Centrality Estimation from Sampled Networks*

Yang Xu†
University of Oxford

November 5, 2026
Click here for the latest version.

Abstract

Can a partially observed network reveal how central everyone is? Centrality shapes economic outcomes and guides policy. However, computing it requires full network data, which is costly and slow to collect. Simple random sampling biases centrality estimates. I propose a sequential sampling design: start with a random household, record its contacts, draw one of them at random, and repeat. The estimated eigenvector centrality is consistent at the rate of the average degree over the largest eigenvalue, for a fixed sampling fraction and nonvanishing spectral gaps. In least-squares regressions of an outcome on the estimated centrality, the robust t-statistics from the sampled and the full network are asymptotically equivalent under the null of a zero slope. In an application to the village networks of Banerjee et al. (2019), sequentially surveying half the households reproduces their full-network regression slopes within about one percent, whereas a simple random sample understates them by eleven to twelve percent.

Keywords: Partially Observed Networks, Sampling Design, Eigenvector Centrality

1 Introduction

An agent's position in a network, as measured by centrality, shapes economic outcomes. Politicians from central families win more votes (Cruz et al., 2017), shocks to central sectors drive aggregate fluctuations (Acemoglu et al., 2012), and microfinance participation diffuses further when the informed households are central (Banerjee et al., 2013). Centrality also guides policy. When a network mediates strategic spillovers, an optimal intervention allocates its budget in proportion to centrality (Galeotti et al., 2020). Computing centrality requires the *full* network. However, budget and time constraints limit network data collection in both empirical research (Breza et al., 2020) and interventions (Banerjee et al., 2019). I ask whether a *sampled network* suffices to estimate centrality.

*I am extremely grateful for the generous support and encouragement of Martin Weidner. I also thank Áureo de Paula, Banu Demir, Cristina Gualdani, Anders Bredahl Kock, and Frank Windmeijer for helpful feedback. I also thank participants at Nuffield Postdoc Seminar Series for useful discussion. All errors are mine.

†Department of Economics and Nuffield College, University of Oxford. Email: yang.xu@economics.ox.ac.uk

This paper makes three contributions. First, I propose a sequential sampling design and show that it consistently estimates the eigenvector centrality of every agent from the partially observed network. Eigenvector centrality measures how central an agent is by how central its contacts are (Section 2). To the best of my knowledge, no existing work provides a consistent estimator of this or any other commonly used centrality measure from a sampled network without modeling how agents form links. Second, for least-squares regressions of an outcome on the estimated centrality, I show that the plug-in slope is consistent and that, under the null of a zero slope, the robust t-statistic from the sampled network is asymptotically equivalent to the statistic from the full network.[1] Third, I quantify the cost saving in network data collection.[2] In the villages of Banerjee et al. (2019), sequentially surveying half the households reproduces the full-network slopes to within about one percent, whereas simple random sampling at the same budget understates them by eleven to twelve percent.

Sampling a network poses two challenges. First, the sampled network generally does *not* inherit the structure of the full network. Under simple random sampling, a connected network can split into several components. The degree distribution of the sampled network differs from that of the full network (Stumpf et al., 2005). Statistics computed from the sampled network, such as betweenness centrality and clustering, are biased as well (Lee et al., 2006). These errors carry over to regression analysis. Chandrasekhar and Lewis (2016) show that the measurement error in network statistics is non-classical and biases regression slopes in either direction. Second, the pattern of missing links is structured. A link between two unsampled agents is never observed. Hence, the missing links form a single block, the links among the unsampled agents, in contrast with a random sample of links, in which each link is missing independently of the others.

When missing data bias an estimator, a prominent approach is to model them. For a partially observed network, a model of the missing links is a model of the complex process in which agents form links. This paper takes a different approach. The sampling design, which the researcher controls, structures what is observed and what is missing. I treat the missing links as a *perturbation* of the full network and bound their effect on the estimated centrality. The size of the perturbation depends on the sample. I therefore seek a sampling design under which the centrality of every agent in the full network can be accurately recovered from the partially observed one. Two regularities of social and economic networks make this possible. First, degrees are heterogeneous, with some agents holding many more links than others (Graham, 2017; Jackson et al., 2017). Second, networks are sparse, with realized links making up a small share of potential links. I propose a sequential sampling design that exploits both. The surveyor draws the first household at random from the roster, records its contacts, draws the next household at random from those contacts, and interviews a household the first time it is drawn (Section 3.1).

Administrative data offer an alternative to the network survey where links leave records. For example, tax records on firm-to-firm transactions yield buyer–seller networks (Bernard et al., 2022; Demir et al., 2024). Nevertheless, such records omit transactions below the reporting threshold. They also exclude the informal sector, which accounts for more than

[1] For the null of nonzero slopes, I propose a bootstrap bias correction.

[2] J-PAL South Asia estimates that a full network survey of 120 villages over 8.2 months costs 189,164 US dollars (Breza et al., 2020, Table 4).

a third of output and employment in developing countries (La Porta and Shleifer, 2014). Importantly, social ties such as friendship seldom appear in any registry. When no record exists, the leading low-cost alternative is aggregated relational data, in which each respondent reports the number of contacts with a given trait, combined with a parametric model of how agents form links (Breza et al., 2020, 2023). This approach cuts the data collection cost substantially. However, it recovers a distribution of networks rather than the realized one, so that no link is ever observed, and it relies on a parametric model of link formation, which I do without.

This paper contributes to the literature on estimating eigenvector centrality from sampled networks. The closest work is Ruggeri and De Bacco (2020), who select nodes by a greedy heuristic initialized with a random walk and compute eigenvector centrality from the *principal submatrix* of the adjacency matrix on the sampled nodes. They establish no consistency result. Breza et al. (2023) establish consistency of node-level centrality under a latent space model using aggregated relational data. Closest to the regression analysis of this paper, Le and Li (2022) and Cai (2026) study regressions on eigenvectors or centralities of a network observed with noise, under a model of network formation. Instead of modeling network formation, I condition on the network.

The model-based literature treats G as a draw from a model of network formation. Graphon models rest on the exchangeability of nodes (Aldous, 1981; Hoover, 1979; Lovász and Szegedy, 2006). Under exchangeability the network is dense in the limit (Orbanz and Roy, 2015), and sparse variants require a modified asymptotic framework (Bickel et al., 2011; Crane and Dempsey, 2018; Graham, 2024). The β-model treats the heterogeneity as fixed effects and conditions on the degree sequence as the sufficient statistic (Chatterjee et al., 2011; Graham, 2017). Strategic models either impose an equilibrium selection rule (Mele, 2017) or partially identify preferences without one (Sheng, 2020; de Paula et al., 2018). This paper imposes no model of link formation. It conditions on the network, treats every link as given, requires no exchangeability assumption, and allows the average degree to stay bounded or grow with n (Section 3.5).

The sequential sampling design has two lineages. In sociology, link-tracing designs run from snowball sampling (Goodman, 1961) to respondent-driven sampling (Heckathorn, 1997). In computer science, random walks sample the web graph (Henzinger et al., 2000; Bar-Yossef et al., 2000). Section 3.1.1 compares the design with the link-tracing designs.

The paper proceeds as follows. Section 2 defines the data structure, the full-network estimand, and the sampled-network estimator, and states the challenges posed by partial network observation. Section 3 states the sampling algorithm, formalizes it as a Markov chain, compares it with alternative sampling designs, and establishes the consistency of the estimated centrality with an explicit rate. Section 4 introduces the regression of an outcome on the estimated centrality and shows that the plug-in slope is consistent. Section 5 shows that the first-stage bias is multiplicative, derives the distribution of the robust t-statistic under the null of a zero slope, and estimates the bias factor by a bootstrap. Section 6 reports simulations on four families of random graphs calibrated to the Karnataka village networks of Banerjee et al. (2019), and Section 7 evaluates the sampling design on those networks. Section 8 concludes. The appendix contains the proofs.

2 Set-up

2.1 Data structure

Consider a population of n households, indexed by $i = 1, \ldots, n$. The households are the nodes of an undirected, unweighted network.[3] The network $G = (V, \mathcal{E})$ consists of the node set V and the link set $\mathcal{E}$. A link is an unordered pair ij of distinct households. Households i and j are linked when $ij \in \mathcal{E}$.

The network is represented by its adjacency matrix $A \in \{0,1\}^{n \times n}$, with $A_{ij} := \mathbf{1}\{ij \in \mathcal{E}\}$. Links are undirected, so A is symmetric. By convention, self-loops are excluded, hence $A_{ii} = 0$. The neighborhood of household i is $N(i) := \{j \in V : A_{ij} = 1\}$, the set of households linked to i. For a set $S \subseteq V$, its neighborhood is $N(S) := \bigcup_{i \in S} N(i)$. The degree of i is the cardinality $d_i := |N(i)| = \sum_{j \in V} A_{ij}$, the ith row sum of A. The degree matrix is $D := \mathrm{diag}(d_1, \ldots, d_n)$. The average degree is $\bar{d} := \frac{1}{n} \sum_{i \in V} d_i$. For a symmetric matrix M, write $\lambda_1(M) \geq \lambda_2(M) \geq \cdots \geq \lambda_n(M)$ for its eigenvalues in decreasing order. Reserve $\lambda := \lambda_1(A)$, the largest eigenvalue of A.

I further assume that G is connected. A path between two households is a sequence of distinct households in which each consecutive pair is linked. The network G is connected when every pair of its households is joined by a path. Connectedness of G implies that A is irreducible.

The network G is fixed and deterministic. Randomness in the partially observed network arises from the sampling design alone. I impose no model of link formation on G.

2.2 Estimand from the full network

The estimand of this paper is the eigenvector centrality of the full network G, a node-level network statistic, or micro statistic in the taxonomy of Jackson et al. (2017). Like any network statistic, it compresses the $\binom{n}{2}$ link indicators of A into a summary of far lower dimension that can serve as a regressor, an outcome, or a targeting criterion. It is defined without reference to the sampling design. Partial network observability enters only through the sample introduced in Section 2.3.

Definition 1 (Eigenvector centrality)**.** *Let v be the entrywise nonnegative vector with $Av = \lambda v$ and $\|v\|_2 = 1$. The eigenvector centrality of household i is the ith coordinate v_i of v.*

Existence and uniqueness follow from the Perron–Frobenius theorem. Because the nonnegative matrix A is irreducible, λ is simple and its eigenvector is strictly positive. The definition is recursive. A household is central in proportion to the centrality of its friends, and the centrality of each household depends on the entire network. An isolated household, one with no links, is assigned centrality zero by convention. Definition 1 then extends to any network that is connected once its isolated households are removed, with v strictly positive at the remaining households and zero at every isolated one.

An eigenvector has no inherent unit. The centrality is therefore defined only up to scale, and a normalization must be fixed. Definition 1 fixes the unit norm $\|v\|_2 = 1$. This convention

[3] Theorem 1 extends to weighted networks (Remark 3).

defines the estimand and carries the consistency analysis. The regression analysis in Section 4 rescales to $\sqrt{n}\,v$, which makes the slope invariant to the network size. The rescaling is a change of units, not a second estimand.

Eigenvector centrality is closely linked to information diffusion. The diffusion centrality of household i (Banerjee et al., 2013) is the expected number of times households receive information originating from i when every informed household tells their friends with diffusion probability $1/\lambda$ per period. Normalized to unit length, it converges to eigenvector centrality as the number of periods increases.

2.3 Estimator from the sampled network

The sampling unit is the household. A sample is a set S of households with $\emptyset \neq S \subseteq V$,[4] with complement $\bar{S} := V \setminus S$ and contact set $N(S) := \bigcup_{i \in S} N(i)$.

Each sampled household $i \in S$ names its social contacts, and the names are matched back to V, the full list of households. Because a link is recorded whenever either household names the other, the nominations of i alone determine the ith row of A and, by symmetry, the ith column, including links to households in $\bar{S}$ that are never interviewed.

Assumption 1 (Link reporting and name matching)**.** *For every sampled household $i \in S$, the list of contacts it reports is $N(i)$, and each reported name is matched to the correct household in V. The network survey therefore reveals*

$$\{\, A_{ij} \,:\, i \in S \text{ or } j \in S \,\}. \tag{1}$$

Assumption 1 describes the star-sampled, OR-symmetrized survey convention of Chandrasekhar and Lewis (2016), Banerjee et al. (2013), Banerjee et al. (2019), and Beaman et al. (2021). It rules out censored reporting (Griffith, 2022), so the survey must not cap the number of contacts a household may name. It also rules out matching error, in which a reported name is attached to the wrong individual on the roster. Matching names to the roster is the final step of data collection and is costly (Breza et al., 2020). One way to match is to go through the full roster with each respondent, one household at a time (Breza et al., 2017). I maintain Assumption 1 throughout, for every sampling design discussed in the paper.

Under Assumption 1, the only entries of A that remain unobserved are those with $i, j \in \bar{S}$, the links between two unsampled households. Because the centrality of Definition 1 is a function of a matrix, I impute zeros for these entries, which erases any latent links among unsampled households. The two reasons for imputing zero are given after the block display below.

Definition 2 (Sample adjacency matrix)**.** *The sample adjacency matrix of S is*

$$\hat{A}_{ij} \;:=\; A_{ij}\,\mathbf{1}\{i \in S \text{ or } j \in S\}, \tag{2}$$

and $\hat{G}$ denotes the network with adjacency matrix $\hat{A}$.

[4] Full-network elicitation corresponds to $S = V$.

Ordering the nodes so that S precedes $\bar{S}$ displays the imputation as a single zero block:

$$A = \begin{pmatrix} A_{SS} & A_{S\bar{S}} \\ A_{S\bar{S}}^\top & A_{\bar{S}\bar{S}} \end{pmatrix}, \qquad \hat{A} = \begin{pmatrix} A_{SS} & A_{S\bar{S}} \\ A_{S\bar{S}}^\top & 0 \end{pmatrix}. \tag{3}$$

The two matrices differ only in $A_{\bar{S}\bar{S}}$, the adjacency matrix of the unobserved subgraph induced on $\bar{S}$. The imputation therefore errs only by omission, and $0 \leq \hat{A} \leq A$ entrywise. There are two reasons for the zero imputation. First, most networks are sparse. Second, under sequential sampling, the omitted links connect low-degree households with high probability so that the unobserved part of the network is sparse with high probability (Section 3.4). As a result, the error of Section 3.3 is small.

Definition 3 (Estimator)**.** *Given a sample S, let $\hat{\lambda} := \lambda_1(\hat{A})$ be the largest eigenvalue of the sample adjacency matrix $\hat{A}$. The estimated eigenvector centrality $\hat{v} := \hat{v}(S)$ is a nonnegative unit vector satisfying*

$$\hat{A}\,\hat{v} \;=\; \hat{\lambda}\,\hat{v}. \tag{4}$$

The estimator is Definition 1 applied to the sampled network $\hat{G}$. Because G is connected, every sampled household has an observed link, and a nonnegative unit eigenvector of $\hat{A}$ exists for every sample S. Two empirical implications follow. First, the estimator computes a centrality for every household in V. A household with no observed link receives centrality zero, and Section 4 uses this when the regression runs on every household in V. Second, whether the computed centrality is unique depends on the sample. It is unique when every sampled household can be reached from every other along observed links (Proposition 5 in the appendix). The sequential sampling design guarantees this on every realization, and simple random sampling does not (Section 3.1.1).

2.4 Challenge

Partial network data poses two statistical challenges for centrality estimation. First, the missingness pattern is structured and depends on the network G. They are exactly the entries $\{\, A_{ij} \,:\, i, j \in \bar{S} \,\}$, so the missingness concentrates among unsampled households. It is not ignorable in the sense of Rubin (1976). Matrix completion (Candes and Tao, 2010) is not suited to the problem either. Its low-rank assumption restricts the network structure and constrains the way in which individuals form links, and its nuclear-norm minimization targets the adjacency matrix rather than the eigenvector of interest.

The second challenge is that the estimand is an eigenvector and hence a *global* functional of A. Every household's centrality depends on every entry of the true adjacency matrix. Every unobserved link therefore distorts $\hat{v}$ at every household, sampled or un-sampled. The bias of the estimated individual centrality may be upward or downward depending on the realized sample.

The analytical corrections of Chandrasekhar and Lewis (2016) cover only a handful of local statistics, such as average degree and clustering. Thirkettle (2019) bounds centralities that are monotone in the network, such as Katz–Bonacich centrality, under a model of network formation. Inverse probability weighting (Horvitz and Thompson, 1952) does not resolve the problem either. It is unbiased for sums of unit-level quantities, such as the number of links

(Frank, 1977). However, the eigenvector centrality v is a nonlinear function of the whole matrix A.[5]

3 Sampled network

The preceding section defines the estimand as a functional of the full network G and the plug-in estimator $\hat{v}$ as a functional of the partially sampled network $\hat{G}$. The sampling design determines the sample S. Throughout the paper, the benchmark is simple random sampling.

Definition 4 (Simple random sampling)**.** *Simple random sampling draws $|S|$ households from population V uniformly at random without replacement, so that*

$$\mathbb{P}(i \in S) \;=\; \frac{|S|}{n}, \qquad i \in V. \tag{5}$$

The simple random sampling (5) depends on G only through its network size n. The sequential sampling design instead selects the next household along one of its links. This section formalizes the sequential sampling design as a Markov chain and establishes the consistency of $\hat{v}$.

Section 3.1 states the sampling algorithm and defines the sequential sampling design. Section 3.2 provides an overview of the analysis strategy. Section 3.3 treats the unobserved subgraph as a perturbation of the adjacency matrix A and bounds the estimation error. Section 3.4 derives the distribution of the sample and contrasts it with simple random sampling. Section 3.5 states the consistency result.

3.1 Sequential sampling design

The sequential sampling design follows the links of the network, in the lineage of the snowball sampling of sociology (Goodman, 1961) and the web sampling of computer science (Henzinger et al., 2000; Bar-Yossef et al., 2000). I adapt the design to the network surveys in economics (Banerjee et al., 2013, 2019).

The surveyor needs a randomization device, a sampling frame such as a household roster, and a sampling fraction set by the budget. The survey proceeds in three steps.

Algorithm 3.1 (Sequential sampling)**.**

(i) ***Start.*** *Interview a household drawn uniformly at random from the roster and record its contacts. Redraw if no contact is named.*

(ii) ***Draw.*** *Draw the next household uniformly at random from the contacts of the current household. It becomes the current household.*

(iii) ***Interview if new.*** *Interview the current household and record its contacts unless it has already been interviewed. Return to (ii) until the sampling fraction of the roster has been interviewed.*

[5] By power iteration, $v = \lim_{k\to\infty} A^k\mathbf{1}/\|A^k\mathbf{1}\|_2$ for every connected and non-bipartite graph G.

Algorithm 3.1 specifies where sequential sampling starts, how it moves from one household to the next, and which households it interviews. Definition 5 formalizes the three steps through the random variable H_t, the current household after t draws, so that H_0 is the initial household and the sample S collects the households interviewed until the budget is exhausted.

Definition 5 (Sequential sampling design)**.** *Let G be a connected graph with adjacency matrix A and degrees $d_1, \dots, d_n$. The sequential sampling design is the Markov chain $H_0, H_1, \dots$ on the state space V with*

$$H_0 \sim \mathrm{Unif}(V), \qquad \mathbb{P}\big(H_{t+1} = j \,\big|\, H_t = i\big) \;=\; \frac{A_{ij}}{d_i}, \qquad t \geq 0. \tag{6}$$

At sampling fraction $\alpha \in (0, 1]$, the sample S is the set of the first $\lceil \alpha n \rceil$ distinct states visited.

Under Assumption 1, the reported contacts of household i are its neighbors $N(i)$, so the uniform draw of step (ii) selects each of the d_i neighbors with probability $1/d_i$ and every other household with probability zero. This is the transition probability in (6). In matrix form, the transition probabilities are the entries of $P := D^{-1}A$, the adjacency matrix A with its ith row divided by the degree d_i. The transition matrix P does not depend on t, so sequential sampling is a time-homogeneous Markov chain.

The researcher chooses the sampling fraction $\alpha \in (0, 1]$, which fixes the budget at $\lceil \alpha n \rceil$ interviews. In other words, the surveyor stops as soon as the interviewed share of the roster reaches α, as in step (iii) of Algorithm 3.1. The budget determines the number of interviews but not the number of draws. The latter is random, because drawing a previously sampled household requires no new interview.[6] Indexing the sampling process by draws maintains the Markov property (6). Indexing it by the sample size would not, because the next distinct household depends on the entire sampling history. The sampling process would then not be a Markov chain on V, and the probabilistic tools of Section 3.4 would not apply.

3.1.1 Comparison with alternative sampling designs

Network sampling designs generally fall into two categories. *Frame-based designs*, such as simple random sampling, draw households from a sampling frame independently of the network G (Banerjee et al., 2013). *Link-tracing designs*, such as snowball and respondent-driven sampling, follow the links from an initial sample and serve populations without a frame (Goodman, 1961; Heckathorn, 1997). In this paper, the sequential sampling design traces links from a frame. Network surveys in economics first enumerate every name to construct a sampling frame and then match every reported contact to it, satisfying Assumption 1 (Breza et al., 2020). There are two main benefits. First, the frame allows for a uniformly drawn initial household. Second, the name-matching reveals the links between sampled and unsampled households, the block $A_{S\bar{S}}$, and restricts the missing links to the block $A_{\bar{S}\bar{S}}$. The consistency result uses both (Theorem 1).

Under simple random sampling, the sampled households can split across several components of the sampled network $\hat{G}$, and the estimator assigns zero centrality to every interviewed

[6] The number of draws is finite with probability one, because the sampling process is a finite irreducible Markov chain.

household outside the component with the largest eigenvalue. Under sequential sampling, the sampled households lie in one component of $\hat{G}$ by construction, and the estimator is unique.

The nearest relatives of the sequential sampling design are the link-tracing designs that build on a random walk. Respondent-driven sampling is modeled as a random walk, but respondents nominate their own recruits and each respondent is drawn without replacement (Gile and Handcock, 2010). In the randomized recruitment-driven sampling of Visokay et al. (2026), the researcher draws the recruits at random from each respondent's contacts. Without a frame, the population V is not enumerated. The named contacts therefore cannot be placed in V, and the block $A_{S\bar{S}}$ is undefined. The random walk designs of Klovdahl (1989) run several short walks, and Thompson (2006) adjusts the transition probabilities so that the walk visits every household with the same long-run frequency, which removes the degree-proportional visiting that Section 3.4 exploits.

The estimators for these designs target population totals by inverse probability weighting from the recruitment links and the reported degrees, which does not extend to nonlinear functions of the entire network such as eigenvector centrality (Section 2.4). Importantly, the sequential sampling design does not apply to every population. In the stigmatized populations that respondent-driven sampling serves, a sampling frame and full contact lists may be infeasible or unethical to collect (Visokay et al., 2026).

3.2 Overview of analysis

Sampling, by construction, leaves part of the population unobserved. This paper neither models nor recovers the unobserved links. Instead, I treat them as a perturbation of the adjacency matrix A and *bound* the deviation of the estimated centrality $\hat{v}$ from the true centrality v.

The analysis has a deterministic part and a stochastic part. The deterministic part bounds the error for every sample by the centrality that the unobserved block $A_{\bar{S}\bar{S}}$ carries, divided by a spectral gap (Section 3.3). The stochastic part is the distribution of the unobserved block, which the sequential sampling design determines through the transition matrix P of its random walk (Section 3.4). The structure of the missing data therefore comes from the sampling design, which the researcher controls, rather than from a model of link formation. Section 3.5 combines the two parts into a rate of convergence.

3.3 Matrix perturbation and missing links

In the block display (3), the adjacency matrix A of the full network is the sum of the sample adjacency matrix $\hat{A}$ of Definition 2 and the unobserved block $A_{\bar{S}\bar{S}}$,

$$A = \hat{A} + E, \qquad E := \begin{pmatrix} 0 & 0 \\ 0 & A_{\bar{S}\bar{S}} \end{pmatrix}. \tag{7}$$

The perturbation E vanishes outside $\bar{S} \times \bar{S}$. Inside the block, it records the links between two unsampled households, the only links the sample leaves unobserved. Figure 1 displays the block on a network of six households. Household 5 is not sampled, yet every one of its links joins it to a sampled household. All of its links are therefore observed, and the

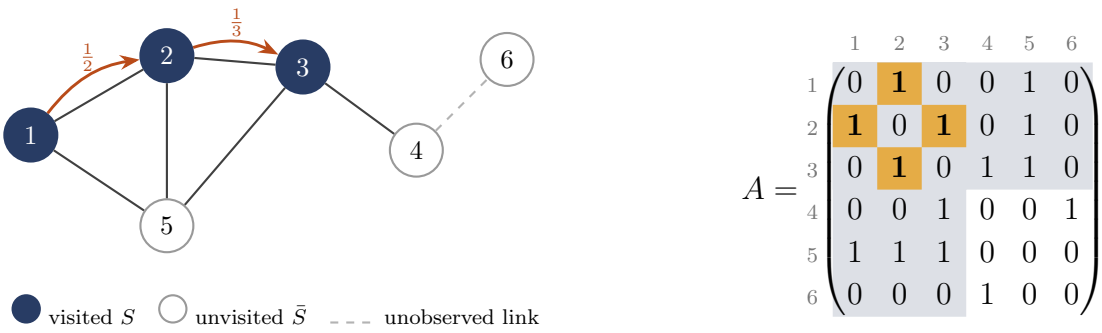


Figure 1: Sequential sampling on a network of six households.

Notes: The sequential sampling design starts at household 1 and makes two draws, so the sample is $S = \{1, 2, 3\}$. Left: colored households are sampled, solid links are observed, and the dashed link is the only link that the sample leaves unobserved. Right: the adjacency matrix A. The shaded rows and columns belong to the sampled households, the highlighted entries are the two draws, and the unshaded block collects the entries with both households unsampled, among them the dashed link.

observed network extends beyond the sample S. The dashed link is the only link joining two unsampled households and the only link that E records. Household 6 lies outside $S \cup N(S)$ and has no observed link, so $\hat{v}_6 = 0$ (Section 2.3).

The regression model of Section 4 uses $\hat{v}$ as a regressor. Thus, the relevant error is the Euclidean distance $\|\hat{v} - v\|_2$. Both v and $\hat{v}$ are nonnegative unit vectors, so this distance is at most $\sqrt{2}\sin\angle(v, \hat{v})$, where $\angle(v, \hat{v})$ is the angle between the eigenvectors of A and $\hat{A}$ for their largest eigenvalues. For every sample, $\sin\angle(v, \hat{v})$ is at most $\|Ev\|_2$ divided by $\lambda - \lambda_2(\hat{A})$ (Lemma 5 in the appendix, a form of the Davis–Kahan theorem, Davis and Kahan, 1970; Chen et al., 2021).

The quantity $\|Ev\|_2$ is the centrality that the unobserved block carries. The quantity $\lambda - \lambda_2(\hat{A})$ is the difference between the largest eigenvalue of A and the second-largest eigenvalue of $\hat{A}$. On the full network the corresponding difference is $\lambda - \lambda_2(A) = \delta\lambda$, where $\delta := 1 - \lambda_2(A)/\lambda$ is the relative spectral gap of A. The sample gap can be narrower than $\delta\lambda$, and how much narrower depends on the unobserved block (Lemma 4). Both parts of the bound are therefore functions of the unobserved block alone. What remains is the distribution of the sample S, which decides how much centrality the block carries and how far it narrows the gap.

3.4 Sampling distribution and network connectivity

The sampling process of Definition 5 is the simple random walk on G, started at a uniformly drawn household. Because G is connected, the random walk has a unique stationary distribution π, which has a closed form. Since $\sum_{i \in V} d_i = n\bar{d}$, every pair of households satisfies

$$\pi_i := \frac{d_i}{n\bar{d}}, \qquad \pi_i P_{ij} = \frac{A_{ij}}{n\bar{d}} = \pi_j P_{ji}, \qquad i, j \in V. \tag{8}$$

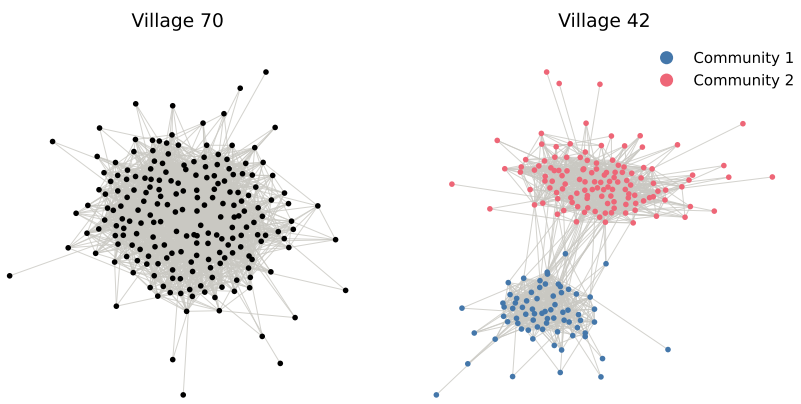


Figure 2: Community structure in two Karnataka village networks.

Notes: Each panel plots the largest connected component of the village network. In Village 42, color marks the two communities, identified by the sign of the second eigenvector of P.

This detailed balance condition makes π stationary and the random walk reversible. In the long run, the fraction of draws spent at household i is π_i, proportional to its degree d_i.[7]

The stationary distribution counts draws, not interviews. A draw that lands on a household already interviewed costs nothing. Indeed, it is a randomization on the surveyor's device. Because the contacts of that household have already been documented, it requires neither a visit nor an interview.

The global connectivity of the network,

$$\gamma := 1 - \lambda_2(P),$$

the spectral gap of the transition matrix P, governs how quickly sequential sampling explores the network. It equals the second-smallest eigenvalue of the normalized Laplacian, the connectivity measure of Jochmans and Weidner (2019).[8] Cheeger's inequality bounds it above and below in terms of the conductance of G, which defines the clusters for network cluster-robust inference in Leung (2023). Hence γ is small when the network has a bottleneck, a small set of households through which every path between two communities must pass. The same bottleneck slows the convergence to consensus when individuals average the beliefs of their friends (Golub and Jackson, 2012).

Under simple random sampling, every household is missed with the same probability. However, under sequential sampling, the probability that a household remains unsampled falls exponentially in its degree (Lemma 2 in the appendix). A link goes unobserved only when both of its households are unsampled. Applying the same lemma, the probability of omitting a link falls exponentially in the sum of their degrees. Hence, a missing link is likely

[7]Replacing the draw from the contacts by a uniform draw from V yields simple random sampling (Definition 4). The draws then form a Markov chain with transition matrix $n^{-1}\mathbf{1}\mathbf{1}^\top$ and stationary distribution $\mathrm{Unif}(V)$, both independent of G.

[8]The transition matrix $P = D^{-1}A$ is similar to the symmetrized matrix $D^{-1/2}AD^{-1/2}$. Because they share the spectrum, the normalized Laplacian $I - D^{-1/2}AD^{-1/2}$ has eigenvalues $1 - \lambda_k(P)$.

to join two households of low degree, which is precisely what the sampling process exploits for consistent estimation of eigenvector centrality.

Figure 2 plots the two Karnataka villages with the largest and the smallest global connectivity γ (Section 7.1). Village 70 has no bottleneck, but Village 42 splits into two communities joined by a few bridging households. The two communities are the Scheduled Caste households and the households that Banerjee et al. (2024) classify as upper caste. Only 2.5 percent of the links join them. Sequential sampling therefore tends to linger in the community where it starts. At a small sampling fraction, households in the other community are likely to remain unsampled, hubs included. The cross-sectional household survey records the caste of every household, which offers a diagnostic. If almost every sampled household belongs to one caste while the village has several, the sampling process is likely confined to one community, and the surveyor should start a second sampling process on the remaining households $V \setminus S$ (Remark 1).

Remark 1 (Parallel sampling)**.** *The global connectivity γ is unknown before the network survey. A sampling process confined to one community produces a long run of draws without a new interview. Two remedies are available.*

1. *A larger sampling fraction offsets a lower global connectivity, because the rate of Theorem 1 depends on the two only through their product $\alpha\gamma$. However, the sampling fraction cannot exceed one.*

2. *The researcher can start Algorithm 3.1 on the remaining households $V \setminus S$.*

Remark 2 (Disconnected components)**.** *A network with several connected components is the limiting case of bottlenecks where no household bridges the components. As a result, the global connectivity is zero, i.e., $\gamma = 0$, and the sampling process never leaves the component where it starts. This is verifiable without prior knowledge of the network. Once every contact of every interviewed household lies in S, no further draw can produce a new interview. The second remedy in Remark 1 is then necessary.*

Connectivity is also desirable for the estimation precision of fixed-effect network regressions (Jochmans and Weidner, 2019). Network cluster-robust inference, by contrast, benefits from weakly connected clusters (Leung, 2023), so the low connectivity that helps cluster-robust inference impedes sequential sampling.

3.5 Main result

The asymptotics of Theorem 1 run along a sequence of connected and undirected networks G_n on n nodes. The network G of Section 2 is one member of the sequence. Every quantity defined from G_n depends on n, including the adjacency matrix A, its largest eigenvalue λ, the average degree $\bar{d}$, the centrality v, and the estimator $\hat{v}$. For ease of reading, I suppress the index on these quantities. It appears only on G_n, on the convergence rate r_n defined below, and on the sequences that the proofs introduce.

The degree balance $\bar{d}/d_h$ is the ratio of the average degree to the harmonic mean degree $d_h := \left(\frac{1}{n}\sum_{i\in V} d_i^{-1}\right)^{-1}$. The relative spectral gap $\delta = 1 - \lambda_2(A)/\lambda$ is the gap between the two

largest eigenvalues of the adjacency matrix A as a fraction of the largest one (Section 3.3). The global connectivity $\gamma = 1 - \lambda_2(P)$ is the same gap for the transition matrix P, whose largest eigenvalue is one (Section 3.4).

With the sampling fraction α of Definition 5, define the convergence rate as

$$r_n := \frac{\bar{d}}{\alpha\gamma\delta\lambda}. \tag{9}$$

The theorem imposes two conditions on the network sequence G_n.

Assumption 2 (Network sequence)**.**

(i) Rate condition. $r_n \log n \to 0$ *as* $n \to \infty$.

(ii) Degree-balance condition. $\bar{d}/d_h \leq \Theta$ *for all* n*, for a constant* $\Theta < \infty$.

Theorem 1 (Consistency and rate of convergence)**.** *Let* $\{G_n\}_{n\geq 2}$ *be a sequence of connected graphs, each sampled by the sequential sampling design of Definition 5 at a fixed sampling fraction* $\alpha \in (0, 1]$*. Let* v *and* $\hat{v}$ *be the eigenvector centrality of Definition 1 and the estimator of Definition 3 on* G_n*. Under Assumption 2,*

$$\left\|\hat{v} - v\right\|_2 \;=\; O_p(r_n). \tag{10}$$

In particular, $\hat{v}$ *is consistent.*

The convergence rate r_n is a ratio of two components. The perturbation component $\bar{d}/(\alpha\gamma)$ bounds the effect of the missing links on the centrality in mean square (Lemma 3 in the appendix). The spectral gap component $\delta\lambda$ is the difference between the two largest eigenvalues of the adjacency matrix A (Section 3.3). The perturbation component is small when the sampling process rarely leaves a household of high degree unsampled. Three quantities decide this. The sampling fraction α, a budget parameter, sets the number of interviews, and every interview requires a draw (Section 3.1). In the long run, a draw lands on a household with probability proportional to its degree, so a household's degree matters relative to the average degree $\bar{d}$ (Section 3.4). Consecutive draws are dependent, and each counts for only a fraction γ of an independent draw from the long-run distribution, where the global connectivity γ measures how quickly the sampling process explores the network (Lemma 2, following Aldous and Fill, 2002, Chapter 3). Together, the three quantities make the missing links likely to join households of low degree (Section 3.4).

Consistency rests on the two conditions of Assumption 2. The rate condition requires the spectral gap component to exceed the perturbation component by more than a logarithmic factor. The missing links enter twice. They perturb the centrality, which the perturbation component bounds on average, and they narrow the spectral gap on the sampled network, which must be controlled in the worst case over all households (Lemma 4). The worst case is the unsampled household with the most unobserved links. Because the probability of remaining unsampled decays exponentially in the degree, this household has a degree of order at most $\log n$ times the perturbation component with probability tending to one, and the logarithm is the price of taking a maximum over n households. The degree-balance condition arises because the researcher has no prior information on the degrees and starts

the sampling process from a household drawn uniformly from the roster. Relative to a start from the long-run distribution, the uniform start costs at most the factor $(\bar{d}/d_h)^{1/2}$, which the condition keeps bounded.

The two conditions of Assumption 2 translate into three properties of the network. First, the degrees are heterogeneous, with some households holding many more links than others, a stylized fact of social and economic networks (Graham, 2017; Jackson et al., 2017). The rate condition requires $\lambda/\bar{d}$ to grow faster than $\log n$ whatever the other quantities do, because the sampling fraction is at most one and the two spectral gaps are at most two. Since $\lambda \leq d_{\max}$ (Lemma 1), the largest degree must then grow faster than $\log n$ times the average degree, which Erdős–Rényi graphs fail at every density. Two lower bounds, $\lambda \geq \sqrt{d_{\max}}$ and $\lambda^2 \geq \bar{d}^2 + \mathrm{Var}(d)$ (Hofmeister, 1988, Theorem 2.2), yield two sufficient conditions: a hub, with $\sqrt{d_{\max}}$ growing faster than $\log n$ times $\bar{d}$, or degree dispersion, with the standard deviation of the degrees growing faster than $\log n$ times $\bar{d}$.

Second, the network is well connected globally, and its largest eigenvalue is well separated from the second-largest. The global connectivity γ and the relative spectral gap δ multiply the rate, and the simplest way to satisfy the rate condition keeps both bounded away from zero, in which case the condition reduces to the first property. A bottleneck that divides the network into communities lowers γ, and the sampling process then tends to stay in the community where it starts (Section 3.4, Remark 1). A second-largest eigenvalue close to the largest lowers δ, and the missing links can then tilt the estimated centrality (Section 3.3).

Third, few households have degrees far below the average. The degree balance $\bar{d}/d_h$ then stays bounded, which is the degree-balance condition.

The theorem imposes no model of link formation and no node exchangeability, because the network G is fixed and the sampling design generates the randomness alone (Section 2). It restricts the average degree $\bar{d}$ only through its ratio to the largest eigenvalue λ, so the average degree may stay bounded or grow with n.

Remark 3 (Weighted networks)**.** *Theorem 1 carries over to weighted networks. In this case, each link ij carries a weight $A_{ij} \in (0, \infty)$ that measures the intensity of the relationship between i and j, and the degree $d_i = \sum_{j \in V} A_{ij}$ is the sum of the weights of the links of i. Assumption 1 then requires each sampled household to report the weights of all its links. Step (ii) of Algorithm 3.1 draws the next household j from the current household i with probability A_{ij}/d_i, as in (6). The proofs in the appendix apply with minor modifications.*

4 Regression model

The downstream analysis studies whether some economic outcome $Y_i \in \mathbb{R}$ of household i relates to its network position in G. There are two data sources. The first is a census that records an outcome and a covariate vector $(Y_i, Z_i) \in \mathbb{R} \times \mathbb{R}^q$ at every household $i \in V$. The first component of Z_i is the intercept term. The second data source is a network survey that covers only the sampled households $S \subseteq V$, because network elicitation is costly and slow.

Throughout this section v denotes the root-n rescaled centrality, normalized by $\frac{1}{\sqrt{n}}\|v\|_2 = 1$ rather than by the unit norm of Definition 1.[9] Under this normalization the average squared

[9]This normalization also appears in Cai (2026).

centrality is one. However, under the unit norm, it becomes $1/n$. Hence, the identifying variation condition of Assumption 4 below would fail mechanically. Under this normalization, Theorem 1 reads $n^{-1/2}\|\hat{v} - v\|_2 = O_p(r_n)$. The regression model is

$$Y_i = \beta\, v_i + Z_i^\top \theta + \epsilon_i, \qquad i \in V, \tag{11}$$

where $\beta \in \mathbb{R}$ is the parameter of interest and $\theta \in \mathbb{R}^q$ collects the coefficients on the covariates. The graph G is fixed and deterministic, so the centrality v is fixed.

4.1 Plug-in regression

Stack the covariate rows as $Z := (Z_1, \ldots, Z_n)^\top \in \mathbb{R}^{n\times q}$ and define the full-network and plug-in regressor matrices

$$X := [\, v \;\; Z\,], \qquad \hat{X} := [\, \hat{v} \;\; Z\,], \qquad X,\ \hat{X} \in \mathbb{R}^{n\times(1+q)}, \tag{12}$$

which differ in the first column only. The column v requires the full network. The plug-in column $\hat{v}$ requires only the sampled adjacency $\hat{A}$. The corresponding least-squares estimators are

$$(\tilde{\beta},\ \tilde{\theta}^\top)^\top := (X^\top X)^{-1} X^\top Y, \qquad (\hat{\beta},\ \hat{\theta}^\top)^\top := (\hat{X}^\top \hat{X})^{-1} \hat{X}^\top Y. \tag{13}$$

The first requires the full network and serves as the benchmark. The second uses only the sampled network, which is the focus of this section.[10]

The second-stage regression (13) uses *every* household. The census records (Y_i, Z_i) on the roster V, and Assumption 1 matches every reported name to that roster. Hence, $\hat{A}$ is indexed by the census, and the plug-in regressor is defined on all of V rather than on the sample S alone (Proposition 5). Restricting the regression to the surveyed households would instead select on degree. By Lemma 2, the probability that household i is left out of S decays exponentially in its degree. The restricted sample therefore tilts toward high-degree households and induces selection bias.

The plug-in regressor $\hat{v}$ is a deterministic function of the realized sample S (Section 3.1). Thus, conditional on (G, Z), the feasible regression is random through S and through the error $\epsilon := (\epsilon_1, \ldots, \epsilon_n)^\top$ only. Every probability statement in this section and in Section 5 is conditional on (G, Z). The controls Z and the error ϵ are defined for each n along the sequence of Section 3.5. The common parameter β does not depend on n. The following assumption restricts the error and its relation to the sample.

Assumption 3 (Regression error: consistency)**.**

(i) Exogenous sampling. *Conditional on the network G, sampling is independent of regression error ϵ and controls Z, $S \perp (\epsilon, Z) \mid G$.*

(ii) Conditional mean zero. *The error has conditional mean zero, $\mathbb{E}[\epsilon \mid G, Z] = 0$.*

[10]The inverse in (13) exists exactly when $\hat{X}$ has full column rank. That is, when the controls Z have full column rank and $\hat{v}$ lies outside their span. Every statement about the feasible regression is conditional on this event, which is a function of (S, Z) and has probability tending to one under the conditions of Proposition 1 (Lemma 10 in the appendix).

(iii) Bounded second moments. *Conditional on (G, Z), the errors are uncorrelated with* $\mathrm{Var}(\epsilon_i \mid G, Z) \le \bar{\sigma}^2$ *for all* $i \in V$*, for a constant* $\bar{\sigma}^2 < \infty$.

Part (i) says that the enumerator selects households independent of their outcomes. It fails, for example, when a surveyor follows referrals toward better-off households. It also means that conditioning on covariates Z does not change the sampling distribution.

Part (ii) is stronger than orthogonality of the error to the regressors, which is sufficient for the full-network regression. The feasible regression replaces v by the plug-in regressor $\hat{v}$, which varies with the sample S. The results of Section 5 hold conditional on the sample and therefore need $\hat{v}^\top \mathbb{E}[\epsilon \mid G, Z] = 0$ for every realization of $\hat{v}$. Orthogonality to v imposes only the single restriction $v^\top \mathbb{E}[\epsilon \mid G, Z] = 0$ on the conditional mean of the error. Parts (i) and (ii) together give $\mathbb{E}[\epsilon \mid S, G, Z] = 0$. Hence, the error is orthogonal to every function of (S, G, Z), in particular to the plug-in regressor $\hat{v}$ and to its residual after partialling out the controls, which is the form Section 5 uses. Part (iii) bounds the conditional variance of the error.

Assumption 4 (Identifying variation)**.**

(i) Full column rank. *The controls Z have full column rank.*

(ii) Residual variation. *The centrality retains variation after partialling out the controls: there exists $\mu_0 > 0$ such that, for all n sufficiently large,*

$$\tfrac{1}{n} v^\top M_Z v \;\ge\; \mu_0, \qquad M_Z := I - Z(Z^\top Z)^{-1} Z^\top. \tag{14}$$

Part (i) rules out collinear controls. Part (ii) requires that the centrality is not asymptotically collinear with them, and the two parts together identify the slope β, since the regressor matrix X in (12) has full column rank exactly when Z has full column rank and $v^\top M_Z v > 0$. The left side of (14) is the average squared residual from the regression of the centrality v on the controls Z, and it depends on the controls only through their span, which makes (14) invariant to their units and parametrization. With an intercept among the controls, (14) fails when the centrality is asymptotically constant across households. This is the case in a regular network, where every household has the same degree and hence the same centrality and $M_Z v = 0$, and in the limit for a sufficiently dense Erdős–Rényi sequence, whose eigenvector centrality converges to the constant vector.

Under Assumptions 3 and 4, the first-stage bound of Theorem 1 delivers consistency of the plug-in slope $\hat{\beta}$.

Proposition 1 (Slope consistency)**.** *Let the conditions of Theorem 1 hold together with Assumptions 3 and 4. Then, conditional on (G, Z), $\hat{\beta} \xrightarrow{p} \beta$.*

5 Inference

The estimated centrality $\hat{v}$ is computed from the sampled network $\hat{G}$. Hence, it measures the eigenvector centrality v of the network G with measurement error. Used as a generated regressor in the second-stage regression, $\hat{v}$ can bias the slope and invalidate conventional inference (Pagan, 1984; Battaglia et al., 2025).

The plug-in slope decomposes exactly as $\hat{\beta} = \rho(S)\,\beta + e(S)$, where the reliability $\rho(S)$ is a function of the realized sample S and the noise term $e(S)$ has mean zero conditional on (S, G, Z) under Assumption 3 (Section 5.1). The bias is therefore multiplicative in β and vanishes when $\beta = 0$. For this reason, the heteroskedasticity-robust t-test of the null $\mathcal{H}_0 : \beta = 0$ is asymptotically exact without any correction for the first stage (Section 5.2). When $\beta \neq 0$, the rate of Theorem 1 does not imply that the bias is asymptotically negligible, and Section 5.3 estimates the bias factor by a bootstrap.

5.1 Bias decomposition

Let $P_Z := I - M_Z$ denote the projector onto the span of the controls Z, with M_Z as in (14). Write $\hat{w} := M_Z \hat{v}$ for the plug-in regressor after partialling out the controls Z. Under part (i) of Assumption 4, the regressor matrix $\hat{X}$ has full column rank exactly when $\hat{w} \neq 0$. On this event, the Frisch–Waugh–Lovell theorem gives the plug-in slope in (13) as $\hat{\beta} = \hat{w}^\top Y / \|\hat{w}\|_2^2$. Define the reliability $\rho(S) := \hat{w}^\top v / \|\hat{w}\|_2^2$ and the noise term $e(S) := \hat{w}^\top \epsilon / \|\hat{w}\|_2^2$. Every statement about $\hat{\beta}$, $\rho(S)$, and $e(S)$ below, including every expectation over S, is conditional on this event. Write $u := \hat{v} - v$ for the measurement error.

Proposition 2 (Bias decomposition)**.** *Let part (i) of Assumption 4 hold. Then, conditional on (G, Z),*

$$\hat{\beta} \;=\; \rho(S)\,\beta \;+\; e(S). \tag{15}$$

(i) Under Assumption 3, the noise term has conditional mean zero, $\mathbb{E}[e(S) \mid S, G, Z] = 0$. Hence $\mathbb{E}[\hat{\beta} \mid S, G, Z] = \rho(S)\,\beta$, and the bias equals

$$\mathbb{E}[\hat{\beta} \mid G, Z] - \beta \;=\; \big(\mathbb{E}[\rho(S) \mid G, Z] - 1\big)\,\beta. \tag{16}$$

(ii) Under the conditions of Theorem 1 and Assumptions 3 and 4, the reliability satisfies $\rho(S) - 1 = O_p(r_n)$ and the noise term satisfies $e(S) = O_p(n^{-1/2})$.

The bias is multiplicative in β. Define the bias factor

$$\bar{\rho} := \mathbb{E}[\rho(S) \mid G, Z].$$

At the null $\beta = 0$, the plug-in slope is exactly unbiased under any sampling design and at any sampling fraction α.

When $\beta \neq 0$, the bias is governed by the reliability $\rho(S)$, the coefficient from the least-squares regression of the unobserved regressor v on the plug-in regressor $\hat{v}$ after partialling out the controls Z. It generalizes the reliability ratio of the classical measurement-error model (Fuller, 1987; Bound et al., 2001), in which the measurement error is orthogonal to the true regressor. Here, the measurement error u is not orthogonal to v, and $\rho(S) - 1$ has two terms (equation (37) in the appendix). The first, $-\frac{1}{2}\|u\|_2^2 / \|\hat{w}\|_2^2$, is nonpositive and of second order in u. The second, $\hat{v}^\top P_Z u / \|\hat{w}\|_2^2$, is of first order in u and has no determinate sign. It is present in every regression of the model of Section 4, whose controls include an intercept.[11]

[11] With an intercept only, as in the simulations of Section 6, the reliability $\rho(S)$ is at most one whenever the mean of the plug-in centrality $\hat{v}$ does not exceed the mean of the centrality v (equation (45) in the appendix).

The two terms of (15) have different rates. The noise term $e(S)$ is of order $n^{-1/2}$, as in the full-network regression. The bias term is bounded by the first-stage rate r_n, and Proposition 6 in the appendix establishes $\sqrt{n}\, r_n \geq \frac{1}{4}$ on every connected graph with at least two nodes. Hence, Theorem 1 does not imply that the bias term is asymptotically negligible relative to the noise term.

Inference therefore proceeds in two cases. When $\beta = 0$, the bias term vanishes and only the noise term remains. When $\beta \neq 0$, the bias term persists and its factor $\bar{\rho}$ must be estimated and removed. The null $\mathcal{H}_0 : \beta = 0$ of Section 5.2 belongs to the first case. A null $\mathcal{H}_0 : \beta = c$ with $c \neq 0$ belongs to the second.

5.2 Asymptotic distribution

I now develop distribution theory for $\hat{\beta}$. Conditional on (S, G, Z), the noise term $e(S)$ of (15) is a weighted sum of the errors with weights $\hat{w}$, and at the null it equals $\hat{\beta}$ exactly. For the central limit theorem, I strengthen the error assumptions as follows.

Assumption 5 (Regression error: inference)**.** *Let* $\sigma_i^2 := \mathrm{Var}(\epsilon_i \mid G, Z)$*. Assume:*

(iv) Conditional independence. *Conditional on* (G, Z)*, the errors* $\epsilon_1, \ldots, \epsilon_n$ *are mutually independent.*

(v) Conditional variance floor. $\sigma_i^2 \geq \underline{\sigma}^2$ *for all* $i \in V$*, for a constant* $\underline{\sigma}^2 > 0$*.*

(vi) Bounded fourth moments. $\mathbb{E}[\epsilon_i^4 \mid G, Z] \leq C$ *for all* $i \in V$*, for a constant* $C < \infty$*.*

Parts (iv) to (vi) are the error conditions of heteroskedasticity-robust inference (Eicker, 1963): independent errors with a variance floor and bounded fourth moments. They place no restriction on the network or on the sampling design. The distribution theory needs one more classical condition on X, which is stated below.

Assumption 6 (Negligible leverage)**.** *The maximum leverage in the full-network regression* (12) *vanishes as* $n \to \infty$*,*

$$h_n := \max_{i \in V} \; X_i^\top (X^\top X)^{-1} X_i \;\longrightarrow\; 0. \tag{17}$$

Assumption 6 restricts the design matrix and gives the Lindeberg condition for the least-squares coefficients (Kline et al., 2020). Because the centrality v is a column of X, it implies that no household carries a fixed share of the centrality mass, $\max_{i \in V} v_i^2 / \sum_{j \in V} v_j^2 \to 0$. This share is at most $d_{\max}/\lambda^2$ (Proposition 7 in the appendix).

The distribution theory therefore restricts the degree sequence of G in two ways. Theorem 1 requires the largest eigenvalue λ to grow faster than $\log n$ times the average degree $\bar{d}$, which calls for heterogeneous degrees. Assumption 6 rules out a single hub that carries the centrality, as in a star graph. Under Assumption 4, it holds when λ grows faster than the square root of the maximum degree $d_{\max}$ and the maximum leverage of the regression on the controls Z alone vanishes (Proposition 7 in the appendix). In other words, consistency of $\hat{v}$ needs heterogeneous degrees, whereas inference on $\hat{\beta}$ needs that no single hub dominates.

When $\beta = 0$, the plug-in slope $\hat{\beta}$ equals the noise term $e(S)$ of (15), a weighted sum of the regression errors with weights determined by the sample S.

Proposition 3 (The t-test at the null)**.** *Let the conditions of Theorem 1 hold, together with Assumptions 3, 4, 5, and 6. Let $\hat{\epsilon} := Y - \hat{X}(\hat{\beta}, \hat{\theta}^\top)^\top$ be the residual vector of the feasible regression (13), and define the heteroskedasticity-robust t-statistic*

$$t := \frac{\hat{\beta}}{\hat{\omega}}, \qquad \hat{\omega}^2 := \frac{\sum_{i \in V} \hat{w}_i^2 \, \hat{\epsilon}_i^2}{(\hat{w}^\top \hat{w})^2}. \tag{18}$$

Under $\mathcal{H}_0 : \beta = 0$, conditional on (G, Z),

$$t \xrightarrow{d} N(0, 1).$$

The statistic t is the robust t-statistic that standard software reports for the regression of Y on $\hat{X}$.[12] Testing $\mathcal{H}_0 : \beta = 0$ therefore requires no correction for the first stage. The sampled-network test is asymptotically equivalent to the full-network test (Proposition 4 in the appendix) and attains its asymptotic size. Against a fixed alternative $\beta \neq 0$ the test is consistent (Corollary 2 in the appendix). The result rests on Assumption 6. If one household carries a non-vanishing share of the weights $\hat{w}$, then $\hat{\beta}$ remains consistent (Proposition 1) but t need not be asymptotically normal.

5.3 Bias correction

When $\beta \neq 0$, the plug-in slope $\hat{\beta}$ estimates $\bar{\rho}\,\beta$ by (16). Removing the bias therefore requires an estimate of the factor $\bar{\rho}$. The reliability depends on the unobserved block $A_{\bar{S}\bar{S}}$, so the factor cannot be computed from the sampled network $\hat{G}$. The factor can be estimated by resampling, because the links omitted from $\hat{G}$ were removed by a known sampling design.

Given a single sampled network $\hat{G}$, existing network bootstrap methods generally resample by imposing a network formation model (Levin and Levina, 2025) or node exchangeability (Bhattacharyya and Bickel, 2015). In the measurement error literature, simulation-extrapolation (Cook and Stefanski, 1994) adds measurement error of known variance incrementally and extrapolates the slope estimates back to the case of no measurement error. The correction proposed below instead reruns the sampling design on $\hat{G}$, which adds a second layer of sampling variation to $\hat{v}$ and replicates the bias arising from missing links.

Algorithm 5.1 (Bootstrap bias correction)**.**

(i) ***Resample.*** *Run Algorithm 3.1 on the sampled network $\hat{G}$ in place of G, with the households in $S \cup N(S)$ as the household roster, at the same sampling fraction α. The bootstrap sample S_b^* is the set of the first $\lceil \alpha \, |S \cup N(S)| \rceil$ distinct households drawn. Compute the bootstrap centrality $\hat{v}_b^*$ of the bootstrap sample. Repeat independently for $b = 1, \ldots, B$.*

(ii) ***Regress.*** *Regress the plug-in centrality $\hat{v}$ on the bootstrap centrality $\hat{v}_b^*$ and the controls Z over the households in $S \cup N(S)$, and record the coefficient ρ_b^* on $\hat{v}_b^*$.*

[12] By the Frisch–Waugh–Lovell theorem, $\hat{\omega}^2$ in (18) is the $(1,1)$ element of the heteroskedasticity-consistent covariance matrix of this regression.

(iii) ***Correct.*** *Average the coefficients,* $\bar{\rho}^* := B^{-1} \sum_{b=1}^{B} \rho_b^*$*, and report the corrected slope* $\check{\beta} := \hat{\beta}/\bar{\rho}^*$.

Algorithm 5.1 is a bootstrap in which the sampled network $\hat{G}$ plays the role of the full network G. The plug-in centrality $\hat{v}$ plays the role of the true centrality v. Similarly, the households in $S \cup N(S)$, the households with observed links, play the role of the roster V. The bootstrap centrality $\hat{v}_b^*$ is the estimator of Definition 3 computed from the sample adjacency matrix (2) of S_b^*, with $\hat{A}$ in place of A, and normalized as $\|\hat{v}_b^*\|_2 = \sqrt{n}$ just like $\hat{v}$. Hence, the coefficient ρ_b^* in Step (ii) relates $\hat{v}_b^*$ to $\hat{v}$ exactly as the reliability $\rho(S)$ relates $\hat{v}$ to v. Because of the sequential sampling design, $\hat{v}_b^*$ is unique on every bootstrap draw (Proposition 5).

The division by $\bar{\rho}^*$ in Step (iii) reflects the multiplicative form of the bias, $\mathbb{E}[\hat{\beta} \mid S, G, Z] = \rho(S)\,\beta$ (Proposition 2), whereas a standard bootstrap subtracts its estimate of an additive bias. The factor $\bar{\rho}^*$ is a function of $(\hat{G}, Z)$ and of bootstrap draws independent of ϵ. Hence, for $\bar{\rho}^* \neq 0$, (15) and part (i) of Proposition 2 give

$$\mathbb{E}[\check{\beta} \mid S, \bar{\rho}^*, G, Z] \;=\; \frac{\rho(S)}{\bar{\rho}^*}\,\beta. \tag{19}$$

At the null, $\check{\beta}$ is therefore exactly unbiased. Away from the null, the corrected slope is conditionally unbiased when the bootstrap reproduces the realized reliability, $\bar{\rho}^* = \rho(S)$, and its relative bias is $(\rho(S) - \bar{\rho}^*)/\bar{\rho}^*$ otherwise. Whether $\bar{\rho}^*$ tracks $\rho(S)$ depends on how closely resampling from $\hat{G}$ reproduces the missingness that produced $\hat{G}$.

6 Simulation

This section evaluates the sequential sampling design of Section 3.1 by Monte Carlo, with the simple random sampling benchmark of Definition 4. The simulated graphs come from four canonical random graph families, calibrated to the size and average degree of the Karnataka village networks of the application (Section 7.1, Table 4). Section 6.1 describes the families and the calibration. Section 6.2 compares first-stage estimation accuracy across sampling budgets. Section 6.3 examines second-stage estimation of the regression slope.

6.1 Random graphs

The four families are Barabási–Albert, Chung–Lu, Erdős–Rényi, and the stochastic block model. The first two generate hubs and heavy-tailed degree distributions, the last two are homogeneous and generate neither. The Barabási–Albert family satisfies the hub condition of Section 3.5 but not the dispersion condition. The Chung–Lu family with tail exponent in $(2, 3)$ and a growing average degree achieves dispersion without a hub, because its largest eigenvalue then tracks the dispersion of the degrees rather than the largest degree (Chung et al., 2003). Table 1 reports summary statistics of the simulated graphs, Table 4 in Section 7.1 summary statistics of the villages, and Figure 3 the distributions of degree, centrality, and the largest eigenvalue.

The Barabási–Albert model (Barabási and Albert, 1999) grows the network one household at a time, each entrant linking to a fixed number of existing nodes with probability proportional

Table 1: Statistics of the random graphs.

	$\bar{d}$	σ_d	$d_{\max}$	$\bar{v}$	σ_v	$v_{\max}$	λ	δ	γ
Barabási–Albert	15.4	10.5	68	0.832	0.555	3.534	22.2	0.64	0.54
Chung–Lu	16.0	10.1	59	0.838	0.545	3.080	22.4	0.64	0.54
Erdős–Rényi	16.0	3.8	27	0.968	0.249	1.723	16.9	0.56	0.54
Stochastic block model	15.9	3.8	27	0.969	0.248	1.716	16.8	0.35	0.34

Notes: Entries are means over the 100 generated graphs per family at $n = 200$. σ_d and σ_v are the standard deviations of degree and centrality across households, and $v_{\max}$ is the largest centrality. Degree and centrality are computed on all n households, with centrality normalized as $n^{-1/2}\|v\|_2 = 1$, and λ, δ, and γ on the largest connected component.

to their degrees, so early entrants accumulate links, the mechanism behind the heavy-tailed degree distributions of many observed networks (Voitalov et al., 2019). The Chung–Lu model (Chung and Lu, 2002) instead fixes a power-law expected degree sequence and links each pair independently with probability proportional to the product of their expected degrees, so the tail exponent is a parameter the researcher chooses rather than a constant fixed by the growth process.

The Erdős–Rényi model links each of the $\binom{n}{2}$ pairs independently with a common probability p. The stochastic block model partitions V into two equal blocks and links pairs independently with probability p_w within a block and $p_b < p_w$ across blocks, the network analogue of homophily, with p_w and p_b set so that the expected degree equals that of the Erdős–Rényi specification. The two homogeneous families therefore differ primarily in community structure.

The calibration is anchored to the 33 villages of Table 4. Village sizes range from 114 households to 375, with a mean of 222 and a standard deviation of 68 across villages. The simulated graphs of Table 1 have $n = 200$ households, the rounded village mean. The village networks are also sparse. Even after aggregating twelve types of relations, the average household is linked to only a small fraction of the village. Hence, every family is calibrated to an expected mean degree of 16, close to the village mean of 17.4.

Beyond the two targeted moments, Tables 1 and 4 and Figure 3 distinguish the homogeneous and heavy-tailed families on three untargeted moments. The first is the presence of hubs. The average village has a maximum degree of 53, and the heavy-tailed families generate hubs of comparable size. However, the homogeneous families reproduce only about half this figure, as seen in the right tail in panel (a). The second is heterogeneity in degree and in centrality. The heavy-tailed families reproduce the village degree dispersion σ_d, whereas the homogeneous families generate about two fifths of it. In the downstream analysis, this weakens the identifying variation of Assumption 4, and a regular graph fails the assumption exactly. Panel (b) shows that the heavy-tailed families also reproduce the right tail and the dispersion of village centrality but not its shape. The village histogram places more mass on peripheral households with low centrality than either family generates. The third is the largest eigenvalue, which enters the rate r_n. The heavy-tailed families match the village average of 22.8 within three percent, while the homogeneous families fall short by a quarter, as panel (c) shows. The sharp contrast between the two types of families in Tables 2 and 3 therefore isolates the role of degree heterogeneity.

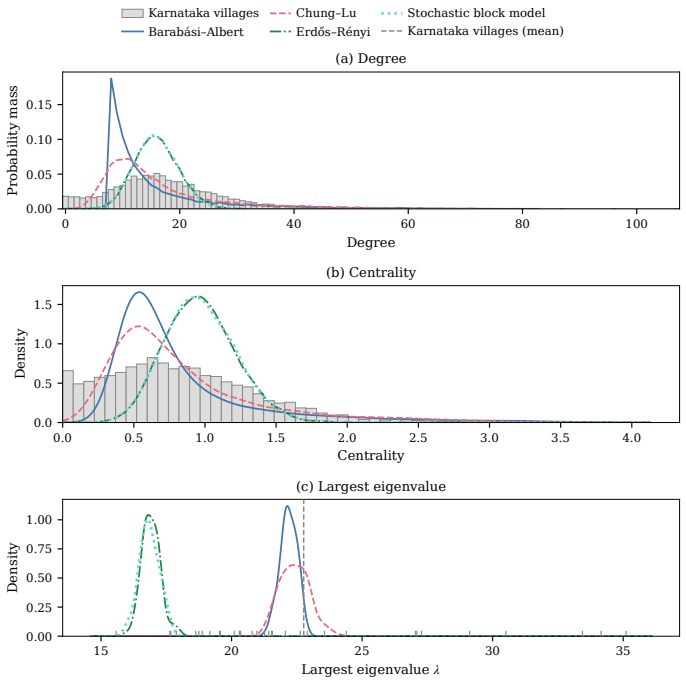


Figure 3: Degree, centrality, and largest-eigenvalue distributions.

Notes: Histograms pool the households of the 33 villages of Table 4, and curves are computed from the 100 generated graphs per family at $n = 200$. Centrality is normalized as $n^{-1/2}\|v\|_2 = 1$ on the full roster, and the largest eigenvalue is computed on the largest connected component. In panel (c), ticks mark the 33 village eigenvalues and the dashed line their mean.

The remaining two quantities in the rate, the relative spectral gap δ and the global connectivity γ, are matched less closely. Every family exceeds the village relative spectral gap of 0.33, the stochastic block model only narrowly at 0.35, and every family exceeds the village global connectivity of 0.22. Hence the simulated graphs are better connected globally than the real villages. Community structure in the village networks is one plausible explanation, as the stochastic block model comes closest on both quantities. Smaller values of both raise the rate r_n in (9), so the villages are the more challenging case for the sequential sampling design, and the application of Section 7 tests it where the theory predicts slower convergence.

6.2 First-stage estimation of network centrality

This section assesses the first stage, the estimation of eigenvector centrality from the partially observed network. I compare the sequential sampling design of Definition 5 with simple random sampling (Definition 4). At every budget the two designs interview the identical number of nodes, so paired entries in Table 2 hold the survey cost fixed. The grid varies the network size over $100, 200, 1000$ and the budget over $10\%, 30\%, 50\%$ of the network size.

The simulation has two sources of randomness, the graph and the sample. For each family, I generate 100 graphs and hold each fixed. Hence, each graph's adjacency matrix A and centrality v are deterministic, matching the design-based framework of Section 3.1. Each design is run 20 times on each graph, and every replication produces a sample adjacency

Table 2: First-stage accuracy of the network centrality.

	$n = 100$			$n = 200$			$n = 1000$		
Budget	10%	30%	50%	10%	30%	50%	10%	30%	50%
Barabási–Albert									
Sequential sampling	0.809	0.950	0.987	0.834	0.962	0.991	0.892	0.984	0.997
Simple random sampling	0.756	0.910	0.965	0.744	0.908	0.964	0.702	0.897	0.960
Chung–Lu									
Sequential sampling	0.795	0.939	0.983	0.822	0.956	0.990	0.895	0.986	0.998
Simple random sampling	0.762	0.912	0.966	0.750	0.909	0.964	0.727	0.904	0.963
Erdős–Rényi									
Sequential sampling	0.728	0.900	0.963	0.712	0.896	0.961	0.698	0.891	0.959
Simple random sampling	0.742	0.905	0.963	0.725	0.899	0.961	0.706	0.895	0.959
Stochastic block model									
Sequential sampling	0.706	0.884	0.955	0.697	0.886	0.957	0.690	0.889	0.959
Simple random sampling	0.731	0.899	0.960	0.717	0.896	0.960	0.702	0.894	0.959

Notes: Entries are the mean cosine similarity $\cos\angle(\hat{v}, v)$ of (20) over 100 graphs and 20 sampling replications per graph and design. The budget is the share of households interviewed, the same under both designs.

matrix $\hat{A}$ and the plug-in centrality $\hat{v}$ of Definition 3. Table 2 reports means over the 100 graphs and 20 replications per graph and design.

Accuracy is measured by the cosine similarity

$$\cos\angle(\hat{v}, v) \;=\; \frac{\hat{v}^\top v}{\|\hat{v}\|_2 \, \|v\|_2}. \tag{20}$$

The ratio is invariant to rescaling of either vector, so the measure requires no normalization convention.[13] By the Perron–Frobenius theorem $\hat{v}$ is entrywise nonnegative and v is strictly positive on a connected graph, so the similarity lies in $(0, 1]$ and equals one exactly when $\hat{v}$ is proportional to v. Importantly, the measure also has a regression interpretation. In the model $Y = \beta v + \epsilon$, the least-squares slope of Y on $\hat{v}$ without an intercept, with both vectors normalized to a common norm, is centered at $\cos\angle(\hat{v}, v) \cdot \beta$ conditional on the sample. Each entry of Table 2 is therefore the fraction of the true slope that the plug-in regression retains in this illustration.

The sequential sampling design outperforms simple random sampling on the Barabási–Albert and Chung–Lu families at every network size and every budget, and the advantage widens with the network size. At the 10% budget on Barabási–Albert graphs, the sequential sampling design improves from 0.809 at $n = 100$ to 0.892 at $n = 1000$, while simple random sampling deteriorates from 0.756 to 0.702. The Chung–Lu columns display the same pattern. At the 30% budget the sequential sampling design reaches at least 0.93 in every heavy-tailed entry, a level simple random sampling does not attain until the budget rises to 50%. A sequential sample at 30% therefore comes close to the accuracy of a simple random sample at 50%, a saving of two fifths of the interviews.

[13] Under a common norm, it is bounded by Theorem 1, since $1 - \cos\angle(\hat{v}, v) = \frac{1}{2}\,\|\hat{v}/\|\hat{v}\|_2 - v/\|v\|_2\|_2^2$.

Table 3: Second-stage bias of the plug-in slope.

	Sequential sampling			Simple random sampling		
Budget	10%	30%	50%	10%	30%	50%
Barabási–Albert	−0.294	−0.108	−0.045	−0.547	−0.269	−0.122
	(0.069)	(0.044)	(0.045)	(0.101)	(0.063)	(0.050)
Chung–Lu	−0.270	−0.089	−0.032	−0.497	−0.237	−0.105
	(0.061)	(0.043)	(0.043)	(0.091)	(0.057)	(0.047)
Erdős–Rényi	−0.894	−0.739	−0.532	−0.901	−0.755	−0.541
	(0.048)	(0.064)	(0.081)	(0.048)	(0.067)	(0.087)
Stochastic block model	−0.896	−0.743	−0.536	−0.903	−0.756	−0.542
	(0.048)	(0.064)	(0.082)	(0.048)	(0.067)	(0.087)

Notes: Entries are computed over 100 graphs and 20 sampling replications per graph and design at $n = 1000$, with $\beta = 1.5$ and $\sigma^2 = 1$. Each entry reports the relative bias $\mathbb{E}[(\hat{\beta} - \beta)/\beta \mid S] = \rho(S) - 1$, computed exactly for each sample and averaged across replications and graphs, with the standard deviation of $\hat{\beta}$ across replications within a graph, averaged over graphs, in parentheses. Budgets are as in Table 2.

On the Erdős–Rényi and stochastic block model families the two designs perform nearly identically, and accuracy depends on the budget alone: within each row of the two lower panels, the entries vary by at most 0.04 as the network size grows tenfold.

The contrast between the two pairs of families follows from the stationary distribution (8). The sampling process visits nodes in proportion to their degrees. In a heavy-tailed graph, the eigenvector centrality tend to concentrate on the hubs, and the process is likely to reach them faster. In the homogeneous families degrees concentrate near the mean, so the stationary distribution is nearly uniform and the sequential sampling design draws nodes almost as simple random sampling does. Under budget constraints, degree-biased sampling is able to leverage degree heterogeneity in the network to find an informative sample.

The results in Table 2 also warrant caution. Even a similarity of 0.96 leaves a four percent bias in the plug-in slope under this interpretation, and Theorem 1 does not imply that this bias vanishes fast enough relative to the standard error as the network grows (Section 5.1). Section 6.3 quantifies the consequences for estimating β.

6.3 Second-stage estimation of the regression slope

This section examines the second-stage regression of the outcome on the estimated network centrality. The outcome model is $Y_i = \beta v_i + \epsilon_i$ with $\beta = 1.5$ and errors that are i.i.d. $N(0, 1)$ and independent of the sampling process. Both the true and the estimated centrality enter the regression under the normalization $n^{-1/2}\|v\|_2 = 1$. The fitted regression is the plug-in least squares of (13) with an intercept and no further controls. Because the plug-in slope is linear in the errors, its mean and variance given the sample are available in closed form (Proposition 2), and I compute them rather than draw errors. The relative bias given the sample, $\mathbb{E}[\hat{\beta} \mid S, G, Z]/\beta - 1 = \rho(S) - 1$, depends on neither β nor the error variance, so the values $\beta = 1.5$ and unit variance matter only for the standard deviation of $\hat{\beta}$.

The network size is fixed at $n = 1000$, and the budget varies over $10\%, 30\%, 50\%$ of

the network size. For each family, I reuse the 100 calibrated graphs of Section 6.2 and run each design 20 times per graph. Monte Carlo error therefore comes from the graphs and the samples. Table 3 reports the relative bias $\rho(S) - 1$ averaged over samples and graphs, with the standard deviation of $\hat{\beta}$ across samples within a graph, averaged over graphs, in parentheses. With $Z = \mathbf{1}$, the first estimates $\bar{\rho} - 1$ averaged over the 100 graphs. The relative bias is negative in every entry: the plug-in slope is biased toward zero, and the entries below are magnitudes.

On the Barabási–Albert and Chung–Lu families the relative bias shrinks as the budget grows. Under the sequential sampling design it falls from 29 and 27 percent at the 10% budget to 4.5 and 3.2 percent at the 50% budget. These magnitudes of bias are not negligible in the downstream regression. A first-stage similarity of 0.96 corresponds to a slope bias of about four percent under the regression interpretation of (20), and the correspondence is approximate because the fitted intercept absorbs the mean of the plug-in regressor. Simple random sampling is about two to three times as biased in every entry, and the interview saving of the first stage carries over: the sequential sampling design at the 30% budget is less biased than simple random sampling at the 50% budget.

On the Erdős–Rényi and stochastic block model families the attenuation is severe under both designs. The relative bias is near 90 percent at the 10% budget and still above 53 percent at the 50% budget, even though the first-stage similarity on the same graphs reaches 0.96 at the 50% budget. The two findings are consistent. When centrality is nearly constant, the constant vector accounts for almost all of the cosine similarity. The fitted intercept removes exactly this component, so the slope is identified from the small residual variation, and the estimation error is large relative to that variation. This is the failure mode that the identifying-variation assumption rules out asymptotically.

Despite this severe bias toward zero, the homogeneous families report the smallest standard deviations at the 10% budget. This is because shrinking the slope toward zero also shrinks its dispersion. Hence, the small standard deviations signal shrinkage, not statistical precision. The informative comparison is on the families with degree heterogeneity. On the Barabási–Albert and Chung–Lu families at the 50% budget, the relative biases of 4.5 and 3.2 percent correspond to absolute biases of 0.068 and 0.048 at $\beta = 1.5$, and both exceed the standard deviations of 0.045 and 0.043. The plug-in slope is consistent as the network grows at a fixed sampling fraction. In a finite sample, the binding obstacle to inference is the convergence rate of $\hat{v}$, and Section 5 develops the inference theory and proposes a bias correction.

7 Application

This section evaluates the sequential sampling design on real village networks, against simple random sampling (Definition 4). The question is whether a researcher with only the *sampled* network would reach the same estimates as one with the *full* network. A positive answer would mean substantial costs and time savings from data collection for empirical network analysis.

Section 7.1 describes the network data. Section 7.2 describes the regressions, their full-network estimates, and the feasible estimates under sampling. Section 7.3 extends the comparison to other network statistics and alternative centrality measures.

Table 4: Statistics of the Karnataka villages.

Statistic	Mean	SD
Network structure		
Network size n	222	68
Share in the largest connected component	0.981	0.013
Average degree $\bar{d}$	17.4	3.8
Degree standard deviation σ_d	9.1	2.2
Maximum degree $d_{\max}$	53	17
Density	0.086	0.030
Clustering	0.300	0.045
Spectral statistics and degree balance		
Largest eigenvalue λ	22.8	5.0
Relative spectral gap δ	0.327	0.099
Global connectivity γ	0.223	0.099
Degree balance $\bar{d}/d_h$	1.70	0.24
Outcome variables		
Loan nominations		
Share of households nominated	0.048	0.029
Nominations received per household	0.455	0.157
Event nominations		
Share of households nominated	0.039	0.021
Nominations received per household	0.347	0.166

Notes: Entries are means and standard deviations across the 33 villages of village-level statistics. Network size is the number of households on the roster. The network structure statistics are computed on the full roster, with isolated households at degree zero. Clustering is the average local clustering coefficient, zero for households with fewer than two links. The spectral statistics and the degree balance are computed on the largest connected component.

7.1 Village networks

The networks come from the second wave of the Karnataka village surveys, collected in 2012 and released with Banerjee et al. (2019) and Banerjee et al. (2024). The first wave, used in Banerjee et al. (2013), surveyed about 46 percent of the households in each village. The second wave surveyed about 89 percent of the households, and because a link is recorded when either of its two households reports it, the undirected network I use contains about 98.8 percent of all links (Banerjee et al., 2024). I treat each second-wave network, itself a star sample in the sense of Assumption 1, as the full network G.

In 33 of the villages, Banerjee et al. (2019) also ask villagers whom to contact to spread information about an event or a new loan product. Villagers disproportionately name central households, and in bivariate regressions eigenvector centrality predicts nominations essentially as well as diffusion centrality. Central households serve as effective seeds for policy intervention: in a randomized trial across 521 Haryana villages, seeding information about monthly immunization camps with households that villagers nominated raised the number of children immunized each month by 22 percent relative to randomly chosen seeds. Table 4 describes the 33 villages.

Table 5: Second-stage bias in the Karnataka villages.

	Full network	Sequential sampling			Simple random sampling		
Outcome	$\tilde{\beta}$	10%	30%	50%	10%	30%	50%
Loan nominations	0.378	−0.386	−0.110	−0.011	−0.528	−0.254	−0.114
		(0.057)	(0.034)	(0.018)	(0.053)	(0.048)	(0.032)
Event nominations	0.283	−0.419	−0.110	−0.005	−0.522	−0.261	−0.121
		(0.044)	(0.028)	(0.017)	(0.050)	(0.033)	(0.026)

Notes: Entries are computed over the 100 sampling replications per design and budget of Section 7.2. Each entry reports the mean of $(\hat{\beta} - \tilde{\beta})/\tilde{\beta}$ across replications, with the standard deviation of $\hat{\beta}$ across replications in parentheses, where $\tilde{\beta}$ is the full-network slope of (13).

7.2 Replication under sampled networks

The regression exercises have exactly the data structure of Section 4. The outcome Y is the number of nominations each household receives, once for the loan question and once for the event question. The two outcomes measure the diffusion of financial and of social information. The only difference lies in the centrality regressor.

The full-network regressor is the eigenvector centrality v computed from the adjacency matrix A of the full network of Section 7.1. Eliciting A is the costly part of the data collection, and v is available only when G is fully elicited. Banerjee et al. (2019) regress the nomination count on centrality normalized by its standard deviation, pooled across villages with standard errors clustered by village. Table C.1 of their online appendix reports the least-squares slopes, 0.378 on the loan question and 0.283 on the event question, with standard errors of 0.074 and 0.064, on 6,466 households, their estimation sample. I adopt the same specification, normalization, and estimation sample.

With sampled network $\hat{G}$, the true centrality v is unobserved. The feasible regression thus replaces it with the plug-in centrality $\hat{v}$ of Definition 3, computed on the sample adjacency matrix $\hat{A}$. This yields the feasible slope estimate $\hat{\beta}$ of (13). The deviation of $\hat{\beta}$ from the full-network slope is the bias induced by partial network data.

Each village network is restricted to its largest connected component, on which v and $\hat{v}$ are computed. Every household outside this component is isolated. Isolated households have centrality zero by Definition 1 and enter the regressions at zero as well. Table 4 reports the share of households in the largest component. An isolated household can be sampled only by the initial draw in sequential sampling or by simple random sampling. In both cases, such a household is interviewed and reports no link while counting toward the budget. Under the sequential sampling design, the surveyor then redraws the initial household, as in step (i) of Algorithm 3.1. Under either sampling design, the estimated centrality of an isolated household is zero, which is also its true value.

How would the conclusions of Banerjee et al. (2019) have changed under partial network data? Partial sampling cuts the data cost, but the saving is worthwhile only if the bias it induces is small. I re-estimate both regressions at interview budgets of 10, 30, and 50 percent of the households in each village, once under the sequential sampling design and once under simple random sampling, with 100 replications of each design per village and budget. The

largest budget matches the original data collection: Banerjee et al. (2013) interviewed about 46 percent of the households in each village, drawn without adapting to the network. Table 5 reports, under each design at every budget, the relative bias of $\hat{\beta}$ against the full-network slope and the standard deviation of $\hat{\beta}$ across replications.

At the 50 percent budget the sequential sampling design reproduces both full-network slopes nearly without bias. On loan nominations the relative bias is -1.1 percent against -11.4 percent under simple random sampling. On event nominations it is -0.5 percent against -12.1 percent. The interview saving of Section 6.2 reappears on the villages: the sequential sampling design at the 30 percent budget is as accurate as simple random sampling at the 50 percent budget.

The standard deviation of $\hat{\beta}$ falls with the budget under both designs and falls faster under the sequential sampling design. The one exception is the loan regression at the 10 percent budget, where the sequential dispersion is slightly larger. The gap is consistent with Monte Carlo error. The heavier bias toward zero under simple random sampling also compresses its dispersion, as in Section 6.3.

A researcher surveying half of the households in each village under the sequential sampling design would therefore have reproduced the findings of Banerjee et al. (2019) at half the network data cost, while simple random sampling at the same budget would have underestimated both slopes by about 12 percent.

7.3 Other network statistics

The focus of this paper so far has been on eigenvector centrality. I examine whether the sequential sampling design also recovers other commonly used network statistics, for which Theorem 1 provides no guarantee. Table 6 reports the relative bias of five statistics of the sampled villages: the largest eigenvalue, transitivity, the average clustering coefficient, the average path length, and the diameter.

At every budget, the sequential sampling design has a smaller relative bias than simple random sampling on every statistic. The gap is largest for the average clustering coefficient, which the sequential sampling design recovers within 8 percent at every budget, whereas simple random sampling understates it by 46 percent at the 10 percent budget and by 15 percent at the 50 percent budget. The largest eigenvalue and transitivity are underestimated under both designs, by about half at the 10 percent budget and by about a tenth at the 50 percent budget under the sequential sampling design. The largest eigenvalue is underestimated by construction, because $\hat{\lambda} \leq \lambda$ for every sample (Proposition 5). Transitivity, the share of pairs of a household's friends that are also connected, is underestimated because sampling any household reveals its links to both friends but not the link between them. The latter further requires sampling at least one of its two friends.

The applied literature also uses diffusion centrality and betweenness centrality. Diffusion centrality (Section 2) requires two parameters, the diffusion probability and the number of periods of diffusion. Following Banerjee et al. (2019), I set them to $1/\lambda$ and the diameter of the largest connected component. Importantly, the sampled-network diffusion centrality depends on parameters computed from the full network. Betweenness centrality is a non-spectral measure of how often a household lies on the shortest path between two other households. On the sampled network, both regressors are computed from the sample adjacency matrix $\hat{A}$

Table 6: Relative bias of other network statistics.

	Sequential sampling			Simple random sampling		
Statistic	10%	30%	50%	10%	30%	50%
Largest eigenvalue	−0.472	−0.219	−0.085	−0.585	−0.339	−0.177
	(0.062)	(0.046)	(0.027)	(0.047)	(0.040)	(0.029)
Transitivity	−0.526	−0.251	−0.105	−0.752	−0.408	−0.201
	(0.115)	(0.060)	(0.035)	(0.067)	(0.050)	(0.035)
Average clustering coefficient	−0.079	−0.040	−0.036	−0.461	−0.223	−0.146
	(0.166)	(0.080)	(0.045)	(0.168)	(0.076)	(0.045)
Average path length	0.234	0.102	0.037	0.413	0.170	0.070
	(0.077)	(0.033)	(0.015)	(0.082)	(0.032)	(0.014)
Diameter	0.083	0.059	0.021	0.335	0.157	0.070
	(0.148)	(0.128)	(0.118)	(0.191)	(0.143)	(0.113)

Notes: Entries are computed over the 33 villages and the 100 sampling replications per design and budget, as in Table 5. Each entry reports the mean of the sampled-network statistic minus its full-network value, divided by the full-network value, with the SD of the ratio in parentheses. Full-network statistics are computed on the largest connected component (LCC). Sampled-network statistics are computed on all non-isolated households, except that average path length and diameter are computed on the LCC of the sampled network.

and are zero for isolated households.

Table 7: Second-stage bias with other centrality measures.

	Full network	Sequential sampling			Simple random sampling		
Outcome	$\tilde{\beta}$	10%	30%	50%	10%	30%	50%
Panel A. Diffusion centrality							
Loan nominations	0.358	−0.368	−0.106	−0.007	−0.555	−0.285	−0.120
		(0.058)	(0.035)	(0.021)	(0.050)	(0.048)	(0.032)
Event nominations	0.265	−0.389	−0.078	0.021	−0.539	−0.270	−0.110
		(0.045)	(0.030)	(0.018)	(0.048)	(0.035)	(0.027)
Panel B. Betweenness centrality							
Loan nominations	0.455	−0.510	−0.218	−0.057	−0.737	−0.434	−0.217
		(0.085)	(0.060)	(0.036)	(0.072)	(0.082)	(0.063)
Event nominations	0.344	−0.528	−0.201	−0.048	−0.730	−0.436	−0.219
		(0.063)	(0.053)	(0.031)	(0.062)	(0.063)	(0.053)

Notes: As Table 5, with eigenvector centrality replaced by the diffusion centrality of Section 7.3 in Panel A and by betweenness centrality in Panel B.

Table 7 reports the second-stage bias for the estimated slopes on diffusion and betweenness centrality. At every budget, simple random sampling is more biased than the sequential sampling design. Remarkably, at the 50 percent budget, the sequential sampling design reproduces the full-network slopes on diffusion centrality within about 2 percent, almost as well as on eigenvector centrality, whereas simple random sampling understates them by 12 and 11 percent. On betweenness centrality, it reproduces the slopes within 6 percent, whereas

simple random sampling understates them by 22 percent. The sampling exercises suggest that the sequential sampling design is useful beyond eigenvector centrality.

8 Conclusion

Under simple random sampling, the sampled network generally does not inherit the structure of the full network (Stumpf et al., 2005; Lee et al., 2006). This contrasts with sampling a cross section or a panel, where a missing observation affects only the unit to which it belongs. In a network, a missing observation carries an externality of its own. For example, a single missing link or node can bias the centrality of every agent. As a result, sampled networks pose a challenge to empirical analysis and to network-based policy intervention alike.

This paper proposes a sequential sampling design for network data and establishes that it consistently estimates eigenvector centrality from the sampled network. The surveyor moves along links, and every draw comes from the surveyor's own randomization device. For regressions on the estimated centrality, I show that the plug-in slope is consistent and that the robust t-test of the null of no effect is asymptotically exact. In the application to the village networks of Banerjee et al. (2019), a sequential sample of half the households reproduces the full-network slopes within about one percent, a substantial saving in the budget and time of network data collection.

Two directions for future research stand out. First, this paper provides theoretical guarantees for eigenvector centrality under the sequential sampling design. The broader question is which network statistics can be estimated accurately, under which sampling design, with which estimator, and at what cost. Second, this paper studies undirected networks by interpreting any reported link as a potential channel of interaction between two agents. However, certain networks, including buyer-seller networks, are directed by nature. My proof strategy rests on the symmetry of the adjacency matrix A, and the directed case requires new tools rather than a simple modification.

Finally, researchers need not restrict themselves to simple random sampling. Breza et al. (2020) lower the cost of network data by collecting novel aggregate relational data instead of detailed links, recovering a wide class of network statistics from a parametric model of link formation. This paper lowers the cost of estimating eigenvector centrality through the sequential sampling design without imposing a model of link formation. This approach turns partial network observability from a data constraint into a choice of sampling design. Where simple random sampling fails, a design adapted to the network may succeed.

A Proofs

The appendix uses the notation of the main text. Throughout, $n \geq 2$, and e_i and $\mathbf{1}$ denote the ith standard basis vector and the all-ones vector.

A.1 Consistency of the estimated centrality

For a symmetric matrix R, write $\|R\|_{\mathrm{op}}$ for its spectral norm. The proofs of this subsection symmetrize the transition matrix P. The normalized adjacency matrix

$$\mathcal{A} := D^{-1/2} A D^{-1/2} = D^{1/2} P D^{-1/2} \tag{21}$$

is symmetric with nonnegative entries and zero diagonal, and it is similar to P. The two matrices therefore have the same eigenvalues $1 = \lambda_1(P) > \lambda_2(P) \geq \cdots \geq \lambda_n(P) \geq -1$. The first inequality is strict because G is connected. The last holds because P is stochastic, and its spectral radius is therefore at most its maximal row sum, which is one. Since $\lambda_2(P) = 1 - \gamma$, the global connectivity satisfies $\gamma > 0$. With π the stationary distribution of (8), the unit eigenvector of $\mathcal{A}$ for the eigenvalue one is $\varphi := (\sqrt{\pi_1}, \ldots, \sqrt{\pi_n})^\top$, because $D^{-1/2}\varphi = \mathbf{1}/\sqrt{n\bar{d}}$ and $A\mathbf{1} = D\mathbf{1}$ give $\mathcal{A}\varphi = D^{-1/2} D\mathbf{1}/\sqrt{n\bar{d}} = \varphi$, and $\|\varphi\|_2^2 = \sum_i \pi_i = 1$. The bounds of this subsection are stated through the sampling intensity per unit degree,

$$c := \frac{\gamma\alpha}{\bar{d}}. \tag{22}$$

Lemma 1 (Norms of symmetric nonnegative matrices)**.** *Let R be a symmetric matrix with nonnegative entries. Then $\|R\|_{\mathrm{op}} = \lambda_1(R) \leq \max_i \sum_j R_{ij}$.*

Proof. For a unit vector y, write $|y|$ for the vector of absolute values, a unit vector. Since the entries of R are nonnegative, $|y^\top R y| \leq |y|^\top R |y| \leq \lambda_1(R)$, the last step by the Rayleigh quotient of R at $|y|$. Hence every eigenvalue of R lies in $[-\lambda_1(R), \lambda_1(R)]$. Since also $\lambda_1(R) \geq e_1^\top R e_1 = R_{11} \geq 0$, it follows that $\|R\|_{\mathrm{op}} = \max_k |\lambda_k(R)| = \lambda_1(R)$.

For the row-sum bound, let $g \neq 0$ be an eigenvector of R for $\lambda_1(R)$ and let i be an index with $|g_i| = \max_j |g_j| > 0$. The ith coordinate of $Rg = \lambda_1(R) g$ and the triangle inequality give

$$\lambda_1(R)\,|g_i| \;=\; \Big|\sum_j R_{ij} g_j\Big| \;\leq\; \sum_j R_{ij} |g_j| \;\leq\; |g_i| \sum_j R_{ij}.$$

Dividing by $|g_i|$ gives the claim. □

Lemma 2 (Set avoidance)**.** *Let $J \subseteq V$ be nonempty with $\bar{J} := V \setminus J \neq \emptyset$. Under sequential sampling,*

$$\mathbb{P}\big(J \subseteq \bar{S}\big) \;\leq\; \Big(\frac{\bar{d}}{d_h}\Big)^{1/2} \exp\Big\{1 - c \sum_{j \in J} d_j\Big\}. \tag{23}$$

In particular $\mathbb{P}(j \in \bar{S}) \leq (\bar{d}/d_h)^{1/2} e^{1 - c d_j}$ for every node j.

Proof. Write $\pi(J) := \sum_{j \in J} \pi_j$ and $\pi(\bar{J}) := 1 - \pi(J)$. The proof establishes, for every $t \geq 0$,

$$\mathbb{P}(H_0 \in \bar{J}, \ldots, H_t \in \bar{J}) \;\leq\; \Big(\frac{\bar{d}}{d_h}\Big)^{1/2} \pi(\bar{J})^{1/2} \big(1 - \gamma\pi(J)\big)^t,$$

and then takes $t = \lceil \alpha n \rceil - 1$.

The bound starts from the walk killed on entering J. Let $P_{\bar{J}\bar{J}}$, $\mathcal{A}_{\bar{J}\bar{J}}$, and $D_{\bar{J}\bar{J}}$ be the principal submatrices of P, $\mathcal{A}$, and D indexed by $\bar{J}$, and let $\mathbf{1}_{\bar{J}} \in \mathbb{R}^{\bar{J}}$ be the all-ones vector.

Fix $t \geq 0$. Then $\mathcal{A}_{\bar{J}\bar{J}} = D_{\bar{J}\bar{J}}^{1/2} P_{\bar{J}\bar{J}} D_{\bar{J}\bar{J}}^{-1/2}$ and $P_{\bar{J}\bar{J}}^t = D_{\bar{J}\bar{J}}^{-1/2} \mathcal{A}_{\bar{J}\bar{J}}^t D_{\bar{J}\bar{J}}^{1/2}$. Since H_0 is uniform on V, summing the path probabilities of the walk over the paths that stay in $\bar{J}$ gives

$$\begin{aligned}\mathbb{P}(H_0 \in \bar{J}, \dots, H_t \in \bar{J}) &= \frac{1}{n} \sum_{i_0, \dots, i_t \in \bar{J}} P_{i_0 i_1} \cdots P_{i_{t-1} i_t} = \frac{1}{n} \mathbf{1}_{\bar{J}}^\top P_{\bar{J}\bar{J}}^t \mathbf{1}_{\bar{J}} \\ &= \frac{1}{n} \big(D_{\bar{J}\bar{J}}^{-1/2} \mathbf{1}_{\bar{J}}\big)^\top \mathcal{A}_{\bar{J}\bar{J}}^t \big(D_{\bar{J}\bar{J}}^{1/2} \mathbf{1}_{\bar{J}}\big).\end{aligned}$$

The two vectors in the last expression have squared norms

$$\big\|D_{\bar{J}\bar{J}}^{-1/2} \mathbf{1}_{\bar{J}}\big\|_2^2 = \sum_{i \in \bar{J}} d_i^{-1} \leq \sum_{i \in V} d_i^{-1} = \frac{n}{d_h}, \qquad \big\|D_{\bar{J}\bar{J}}^{1/2} \mathbf{1}_{\bar{J}}\big\|_2^2 = \sum_{i \in \bar{J}} d_i = n \bar{d}\, \pi(\bar{J}),$$

the first by the definition of the harmonic mean degree d_h in Section 3.5 and the second by (8). The Cauchy–Schwarz inequality and $\|\mathcal{A}_{\bar{J}\bar{J}}^t\|_{\mathrm{op}} \leq \|\mathcal{A}_{\bar{J}\bar{J}}\|_{\mathrm{op}}^t$ therefore give

$$\mathbb{P}(H_0 \in \bar{J}, \dots, H_t \in \bar{J}) \;\leq\; \Big(\frac{\bar{d}}{d_h}\Big)^{1/2} \pi(\bar{J})^{1/2}\, \|\mathcal{A}_{\bar{J}\bar{J}}\|_{\mathrm{op}}^t. \tag{24}$$

Inequality (24) adapts the stationary-start bound (3.68) of Aldous and Fill (2002, Chapter 3, Proposition 3.21) to discrete time and to the uniform start.

It remains to bound the largest eigenvalue of the killed walk. The matrix $\mathcal{A}_{\bar{J}\bar{J}}$ is symmetric with nonnegative entries, and Lemma 1 gives $\|\mathcal{A}_{\bar{J}\bar{J}}\|_{\mathrm{op}} = \lambda_1(\mathcal{A}_{\bar{J}\bar{J}})$. Let $y \in \mathbb{R}^V$ be a unit vector supported on $\bar{J}$ and let $y_{\bar{J}} \in \mathbb{R}^{\bar{J}}$ collect its coordinates in $\bar{J}$, so that $y^\top \mathcal{A} y = y_{\bar{J}}^\top \mathcal{A}_{\bar{J}\bar{J}}\, y_{\bar{J}}$, and every unit vector of $\mathbb{R}^{\bar{J}}$ arises as such a $y_{\bar{J}}$. Expand y in an orthonormal eigenbasis $\varphi_1 = \varphi, \varphi_2, \dots, \varphi_n$ of $\mathcal{A}$ with coefficients $\zeta_k := \langle y, \varphi_k \rangle$. Since $\lambda_k(P) \leq \lambda_2(P) = 1 - \gamma$ for $k \geq 2$ and the weights ζ_k^2 are nonnegative, the sign of $1 - \gamma$ playing no role,

$$y^\top \mathcal{A} y = \sum_k \lambda_k(P)\, \zeta_k^2 \leq \zeta_1^2 + (1 - \gamma)(1 - \zeta_1^2) = (1 - \gamma) + \gamma \zeta_1^2.$$

By the Cauchy–Schwarz inequality, the first coefficient satisfies

$$|\zeta_1| = \Big|\sum_{i \in \bar{J}} y_i \sqrt{\pi_i}\Big| \leq \Big(\sum_{i \in \bar{J}} \pi_i\Big)^{1/2} \|y\|_2 = (1 - \pi(J))^{1/2}.$$

Hence $\zeta_1^2 \leq 1 - \pi(J)$. Since $\gamma > 0$, the two displays give

$$\lambda_1(\mathcal{A}_{\bar{J}\bar{J}}) = \max_{\|y_{\bar{J}}\|_2 = 1} y_{\bar{J}}^\top \mathcal{A}_{\bar{J}\bar{J}}\, y_{\bar{J}} \;\leq\; (1 - \gamma) + \gamma(1 - \pi(J)) = 1 - \gamma\pi(J). \tag{25}$$

The base is nonnegative, because for any $i \in \bar{J}$ the zero diagonal of $\mathcal{A}$ gives

$$0 = e_i^\top \mathcal{A} e_i \leq \lambda_1(\mathcal{A}_{\bar{J}\bar{J}}) \leq 1 - \gamma\pi(J).$$

Inequality (25) is the eigenvalue form of Aldous and Fill (2002, Chapter 3, Corollary 3.34), through their Theorem 3.33, with relaxation time $1/\gamma$. The argument above does not need the walk restricted to $\bar{J}$ to be irreducible, which their quasistationary argument assumes. Together, (24) and (25) give the bound stated at the start of the proof.

Now take $t = \lceil \alpha n \rceil - 1$. The distinct nodes among $H_0, \dots, H_t$ number at most $t+1 = \lceil \alpha n \rceil$ and are the first distinct nodes of the sequence $H_0, H_1, \dots$. Hence, S contains $H_0, \dots, H_t$ by Definition 5, and if $J \subseteq \bar{S}$, none of these nodes lies in J. Write $x := \gamma\pi(J)$, which lies in $[0,1]$ because $\gamma > 0$ and the base $1 - x$ is nonnegative. Then

$$\mathbb{P}\big(J \subseteq \bar{S}\big) \;\le\; \mathbb{P}(H_0 \in \bar{J}, \dots, H_t \in \bar{J}) \;\le\; \Big(\frac{\bar{d}}{d_h}\Big)^{1/2} \pi(\bar{J})^{1/2}(1-x)^t \;\le\; \Big(\frac{\bar{d}}{d_h}\Big)^{1/2} e^{1-x\alpha n}.$$

The last inequality uses $\pi(\bar{J})^{1/2} \le 1$, $1 - x \le e^{-x}$, $t \ge \alpha n - 1$ because $\lceil \alpha n \rceil \ge \alpha n$, and $x \le 1$. Since $x\alpha n = \gamma\alpha \sum_{j\in J} d_j/\bar{d} = c\sum_{j\in J} d_j$ by (22), this is (23). The last claim is the case $J = \{j\}$, for which $\bar{J} \neq \emptyset$ because $n \ge 2$. □

Lemma 3 (Second moment of the perturbation)**.** *Let v be the unit Perron vector of A and $E = A - \hat{A}$ the perturbation of (7). Under sequential sampling,*

$$\mathbb{E}\|Ev\|_2^2 \;\le\; \frac{1}{e}\Big(\frac{\bar{d}}{d_h}\Big)^{1/2}\Big(\frac{\bar{d}}{\gamma\alpha}\Big)^2. \tag{26}$$

Proof. Since the perturbation E has support $\bar{S} \times \bar{S}$ with entries A_{ik}, $(Ev)_i = \mathbf{1}\{i \in \bar{S}\}\sum_{k\sim i} \mathbf{1}\{k \in \bar{S}\}\, v_k$, where $k \sim i$ denotes $A_{ik} = 1$. The Cauchy–Schwarz inequality over the d_i neighbors of i gives

$$(Ev)_i^2 \;\le\; d_i \sum_{k\sim i} v_k^2\, \mathbf{1}\{i \in \bar{S},\; k \in \bar{S}\}.$$

For adjacent i and k, put $J := \{i, k\}$. If $J = V$, then $\mathbb{P}(J \subseteq \bar{S}) = 0$ because $H_0 \in S$. Otherwise $\bar{J} \neq \emptyset$, and Lemma 2 gives $\mathbb{P}(J \subseteq \bar{S}) \le e\,(\bar{d}/d_h)^{1/2} e^{-c(d_i+d_k)}$. Taking expectations, applying this bound, and exchanging the order of summation,

$$\begin{aligned}\mathbb{E}\|Ev\|_2^2 &\le e\Big(\frac{\bar{d}}{d_h}\Big)^{1/2} \sum_k e^{-cd_k} v_k^2 \sum_{i\sim k} d_i e^{-cd_i} \;\le\; e\Big(\frac{\bar{d}}{d_h}\Big)^{1/2} \frac{1}{ce} \sum_k d_k e^{-cd_k} v_k^2 \\ &\le e\Big(\frac{\bar{d}}{d_h}\Big)^{1/2} \frac{1}{(ce)^2} \sum_k v_k^2 \;=\; \frac{1}{e}\Big(\frac{\bar{d}}{d_h}\Big)^{1/2}\Big(\frac{\bar{d}}{\gamma\alpha}\Big)^2.\end{aligned}$$

The second and third inequalities apply $\sup_{x\ge 0} xe^{-cx} = 1/(ce)$, first to each of the d_k summands $d_i e^{-cd_i}$ with $i \sim k$ and then to $d_k e^{-cd_k}$. The equality uses $\sum_k v_k^2 = 1$ and $c = \gamma\alpha/\bar{d}$ from (22). □

For a threshold $\tau > 0$, the heavy-tail sum is

$$\kappa_\tau := \sum_{i\in V:\; d_i > \tau} e^{-cd_i}. \tag{27}$$

Lemma 4 (Spectral separation)**.** *For $\tau > 0$ let $\mathcal{B}_\tau := \{\|E\|_{\mathrm{op}} \le \tau\}$, and write $\hat{\lambda}_2 := \lambda_2(\hat{A})$. Under sequential sampling,*

$$\mathbb{P}(\mathcal{B}_\tau^c) \;\le\; e\Big(\frac{\bar{d}}{d_h}\Big)^{1/2} \kappa_\tau, \tag{28}$$

and on $\mathcal{B}_\tau$,

$$\lambda - \hat{\lambda}_2 \;\ge\; \delta\lambda - \tau. \tag{29}$$

The event $\mathcal{B}_\tau$ contains the event that no unsampled node has more than τ unsampled neighbors.

Proof. Weyl's inequality for $\hat{A} = A + (-E)$ gives $\hat{\lambda}_2 \le \lambda_2(A) + \lambda_1(-E) \le \lambda_2(A) + \|E\|_{\mathrm{op}}$. By the definition of the relative spectral gap δ, $\lambda_2(A) = (1-\delta)\lambda$. Hence, on $\mathcal{B}_\tau$, $\lambda - \hat{\lambda}_2 \ge \delta\lambda - \|E\|_{\mathrm{op}} \ge \delta\lambda - \tau$, which is (29).

The other two claims follow from a row-sum bound on E. By (7), E equals A on the unobserved block $\bar{S} \times \bar{S}$ and is zero elsewhere. Hence E is symmetric with nonnegative entries, and its ith row sum is $\mathbf{1}\{i \in \bar{S}\}\, d_i^{\bar{S}}$, where $d_i^{\bar{S}} := |N(i) \cap \bar{S}|$ is the number of unsampled neighbors of i. Lemma 1 gives $\|E\|_{\mathrm{op}} \le \max_{i \in V} \mathbf{1}\{i \in \bar{S}\}\, d_i^{\bar{S}}$. Thus $\mathcal{B}_\tau$ contains the event that $d_i^{\bar{S}} \le \tau$ for every $i \in \bar{S}$, which is the last claim.

On $\mathcal{B}_\tau^c$, this bound gives an $i \in \bar{S}$ with $d_i \ge d_i^{\bar{S}} > \tau$. Therefore

$$\mathcal{B}_\tau^c \;\subseteq \bigcup_{i \in V:\ d_i > \tau} \{i \in \bar{S}\}.$$

The union bound and the single-node case of Lemma 2 give $\mathbb{P}(\mathcal{B}_\tau^c) \le \sum_{i:\ d_i > \tau} (\bar{d}/d_h)^{1/2} e^{1-cd_i}$. By (27), the right side equals $e\,(\bar{d}/d_h)^{1/2}\kappa_\tau$, which is (28). □

Lemma 5 (Perturbation bound)**.** *Let A and $\hat{A}$ be symmetric $n \times n$ matrices and $E := A - \hat{A}$. Let (λ, v) be an eigenpair of A with $\|v\|_2 = 1$, let $\hat{v}$ be a unit eigenvector of $\hat{A}$ for its largest eigenvalue, and write $\angle(v, \hat{v}) \in [0, \pi/2]$ for the angle between the lines spanned by v and $\hat{v}$, so that $\cos\angle(v, \hat{v}) = |\langle v, \hat{v}\rangle|$. If $\lambda > \lambda_2(\hat{A})$, then*

$$\sin\angle(v, \hat{v}) \;\le\; \frac{\|Ev\|_2}{\lambda - \lambda_2(\hat{A})}. \tag{30}$$

If moreover $\langle v, \hat{v}\rangle \ge 0$, then $\|\hat{v} - v\|_2 \le \sqrt{2}\,\sin\angle(v, \hat{v})$.

Proof. Let $\hat{u}_1 = \hat{v}, \hat{u}_2, \ldots, \hat{u}_n$ be an orthonormal eigenbasis of $\hat{A}$ with eigenvalues $\lambda_1(\hat{A}) \ge \lambda_2(\hat{A}) \ge \cdots \ge \lambda_n(\hat{A})$. Such a basis exists because $\hat{v}$ is an eigenvector for $\lambda_1(\hat{A})$ and the orthogonal complement of $\hat{v}$ is invariant under $\hat{A}$, and the eigenvalues of $\hat{A}$ on this complement are $\lambda_2(\hat{A}), \ldots, \lambda_n(\hat{A})$ also when $\lambda_1(\hat{A})$ is not simple. Since $\hat{A}v = Av - Ev = \lambda v - Ev$,

$$\|Ev\|_2^2 = \|(\hat{A} - \lambda I)v\|_2^2 = \sum_{i=1}^{n} \big(\lambda_i(\hat{A}) - \lambda\big)^2 \langle \hat{u}_i, v\rangle^2 \;\ge\; \big(\lambda - \lambda_2(\hat{A})\big)^2 \sum_{i \ge 2} \langle \hat{u}_i, v\rangle^2,$$

because the $i = 1$ term is nonnegative and $\lambda - \lambda_i(\hat{A}) \ge \lambda - \lambda_2(\hat{A}) > 0$ for $i \ge 2$ by the hypothesis $\lambda > \lambda_2(\hat{A})$.

For the angle, the $\hat{u}_i$ form an orthonormal basis and $\|v\|_2 = 1$, so

$$\sum_{i \ge 2} \langle \hat{u}_i, v\rangle^2 = 1 - \langle \hat{v}, v\rangle^2 = 1 - \cos^2\angle(v, \hat{v}) = \sin^2\angle(v, \hat{v}).$$

Dividing the previous display by $(\lambda - \lambda_2(\hat{A}))^2 > 0$ and taking square roots gives (30).

For the distance, if $\langle v, \hat{v}\rangle \ge 0$, then $\cos\angle(v, \hat{v}) = \langle v, \hat{v}\rangle$ and $\|\hat{v} - v\|_2^2 = 2 - 2\langle v, \hat{v}\rangle = 2(1 - \cos\angle(v, \hat{v}))$. For $\psi \in [0, \pi/2]$, $1 - \cos\psi = \sin^2\psi/(1 + \cos\psi) \le \sin^2\psi$. Hence $\|\hat{v} - v\|_2^2 \le 2\sin^2\angle(v, \hat{v})$. □

Since λ is simple (Section 2), $\lambda_2(A) < \lambda$. Moreover, $\lambda > 0$, because $\lambda v_1 = \sum_j A_{1j} v_j > 0$ by $v > 0$ and $d_1 \geq 1$. Hence $\delta > 0$, and the interval $(0, \delta\lambda)$ in the next lemma is nonempty.

Lemma 6 (Tail bound)**.** *Under sequential sampling, for every $\tau \in (0, \delta\lambda)$ and every $x > 0$,*

$$\mathbb{P}\big(\sin \angle(v, \hat{v}) > x\big) \;\leq\; \Big(\frac{\bar{d}}{d_h}\Big)^{1/2} \Big[e\,\kappa_\tau + \frac{r_n^2}{e\,x^2\,(1 - \tau/(\delta\lambda))^2} \Big]. \tag{31}$$

Proof. On $\mathcal{B}_\tau$, Lemma 4 gives $\lambda - \hat{\lambda}_2 \geq \delta\lambda - \tau > 0$. The hypothesis of Lemma 5 therefore holds for the symmetric matrices A and $\hat{A}$, the Perron pair (λ, v), and the unit eigenvector $\hat{v}$ of $\hat{A}$ for $\hat{\lambda} = \lambda_1(\hat{A})$. Its conclusion (30) and (29) give $\sin \angle(v, \hat{v}) \leq \|Ev\|_2 / (\lambda - \hat{\lambda}_2) \leq \|Ev\|_2 / (\delta\lambda - \tau)$. Hence, pointwise,

$$\sin^2 \angle(v, \hat{v})\, \mathbf{1}\{\mathcal{B}_\tau\} \;\leq\; \frac{\|Ev\|_2^2}{(\delta\lambda - \tau)^2},$$

the left side vanishing off $\mathcal{B}_\tau$. Taking expectations and applying (26) of Lemma 3,

$$\mathbb{E}\big[\sin^2 \angle(v, \hat{v})\, \mathbf{1}\{\mathcal{B}_\tau\}\big] \;\leq\; \frac{1}{e}\Big(\frac{\bar{d}}{d_h}\Big)^{1/2} \Big(\frac{\bar{d}}{\gamma\alpha(\delta\lambda - \tau)}\Big)^2 = \Big(\frac{\bar{d}}{d_h}\Big)^{1/2} \frac{r_n^2}{e\,(1 - \tau/(\delta\lambda))^2},$$

the equality by the definition $r_n = \bar{d}/(\alpha\gamma\delta\lambda)$ in (9).

Since $\sin \angle(v, \hat{v}) \geq 0$, $\{\sin \angle(v, \hat{v}) > x\} \cap \mathcal{B}_\tau = \{\sin^2 \angle(v, \hat{v})\, \mathbf{1}\{\mathcal{B}_\tau\} > x^2\}$. Markov's inequality bounds the probability of this event by the right side of the previous display divided by x^2. Adding $\mathbb{P}(\mathcal{B}_\tau^c)$ from (28) gives (31). □

Proof of Theorem 1 (Consistency and rate of convergence). Let Θ be the constant of part (ii) of Assumption 2, and set $\tau_n := 2c^{-1} \log n = 2\bar{d} \log n / (\gamma\alpha)$. Each of the at most n summands of κ_{τ_n} in (27) is below $e^{-c\tau_n}$, so $\kappa_{\tau_n} \leq n e^{-c\tau_n} = n^{-1}$. Moreover, $\tau_n/(\delta\lambda) = 2 r_n \log n \to 0$ by part (i) of Assumption 2 and (9). Hence $\tau_n \in (0, \delta\lambda)$ and $1 - \tau_n/(\delta\lambda) \geq 1/2$ for all large n.

Fix $K > 0$ and put $x_n := K r_n / (e(1 - \tau_n/(\delta\lambda)))$. Lemma 6 with $\tau = \tau_n$ and $x = x_n$, together with $(\bar{d}/d_h)^{1/2} \leq \Theta^{1/2}$, gives

$$\mathbb{P}\big(\sin \angle(v, \hat{v}) > x_n\big) \;\leq\; e\,\Theta^{1/2}\big(n^{-1} + K^{-2}\big)$$

for all large n. Since $x_n \leq 2Kr_n/e$ for all large n, $\limsup_n \mathbb{P}(\sin \angle(v, \hat{v}) > 2Kr_n/e) \leq e\,\Theta^{1/2} K^{-2}$, which tends to zero as $K \to \infty$. That is, $\sin \angle(v, \hat{v}) = O_p(r_n)$.

Both v and $\hat{v}$ are nonnegative, so $\langle v, \hat{v} \rangle \geq 0$, and the second clause of Lemma 5 gives $\|\hat{v} - v\|_2 \leq \sqrt{2} \sin \angle(v, \hat{v}) = O_p(r_n)$, which is (10). Since $r_n \log n \to 0$ implies $r_n \to 0$, $\hat{v}$ is consistent. □

A.2 Regression and inference

In the proofs for Sections 4 and 5, v and $\hat{v}$ denote $\sqrt{n}$ times the unit vectors of Definitions 1 and 3, so that $\|v\|_2 = \|\hat{v}\|_2 = \sqrt{n}$, and the measurement error is $u = \hat{v} - v$ of Section 5. Under this normalization Theorem 1 reads

$$\|u\|_2 \;=\; O_p(\sqrt{n}\, r_n),$$

with $r_n \to 0$, because $r_n \log n \to 0$. With M_Z as in Assumption 4 and P_Z as in Section 5, write P_X, M_X and $P_{\hat{X}}, M_{\hat{X}}$ for the corresponding projections for the full-network and feasible regressor matrices X and $\hat{X}$ of (12). The feasible residualized regressor is $\hat{w} = M_Z \hat{v}$ of Section 5, a function of S, and its full-network counterpart is

$$\tilde{w} \;:=\; M_Z v,$$

which is nonrandom. Throughout, a hat marks a feasible object and a tilde marks its full-network counterpart. The two regressors differ by $\hat{w} - \tilde{w} = M_Z u$. Since M_Z is a contraction,

$$\|\hat{w} - \tilde{w}\|_2 \;\le\; \|u\|_2.$$

The residual variation in Assumption 4 is $\mu_n := n^{-1} v^\top M_Z v = n^{-1}\|\tilde{w}\|_2^2$, and $\mu_0 > 0$ is the constant of Assumption 4, so that $\mu_n \ge \mu_0$ for all large n by (14). Only such n are considered below, except in the proof of (15) and part (i) of Proposition 2, which uses only part (i) of Assumption 4 and holds for every n. For such n, $\tilde{w} \neq 0$ and $v \notin \mathrm{col}(Z)$. Since Z has full column rank by part (i) of Assumption 4, $X = [v\ Z]$ has full column rank, $X^\top X$ is nonsingular, and P_Z and P_X are defined. The full-network and feasible leverages are

$$h_{ii} := X_i^\top (X^\top X)^{-1} X_i, \qquad \hat{h}_{ii} := \hat{X}_i^\top (\hat{X}^\top \hat{X})^{-1} \hat{X}_i, \qquad h_n = \max_i h_{ii}, \qquad \hat{h}_n := \max_i \hat{h}_{ii},$$

where h_n is the maximum leverage of Assumption 6, and $\hat{h}_{ii}$ is defined on the event $\{\hat{w} \neq 0\}$, on which $\hat{X}$ has full column rank. The error variances are $\sigma_i^2 = \mathrm{Var}(\epsilon_i)$, collected in $\Sigma := \mathrm{diag}(\sigma_1^2, \ldots, \sigma_n^2)$, which is the covariance matrix of ϵ because the errors are uncorrelated by part (iii) of Assumption 3. For a nonzero vector $w \in \mathbb{R}^n$, write

$$m(w) := \max_{i \in V} \frac{w_i^2}{\|w\|_2^2}, \qquad s(w)^2 := \sum_{i \in V} w_i^2 \, \sigma_i^2,$$

so that $\|w\|_\infty^2 = m(w)\,\|w\|_2^2$, $s(w)^2 \le \bar{\sigma}^2 \|w\|_2^2$ under part (iii) of Assumption 3, and $s(w)^2 \ge \underline{\sigma}^2 \|w\|_2^2$ under part (v) of Assumption 5.

The feasible coefficients and every quantity computed from the feasible regression are defined on the event $\{\hat{w} \neq 0\}$, which is S-measurable and has probability tending to one by Lemma 10(a). Statements about them are on this event, with the conditioning left implicit, as in Sections 4 and 5. The t-test of Proposition 3 rejects $\mathcal{H}_0$ when $|\hat{\beta}| > z\,\hat{\omega}$ for a critical value $z > 0$, which is the event $\{|t| > z\}$ whenever $\hat{\omega} > 0$. Set $t := 0$ when $\hat{\omega} = 0$. All probability statements in this subsection are conditional on (G, Z), as in Sections 4 and 5. The conditional expectation $\mathbb{E}[\,\cdot \mid G, Z]$ is abbreviated to $\mathbb{E}[\,\cdot\,]$ and $\mathbb{E}[\,\cdot \mid S, G, Z]$ to $\mathbb{E}[\,\cdot \mid S]$, and likewise for $\mathbb{P}$, Var, and F_U. Because S takes finitely many values, every conditional expectation given S that appears below exists whenever the corresponding moment of ϵ does. For a real random variable U, F_U denotes its distribution function and $F_U(\cdot \mid S)$ its conditional distribution function given S. For distribution functions F_1 and F_2, $\|F_1 - F_2\|_\infty := \sup_{x \in \mathbb{R}} |F_1(x) - F_2(x)|$ is the Kolmogorov distance. Convergence in distribution to $N(0,1)$ means $\|F_{U_n} - \Phi\|_\infty \to 0$, where Φ is the standard normal distribution function.

Lemma 7 (The errors given the sample)**.** *Let part (i) of Assumption 3 hold. Conditional on (G, Z), the sample S has the distribution of Definition 5 and is independent of ϵ, so the*

conditional distribution of ϵ given (S, G, Z) is its distribution given (G, Z). In particular, whenever parts (ii) and (iii) of Assumption 3 or Assumption 5 are imposed, they hold with $\mathbb{E}[\cdot \mid S, G, Z]$ in place of $\mathbb{E}[\cdot \mid G, Z]$.

Proof. Because G is deterministic, part (i) of Assumption 3 says that S is independent of the pair (ϵ, Z). Hence S is independent of Z, and its distribution given Z is its unconditional distribution, which is that of Definition 5 by construction of the design.

For the conditional independence, let f_1, f_2, f_3 be bounded measurable functions. The independence of S and (ϵ, Z), the tower property, and the independence of S and Z give

$$\mathbb{E}\big[f_1(S)\, f_2(\epsilon)\, f_3(Z)\big] \;=\; \mathbb{E}[f_1(S)]\; \mathbb{E}\big[f_3(Z)\, \mathbb{E}[f_2(\epsilon) \mid Z]\big] \;=\; \mathbb{E}\big[f_1(S)\, f_3(Z)\, \mathbb{E}[f_2(\epsilon) \mid Z]\big],$$

which says that S and ϵ are independent given Z. Since S takes finitely many values, this means $\mathbb{P}(\epsilon \in \Gamma \mid S = S_0) = \mathbb{P}(\epsilon \in \Gamma)$ for every Borel set $\Gamma \subseteq \mathbb{R}^n$ and every S_0 with $\mathbb{P}(S = S_0) > 0$, all probabilities being conditional on (G, Z) under the abbreviation above. The last claim follows because parts (ii) and (iii) of Assumption 3 and Assumption 5 are statements about the distribution of ϵ given (G, Z). □

Lemma 8 (Projections, leverage, and the shared norm)**.** *Parts (a) and (b) are stated on the event $\{\hat{w} \neq 0\}$. Parts (c) and (d) hold on every realization.*

(a) The projections decompose as

$$P_X \;=\; P_Z + \frac{\tilde{w}\tilde{w}^\top}{\|\tilde{w}\|_2^2}, \qquad P_{\hat{X}} \;=\; P_Z + \frac{\hat{w}\hat{w}^\top}{\|\hat{w}\|_2^2}, \tag{32}$$

so that the leverages decompose as

$$h_{ii} \;=\; (P_Z)_{ii} + \frac{\tilde{w}_i^2}{\|\tilde{w}\|_2^2}, \qquad \hat{h}_{ii} \;=\; (P_Z)_{ii} + \frac{\hat{w}_i^2}{\|\hat{w}\|_2^2}, \tag{33}$$

and in particular $\tilde{w}_i^2/\|\tilde{w}\|_2^2 \le h_{ii}$ and $\hat{w}_i^2/\|\hat{w}\|_2^2 \le \hat{h}_{ii}$ for every i.

(b) (Frisch–Waugh–Lovell.) For every $Y \in \mathbb{R}^n$, the first coordinate of $(\hat{X}^\top \hat{X})^{-1}\hat{X}^\top Y$ is $\hat{w}^\top Y/\|\hat{w}\|_2^2$. That is, the first row of $(\hat{X}^\top \hat{X})^{-1}\hat{X}^\top$ is

$$e_1^\top (\hat{X}^\top \hat{X})^{-1}\hat{X}^\top \;=\; \frac{\hat{w}^\top}{\|\hat{w}\|_2^2}. \tag{34}$$

The same holds for X with $\tilde{w}$ in place of $\hat{w}$.

(c) (Shared norm.) Since $\|v\|_2 = \|\hat{v}\|_2 = \sqrt{n}$,

$$\hat{v}^\top u \;=\; \tfrac{1}{2}\|u\|_2^2, \qquad v^\top u \;=\; -\tfrac{1}{2}\|u\|_2^2, \qquad \hat{v}^\top v \;=\; n - \tfrac{1}{2}\|u\|_2^2. \tag{35}$$

(d) (Norm bounds.) $\|\tilde{w}\|_2 \le \sqrt{n}$, $\|\hat{w}\|_2 \le \sqrt{n}$, and

$$\|\tilde{w}\|_2^2 \;=\; v^\top M_Z v \;=\; n\,\mu_n. \tag{36}$$

Proof. (a) Since $\hat{v} = P_Z\hat{v} + \hat{w}$ with $\hat{w} \perp \mathrm{col}(Z)$, the column space of $\hat{X} = [\hat{v}\ Z]$ is the orthogonal direct sum $\mathrm{col}(Z) \oplus \mathrm{span}(\hat{w})$, and the orthogonal projection onto an orthogonal direct sum is the sum of the projections onto the summands. This is the second identity in (32), and the first is the same argument for X with $\tilde{w} \neq 0$. Taking the ith diagonal entry gives (33), and the inequalities follow because $(P_Z)_{ii} = \|P_Z e_i\|_2^2 \geq 0$.

(b) Write $(\hat{\beta}, \hat{\theta}^\top)^\top := (\hat{X}^\top\hat{X})^{-1}\hat{X}^\top Y$, so the fitted vector is $P_{\hat{X}}Y = \hat{\beta}\hat{v} + Z\hat{\theta}$. Applying M_Z to both sides and using $M_Z Z = 0$ and $M_Z\hat{v} = \hat{w}$ gives $M_Z P_{\hat{X}} Y = \hat{\beta}\,\hat{w}$. By (32), $M_Z P_Z = 0$, and $M_Z\hat{w} = \hat{w}$, $M_Z P_{\hat{X}} = \hat{w}\hat{w}^\top/\|\hat{w}\|_2^2$. Hence $\hat{\beta}\,\hat{w} = \hat{w}\,(\hat{w}^\top Y)/\|\hat{w}\|_2^2$, and $\hat{w} \neq 0$ gives $\hat{\beta} = \hat{w}^\top Y/\|\hat{w}\|_2^2$. Since this holds for every Y, (34) follows.

(c) Expanding the squared norm, $\|u\|_2^2 = \|\hat{v}\|_2^2 - 2\hat{v}^\top v + \|v\|_2^2 = 2n - 2\hat{v}^\top v$, which is the third identity. Then $\hat{v}^\top u = \hat{v}^\top\hat{v} - \hat{v}^\top v = n - (n - \frac{1}{2}\|u\|_2^2)$ and $v^\top u = \hat{v}^\top v - v^\top v = -\frac{1}{2}\|u\|_2^2$.

(d) Projections are contractions, so $\|M_Z v\|_2 \leq \|v\|_2 = \sqrt{n}$ and likewise for $\hat{v}$. Since M_Z is symmetric and idempotent, $\|\tilde{w}\|_2^2 = v^\top M_Z^\top M_Z v = v^\top M_Z v$, which is $n\mu_n$ by the definition of μ_n. □

Lemma 9 (Reliability identity)**.** *On* $\{\hat{w} \neq 0\}$,

$$\big(\rho(S) - 1\big)\,\|\hat{w}\|_2^2 \;=\; -\,\hat{w}^\top u \;=\; \hat{v}^\top P_Z u \;-\; \tfrac{1}{2}\|u\|_2^2. \tag{37}$$

Proof. Because M_Z is symmetric and idempotent, $\hat{w}^\top\hat{v} = \|\hat{w}\|_2^2$ and $\hat{w}^\top u = \hat{v}^\top M_Z u$. Substituting $v = \hat{v} - u$ into $\rho(S) = \hat{w}^\top v/\|\hat{w}\|_2^2$ gives $(\rho(S) - 1)\,\|\hat{w}\|_2^2 = -\hat{w}^\top u$. The second equality is $\hat{w}^\top u = \hat{v}^\top u - \hat{v}^\top P_Z u = \frac{1}{2}\|u\|_2^2 - \hat{v}^\top P_Z u$, by $M_Z = I - P_Z$ and (35). □

Proof of (15) and part (i) of Proposition 2 (Bias decomposition). Because M_Z is symmetric and idempotent, $\hat{w}^\top = \hat{v}^\top M_Z$ and $\|\hat{w}\|_2^2 = \hat{v}^\top M_Z\hat{v}$. Hence the expression $\hat{\beta} = \hat{w}^\top Y/\|\hat{w}\|_2^2$ of Section 5.1 is (34) applied to Y. On $\{\hat{w} \neq 0\}$, substituting the model (11) into $\hat{w}^\top Y$ and using $M_Z Z = 0$ gives $\hat{w}^\top Y = \beta\,\hat{w}^\top v + \hat{w}^\top\epsilon$. Dividing by $\|\hat{w}\|_2^2$ gives (15).

For part (i), the event $\{\hat{w} \neq 0\}$ is S-measurable. On it, Lemma 7 and part (ii) of Assumption 3 give

$$\mathbb{E}[\hat{w}^\top\epsilon \mid S] = \hat{w}^\top\mathbb{E}[\epsilon \mid S] = \hat{w}^\top\mathbb{E}[\epsilon] = 0.$$

Hence $\mathbb{E}[e(S) \mid S] = 0$ and, by (15), $\mathbb{E}[\hat{\beta} \mid S] = \rho(S)\beta$. These expectations exist because S takes finitely many values and $\mathbb{E}|\epsilon_i| \leq (\mathbb{E}\epsilon_i^2)^{1/2} \leq \bar{\sigma}$ by the Cauchy–Schwarz inequality and parts (ii) and (iii) of Assumption 3. Taking expectations over S on the event $\{\hat{w} \neq 0\}$ with positive probability gives $\mathbb{E}[\hat{\beta} \mid G, Z] - \beta = (\mathbb{E}[\rho(S) \mid G, Z] - 1)\beta$, written here without the abbreviation, which is (16). □

Lemma 10 (Weights and leverage of the feasible regression)**.** *Let part (i) of Assumption 3, Assumption 4, and the conditions of Theorem 1 hold. Then:*

(a) $\big|\,\|\hat{w}\|_2 - \|\tilde{w}\|_2\,\big| \;\leq\; \|u\|_2$, *so that* $\|\hat{w}\|_2/\sqrt{n} \;\geq\; \sqrt{\mu_n} - \|u\|_2/\sqrt{n}$. *In particular* $\mathbb{P}\big(\|\hat{w}\|_2 \geq \sqrt{n\mu_0}/2\big) \to 1$, *hence* $\mathbb{P}(\hat{w} \neq 0) \to 1$ *and* $\|\hat{w}\|_2^{-2} = O_p(n^{-1})$.

(b) If in addition Assumption 6 holds, then on $\{\hat{w} \neq 0\}$

$$m(\hat{w})^{1/2} \;\leq\; \frac{\sqrt{h_n}\,\|\tilde{w}\|_2 + \|u\|_2}{\|\hat{w}\|_2},$$

and consequently

$$m(\hat{w}) \;\xrightarrow{p}\; 0, \qquad \|\hat{w}\|_\infty^2 \;=\; o_p(n)\,, \qquad \hat{h}_n \;\leq\; h_n + m(\hat{w}) \;\xrightarrow{p}\; 0.$$

Proof. (a) The triangle inequality and $\|\hat{w} - \tilde{w}\|_2 \le \|u\|_2$ give $|\|\hat{w}\|_2 - \|\tilde{w}\|_2| \le \|u\|_2$, and (36) gives $\|\tilde{w}\|_2 \ge \sqrt{n\mu_n}$. Hence $\|\hat{w}\|_2/\sqrt{n} \ge \sqrt{\mu_n} - \|u\|_2/\sqrt{n}$. Theorem 1 applies under the conditional distribution by Lemma 7 and gives $\|u\|_2/\sqrt{n} = O_p(r_n)$ with $r_n \to 0$, so $\mathbb{P}(\|u\|_2/\sqrt{n} \le \sqrt{\mu_0}/2) \to 1$. On this event, since $\mu_n \ge \mu_0$, $\|\hat{w}\|_2 \ge \sqrt{n\mu_0}/2 > 0$ and $\|\hat{w}\|_2^{-2} \le 4/(n\mu_0)$, which gives the last two claims.

(b) On $\{\hat{w} \ne 0\}$, the triangle inequality in ℓ_∞, the bound $\tilde{w}_i^2 \le h_{ii}\|\tilde{w}\|_2^2 \le h_n\|\tilde{w}\|_2^2$ from (33), and $\|\cdot\|_\infty \le \|\cdot\|_2$ give

$$\|\hat{w}\|_\infty \;\le\; \|\tilde{w}\|_\infty + \|\hat{w} - \tilde{w}\|_\infty \;\le\; \sqrt{h_n}\,\|\tilde{w}\|_2 + \|u\|_2 \;\le\; \sqrt{h_n}\,\sqrt{n} + \|u\|_2,$$

the last step by Lemma 8(d). Dividing by $\|\hat{w}\|_2$ gives the displayed bound on $m(\hat{w})^{1/2} = \|\hat{w}\|_\infty/\|\hat{w}\|_2$. On the event $\{\|\hat{w}\|_2 \ge \sqrt{n\mu_0}/2\}$ of (a), the bound is at most $2(\sqrt{h_n} + \|u\|_2/\sqrt{n})/\sqrt{\mu_0}$. This tends to zero in probability because $h_n \to 0$ by Assumption 6 and $\|u\|_2/\sqrt{n} = O_p(r_n)$ with $r_n \to 0$. Hence $m(\hat{w}) \xrightarrow{p} 0$. Lemma 8(d) then gives

$$\|\hat{w}\|_\infty^2 \;=\; m(\hat{w})\,\|\hat{w}\|_2^2 \;\le\; n\,m(\hat{w}) \;=\; o_p(n)\,.$$

On $\{\hat{w} \ne 0\}$, (33) gives $\hat{h}_{ii} = (P_Z)_{ii} + \hat{w}_i^2/\|\hat{w}\|_2^2 \le h_{ii} + m(\hat{w})$, using $(P_Z)_{ii} \le h_{ii}$ from the full-network identity. Hence $\hat{h}_n \le h_n + m(\hat{w})$, which tends to zero in probability. □

Proof of part (ii) of Proposition 2 and of Proposition 1 (Slope consistency). For the reliability, on $\{\hat{w} \ne 0\}$, (37) gives $\rho(S) - 1 = -\hat{w}^\top u/\|\hat{w}\|_2^2$. The Cauchy–Schwarz inequality then gives $|\rho(S) - 1| \le \|u\|_2/\|\hat{w}\|_2$. By Lemma 10(a), the event $\{\|\hat{w}\|_2 \ge \sqrt{n\mu_0}/2\}$, on which $|\rho(S) - 1| \le 2\|u\|_2/\sqrt{n\mu_0}$, has probability tending to one. Moreover, $\|u\|_2/\sqrt{n} = O_p(r_n)$ by Theorem 1, which applies under the conditional distribution by Lemma 7. Hence $\rho(S) - 1 = O_p(r_n)$.

For the noise term, $e(S) = \hat{w}^\top\epsilon/\|\hat{w}\|_2^2$ on $\{\hat{w} \ne 0\}$ by definition. By Lemma 7 and part (ii) of Assumption 3, $\mathbb{E}[\hat{w}^\top\epsilon \mid S] = 0$, and by part (iii),

$$\mathrm{Var}(\hat{w}^\top\epsilon \mid S) = \hat{w}^\top\Sigma\hat{w} = \sum_i \hat{w}_i^2\sigma_i^2 \le \bar{\sigma}^2\|\hat{w}\|_2^2 \le \bar{\sigma}^2 n,$$

the last step by Lemma 8(d). Chebyshev's inequality gives $\mathbb{P}(|\hat{w}^\top\epsilon| > K\sqrt{n} \mid S) \le \bar{\sigma}^2/K^2$ for every $K > 0$. The bound does not depend on S, so integrating over S gives $\mathbb{P}(|\hat{w}^\top\epsilon| > K\sqrt{n}) \le \bar{\sigma}^2/K^2$, that is, $\hat{w}^\top\epsilon = O_p(\sqrt{n})$. On the event $\{\|\hat{w}\|_2 \ge \sqrt{n\mu_0}/2\}$, $|e(S)| \le 4|\hat{w}^\top\epsilon|/(n\mu_0)$. Hence $e(S) = O_p(n^{-1/2})$.

Proposition 1 follows. By (15), $\hat{\beta} - \beta = (\rho(S) - 1)\beta + e(S)$ on $\{\hat{w} \ne 0\}$. The first term is $O_p(r_n)$ for the fixed parameter β and the second is $O_p(n^{-1/2})$. Both tend to zero in probability because $r_n \to 0$. □

Lemma 11 (Conditional central limit theorem and variance estimation)**.** *Let Assumptions 3 and 5 hold. Let $w \in \mathbb{R}^n$, $w \ne 0$, and let Π be an orthogonal projection matrix, both functions of S. Write*

$$\xi_i \;:=\; \frac{w_i\,\epsilon_i}{s(w)}, \qquad W \;:=\; \sum_{i\in V}\xi_i \;=\; \frac{w^\top\epsilon}{s(w)},$$

and let C_{BE} be the absolute constant of the Berry–Esseen inequality. Then:

(a) $\big\|F_W(\cdot \mid S) - \Phi\big\|_\infty \;\le\; C_{\mathrm{BE}} \sum_{i \in V} \mathbb{E}\big[|\xi_i|^3 \,\big|\, S\big] \;\le\; \dfrac{C_{\mathrm{BE}}\, C^{3/4}}{\underline{\sigma}^3}\, m(w)^{1/2}.$

(b) for every $\eta \in (0, 5]$,

$$\mathbb{P}\Big(\Big|\frac{\sum_i w_i^2 \big((I-\Pi)\epsilon\big)_i^2}{s(w)^2} - 1\Big| > \eta \,\Big|\, S\Big) \;\le\; \frac{25}{\eta^2}\Big(\frac{C\, m(w)}{\underline{\sigma}^4} + \frac{\bar{\sigma}^2}{\underline{\sigma}^2}\, \max_i \Pi_{ii}\Big).$$

Proof. Part (a) is the Berry–Esseen inequality applied conditionally on S. For part (b), the proof splits the normalized variance estimate into a sum of independent squares and two terms driven by the projection Π. Chebyshev's inequality controls the first, and Markov's inequality controls the projection terms.

By Lemma 7, conditional on S the coordinates of ϵ are independent with mean zero, variances $\sigma_i^2 \in [\underline{\sigma}^2, \bar{\sigma}^2]$, and $\mathbb{E}[\epsilon_i^4 \mid S] \le C$. The mean and the upper bound on the variance are parts (ii) and (iii) of Assumption 3. The independence, the lower bound, and the fourth moment are parts (iv) to (vi) of Assumption 5. Since w and Π are functions of S, they are constants under $\mathbb{P}(\cdot \mid S)$, and $s := s(w) \ge \underline{\sigma}\|w\|_2 > 0$. Hence, conditional on S, the ξ_i are independent with $\mathbb{E}[\xi_i \mid S] = 0$ and $\sum_i \mathrm{Var}(\xi_i \mid S) = \sum_i w_i^2 \sigma_i^2 / s^2 = 1$, which is the setting of Chen et al. (2011, Chapter 3).

Two moment bounds serve both parts. By Lyapunov's inequality, $\mathbb{E}[|\epsilon_i|^3 \mid S] \le \mathbb{E}[\epsilon_i^4 \mid S]^{3/4} \le C^{3/4}$, and $\sum_i |w_i|^k \le \|w\|_\infty^{k-2} \|w\|_2^2 = m(w)^{(k-2)/2} \|w\|_2^k$ for $k \ge 2$. Hence

$$\begin{aligned} \sum_i \mathbb{E}\big[|\xi_i|^3 \,\big|\, S\big] \;&\le\; \frac{C^{3/4} \sum_i |w_i|^3}{\underline{\sigma}^3\, \|w\|_2^3} \;\le\; \frac{C^{3/4}}{\underline{\sigma}^3}\, m(w)^{1/2}, \\ \sum_i \mathbb{E}\big[\xi_i^4 \,\big|\, S\big] \;&\le\; \frac{C \sum_i w_i^4}{\underline{\sigma}^4\, \|w\|_2^4} \;\le\; \frac{C}{\underline{\sigma}^4}\, m(w). \end{aligned} \tag{38}$$

For part (a), the Berry–Esseen inequality for sums of independent random variables (Chen et al., 2011, Chapter 3) bounds $\|F_W(\cdot \mid S) - \Phi\|_\infty$ by $C_{\mathrm{BE}} \sum_i \mathbb{E}[|\xi_i|^3 \mid S]$, and (38) bounds the sum.

For part (b), expanding $((I-\Pi)\epsilon)_i = \epsilon_i - (\Pi\epsilon)_i$ gives

$$\frac{\sum_i w_i^2 \big((I-\Pi)\epsilon\big)_i^2}{s^2} \;=\; L_1 - 2L_2 + L_3,$$

where $L_1 := \sum_i \xi_i^2$, $L_2 := \sum_i w_i^2\, \epsilon_i\, (\Pi\epsilon)_i / s^2$, and $L_3 := \sum_i w_i^2\, (\Pi\epsilon)_i^2 / s^2$, with $|L_2| \le L_1^{1/2} L_3^{1/2}$ by the Cauchy–Schwarz inequality. Conditional on S, L_1 is a sum of independent terms with $\mathbb{E}[L_1 \mid S] = 1$ and

$$\mathrm{Var}(L_1 \mid S) = \sum_i \mathrm{Var}(\xi_i^2 \mid S) \le \sum_i \mathbb{E}[\xi_i^4 \mid S] \le \frac{C\, m(w)}{\underline{\sigma}^4}$$

by (38). For L_3, $\mathbb{E}[\epsilon \mid S] = 0$ and $\mathrm{Var}(\epsilon \mid S) = \Sigma$ is diagonal, so $\mathbb{E}[\epsilon\epsilon^\top \mid S] = \Sigma$. Every orthogonal projection Π that is a function of S therefore satisfies

$$\mathbb{E}\big[(\Pi\epsilon)_i^2 \mid S\big] = (\Pi\Sigma\Pi)_{ii} = \sum_j \Pi_{ij}^2 \sigma_j^2 \le \bar{\sigma}^2 \sum_j \Pi_{ij}^2 = \bar{\sigma}^2 (\Pi^2)_{ii} = \bar{\sigma}^2 \Pi_{ii}, \tag{39}$$

the last two equalities because Π is symmetric and idempotent. Hence

$$\mathbb{E}[L_3 \mid S] \;\le\; \bar{\sigma}^2 \max_i \Pi_{ii} \, \frac{\|w\|_2^2}{s^2} \;\le\; \frac{\bar{\sigma}^2}{\underline{\sigma}^2} \max_i \Pi_{ii}.$$

It remains to convert the two bounds into the tolerance η. Fix $\eta \in (0, 1]$. Chebyshev's inequality for L_1, Markov's inequality for L_3, and the union bound give

$$\mathbb{P}\big(|L_1 - 1| > \eta \;\text{ or }\; L_3 > \eta^2 \,\big|\, S\big) \;\le\; \frac{1}{\eta^2}\Big(\frac{C\, m(w)}{\underline{\sigma}^4} + \frac{\bar{\sigma}^2}{\underline{\sigma}^2} \max_i \Pi_{ii}\Big).$$

On the complementary event, $|L_1 - 1| \le \eta$ and $L_3 \le \eta^2$. Then $L_1 \le 1 + \eta \le 2$, $|L_2| \le \sqrt{2}\,\eta$, and

$$|L_1 - 2L_2 + L_3 - 1| \;\le\; \eta + 2\sqrt{2}\,\eta + \eta^2 \;\le\; (2 + 2\sqrt{2})\,\eta \;\le\; 5\,\eta,$$

because $\eta^2 \le \eta$. Hence $\mathbb{P}(|L_1 - 2L_2 + L_3 - 1| > 5\eta \mid S)$ is at most the bound of the previous display but one. Replacing η by $\eta/5$ gives (b) for $\eta \in (0, 5]$. □

Proof of Proposition 3 (The t-test at the null). The proof writes t as the ratio of a standardized numerator W_n to the square root of a standardized variance estimate $\mathcal{V}_n$, and shows that $W_n \xrightarrow{d} N(0,1)$ and $\mathcal{V}_n \xrightarrow{p} 1$. Slutsky's theorem then gives the claim.

Under $\mathcal{H}_0$, $Y = Z\theta + \epsilon$. On $\{\hat{w} \neq 0\}$, Lemma 8(b) and $\hat{w}^\top Z = 0$ give $\hat{\beta} = \hat{w}^\top \epsilon / \|\hat{w}\|_2^2$, and $M_{\hat{X}} Z = 0$ gives $\hat{\epsilon} = M_{\hat{X}} Y = M_{\hat{X}} \epsilon = (I - P_{\hat{X}})\epsilon$. By Lemma 7 and part (iii) of Assumption 3, $\mathrm{Var}(\hat{w}^\top \epsilon \mid S) = \hat{w}^\top \Sigma \hat{w} = s(\hat{w})^2$. Hence $s(\hat{w})$ is the conditional standard deviation of the numerator, and $s(\hat{w}) \ge \underline{\sigma} \|\hat{w}\|_2 > 0$ by part (v) of Assumption 5. On $\{\hat{w} \neq 0\} \cap \{\hat{\omega} \neq 0\}$, the factor $\|\hat{w}\|_2^2$ cancels between $\hat{\beta}$ and $\hat{\omega}$ in (18). Dividing the numerator and the denominator by $s(\hat{w})$ gives

$$t \;=\; \frac{\hat{w}^\top \epsilon}{\big(\sum_i \hat{w}_i^2\, \hat{\epsilon}_i^2\big)^{1/2}} \;=\; \frac{W_n}{\mathcal{V}_n^{1/2}}, \qquad W_n := \frac{\hat{w}^\top \epsilon}{s(\hat{w})}, \qquad \mathcal{V}_n := \frac{\sum_i \hat{w}_i^2 \big((I - P_{\hat{X}})\epsilon\big)_i^2}{s(\hat{w})^2}. \tag{40}$$

Each conditional bound below is integrated over S by one argument. The vector $\hat{w}$ and the matrix $P_{\hat{X}}$ are functions of S. Hence Lemma 11 applies conditional on S with $(w, \Pi) = (\hat{w}, P_{\hat{X}})$ on $\{\hat{w} \neq 0\}$, and $\max_i (P_{\hat{X}})_{ii} = \hat{h}_n$ by definition of the feasible leverage. By Lemma 10(b), $m(\hat{w}) \xrightarrow{p} 0$ and $\hat{h}_n \xrightarrow{p} 0$. Let $\mathcal{K}_n$ be events and $a_n(S) \ge 0$ be S-measurable, with $a_n(S) \xrightarrow{p} 0$ and $\mathbb{P}(\mathcal{K}_n \mid S) \le a_n(S)$ on $\{\hat{w} \neq 0\}$. Then,

$$\mathbb{P}(\mathcal{K}_n) \;\le\; \mathbb{E}\big[a_n(S) \wedge 1\big] + \mathbb{P}(\hat{w} = 0) \;\to\; 0, \tag{41}$$

because a sequence bounded by one that tends to zero in probability has expectation tending to zero, and $\mathbb{P}(\hat{w} = 0) \to 0$ by Lemma 10(a).

For the numerator, on $\{\hat{w} \neq 0\}$, W_n is the variable W of Lemma 11 with $w = \hat{w}$. Part (a) gives $\|F_{W_n}(\cdot \mid S) - \Phi\|_\infty \le C_{\mathrm{BE}} C^{3/4} \underline{\sigma}^{-3}\, m(\hat{w})^{1/2}$, a bound that tends to zero in probability. Since $F_{W_n}(x) = \mathbb{E}[F_{W_n}(x \mid S)]$ for every x, $\|F_{W_n} - \Phi\|_\infty \le \mathbb{E}\|F_{W_n}(\cdot \mid S) - \Phi\|_\infty$. The Kolmogorov distance is at most one on $\{\hat{w} = 0\}$, and the argument of (41) gives

$$\|F_{W_n} - \Phi\|_\infty \;\le\; \mathbb{E}\big[C_{\mathrm{BE}} C^{3/4} \underline{\sigma}^{-3} m(\hat{w})^{1/2} \wedge 1\big] + \mathbb{P}(\hat{w} = 0) \;\to\; 0.$$

Hence $W_n \xrightarrow{d} N(0,1)$.

For the denominator, fix $\eta \in (0,5]$. Lemma 11(b) with $(w,\Pi) = (\hat{w}, P_{\hat{X}})$ gives $\mathbb{P}(|\mathcal{V}_n - 1| > \eta \mid S) \le a_n(S)$ on $\{\hat{w} \neq 0\}$ for the S-measurable bound

$$a_n(S) \;:=\; \frac{25}{\eta^2}\Big(\frac{C\, m(\hat{w})}{\underline{\sigma}^4} + \frac{\bar{\sigma}^2}{\underline{\sigma}^2}\,\hat{h}_n\Big) \;\xrightarrow{p}\; 0.$$

Then (41) gives $\mathbb{P}(|\mathcal{V}_n - 1| > \eta) \to 0$. Since $\hat{\omega} = 0$ exactly when $\mathcal{V}_n = 0$, it follows that

$$\mathcal{V}_n \;\xrightarrow{p}\; 1, \qquad \mathbb{P}(\hat{\omega} = 0) \;\to\; 0. \tag{42}$$

In particular, the convention $t := 0$ on the event $\{\hat{\omega} = 0\}$ affects no limit.

By (40), $t = W_n / \mathcal{V}_n^{1/2}$, with $W_n \xrightarrow{d} N(0,1)$ and $\mathcal{V}_n^{1/2} \xrightarrow{p} 1$ by (42). Slutsky's theorem gives $t \xrightarrow{d} N(0,1)$. □

The full-network regression of Y on X has slope $\tilde{\beta}$, residual vector $\tilde{\epsilon} := M_X Y$, variance estimate $\tilde{\omega}^2 := \sum_i \tilde{w}_i^2 \tilde{\epsilon}_i^2 / \|\tilde{w}\|_2^4$, and t-statistic $\tilde{t} := \tilde{\beta}/\tilde{\omega}$, with $\tilde{t} := 0$ when $\tilde{\omega} = 0$.

Corollary 1 (The full-network regression)**.** *Let the hypotheses of Proposition 3 hold and write* $\tilde{W}_n := \tilde{w}^\top \epsilon / s(\tilde{w})$. *Then*

$$\|F_{\tilde{W}_n} - \Phi\|_\infty \;\le\; \frac{C_{\mathrm{BE}}\, C^{3/4}}{\underline{\sigma}^3}\, h_n^{1/2} \;\to\; 0,$$

so $\tilde{W}_n \xrightarrow{d} N(0,1)$, *and under* $\mathcal{H}_0 : \beta = 0$,

$$\tilde{\mathcal{V}}_n \;:=\; \frac{\sum_i \tilde{w}_i^2\, \tilde{\epsilon}_i^2}{s(\tilde{w})^2} \;\xrightarrow{p}\; 1, \qquad \mathbb{P}(\tilde{\omega} = 0) \;\to\; 0, \qquad \tilde{t} \;\xrightarrow{d}\; N(0,1).$$

Proof. The vector $\tilde{w}$ and the projection P_X are nonrandom, hence functions of S, and $\tilde{w} \neq 0$ because $\|\tilde{w}\|_2^2 \ge n\mu_n > 0$ by (36). By (33) and the definition of h_n, $m(\tilde{w}) = \max_i \tilde{w}_i^2 / \|\tilde{w}\|_2^2 \le h_n$ and $\max_i (P_X)_{ii} = h_n$, and $h_n \to 0$ by Assumption 6. Lemma 11 therefore applies with $(w,\Pi) = (\tilde{w}, P_X)$, and its bounds do not depend on S.

For the numerator, Lemma 11(a) with $w = \tilde{w}$ gives $\|F_{\tilde{W}_n}(\cdot \mid S) - \Phi\|_\infty \le C_{\mathrm{BE}} C^{3/4} \underline{\sigma}^{-3} h_n^{1/2}$. Since $F_{\tilde{W}_n}(x) = \mathbb{E}[F_{\tilde{W}_n}(x \mid S)]$ for every x, the same bound holds for $\|F_{\tilde{W}_n} - \Phi\|_\infty$, and it tends to zero.

For the denominator and the ratio, take $\beta = 0$. The derivation of (40), with X and $\tilde{w}$ in place of $\hat{X}$ and $\hat{w}$, applies because Lemma 8(b) holds for X, $\tilde{w}^\top Z = 0$, $M_X Z = 0$, and $s(\tilde{w}) \ge \underline{\sigma}\|\tilde{w}\|_2 > 0$. It gives $\tilde{\beta} = \tilde{w}^\top \epsilon / \|\tilde{w}\|_2^2$ and $\tilde{\epsilon} = M_X \epsilon = (I - P_X)\epsilon$, and on $\{\tilde{\omega} \neq 0\}$

$$\tilde{t} \;=\; \frac{\tilde{W}_n}{\tilde{\mathcal{V}}_n^{1/2}}. \tag{43}$$

Lemma 11(b) with $(w,\Pi) = (\tilde{w}, P_X)$ gives, for every $\eta \in (0,5]$, $\mathbb{P}(|\tilde{\mathcal{V}}_n - 1| > \eta \mid S) \le 25\eta^{-2}(C h_n / \underline{\sigma}^4 + \bar{\sigma}^2 h_n / \underline{\sigma}^2)$, a bound that does not depend on S and tends to zero. Hence $\tilde{\mathcal{V}}_n \xrightarrow{p} 1$. Since $\mathbb{P}(\tilde{\omega} = 0) = \mathbb{P}(\tilde{\mathcal{V}}_n = 0) \to 0$, the convention $\tilde{t} := 0$ on $\{\tilde{\omega} = 0\}$ affects no limit, and Slutsky's theorem gives $\tilde{t} \xrightarrow{d} N(0,1)$. □

Proposition 4 (Full-network equivalence at the null)**.** *Under the hypotheses of Proposition 3 and $\mathcal{H}_0 : \beta = 0$, the feasible t-statistic t and the full-network t-statistic $\tilde{t}$ are asymptotically equivalent,*

$$t - \tilde{t} \xrightarrow{p} 0.$$

In particular, for every $z > 0$, $\mathbb{P}\big(\mathbf{1}\{|t| > z\} \neq \mathbf{1}\{|\tilde{t}| > z\}\big) \to 0$.

Proof. Take $\beta = 0$. On $\{\hat{w} \neq 0\} \cap \{\hat{\omega} \neq 0\}$, $t = W_n/\mathcal{V}_n^{1/2}$ by (40), and on $\{\tilde{\omega} \neq 0\}$, $\tilde{t} = \tilde{W}_n/\tilde{\mathcal{V}}_n^{1/2}$ by (43). The complements of these events have probability tending to zero, by Lemma 10(a), (42), and Corollary 1. On their intersection,

$$t - \tilde{t} \;=\; \frac{W_n - \tilde{W}_n}{\mathcal{V}_n^{1/2}} \;+\; \tilde{W}_n\Big(\frac{1}{\mathcal{V}_n^{1/2}} - \frac{1}{\tilde{\mathcal{V}}_n^{1/2}}\Big).$$

Both $\mathcal{V}_n$ and $\tilde{\mathcal{V}}_n$ converge in probability to one, by (42) and Corollary 1, and $\tilde{W}_n = O_p(1)$ because it converges in distribution by Corollary 1. The second term is therefore $o_p(1)$, and it suffices to show $W_n - \tilde{W}_n \xrightarrow{p} 0$.

For the numerators, write $W_n - \tilde{W}_n = \Delta^\top \epsilon$ with the S-measurable vector

$$\Delta \;:=\; \frac{\hat{w}}{s(\hat{w})} - \frac{\tilde{w}}{s(\tilde{w})}.$$

Under part (v) of Assumption 5, $s(\cdot)$ is the norm $w \mapsto \|\Sigma^{1/2} w\|_2$, so $\underline{\sigma}\|w\|_2 \le s(w) \le \bar{\sigma}\|w\|_2$ and, by the triangle inequality for this norm, $|s(\hat{w}) - s(\tilde{w})| \le s(\hat{w} - \tilde{w}) \le \bar{\sigma}\|\hat{w} - \tilde{w}\|_2$. Hence

$$\|\Delta\|_2 \;\le\; \frac{\|\hat{w} - \tilde{w}\|_2}{s(\hat{w})} + \frac{\|\tilde{w}\|_2\,|s(\tilde{w}) - s(\hat{w})|}{s(\hat{w})\,s(\tilde{w})} \;\le\; \Big(1 + \frac{\bar{\sigma}}{\underline{\sigma}}\Big)\frac{\|\hat{w} - \tilde{w}\|_2}{s(\hat{w})}.$$

On the event $\{\|\hat{w}\|_2 \ge \sqrt{n\mu_0}/2\}$ of Lemma 10(a), $s(\hat{w}) \ge \underline{\sigma}\sqrt{n\mu_0}/2$, and $\|\hat{w} - \tilde{w}\|_2 \le \|u\|_2 = O_p(\sqrt{n}\, r_n)$ by Theorem 1. Hence $\|\Delta\|_2 = O_p(r_n)$. By Lemma 7 and parts (ii) and (iii) of Assumption 3, $\mathbb{E}[\Delta^\top \epsilon \mid S] = 0$ and $\mathrm{Var}(\Delta^\top \epsilon \mid S) = \sum_i \Delta_i^2 \sigma_i^2 \le \bar{\sigma}^2 \|\Delta\|_2^2$. Chebyshev's inequality gives $\mathbb{P}(|\Delta^\top \epsilon| > K\|\Delta\|_2 \mid S) \le \bar{\sigma}^2/K^2$ for every $K > 0$. The bound is free of S, so $W_n - \tilde{W}_n = O_p(\|\Delta\|_2) = O_p(r_n) \xrightarrow{p} 0$.

For the rejection regions, fix $z > 0$ and $\eta \in (0, z)$. If $|t - \tilde{t}| \le \eta$ and $||\tilde{t}| - z| > \eta$, then $|t|$ and $|\tilde{t}|$ lie on the same side of z. The two indicators therefore differ only if $|t - \tilde{t}| > \eta$ or $||\tilde{t}| - z| \le \eta$, and the union bound gives

$$\mathbb{P}\big(\mathbf{1}\{|t| > z\} \neq \mathbf{1}\{|\tilde{t}| > z\}\big) \;\le\; \mathbb{P}\big(|t - \tilde{t}| > \eta\big) + \mathbb{P}\big(z - \eta \le |\tilde{t}| \le z + \eta\big).$$

The first term tends to zero because $t - \tilde{t} \xrightarrow{p} 0$. Since $\tilde{t} \xrightarrow{d} N(0,1)$ by Corollary 1 and the limit distribution function is continuous, the second term tends to $2[\Phi(z+\eta) - \Phi(z-\eta)] \le 4\eta/\sqrt{2\pi}$. Letting $\eta \to 0$ completes the proof. □

Corollary 2 (Consistency against fixed alternatives)**.** *Let the hypotheses of Proposition 1 and Assumption 6 hold, and let $\beta \neq 0$ be fixed. Then $\hat{\omega} \xrightarrow{p} 0$ and, for every $z > 0$, $\mathbb{P}\big(|\hat{\beta}| > z\,\hat{\omega}\big) \to 1$. That is, the t-test of Proposition 3 is consistent.*

Proof. Proposition 1 gives $\hat{\beta} \xrightarrow{p} \beta \neq 0$, so it suffices to show $\hat{\omega} \xrightarrow{p} 0$. On $\{\hat{w} \neq 0\}$, $M_{\hat{X}} Z = 0$ and $M_{\hat{X}} \hat{v} = 0$ give $\hat{\epsilon} = M_{\hat{X}} Y = M_{\hat{X}} \epsilon - \beta M_{\hat{X}} u$. The inequality $(x+y)^2 \leq 2x^2 + 2y^2$ then gives

$$\hat{\omega}^2 \|\hat{w}\|_2^4 \;=\; \sum_i \hat{w}_i^2 \, \hat{\epsilon}_i^2 \;\leq\; 2 \sum_i \hat{w}_i^2 \big(M_{\hat{X}} \epsilon\big)_i^2 + 2\beta^2 \|\hat{w}\|_\infty^2 \, \|u\|_2^2.$$

By Lemma 7, parts (ii) and (iii) of Assumption 3, and (39) with $\Pi = M_{\hat{X}}$, the first sum has conditional mean

$$\sum_i \hat{w}_i^2 \big(M_{\hat{X}} \Sigma M_{\hat{X}}\big)_{ii} \;\leq\; \bar{\sigma}^2 \sum_i \hat{w}_i^2 (M_{\hat{X}})_{ii} \;\leq\; \bar{\sigma}^2 \|\hat{w}\|_2^2 \;\leq\; \bar{\sigma}^2 n.$$

The second inequality uses $(M_{\hat{X}})_{ii} \leq 1$, and the last is Lemma 8(d). The first sum is therefore $O_p(n)$ by Markov's inequality. The second term is $o_p(n) \cdot O_p(n r_n^2)$ by Lemma 10(b) and Theorem 1, and $\|\hat{w}\|_2^{-4} = O_p(n^{-2})$ by Lemma 10(a). Since also $\mathbb{P}(\hat{w} = 0) \to 0$ by Lemma 10(a), $\hat{\omega}^2 = O_p(n^{-1}) + o_p(r_n^2) \xrightarrow{p} 0$, because $r_n \to 0$.

For the rejection event, fix $z > 0$. Then $\mathbb{P}(|\hat{\beta}| > z\hat{\omega}) \geq \mathbb{P}(|\hat{\beta}| \geq |\beta|/2, \; z\hat{\omega} < |\beta|/2) \to 1$. □

A.3 Further results

Proposition 5 uses the unit normalization of Definition 3, and the other results of this subsection use the normalization of the previous subsection.

Proposition 5 (Uniqueness and support)**.** *For every sample S, $0 < \hat{\lambda} \leq \lambda$. If the sampled nodes lie in a single component of the sampled network $\hat{G}$, then the estimator $\hat{v}$ of Definition 3 is unique, and its support is exactly $S \cup N(S)$:*

$$\hat{v}_i \;>\; 0 \quad \textit{if } i \in S \cup N(S), \qquad \hat{v}_i \;=\; 0 \quad \textit{otherwise.} \tag{44}$$

Proof. For a unit vector y, write $|y|$ for the vector of absolute values, a unit vector. Since the sample adjacency matrix $\hat{A}$ has nonnegative entries and $0 \leq \hat{A} \leq A$ entrywise by (2),

$$y^\top \hat{A} y \leq |y|^\top \hat{A} |y| \leq |y|^\top A |y| \leq \lambda,$$

the last step by the Rayleigh quotient of A at $|y|$. Maximizing over y gives $\hat{\lambda} \leq \lambda$. For the lower bound, pick $i \in S$, which exists because $S \neq \emptyset$. Connectedness with $n \geq 2$ gives $d_i \geq 1$, so there is a node j with $A_{ij} = 1$, and $\hat{A}_{ij} = A_{ij} = 1$ because $i \in S$. The unit vector $y := (e_i + e_j)/\sqrt{2}$ gives $y^\top \hat{A} y = \hat{A}_{ij} = 1$. Hence $\hat{\lambda} \geq 1 > 0$.

For uniqueness and support, write $S^+ := S \cup N(S)$ for the observed nodes. Under Assumption 1, a node has a link in the sampled network $\hat{G}$ if and only if it lies in S^+. Indeed, every $i \in S$ has its $d_i \geq 1$ links in $\hat{G}$, every $j \in N(S)$ has a link to S in $\hat{G}$, and a node j outside S^+ has $\hat{A}_{jk} = 0$ for every k by (2). Ordering the nodes so that S^+ comes first displays $\hat{A}$ as the block diagonal matrix $\operatorname{diag}(\hat{A}_{S^+S^+}, 0)$, where the principal submatrix $\hat{A}_{S^+S^+}$ is symmetric with nonnegative entries and contains every link of $\hat{G}$. The spectrum of $\hat{A}$ is the spectrum of $\hat{A}_{S^+S^+}$ together with zeros. Since $\hat{\lambda} > 0$, $\hat{\lambda} = \max\{\lambda_1(\hat{A}_{S^+S^+}), 0\} = \lambda_1(\hat{A}_{S^+S^+})$.

Every node in $N(S)$ has a link to S in $\hat{G}$. Hence the component of the sampled network $\hat{G}$ that contains the sampled nodes contains S^+, and no node outside S^+ has a link in $\hat{G}$.

If the sampled nodes lie in a single component, that component is therefore exactly S^+, and $\hat{A}_{S^+S^+}$ is irreducible. By the Perron–Frobenius theorem, $\hat{\lambda} = \lambda_1(\hat{A}_{S^+S^+})$ is a simple eigenvalue of $\hat{A}_{S^+S^+}$ with an eigenvector g that is positive on S^+. An eigenvector y of $\hat{A}$ for $\hat{\lambda}$, partitioned as $y = (y_1, y_2)$ along S^+ and its complement, satisfies $\hat{A}_{S^+S^+} y_1 = \hat{\lambda} y_1$ and $0 = \hat{\lambda} y_2$. Hence $y_2 = 0$ because $\hat{\lambda} > 0$, and y_1 is a multiple of g. The eigenspace of $\hat{A}$ for $\hat{\lambda}$ is the line spanned by $(g, 0)$, which contains exactly one nonnegative unit vector. This vector is $\hat{v}$, and its support is $S^+ = S \cup N(S)$, which is (44). □

With an intercept only, $Z = \mathbf{1}$, write $\bar{\hat{v}} := n^{-1} \sum_i \hat{v}_i$, $\bar{v} := n^{-1} \sum_i v_i$, and $\bar{u} := \bar{\hat{v}} - \bar{v}$ for the means of the plug-in centrality $\hat{v}$, the centrality v, and the measurement error u.

Corollary 3 (Intercept only)**.** *Let $Z = \mathbf{1}$. On the event $\{\hat{w} \neq 0\}$,*

$$\rho(S) - 1 \;=\; \frac{\bar{\hat{v}}\,\bar{u} \;-\; \frac{1}{2}\|u\|_2^2/n}{1 - \bar{\hat{v}}^{\,2}}, \qquad \bar{u} \;=\; \frac{\|\hat{v}\|_1 - \|v\|_1}{n}. \tag{45}$$

In particular, $\rho(S) \leq 1$ whenever $\bar{u} \leq 0$.

Proof. With $Z = \mathbf{1}$, $P_Z = n^{-1}\mathbf{1}\mathbf{1}^\top$, so $\hat{v}^\top P_Z u = n\,\bar{\hat{v}}\,\bar{u}$. Since $\|\hat{v}\|_2^2 = n$, $\|\hat{w}\|_2^2 = \|\hat{v}\|_2^2 - \|P_Z \hat{v}\|_2^2 = n(1 - \bar{\hat{v}}^{\,2})$, which is positive on $\{\hat{w} \neq 0\}$. Dividing (37) of Lemma 9 by $\|\hat{w}\|_2^2$ gives the first identity in (45). The second holds because v and $\hat{v}$ are nonnegative. If $\bar{u} \leq 0$, the numerator is nonpositive because $\bar{\hat{v}} \geq 0$, and $\rho(S) \leq 1$. □

Proposition 6 (Rate floor)**.** *For every connected graph on $n \geq 2$ nodes, $\sqrt{n}\, r_n \;\geq\; \frac{1}{4}$.*

Proof. Every eigenvalue contributes a nonnegative square to the trace of A^2, so

$$\lambda^2 \;=\; \lambda_1(A)^2 \;\leq\; \sum_{i=1}^{n} \lambda_i(A)^2 \;=\; \operatorname{tr}(A^2) \;=\; \sum_{i,j} A_{ij} A_{ji} \;=\; \sum_{i,j} A_{ij} \;=\; n\bar{d},$$

the last two equalities by symmetry and $A_{ij} \in \{0, 1\}$. Hence $\lambda \leq \sqrt{n\bar{d}}$.

The remaining factors of r_n satisfy

$$\bar{d} \;\geq\; 1, \qquad \gamma \;=\; 1 - \lambda_2(P) \;\leq\; 2, \qquad \delta \;=\; 1 - \lambda_2(A)/\lambda \;\leq\; 2, \qquad \alpha \;\leq\; 1.$$

The first holds because connectedness with $n \geq 2$ gives $d_i \geq 1$ at every node. The second holds because $\lambda_2(P) \geq \lambda_n(P) \geq -1$ by the eigenvalue ordering recorded after (21). The third holds because $\lambda_2(A) \geq \lambda_n(A) \geq -\lambda_1(A) = -\lambda$ by Lemma 1. The last is Definition 5. Substituting $\lambda \leq \sqrt{n\bar{d}}$ into $r_n = \bar{d}/(\alpha\gamma\delta\lambda)$ and then these bounds,

$$\sqrt{n}\, r_n \;=\; \frac{\sqrt{n}\,\bar{d}}{\delta\gamma\alpha\lambda} \;\geq\; \frac{\sqrt{n}\,\bar{d}}{\delta\gamma\alpha\sqrt{n\bar{d}}} \;=\; \frac{\sqrt{\bar{d}}}{\delta\gamma\alpha} \;\geq\; \frac{1}{4\alpha} \;\geq\; \frac{1}{4}.$$

□

Proposition 7 (Graph content of Assumption 6)**.** *(a) For every $i \in V$, $v_i^2/n \leq 2\,h_{ii}$ and $(P_Z)_{ii} \leq h_{ii}$. Moreover,*

$$h_n \;\leq\; \max_i (P_Z)_{ii} \left(1 + \frac{2}{\mu_n}\right) + \frac{2}{\mu_n} \max_i \frac{v_i^2}{n}.$$

Hence, under Assumption 4, Assumption 6 holds if and only if $\max_i v_i^2/n \to 0$ and $\max_i (P_Z)_{ii} \to 0$.

(b) For every connected graph, $\max_i v_i^2/n \le d_{\max}/\lambda^2$*, where* $d_{\max} := \max_i d_i$.

Proof. (a) By (33) and Lemma 8(d), $\tilde{w}_i^2 = (h_{ii} - (P_Z)_{ii})\,\|\tilde{w}\|_2^2 \le (h_{ii} - (P_Z)_{ii})\,n$. Since P_Z is symmetric and idempotent, $\|P_Z e_i\|_2^2 = (P_Z)_{ii}$, and the Cauchy–Schwarz inequality gives

$$|(P_Z v)_i| = |(P_Z e_i)^\top v| \le \|P_Z e_i\|_2 \|v\|_2 = \sqrt{(P_Z)_{ii}}\,\sqrt{n}.$$

As $v = \tilde{w} + P_Z v$, the triangle inequality gives

$$\frac{|v_i|}{\sqrt{n}} \;\le\; \sqrt{h_{ii} - (P_Z)_{ii}} + \sqrt{(P_Z)_{ii}} \;\le\; \sqrt{2h_{ii}},$$

the last step because $(\sqrt{x} + \sqrt{y})^2 \le 2(x+y)$. The inequality $(P_Z)_{ii} \le h_{ii}$ is (33) with $\tilde{w}_i^2 \ge 0$. For the bound on h_n, (33) and (36) give $h_{ii} = (P_Z)_{ii} + \tilde{w}_i^2/\|\tilde{w}\|_2^2 \le (P_Z)_{ii} + \tilde{w}_i^2/(n\mu_n)$. The inequality $(x+y)^2 \le 2x^2 + 2y^2$ and the bound on $|(P_Z v)_i|$ give

$$\tilde{w}_i^2 = (v_i - (P_Z v)_i)^2 \le 2v_i^2 + 2(P_Z v)_i^2 \le 2v_i^2 + 2n(P_Z)_{ii}.$$

Taking the maximum over i gives the bound on h_n. For the equivalence, the two inequalities for each i give $\max_i v_i^2/n \le 2h_n$ and $\max_i (P_Z)_{ii} \le h_n$. The bound on h_n gives the converse, because $\mu_n \ge \mu_0 > 0$ under Assumption 4.

(b) The vector $v/\sqrt{n}$ is the unit Perron vector. For every i, the eigenvector equation and the Cauchy–Schwarz inequality give

$$\frac{\lambda v_i}{\sqrt{n}} = \sum_j A_{ij} \frac{v_j}{\sqrt{n}} \le \Big(\sum_j A_{ij}\Big)^{1/2} \Big(\sum_j A_{ij} \frac{v_j^2}{n}\Big)^{1/2} \le \sqrt{d_i},$$

the last step by $A_{ij} \in \{0,1\}$ and $\sum_j v_j^2/n = 1$. Hence $v_i^2/n \le d_i/\lambda^2 \le d_{\max}/\lambda^2$. □

References

Acemoglu, D., V. M. Carvalho, A. Ozdaglar, and A. Tahbaz-Salehi (2012): "The Network Origins of Aggregate Fluctuations," *Econometrica*, 80, 1977–2016.

Aldous, D. and J. A. Fill (2002): "Reversible Markov Chains and Random Walks on Graphs," .

Aldous, D. J. (1981): "Representations for partially exchangeable arrays of random variables," *Journal of Multivariate Analysis*, 11, 581–598.

Banerjee, A., E. Breza, A. G. Chandrasekhar, E. Duflo, M. O. Jackson, and C. Kinnan (2024): "Changes in Social Network Structure in Response to Exposure to Formal Credit Markets," *The Review of Economic Studies*, 91, 1331–1372.

Banerjee, A., A. G. Chandrasekhar, E. Duflo, and M. O. Jackson (2013): "The Diffusion of Microfinance," *Science*, 341, 1236498.

——— (2019): "Using Gossips to Spread Information: Theory and Evidence from Two Randomized Controlled Trials," *The Review of Economic Studies*, 86, 2453–2490.

Bar-Yossef, Z., A. Berg, S. Chien, J. Fakcharoenphol, and D. Weitz (2000): "Approximating aggregate queries about web pages via random walks," in *Proceedings of the 26th international conference on very large data bases (VLDB 2000)*, ed. by A. El Abbadi, M. L. Brodie, S. Chakravarthy, U. Dayal, N. Kamel, G. Schlageter, and K.-Y. Whang, Cairo, Egypt: Morgan Kaufmann, 535–544.

Barabási, A.-L. and R. Albert (1999): "Emergence of Scaling in Random Networks," *Science*, 286, 509–512.

Battaglia, L., T. Christensen, S. Hansen, and S. Sacher (2025): "Inference for Regression with Variables Generated by AI or Machine Learning," ArXiv:2402.15585 [econ.EM].

Beaman, L., A. BenYishay, J. Magruder, and A. M. Mobarak (2021): "Can Network Theory-Based Targeting Increase Technology Adoption?" *American Economic Review*, 111, 1918–1943.

Bernard, A. B., E. Dhyne, G. Magerman, K. Manova, and A. Moxnes (2022): "The Origins of Firm Heterogeneity: A Production Network Approach," *Journal of Political Economy*, 130, 1765–1804.

Bhattacharyya, S. and P. J. Bickel (2015): "Subsampling Bootstrap of Count Features of Networks," *The Annals of Statistics*, 43, 2384–2411.

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

Bound, J., C. Brown, and N. Mathiowetz (2001): "Measurement error in survey data," in *Handbook of econometrics*, ed. by J. J. Heckman and E. Leamer, Amsterdam: Elsevier, vol. 5, 3705–3843, section: 59.

Breza, E., A. G. Chandrasekhar, S. Lubold, T. H. McCormick, and M. Pan (2023): "Consistently estimating network statistics using aggregated relational data," *Proceedings of the National Academy of Sciences*, 120, e2207185120.

Breza, E., A. G. Chandrasekhar, T. H. McCormick, and M. Pan (2017): "Using Aggregated Relational Data to Feasibly Identify Network Structure without Network Data," .

——— (2020): "Using Aggregated Relational Data to Feasibly Identify Network Structure without Network Data," *American Economic Review*, 110, 2454–2484.

Cai, Y. (2026): "Linear Regression with Centrality Measures," ArXiv:2210.10024 [econ.EM].

Candes, E. J. and T. Tao (2010): "The Power of Convex Relaxation: Near-Optimal Matrix Completion," *IEEE Transactions on Information Theory*, 56, 2053–2080.

Chandrasekhar, A. G. and R. Lewis (2016): "Econometrics of Sampled Networks," .

Chatterjee, S., P. Diaconis, and A. Sly (2011): "Random graphs with a given degree sequence," *The Annals of Applied Probability*, 21, 1400–1435.

Chen, L. H. Y., L. Goldstein, and Q.-M. Shao (2011): "Berry–Esseen Bounds for Independent Random Variables," in *Normal Approximation by Stein's Method*, ed. by L. H. Chen, L. Goldstein, and Q.-M. Shao, Berlin, Heidelberg: Springer, 45–62.

Chen, Y., Y. Chi, J. Fan, and C. Ma (2021): "Spectral Methods for Data Science: A Statistical Perspective," *Foundations and Trends® in Machine Learning*, 14, 566–806, arXiv:2012.08496 [stat.ML].

Chung, F. and L. Lu (2002): "Connected Components in Random Graphs with Given Expected Degree Sequences," *Annals of Combinatorics*, 6, 125–145.

Chung, F., L. Lu, and V. Vu (2003): "Eigenvalues of Random Power law Graphs," *Annals of Combinatorics*, 7, 21–33.

Cook, J. R. and L. A. Stefanski (1994): "Simulation-Extrapolation Estimation in Parametric Measurement Error Models," *Journal of the American Statistical Association*, 89, 1314–1328.

Crane, H. and W. Dempsey (2018): "EDGE EXCHANGEABLE MODELS FOR INTERACTION NETWORKS," *Journal of the American Statistical Association*, 113, 1311–1326.

Cruz, C., J. Labonne, and P. Querubín (2017): "Politician Family Networks and Electoral Outcomes: Evidence from the Philippines," *American Economic Review*, 107, 3006–3037.

Davis, C. and W. M. Kahan (1970): "The Rotation of Eigenvectors by a Perturbation. III," *SIAM Journal on Numerical Analysis*, 7, 1–46.

de Paula, A., S. Richards-Shubik, and E. Tamer (2018): "Identifying Preferences in Networks With Bounded Degree," *Econometrica*, 86, 263–288.

Demir, B., A. C. Fieler, D. Y. Xu, and K. K. Yang (2024): "O-Ring Production Networks," *Journal of Political Economy*, 132, 200–247.

Eicker, F. (1963): "Asymptotic Normality and Consistency of the Least Squares Estimators for Families of Linear Regressions," *The Annals of Mathematical Statistics*, 34, 447–456.

Frank, O. (1977): "Survey sampling in graphs," *Journal of Statistical Planning and Inference*, 1, 235–264.

Fuller, W. A. (1987): *Measurement error models*, Wiley series in probability and mathematical statistics, New York: John Wiley & Sons.

Galeotti, A., B. Golub, and S. Goyal (2020): "Targeting Interventions in Networks," *Econometrica*, 88, 2445–2471.

Gile, K. J. and M. S. Handcock (2010): "Respondent-Driven Sampling: An Assessment of Current Methodology," *Sociological Methodology*, 40, 285–327.

Golub, B. and M. O. Jackson (2012): "How Homophily Affects the Speed of Learning and Best-Response Dynamics," *The Quarterly Journal of Economics*, 127, 1287–1338.

Goodman, L. A. (1961): "Snowball Sampling," *The Annals of Mathematical Statistics*, 32, 148–170.

Graham, B. S. (2017): "An Econometric Model of Network Formation With Degree Heterogeneity," *Econometrica*, 85, 1033–1063.

——— (2024): "Sparse Network Asymptotics for Logistic Regression Under Possible Misspecification," *Econometrica*, 92, 1837–1868.

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, 779–805.

Heckathorn, D. D. (1997): "Respondent-Driven Sampling: A New Approach to the Study of Hidden Populations," *Social Problems*, 44, 174–199.

Henzinger, M. R., A. Heydon, M. Mitzenmacher, and M. Najork (2000): "On near-uniform URL sampling," *Computer Networks*, 33, 295–308.

Hofmeister, M. (1988): "Spectral Radius and Degree Sequence," *Mathematische Nachrichten*, 139, 37–44.

Hoover, D. N. (1979): "Relations on probability spaces and arrays of random variables," Preprint, Institute for Advanced Study, Princeton, NJ.

Horvitz, D. G. and D. J. Thompson (1952): "A Generalization of Sampling Without Replacement from a Finite Universe," *Journal of the American Statistical Association*, 47, 663–685.

Jackson, M. O., B. W. Rogers, and Y. Zenou (2017): "The Economic Consequences of Social-Network Structure," *Journal of Economic Literature*, 55, 49–95.

Jochmans, K. and M. Weidner (2019): "Fixed-Effect Regressions on Network Data," *Econometrica*, 87, 1543–1560.

Kline, P., R. Saggio, and M. Sølvsten (2020): "Leave-Out Estimation of Variance Components," *Econometrica*, 88, 1859–1898.

Klovdahl, A. S. (1989): "Urban social networks: Some methodological problems and possibilities," in *The small world*, ed. by M. Kochen, Norwood, NJ: Ablex Publishing Corporation, Communication and information science, 176–210, section: 10.

La Porta, R. and A. Shleifer (2014): "Informality and Development," *Journal of Economic Perspectives*, 28, 109–126.

Le, C. M. and T. Li (2022): "Linear Regression and Its Inference on Noisy Network-Linked Data," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84, 1851–1885.

Lee, S. H., P.-J. Kim, and H. Jeong (2006): "Statistical properties of sampled networks," *Physical Review E*, 73, 016102.

Leung, M. P. (2023): "Network Cluster-Robust Inference," *Econometrica*, 91, 641–667.

Levin, K. and E. Levina (2025): "Bootstrapping networks with latent space structure," *Electronic Journal of Statistics*, 19, 745–791.

Lovász, L. and B. Szegedy (2006): "Limits of dense graph sequences," *Journal of Combinatorial Theory, Series B*, 96, 933–957.

Mele, A. (2017): "A Structural Model of Dense Network Formation," *Econometrica*, 85, 825–850.

Orbanz, P. and D. M. Roy (2015): "Bayesian Models of Graphs, Arrays and Other Exchangeable Random Structures," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37, 437–461.

Pagan, A. (1984): "Econometric Issues in the Analysis of Regressions with Generated Regressors," *International Economic Review*, 25, 221–247.

Rubin, D. B. (1976): "Inference and missing data," *Biometrika*, 63, 581–592.

Ruggeri, N. and C. De Bacco (2020): "Sampling on Networks: Estimating Eigenvector Centrality on Incomplete Networks," in *Complex Networks and Their Applications VIII*, ed. by H. Cherifi, S. Gaito, J. F. Mendes, E. Moro, and L. M. Rocha, Cham: Springer International Publishing, 90–101.

Sheng, S. (2020): "A Structural Econometric Analysis of Network Formation Games Through Subnetworks," *Econometrica*, 88, 1829–1858.

Stumpf, M. P. H., C. Wiuf, and R. M. May (2005): "Subnets of scale-free networks are not scale-free: Sampling properties of networks," *Proceedings of the National Academy of Sciences of the United States of America*, 102, 4221–4224.

Thirkettle, M. (2019): "Identification and Estimation of Network Statistics with Missing Link Data," .

Thompson, S. K. (2006): "Targeted random walk designs," *Survey Methodology*, 32, 11–24.

Visokay, A., L. Boudreau, R. M. Heath, and T. H. McCormick (2026): "Randomized Recruitment Driven Sampling," ArXiv:2603.00365 [stat.AP].

Voitalov, I., P. van der Hoorn, R. van der Hofstad, and D. Krioukov (2019): "Scale-free networks well done," *Physical Review Research*, 1, 033034.