Linear Regression with Centrality Measures

Yong Cai

Jan 13, 2023

Northwestern University

Linear Regression with Centrality Measures

- Suppose researchers observe agents in a network
- Want to know how network position of an agent affects economic outcome
- Network is high dimensional $\Rightarrow$ summarize using centrality measures
- Use OLS to study relationship between outcome and centrality

Linear Regression with Centrality Measures

- Suppose researchers observe agents in a network
- Want to know how network position of an agent affects economic outcome
- Network is high dimensional $\Rightarrow$ summarize using centrality measures
- Use OLS to study relationship between outcome and centrality
- When is linear regression with centrality measures statistically valid?

Examples

- Better-networked venture capital firms are more profitable (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 in the Philippines (Cruz, Labonne, and Querubin, 2017)

Linear Regression with Centrality Measures

- Researchers study:

$$Y_i = \underbrace{C_i}_{\text{centrality}} \beta + \underbrace{X_i'}_{\text{controls}} \gamma + \varepsilon_i \quad , \quad E[X_i \varepsilon_i] = 0, E[C_i \varepsilon_i] = 0$$

- β is parameter of interest
- Different centrality measures capture different ways of being important
- Estimate β by OLS; conduct inference using t-test

Statistical Challenges

- Networks may be sparse
  - Many more agents than links per agent
  - Not enough variation to identify β
  - Sparsity is "stylized fact"

Statistical Challenges

- Networks may be sparse
  - Many more agents than links per agent
  - Not enough variation to identify β
  - Sparsity is "stylized fact"
- Networks may be constructed using proxies
  - Interaction of interest may not be observed
  - Use a related network instead
  - E.g. want risk sharing network, have friendship network

Statistical Challenges

- Networks may be sparse
  - Many more agents than links per agent
  - Not enough variation to identify β
  - Sparsity is "stylized fact"
- Networks may be constructed using proxies
  - Interaction of interest may not be observed
  - Use a related network instead
  - E.g. want risk sharing network, have friendship network

$\Rightarrow$ Weaker signals harder to separate from noise

When is linear regression on centrality measures statistically valid?

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

Contributions

1. Show that OLS can become inconsistent with sparse, proxy networks
    - Characterize threshold at which inconsistency occurs
    - Show that eigenvector is less robust than degree and diffusion
    - Comparing significance involves both economic *and* statistical properties
    - Rule-of-Thumb for sparsity regime

Contributions

1. Show that OLS can become inconsistent with sparse, proxy networks
    - Characterize threshold at which inconsistency occurs
    - Show that eigenvector is less robust than degree and diffusion
    - Comparing significance involves both economic *and* statistical properties
    - Rule-of-Thumb for sparsity regime

2. Distributional theory with sparse, proxy networks
    - 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
    - Usual confidence intervals and tests may not be valid

Contributions

1. Show that OLS can become inconsistent with sparse, proxy networks
    - Characterize threshold at which inconsistency occurs
    - Show that eigenvector is less robust than degree and diffusion
    - Comparing significance involves both economic *and* statistical properties
    - Rule-of-Thumb for sparsity regime

2. Distributional theory with sparse, proxy networks
    - 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
    - Usual confidence intervals and tests may not be valid

3. Novel bias correction and inference methods

Related Literature

- Linear Regression with Centrality Measures
  - **Eigenvectors:** Le and Li (2020); Cai, Yang, Zhu, Shen, and Zhao (2021)

- Estimation of Centrality Statistics with Proxy Networks
  - **Simulations:** Costenbader and Valente (2003); Borgatti, Carley, and Krackhardt (2006)
  - **Theory:** Dasaratha (2020); Avella-Medina, Parise, Schaub, and Segarra (2020)

- Econometrics of Sparse Networks
  - **Network Formation:** Graham (2017); Jochmans (2018); Graham (2020b); De Paula et al. (2018); Menzel (2022)
  - **Network Recovery:** Manresa (2016); Rose (2016); Wang (2018); De Paula et al. (2020)
  - **Network Moments:** Bickel et al. (2011); Bhattacharyya and Bickel (2015); Matsushita and Otsu (2021); Green and Shalizi (2022); Leung and Moon (2019); Menzel (2021)

Table of Contents

Centrality Measures

Centrality Measures

- Network data is symmetric adjacency matrix A
- A_{ij} records intensity of relationship between i and j

Centrality Measures

- Network data is symmetric adjacency matrix A
- A_{ij} records intensity of relationship between i and j
- Centrality measures summarize network positions into "importance"
- When A is known, centrality measures exactly computable

Centrality Measures

- Network data is symmetric adjacency matrix A
- A_{ij} records intensity of relationship between i and j
- Centrality measures summarize network positions into "importance"
- When A is known, centrality measures exactly computable
- Many different ways to measure importance $\Rightarrow$ many centrality measures
- Focus on degree, diffusion and eigenvector

Degree and Diffusion Centralities

- Degree is the total intensity of direct connections:

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

Degree and Diffusion Centralities

- Degree is the total intensity of direct connections:

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

- Diffusion reflects ability to broadcast messages in a network:

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

- T, δ are chosen by researchers

Eigenvector Centrality

- Eigenvector is the leading eigenvector of A, scaled

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

- Friends of important agents are themselves more important

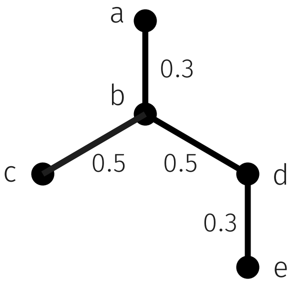
Eigenvector Centrality

- Various choices of a_n in literature
- a_n turns out to matter for statistical properties
- Today: $a_n = \sqrt{\lambda_1(A)}$

	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 1: Examples of a_n in econometric theory and empirical work

Centrality Statistics

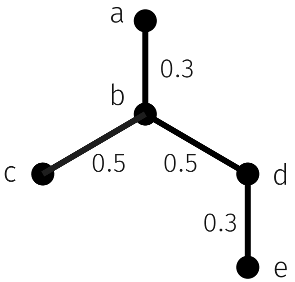


$$\begin{pmatrix} 0 & 0.3 & 0 & 0 & 0 \\ 0.3 & 0 & 0.5 & 0.5 & 0 \\ 0 & 0.5 & 0 & 0 & 0 \\ 0 & 0.5 & 0 & 0 & 0.3 \\ 0 & 0 & 0 & 0.3 & 0 \end{pmatrix}$$

Centrality Statistics

$$\begin{pmatrix} 0.3 \\ 1.3 \\ 0.5 \\ 0.8 \\ 0.3 \end{pmatrix}$$

Degree
$C^{(1)}$



$$\begin{pmatrix} 0 & 0.3 & 0 & 0 & 0 \\ 0.3 & 0 & 0.5 & 0.5 & 0 \\ 0 & 0.5 & 0 & 0 & 0 \\ 0 & 0.5 & 0 & 0 & 0.3 \\ 0 & 0 & 0 & 0.3 & 0 \end{pmatrix}$$

Centrality Statistics

$$\begin{pmatrix} 0.3 \\ 1.3 \\ 0.5 \\ 0.8 \\ 0.3 \end{pmatrix} \qquad \begin{pmatrix} 0.28 \\ 0.94 \\ 0.46 \\ 0.64 \\ 0.24 \end{pmatrix}$$

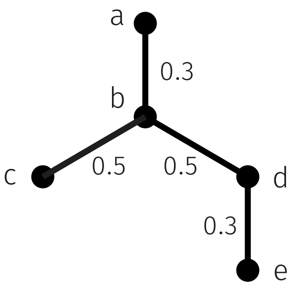
Degree

$C^{(1)}$

Diffusion

$C^{(T)}$

$(T = 3, \delta = 0.5)$



$$\begin{pmatrix} 0 & 0.3 & 0 & 0 & 0 \\ 0.3 & 0 & 0.5 & 0.5 & 0 \\ 0 & 0.5 & 0 & 0 & 0 \\ 0 & 0.5 & 0 & 0 & 0.3 \\ 0 & 0 & 0 & 0.3 & 0 \end{pmatrix}$$

Centrality Statistics

$$\begin{pmatrix} 0.3 \\ 1.3 \\ 0.5 \\ 0.8 \\ 0.3 \end{pmatrix} \qquad \begin{pmatrix} 0.28 \\ 0.94 \\ 0.46 \\ 0.64 \\ 0.24 \end{pmatrix} \qquad 0.89 \cdot \begin{pmatrix} 0.25 \\ 0.68 \\ 0.42 \\ 0.50 \\ 0.18 \end{pmatrix}$$

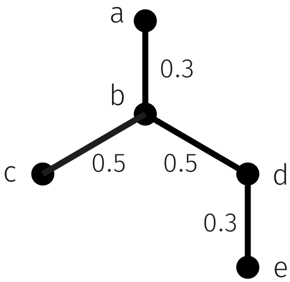
Degree $C^{(1)}$

Diffusion $C^{(T)}$

$(T = 3, \delta = 0.5)$

Eigenvector $C^{(\infty)}$

$\left(a_n = \sqrt{\lambda_1(A)}\right)$



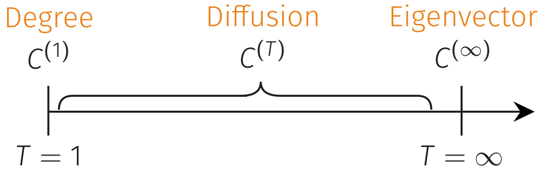
$$\begin{pmatrix} 0 & 0.3 & 0 & 0 & 0 \\ 0.3 & 0 & 0.5 & 0.5 & 0 \\ 0 & 0.5 & 0 & 0 & 0 \\ 0 & 0.5 & 0 & 0 & 0.3 \\ 0 & 0 & 0 & 0.3 & 0 \end{pmatrix}$$

Centrality Statistics

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



Model and Assumptions

Model and Assumptions

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

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

- Will make enough assumptions so $\beta^{(d)}$ is slope of CEF
- DGP yields i.i.d. draws of $\{(\varepsilon_i, U_i)\}$
- $U_i \sim U[0, 1]$ is unobserved latent type used to construct network

True Network

- 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

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

  Symmetry: $A_{ij} = A_{ji}$; normalisation: $A_{ii} = 0$

- When A_{ij} is large, agents i and j have a strong relationship

- $p_n \to 0$ reflects sparsity

Examples

Informal Insurance and Consumption Smoothing

- Y_i: variance in consumption expenditure
- A_{ij}: probability that i lends j money or vice versa
- U_i: social class

Examples

Informal Insurance and Consumption Smoothing

- Y_i: variance in consumption expenditure
- A_{ij}: probability that i lends j money or vice versa
- U_i: social class

Graphons (f)

- "Full" Insurance: $f\left(U_i, U_j\right) = 1$
- Assortative: $f\left(U_i, U_j\right) = 1 - \left(U_i - U_j\right)^2$

True Network

- When A is observed, can exactly compute centrality measures
- Form the estimator

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

Proxy Network

- When A is not observed, use $\hat{A}$: for $j > i$,

$$\hat{A}_{ij} \,\big|\, \mathbf{U} \overset{\text{iid}}{\sim} \text{Bernoulli}\left(A_{ij}\right)$$

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

- Proxy error is “white noise”

Proxy Network

- When A is not observed, use $\hat{A}$: for $j > i$,

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

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

- Proxy error is "white noise"
- Use $\hat{A}$ as plug-in for A to compute $\hat{C}^{(d)}$
- Estimate

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

Examples

Informal Insurance and Consumption Smoothing

- Y_i: variance in consumption expenditure
- A_{ij}: probability that i lends j money or vice versa
- U_i: social class

Proxy Networks

- Social: $\hat{A}_{ij} = 1$ if either i or j reports the other as a friend
- Financial: $\hat{A}_{ij} = 1$ if i borrows from or lends to j

Sparsity

- Want to understand the properties of $\tilde{\beta}^{(d)}$ and $\hat{\beta}^{(d)}$ when networks are sparse
- Theoretical device: $p_n \to 0$ as $n \to \infty$:

$$A_{ij} := p_n f(U_i, U_j) \quad , \quad \hat{A}_{ij} = \text{Bernoulli}\left(p_n f(U_i, U_j)\right)$$

- Weak interaction in true network: $A_{ij} \to 0$
- Sparse proxy networks: many $\hat{A}_{ij}$'s are 0
- Rate at which $p_n \to 0$ reflects different amounts of sparsity

▸ Weak Ties ▸ Local Asymp.

Model and Assumptions

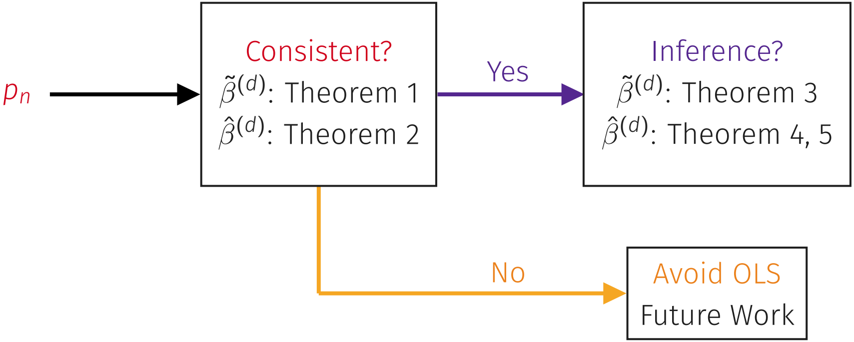
- Proxy Networks: $\hat{A}$ is observed but A economically meaningful
  - **Node-Level Regressions**: Auerbach (2022), Le and Li (2020), Cai et al. (2021)
  - **Dyadic Models**: see e.g. De Paula (2017) Section 3, Graham (2020a) Section 6
- Bickel-Chen Model of Sparsity (Bickel and Chen 2009)
  - Graham (2020a): "The Bickel-Chen model is the default one in the nonparametric statistics and machine learning literatures on random graphs."
  - **Dyadic Models**: Jochmans (2018), Graham (2020b)
- Standard models of proxy networks and sparsity to study linear regression

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

Consistency with True Network A

Theorem 1

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

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

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

- $n \cdot n^{-3/2} \to 0$
- Inconsistent only under extreme sparsity
- Consistency requires $\|C^{(d)}\|_2 \to \infty$ w.p.a. 1
- a_n inflates eigenvector, counters sparsity

Consistency with Proxy Network $\hat{A}$

Theorem 2

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

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

- $\hat{\beta}^{(1)}$, $\hat{\beta}^{(T)}$ less robust to sparsity than $\tilde{\beta}^{(1)}$, $\tilde{\beta}^{(T)}$
- Consistency with proxy networks in dense regime

▸ More Phases

Consistency with Proxy Network $\hat{A}$

Theorem 2

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

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

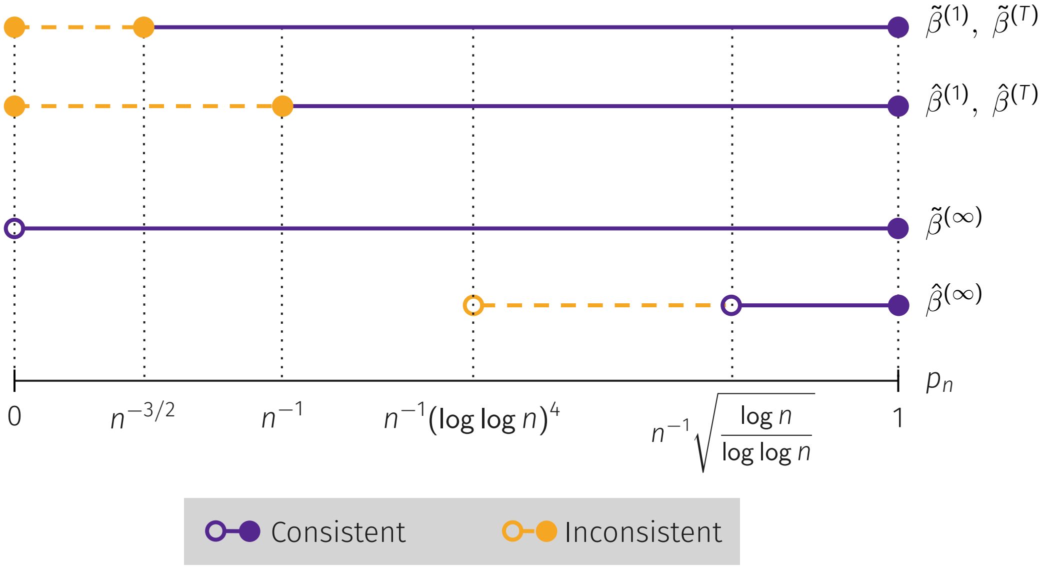
It is inconsistent if

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

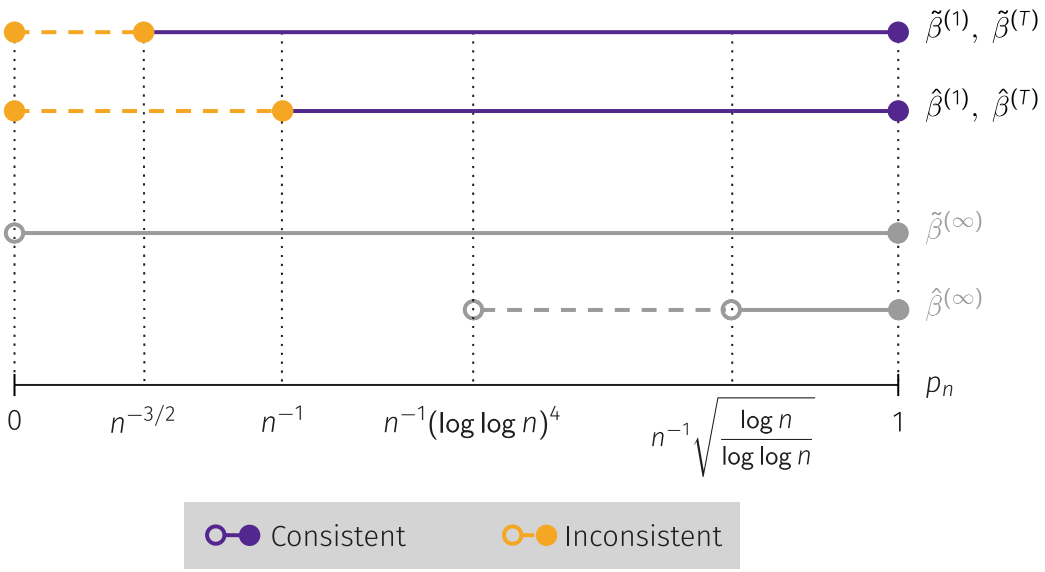
Consistency with Proxy Networks

- With proxy networks, p_n now matters
- Within the inconsistency thresholds, eigenvalues of $\hat{A}$ corresponding to informative eigenvectors are small (Alt et al. 2021a,b)
- Leading eigenvector not informative about eigenvectors of A
- Not clear if eigenvectors of $\hat{A}$ have structure below lower threshold

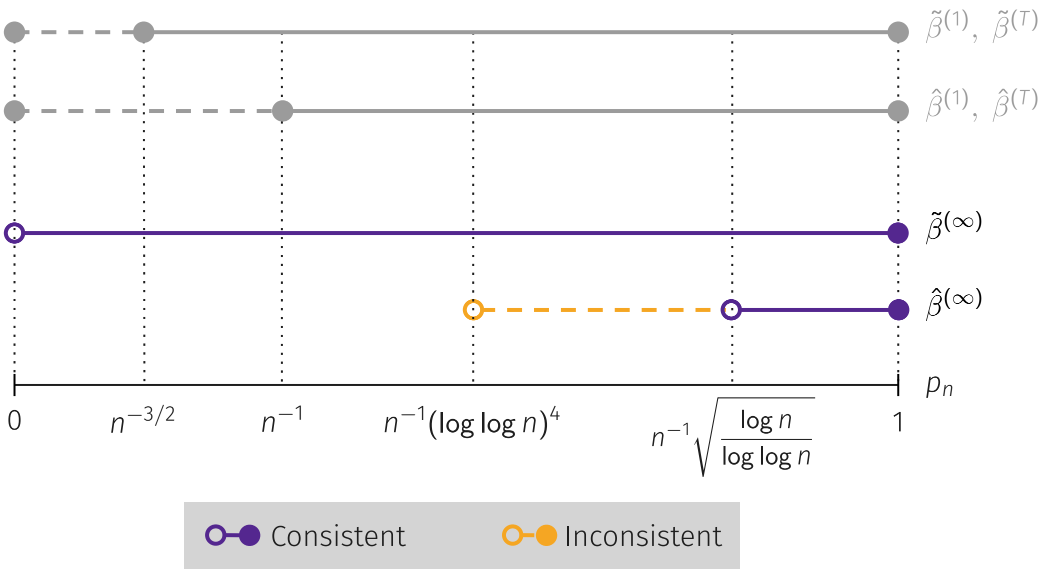
Consistency Thresholds



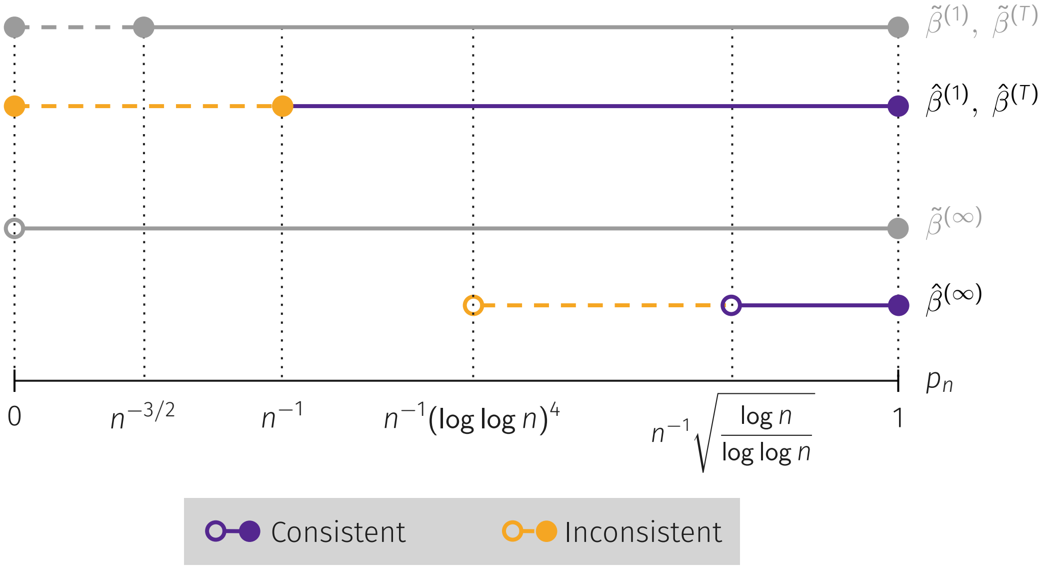
Consistency Thresholds - Degree and Diffusion



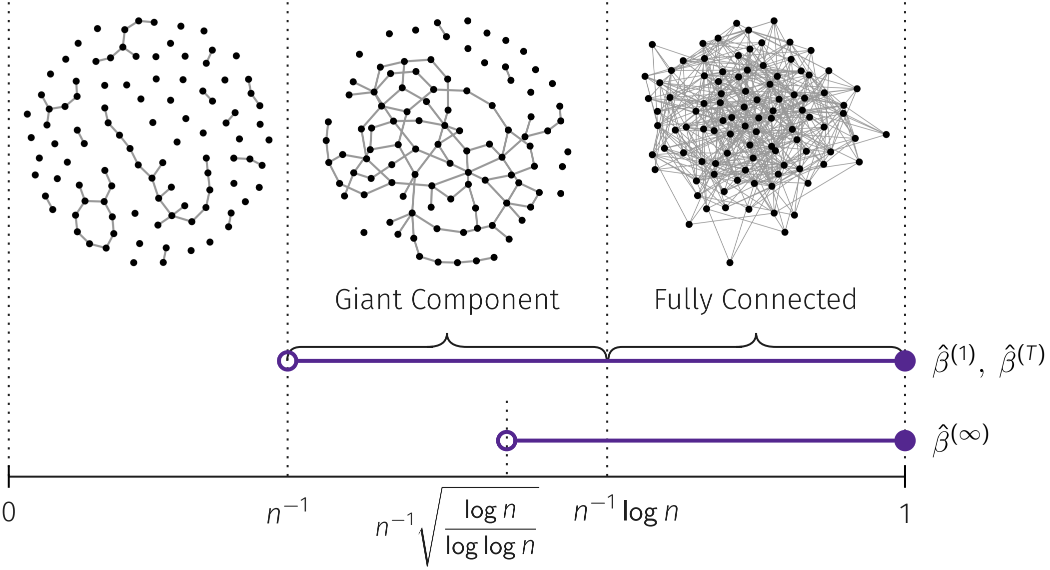
Consistency Thresholds - Eigenvector



Consistency Thresholds - Proxy Networks



Rule of Thumb for Consistency



▸ Formal Statement

Theoretical Results

Distributional Theory

Distributional Theory with True Network A

Theorem 3

For $d \in \{1, T, \infty\}$, suppose $\tilde{\beta}^{(d)}$ is consistent. Then,

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

where $V_0^{(d)} = E\left[\left(C_i^{(d)}\right)^2\right]^{-2} E\left[\left(C_i^{(d)}\varepsilon_i\right)^2\right].$

Theorem 4

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 $V_0^{(d)} = E\left[\left(C_i^{(d)}\right)^2\right]^{-2} E\left[\left(C_i^{(d)}\varepsilon_i\right)^2\right]$.

- 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 $\beta^{(d)} = 0$.

Distributional Theory with Proxy Network $\hat{A}$

Theorem 4

For $d \in \{1, T\}$, suppose $\hat{\beta}^{(d)}$ is consistent. If $\beta^{(d)} \neq 0$,

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

Distributional Theory with Proxy Network $\hat{A}$

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

- Using proxy network 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 with Proxy Network $\hat{A}$

- $\hat{B}^{(d)}$ and $\hat{V}^{(d)}$ in paper
- Bias-corrected estimator:

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

- Adjusted tests and confidence intervals in paper
- No additional data requirement; p_n need not be specified
- Bias estimation is challenging since $B^{(d)}$ much larger than $\sqrt{V^{(d)}}$

▸ Bias Corr. Details

Distributional Theory with Proxy Network $\hat{A}$

Theorem 5

Suppose f has rank $R < \infty$. Suppose also for $\eta > 0$ that

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

and that $a_n = \sqrt{\lambda_1(A)}$. Then,

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

where

$$V_0^{(\infty)} = E\left[(C_i^{(\infty)})^2 \right]^{-2} E\left[(C_i^{(\infty)} \varepsilon_i)^2 \right]$$

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
- Using $\sqrt{\lambda_1(\hat{A})}$ does not change result

Inference under Sparsity

	Proxy Network		True Network
	$\beta^{(1)} / \beta^{(T)}$	$\beta^{(\infty)}$	$\beta^{(1)} / \beta^{(T)} / \beta^{(\infty)}$
$H_0 : \beta^{(d)} = 0$		Dense: t-test*	
$H_0 : \beta^{(d)} = b$			
Confidence Intervals			

Table 2: *: $p_n \gg n^{-1/2}$ (Le and Li, 2020). For $\beta^{(\infty)}$, $a_n = \sqrt{\lambda_1(A)}$.

Inference under Sparsity

	Proxy Network		True Network
	$\beta^{(1)}$ / $\beta^{(T)}$	$\beta^{(\infty)}$	$\beta^{(1)}$ / $\beta^{(T)}$ / $\beta^{(\infty)}$
$H_0 : \beta^{(d)} = 0$		Dense: t-test*	t-test
$H_0 : \beta^{(d)} = b$			
Confidence Intervals			t-stat based

Table 2: *: $p_n \gg n^{-1/2}$ (Le and Li, 2020). For $\beta^{(\infty)}$, $a_n = \sqrt{\lambda_1(A)}$.

Key: Theorem 3,

	Proxy Network		True Network
	$\beta^{(1)}$ / $\beta^{(T)}$	$\beta^{(\infty)}$	$\beta^{(1)}$ / $\beta^{(T)}$ / $\beta^{(\infty)}$
$H_0 : \beta^{(d)} = 0$	t-test	Dense: t-test*	t-test
$H_0 : \beta^{(d)} = b$	New Method		
Confidence Intervals	New Method		t-stat based

Table 2: *: $p_n \gg n^{-1/2}$ (Le and Li, 2020). For $\beta^{(\infty)}$, $a_n = \sqrt{\lambda_1(A)}$.

Key: Theorem 3, Theorem 4,

	Proxy Network		True Network
	$\beta^{(1)}$ / $\beta^{(T)}$	$\beta^{(\infty)}$	$\beta^{(1)}$ / $\beta^{(T)}$ / $\beta^{(\infty)}$
$H_0 : \beta^{(d)} = 0$	t-test	Dense: t-test* Sparse: t-test	t-test
$H_0 : \beta^{(d)} = b$	New Method	t-test	
Confidence Intervals	New Method	t-stat based	t-stat based

Table 2: *: $p_n \gg n^{-1/2}$ (Le and Li, 2020). For $\beta^{(\infty)}$, $a_n = \sqrt{\lambda_1(A)}$.

Key: Theorem 3, Theorem 4, Theorem 5

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:

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

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

Inconsistency with Proxy Networks

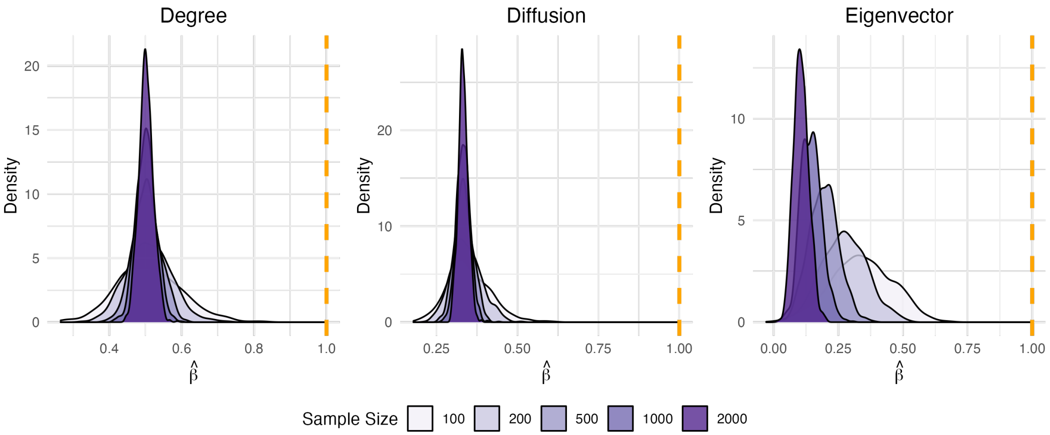


Figure 1: Distribution of $\hat{\beta}^{(d)}$ for $p_n = 1/n$. For $\hat{\beta}^{(\infty)}$, $a_n = \sqrt{n}$. $\beta = 1$ (orange dashed line).

Consistency with True Networks

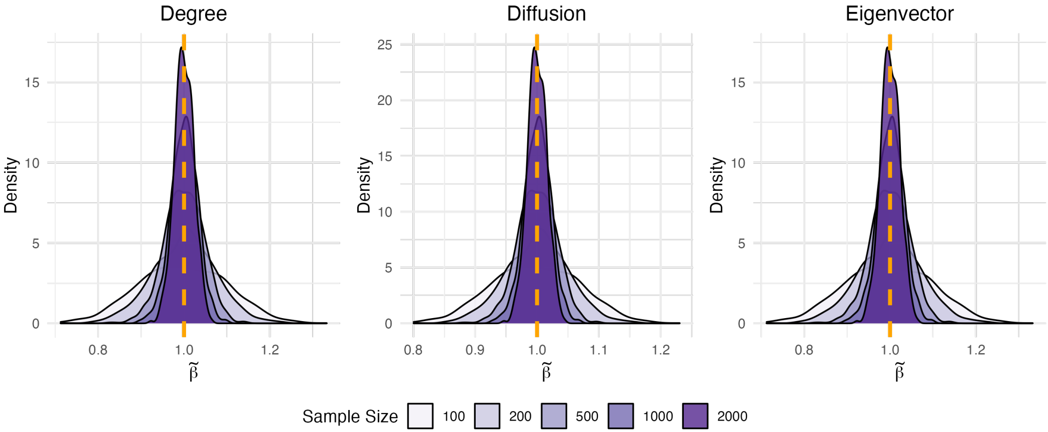


Figure 2: Distribution of $\tilde{\beta}^{(d)}$ for $p_n = 1/n$. For $\tilde{\beta}^{(\infty)}$, $a_n = \sqrt{n}$. $\beta = 1$ (orange dashed line).

Bias correction is effective

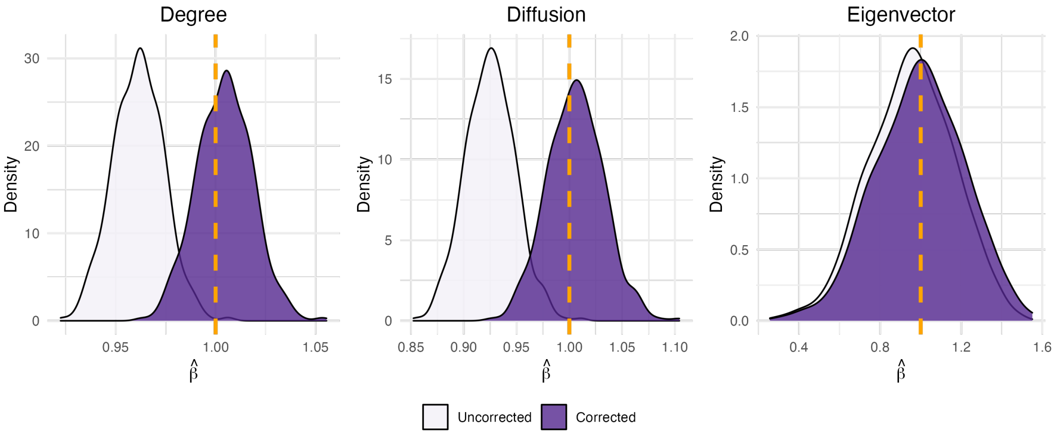


Figure 3: Distributions of $\hat{\beta}^{(d)}$ and their bias corrected versions $\check{\beta}^{(d)}$ for $p_n = 1/\sqrt{n}$, $n = 500$, $a_n = \sqrt{\lambda_1(\hat{A})}$. $\beta = 1$ (orange dashed line).

Distributional theory is accurate

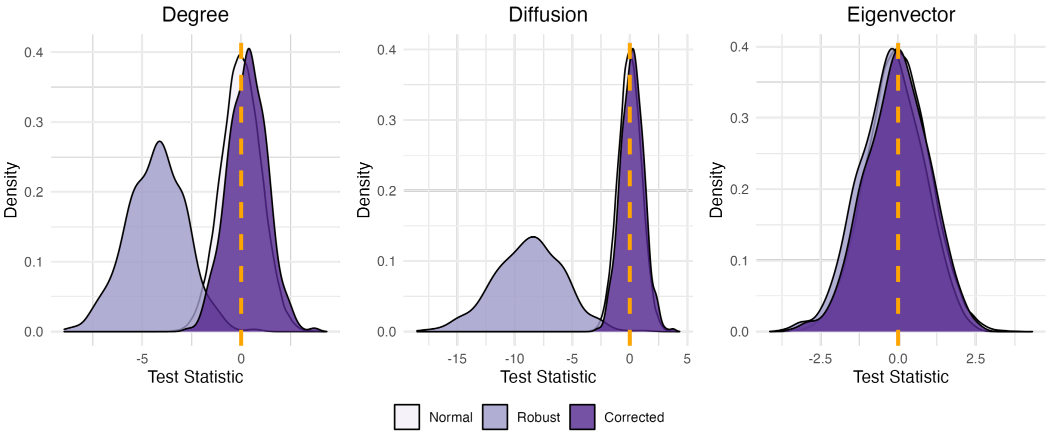


Figure 4: Distribution of the centered and scaled test statistics. Robust refers to tests based on t-statistic with robust (hc) standard errors. $p_n = 1/\sqrt{n}$, $n = 500$, $a_n = \sqrt{\lambda_1(\hat{A})}$.

Adjusted tests are better

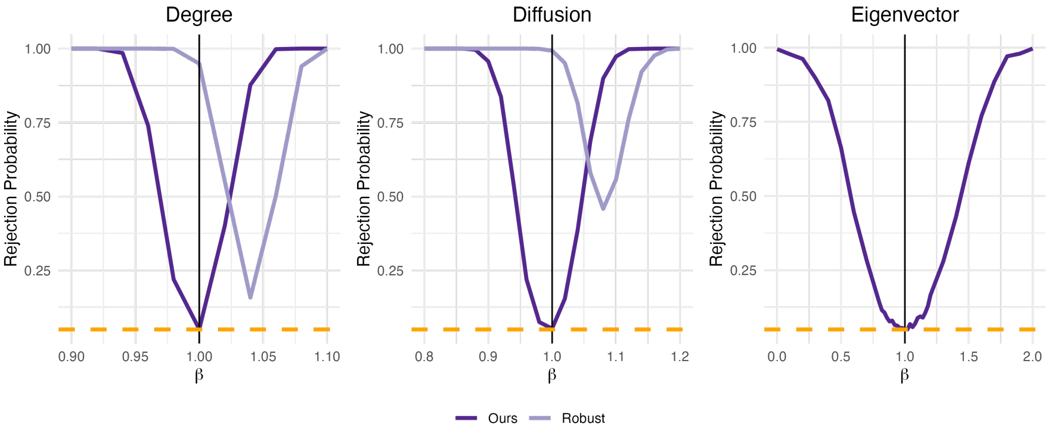


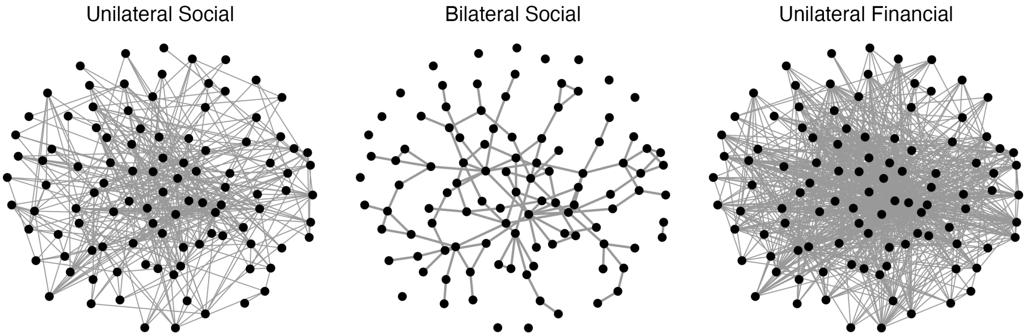
Figure 5: Power of the two-sided test of $H_0 : \beta = 1$ under various alternatives. Test at 5% level of significance (orange dashed line). $p_n = 1/\sqrt{n}$, $n = 500$, $a_n = \sqrt{\lambda_1(\hat{A})}$.

Empirical Demonstration

Empirical Demonstration

- De Weerdt and Dercon (2006): want to know if informal insurance can help consumption smoothing
- Village of 119 households in Nyakatoke, Tanzania
- Regress variance in food expenditure on centrality in network

Empirical Demonstration



$(n = 119)$	Mean	Median	Min	Max
Unilateral Social	8.02	7	1	31
Bilateral Social	2.30	2	0	10
Unilateral Financial	16.53	14	3	79

Table 3: Degree distributions of various networks in Nyakatoke

Empirical Demonstration

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

▸ Conf. Int

Conclusion

Conclusion

1. Show that OLS can become inconsistent with sparse, proxy networks
    - Characterize threshold at which inconsistency occurs
    - Show that eigenvector is less robust than degree and diffusion
    - Comparing significance involves both economic *and* statistical properties
    - Rule-of-Thumb for sparsity regime

2. Distributional theory with sparse, proxy networks
    - 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
    - Usual confidence intervals and tests may not be valid

3. Novel bias correction and 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.

Bhattacharyya, S. and P. J. Bickel (2015). Subsampling bootstrap of count features of networks. *The Annals of Statistics 43*(6), 2384–2411.

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.

Bickel, P. J., A. Chen, and E. Levina (2011). The method of moments and degree distributions for network models. *The Annals of Statistics 39*(5), 2280–2301.

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.

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

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 Paula, Á., S. Richards-Shubik, and E. Tamer (2018). Identifying preferences in networks with bounded degree. *Econometrica 86*(1), 263–288.

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.

Graham, B. S. (2017). An econometric model of network formation with degree heterogeneity. *Econometrica 85*(4), 1033–1063.

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.

Green, A. and C. R. Shalizi (2022). Bootstrapping exchangeable random graphs. *Electronic Journal of Statistics 16*(1), 1058–1095.

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*.

Leung, M. P. and H. R. Moon (2019). Normal approximation in large network models. *arXiv preprint arXiv:1904.11060*.

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

Matsushita, Y. and T. Otsu (2021). Jackknife empirical likelihood: small bandwidth, sparse network and high-dimensional asymptotics. *Biometrika 108*(3), 661–674.

Menzel, K. (2021). Central limit theory for models of strategic network formation. *arXiv preprint arXiv:2111.01678*.

Menzel, K. (2022). Strategic network formation with many agents.

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.

Wang, Y. (2018). Panel data with high-dimensional factors: inference on treatment effects with an application to sampled networks. Technical report, 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 Proxy Networks

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

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

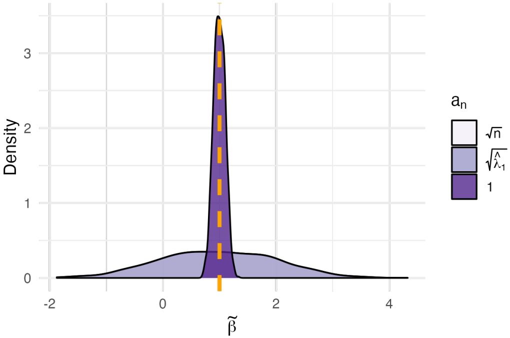


Figure 6: Distribution of $\tilde{\beta}^{(\infty)}$ for $n = 100$, $p_n = 1/n$ under various a_n. $\beta = 1$ (orange dashed line).

Eigenvector is more sensitive to sparsity than Degree and Diffusion

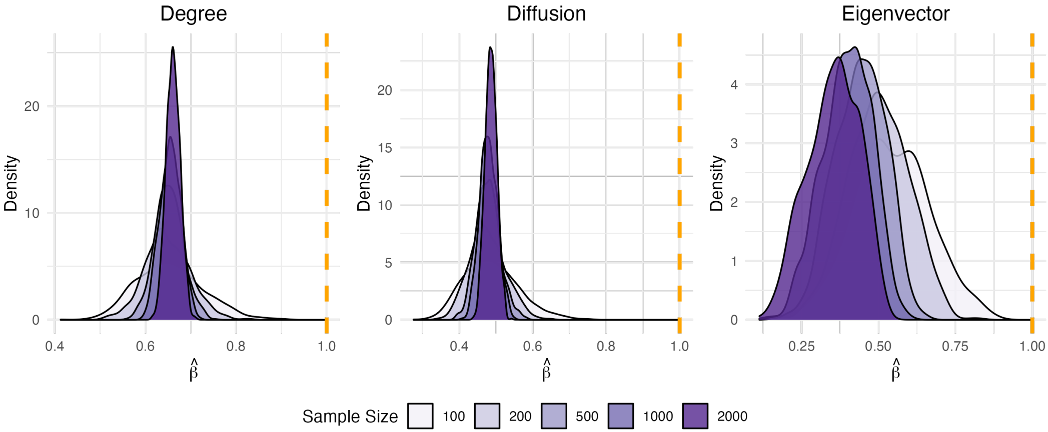


Figure 7: Distribution of $\hat{\beta}^{(d)}$ for $p_n = n^{-1}\sqrt{\log n / \log\log n}$. For $\tilde{\beta}^{(\infty)}$, $a_n = \sqrt{n}$. $\beta = 1$ (orange dashed line).

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 6: 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.

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

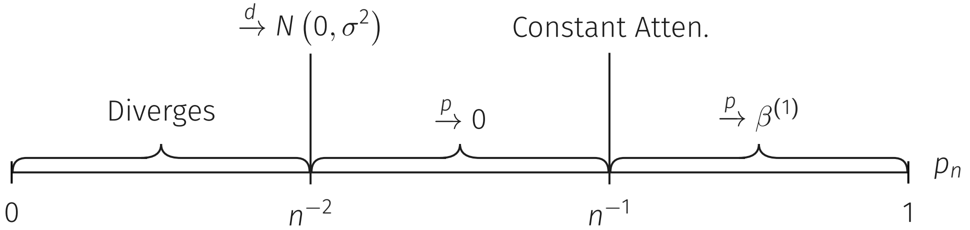
Weak Ties Theory

- 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

All Phase Transitions in $\hat{\beta}^{(1)}$



▸ Back

Rule of Thumb for Consistency

Rule of Thumb

(a) Treat $\hat{\beta}^{(1)}$ and $\hat{\beta}^{(T)}$ as consistent if there exists a giant component with at least $N/2$ nodes.

(b) Treat $\hat{\beta}^{(\infty)}$ as consistent if the network is fully connected.

Back

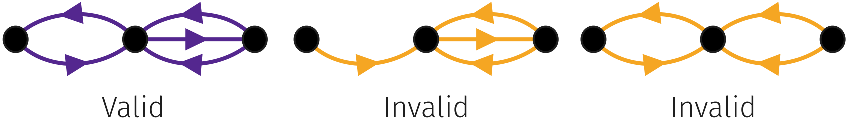
Bias Correction

- Bias correction reduces to the problem of
$$E\left[\iota'\left(\hat{A}-A\right)^{2}t\iota \mid U\right]$$
at a sufficiently fast rate.

- Given that $E[(\hat{A}_{ij}-A_{ij})^2|U] \approx A_{ij}$, an intuitive estimator is $\iota' A^t \iota$

- Consistent, but does not converge fast enough

- Difference between the two relates to the number of paths of a given length on a line in which each edge is traversed at least twice.



Back

Low Rank Assumption

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

Empirical Demonstration

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

Table 7: One-sided confidence intervals for degree and diffusion in Bilateral Social network.

Back