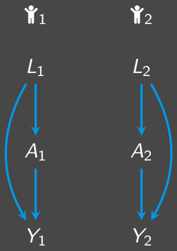
Why Might Data Be Dependent? [4]

$L =$ confounders; $A =$ therapy sessions; $Y =$ job satisfaction.

[4] Shalizi and Thomas 2011, Ogburn and VanderWeele 2014, Lauritzen and Richardson 2002, Shpitser 2015

Why Might Data Be Dependent? [4]

L = confounders; A = therapy sessions; Y = job satisfaction.

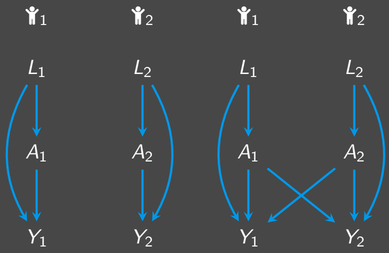


No Interference

[4] Shalizi and Thomas 2011, Ogburn and VanderWeele 2014, Lauritzen and Richardson 2002, Shpitser 2015

Why Might Data Be Dependent? [4]

L = confounders; A = therapy sessions; Y = job satisfaction.



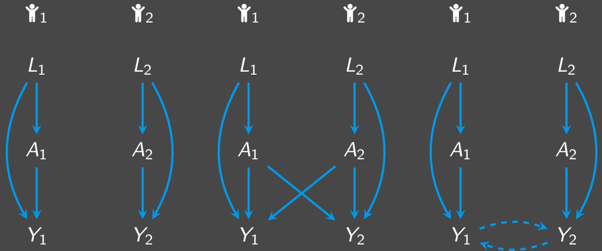
No Interference

Direct Interference

[4] Shalizi and Thomas 2011, Ogburn and VanderWeele 2014, Lauritzen and Richardson 2002, Shpitser 2015

Why Might Data Be Dependent? [4]

L = confounders; A = therapy sessions; Y = job satisfaction.



No Interference

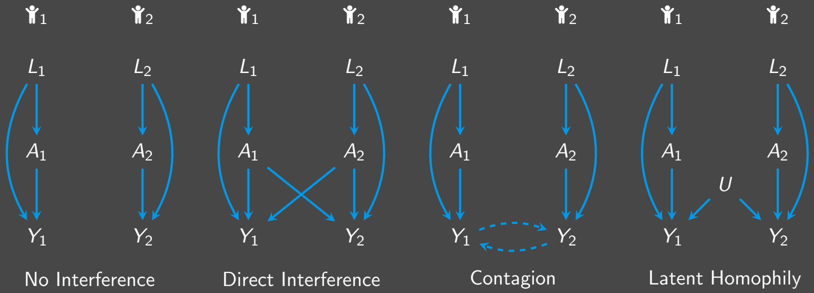
Direct Interference

Contagion

[4] Shalizi and Thomas 2011, Ogburn and VanderWeele 2014, Lauritzen and Richardson 2002, Shpitser 2015

Why Might Data Be Dependent? [4]

L = confounders; A = therapy sessions; Y = job satisfaction.



No Interference

Direct Interference

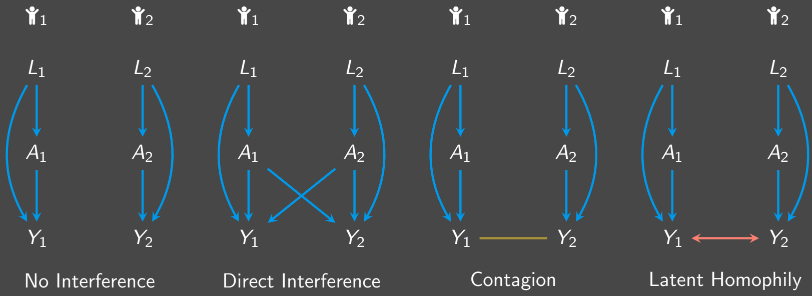
Contagion

Latent Homophily

[4] Shalizi and Thomas 2011, Ogburn and VanderWeele 2014, Lauritzen and Richardson 2002, Shpitser 2015

Why Might Data Be Dependent? [5]

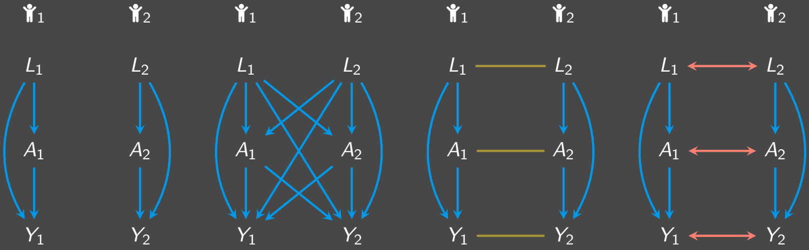
L = confounders; A = therapy sessions; Y = job satisfaction.



[5] Shalizi and Thomas 2011, Ogburn and VanderWeele 2014, Lauritzen and Richardson 2002, Shpitser 2015

Why Might Data Be Dependent? [6]

$L =$ confounders; $A =$ therapy sessions; $Y =$ job satisfaction.



No Interference

Direct Interference

Contagion

Latent Homophily

[6] Shalizi and Thomas 2011, Ogburn and VanderWeele 2014, Lauritzen and Richardson 2002, Shpitser 2015

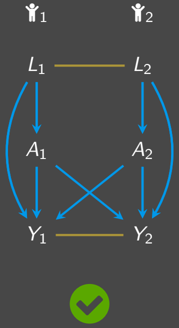
Previous Work (1)

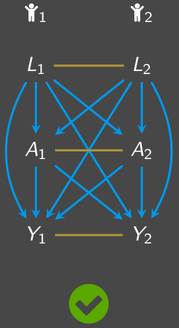
Auto-g computation[7]: can estimate causal effects under full interference, **as long as there is no latent homophily** ($\leftrightarrow$).

[7] Tchetgen Tchetgen, Fulcher, and Shpitser 2021

Previous Work (1)

Auto-g computation[7]: can estimate causal effects under full interference, **as long as there is no latent homophily** ($\leftrightarrow$).

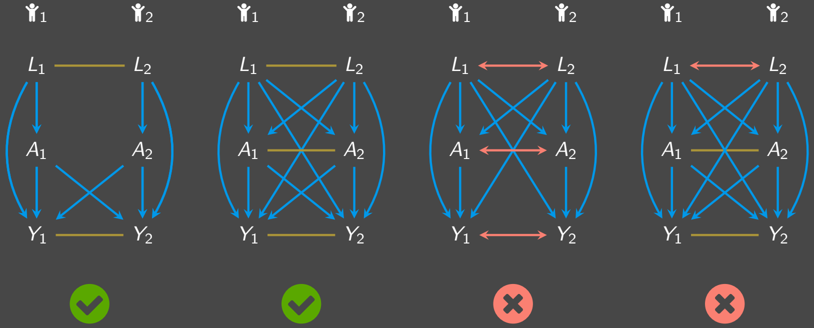




[7] Tchetgen Tchetgen, Fulcher, and Shpitser 2021

Previous Work (1)

Auto-g computation[7]: can estimate causal effects under full interference, **as long as there is no latent homophily** ($\leftrightarrow$).



[7] Tchetgen Tchetgen, Fulcher, and Shpitser 2021

Previous Work (2)

Causal Inference for Social Network Data [8].

- allows for direct interference and latent homophily between individuals.

[8] **ogburn2024causal**

Open Problems

1. A causal effect estimation method that simultaneously accounts for all three mechanisms:

   direct interference (→), contagion (–), and latent homophily (↔).

Open Problems

1. A causal effect estimation method that simultaneously accounts for all three mechanisms:

   direct interference ($\rightarrow$), contagion ($-$), and latent homophily ($\leftrightarrow$).

2. Current methods rely on prior knowledge & belief.

   We want a test to distinguish between contagion ($-$) and latent homophily ($\leftrightarrow$).

Open Problems

1. A causal effect estimation method that simultaneously accounts for all three mechanisms:

   direct interference ($\rightarrow$), contagion ($-$), and latent homophily ($\leftrightarrow$).

2. Current methods rely on prior knowledge & belief.

   We want a test to distinguish between contagion ($-$) and latent homophily ($\leftrightarrow$).

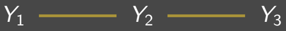
Intuition

Claim: contagion vs. latent homophily is distinguishable using an independence test.

Intuition

Claim: contagion vs. latent homophily is distinguishable using an independence test.

Undirected Edge:

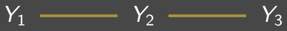


$$Y_1 \not\perp\!\!\!\perp Y_3 \text{ and } Y_1 \perp\!\!\!\perp Y_3 \mid Y_2$$

Intuition

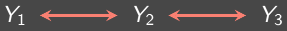
Claim: contagion vs. latent homophily is distinguishable using an independence test.

Undirected Edge:



$$Y_1 \not\perp\!\!\!\perp Y_3 \text{ and } Y_1 \perp\!\!\!\perp Y_3 \mid Y_2$$

Bidirected Edge:

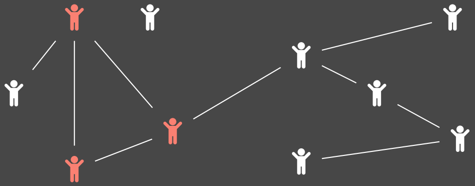


$$Y_1 \perp\!\!\!\perp Y_3 \text{ and } Y_1 \not\perp\!\!\!\perp Y_3 \mid Y_2$$

How To Get i.i.d. Samples for Independence Tests

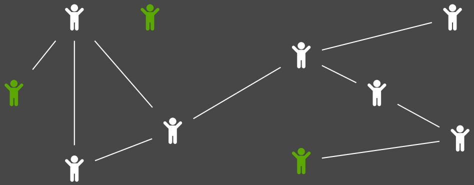
How To Get i.i.d. Samples for Independence Tests

Intuition: further away in network $\approx$ less dependent.



How To Get i.i.d. Samples for Independence Tests

Intuition: further away in network $\approx$ less dependent.

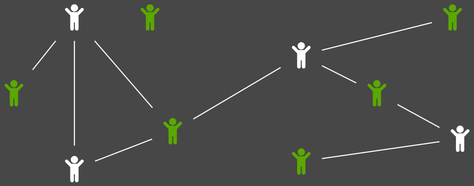


How To Get i.i.d. Samples for Independence Tests

Independent Set: a set of vertices in a graph, no two of which are adjacent.

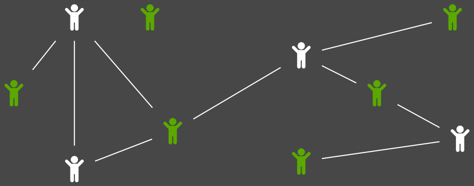
How To Get i.i.d. Samples for Independence Tests

Independent Set: a set of vertices in a graph, no two of which are adjacent.



How To Get i.i.d. Samples for Independence Tests

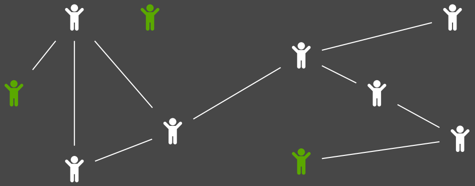
Independent Set: a set of vertices in a graph, no two of which are adjacent.



General version: “k-hop” independent set.

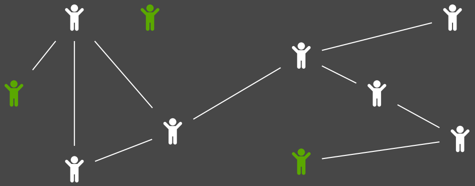
Our Proposed Test

Step 1: find a maximal 5-hop independent set $\mathcal{I}$ from the network.



Our Proposed Test

Step 2: for each person in $\mathcal{I}$, collect information on their neighbors and their 2nd-order neighbors (i.e., neighbors' neighbors).



Our Proposed Test

Step 3: Is $\text{person}_i \perp\!\!\!\perp \text{2nd-order nb}(i) \mid \text{nb}(i)$?

Our Proposed Test

Step 3: Is $\text{person}_i \perp\!\!\!\perp \text{2nd-order nb}(i) \mid \text{nb}(i)$?

Likelihood ratio test:

- Model 1: $\text{person}_i \sim \text{nb}(i)$
- Model 2: $\text{person}_i \sim \text{nb}(i) + \text{2nd-order nb}(i)$

Our Proposed Test

Step 3: Is $\text{person}_i \perp\!\!\!\perp \text{2nd-order nb}(i) \mid \text{nb}(i)$?

Likelihood ratio test:

- Model 1: $\text{person}_i \sim \text{nb}(i)$
- Model 2: $\text{person}_i \sim \text{nb}(i) + \text{2nd-order nb}(i)$

If $\perp\!\!\!\perp$, conclude contagion ($-$).

If $\not\perp\!\!\!\perp$, conclude latent homophily ($\leftrightarrow$).

Evaluating Our Test

Power: how often it correctly detects homophily.

Evaluating Our Test

Power: how often it correctly detects homophily.

Type 1 Error Rate: how often it incorrectly concludes contagion as homophily.

Evaluating Our Test

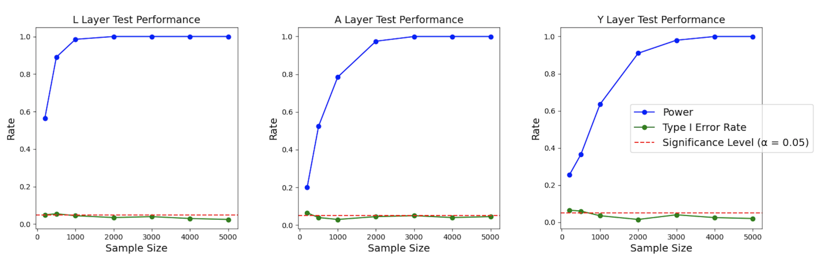
Power: how often it correctly detects homophily.

Approach 1 as sample size increases.

Type 1 Error Rate: how often it incorrectly concludes contagion as homophily.

Less than significance level α.

Evaluating Our Test



Recap

1. A causal effect estimation method that simultaneously accounts for all three mechanisms:

   direct interference (→), contagion (–), and latent homophily (↔).

2. Current methods rely on prior knowledge & belief.

   We want a test to distinguish between contagion (–) and latent homophily (↔).

Intuition

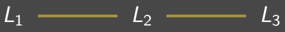
Why do we even need a new method when latent homophily ($\leftrightarrow$) is present?

[9] Lauritzen and Richardson 2002

Intuition

Why do we even need a new method when latent homophily ($\leftrightarrow$) is present?

Undirected Edge:



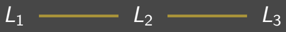
Gibbs factors [9]: $p(L_1 \mid L_2)$, $p(L_2 \mid L_1, L_3)$, and $p(L_3 \mid L_2)$

[9] Lauritzen and Richardson 2002

Intuition

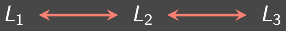
Why do we even need a new method when latent homophily ($\leftrightarrow$) is present?

Undirected Edge:



Gibbs factors [9]: $p(L_1 \mid L_2)$, $p(L_2 \mid L_1, L_3)$, and $p(L_3 \mid L_2)$

Bidirected Edge:



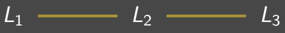
$p(L_1, L_2, L_3)$

[9] Lauritzen and Richardson 2002

Intuition

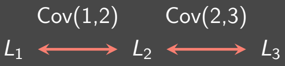
Why do we even need a new method when latent homophily ($\leftrightarrow$) is present?

Undirected Edge:



$$\text{Gibbs factors }^{9}\text{: } p(L_1 \mid L_2),\ p(L_2 \mid L_1, L_3), \text{ and } p(L_3 \mid L_2)$$

Bidirected Edge:



$$p(L_1, L_2, L_3) \sim MVN(\mu, \Sigma)$$

[9] Lauritzen and Richardson 2002

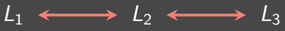
Estimate Parameters of a MVN

[10] Drton, Eichler, and Richardson 2009

[11] Moon 1996

Estimate Parameters of a MVN

If we have i.i.d. samples from $p(L_1, L_2, L_3) \sim MVN(\mu, \Sigma)$:

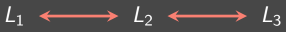


[10] Drton, Eichler, and Richardson 2009

[11] Moon 1996

Estimate Parameters of a MVN

If we have i.i.d. samples from $p(L_1, L_2, L_3) \sim MVN(\mu, \Sigma)$:



Residual Iterative Conditional Fitting (RICF).[10]

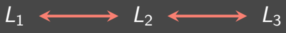
Similar to the Expectation Maximization (EM) algorithm [11].

[10] Drton, Eichler, and Richardson 2009

[11] Moon 1996

Estimate Parameters of a MVN

If we have i.i.d. samples from $p(L_1, L_2, L_3) \sim MVN(\mu, \Sigma)$:



Residual Iterative Conditional Fitting (RICF).[10]

Similar to the Expectation Maximization (EM) algorithm [11].

Iteratively finds the best-fitting $\widehat{\mu}$ and $\widehat{\Sigma}$ for our samples.

[10] Drton, Eichler, and Richardson 2009

[11] Moon 1996

New Method

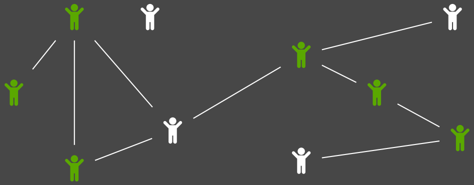
Able to estimate network causal effects when latent homophily ($\leftrightarrow$) is present.

New Method

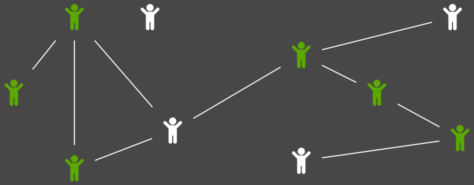
Step 1: find **connected triplets** s.t. no one in one triplet is adjacent to anyone in another triplet.

New Method

Step 1: find **connected triplets** s.t. no one in one triplet is adjacent to anyone in another triplet.

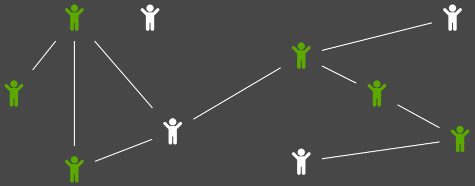


New Method



Step 2: collect data from these triplets

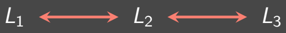
New Method



Step 2: collect data from these triplets, which can be seen as i.i.d. samples from the following graph:

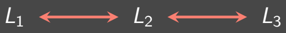
$$\begin{tikzcd} L_1 \arrow[r, leftrightarrow] & L_2 \arrow[r, leftrightarrow] & L_3 \end{tikzcd}$$

New Method



Step 3: estimate $\widehat{\mu}$ and $\widehat{\Sigma}$ using RICF.

New Method



Step 3: estimate $\widehat{\mu}$ and $\widehat{\Sigma}$ using RICF.

MVN($\widehat{\mu}$, $\widehat{\Sigma}$) $\approx$ the DGP of bidirected edges ($\leftrightarrow$).

New Method

We now can recover all kinds of DGPs under full interference:

New Method

We now can recover all kinds of DGPs under full interference:

✓ bidirected edges ($\leftrightarrow$): use thesis method

New Method

We now can recover all kinds of DGPs under full interference:

- ✓ bidirected edges ($\leftrightarrow$): use thesis method
- ✓ undirected edges ($-$): use auto-g method

New Method

We now can recover all kinds of DGPs under full interference:

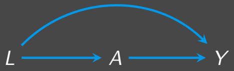
- ✓ bidirected edges ($\leftrightarrow$): use thesis method
- ✓ undirected edges ($-$): use auto-g method
- ✓ directed edges ($\rightarrow$): use auto-g method

Why Estimate DGP = Estimate Causal Effects?

Why Estimate DGP = Estimate Causal Effects?

A DGP is like a computer program:

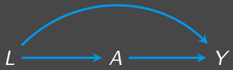
1. L receives a value;
2. $A \leftarrow f_A(L) + \text{noise}$;
3. $Y \leftarrow f_Y(A, L) + \text{noise}$;



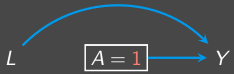
Why Estimate DGP = Estimate Causal Effects?

A DGP is like a computer program:

1. L receives a value;
2. $A \leftarrow f_A(L) + \text{noise}$;
3. $Y \leftarrow f_Y(A, L) + \text{noise}$;



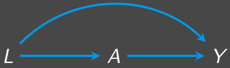
1. L receives a value;
2. $A \leftarrow 1$;
3. $Y \leftarrow f_Y(1, L) + \text{noise}$;



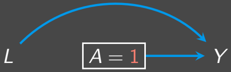
Why Estimate DGP = Estimate Causal Effects?

A DGP is like a computer program:

1. L receives a value;
2. $A \leftarrow f_A(L) + \text{noise}$;
3. $Y \leftarrow f_Y(A, L) + \text{noise}$;



1. L receives a value;
2. $A \leftarrow 1$;
3. $Y \leftarrow f_Y(1, L) + \text{noise}$;



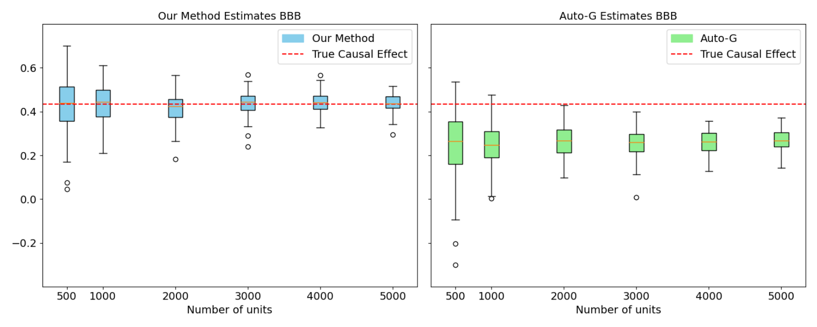
Similar for undirected (–) and bidirected (↔) edges.

Simulation Study 1

Latent homophily ($\leftrightarrow$) in all three (L, A, and Y) layers.

Simulation Study 1

Latent homophily ($\leftrightarrow$) in all three (L, A, and Y) layers.



Simulation Study 2

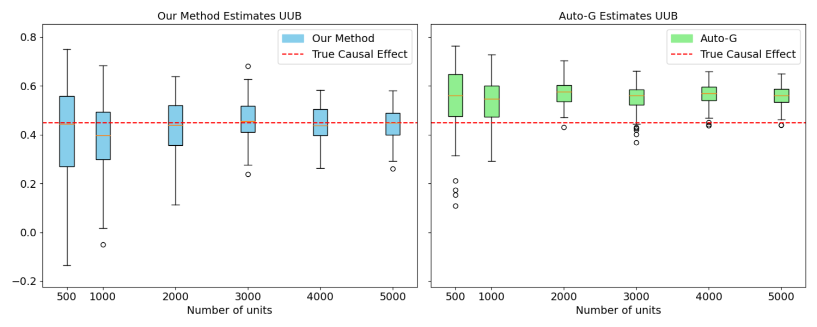
Contagion (−) in the L and A layers.

Latent homophily ($\leftrightarrow$) in the Y layers.

Simulation Study 2

Contagion (–) in the L and A layers.

Latent homophily ($\leftrightarrow$) in the Y layers.



Potential Broader Impact

New method for causal inference in network data with a more flexible set of assumptions:

- New opportunities for application of causal inference.

Potential Broader Impact

New method for causal inference in network data with a more flexible set of assumptions:

- New opportunities for application of causal inference.

Tests to distinguish contagion vs. latent homophily:

- Tool to verify model assumptions.
- Tool for causal discovery.

Sample Use Case of Our Method

$L =$ coursework & career preparation

$A =$ screen time

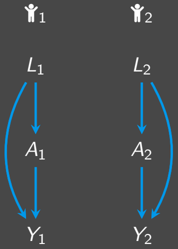
$Y =$ sleep disorder

Sample Use Case of Our Method

L = coursework & career preparation

A = screen time

Y = sleep disorder

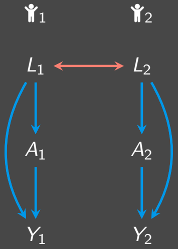


Sample Use Case of Our Method

L = coursework & career preparation

A = screen time

Y = sleep disorder



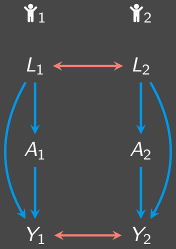
$L_1 \leftrightarrow L_2$: similar values, interests, and goals

Sample Use Case of Our Method

L = coursework & career preparation

A = screen time

Y = sleep disorder



$L_1 \leftrightarrow L_2$: similar values, interests, and goals

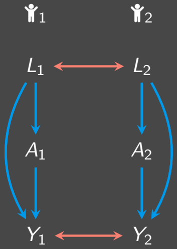
$Y_1 \leftrightarrow Y_2$: similar lifestyle (e.g. diet, exercise, etc.)

Sample Use Case of Our Method

L = coursework & career preparation

A = screen time

Y = sleep disorder



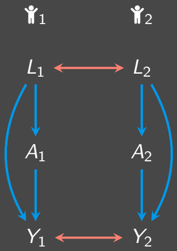
Before: can't apply the auto-g method

Sample Use Case of Our Method

L = coursework & career preparation

A = screen time

Y = sleep disorder



Before: can't apply the auto-g method

Thesis method:

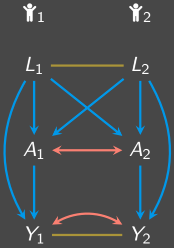
- hypothesis tests to confirm our model set up
- identify and estimate network causal effects

Limitation and Open Problems for Future Work

Contagion (–) and latent homophily (↔) cannot exist between two variables at the same time.

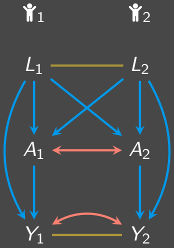
Limitation and Open Problems for Future Work

Contagion (–) and latent homophily (↔) cannot exist between two variables at the same time.



Limitation and Open Problems for Future Work

Contagion (–) and latent homophily (↔) cannot exist between two variables at the same time.



Can certainly happen in real life: e.g. $Y =$ stress level.

Acknowledgement

Acknowledgement

- My advisor Prof. Rohit Bhattacharya

Acknowledgement

- My advisor Prof. Rohit Bhattacharya
- Second reader Prof. Sam McCauley

Acknowledgement

- My advisor Prof. Rohit Bhattacharya
- Second reader Prof. Sam McCauley
- Prof. Aaron Williams

Acknowledgement

- My advisor Prof. Rohit Bhattacharya
- Second reader Prof. Sam McCauley
- Prof. Aaron Williams
- Limia and Brownswiss

Acknowledgement

- My advisor Prof. Rohit Bhattacharya
- Second reader Prof. Sam McCauley
- Prof. Aaron Williams
- Limia and Brownswiss
- Family and friends 🧡

References I

- Bhattacharya, Rohit, Daniel Malinsky, and Ilya Shpitser (2020). "Causal inference under interference and network uncertainty". In: *Uncertainty in Artificial Intelligence*. PMLR, pp. 1028–1038.
- Drton, Mathias, Michael Eichler, and Thomas S Richardson (2009). "Computing Maximum Likelihood Estimates in Recursive Linear Models with Correlated Errors.". In: *Journal of Machine Learning Research* 10.10.
- Hudgens, Michael G and M Elizabeth Halloran (2008). "Toward causal inference with interference". In: *Journal of the American Statistical Association* 103.482, pp. 832–842.
- Kang, Hyunseung and Guido Imbens (2016). "Peer encouragement designs in causal inference with partial interference and identification of local average network effects". In: *arXiv preprint arXiv:1609.04464*.
- Lauritzen, Steffen L and Thomas S Richardson (2002). "Chain graph models and their causal interpretations". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 64.3, pp. 321–348.
- Moon, Todd K (1996). "The expectation-maximization algorithm". In: *IEEE Signal processing magazine* 13.6, pp. 47–60.

References II

- Ogburn, Elizabeth L and Tyler J VanderWeele (2014). "Causal diagrams for interference". In.
- Shalizi, Cosma Rohilla and Andrew C Thomas (2011). "Homophily and contagion are generically confounded in observational social network studies". In: *Sociological methods & research* 40.2, pp. 211–239.
- Shpitser, Ilya (2015). "Segregated graphs and marginals of chain graph models". In: *Advances in neural information processing systems* 28.
- Tchetgen, Eric J Tchetgen and Tyler J VanderWeele (2012). "On causal inference in the presence of interference". In: *Statistical methods in medical research* 21.1, pp. 55–75.
- Tchetgen Tchetgen, Eric J, Isabel R Fulcher, and Ilya Shpitser (2021). "Auto-g-computation of causal effects on a network". In: *Journal of the American Statistical Association* 116.534, pp. 833–844.

Thanks!

Questions?