Asymptotic unidentifiability of peer effects in the linear-in-means model

Alex Hayes | PhD Defense

2024-04-04 @ 12:30 pm
Service Memorial Institute 133, Medical Sciences Center, UW-Madison

Department of Statistics, University of Wisconsin-Madison

What are peer effects?

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

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

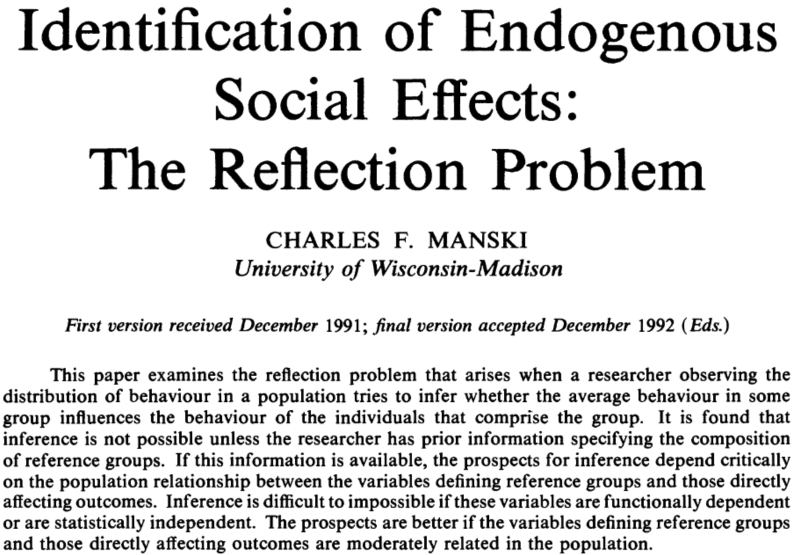
Interference: if my friends get vaccinated, I am less likely to get sick

* Not a causal talk. But causally inspired.

The canonical linear model for peer effects

$$\underbrace{Y_i}_{\text{sick?}} = \underbrace{\alpha}_{\substack{\text{base}\\\text{rate}}} + \underbrace{\beta}_{\substack{\text{contagion}\\\text{effect}}} \underbrace{\sum_{i\neq j} \frac{A_{ij}}{d_i} Y_j}_{\substack{\text{portion}\\\text{sick}\\\text{peers}}} + \underbrace{\gamma}_{\substack{\text{direct}\\\text{effect}}} \underbrace{T_i}_{\text{vaccinated?}} + \underbrace{\delta}_{\substack{\text{interference}\\\text{effect}}} \underbrace{\sum_{i\neq j} \frac{A_{ij}}{d_i} T_j}_{\substack{\text{portion}\\\text{vaccinated}\\\text{peers}}} + \underbrace{\varepsilon_i}_{\text{error}}$$

Outcome	Y_i	$\in \{0,1\}$
Treatment	T_i	$\in \{0,1\}$
Adjacency matrix	A	$\in \{0,1\}^{n\times n}$
Edge $i \sim j$	A_{ij}	$\in \{0,1\}$
Node degree	d_i	$\in \mathbb{Z}^+$



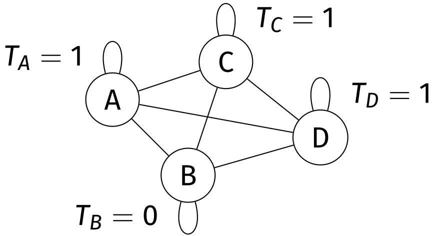
Identification of Endogenous Social Effects: The Reflection Problem

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.

Manski's reflection problem: highly structured networks break the model

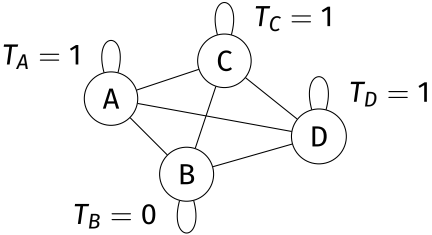


Average value of *T* amongst peers is the same for all nodes!

$$GT_A = 3/4$$
$$GT_B = 3/4$$
$$GT_C = 3/4$$
$$GT_D = 3/4$$

Problem: cannot distinguish base rate α from interference effect δ

Manski's reflection problem: highly structured networks break the model



$$\begin{bmatrix}Y_A\\Y_B\\Y_C\\Y_D\end{bmatrix} = \begin{array}{c}\begin{array}{cccc}1_n & GY & T & GT\end{array}\\ \underbrace{\begin{bmatrix}1 & GY_A & 1 & 3/4\\ 1 & GY_B & 0 & 3/4\\ 1 & GY_C & 1 & 3/4\\ 1 & GY_D & 1 & 3/4\end{bmatrix}}_{W}\end{array} \begin{bmatrix}\alpha\\ \beta\\ \gamma\\ \delta\end{bmatrix} + \begin{bmatrix}\varepsilon_A\\ \varepsilon_B\\ \varepsilon_C\\ \varepsilon_D\end{bmatrix}$$

Problem: Design matrix W becomes collinear!

Identification

Define the degree matrix $D = \mathrm{diag}(d_1, d_2, \ldots, d_n)$, where $d_i = \sum_j A_{ij}$. Let $G = D^{-1}A$ be the row-normalized adjacency matrix. Then

$$Y = \alpha 1_n + \beta GY + T\gamma + GT\delta + \varepsilon.$$

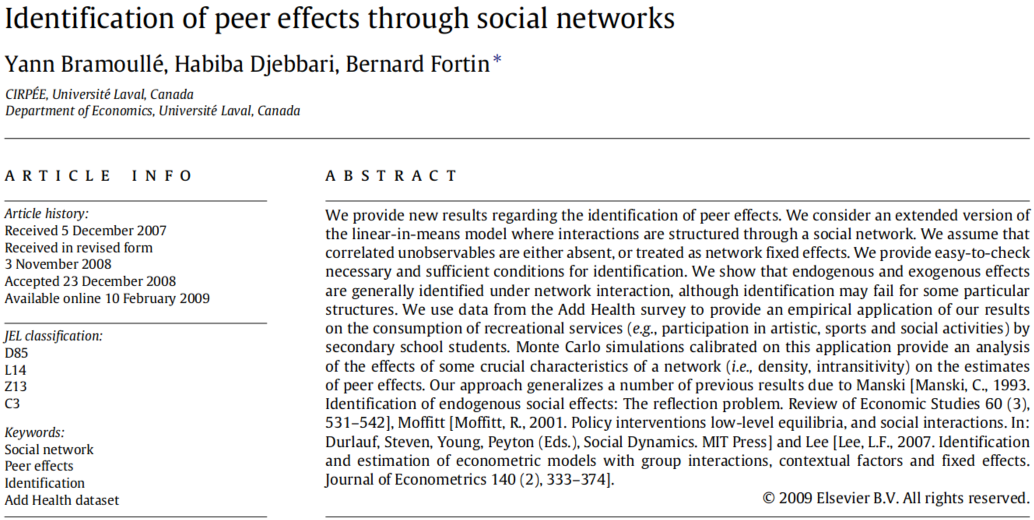
Definition

We say that $(\alpha, \beta, \gamma, \delta)$ are identified when the columns of the design matrix

$$W_n = \begin{bmatrix} 1_n & GY & T & GT \end{bmatrix}.$$

are linearly independent. Otherwise, we say that $(\alpha, \beta, \gamma, \delta)$ are unidentified.

We assume that $\mathbb{E}\left[\varepsilon | W_n\right] = 0$.



Identification of peer effects through social networks

Yann Bramoullé, Habiba Djebbari, Bernard Fortin*

CIRPÉE, Université Laval, Canada
Department of Economics, Université Laval, Canada

ARTICLE INFO

Article history:
Received 5 December 2007
Received in revised form
3 November 2008
Accepted 23 December 2008
Available online 10 February 2009

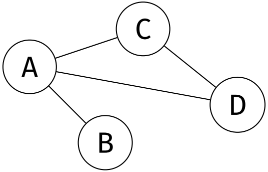
JEL classification:
D85
L14
Z13
C3

Keywords:
Social network
Peer effects
Identification
Add Health dataset

ABSTRACT

We provide new results regarding the identification of peer effects. We consider an extended version of the linear-in-means model where interactions are structured through a social network. We assume that correlated unobservables are either absent, or treated as network fixed effects. We provide easy-to-check necessary and sufficient conditions for identification. We show that endogenous and exogenous effects are generally identified under network interaction, although identification may fail for some particular structures. We use data from the Add Health survey to provide an empirical application of our results on the consumption of recreational services (*e.g.*, participation in artistic, sports and social activities) by secondary school students. Monte Carlo simulations calibrated on this application provide an analysis of the effects of some crucial characteristics of a network (*i.e.*, density, intransitivity) on the estimates of peer effects. Our approach generalizes a number of previous results due to Manski [Manski, C., 1993. Identification of endogenous social effects: The reflection problem. Review of Economic Studies 60 (3), 531–542], Moffitt [Moffitt, R., 2001. Policy interventions low-level equilibria, and social interactions. In: Durlauf, Steven, Young, Peyton (Eds.), Social Dynamics. MIT Press] and Lee [Lee, L.F., 2007. Identification and estimation of econometric models with group interactions, contextual factors and fixed effects. Journal of Econometrics 140 (2), 333–374].

Bramoullé: intransivity (i.e, open triangles) fixes the problem



A ↔ *B* ↔ *D* is an intransitive triangle. If *B* were friends with *D* it would “close” the triangle

Since 2009: as long as there is intransitivity in the network, $(\alpha, \beta, \gamma, \delta)$ are identified, practitioners are good to use the linear-in-means model

Reflection problem solved!

Proposition (Finite sample identification, Bramoullé et al. (2009))

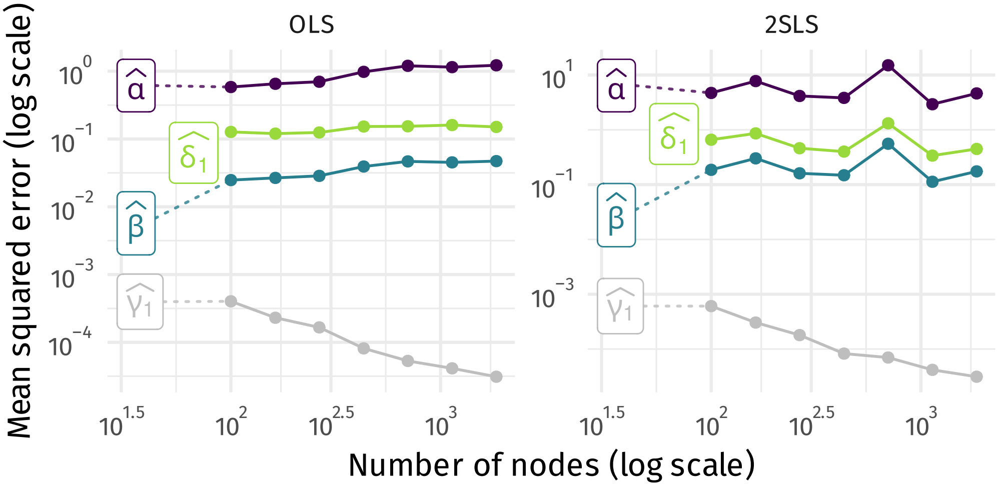
Let ε be mean zero, i.i.d. noise and let

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

Suppose that $|\beta| < 1$ and $\gamma\beta + \delta \neq 0$. If I, G and G^2 are linearly independent, in the sense that $aI + bG + cG^2 = 0$ requires $a = b = c = 0$, then α, β, γ and δ are identified.

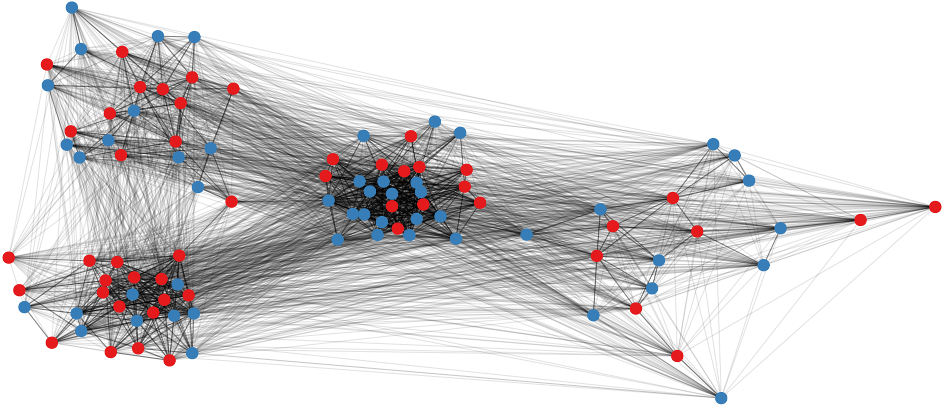
$\gamma\beta + \delta \neq 0$ means that there is either some interference effect, or some direct effect and some contagion effect, and if there are both, they don't cancel each other out.

Simulations on an intransitive network



Randomized experiment on a stochastic blockmodel

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



SBM has four blocks and mild degree correction

The problem: interference effect is an average of i.i.d. random variables

Even though the network is intransitive, heading towards a situation like

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ \vdots \\ Y_n \end{bmatrix} = \begin{array}{c} \begin{matrix} 1_n & GY & T & GT \end{matrix} \\ \begin{bmatrix} 1 & GY_1 & 1 & 1/2 \\ 1 & GY_2 & 0 & 1/2 \\ 1 & GY_3 & 1 & 1/2 \\ 1 & GY_4 & 0 & 1/2 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & GY_n & 1 & 1/2 \end{bmatrix} \end{array} \begin{bmatrix} \alpha \\ \beta \\ \gamma \\ \delta \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

The *GT* term converges to a constant

$$\sum_{i \neq j} \frac{A_{ij}}{d_i} T_j \to \frac{1}{2}$$

as $d_i \to \infty$

GY term also becomes colinear with intercept

Recall

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

If $|\beta| < 1$, then $I - \beta G$ is invertible and there is a unique solution

$$\begin{aligned} Y &= (I - \beta G)^{-1}(\alpha 1_n + \gamma T + \delta GT + \varepsilon) \\ &= \sum_{k=0}^{\infty} \beta^k G^k (\alpha 1_n + \gamma T + \delta GT + \varepsilon) \qquad \text{since } (I - \beta G)^{-1} = \sum_{k=0}^{\infty} \beta^k G^k \end{aligned}$$

* This “reduced form” of the linear-in-means model also characterizes the data generating process

GY term also becomes colinear with intercept

Suppose no node is isolated. Then, the reduced form of Y is given by

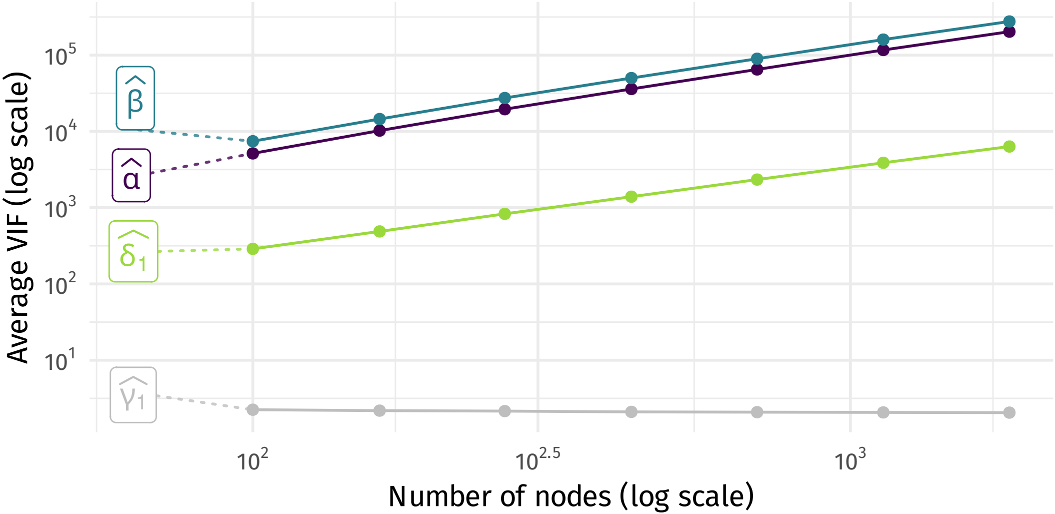
$$Y = \frac{\alpha}{1-\beta} 1_n + \gamma T + (\gamma\beta + \delta) \sum_{k=0}^{\infty} \beta^k G^{k+1} T + \sum_{k=0}^{\infty} \beta^k G^k \varepsilon$$

and further

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

If neighborhood averages GT are converging, repeated neighborhood averages converge as well, because $G1_n = 1_n$ (neighborhood average of constants is constant)

1_n, GY, and GT become collinear as n → ∞



Definition

We say that $(\alpha, \beta, \gamma, \delta)$ are asymptotically identified when the design matrix (1) converges to a limit object W in the sense that

$$\max_{ij}\left|\begin{bmatrix}1_n & GY & T & GT\end{bmatrix}_{ij} - W_{ij}\right| = o(1)$$

and the columns of the W are linearly independent. If the columns of W are linearly dependent, we say that $(\alpha, \beta, \gamma, \delta)$ are asymptotically unidentified.

Assumption (1)

Suppose that

1. $T_1, T_2, \ldots, T_n$ *are independent with shared mean* $\zeta \in \mathbb{R}$*, and* T *is independent of* A*.*
2. $\{T_i - \zeta : i \in [n]\}$ *are independent* (ν, b)*-subgamma random variables.*
3. $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ *are independent subgamma random variables with parameters not depending on* n*.*
4. A *is such that*

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

The fourth condition requires that the size of each neighborhood is growing. When the network is binary, such that $A_{ij} \in \{0, 1\}$, it reduces to $\min_{i\in[n]} d_i = \omega(\log n)$. That is, the size of the smallest neighborhood must grow at

Lemma (1)

Under Assumption 1,

$$\max_{i\in[n]} \left\| [GT]_i - \zeta \right\| = o(1) \ \ \textit{almost surely}$$

and there exists $\eta = \eta(\zeta, \alpha, \beta, \gamma, \delta) \in \mathbb{R}$ *such that*

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

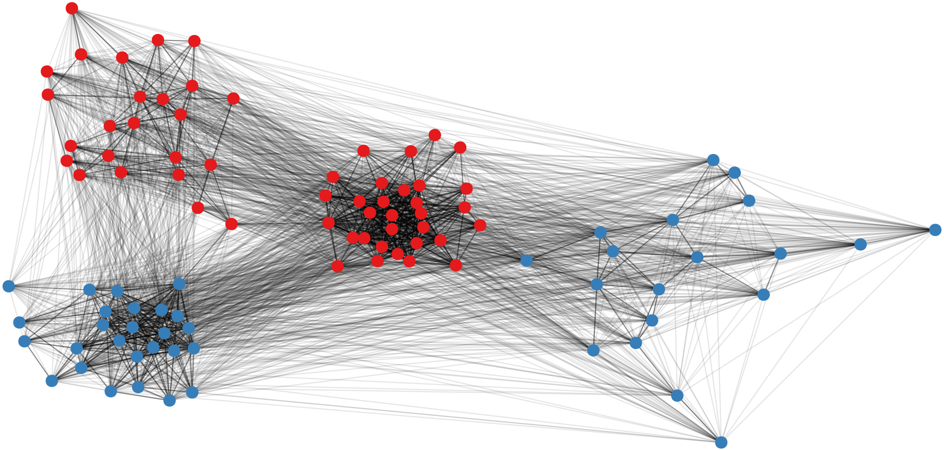
Theorem (1)

Under Assumption 1, α, β *and* δ *are asymptotically unidentified.*

Caveats

1. Isolated nodes: If all the connected components that are not singletons satisfying Assumption 1, can recover α but β and δ are still aliased.
2. Sparse networks: Our results don't cover sparse networks!
3. Non-random covariates: Our results don't cover fixed *T*!
4. Independence: What if *T* and the network *A* are dependent?

Treatments dependent on block membership



SBM has four blocks and mild degree correction

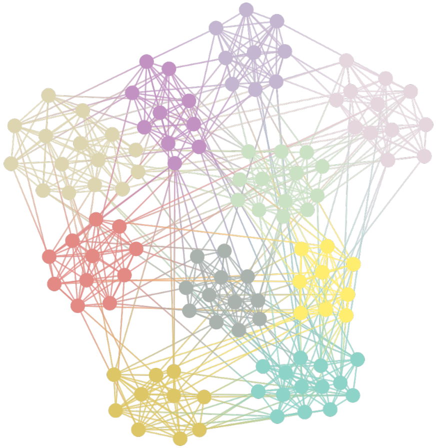
Why dependence might help

The *GT* term is not longer an average of independent observations

$$\sum_{i \neq j} \frac{A_{ij}}{d_i} T_j$$

so *GT* might not converge to $\mathbb{E}[T]$ for every node.

Intuition: stochastic blockmodels



d communities or "blocks"

$X_{i\cdot} \in \{0,1\}^d$ one-hot indicator of node i's block

X is latent (i.e. unobserved)

$B \in [0,1]^{d \times d}$ inter-block edge probabilities

Friendships depend on group memberships and B

$$\mathbb{P}\left(A_{ij} = 1 \,\middle|\, X\right) = X_{i\cdot} B X_{j\cdot}^T$$

Theorem

Suppose that (A, X) are sampled from a random dot product model where X is rank d with probability 1. Let ε be a vector of mean zero, i.i.d. $(\nu_\varepsilon, b_\varepsilon)$-subgamma random variables, with $(\nu_\varepsilon, b_\varepsilon)$ not depending on n, and let

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

for $\alpha, \beta \in \mathbb{R}$ and $\gamma, \delta \in \mathbb{R}^d$. Suppose that X has $k \geq 2d$ distinct rows. Then, under suitable technical 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 $(\alpha, \beta, \delta_1, ..., \delta_d)$ are set to zero in the data generating process, the limit object of W_n is a matrix with full rank, and $\alpha, \beta, \gamma, \delta$ are thus asymptotically identifiable.

Simulation details

All networks in the simulations are generated from a Poisson degree-corrected stochastic blockmodel with n nodes and four equally probably blocks (see Example **??**). The edge formation matrix $B \in [0, 1]^{4 \times 4}$ is set to

$$B = \begin{bmatrix} 0.5 & 0.05 & 0.05 & 0.05 \\ 0.05 & 0.5 & 0.05 & 0.05 \\ 0.05 & 0.05 & 0.5 & 0.05 \\ 0.05 & 0.05 & 0.05 & 0.5 \end{bmatrix},$$

and ρ is such that the expected mean degree of the network $2n^{0.7}$. $\theta_1, ..., \theta_n$ are sampled independently from a continuous uniform distribution supported on $[1, 2]$.

Simulation details, cont'd

We consider three distinct generative models for nodal outcomes $Y_1, ..., Y_n$.

1. *Randomized controlled trial*: The linear-in-means model with a single nodal covariate $T_i \overset{\text{iid}}{\sim}$ Bernoulli(0.5) for all nodes, which are sampled independently of the network. The regression model is thus

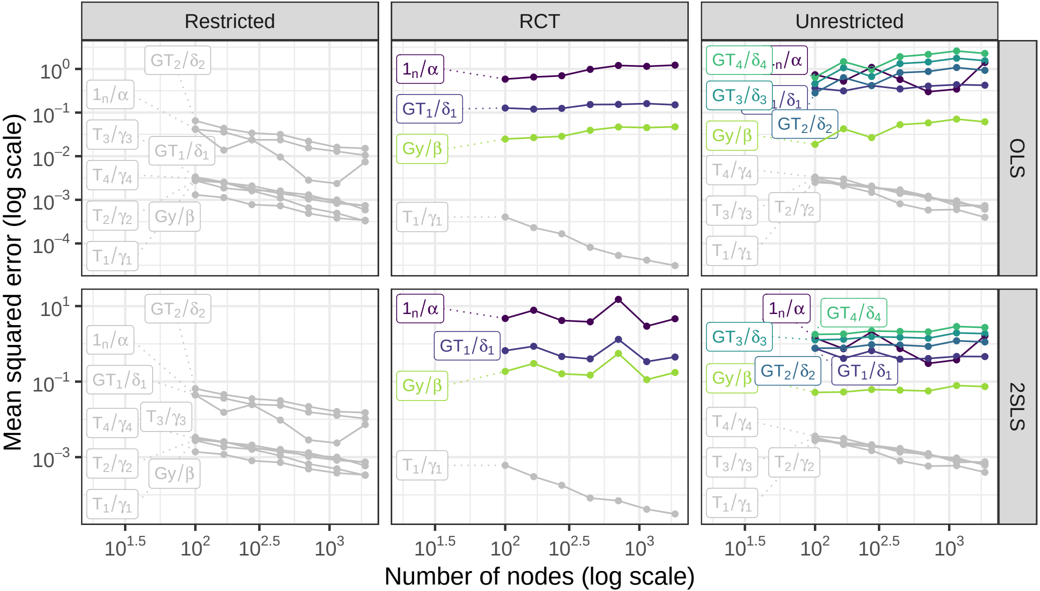
$$Y = 1_n\alpha + GY\beta + T\gamma + GT\delta + \varepsilon,$$

and we fix $\alpha = 3, \beta = 0.2, \gamma = 4, \delta = 2$ and $\varepsilon \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$ with $\sigma = 0.1$. α, β and δ are asymptotically unidentified in this model.

2. *Unrestricted*: The linear-in-means model, where the nodal covariates at the latent positions of the stochastic blockmodel, that is $T_i = X_i \in \mathbb{R}^4$. Recall that $X_i \in R^4$ where $X = US^{1/2}$ and USU^T is the eigendecomposition of $\mathbb{E}[A \,|\, z(1), ..., z(n), \theta]$. The nodal regression model is thus

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

where we again set $\alpha = 3$, $\beta = 0.2$ and $\varepsilon \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$ with $\sigma = 0.1$. Since


Restricted
RCT
Unrestricted
OLS
2SLS
Mean squared error (log scale)
Number of nodes (log scale)
10^0
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^1
10^{-1}
10^{-3}
$10^{1.5}$
10^2
$10^{2.5}$
10^3
GT_2/δ_2
$1_n/\alpha$
T_3/γ_3
GT_1/δ_1
T_4/γ_4
T_2/γ_2
Gy/β
T_1/γ_1
$1_n/\alpha$
GT_1/δ_1
Gy/β
T_1/γ_1
GT_4/δ_4
$_n/\alpha$
GT_3/δ_3
$1/\delta_1$
GT_2/δ_2
Gy/β
T_4/γ_4
T_3/γ_3
T_2/γ_2
T_1/γ_1
GT_2/δ_2
$1_n/\alpha$
GT_1/δ_1
T_4/γ_4
T_3/γ_3
T_2/γ_2
Gy/β
T_1/γ_1
$1_n/\alpha$
GT_1/δ_1
Gy/β
T_1/γ_1
$1_n/\alpha$
GT_4/δ_4
GT_3/δ_3
Gy/β
GT_2/δ_2
GT_1/δ_1
T_4/γ_4
T_3/γ_3
T_2/γ_2
T_1/γ_1

Why this matters

Paul et al. (2022)

1. Shalizi and Thomas (2011); McFowland and Shalizi (2021): homophily and contagion are non-parametrically confounded. Must make parametric assumptions to estimate contagion effects.
2. Bramoullé et al. (2020): most parametric contagion identification theory considers fixed n setting but not the $n \to \infty$ limit. Many results won't hold under stochastic blockmodels.

Open questions

- Do *GT* and *GY* ever not converge to constants?
- Are longitudinal contagion models also affected?
- What happens in sparse networks asymptotically?

Thank you! Questions?

Stay in touch

- @alexpghayes
- alex.hayes@wisc.edu
- *https://www.alexpghayes.com*
- *https://github.com/alexpghayes*

References

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

Bramoullé, Y., H. Djebbari, and B. Fortin (2020). Peer Effects in Networks: A Survey. *Annual Review of Economics 12*(1), 603–629.

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.

Paul, S., S. Nath, and K. Warren (2022, August). Network Influence with Latent Homophily and Measurement Error.

Shalizi, C. R. and A. C. Thomas (2011, May). Homophily and Contagion Are Generically Confounded in Observational Social Network Studies. *Sociological Methods & Research 40*(2), 211–239.