Linear Regression with Centrality Measures

Yong Cai

Oct 26, 2022

Northwestern University

Introduction

- Network positions matter:
    - Better-networked VC firms successfully exit greater proportion of investments (Hochberg, Ljungqvist, and Lu, 2007)
    - Greater take-up when seeding microfinance to central villagers (Banerjee, Chandrasekhar, Duflo, and Jackson, 2013)
    - Central families overrepresented in political offices (Cruz, Labonne, and Querubin, 2017)

Introduction

- Motivates the regression:

$$Y_i = C_i\beta + X_i'\gamma + \varepsilon_i \quad , \quad E[X_i\varepsilon_i] = 0, E[C_i\varepsilon_i] = 0$$

  - Y_i is outcome
  - C_i is centrality – agent-level measure of importance in network
  - C_i not directly observed, but calculated on network data
  - β is parameter of interest

- Researchers typically estimate β by OLS; conduct inference using t-test
- Reasonable if procedure has good finite sample properties

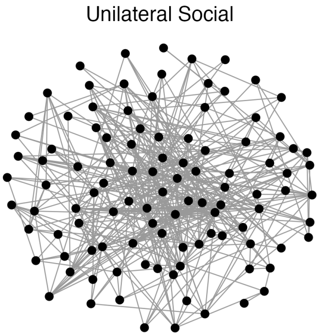
Introduction

- Networks may be sparse
  - Many more agents than links per agent
  - Because interactions or observations are rare
  - Not enough variation to identify β
  - Chandrasekhar (2016): sparsity is "stylized fact"
- Networks may be noisy
  - Often obtained by surveys or constructed using proxies
  - Analyses often treat measured network as true network
  - Ignoring measurement error leads to poor statistical properties

Introduction

- Measurement error and sparsity are common in economic networks
- E.g. Informal Insurance Network (De Weerdt and Dercon, 2006)
    - Want to know if informal insurance helps consumption smoothing
    - 119 households in rural Tanzania
    - A_{ij} is probability that i borrows from j
    - Different ways of defining proxies, with different amount of sparsity

Networks in Nyakatoke, Tanzania



$(n = 119)$	Mean	Median	Min	Max
Unilateral Social	8.02	7	1	31
Bilateral Social	2.30	2	0	10
Unilateral Financial	6.51	5	0	43

Table 1: Degree distributions of various networks in Nyakatoke. $\sqrt{119} = 10.9$.

This Paper

- Degree, diffusion and eigenvector centralities
- Cross-sectional: one large network
- Novel asymptotic framework with sparsity and measurement error
- More similar to data $\Rightarrow$ asymp. approx. more accurate in finite sample

Contributions

1. Show that OLS can become inconsistent under sparsity
    - Characterize threshold at which inconsistency occurs
    - Show that eigenvector less robust than degree and diffusion
2. Distributional theory under measurement error and sparsity
    - Even when consistent, OLS estimators are asymptotically biased
    - Asymptotic bias can be large relative to variance
    - Slower rate of convergence than reflected by robust standard errors
3. Novel bias correction and inference methods

Takeaways

- Sparsity and measurement error are challenges for network data
- Comparing significance involves both economic *and* statistical properties
- Eigenvector centrality particularly fragile

- Even when consistent, bias can be large and rate of convergence slow
- Use different estimators and inference methods

Related Literature

Related Literature

- Regression with network position on RHS:
  - **Eigenvectors:** denser than this paper: Le and Li (2020);
    sparser but i.i.d. Gaussian errors: Cai, Yang, Zhu, Shen, and Zhao (2021)
  - **Nonparametric:** dense case: Auerbach (2022)

- Centrality Statistics under Classical Measurement Error in Network
  - **Simulations:** Costenbader and Valente (2003);
    Borgatti, Carley, and Krackhardt (2006)
  - **Estimation of Centrality:** Dasaratha (2020);
    Avella-Medina, Parise, Schaub, and Segarra (2020)

Related Literature

- Non-Classical Measurement Error in Network
  - **Partial observation:** Chandrasekhar and Lewis (2016); Thirkettle (2019); Griffith (2022)
  - **"Small" error:** With 2SLS: Lewbel, Qu, Tang, et al. (2021)

- Econometrics on Sparse Networks
  - **Network Formation:** Jochmans (2018); Graham (2020b)
  - **Network Recovery:** Manresa (2016); Rose (2016); De Paula et al. (2020)

Table of Contents

Centrality Statistics

Centrality Statistics

- Network data is symmetric adjacency matrix A
- A_{ij} records intensity of relationship between i and j
- More intuitive when binary, but need not be

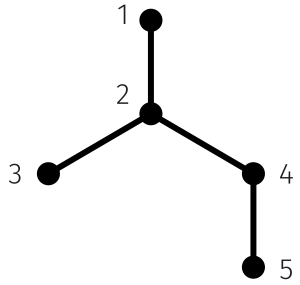
- When A is known, centrality statistics exactly computable
- Many different ways to measure importance $\Rightarrow$ many centrality statistics
- Focus on degree, diffusion and eigenvector

Degree:

$$C^{(1)}(A) = A\iota \ .$$

- $C_i^{(1)}$ is sum of row i of adjacency matrix
- If A binary, degree of i is number of links

$$C^{(1)} = \begin{pmatrix} 1 \\ 3 \\ 1 \\ 2 \\ 1 \end{pmatrix}$$



$$A = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

Centrality Statistics

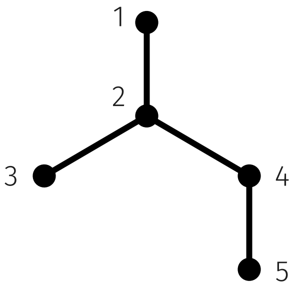
Diffusion: For a given $T \in \mathbb{N}$ and $\delta > 0$,

$$C^{(T)}(A) = \left(\sum_{t=1}^{T} \delta^t A^t \right) \iota$$

- If A binary, $(A^t)_{ij}$ is number of walks from i to j in t-steps
- Number of ways i can reach j over t periods
- δ^t is decay of influence over time
- $C_i^{(T)}$ is weighted sum of the agents that i can reach over T periods

Let $\delta = 0.5$, $T = 3$.

$$A^2 = \begin{pmatrix} 1 & 0 & 1 & 1 & 0 \\ 0 & 3 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 2 & 0 \\ 0 & 1 & 0 & 0 & 1 \end{pmatrix} \quad , \quad A^3 = \begin{pmatrix} 0 & 3 & 0 & 0 & 1 \\ 3 & 0 & 3 & 4 & 0 \\ 0 & 3 & 0 & 0 & 1 \\ 0 & 4 & 0 & 0 & 2 \\ 1 & 0 & 1 & 2 & 0 \end{pmatrix}$$

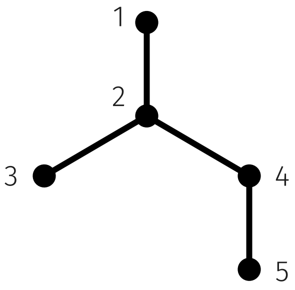


$$A = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

Centrality Statistics

Let $\delta = 0.5$, $T = 3$.

$$C^{(T)} = \begin{pmatrix} 1.75 \\ 3.75 \\ 1.75 \\ 2.75 \\ 1.5 \end{pmatrix}$$



$$A = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

Centrality Statistics

Eigenvector: Eigenvector centrality is the leading eigenvector of A, scaled

$$C^{(\infty)}(A) = a_n v_1(A)$$

- Want influence of agent proportional to influence of friends: For $k > 0$, seek

$$C_i^{(\infty)} \overset{\text{want}}{=} k \sum_{j \in [n]} A_{ij} C_j^{(\infty)}$$

- Eigenvectors of A solve the above equation; k is corresponding eigenvalue
- Perron-Frobenius Theorem: leading eigenvector uniquely non-negative

Centrality Statistics

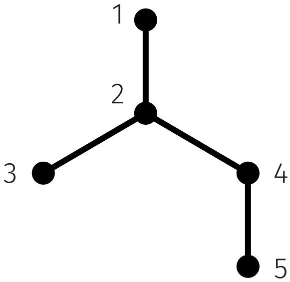
- Eigenvectors defined up to scale
- Various normalizations in literature, examples below
- Turns out that scaling matters for statistical properties
- Today: focus on $a_n = \lambda_1(A)$ (more in paper)

	Applied Work	Econometrics
$a_n = 1$	Banerjee et al. (2013) Cruz et al. (2017)	Dasaratha (2020)
$a_n = \sqrt{n}$	Chandrasekhar et al. (2018) Banerjee et al. (2019)	Avella-Medina et al. (2020) Cai et al. (2021)

Table 2: Examples of a_n in econometric theory and empirical work

$$C^{(\infty)} = a_n \begin{pmatrix} 0.35 \\ 0.65 \\ 0.35 \\ 0.50 \\ 0.27 \end{pmatrix}$$

- 1,3 and 5 all have 1 friend
- 1 and 3 are more central than 5 because they are friends with 2



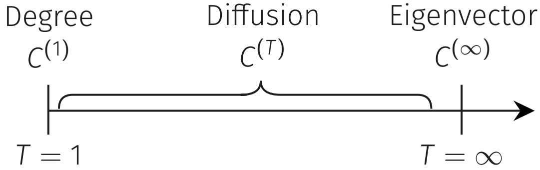
$$A = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

Centrality Statistics

- When $T = 1$, $C^{(1)}(A) \propto C^{(T)}(A)$
- Banerjee et al. (2019): If δ is larger than the leading eigenvalue of A,

$$\lim_{T\to\infty} C^{(T)}(A) \propto C^{(\infty)}(A)$$

- The statistics we study can thus be represented on a line (Bloch et al. 2021):



Model and Assumptions

Model and Assumptions

- For simplicity, consider regression without other covariates
- For $d \in \{1, T, \infty\}$:

$$Y_i = \beta^{(d)} C_i^{(d)} + \varepsilon_i^{(d)}$$

- $C^{(d)}$ is centrality statistic computed using network data
- $\beta^{(d)}$ is parameter of interest
- Will make enough assumptions so $\beta^{(d)}$ is slope of CEF

Model and Assumptions

- DGP yields i.i.d. draws of $\{(Y_i, U_i)\}$
- Y_i is observed outcome
- $U_i \sim U[0,1]$ is unobserved latent type to construct network

Model and Assumptions

- Let A be the $n \times n$ symmetric adjacency matrix
- For $f : [0,1]^2 \to [0,1]$, $p_n \in (0,1]$ and $j > i$, let

$$A_{ij} := p_n f(U_i, U_j)$$

- By symmetry, $A_{ij} = A_{ji}$; normalisation: $A_{ii} = 0$
- When A_{ij} is large, agents i and j have a strong relationship

- $p_n \to 0$ is sparsity parameter

Model and Assumptions

- In the set-up with no measurement error, A is observed
- Can exactly compute centrality measures
- Form the estimator

$$\tilde{\beta}^{(d)} = \frac{Y' C^{(d)}}{\left(C^{(d)}\right)' C^{(d)}}$$

Model and Assumptions

- With measurement error, observe $\hat{A}$, where for $j > i$,

$$\hat{A}_{ij} \,\middle|\, \mathbf{U} \overset{\text{iid}}{\sim} \text{Bernoulli}\left(p_n f(U_i, U_j)\right)$$

  Below the diagonal, $\hat{A}_{ij} = \hat{A}_{ji}$

- Use $\hat{A}$ as plug-in for A to compute $\hat{C}^{(d)}$
- Estimate

$$\hat{\beta}^{(d)} = \frac{Y'\hat{C}^{(d)}}{\left(\hat{C}^{(d)}\right)'\hat{C}^{(d)}}$$

Example

Suppose we want to study consumption smoothing and informal insurance network.

- Y_i: variance in consumption expenditure
- A_{ij}: probability that i lends j money or vice versa
- $\hat{A}_{ij}$: event in which i is observed lending j money

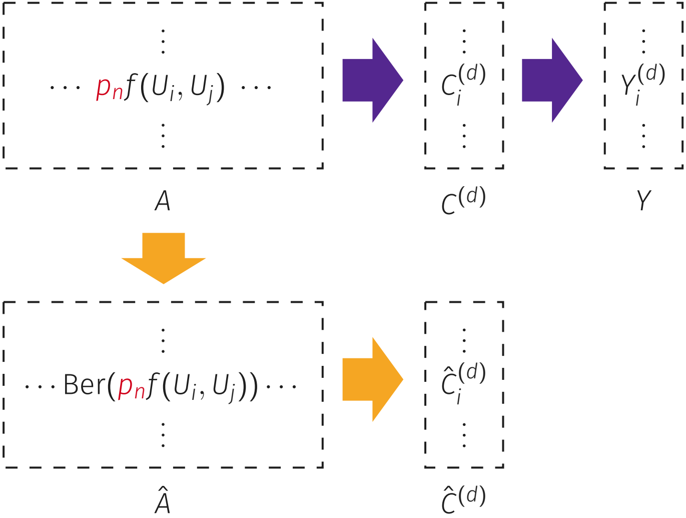
Model and Assumptions

- Want to understand the properties of $\tilde{\beta}^{(d)}$ and $\hat{\beta}^{(d)}$ when networks are sparse
- Theoretical device: $p_n \to 0$ as $n \to \infty$:
    - $A_{ij} \to 0$, $E\left[C_i^{(1)}\right] \ll n$
    - Many $\hat{A}_{ij}$'s are 0, $E\left[\hat{C}_i^{(1)}\right] \ll n$
- Standard model without p_n (see e.g. Graham 2020a, De Paula 2017)
- Use of p_n is common
    - **Statistics**: Bollobás, Janson, and Riordan (2007), Bickel and Chen (2009), etc
    - **Network Formation**: Jochmans (2018), Graham (2020b)

$\Rightarrow$ Use asymptotic framework that describes data to obtain better finite sample approximations

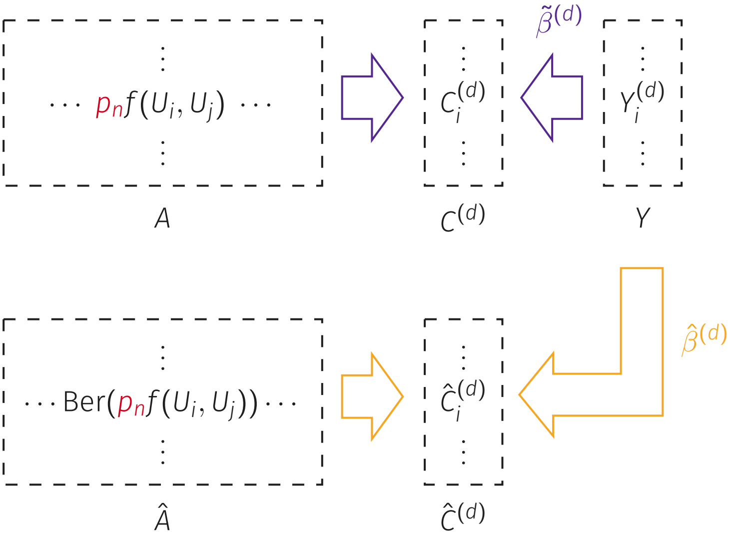
▸ Weak Ties ▸ Local Asymp.

DGP



$\cdots \; p_n f(U_i, U_j) \; \cdots$
A
$C_i^{(d)}$
$C^{(d)}$
$Y_i^{(d)}$
Y
$\cdots \mathrm{Ber}(p_n f(U_i, U_j)) \cdots$
$\hat{A}$
$\hat{C}_i^{(d)}$
$\hat{C}^{(d)}$

Observed Data

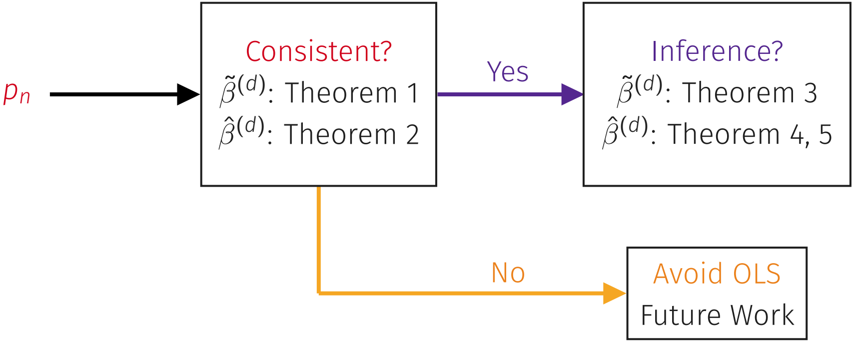


Preview of Results

- Main question: how do $\tilde{\beta}^{(d)}$ and $\hat{\beta}^{(d)}$ behave at as we vary p_n?

Preview of Results

- Main question: how do $\tilde{\beta}^{(d)}$ and $\hat{\beta}^{(d)}$ behave at as we vary p_n?



Theoretical Results

Theoretical Results

Consistency

Consistency without Measurement Error

Theorem 1 (No Measurement Error)

$\tilde{\beta}^{(1)}$ and $\tilde{\beta}^{(T)}$ are consistent if and only

$$p_n \gg n^{-\frac{3}{2}} \ .$$

$\tilde{\beta}^{(\infty)}$ is consistent if $a_n = \lambda_1(A)$.

(if $n = 100, E[C_i^{(1)}] \gg 0.1$)

- Inconsistent under extreme sparsity
- Consistency requires $\|C^{(d)}\|_2 \to \infty$ w.p.a. 1
- a_n inflates eigenvector, counters sparsity

Consistency with Measurement Error

Theorem 2 (Measurement Error)

$\hat{\beta}^{(1)}$ and $\hat{\beta}^{(T)}$ are consistent if and only if

$$p_n \gg n^{-1} .$$

(*if* $n = 100$, $E[C_i^{(1)}] \gg 1$)

- $\hat{\beta}^{(1)}$, $\hat{\beta}^{(T)}$ less robust to sparsity than $\tilde{\beta}^{(1)}$, $\tilde{\beta}^{(T)}$

Consistency with Measurement Error

Theorem 2 (Measurement Error)

$\hat{\beta}^{(\infty)}$ consistent if

$$p_n \gg n^{-1}\sqrt{\frac{\log n}{\log\log n}}\ .$$

It is inconsistent if

$$n^{-1}(\log\log n)^4 \ll p_n \ll n^{-1}\sqrt{\frac{\log n}{\log\log n}}\ .$$

(if $n = 100, E[C_i^{(1)}] \gg 2.6$*)*

Consistency with Measurement Error

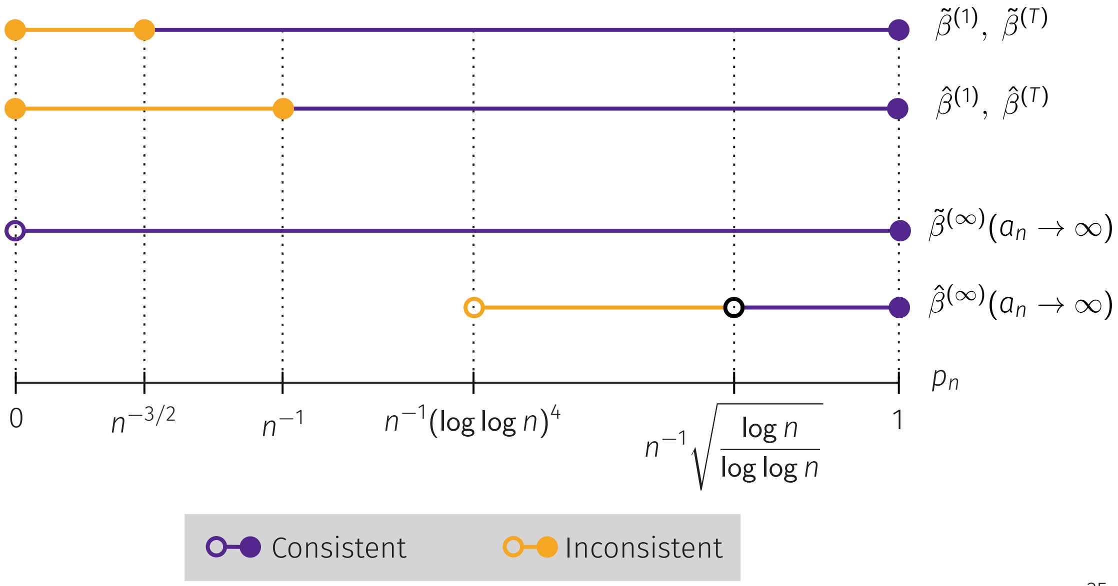
- With measurement error, p_n now matters
- Eigenvector centrality less robust to sparsity than degree and diffusion
- In sparse regimes, centralities differ in economic significance *and* statistical properties
- Compare significance with caution

Consistency with Measurement Error

- Eigenvector localization occurs at the consistency threshold(Alt et al. 2021a,b)
- Eigenvectors of $\hat{A}$ concentrate on agent with the highest realized degree
- Pure noise; not informative about eigenvectors of A

- Not clear if eigenvectors of $\hat{A}$ have describable structure below lower threshold

Summary



$\tilde{\beta}^{(1)}, \tilde{\beta}^{(T)}$
$\hat{\beta}^{(1)}, \hat{\beta}^{(T)}$
$\tilde{\beta}^{(\infty)}(a_n \to \infty)$
$\hat{\beta}^{(\infty)}(a_n \to \infty)$
p_n
0
$n^{-3/2}$
n^{-1}
$n^{-1}(\log\log n)^4$
$n^{-1}\sqrt{\frac{\log n}{\log\log n}}$
1
Consistent
Inconsistent

Theoretical Results

Distributional Theory

Theorem 3 (No Measurement Error)

Let $V_0^{(d)} = \sum_{i=1}^{n} \left(C_i^{(d)}\right)^2 \varepsilon_i^2$.

(a) *Suppose* $p_n \succ n^{-3/2}$. *Then, for* $d \in \{1, T\}$,

$$\frac{\tilde{\beta}^{(d)} - \beta^{(d)}}{\sqrt{V_0^{(d)}}} \xrightarrow{d} N(0,1)\ .$$

(b) *Suppose* $a_n = \lambda_1(A)$. *Then,*

$$\frac{\tilde{\beta}^{(\infty)} - \beta^{(\infty)}}{\sqrt{V_0^{(\infty)}}} \xrightarrow{d} N(0,1)\ .$$

Theorem 4 (With Measurement Error)

Suppose for $d \in \{1, T, \infty\}$ that $\hat{\beta}^{(d)}$ is consistent. Then if $\beta^{(d)} = 0$,

$$\frac{\hat{\beta}^{(d)}}{\sqrt{V_0^{(d)}}} \xrightarrow{d} N(0,1)$$

where $V_0^{(d)} = \sum_{i=1}^{n} \left(c_i^{(d)}\right)^2 \varepsilon_i^2$

- Plug-in estimation works for $V_0^{(d)}$ in the cases above
- Robust/hc t-statistic appropriate for $\tilde{\beta}^{(d)}$ for any null
- Appropriate for $\hat{\beta}^{(d)}$ if null is $\beta^{(d)} = 0$.

Theorem 4 (With Measurement Error)

For $d \in \{1, T\}$, let $p_n \gg n^{-1}$. Suppose $\beta^{(d)} \neq 0$. Then,

$$\frac{\hat{\beta}^{(d)} - \beta^{(d)}\left(1 - B^{(d)}\right)}{\sqrt{V^{(d)}}} \xrightarrow{d} N(0,1)$$

Distributional Theory

$$\underbrace{\sqrt{V_0^{(d)}}}_{O_p\left(n^{-3/2}p_n^{-1}\right)} \ll \underbrace{\sqrt{V^{(d)}}}_{O_p\left(n^{-1}p_n^{-1/2}\right)} \overset{p_n \to 0}{\ll} \underbrace{B^{(d)}}_{O_p\left(n^{-1}p_n^{-1}\right)}$$

- Measurement error slows down rate of convergence
- Bias can be much larger than variance if $p_n \to 0$

- Bias correction necessary for obtaining non-degenerate limit distribution
- hc/robust t-statistic not appropriate when $\beta^{(d)} \neq 0$
- t-statistic based confidence intervals are invalid

Distributional Theory

- $\hat{B}^{(d)}$ and $\hat{V}^{(d)}$ in paper
- Also adjusted tests and confidence intervals
- Bias-corrected estimator:

$$\check{\beta}^{(d)} = \frac{\hat{\beta}^{(d)}}{1 - \hat{B}^{(d)}}$$

Theorem 5

Under regularity conditions, suppose for $\eta > 0$ that

$$p_n \gg n^{-1} \left(\frac{\log n}{\log \log n} \right)^{\frac{1}{2}+\eta} \tag{1}$$

and $a_n = \lambda_1(A)$. Then,

$$\frac{\hat{\beta}^{(\infty)} - \beta^{(\infty)}}{\sqrt{V_0^{(\infty)}}} \xrightarrow{d} N(0,1) \ , \tag{2}$$

where

$$V_0^{(\infty)} = \sum_{i=1}^{n} \left(C_i^{(\infty)} \right)^2 \varepsilon_i^2$$

▸ Rank Assump.

Eigenvector Centrality

- Chose a_n so that distribution is easy to characterize
- Usual hc/robust t-statistic valid for all null hypotheses
- Trade-off: choosing a model with slower, known rate of convergence for one with faster, unknown rate

	Meas. Error		No Error
	$\beta^{(1)}\,/\,\beta^{(T)}$	$\beta^{(\infty)}$	$\beta^{(1)}\,/\,\beta^{(T)}\,/\,\beta^{(\infty)}$
$H_0 : \beta^{(d)} = 0$	t-test	t-test	t-test
$H_0 : \beta^{(d)} = b$	New Method		
Confidence Interval	New Method	t-stat based	t-stat based

Table 3: Inference under Sparsity. For $\hat{\beta}^{(\infty)}$, $a_n = \sqrt{\lambda_1(A)}$.

Simulations

Simulations

- Suppose $f = 1$ so that

$$A_{ij} = \begin{cases} p_n & \text{if } i \neq j, \\ 0 & \text{otherwise.} \end{cases}$$

- Our regression model is:

$$Y_i = \beta C_i^{(d)} + \varepsilon_i^{(d)}$$

$\varepsilon_i^{(d)} \overset{\text{i.i.d.}}{\sim} N(0, 1)$ and $\varepsilon_i^{(d)} \perp\!\!\!\perp \hat{A}_{jk}$ for all $i, j, k \in [n]$.

Inconsistency with Measurement Error

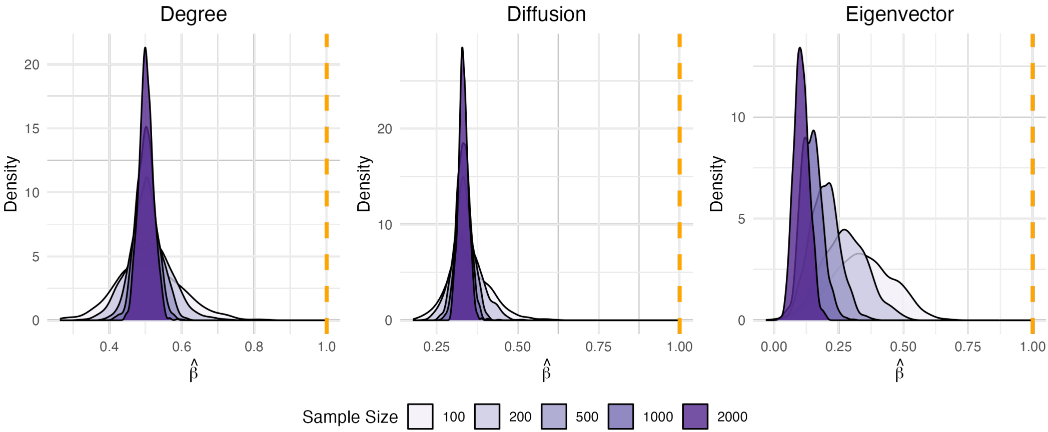


Figure 1: Distribution of $\hat{\beta}^{(d)}$ for $p_n = 1/n$. For $\tilde{\beta}^{(\infty)}$, $a_n = \sqrt{n}$. $\beta = 1$ (orange dashed line).

Bias correction is effective

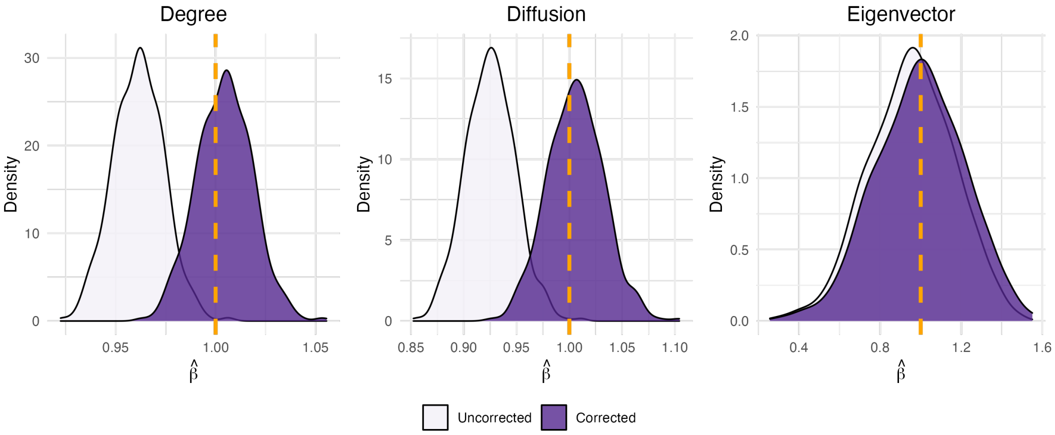


Figure 2: Distributions of $\hat{\beta}^{(d)}$ and their bias corrected versions $\check{\beta}^{(d)}$ for $p_n = 1/\sqrt{n}$, $n = 500$, $a_n = \sqrt{\hat{\lambda}_1(\hat{A})}$. $\beta = 1$ (orange dashed line).

Distributional theory is accurate

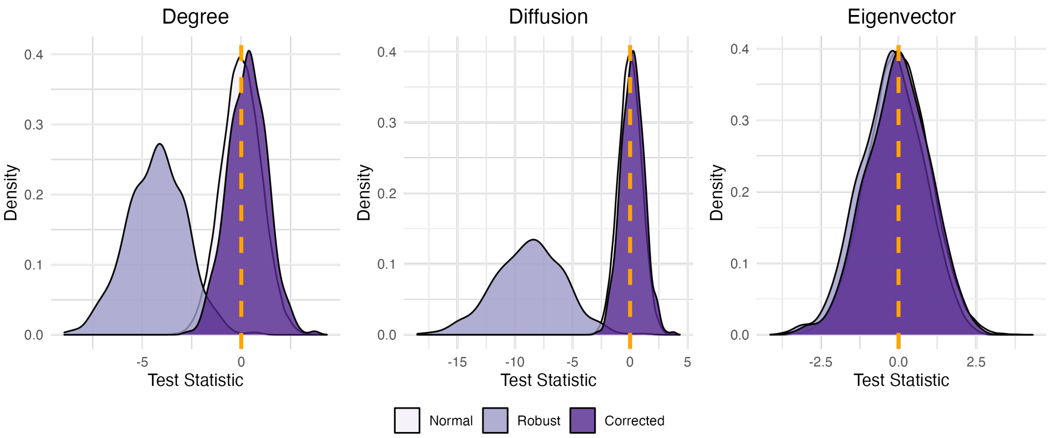


Figure 3: Distribution of the centered and scaled test statistics. Robust refers to tests based on t-statistic with robust (hc) standard errors. $p_n = 1/\sqrt{n}$, $n = 500$, $a_n = \sqrt{\hat{\lambda}_1(\hat{A})}$.

Adjusted tests are better

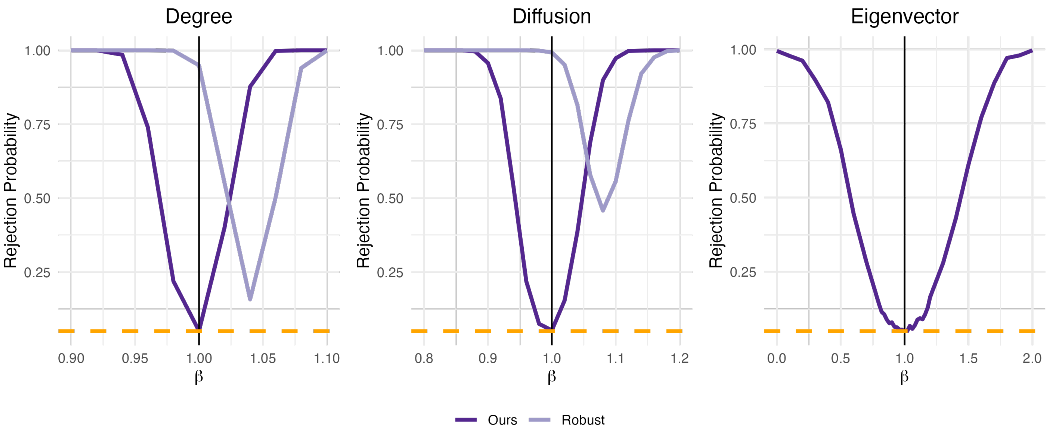


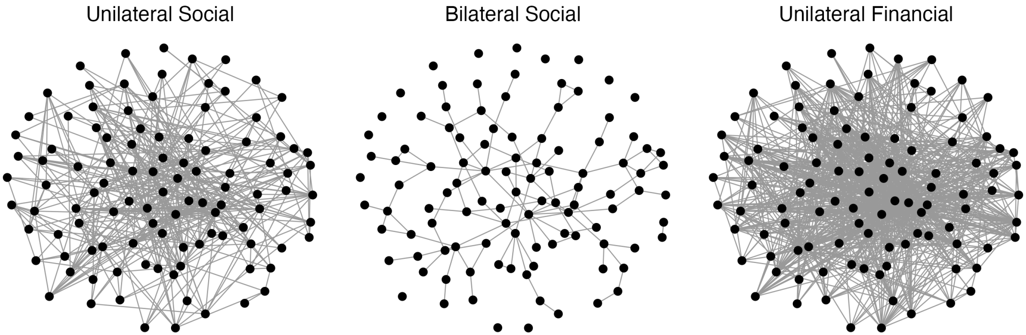
Figure 4: Power of the two-sided test of $H_0 : \beta = 1$ under various alternatives. Test at 5% level of significance (orange dashed line). $p_n = 1/\sqrt{n}$, $n = 500$, $a_n = \sqrt{\hat{\lambda}_1(\hat{A})}$.

Application

Application

- De Weerdt and Dercon (2006): want to know if informal insurance can help consumption smoothing
- Regress variance in food expenditure on centrality in network

Application



$(n = 119)$	Mean	Median	Min	Max
Unilateral Social	8.02	7	1	31
Bilateral Social	2.30	2	0	10
Unilateral Financial	6.51	5	0	43

Table 4: Degree distributions of various networks in Nyakatoke

		Estimate	p-value	Atten.	Bias Corr.
Unilateral Social	Degree	-1064	0.67	0.91	-1172
	Diffusion	-4274	0.77	1.00	-4290
	Eigenvector	-12353	0.86	0.91	-13548
Bilateral Social	Degree	-11604	0.06	0.74	-15592
	Diffusion	-23672	0.16	0.95	-24883
	Eigenvector	-10543	0.93	0.78	-13434
Unilateral Financial	Degree	-412	0.70	0.96	-429
	Diffusion	-4559	0.74	1.00	-4561
	Eigenvector	-15040	0.77	0.96	-15699

Table 5: Regression results. For diffusion, $\delta = 1/\sqrt{\lambda_1(\hat{A})}$, $T = 2$. For eigenvector, $a_n = \sqrt{\lambda_1(\hat{A})}$.

		90%	95%	99%
Degree	Robust	(-19500, ∞)	(-21700, ∞)	(-25900, ∞)
	Ours	(-18800, ∞)	(-20000, ∞)	(-22700, ∞)
Diffusion	Robust	(-45000, ∞)	(-51000, ∞)	(-62400, ∞)
	Ours	(-25200, ∞)	(-25300, ∞)	(-25500, ∞)

Table 6: One-sided confidence intervals for degree and diffusion in Bilateral Social network.

Application

- Bias correction increased estimate by as much as 25%; effect largest for sparsest network
- One-sided CI are tighter than robust CI
- For comparison, robust CI $\subset$ our two-sided CI

Conclusion

Conclusion

- Regression on degree, diffusion and eigenvector centrality
- Asymptotic framework with sparsity and measurement error

- Sparsity and measurement error bad for regression on all measures
- Eigenvector centrality most fragile + sensitive to scaling
- Caution when comparing across regressors in sparse networks

- OLS asymptotically biased even when consistent; converges more slowly
- Use bias correction and alternative inference methods

Thank you!

References

Alt, J., R. Ducatez, and A. Knowles (2021a). Extremal eigenvalues of critical Erdős–Rényi graphs. *The Annals of Probability 49*(3), 1347–1401.

Alt, J., R. Ducatez, and A. Knowles (2021b). Poisson statistics and localization at the spectral edge of sparse Erdős–Rényi graphs. *arXiv preprint arXiv:2106.12519*.

Athey, S., M. Bayati, N. Doudchenko, G. Imbens, and K. Khosravi (2021). Matrix completion methods for causal panel data models. *Journal of the American Statistical Association 116*(536), 1716–1730.

Auerbach, E. (2022). Identification and estimation of a partially linear regression model using network data. *Econometrica 90*(1), 347–365.

Avella-Medina, M., F. Parise, M. T. Schaub, and S. Segarra (2020). Centrality measures for graphons: Accounting for uncertainty in networks. *IEEE Transactions on Network Science and Engineering 7*(1), 520–537.

Banerjee, A., A. G. Chandrasekhar, E. Duflo, and M. O. Jackson (2013). The diffusion of microfinance. *Science 341*(6144), 1236498.

Banerjee, A., A. G. Chandrasekhar, E. Duflo, and M. O. Jackson (2019). Using gossips to spread information: Theory and evidence from two randomized controlled trials. *The Review of Economic Studies 86*(6), 2453–2490.

Bickel, P. J. and A. Chen (2009). A nonparametric view of network models and Newman–Girvan and other modularities. *Proceedings of the National Academy of Sciences 106*(50), 21068–21073.

Bloch, F., M. O. Jackson, and P. Tebaldi (2021). Centrality measures in networks. *arXiv preprint arXiv:1608.05845*.

Bollobás, B., S. Janson, and O. Riordan (2007). The phase transition in inhomogeneous random graphs. *Random Structures & Algorithms 31*(1), 3–122.

Borgatti, S. P., K. M. Carley, and D. Krackhardt (2006). On the robustness of centrality measures under conditions of imperfect data. *Social networks 28*(2), 124–136.

Cai, J., D. Yang, W. Zhu, H. Shen, and L. Zhao (2021). Network regression and supervised centrality estimation. *Available at SSRN 3963523*.

Candès, E. J. and T. Tao (2010). The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory 56*(5), 2053–2080.

Carvalho, V. M., M. Nirei, Y. U. Saito, and A. Tahbaz-Salehi (2021). Supply chain disruptions: Evidence from the great east japan earthquake. *The Quarterly Journal of Economics 136*(2), 1255–1321.

Chan, N. H. and C.-Z. Wei (1987). Asymptotic inference for nearly nonstationary ar (1) processes. *The Annals of Statistics*, 1050–1063.

Chandrasekhar, A. (2016). Econometrics of network formation. *The Oxford handbook of the economics of networks*, 303–357.

Chandrasekhar, A. and R. Lewis (2016). Econometrics of sampled networks. *Working Paper*.

Chandrasekhar, A. G., C. Kinnan, and H. Larreguy (2018). Social networks as contract enforcement: Evidence from a lab experiment in the field. *American Economic Journal: Applied Economics 10*(4), 43–78.

Chatterjee, S. (2015). Matrix estimation by universal singular value thresholding. *The Annals of Statistics 43*(1), 177–214.

Costenbader, E. and T. W. Valente (2003). The stability of centrality measures when networks are sampled. *Social networks 25*(4), 283–307.

Cruz, C., J. Labonne, and P. Querubin (2017). Politician family networks and electoral outcomes: Evidence from the philippines. *American Economic Review 107*(10), 3006–37.

Dasaratha, K. (2020). Distributions of centrality on networks. *Games and Economic Behavior 122*, 1–27.

De Paula, A. (2017). Econometrics of network models. In *Advances in economics and econometrics: Theory and applications, eleventh world congress*, pp. 268–323. Cambridge University Press Cambridge.

De Paula, A., I. Rasul, and P. Souza (2020). Recovering social networks from panel data: identification, simulations and an application.

De Weerdt, J. and S. Dercon (2006). Risk-sharing networks and insurance against illness. *Journal of development Economics 81*(2), 337–356.

Eagle, N., M. Macy, and R. Claxton (2010). Network diversity and economic development. *Science 328*(5981), 1029–1031.

Frydman, C. and E. Hilt (2017). Investment banks as corporate monitors in the early twentieth century United States. *American Economic Review 107*(7), 1938–70.

Graham, B. S. (2020a). Network data. In *Handbook of Econometrics*, Volume 7, pp. 111–218. Elsevier.

Graham, B. S. (2020b). Sparse network asymptotics for logistic regression.

Granovetter, M. S. (1973). The strength of weak ties. *American journal of sociology 78*(6), 1360–1380.

Griffith, A. (2022). Name your friends, but only five? the importance of censoring in peer effects estimates using social network data. *Journal of Labor Economics 40*(4), 779–805.

Hochberg, Y. V., A. Ljungqvist, and Y. Lu (2007). Whom you know matters: Venture capital networks and investment performance. *The Journal of Finance 62*(1), 251–301.

Holland, P. W., K. B. Laskey, and S. Leinhardt (1983). Stochastic blockmodels: First steps. *Social networks 5*(2), 109–137.

Jochmans, K. (2018). Semiparametric analysis of network formation. *Journal of Business & Economic Statistics 36*(4), 705–713.

Le, C. M., E. Levina, and R. Vershynin (2017). Concentration and regularization of random graphs. *Random Structures & Algorithms 51*(3), 538–561.

Le, C. M. and T. Li (2020). Linear regression and its inference on noisy network-linked data. *arXiv preprint arXiv:2007.00803*.

Lewbel, A., X. Qu, X. Tang, et al. (2021). *Social Networks with Mismeasured Links*. Boston College.

Manresa, E. (2016). Estimating the structure of social interactions using panel data. *Working Paper*.

Negahban, S. and M. J. Wainwright (2012). Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *The Journal of Machine Learning Research 13*(1), 1665–1697.

Rajkumar, K., G. Saint-Jacques, I. Bojinov, E. Brynjolfsson, and S. Aral (2022). A causal test of the strength of weak ties. *Science 377*(6612), 1304–1310.

Reagans, R. and E. W. Zuckerman (2001). Networks, diversity, and productivity: The social capital of corporate r&d teams. *Organization science 12*(4), 502–517.

Rose, C. (2016). Identification of spillover effects using panel data. Technical report, Working Paper.

Rothenberg, T. J. (1984). Approximating the distributions of econometric estimators and test statistics. *Handbook of econometrics 2*, 881–935.

Staiger, D. and J. H. Stock (1997). Instrumental variables regression with weak instruments. *Econometrica 65*(3), 557–586.

Thirkettle, M. (2019). Identification and estimation of network statistics with missing link data. *Working Paper*.

Young, S. J. and E. R. Scheinerman (2007). Random dot product graph models for social networks. In *International Workshop on Algorithms and Models for the Web-Graph*, pp. 138–149. Springer.

Regularized Eigenvectors

Regularized Eigenvectors

- Suppose $d = np_n$ is known. Define:

$$w_i := \min\left\{\frac{2d}{C_i^1(\hat{A})}, 1\right\}$$

- w_i is the ratio by which the degree of i exceeds $2d$.
- Let the regularized matrix $\tilde{A}$ be defined as follows:

$$\tilde{A}_{ij} = \sqrt{w_i w_j} \cdot \hat{A}_{ij}$$

- $\tilde{A}$ is the adjacency matrix in which we down-weight the links of high-degree agents so that degree is windsorized at $2d$.

Consistency with Measurement Error

- Le et al. (2017) show that this regularized matrix concentrates to A in spectral norm even under sparsity.
- Leads naturally to the following:

Proposition 1 (Regularized Eigenvector)

Suppose $a_n \to \infty$. The linear regressions of Y on $C^{(\infty)}(\tilde{A})$ is consistent if and only if

$$p_n \gg n^{-1}\ .$$

Additional Simulations

Scaling of Eigenvector Matters

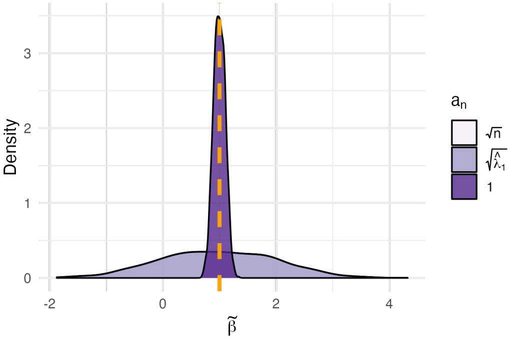


Figure 5: Distribution of $\tilde{\beta}^{(\infty)}$ for $n = 100$, $p_n = 1/n$ under various a_n. $\beta = 1$ (orange dashed line).

Consistency without Measurement Error

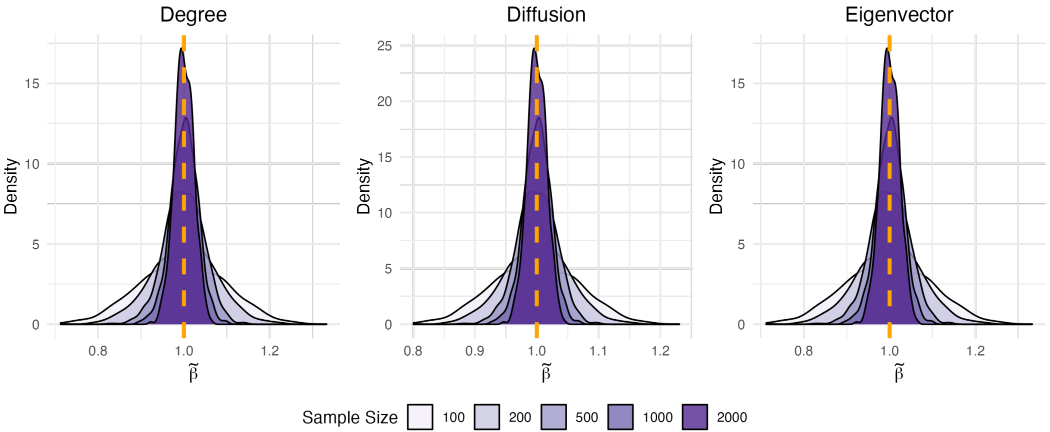


Figure 6: Distribution of $\tilde{\beta}^{(d)}$ for $p_n = 1/n$. For $\tilde{\beta}^{(\infty)}$, $a_n = \sqrt{n}$. $\beta = 1$ (orange dashed line).

Eigenvector is more sensitive to sparsity than Degree and Diffusion

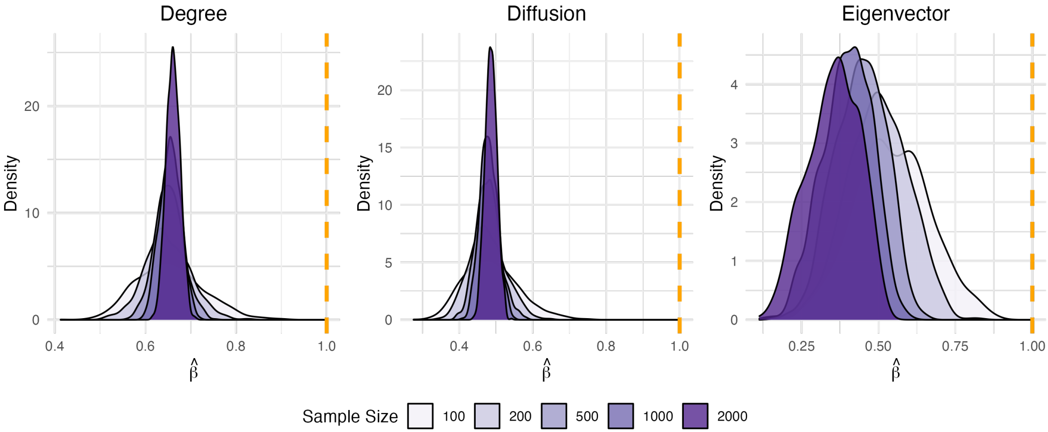


Figure 7: Distribution of $\hat{\beta}^{(d)}$ for $p_n = n^{-1}\sqrt{\log n / \log\log n}$. For $\tilde{\beta}^{(\infty)}$, $a_n = \sqrt{n}$. $\beta = 1$ (orange dashed line).

Size of $H_0 : \beta = 1$

p_n	Statistic		100	200	500	1000	2000
			Sample Size				
0.1	Degree	Ours	0.055	0.052	0.067	0.062	0.065
		Robust	0.656	0.673	0.690	0.668	0.674
	Diffusion	Ours	0.049	0.053	0.064	0.059	0.060
		Robust	0.889	0.894	0.887	0.871	0.898
	Eigenvector		0.045	0.043	0.037	0.056	0.044
$n^{-1/3}$	Degree	Ours	0.066	0.065	0.067	0.058	0.065
		Robust	0.330	0.450	0.573	0.705	0.783
	Diffusion	Ours	0.080	0.070	0.074	0.057	0.064
		Robust	0.645	0.734	0.813	0.888	0.934
	Eigenvector		0.045	0.042	0.051	0.042	0.058
$n^{-1/2}$	Degree	Ours	0.072	0.049	0.051	0.037	0.062
		Robust	0.659	0.801	0.949	0.993	0.999
	Diffusion	Ours	0.071	0.045	0.053	0.037	0.059
		Robust	0.881	0.948	0.993	1.000	1.000
	Eigenvector		0.077	0.045	0.050	0.050	0.047

Power of $H_0 : \beta = 0$ under the alternative $H_1 : \beta = 1$

p_n	Statistic	Sample Size				
		100	200	500	1000	2000
0.1	Degree - Robust	1.000	1.000	1.000	1.000	1.000
	Diffusion - Robust	1.000	1.000	1.000	1.000	1.000
	Eigenvector	0.845	0.995	1.000	1.000	1.000
$n^{-1/3}$	Degree - Robust	1.000	1.000	1.000	1.000	1.000
	Diffusion - Robust	1.000	1.000	1.000	1.000	1.000
	Eigenvector	0.998	1.000	1.000	1.000	1.000
$n^{-1/2}$	Degree - Robust	1.000	1.000	1.000	1.000	1.000
	Diffusion - Robust	1.000	1.000	1.000	1.000	1.000
	Eigenvector	0.832	0.947	0.994	1.000	1.000

Table 8: Power of 5% level two-sided tests of $H_0 : \beta = 0$ when $\beta = 1$. Under this H_0, the our test statistics is the usual t-statistic with robust (heteroskedasticity-consistent) standard errors.

Additional Example 1

- E.g. Production Networks (Carvalho, Nirei, Saito, and Tahbaz-Salehi, 2021)
  - Want to know how shocks propagate through production networks
  - Measurement error:

    "First, it only reports a binary measure of interfirm supplier- customer relations... we do not observe a yen measure associated with their transactions."
  - Sparsity: (out of 750,000 firms)

    "Second, the forms used by [the credit agency] limit the number of suppliers and customers that firms can report to 24 each."

Back

Additional Example 2

- E.g. Board-of-Director Network (Frydman and Hilt, 2017)
  - Clayton Antitrust Act of 1914 prohibits bankers from serving on boards of railroads
  - Want to know if policy reduced bank lending due to greater monitoring costs
  - A_{ij} records probability that bank i detects fraud in firm j
  - Proxied using board interlocks
  - Interlocks are few relative to firms (395 firms, 6 links between banks and utilities, 14 links between banks and railroads).

Back

Local Asymptotics

- Similar in spirit to modeling:
  - Correlation of weak instruments and endogenous variables decaying to 0 (e.g. Staiger and Stock 1997).
  - Power of tests using local alternatives (Pitman drift, see e.g. Rothenberg 1984).
  - Local to unity asymptotics for time series (e.g. Chan and Wei 1987)

Back

Weak Ties Theory

- Granovetter (1973): Weak ties which are more numerous are key drivers of outcomes
  - Weak ties: A_{ij} is small
  - Numerous: most A_{ij} non-zero ($O(n^2)$)

- Job referrals in Newton, MA:
  - Most recent job changers found jobs through friends "marginally included in the current network of contacts".
  - *"It is remarkable that people receive crucial information from individuals whose very existence they have forgotten."*
- Other examples: innovation (e.g. Reagans and Zuckerman 2001), economic development (e.g. Eagle et al. 2010), job referrals (e.g. Rajkumar et al. 2022).

▸ Back

Assumption 1 (Rank R Graphon)

Suppose f has rank $R < \infty$:

$$f(u,v) = \sum_{r=1}^{R} \tilde{\lambda}_r \phi_r(u)\phi_r(v) \quad , \tag{3}$$

where $\|\phi_r\| = 1$ for all $r \in [R]$ and if $r \neq s$,

$$\int_{[0,1]} \phi_r(u)\phi_s(u)du = 0 \ .$$

Furthermore, suppose that

$$\Delta_{\min} = \min_{1 \geq r \geq R-1} \left|\tilde{\lambda}_r - \tilde{\lambda}_{r+1}\right| > 0$$

Low Rank Assumption

- The rank assumption means the networks have “structure” (Chatterjee 2015).
- Many popular network models are low rank
    - Stochastic Block Model (Holland et al. 1983)
    - Random Dot Product Graphs (Young and Scheinerman 2007)
- Also common in the matrix completion literature (e.g. Candès and Tao 2010, Negahban and Wainwright 2012, Athey et al. 2021).

Back

Stochastic Block Model

- The stochastic block model Holland et al. (1983) is a popular model of networks.
- It assumes that agents fall into groups $g \in \{1, ..., B\}$.
- Link probability between agents depend only on the groups to which they belong.

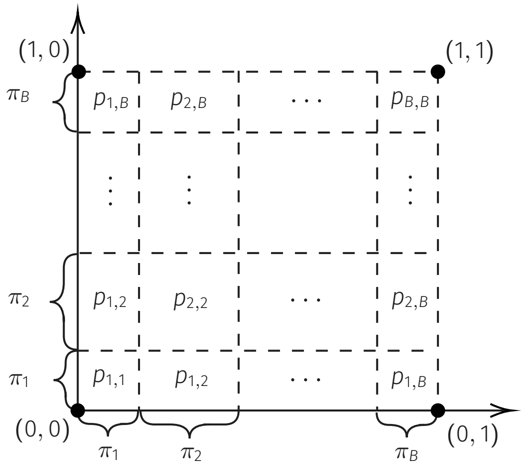


Figure 8: The graphon f of a stochastic block model with B blocks. f is a step-function with B^2 steps and is of rank B.

Back