

Understanding linear regression for social networks

Alex Hayes

February 19, 2025 @ Oregon State University

Department of Statistics
University of Wisconsin-Madison

My background and research

My research interests and professional experience

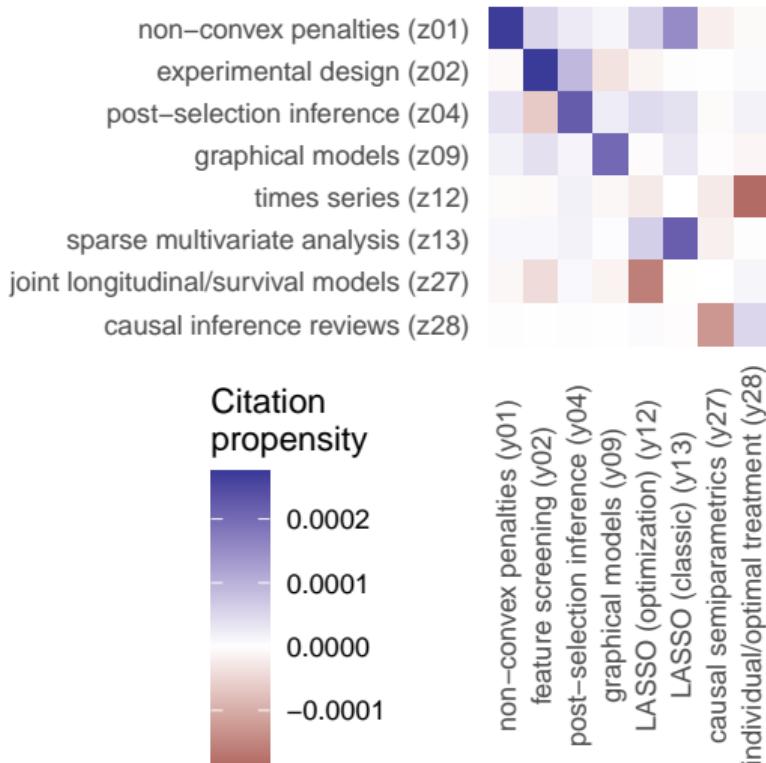
Research interests:

- (Social) networks
- Causal inference
- Peer effects
- Social processes in science
- Software design

Professional background:

- PhD Statistics, UW-Madison
- Facebook Core Data Science (Intern)
- RStudio (Intern)
- B.A. Statistics, Rice University

Past work: methods for community detection in networks (JCGS)



JOURNAL OF COMPUTATIONAL AND GRAPHICAL STATISTICS
2024, VOL. 00, NO. 0, 1–14
<https://doi.org/10.1080/10618600.2024.2394464>

Co-Factor Analysis of Citation Networks

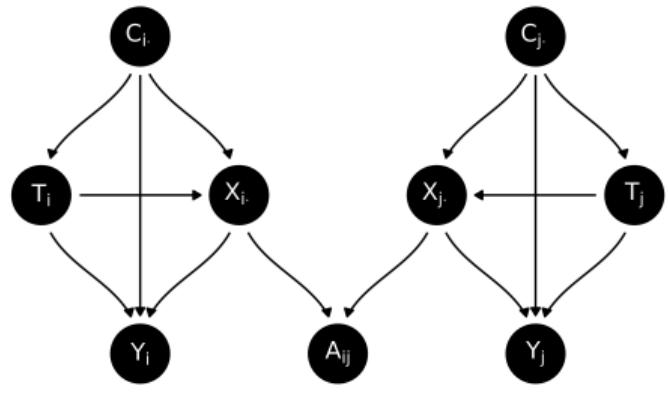
Alex Hayes  and Karl Rohe

Department of Statistics, University of Wisconsin-Madison, Madison, WI

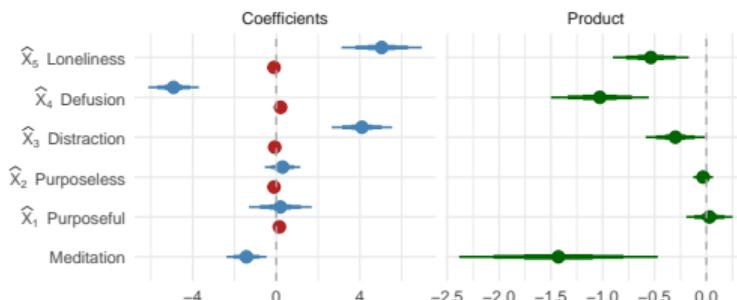
ABSTRACT

One compelling use of citation networks is to characterize papers by their relationships to the surrounding literature. We propose a method to characterize papers by embedding them into two distinct “co-factor” spaces: one describing how papers send citations, and the other describing how papers receive citations. This approach presents several challenges. First, older documents cannot cite newer documents, and thus it is not clear that co-factors are even identifiable. We resolve this challenge by developing a co-factor model for asymmetric adjacency matrices with missing lower triangles and showing that identification is possible. We then frame estimation as a matrix completion problem and develop a specialized implementation of matrix completion because prior implementations are memory bound in our setting. Simulations show that our estimator has promising finite sample properties, and that naive approaches fail to recover latent co-factor structure. We leverage our estimator to investigate 255,780 papers published in statistics journals from 1898 to 2024, resulting in the most comprehensive topic model of the statistics literature to date. We find interpretable co-factors corresponding to many statistical subfields, including time series, variable selection, spatial methods, graphical models, GLM(M)s, causal inference, multiple testing, quantile regression, semiparametrics, dimension reduction, and several more. Supplementary materials for this article are available online.

Past work: causal inference on networks (JMLR)



◆ Mediator ◆ Outcome ◆ Product



Estimating Network-Mediated Causal Effects via Principal Components Network Regression

Alex Hayes

Department of Statistics
University of Wisconsin-Madison
Madison, WI, USA

ALEX.HAYES@WISC.EDU

Mark M. Fredrickson

Department of Statistics
University of Michigan
Ann Arbor, MI, USA

MFREDRIC@UMICH.EDU

Keith Levin

Department of Statistics
University of Wisconsin-Madison
Madison, WI, USA

KDLEVIN@WISC.EDU

Editor: Ilya Shpitser

Abstract

We develop a method to decompose causal effects on a social network into an indirect effect mediated by the network, and a direct effect independent of the social network. To handle the complexity of network structures, we assume that latent social groups act as causal mediators. We develop principal components network regression models to differentiate the social effect from the non-social effect. Fitting the regression models is as simple as principal components analysis followed by ordinary least squares estimation. We prove asymptotic theory for regression coefficients from this procedure and show that it is widely applicable, allowing for a variety of distributions on the regression errors and network edges. We carefully characterize the counterfactual assumptions necessary to use the regression models for causal inference, and show that current approaches to causal network regression may result in over-control bias. The method is very general, so that it is applicable to many types of structured data beyond social networks, such as text, areal data, psychometrics, images and omics.

Past work: software development (JOSS)

- Developing R packages in the tidyverse with RStudio
- Collaborations with ROpenSci to develop standards for stats modeling
- Numerous R packages on CRAN, most notably broom



Welcome to the Tidyverse

Hadley Wickham¹, Mara Averick¹, Jennifer Bryan¹, Winston Chang¹, Lucy D'Agostino McGowan⁸, Romain François¹, Garrett Grolemund¹, Alex Hayes¹², Lionel Henry¹, Jim Hester¹, Max Kuhn¹, Thomas Lin Pedersen¹, Evan Miller¹³, Stephan Milton Bache³, Kirill Müller², Jeroen Ooms¹⁴, David Robinson⁵, Dana Paige Seidel¹⁰, Vitalie Spinu⁴, Kohske Takahashi⁹, Davis Vaughan¹, Claus Wilke⁶, Kara Woo⁷, and Hiroaki Yutani¹¹

1 RStudio 2 cynkra 3 Redbubble 4 Erasmus University Rotterdam 5 Flatiron Health 6 Department of Integrative Biology, The University of Texas at Austin 7 Sage Bionetworks 8 Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health 9 Chukyo University, Japan 10 Department of Environmental Science, Policy, & Management, University of California, Berkeley 11 LINE Corporation 12 University of Wisconsin, Madison 13 None 14 University of California, Berkeley

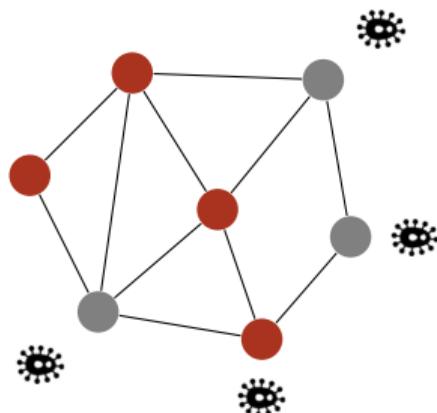
Summary



At a high level, the tidyverse is a language for solving data science challenges with R code. Its primary goal is to facilitate a conversation between a human and a computer about data. Less abstractly, the tidyverse is a collection of R packages that share a high-level design philosophy and low-level grammar and data structures, so that learning one package makes it easier to learn the next.

Recent work on network regression

Understanding social influence is fundamental in a highly connected society

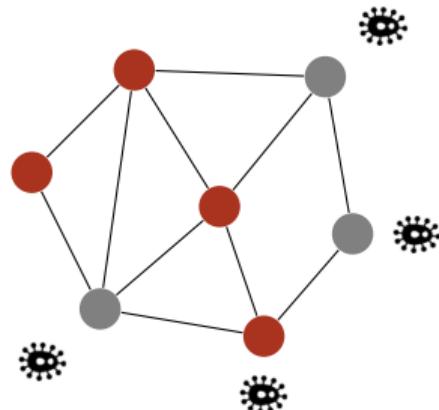


Direct effect: if I get **vaccinated**, I am less likely to get sick 🦠

Contagion: if my friends get sick 🦠, I am more likely to get sick 🦠

Interference: if my friends get **vaccinated**, I am less likely to get sick 🦠

Understanding social influence is fundamental in a highly connected society



Direct effect: if I get **vaccinated**, I am less likely to get sick 🦠

Contagion: if my friends get sick 🦠, I am more likely to get sick 🦠

Interference: if my friends get **vaccinated**, I am less likely to get sick 🦠

* Can be defined counterfactually (Vazquez-Bare, 2023), but we do not consider counterfactual inference in this talk.

This talk is about the linear-in-means model

Very popular tool for estimating social influence

Used in education, crime, health, social policy, etc¹

I discovered an issue with this model

¹Sacerdote (2001); Epple and Romano (2011); Soeteven and Kooreman (2007); Trogdon et al. (2008); Duflo and Saez (2003); Bertrand et al. (2000); Glaeser et al. (1996); Patacchini and Zenou (2012); Carrell et al. (2013), etc

Linear-in-means models are a canonical tool to estimate social influence

Outcome (sick?) $Y_i \in \{0, 1\}$ Base rate $\alpha \in \mathbb{R}$

$$\underbrace{Y_i}_{\text{sick?}} = \alpha$$

Linear-in-means models are a canonical tool to estimate social influence

Outcome	(sick?)	$Y_i \in \{0, 1\}$	Base rate	$\alpha \in \mathbb{R}$
Node degree	(num friends)	$d_i \in \{0, 1, 2, \dots\}$	Contagion	$\beta \in (-1, 1)$
Edge $i \sim j$	(friends?)	$A_{ij} \in \{0, 1\}$		

$$\underbrace{Y_i}_{\text{sick?}} = \alpha + \beta \underbrace{\frac{1}{d_i} \sum_{j : A_{ij}=1} Y_j}_{\text{fraction sick friends}}$$

Linear-in-means models are a canonical tool to estimate social influence

Outcome	(sick?)	$Y_i \in \{0, 1\}$	Base rate	$\alpha \in \mathbb{R}$
Node degree	(num friends)	$d_i \in \{0, 1, 2, \dots\}$	Contagion	$\beta \in (-1, 1)$
Edge $i \sim j$	(friends?)	$A_{ij} \in \{0, 1\}$	Direct effect	$\gamma \in \mathbb{R}$
Treatment	(vaccinated?)	$T_i \in \{0, 1\}$		

$$\underbrace{Y_i}_{\text{sick?}} = \alpha + \beta \underbrace{\frac{1}{d_i} \sum_{j : A_{ij}=1} Y_j}_{\text{fraction sick friends}} + \gamma \underbrace{T_i}_{\text{vaccinated?}}$$

Linear-in-means models are a canonical tool to estimate social influence

Outcome	(sick?)	$Y_i \in \{0, 1\}$	Base rate	$\alpha \in \mathbb{R}$
Node degree	(num friends)	$d_i \in \{0, 1, 2, \dots\}$	Contagion	$\beta \in (-1, 1)$
Edge $i \sim j$	(friends?)	$A_{ij} \in \{0, 1\}$	Direct effect	$\gamma \in \mathbb{R}$
Treatment	(vaccinated?)	$T_i \in \{0, 1\}$	Interference	$\delta \in \mathbb{R}$

$$\underbrace{Y_i}_{\text{sick?}} = \alpha + \beta \underbrace{\frac{1}{d_i} \sum_{j : A_{ij}=1} Y_j}_{\text{fraction sick friends}} + \gamma \underbrace{T_i}_{\text{vaccinated?}} + \delta \underbrace{\frac{1}{d_i} \sum_{j : A_{ij}=1} T_j}_{\text{fraction vaccinated friends}}$$

Linear-in-means models are a canonical tool to estimate social influence

Outcome	(sick?)	$Y_i \in \{0, 1\}$	Base rate	$\alpha \in \mathbb{R}$
Node degree	(num friends)	$d_i \in \{0, 1, 2, \dots\}$	Contagion	$\beta \in (-1, 1)$
Edge $i \sim j$	(friends?)	$A_{ij} \in \{0, 1\}$	Direct effect	$\gamma \in \mathbb{R}$
Treatment	(vaccinated?)	$T_i \in \{0, 1\}$	Interference	$\delta \in \mathbb{R}$

$$\underbrace{Y_i}_{\text{sick?}} = \alpha + \beta \underbrace{\frac{1}{d_i} \sum_{j : A_{ij}=1} Y_j}_{\text{fraction sick friends}} + \gamma \underbrace{T_i}_{\text{vaccinated?}} + \delta \underbrace{\frac{1}{d_i} \sum_{j : A_{ij}=1} T_j}_{\text{fraction vaccinated friends}} + \underbrace{\varepsilon_i}_{\text{error}}$$

Linear-in-means models are a canonical tool to estimate social influence

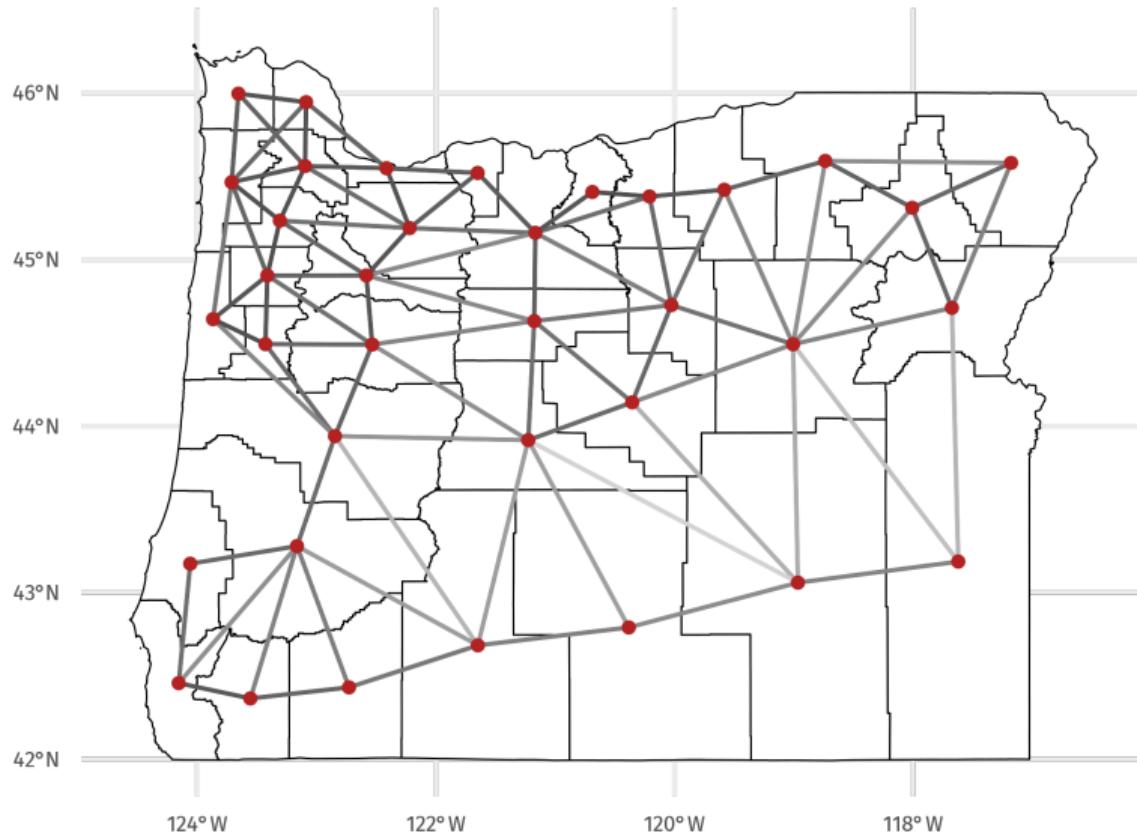
Outcome	(sick?)	$Y_i \in \{0, 1\}$	Base rate	$\alpha \in \mathbb{R}$
Node degree	(num friends)	$d_i \in \{0, 1, 2, \dots\}$	Contagion	$\beta \in (-1, 1)$
Edge $i \sim j$	(friends?)	$A_{ij} \in \{0, 1\}$	Direct effect	$\gamma \in \mathbb{R}$
Treatment	(vaccinated?)	$T_i \in \{0, 1\}$	Interference	$\delta \in \mathbb{R}$

$$\underbrace{Y_i}_{\text{sick?}} = \alpha + \beta \underbrace{\frac{1}{d_i} \sum_{j : A_{ij}=1} Y_j}_{\text{fraction sick friends}} + \gamma \underbrace{T_i}_{\text{vaccinated?}} + \delta \underbrace{\frac{1}{d_i} \sum_{j : A_{ij}=1} T_j}_{\text{fraction vaccinated friends}} + \underbrace{\varepsilon_i}_{\text{error}}$$

Letting $G = D^{-1}A$ be the row-normalized adjacency matrix, can write in matrix-vector form:

$$Y = \alpha \mathbf{1}_n + \beta GY + T\gamma + GT\delta + \varepsilon$$

Linear-in-means models are closely connected to spatial autoregression models



Identification in the linear-in-means model can be subtle

Review of Economic Studies (1993) 60, 531–542
© 1993 The Review of Economic Studies Limited

0034-6527/93/00270531\$02.00

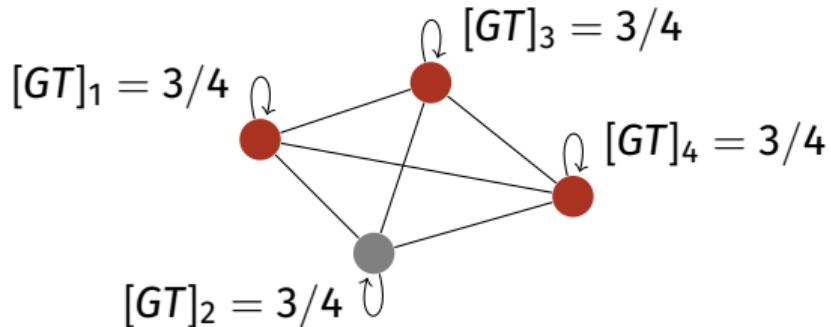
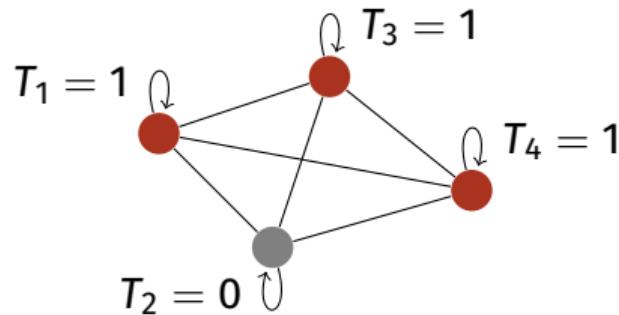
Identification of Endogenous Social Effects: The Reflection Problem

CHARLES F. MANSKI
University of Wisconsin-Madison

First version received December 1991; final version accepted December 1992 (Eds.)

This paper examines the reflection problem that arises when a researcher observing the distribution of behaviour in a population tries to infer whether the average behaviour in some group influences the behaviour of the individuals that comprise the group. It is found that inference is not possible unless the researcher has prior information specifying the composition of reference groups. If this information is available, the prospects for inference depend critically on the population relationship between the variables defining reference groups and those directly affecting outcomes. Inference is difficult to impossible if these variables are functionally dependent or are statistically independent. The prospects are better if the variables defining reference groups and those directly affecting outcomes are moderately related in the population.

Linear-in-means models are famously susceptible to perfect collinearity



$$\begin{matrix} 1_n & GY & T & GT \\ \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \end{bmatrix} & = & \begin{bmatrix} 1 & GY_1 & 1 & 3/4 \\ 1 & GY_2 & 0 & 3/4 \\ 1 & GY_3 & 1 & 3/4 \\ 1 & GY_4 & 1 & 3/4 \end{bmatrix} & \begin{bmatrix} \alpha \\ \beta \\ \gamma \\ \delta \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \end{bmatrix} \end{matrix}$$

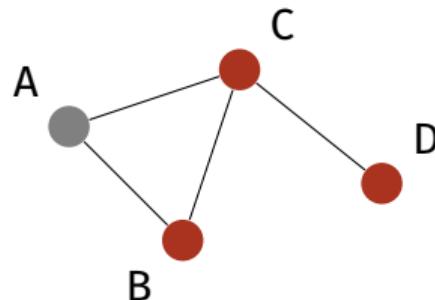
Can't distinguish base rate α from interference δ due to collinearity

It's widely believed that this “reflection problem” is rarely a problem in practice

Proposition (Bramoullé et al. 2009)

Suppose $\gamma\beta + \delta \neq 0$. If I , G and G^2 are linearly independent, i.e., that $aI + bG + cG^2 = 0$ requires $a = b = c = 0$, then α, β, γ and δ are identified.

There is no perfect collinearity and peer influence is identified when there are open triangles (“intransitivity”) in the network



Open: $B \leftrightarrow C \leftrightarrow D \leftrightarrow B$

Closed: $A \leftrightarrow B \leftrightarrow C \leftrightarrow A$

Standard wisdom is that collinearity is **not a problem** because most networks have open triangles

We came up with a new estimator for the linear-in-means model

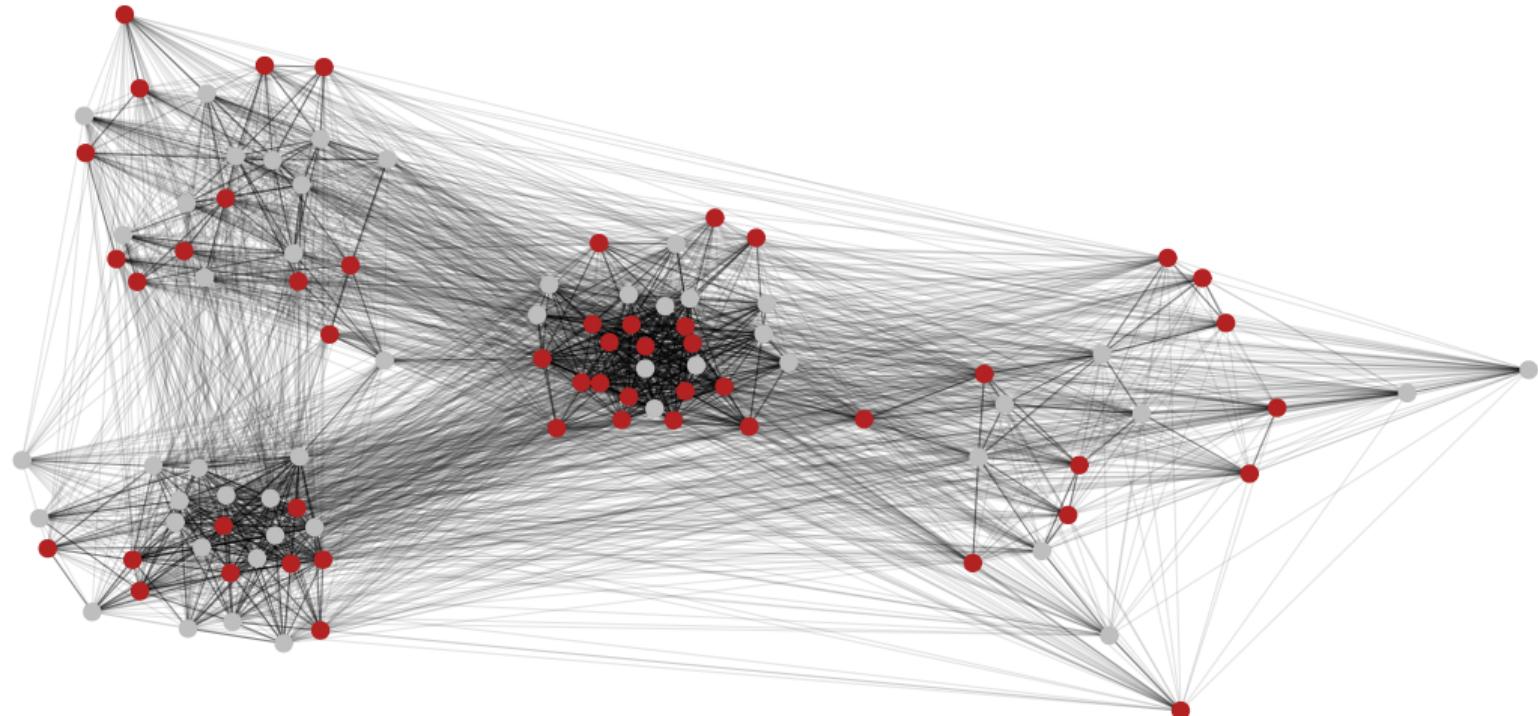
Setting: Treatment random and independent of network. $T_i \stackrel{\text{iid}}{\sim} \text{Bern}(0.5)$

$$Y = \alpha \mathbf{1}_n + \beta GY + T\gamma + GT\delta + \varepsilon$$

We started to run a simulation study² to confirm that our estimator worked...

²Generate Y via the reduced-form specification $Y = (I - \beta G)^{-1}(\alpha \mathbf{1}_n + \gamma T + \delta GT + \varepsilon)$

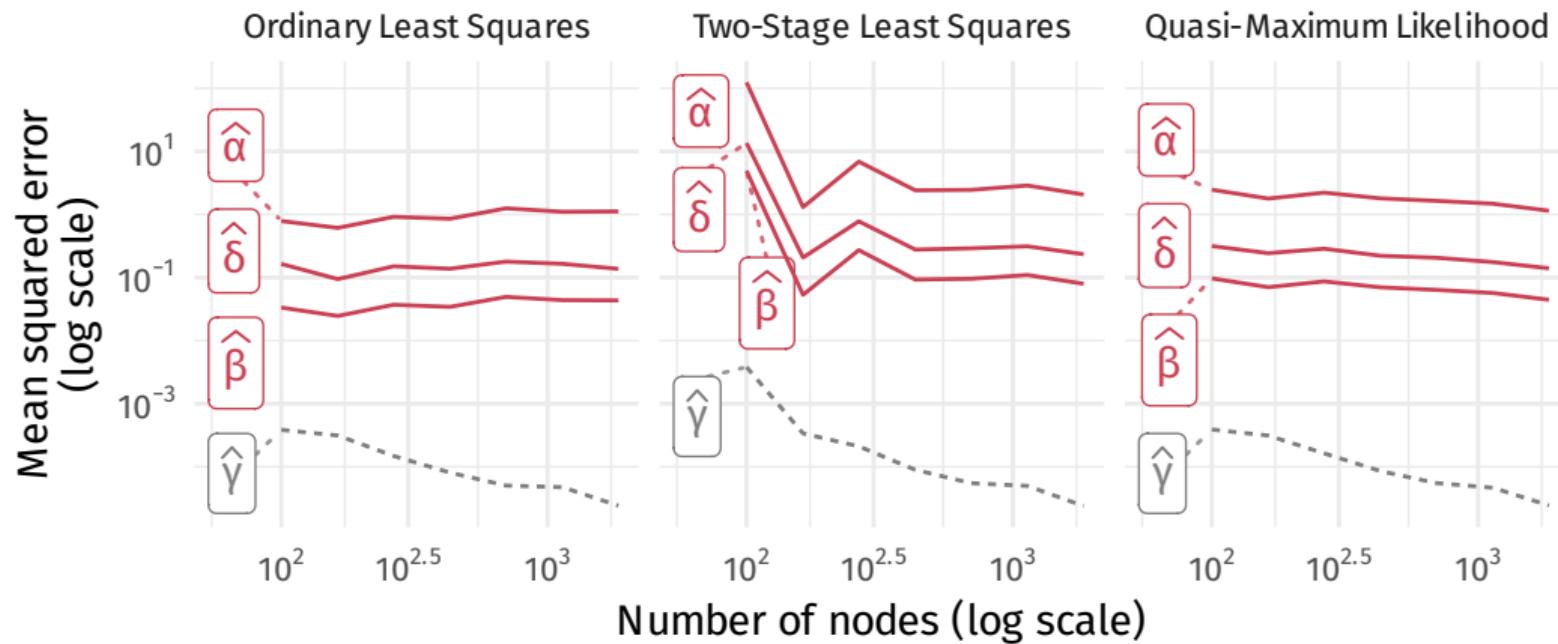
In our simulations, the network had many open triangles...



Treatments are assigned by coin flip and 45% of triangles are open

...but we couldn't estimate peer effects!

It wasn't just us, none of the standard estimators worked!



The issue: the interference column converges to a constant in large samples

$[GT]_i$
fraction
vaccinated
friends

The issue: the interference column converges to a constant in large samples

$$\underbrace{[GT]_i}_{\text{fraction vaccinated friends}} = \frac{1}{d_i} \underbrace{\sum_{j : A_{ij}=1} T_j}_{\text{average of } d_i \text{ i.i.d. coin flips}}$$

The issue: the interference column converges to a constant in large samples

When the network grows ($n \rightarrow \infty$),

$$\lim_{n \rightarrow \infty} \underbrace{[GT]_i}_{\text{fraction vaccinated friends}} = \lim_{n \rightarrow \infty} \frac{1}{d_i} \underbrace{\sum_{j : A_{ij}=1} T_j}_{\substack{\text{average of } d_i \\ \text{i.i.d. coin flips}}}$$

The issue: the interference column converges to a constant in large samples

When the network grows ($n \rightarrow \infty$), if everyone makes more friends ($d_i \rightarrow \infty$)

$$\lim_{n \rightarrow \infty} \underbrace{[GT]_i}_{\text{fraction vaccinated friends}} = \lim_{n \rightarrow \infty} \frac{1}{d_i} \underbrace{\sum_{j : A_{ij}=1} T_j}_{\substack{\text{average of } d_i \\ \text{i.i.d. coin flips}}} = \frac{1}{2}$$

The issue: the interference column converges to a constant in large samples

When the network grows ($n \rightarrow \infty$), if everyone makes more friends ($d_i \rightarrow \infty$)

$$\lim_{n \rightarrow \infty} \underbrace{[GT]_i}_{\text{fraction vaccinated friends}} = \lim_{n \rightarrow \infty} \frac{1}{d_i} \underbrace{\sum_{j : A_{ij}=1} T_j}_{\substack{\text{average of } d_i \\ \text{i.i.d. coin flips}}} = \frac{1}{2}$$

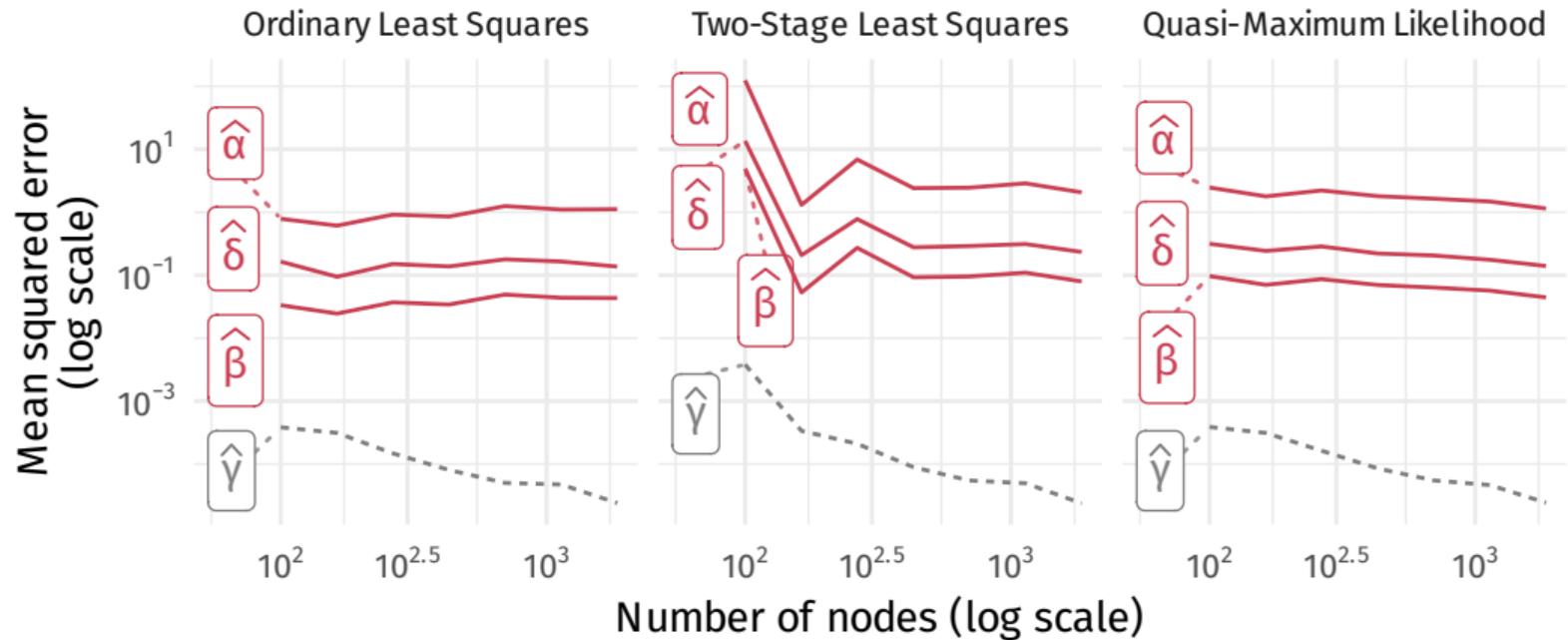
For every single node $i = 1, \dots, n$

Base rates and interference are collinear in large samples

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \underbrace{\begin{bmatrix} 1_n & GY & T & GT \\ 1 & GY_1 & 1 & 1/2 \\ 1 & GY_2 & 0 & 1/2 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & GY_n & 1 & 1/2 \end{bmatrix}}_{\text{as } n \rightarrow \infty} \begin{bmatrix} \alpha \\ \beta \\ \gamma \\ \delta \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Sometimes can't distinguish between base rate α and interference δ

In simulations, we couldn't estimate β either



Outcomes are generated by diffusing the treatment over the network

Why is β also affected?

$$Y = \alpha \mathbf{1}_n + \beta GY + \gamma T + \delta GT + \varepsilon$$

Outcomes are generated by diffusing the treatment over the network

Why is β also affected?

$$Y = \alpha \mathbf{1}_n + \beta GY + \gamma T + \delta GT + \varepsilon$$

$$Y - \beta GY = \alpha \mathbf{1}_n + \gamma T + \delta GT + \varepsilon$$

Outcomes are generated by diffusing the treatment over the network

Why is β also affected?

$$Y = \alpha \mathbf{1}_n + \beta GY + \gamma T + \delta GT + \varepsilon$$

$$Y - \beta GY = \alpha \mathbf{1}_n + \gamma T + \delta GT + \varepsilon$$

$$Y = (I - \beta G)^{-1}(\alpha \mathbf{1}_n + \gamma T + \delta GT + \varepsilon)$$

Outcomes are generated by diffusing the treatment over the network

Why is β also affected?

$$Y = \alpha \mathbf{1}_n + \beta GY + \gamma T + \delta GT + \varepsilon$$

$$Y - \beta GY = \alpha \mathbf{1}_n + \gamma T + \delta GT + \varepsilon$$

$$Y = (I - \beta G)^{-1}(\alpha \mathbf{1}_n + \gamma T + \delta GT + \varepsilon)$$

$$\stackrel{*}{=} \underbrace{\sum_{k=0}^{\infty} \beta^k G^k}_{\text{repeated neighborhood averaging}} (\alpha \mathbf{1}_n + \gamma T + \delta GT + \varepsilon)$$

* Must have $|\beta| < 1$, so effect of averaging decays with repetition

The contagion column converges to a constant in large samples

$$GY = \frac{\alpha}{1-\beta} \mathbf{1}_n + \underbrace{\gamma GT}_{\text{neighborhood average } \rightarrow \gamma/2} + \underbrace{(\gamma\beta + \delta) \sum_{k=0}^{\infty} \beta^k G^{k+2} T}_{\text{repeated neighborhood averages of } T^*} + \underbrace{\sum_{k=0}^{\infty} \beta^k G^{k+1} \varepsilon}_{\text{repeated neighborhood averages of } \varepsilon \rightarrow 0}$$

Each term in the sum converges to a constant

$$GY \rightarrow \eta$$

* Neighborhood average of a constant is that same constant

Base rates, interference and contagion are collinear in large samples

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \underbrace{\begin{bmatrix} 1_n & GY & T & GT \\ 1 & \eta & 1 & 1/2 \\ 1 & \eta & 0 & 1/2 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & \eta & 1 & 1/2 \end{bmatrix}}_{\text{as } n \rightarrow \infty} \begin{bmatrix} \alpha \\ \beta \\ \gamma \\ \delta \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Sometimes can't distinguish between base rate α , interference δ and contagion β

Peer effects are asymptotically collinear under very general circumstances

Assumption

1. T_1, T_2, \dots, T_n are independent with shared mean $\tau \in \mathbb{R}$, and T is independent of A .
2. $\{T_i - \tau : i \in [n]\}$ are independent subgamma random variables.
3. $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$ are independent subgamma random variables.
4. The minimum degree grows strictly faster than $\log n$, such that

$$\lim_{n \rightarrow \infty} \frac{\min_{i \in [n]} d_i}{\log n} = \infty.$$

Recall:

$$Y = \alpha \mathbf{1}_n + \beta GY + T\gamma + GT\delta + \varepsilon$$

These assumptions are distributionally agnostic

Definition (Boucheron et al. 2013)

Let Z be a mean-zero random variable with cumulant generating function $\psi_Z(t) = \log \mathbb{E}[e^{tZ}]$. Z is *subgamma* with parameters $\nu \geq 0$ and $b \geq 0$ if

$$\psi_Z(t) \leq \frac{t^2\nu}{2(1-bt)} \text{ and } \psi_{-Z}(t) \leq \frac{t^2\nu}{2(1-bt)} \text{ for all } t < 1/b.$$

We then write that Z is (ν, b) -subgamma.

Examples: Bernoulli, Poisson, Exponential, Gamma, Gaussian, sub-Gaussian, squared sub-Gaussians, bounded distributions, etc

The interference and contagion columns converge uniformly to constants

Lemma

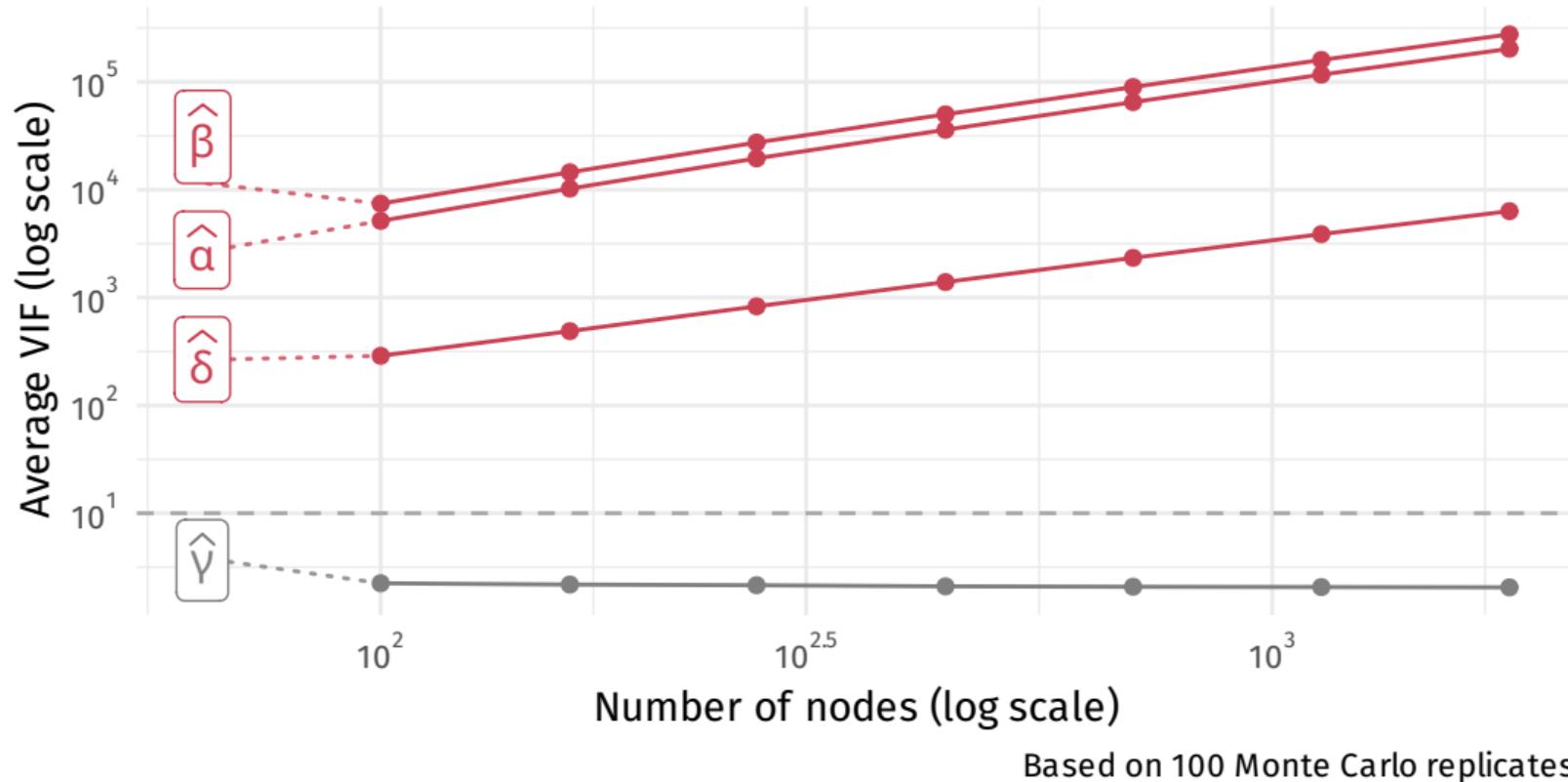
Under the previous assumptions,

$$\max_{i \in [n]} |[GT]_i - \tau| = o(1) \text{ almost surely}$$

and there exists $\eta \in \mathbb{R}$ such that

$$\max_{i \in [n]} |[GY]_i - \eta| = o(1) \text{ almost surely.}$$

Collinearity shows up quickly in finite samples



Asymptotic collinearity can lead to inconsistency

Theorem (Hayes and Levin 2024)

Let $(\hat{\alpha}, \hat{\beta}, \hat{\gamma}, \hat{\delta})$ be the vector of ordinary least squares estimates of $(\alpha, \beta, \gamma, \delta)$. Recall that $G = D^{-1}A$ is the row-normalized adjacency matrix. Suppose that the degrees of the network are such that $\|G\|_F^2 = o(n)$. Then if $\beta = 0$,

$$\min\{|\hat{\alpha} - \alpha|, |\hat{\beta} - \beta|\} = \Omega_P(1)$$

and

$$|\hat{\delta} - \delta| = \Omega_P\left(\frac{1}{\|G\|_F}\right). \quad (1)$$

If $\beta \neq 0$,

$$\min\{|\hat{\alpha} - \alpha|, |\hat{\beta} - \beta|\} = \Omega_P\left(\frac{1}{\|G\|_F}\right).$$

Under the stronger assumption $\|G\|_F^2 = o(\sqrt{n})$, eq. (1) holds for all β .

Asymptotic collinearity can lead to inconsistency

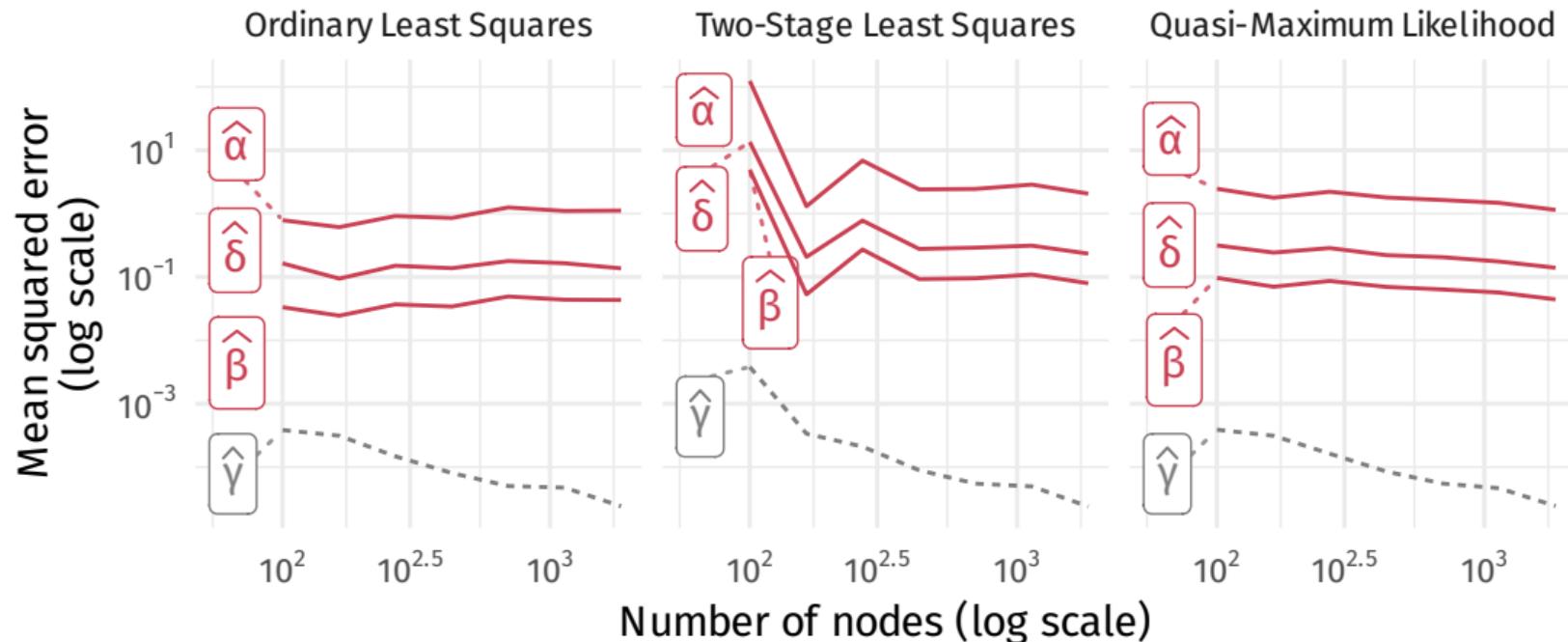
Theorem (Intuitive)

When minimum degree grows, ordinary least squares estimates of α , β and δ are either inconsistent, or at best consistent at $\sqrt{n/d_{\min}}$ rates, where $d_{\min} = \min_{i \in [n]} d_i$.

This is because the signal-to-noise ratio depends on the minimum degree³.

³Our lower bound matches the upper bound on estimation error in Lee (2004).

Asymptotic collinearity affects other estimators as well



Details: weighted and directed networks

Weighted networks: If $A \in \mathbb{R}_{\geq 0}^{n \times n}$ is a positive, weighted network with (ν, b) -subgamma edges A_{ij} , we require that

$$\max_{i \in [n]} \frac{1}{d_i^2} \sum_{j=1}^n A_{ij}^2 = o\left(\frac{1}{\nu \log^2 n}\right) \text{ and } \max_{j \in [n]} \frac{A_{ij}}{d_j} = o\left(\frac{1}{b \log n}\right).$$

Roughly: no one edge can be too important for a given node

Directed networks: extension possible, but slightly more involved

Isolated nodes: can allow a vanishing fraction of nodes to be isolated

Standard wisdom states that collinearity isn't a problem in linear-in-means in networks with open triangles.

Takeaway 1

When nodal covariates are independent of the network, we show that peer effects may not be estimable, due to collinearity, even when there are many open triangles.

Takeaway 2

Direct effects are estimable even when there is asymptotic collinearity.

Why hasn't the spatial econometrics literature encountered this issue before?

1. Often treat covariates T and network G as fixed rather than random
2. Theory directly assumes there is no asymptotic collinearity
3. Simulations rarely investigate consistency of estimators

The interference literature is well-aware of issues with Bernoulli designs

The Conflict Graph Design: Estimating Causal Effects under Arbitrary Neighborhood Interference

Vardis Kandiros¹, Charilaos Pipis¹, Constantinos Daskalakis¹, and Christopher Harshaw²

¹*Massachusetts Institute of Technology*

²*Columbia University*

November 19, 2024

Abstract

A fundamental problem in network experiments is selecting an appropriate experimental design in order to precisely estimate a given causal effect of interest. In fact, optimal rates of estimation remain unknown for essentially all causal effects in network experiments. In this work, we propose a general approach for constructing experiment designs under network interference with the goal of precisely estimating a pre-specified causal effect. A central aspect of our approach is the notion of a *conflict graph*, which captures the fundamental unobservability associated with the causal effect and the underlying network. We refer to our experimental design as the Conflict Graph Design. In order to estimate effects, we propose a modified Horvitz-Thompson estimator. We show that its variance under the Conflict Graph Design is bounded as $\mathcal{O}(\lambda(\mathcal{H})/n)$, where $\lambda(\mathcal{H})$ is the largest eigenvalue of the adjacency matrix of the conflict graph. These rates depend on both the underlying network and the particular causal effect under investigation. Not only does this yield the best known rates of estimation for several well-studied causal effects (e.g. the global and direct effects) but it also provides new methods for effects which have received less attention from the perspective of experiment design (e.g. spill-over effects). Our results corroborate two implicitly understood points in the literature: (1) that in order to increase precision, experiment designs should be tailored to specific causal effects of interest and (2) that “more local” effects are easier to estimate than “more global” effects. In addition to point estimation, we construct conservative variance estimators which facilitate the construction of asymptotically valid confidence intervals for the causal effect of interest.

Graph Cluster Randomization: Network Exposure to Multiple Universes

Johan Ugander
Cornell University
jhu5@cornell.edu

Brian Karrer
Facebook
{briankarrer,lars}@fb.com

Lars Backstrom
Facebook
larsbackstrom@fb.com

Jon Kleinberg
Cornell University
kleinber@cs.cornell.edu

ABSTRACT

A/B testing is a standard approach for evaluating the effect of online experiments; the goal is to estimate the ‘average treatment effect’ of a new feature or condition by exposing a sample of the overall population to it. A drawback with A/B testing is that it is poorly suited for experiments involving social interference, when the treatment of individuals spills over to neighboring individuals along an underlying social network. In this work, we propose a novel methodology using graph clustering to analyze average treatment effects under social interference. To begin, we characterize graph-theoretic conditions under which individuals can be considered to be ‘network exposed’ to an experiment. We then show how graph cluster randomization admits an efficient exact algorithm to compute the probabilities for each vertex being network exposed under several of these exposure conditions. Using these probabilities as inverse weights, a Horvitz-Thompson estimator can then provide an effect estimate that is unbiased, provided that the exposure model has been properly specified.

Given an estimator that is unbiased, we focus on minimizing the variance. First, we develop simple sufficient conditions for the variance of the estimator to be asymptotically small in n , the size of the graph. However, for general randomization schemes, this variance can be *lower bounded* by an *exponential* function of the degrees of a graph. In contrast, we show that if a graph satisfies a *restricted-growth* condition on the growth rate of neighborhoods, then there exists a natural clustering algorithm, based on vertex neighborhoods, for which the variance of the estimator can be *upper bounded* by a *linear* function of the degrees. Thus we show that proper cluster randomization can lead to exponentially lower estimator variance when experimentally measuring average treatment effects under interference.

1. INTRODUCTION

Social products and services – from fax machines and cell phones to online social networks – inherently exhibit ‘network effects’ with regard to their value to users. The value of these products to a user is inherently non-local, since it typically grows as members of the user’s social neighborhood use the product as well. Yet randomized experiments (or ‘A/B tests’), the standard machinery of testing frameworks including the Rubin causal model [14], critically assume what is known as the ‘stable unit treatment value assumption’ (SUTVA), that each individual’s response is affected only by their own treatment and not by the treatment of any other individual. Addressing this tension between the formalism of A/B testing and the non-local effects of network interaction has emerged as a key open question in the analysis of on-line behavior and the design of network experiments [6].

Under ordinary randomized trials where the stable unit treatment value assumption is a reasonable approximation — for example when a search engine A/B tests the effect of their color scheme upon the visitation time of their users — the population is divided into two groups: those in the ‘treatment’ group who see the new color scheme A and those in the control group who see the default color scheme B. Assuming there are negligible interference effects between users, each individual in the treated group responds just as he or she would if the entire population were treated, and each individual in the control group responds just as he or she would if the entire population were in control. In this manner, we can imagine that we are observing results from samples of two distinct ‘parallel universes’ at the same time — ‘Universe A’ in which color scheme A is used for everyone, and ‘Universe B’ in which color scheme B is used for everyone — and we can make inferences about the properties of user behavior in each of these universes.

Causal folks are interested in treatments that depend on position in network

Estimating Causal Peer Influence in Homophilous Social Networks by Inferring Latent Locations

Edward McFowland III^a  and Cosma Rohilla Shalizi^b

^aDepartment of Information and Decision Sciences, Carlson School of Management, University of Minnesota, Minneapolis, MN; ^bStatistics Department, Carnegie Mellon University, and the Santa Fe Institute, Pittsburgh, PA

ABSTRACT

Social influence cannot be identified from purely observational data on social networks, because such influence is generically confounded with latent homophily, that is, with a node's network partners being informative about the node's attributes and therefore its behavior. If the network grows according to either a latent community (stochastic block) model, or a continuous latent space model, then latent homophilous attributes can be consistently estimated from the global pattern of social ties. We show that, for common versions of those two network models, these estimates are so informative that controlling for estimated attributes allows for asymptotically unbiased and consistent estimation of social-influence effects in linear models. In particular, the bias shrinks at a rate that directly reflects how much information the network provides about the latent attributes. These are the first results on the consistent nonexperimental estimation of social-influence effects in the presence of latent homophily, and we discuss the prospects for generalizing them.

ARTICLE HISTORY

Received February 2018

Accepted June 2021

KEYWORDS

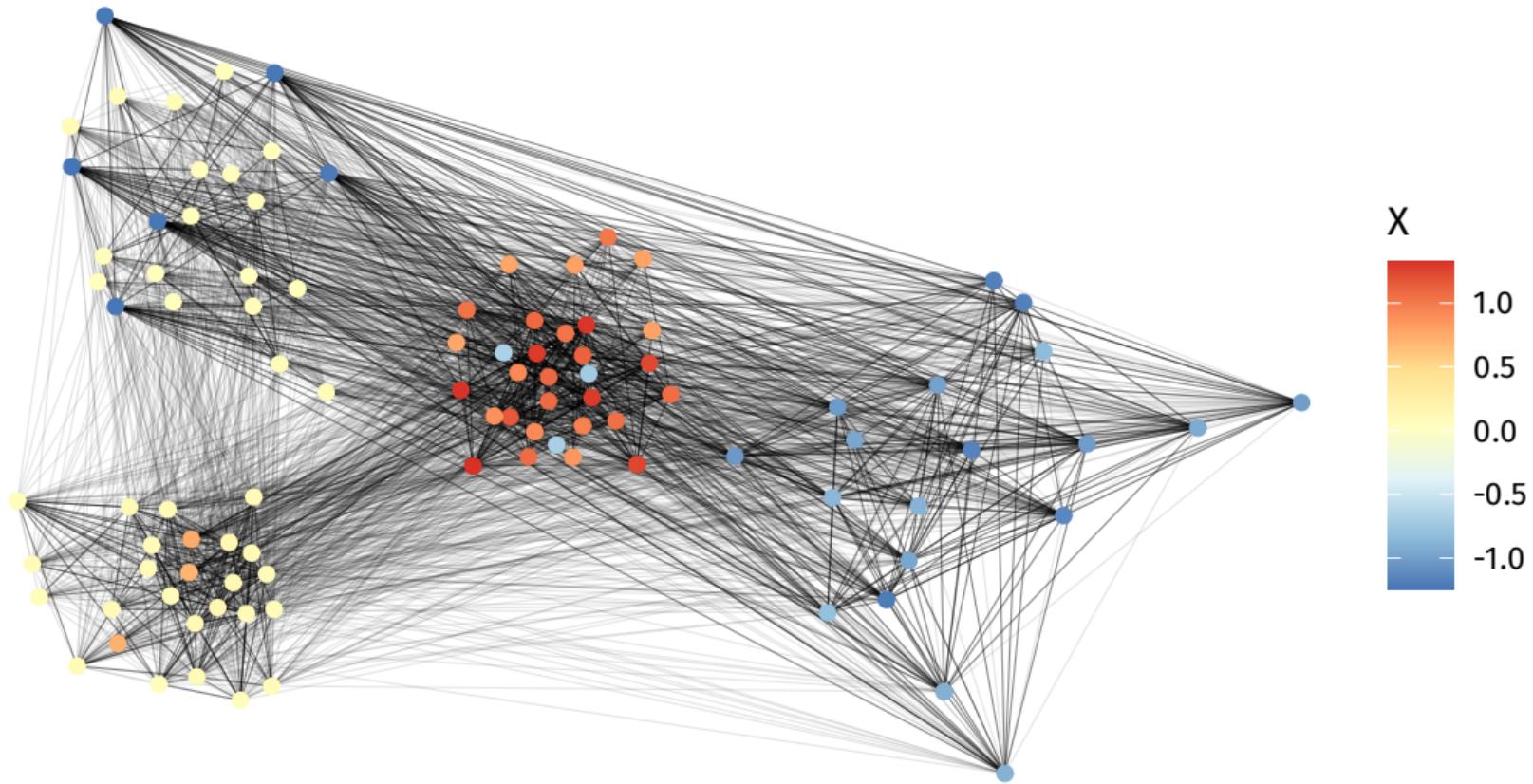
Causal Inference; Homophily; Social Networks; Peer Influence

Dependence between treatment and network might resolve collinearity issues

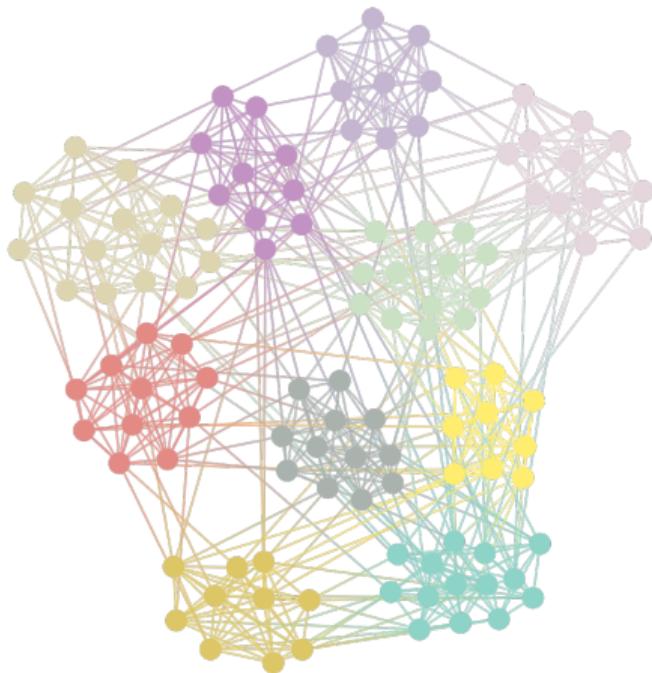
$$\underbrace{[GT]_i}_{\text{fraction vaccinated friends}} = \underbrace{\frac{1}{d_i} \sum_{j : A_{ij}=1} T_j}_{\text{average of dependent treatments}}$$

GT might not converge, or might converge to non-constant value

We considered models where treatment depended on position in network



We considered models where treatment depended on position in network



Stochastic blockmodels are an intuitive way to induce dependence

Block indicators Z_i

Popularity parameters θ_j

Mixing matrix $B \in [0, 1]^{d \times d}$

$$\mathbb{P}(A_{ij} = 1 \mid Z, \theta) = \theta_i Z_i B Z_j^T \theta_j$$

Linear-in-means models on blockmodels are of independent interest!

Causal Network Influence with Latent Homophily and Measurement Error: An Application to Therapeutic Community

Subhadeep Paul^{1*}, Shanjukta Nath², Keith Warren¹

¹The Ohio State University and ²Stanford University

Abstract

The Spatial or Network Autoregressive model (SAR, NAM) is popular for modeling the influence network connected neighbors exert on the outcome of individuals. However, many authors have noted that the *causal* network influence or contagion cannot be identified from observational data due to the presence of homophily. We propose a latent homophily-adjusted spatial autoregressive model for networked responses to identify the causal contagion and contextual effects. The latent homophily is estimated from the spectral embedding of the network's adjacency matrix. Separately, we develop maximum likelihood estimators for the parameters of the SAR model correcting for measurement error when covariates are measured with error. We show that the bias corrected MLE are consistent and derive its asymptotic limiting distribution. We propose to estimate network influence using the bias corrected MLE in a SAR model with the estimated latent homophily added as a covariate. Our simulations show that the methods perform well in finite sample. We apply our methodology to a data-set of female criminal offenders in a therapeutic community (TC) for substance abuse and criminal behavior. We provide causal estimates of network influence on graduation from TC and re-incarceration after accounting for latent homophily.

We prove a partial asymptotic collinearity result for these models

Theorem (Hayes and Levin 2024)

Suppose that A is sampled from a degree-corrected stochastic blockmodel.

Define $X_i = \theta_i Z_i$. Let

$$Y = \alpha 1_n + \beta GY + X\gamma + GX\delta + \varepsilon$$

for $\alpha, \beta \in \mathbb{R}$ and $\gamma, \delta \in \mathbb{R}^d$. Suppose that X has $k \geq 2d$ distinct rows. Then, under some conditions,

$$W_n = \begin{bmatrix} 1_n & GY & X & GX \end{bmatrix}$$

converges uniformly to a limit object with rank $2d$ out of $2d + 2$. If any two entries of $(\alpha, \beta, \delta_1, \dots, \delta_d)$ are set to zero in the data generating process, the limit object of W_n is a matrix with full rank.

We prove a partial asymptotic collinearity result for these models

Key condition to avoid collinearity: sufficient degree heterogeneity such that X and $D^{-1}X$ are linearly independent

General low-rank networks: if $\mathbb{E}[A_{ij} | X] = X_i^T X_j$, a similar result holds, broadly generalizing the partial identification result

Explains **counterintuitive** simulations in Paul et al. (2022); had to drop intercept column to avoid asymptotic collinearity!

We performed a simulation study to confirm the theoretical results

- **Bernoulli:** Treatment random and independent of network. $T_i \stackrel{\text{iid}}{\sim} \text{Bern}(0.5)$

$$Y = \alpha \mathbf{1}_n + \beta GY + T\gamma + GT\delta + \varepsilon,$$

with $\alpha = 3$, $\beta = 0.2$, $\gamma = 4$, $\delta = 2$ and $\varepsilon \stackrel{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$ with $\sigma = 0.1$.

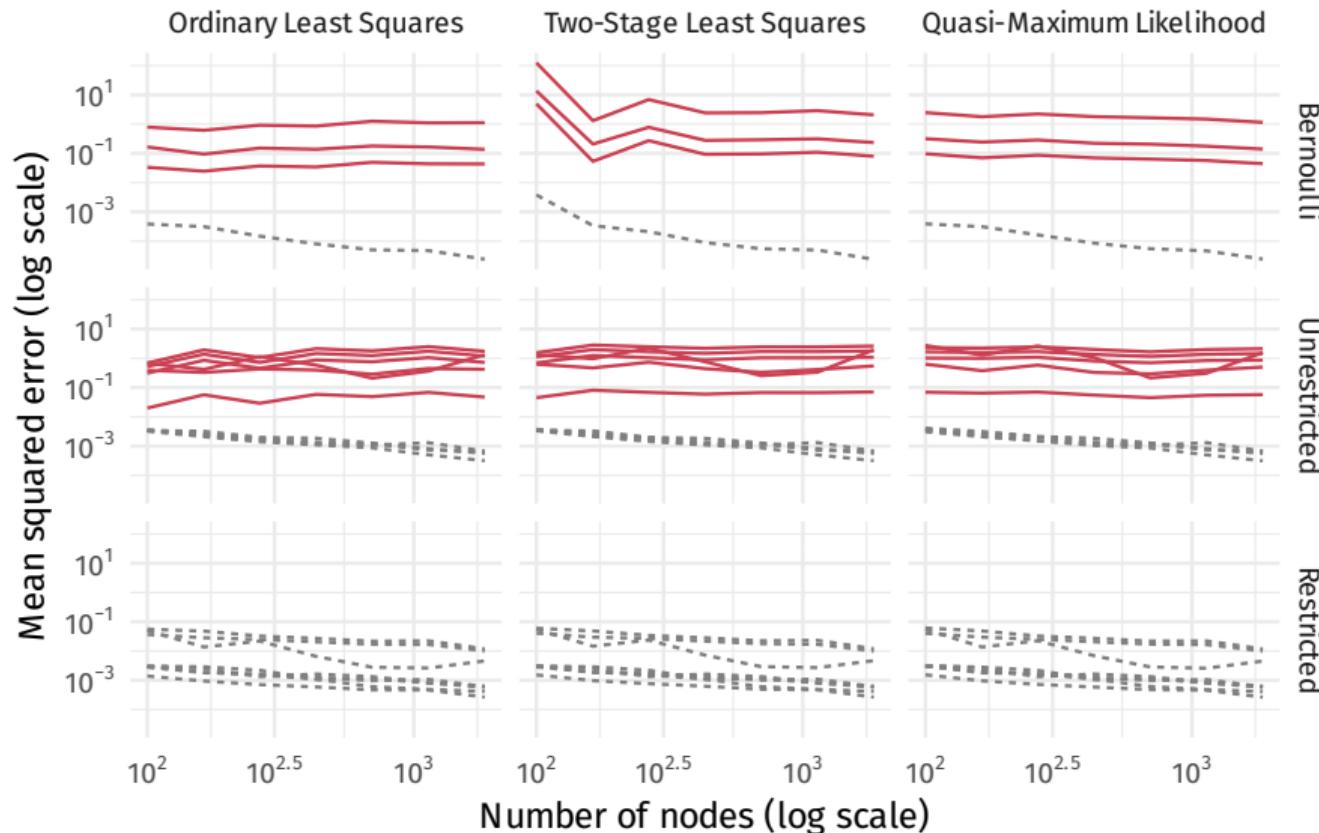
- **Unrestricted model:** Treatment random and dependent on network. Define $X_i = \theta_i Z_i \in \mathbb{R}^4$

$$Y = \alpha \mathbf{1}_n + \beta GY + X\gamma + GX\delta + \varepsilon,$$

where $\alpha = 3$, $\beta = 0.2$ and $\varepsilon \stackrel{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$ with $\sigma = 0.1$. Since $X_i \in \mathbb{R}^4$, $\gamma, \delta \in \mathbb{R}^4$ and we fix $\delta = (2, 2, 2, 2)$ and $\gamma = (1.5, 2.5, 3.5, 4.5)$.

- Restricted model: The unrestricted model, but $\delta = (0, 0, 2, 2)$, so there's no asymptotic collinearity.

Dependence prevented asymptotic collinearity and estimation challenges



Takeaway 3

Explicitly modelling dependence between nodal covariates and network structure can aid identifiability and resolve asymptotic collinearity issues.

Takeaway 4

Treatments dependent on network must be considered on a case-by-case basis, considering both the treatment and the network model.

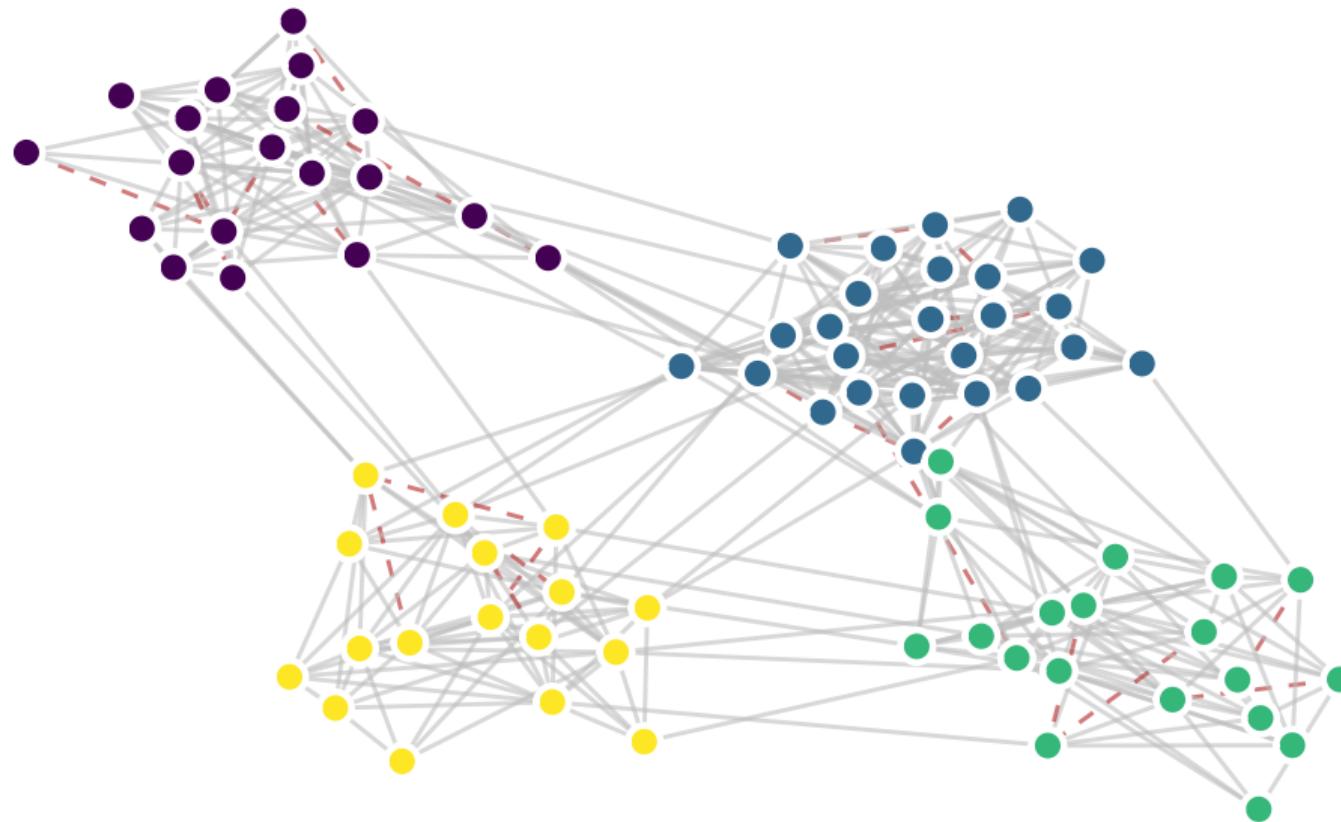
Open questions about linear-in-means models

Is asymptotic collinearity an issue in **longitudinal models** of peer effects like Zhu et al. (2017); McFowland and Shalizi (2021); Katsouris (2024)?

Can asymptotic collinearity be avoided via additional assumptions on the **error covariance** ε , as in Rose (2017)?

Future work

Estimating peer effects in network with missing and noisy edges



Estimating peer effects in network with missing and noisy edges



Figure 1: A

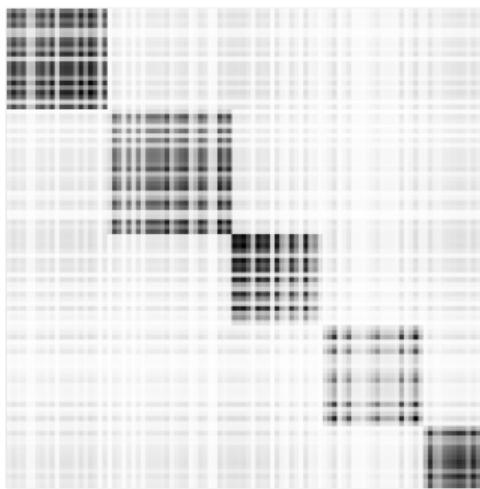


Figure 2: $\mathbb{E}[A | X] = XX^T$

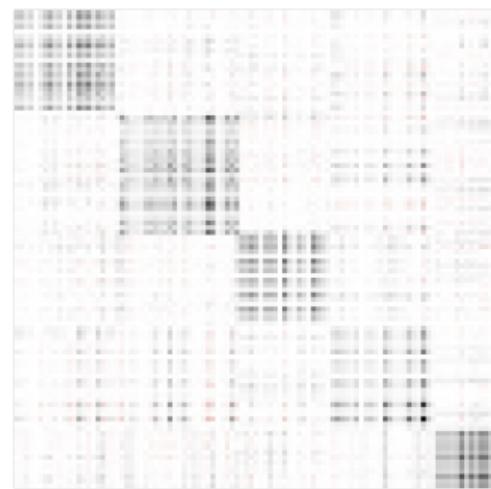
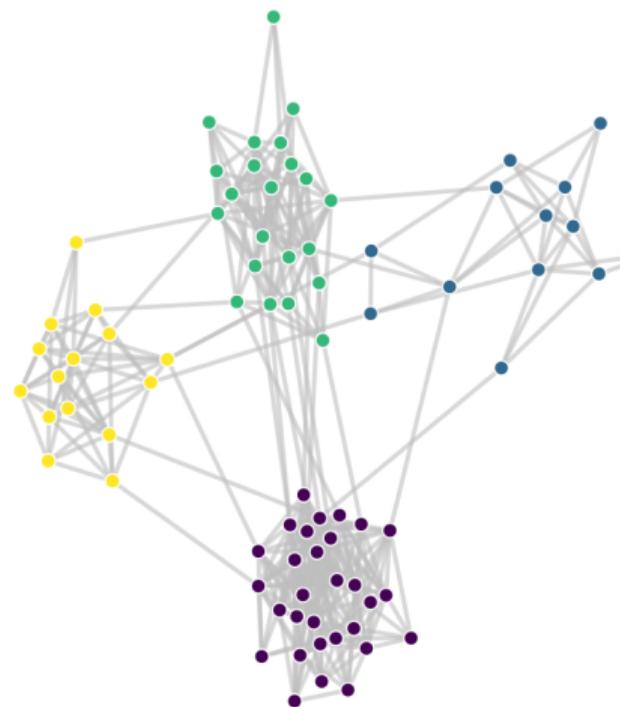


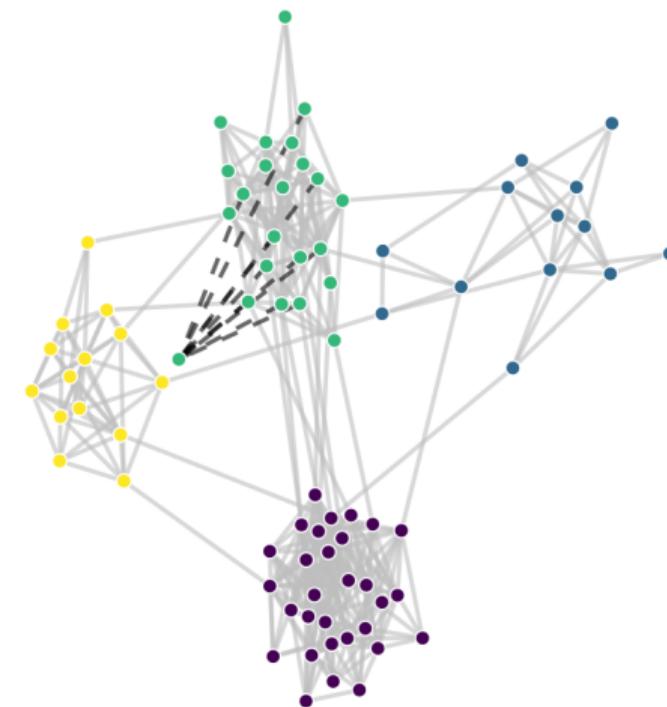
Figure 3: $\hat{X}\hat{X}^T$

Causal inference for co-evolution models

$T = 1$

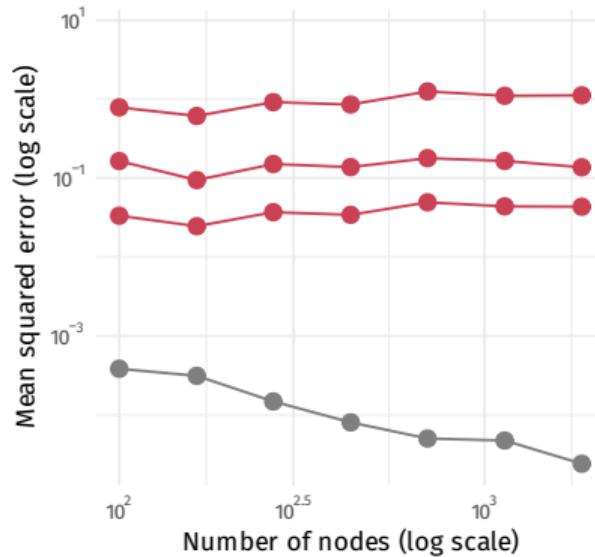


$T = 2$



Thank you! Questions?

✉ alex.hayes@wisc.edu
🌐 alexpghayes.com
GitHub 🐳 github.com/alexpghayes



Pre-print

Alex Hayes and Keith Levin. “Peer Effects in the Linear-in-Means Model May Be Inestimable Even When Identified.” arXiv, October 14, 2024. <http://arxiv.org/abs/2410.10772>.

A formal definition for identifiability

Definition (Maclaren and Nicholson 2020)

A model $\mathcal{M} = \{P_\theta : \theta \in \Theta\}$ is a collection of probability measures P_θ , indexed by a set Θ . A parameter $q(\theta)$ is *identifiable* if and only if $q(\theta_1) \neq q(\theta_2)$ implies $P_{\theta_1} \neq P_{\theta_2}$.

Several equivalent conditions for identifiability in linear models

In linear models, where $Y_i = X_i\theta + \varepsilon_i$ and $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$, the following are equivalent (Lewbel, 2019):

1. θ is identified
2. X is full-rank (i.e., there is no perfect collinearity)
3. the covariance matrix $X^T X / n$ is full-rank
4. the log-likelihood

$$-\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - x_i\theta)^2$$

has a unique maximizer.

A linear model that is identified, asymptotically collinear, and inestimable

Suppose that all data points except for the first data point are exactly equal:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & 2 \\ 1 & 1 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Then α and β are identified but cannot be estimated

Estimators

- OLS: $\text{lm}(y \sim Gy + T + GT)$
- TSLS: $\text{ivreg}(y \sim Gy + T + GT \mid \underbrace{T + GT + G^2T}_{\text{instruments}})$

Definition (Random Dot Product Graph, Young and Scheinerman 2007)

Let F be a distribution on \mathbb{R}^d such that $0 \leq x^T y$ for all $x, y \in \text{supp } F$ and the convex cone of $\text{supp } F$ is d -dimensional. Draw $X_1, X_2, \dots, X_n \stackrel{\text{iid}}{\sim} F$, and collect these in the rows of $X \in \mathbb{R}^{n \times d}$ for ease of notation. Conditional on these n vectors, which we call *latent positions*, generate edges by drawing the edges $\{A_{ij} : 1 \leq i < j \leq n\}$ as independent (ν, b) -subgamma random variables with $\mathbb{E}[A_{ij} | X] = \rho X_i^T X_j$, where $\rho \in [0, 1]$. Then we say that A is distributed according to an n -vertex random dot product graph with latent position distribution F , (ν, b) -subgamma edges and sparsity factor ρ . We write $(A, X) \sim \text{RDPG}(F, n)$, with the subgamma and sparsity parameters made clear from the context.

Proposition

Let $\mu = \mathbb{E}[X] \in \mathbb{R}^d$ and suppose that $Y_1, Y_2, \dots, Y_d, Z_1, Z_2, \dots, Z_d \in \mathbb{R}^d$ are rows of $X \in \mathbb{R}^{n \times d}$ such that Y_1, Y_2, \dots, Y_d are linearly independent and Z_1, Z_2, \dots, Z_d are linearly independent.

$$H_Y = \text{diag} \left(Y_1^T \mu, Y_2^T \mu, \dots, Y_d^T \mu \right) \quad \text{and} \quad H_Z = \text{diag} \left(Z_1^T \mu, Z_2^T \mu, \dots, Z_d^T \mu \right).$$

Provided that $Z^{-1}H_Z^{-1}Z - Y^{-1}H_Y^{-1}Y \in \mathbb{R}^{d \times d}$ is invertible, then the matrix

$$M = \begin{bmatrix} X & H^{-1}X \end{bmatrix} \in \mathbb{R}^{n \times 2d}$$

has rank $2d$.

Morally: need degree heterogeneity so that X and $H^{-1}X$ are linearly independent

Technical conditions for partial identification result

- $\rho = \omega\left(\frac{\log^2 n}{\sqrt{n}}\right)$ and $\frac{\nu + b^2}{\rho} = \Theta(1)$
- $\min_{i \in [n]} |X_i^T \mathbb{E}[X_1]| = \omega\left(\frac{\log^2 n}{\sqrt{n}\rho}\right)$ almost surely.
- $\max_{i \in [n]} \|X_i\| = o(\sqrt{n})$ almost surely.
- $\mathbb{E}\|X_1\|^2 < \infty$.

References

Bertrand, M., E. F. P. Luttmer, and S. Mullainathan (2000, August). Network Effects and Welfare Cultures. *The Quarterly Journal of Economics* 115(3), 1019–1055.

Boucheron, S., G. Lugosi, and P. Massart (2013, February). *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press.

Bramoullé, Y., H. Djebbari, and B. Fortin (2009, May). Identification of peer effects through social networks. *Journal of Econometrics* 150(1), 41–55.

Carrell, S. E., B. I. Sacerdote, and J. E. West (2013). From Natural Variation to Optimal Policy? The Importance of Endogenous Peer Group Formation. *Econometrica* 81(3), 855–882.

Duflo, E. and E. Saez (2003, August). The Role of Information and Social Interactions in Retirement Plan Decisions: Evidence from a Randomized Experiment. *The Quarterly Journal of Economics* 118(3), 815–842.

Epple, D. and R. E. Romano (2011, January). Chapter 20 - Peer Effects in Education:

A Survey of the Theory and Evidence. In J. Benhabib, A. Bisin, and M. O. Jackson (Eds.), *Handbook of Social Economics*, Volume 1, pp. 1053–1163. North-Holland.

Glaeser, E. L., B. Sacerdote, and J. A. Scheinkman (1996, May). Crime and Social Interactions*. *The Quarterly Journal of Economics* 111(2), 507–548.

Hayes, A. and K. Levin (2024, October). Peer effects in the linear-in-means model may be inestimable even when identified.

Katsouris, C. (2024, January). Robust Estimation in Network Vector Autoregression with Nonstationary Regressors.

Lee, L.-F. (2004, November). Asymptotic Distributions of Quasi-Maximum Likelihood Estimators for Spatial Autoregressive Models. *Econometrica* 72(6), 1899–1925.

Lewbel, A. (2019, December). The Identification Zoo: Meanings of Identification in Econometrics. *Journal of Economic Literature* 57(4), 835–903.

- Maclaren, O. J. and R. Nicholson (2020, July). What can be estimated? Identifiability, estimability, causal inference and ill-posed inverse problems. *arXiv:1904.02826 [cs, math, stat]*.
- McFowland, E. and C. R. Shalizi (2021, July). Estimating Causal Peer Influence in Homophilous Social Networks by Inferring Latent Locations. *Journal of the American Statistical Association* 0(0), 1–12.
- Patacchini, E. and Y. Zenou (2012, April). Juvenile Delinquency and Conformism. *The Journal of Law, Economics, and Organization* 28(1), 1–31.
- Paul, S., S. Nath, and K. Warren (2022, March). Causal Network Influence with Latent Homophily and Measurement Error: An Application to Therapeutic Community. *arXiv:2203.14223 [stat]*.
- Rose, C. D. (2017, October). Identification of peer effects through social networks using variance restrictions. *The Econometrics Journal* 20(3), S47–S60.

- Sacerdote, B. (2001, May). Peer effects with random assignment: Results for dartmouth roommates. *The Quarterly Journal of Economics* 16(2), 681–704.
- Soeteven, A. R. and P. Kooreman (2007, April). A discrete-choice model with social interactions: With an application to high school teen behavior. *Journal of Applied Econometrics* 22(3), 599–624.
- Trogdon, J. G., J. Nonnemaker, and J. Pais (2008, September). Peer effects in adolescent overweight. *Journal of Health Economics* 27(5), 1388–1399.
- Vazquez-Bare, G. (2023, November). Identification and estimation of spillover effects in randomized experiments. *Journal of Econometrics* 237(1), 105237.
- Young, S. J. and E. R. Scheinerman (2007). Random Dot Product Graph Models for Social Networks. In A. Bonato and F. R. K. Chung (Eds.), *Algorithms and Models for the Web-Graph*, Volume 4863, pp. 138–149. Berlin, Heidelberg: Springer Berlin Heidelberg.

Zhu, X., R. Pan, G. Li, Y. Liu, and H. Wang (2017, June). Network vector autoregression. *The Annals of Statistics* 45(3), 1096–1123.