Network Sampling and Missing Vertices (Edges) Problems

Chih-Sheng Hsieh

Department of Economics
National Taiwan University

April 29, 2024

Sampling Methods in Social Network Studies

There are different sampling approaches for network data proposed in sociological, statistical, and computer science literature (Hu and Lau, 2013). Here we summarize three main directions:

- Random node (or edge) sampling
- Random walk sampling (Klovdahl, 1989; Lovász, 1993)
- Graph traversal sampling:
  - forest-fire sampling (Leskovec and Faloutsos, 2006; Gjoka et al., 2010);
  - snowball sampling (Coleman, 1958; Goodman, 1961);
  - respondent-driven sampling (Heckathorn, 1997; Salganik and Heckathorn, 2004).

Sampling Methods in Social Network Studies

Network sampling usually comes with three tasks:

Task 1: sample a set of nodes (or edges) and estimate nodal or edge properties of the full (population) network, e.g., average degree, degree distribution.

Task 2: sample a small subnetwork similar to the full network to maintain global structure characteristics such as clustering coefficient, community structure, degree distribution.

Task 3: sample local structures (such as triangles, k-circles) to estimate their relative frequencies or counts.

Sampling Methods in Social Network Studies

In Task 1,

- Population is the entire vertex set (for vertex sampling) and the entire edge set (for edge sampling).
- When a vertex is sampled, we observe its complete edges.

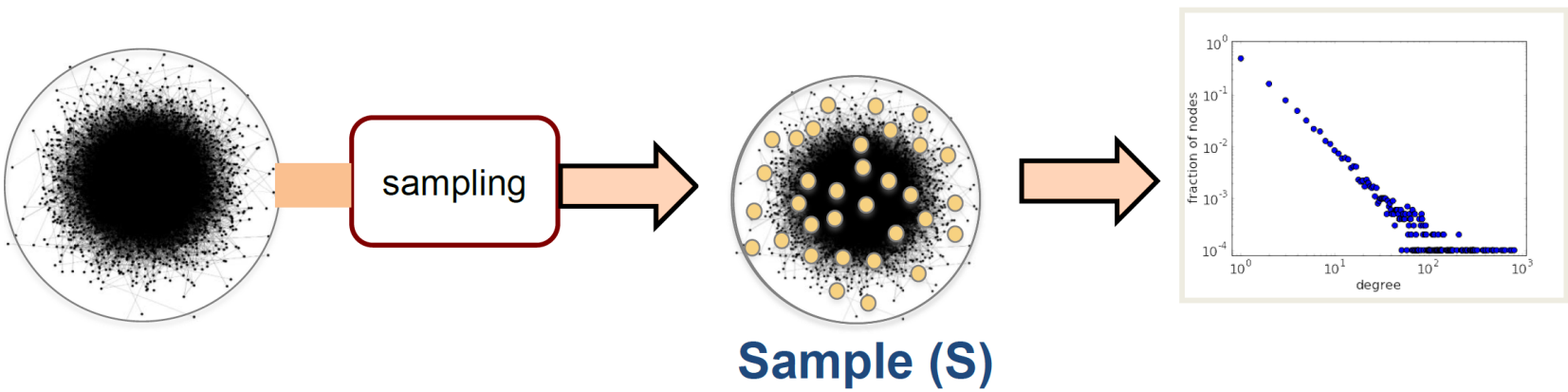
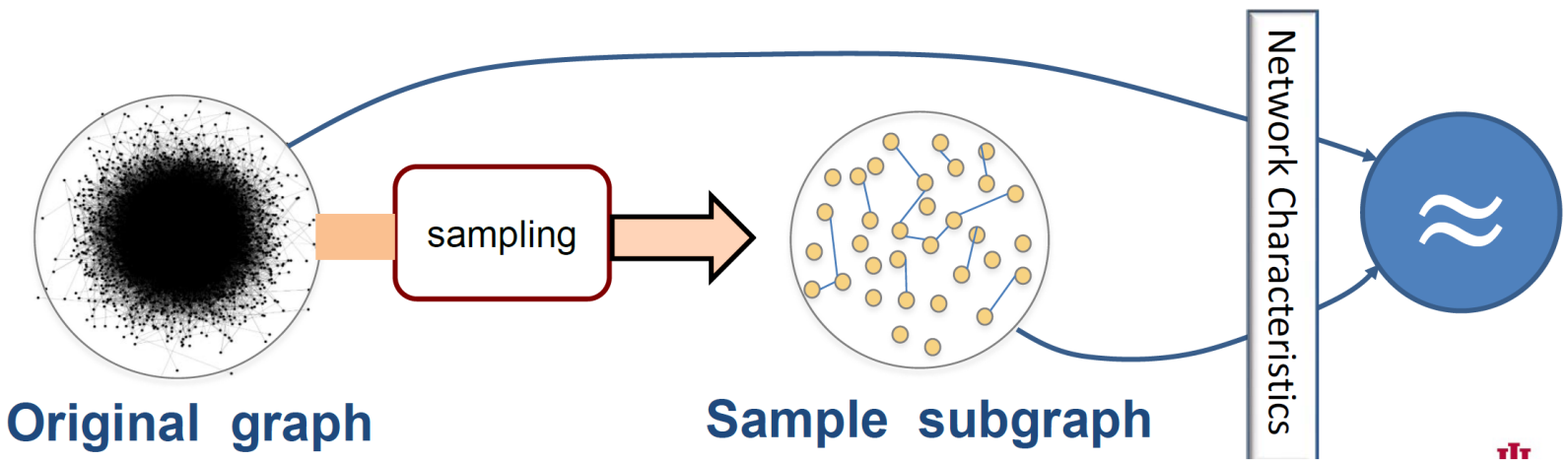


Figure: Estimating degree distribution of the original network

Sampling Methods in Social Network Studies

In Task 2,

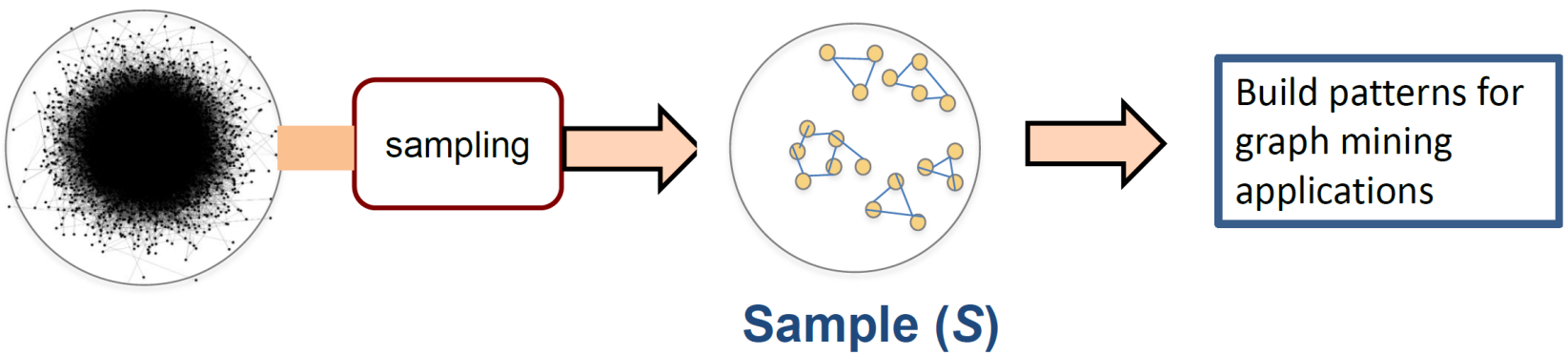
- From the full graph, researchers sample a subgraph with k nodes which preserves the value of key network characteristics, such as clustering coefficient, degree distribution, diameter, centrality, and community structure.
- The sampled network is smaller, so there is a scaling effect on some of the statistics. For example, average degree of the sampled network is smaller.



Sampling Methods in Social Network Studies

In Task 3,

- Sample sub-structures (also called motifs) of interest, such as triangles, k-stars, k-cycles for the purposes of counting, modeling, and making inferences.



Sampling Methods in Social Network Studies

Corresponding to these three sampling methods, there are two general data access assumptions:

- Full access:
  - The entire network is visible.
  - A random node or a random edge in the network can be selected.
- Restricted access:
  - The network is hidden, however, it supports crawling, i.e., it allows to explore the neighbors of a given node.
  - Access to one seed node or a collection of seed nodes are given.
  - Random walk and graph traversal sampling are examples.

Random Node (Edge) Sampling

(Uniform) random node sampling

- In random node sampling, a node is selected uniformly and independently from the set of all nodes.
- a random sample obviously has the advantage of being more representative of the population of the study if the network is fully accessible (if one node is picked, all of its network links are observed).
- It provides unbiased estimates for any nodal attributes – such as average degree and $f(u)$ where f is a function that is defined over the node attributes u.
- however, this advantage may disappear when the topology of a social network matters. This is because the random sampling does not preserve the topology of a network. For example, random node sampling would not preserve the power-law degree distribution in the population (Stumpf et al., 2005).

Random Node (Edge) Sampling

(Proportional) random node sampling (to degree or Page Rank)

- high degree nodes have higher chances to be selected.
- any nodal estimate is biased towards high degree nodes; average degree estimation is higher than actual; and degree distribution is biased towards high degree nodes.

Random Node (Edge) Sampling

(Uniform) random edge sampling

- In a random edge selection, we uniformly select a set of edges.
- As a result, a vertex will be selected in proportion to its degree.
- edge statistics is unbiased due to the uniform edge selection
- nodal statistics will be biased to high degree nodes.

Random Walk Sampling

- Random walk sampling is an example of restricted access.
- first choose a node v uniformly at random.
- for a selected node, all of its neighbors are discovered (not yet explored though).
- then choose one connected node (neighbour) of v at random and add this new node and edge into the sample.
- repeat the above step for the new chosen node. Stop until the desired sample size is reached.
- random walk sampling is memory-less, i.e., the same node can be re-visited (i.e., exploration with replacement).
- it is a simple, resource-efficient method, but may get stuck at the local network neighborhoods.
- Sample statistics from random walk sampling will be biased toward high degree nodes.

Forest-fire Sampling (Leskovec and Faloutsos, 2006)

- first choose a node v uniformly at random.
- then generate a random number x that is geometrically distributed with mean $\frac{p_f}{1-p_f}$. The parameter p_f is called the forward burning probability.
- node v selects x out-links incident to nodes that were not yet visited. Let $w_1, w_2, \cdots, w_x$ denote these connected nodes.
- we then apply this step recursively to each of $w_1, w_2, \cdots, w_x$ until enough nodes have been burned.
- nodes cannot be visited a second time, preventing the construction from cycling (i.e., exploration without replacement).
- if the fire dies, then restart it, i.e., select a new node randomly.
- forest-fire Sampling can be regarded as the probabilistic version of snowball sampling (to be introduced next).

Snowball Sampling (Goodman, 1961)

- Snowball sampling is a non-probability sampling technique where existing study subjects recruit future subjects among their acquaintances.
- Snowball sampling uses a small pool of initial informants to nominate, through their social networks, other participants who meet the eligibility criteria to potentially contribute to a specific study.
- Snowball sampling is frequently adopted by researchers in sociology and marketing to sample from connected subjects (Frenzen and Davis, 1990; Salganik and Heckathorn, 2004; Henry, 2005; Tepper, 1994).

Snowball Sampling (Goodman, 1961)

Advantages:

- locate hidden and hard-to-reach population, e.g., AIDS patients, homeless, illegal drug users.
- researchers invest less money and time in sampling.

Disadvantages:

- community bias: the first seed participants have a strong impact on the sample.
- non-random, statistical properties are hard to derived.
- samples tend to be biased toward subjects who are more cooperative and subjects with larger personal networks.

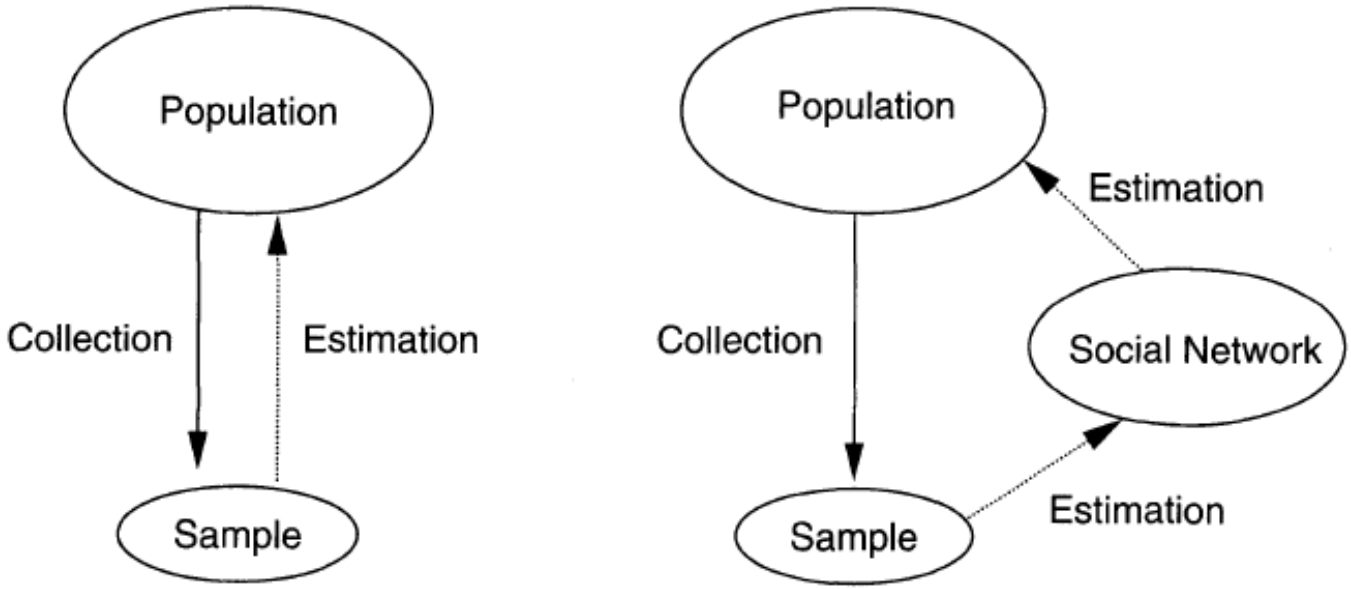
Respondent-Driven Sampling

- Respondent-driven sampling (RDS) (Heckathorn, 1997; Salganik and Heckathorn, 2004), combines "snowball sampling" with a mathematical model that **weights the sample** to compensate for the fact that the sample was collected in a non-random way.
- The main difference between snowball sampling and RDS is that RDS is mathematically tweaked to add an element of randomness.
- In other words, although it starts off as a non-probabilistic method, it ends up as a mixture between non-probabilistic and probabilistic methods.

Respondent-Driven Sampling

- To generate an RDS sample, one begins by selecting a small number of initial participants from the target population who are asked – and typically provided financial incentive – to recruit their contacts in the population.
- The sampling proceeds with current sample members recruiting the next wave of sample members, continuing until the desired sample size is reached.
- With the RDS sampling design, individuals with more contacts in the target population are more likely to be recruited.

Respondent-driven sampling



(a) Schematic of traditional sampling and estimation

(b) Schematic of respondent-driven sampling

FIGURE 1. Schematic representations showing the differences between traditional sampling and estimation and respondent-driven sampling. By not attempting to estimate directly to from the sample to the population, respondent-driven sampling avoids many of the well-known problems with estimation from a chain-referral sample.

Respondent-Driven Sampling

- To adjust for this selection bias, respondents are weighted inversely proportional to their network degree or number of contacts.
- Specifically, for any individual trait f (e.g., with AIDS, using drugs), the RDS estimate $\hat{\mu}_f$ of the population mean of f is defined to be

$$\hat{\mu}_f = \frac{1}{\sum_{i=1}^{n} 1/\text{degree}_i} \sum_{i=1}^{n} \frac{f_i}{\text{degree}_i}$$

- The accuracy of RDS estimates is affected by the structure of the underlying social network, the distribution of traits within the network, and the recruitment dynamics.
- While RDS is widely used in the public health community (in more than 20 counties) and the US CDC for AIDS Relief, Goel and Salganik (2010) use simulations to show RDS is substantially less accurate than generally acknowledged.

Problems of missing data in social networks

- The literature across many disciplines has long noticed that using sampled network data which is incomplete, i.e., some actors or links are missing from the data, may lead to estimation biases in the structural properties of networks.
- Incompleteness of sampled network data may arise from other causes: boundary specification, non-response in network survey, and survey fixed choice design.
- Kossinets (2006) uses simulation methods to examine impacts of missing links due to these causes and finds that biases of missing links on estimated network properties due to the network boundary specification and the fixed choice design are dramatic.

Problems of missing data in social networks

The boundary specification problem (Laumann et al., 1989):

- Boundary specification refers to the task of specifying inclusion rules for actors or relations in a network study.
- For example, when studying friendship network of students in a school, one can define the network boundary by classroom, by year (grade), or by school.
- When choosing inclusion rules for a network study, a researcher is in fact drawing a non-probability sample from all possible networks of its kind.
- The boundary specification problem may be avoided to a certain extent if the community is isolated from the rest of the world.

Effects of missing data in social networks

The boundary specification problem (Laumann et al., 1989)

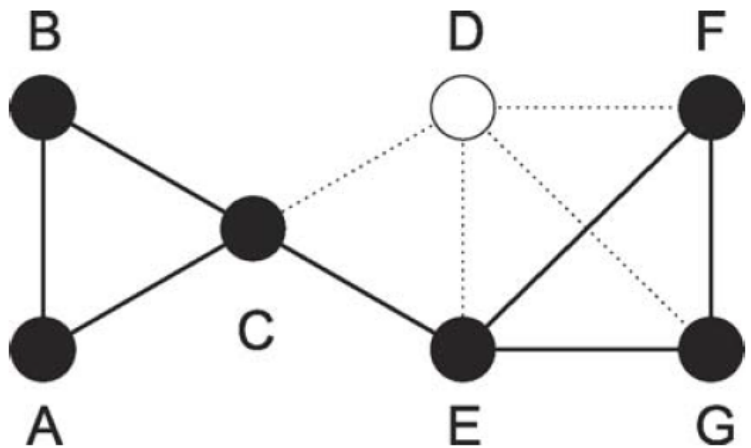


Fig. 1. Illustration of the boundary specification problem. Omission of actors may lead to significant changes in network statistics. In the above example, as a result of exclusion of actor D, the mean network degree z went down 25% from $3\frac{1}{7}$ to $2\frac{1}{3}$.

source: Kossinets (2006)

Problems of missing data in social networks

The fixed choice design problem (Holland and Leinhardt, 1973):

- A significant amount of network studies employ the "fixed choice design (FCD)" to collect information on social relationships. In FCD, a number of possible nominations that each person in the network can make is capped at a maximum M, which typically induces missing edges.
- For example, students can only nominate up to 5 male and 5 female friends in Add Health.
- FCD introduces right-censoring on vertex degree.
- Fixed choice nominations can easily lead to a non-random missing data pattern.

Effects of missing data in social networks

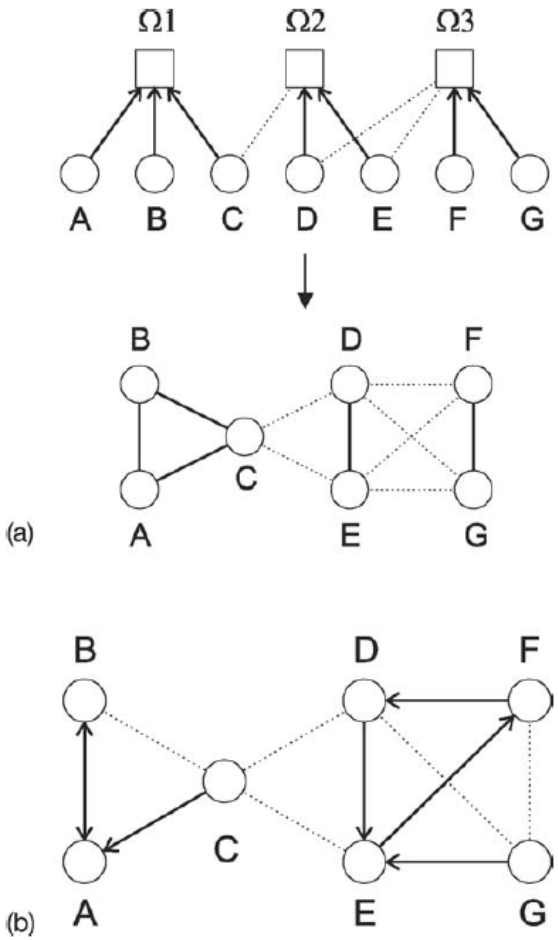


Fig. 4. Illustration of a fixed choice design. (a) Bipartite case: each actor nominates up to a fixed number K from his affiliations. Nominations are shown as arrows. (b) One-mode case: each actor nominates up to a fixed number X from his list of acquaintances. In the hypothetical example pictured above $K = X = 1$. Note that there is only one reciprocated nomination (between actors A and B).

Source: Kossinets (2006)

Problems of missing data in social networks

The survey non-response problem:

- The effects of non-response on some network properties have been described in Stork and Richards (1992) and Rumsey (1993).
- Handcock and Gile (2010), Robins et al. (2004), Koskinen et al. (2010) study the sampled network data with missing link (non-responses) in Exponential Random Graph (ERG) Modeling.
- Huisman and Steglich (2008) study the issue of survey non-response in longitudinal network data with stochastic actor-based dynamic link formation model.

Effects of missing data in social networks

Non-response problem (Stork and Richards, 1992):

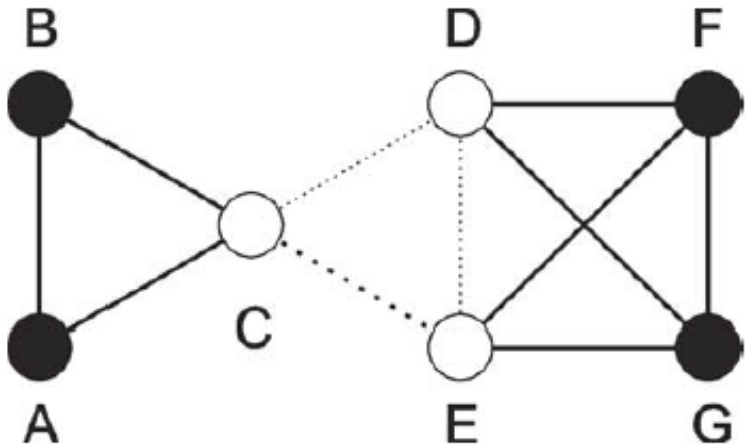


Fig. 3. Non-response in network surveys. Suppose that actors C, D and E did not report their links. However, nominations made by actors A, B, F and G help reconstruct the structure of interactions to a large extent, with a decrease in average degree less than 15%. Compare with the Boundary Specification example (Fig. 1), in which a single missing node caused a 25% deviation in the mean degree.

Source: Kossinets (2006)

Effects of missing data in social networks (Kossinets, 2006)

Table 1
Properties of the network dataset

Quantity	Notation	Cond-mat	Random[a]
Number of authors	N	16726	16726
Number of papers	M	22016	22016
Mean papers per author	μ	3.50	3.50
Mean authors per paper	ν	2.66	2.66
Assortativity (degree correlation)	r_B	−0.18	−0.054 (4)
Unipartite projection (collaborators)			
Mean degree	z	5.69	9.31 (3)
Degree variance	V	41.2	33.9 (6)
Clustering	C	0.36	0.223 (1)
Assortativity	r_U	0.18	0.071 (5)
Number of components	N_C	1188	652 (18)
Size of largest component	S_L	13861	16064 (18)
Mean path in largest component	ℓ_L	6.63	4.728 (8)

[a] A random bipartite graph of the same size and mean degree as the original network. Numbers in parentheses are standard deviations on the least significant figures calculated in an ensemble of 100 such graphs.

Effects of missing data in social networks (Kossinets, 2006)

Table 2
Simulation algorithms for sensitivity analysis

Label	Problem	Model[a]
BSPC	Boundary specification problem for contexts	Remove a fraction of contexts at random
BSPA	Boundary specification problem for actors	Remove a fraction of actors at random
NRE	Non-response effect	Remove links within subgraph induced by a specified fraction of actors
FCC	Fixed choice (contexts)	Apply censoring by degree to actors
FCA	Fixed choice (actors)	Create unipartite projection; apply censoring by degree; keep non-reciprocated links
FCR	Fixed choice (actors), reciprocated nominations only	Create unipartite projection; apply censoring by degree; keep only reciprocated links

[a] We measure properties of the unipartite projection in all models.

Effects of missing data in social networks (Kossinets, 2006)

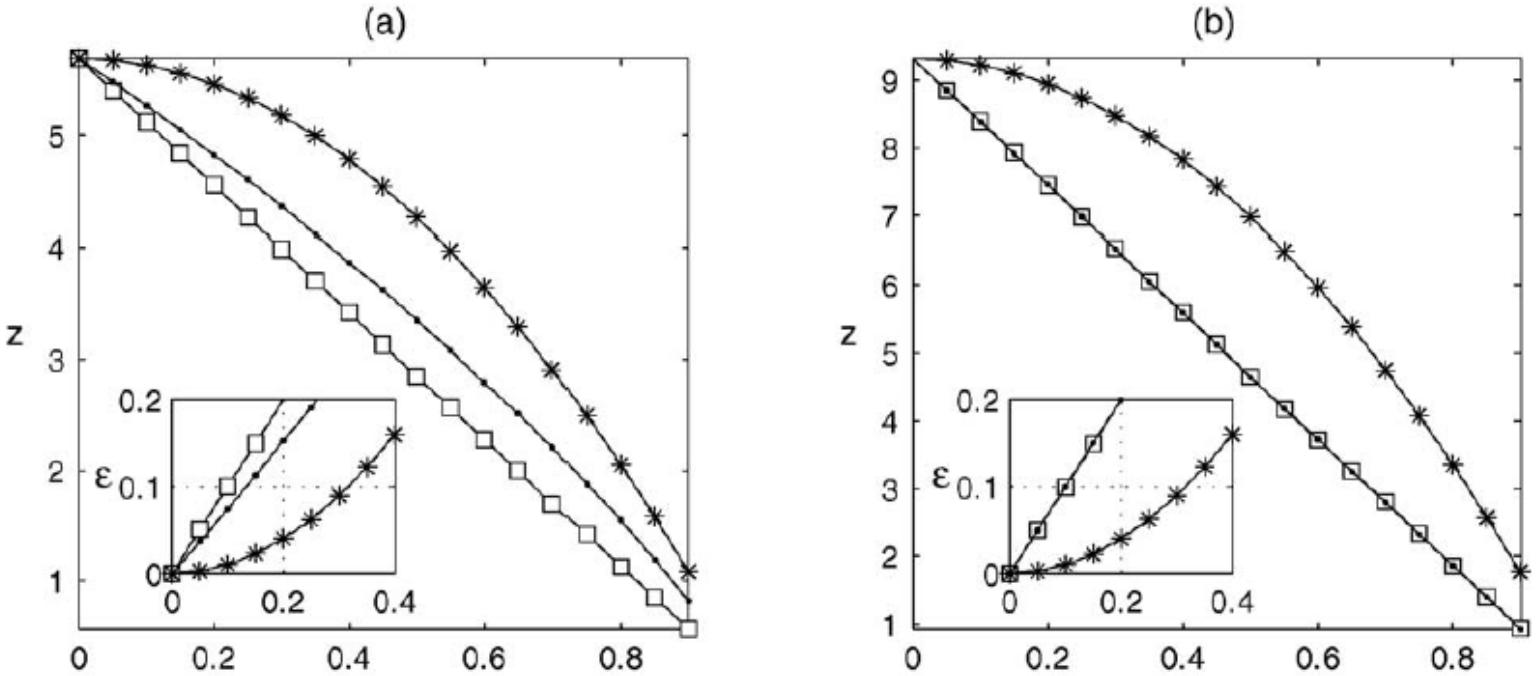


Fig. 6. Sensitivity of mean vertex degree in the unipartite projection z to different missing data mechanisms: (a) in the Condensed Matter graph; (b) in a bipartite random graph. Dots: boundary specification (non-inclusion) effect for interaction contexts (BSPC); the horizontal axis corresponds to the fraction of papers missing from the database. Squares: non-inclusion effect for actors (BSPA) with the x-axis corresponding to the fraction of authors missing from the database. Note that in panel (b) dots overlap with squares. Stars: simulation of survey non-response among authors (NRE); vertices are assumed non-responding at random. The x-axis indicates the fraction of non-respondents. Insets: relative error $\varepsilon = |z - z_0|/z_0$, where z_0 is the true value. Each data point is an average over 50 iterations. Lines connecting datapoints are meant as a guide for the eye.

Effects of missing data in social networks (Kossinets, 2006)

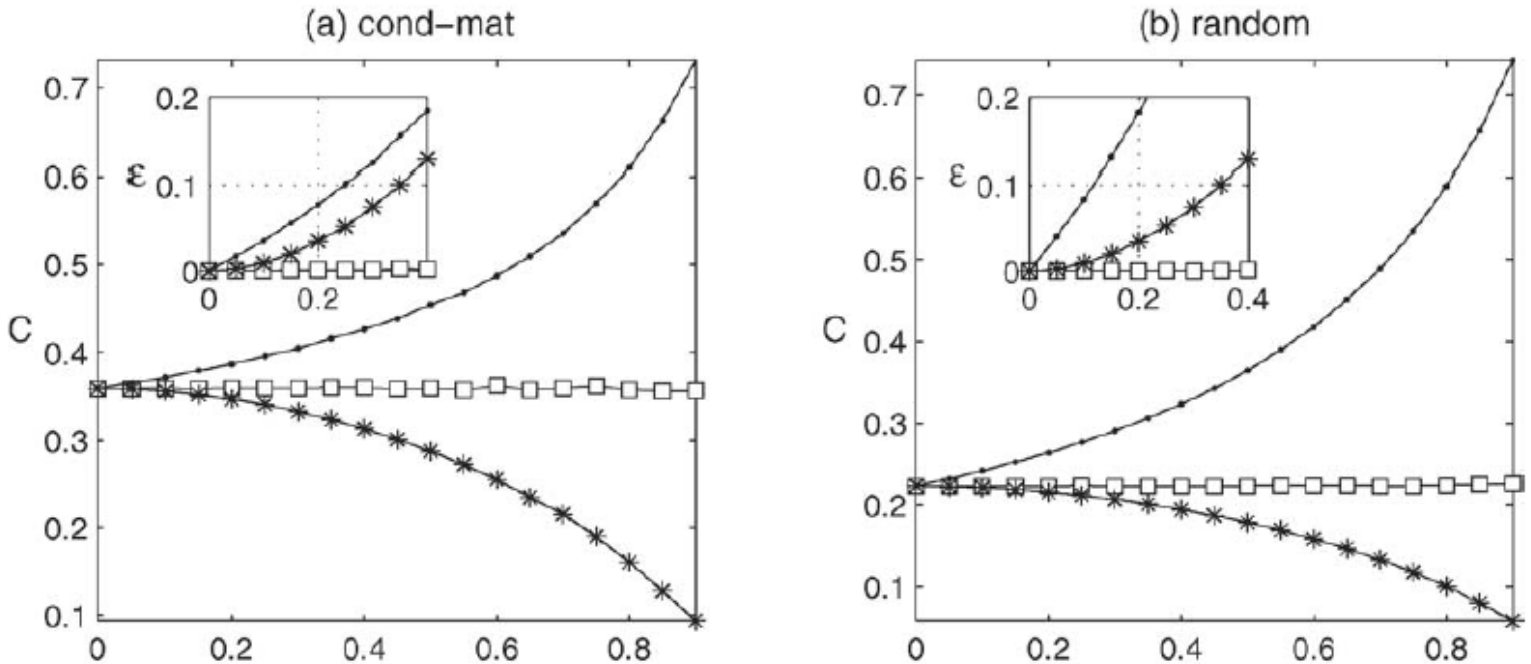


Fig. 8. Sensitivity of clustering C in the unipartite projection: omission of interaction contexts (dots); omission of actors (squares); survey non-response (stars).

Effects of missing data in social networks (Kossinets, 2006)

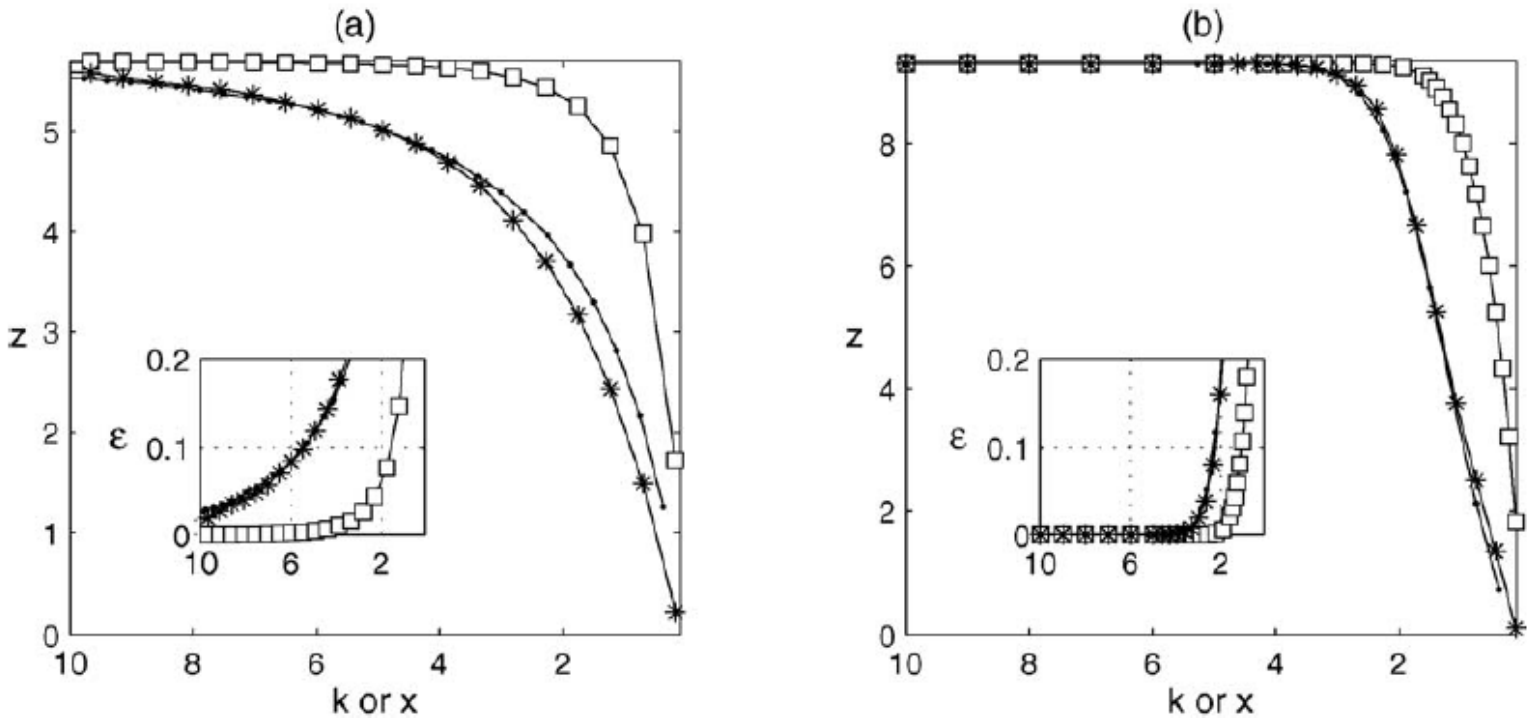


Fig. 12. Fixed choice effect on the mean degree of the unipartite projection z in the Condensed Matter collaboration graph (a) and a comparable random graph (b). Dots: censoring collaborations. The question asked of each author would be to "nominate" up to K papers coauthored by him. The horizontal axis represents the relative degree cutoff $k = K/\mu$, where $\mu = 3.5$ is the mean number of affiliations per actor. Note that the amount of missing data increases as we lower the threshold value. For example, $k = 5$ means that the actual cutoff is $K = 5\mu$, five times the mean actor degree in the bipartite network. Squares: censoring coauthors, no reciprocation required. The question asked of each author would be to nominate up to X coauthors. The horizontal axis represents relative degree cutoff $x = X/z$ in units of z, the mean number of collaborators per author, where (a) $z = 5.69$ in the Physics collaboration graph and (b) $z = 9.31$ in a random network. Stars: only reciprocated nominations, relative cutoff $x = X/z$ in units of z. Insets: relative error $\varepsilon = |z - z_0|/z_0$, where z_0 is the true value. Each data point is an average over 50 iterations. Lines connecting datapoints are a guide for the eye only.

The Impact of Sampling and Network Topology on the Estimation of Social Inter-correlations

- Chen et al. (2013) use simulations to study the impact of sampling networks on the estimation of social inter-correlation.
- They capture social inter-correlation by the following spatial error model (SEM):

$$y_{it} = \alpha + \beta x_i + \theta_{it}$$
$$\theta_{it} = \rho \sum_j W_{ij}\theta_{jt} + u_{it}.$$

- They find the magnitude of social inter-correlations in consumer networks tends to be underestimated if samples of the networks are used (as opposed to using the entire population of the network).
- The sampling methods that better preserve the network structure (such as snowball sampling and forest-fire sampling with high forward-burning probability) perform better in recovering the social inter-correlations.

Statistical Correction Methods for Sampled Networks

- Based on Hsieh et al. (2018a) "Unequally Sampled Networks: Biases and Corrections".
- In a conventional sampling (identically and independently) of variable y, the sample average from the sample S of size m, $\hat{\mu} = \frac{1}{m}\sum_{i \in S} y_i$, is an unbiased estimator of population mean μ_y, i.e., $\mathrm{E}(\hat{\mu}) = \mu_y$, and $\mathrm{Var}(\hat{\mu}) = \sigma_y^2/m$, where σ_y^2 is the population variance of y.
- However, if the sampling probability is not identical, the sample average may be a biased estimator for the population mean.
- In this case, the Horvitz-Thomptson (H-T) estimator (Horvitz and Thompson, 1952) can be applied to alleviate the problem.

Statistical Correction Methods for Sampled Networks

Consider random sampling without replacement:

- denote $\psi_i = \mathrm{E}[\mathcal{I}(i \in S)] = \Pr(i \in S)$ the sampling probability of unit i, where $\mathcal{I}(\cdot)$ is an indicator function.
- there are $\binom{n}{m}$ possible samples of size m that may be chosen from the population P of size n, and $\binom{n-1}{m-1}$ possible samples that may be chosen to include a given unit i, so it follows that $\psi_i = \frac{\binom{n-1}{m-1}}{\binom{n}{m}} = \frac{m}{n}$.

$$\mathrm{E}(\hat{\mu}|P) = \mathrm{E}\left(\frac{1}{m}\sum_{i\in S} y_i \middle| P\right) = \mathrm{E}\left(\frac{1}{m}\sum_{i\in P} y_i \mathcal{I}(i \in S)\right)$$
$$= \frac{1}{m}\sum_{i\in P} y_i \mathrm{E}[\mathcal{I}(i \in S)] = \frac{1}{m}\sum_{i\in P} y_i \psi_i = \frac{1}{n}\sum_{i\in P} y_i$$

- in this case, $\hat{\mu}$ is unbiased. However, in other general cases where ψ_i is related to i's characteristics, $\psi_i \neq \frac{m}{n}$, $\hat{\mu}$ would be biased.

Statistical Correction Methods for Sampled Networks

- The Horvitz-Thompson (H-T) estimator is a kind of weighted average using the inverse of the inclusion probability as the weight and the population size n for normalization.

$$\tilde{\mu}_{\psi} = \frac{1}{n} \sum_{i \in S} \frac{y_i}{\psi_i}$$

- How does it work?

$$\begin{aligned} \mathrm{E}(\tilde{\mu}_{\psi}) &= \mathrm{E}\left(\frac{1}{n} \sum_{i \in S} \frac{y_i}{\psi_i} \right) = \mathrm{E}\left(\frac{1}{n} \sum_{i \in P} \frac{y_i}{\psi_i} \mathcal{I}(i \in S) \right) \\ &= \frac{1}{n} \sum_{i \in P} \frac{y_i}{\psi_i} \mathrm{E}[\mathcal{I}(i \in S)] = \frac{1}{n} \sum_{i \in P} y_i = \mu. \end{aligned}$$

Statistical Correction Methods for Sampled Networks

- Chandrasekhar and Lewis (2011) (henceforth, CL) apply the H-T estimator to correct the sampling bias in network data. They assume simple random sampling in networks.
- A population network (graph) is a pair $G = (V, E)$ of a set of nodes V and edges E. Denote $n = |V|$ the size of population network. The network is represented with an $n \times n$ adjacency matrix $W(G)$.
- Hsieh et al. (2018a) extend the correction to the unequal network sampling.
- Assume the population can be classified into T disjoint types with a generic type $t \in \{1, 2, ..., T\}$. Let V_t be the set of nodes of type t, $n_t = |V_t|$ is the size of subpopulation t and $\sum_t n_t = n$.
- We write $t_i = t$ if individual i is of type t. Then, $t_i = t_j$ ($t_i \neq t_j$) indicates that i and j are (not) of the same type.

Statistical Correction Methods for Sampled Networks

- Researchers only observe the sampled network. Let S be the set of sampled nodes of size $m = |S| = |\psi n|$, $\psi = \frac{m}{n}$ be the sampling rate. Analogously, $m_t = \psi_t n_t$ is the sampled number of individuals of type t and ψ_t is type t's sampling rate.
- We assume that the sampling is unequal in the following sense: the researcher observes $m_t \leq n_t$ individuals of type t, with $m_t = \psi_t n_t$ and $\sum_t m_t = m$. If $\psi_t = \psi$ for all $t \in T$, the sampling is simple random as in CL.
- Conditional on the type, sampling is simple random at the individual level.

Statistical Correction Methods for Sampled Networks

- There are two types of sampled networks. The first one is the *induced subgraph* denoted by $G^{|s}$. The induced graph restricts the network links among the m sampled nodes. The second sample scheme considered is the *star subgraph*, denoted by G^s.

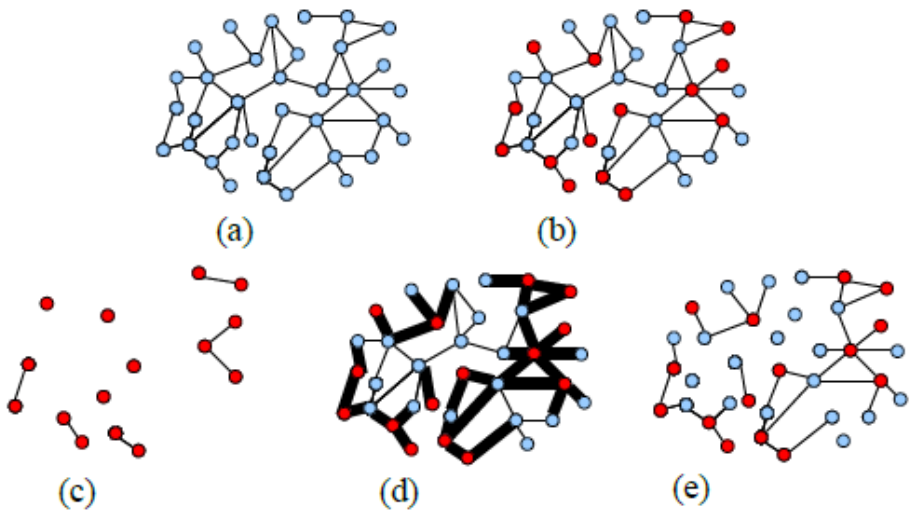


FIGURE 1. (a) $G = (V, E)$. (b) the sampled nodes, S, in red. (c) the induced subgraph, $G^{|S} = (S, E^{|S})$. (d) G, highlighting the sampled nodes and the edges that are induced if each sampled node reports all of its links from the census, E^S. (e) the star subgraph, G^S.

Source: Chandrasekhar and Lewis (2011)

Statistical Correction Methods for Sampled Networks

We use the average network degree as the example to show our correction approach:

- the average degree of population graph G is simply the average number of network links per node in the network, defined as $d(G) = \frac{1}{n}\sum_{i\in V}\sum_{j\in V} W_{ij}$.
- for the induced subnetwork, $d(G^{|s}) = \frac{1}{m}\sum_{i\in S}\sum_{j\in S} W_{ij}^{|s}$. Hence,

$$\begin{aligned}\mathrm{E}(d(G^{|s})|G) &= \mathrm{E}\left(\frac{1}{m}\sum_{i\in S}\sum_{j\in S} W_{ij}^{|s}\,\middle|\, G\right)\\ &= \frac{1}{m}\sum_{i\in V}\sum_{j\in V} W_{ij}\mathrm{E}(\mathcal{I}(i,j\in S)|G).\end{aligned}$$

Statistical Correction Methods for Sampled Networks

- In the case of simple random network sampling,
$$\mathrm{E}(\mathcal{I}(i,j \in S)|G) = \frac{\binom{n-2}{m-2}}{\binom{n}{m}} = \frac{m(m-1)}{n(n-1)} = \frac{m}{n}\left(\frac{m}{n} + o(1)\right) = \psi(\psi + o(1)).$$
$$\mathrm{E}(d(G^{|s})|G) = \frac{1}{m}\sum_{i\in V}\sum_{j\in V} W_{ij}\psi\left(\psi + o(1)\right)$$
$$= \frac{1}{n}\sum_{i\in V}\sum_{j\in V} W_{ij}\left(\psi + o(1)\right) \neq d(G),$$
i.e., $d(G^{|s})$ is a downward biased estimator of $d(G)$.
- According to H-T estimator, we can use the inverse of the inclusion probability (ψ^2) as weight and change the sample average degree estimator to $\tilde{d}(G^{|s}) = \frac{1}{n}\sum_{i\in S}\sum_{j\in S}\frac{W_{ij}^{|s}}{\psi^2}$ and show $\tilde{d}(G^{|s})$ is asymptotically unbiased.
- However, when the sampling probability is unequal, i.e., different types have different sampling rates, the above H-T estimator based on ψ is still biased.

Statistical Correction Methods for Sampled Networks

- In a case that $T = 2$, i.e., individuals are of two types,

$$\mathrm{E}(d(G^{|s})|G) = \mathrm{E}\left(\frac{1}{m} \sum_{i,j \in S} W_{ij}^{|s} \middle| G \right)$$

$$= \mathrm{E}\left(\frac{1}{m} \sum_{\substack{i,j \in S \\ t_i \neq t_j}} W_{ij}^{|s} \middle| G \right) + \mathrm{E}\left(\frac{1}{m} \sum_{\substack{i,j \in S \\ t_i = t_j = 1}} W_{ij}^{|s} \middle| G \right) + \mathrm{E}\left(\frac{1}{m} \sum_{\substack{i,j \in S \\ t_i = t_j = 2}} W_{ij}^{|s} \middle| G \right)$$

$$= \left(\frac{1}{m} \cdot \sum_{\substack{i,j \in V \\ t_i \neq t_j}} W_{ij} \left(\frac{\binom{n_1 - 1}{m_1 - 1}\binom{n_2 - 1}{m_2 - 1}}{\binom{n_1}{m_1}\binom{n_2}{m_2}} \right) \right) + \left(\frac{1}{m} \cdot \sum_{\substack{i,j \in V \\ t_i = t_j = 1}} W_{ij} \left(\frac{\binom{n_1 - 2}{m_1 - 2}\binom{n_2}{m_2}}{\binom{n_1}{m_1}\binom{n_2}{m_2}} \right) \right)$$

$$+ \left(\frac{1}{m} \cdot \sum_{\substack{i,j \in V \\ t_i = t_j = 2}} W_{ij} \left(\frac{\binom{n_2 - 2}{m_2 - 2}\binom{n_1}{m_1}}{\binom{n_1}{m_1}\binom{n_2}{m_2}} \right) \right)$$

$$= \frac{1}{m} \left(\psi_1 \psi_2 \sum_{\substack{ij \in V \\ t_i \neq t_j}} W_{ij} \right) + \frac{1}{m} \left(\psi_1 \left(\psi_1 + o(1) \right) \sum_{\substack{i,j \in V \\ t_i = t_j = 1}} W_{ij} \right) + \frac{1}{m} \left(\psi_2 \left(\psi_2 + o(1) \right) \sum_{\substack{i,j \in V \\ t_i = t_j = 2}} W_{ij} \right).$$

Statistical Correction Methods for Sampled Networks

- The corresponding H-T estimator $\tilde{d}(G^{|s})$ is

$$\tilde{d}(G^{|s}) = \frac{1}{n}\left((\psi_1\psi_2)^{-1}\sum_{\substack{ij\in S\\ t_i\neq t_j}} W_{ij}^{|s}\right) + \frac{1}{n}\left((\psi_1\psi_1)^{-1}\sum_{\substack{i,j\in S\\ t_i=t_j=1}} W_{ij}^{|s}\right) + \frac{1}{n}\left((\psi_2\psi_2)^{-1}\sum_{\substack{i,j\in S\\ t_i=t_j=2}} W_{ij}^{|s}\right)$$

  and it is a asymptotically unbiased estimator for $d(G)$.
- In a general case that $T > 2$,

$$\tilde{d}(G^{|s}) = \frac{1}{n}\sum_{t=1}^{T}\left((\psi_t\psi_t)^{-1}\sum_{\substack{i,j\in S\\ t_i=t_j=t}} W_{ij}^{|s}\right) + \frac{1}{n}\sum_{t=1}^{T}\sum_{\ell\neq t}^{T}\left((\psi_t\psi_\ell)^{-1}\sum_{\substack{i,j\in S\\ t_i=t, t_j=\ell}} W_{ij}^{|s}\right).$$

- Our proposed H-T estimator is based on the post-stratification weights.

Statistical Correction Methods for Sampled Networks

- For the star subgraph, $d(G^s) = \frac{1}{m'} \sum_{i \in S} \sum_{j \in S} W^s_{ij}$. Hence,

$$\mathrm{E}(d(G^s)|G) = \mathrm{E}\left(\frac{1}{m'} \sum_{i \in S} \sum_{j \in S} W^s_{ij} \middle| G \right) = \frac{1}{m'} \sum_{i \in V} \sum_{j \in V} W_{ij} \mathrm{E}(\mathcal{I}(i, j \in S)|G).$$

- In the case of simple random network sampling,
$\mathrm{E}(\mathcal{I}(i, j \in S)|G) = \frac{2\binom{n-1}{m-1} - \binom{n-2}{m-2}}{\binom{n}{m}} = \psi^2 + 2\psi(1-\psi) + o(1).$
- $\mathrm{E}(d(G^s)|G) = \frac{1}{m'} \sum_{i \in V} \sum_{j \in V} W_{ij} \left(\psi^2 + 2\psi(1-\psi) + o(1)\right).$
- According to H-T estimator, we can use the inverse of the inclusion probability $\left(\psi^2 + 2\psi(1-\psi)\right)$ as weight and change the sample average degree estimator to
$\tilde{d}(G^s) = \frac{1}{n} \sum_{i \in S} \sum_{j \in S} \left(\psi^2 + 2\psi(1-\psi)\right)^{-1} W^s_{ij}$ and show $\tilde{d}(G^s)$ is asymptotically unbiased.

Statistical Correction Methods for Sampled Networks

- However, if the sampling is unequal, i.e., different types have different sampling rates, the above H-T estimator based on ψ will be still biased.
- We can show, in a case that $T = 2$, i.e., individuals are in two types,

$$\mathrm{E}(d(G^s)|G) = \mathrm{E}\left(\frac{1}{m'} \sum_{i,j \in S} W_{ij}^s \middle| G \right)$$

$$= \frac{1}{m'} \left(\psi_1 \psi_2 \left(\frac{1}{\psi_2} + \frac{1}{\psi_1} - 1 \right) \sum_{\substack{i,j \in V \\ t_i=1, t_j=2}} W_{ij} \right) + \frac{1}{m'} \left(\psi_1 \left(2 - \psi_1 + o(1)\right) \sum_{\substack{i,j \in V \\ t_i=t_j=1}} W_{ij} \right)$$

$$+ \frac{1}{m'} \left(\psi_2 \left(2 - \psi_2 + o(1)\right) \sum_{\substack{i,j \in V \\ t_i=t_j=2}} W_{ij} \right).$$

Statistical Correction Methods for Sampled Networks

- The corresponding asymptotically unbiased estimator for $d(G)$ is

$$\tilde{d}(G^s) = \frac{1}{n}\left(\left(\psi_1\psi_2\left(\frac{1}{\psi_2}+\frac{1}{\psi_1}-1\right)\right)^{-1}\sum_{\substack{i,j\in S\\ t_i=1,t_j=2}} W^s_{ij}\right) + \frac{1}{n}\left(\left(\psi_1\left(2-\psi_1\right)\right)^{-1}\sum_{\substack{i,j\in S\\ t_i=t_j=1}} W^s_{ij}\right)$$
$$+ \frac{1}{n}\left(\left(\psi_2\left(2-\psi_2\right)\right)^{-1}\sum_{\substack{i,j\in S\\ t_i=t_j=2}} W^s_{ij}\right).$$

- In a general case that $T > 2$,

$$\tilde{d}(G^s) = \frac{1}{n}\sum_{t=1}^{T}\left(\left(\psi_t^2+2\psi_t(1-\psi_t)\right)^{-1}\sum_{\substack{i,j\in S\\ t_i=t_j=t}} W^s_{ij}\right)$$
$$+ \frac{1}{n}\sum_{t=1}^{T}\sum_{\ell\neq t}\left(\left(\psi_t\psi_\ell+\psi_\ell(1-\psi_t)+\psi_t(1-\psi_\ell)\right)^{-1}\sum_{\substack{i,j\in S\\ t_i=t,t_j=\ell}} W^s_{ij}\right).$$

- We can similarly correct for other sample network statistics, e.g., clustering coefficient, graph span, epidemic threshold, homophily index, etc.

Statistical Correction Methods for Sampled Networks

We conduct a Monte Carlo simulation study to examine the performance of the proposed correction approach:

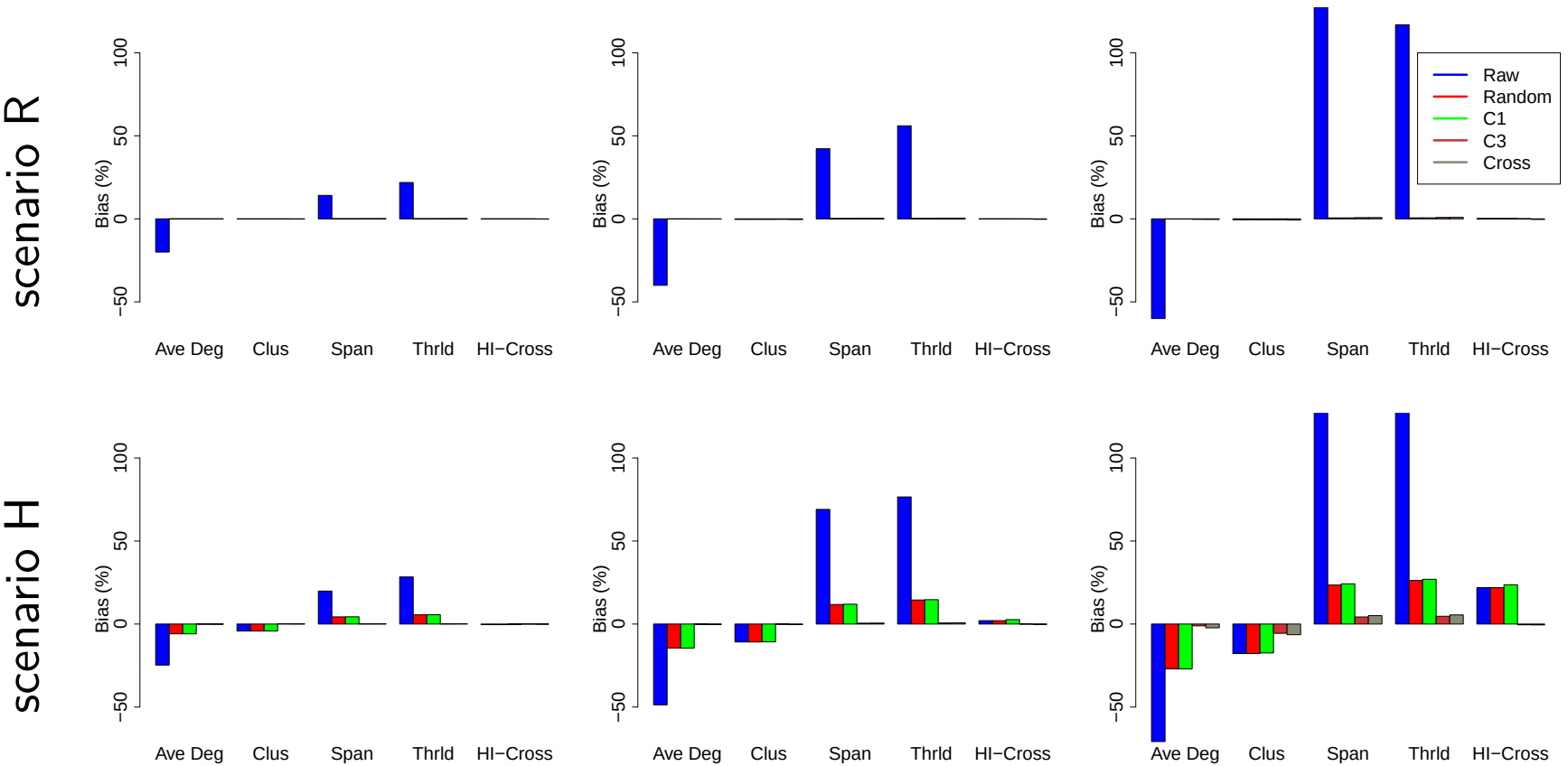
- We take one large Add Health school as a prototype and slightly prune it to be a network population (size: 1500).
- We look at three demographic variables, including senior dummy (C1), gender (C2), and race (C3), to form individual types.
- For C1, seniority equals 1 if older than the school population average and 0 otherwise. For C2, we have male and female. For C3, we have White, Black, and other races.
- Strong correlation between C3 and network connectivity: average degree in each group of C3 is 9.6060, 7.3800, and 4.3960.
- Other than weighting on each characteristics separately, we can also combine these three characteristics to from $2 \times 2 \times 3$ cross-characteristics, denoted by *Cross*.

Statistical Correction Methods for Sampled Networks

We conduct a Monte Carlo simulation to examine the performance of the correction approach:

- Two dimensions of removing strategies to generate artificial network samples:
  - remove 20%, 40%, and 60% of nodes.
  - (i) Scenario R: Random removal, (ii) Scenario H: Remove more high degree nodes, i.e., whites, (iii) Scenario M: Remove more intermediate degree nodes, i.e., blacks, (iv) Scenario L: Remove more low degree nodes, i.e., other races.
- We calculate five sample network statistics (average degree, clustering coefficient, graph span, epidemic threshold, homophily index) based on (i) raw sample (ii) weights based on simple random sampling assumption (iii) post-stratified weights by $C1$ (iv) post-stratified weights by $C3$ (v) post-stratified weights by *Cross*.
- 1000 repetitions of each constellation. Average of biases reported.

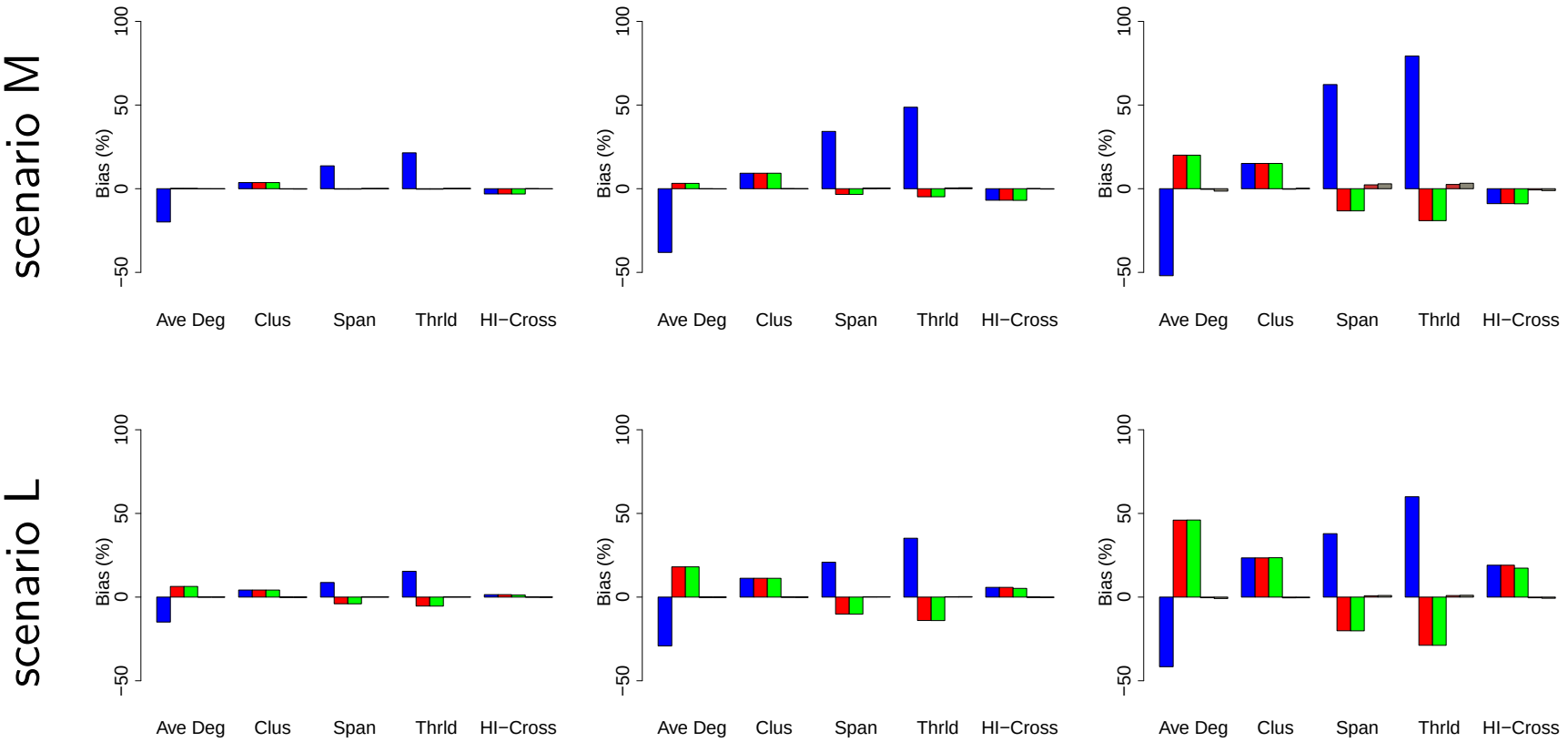
Statistical Correction Methods for Sampled Networks



(a) 20% (b) 40% (c) 60%

Figure: Scenarios R and H on induced subgraph

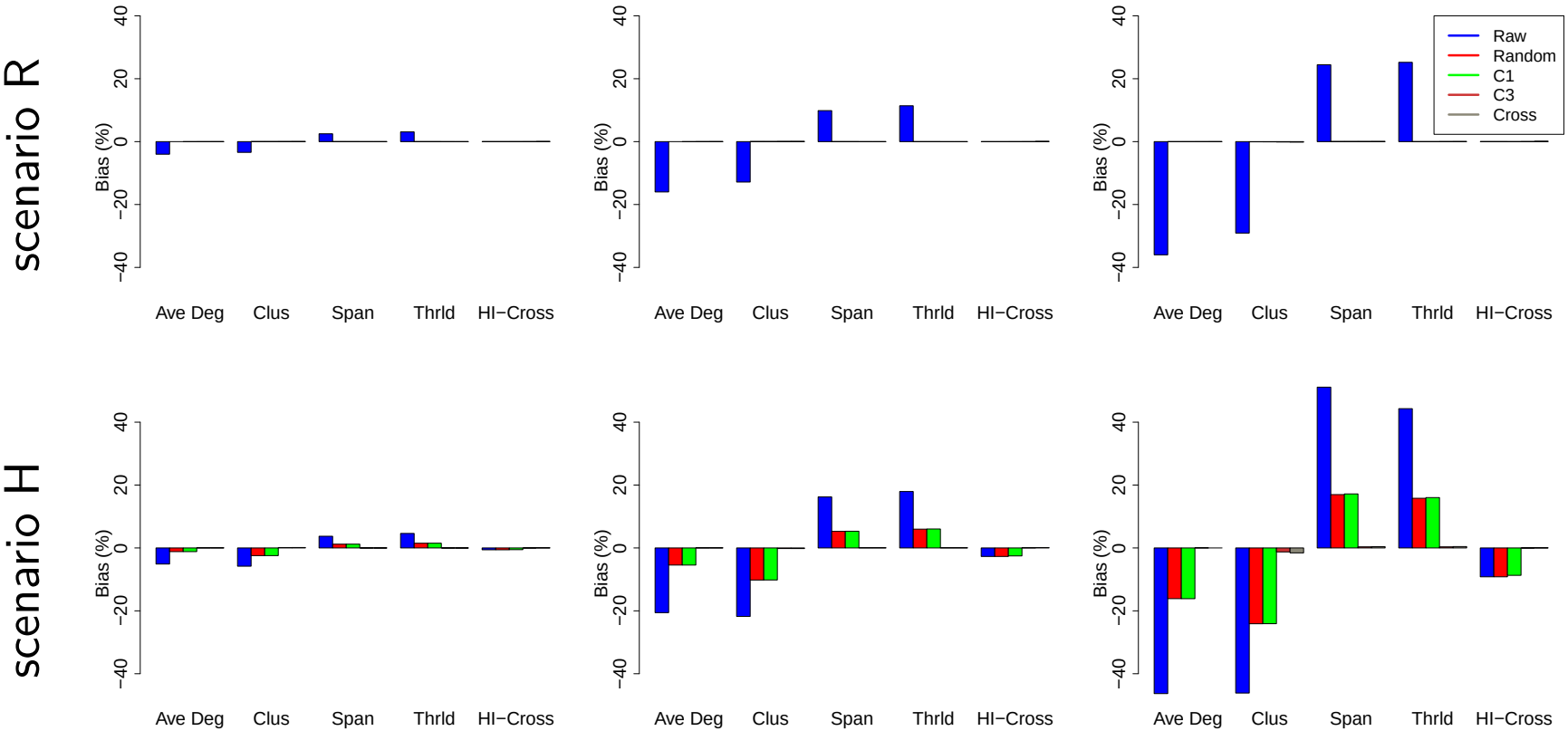
Statistical Correction Methods for Sampled Networks



(a) 20% (b) 40% (c) 60%

Figure: Scenarios M and L on induced subgraph

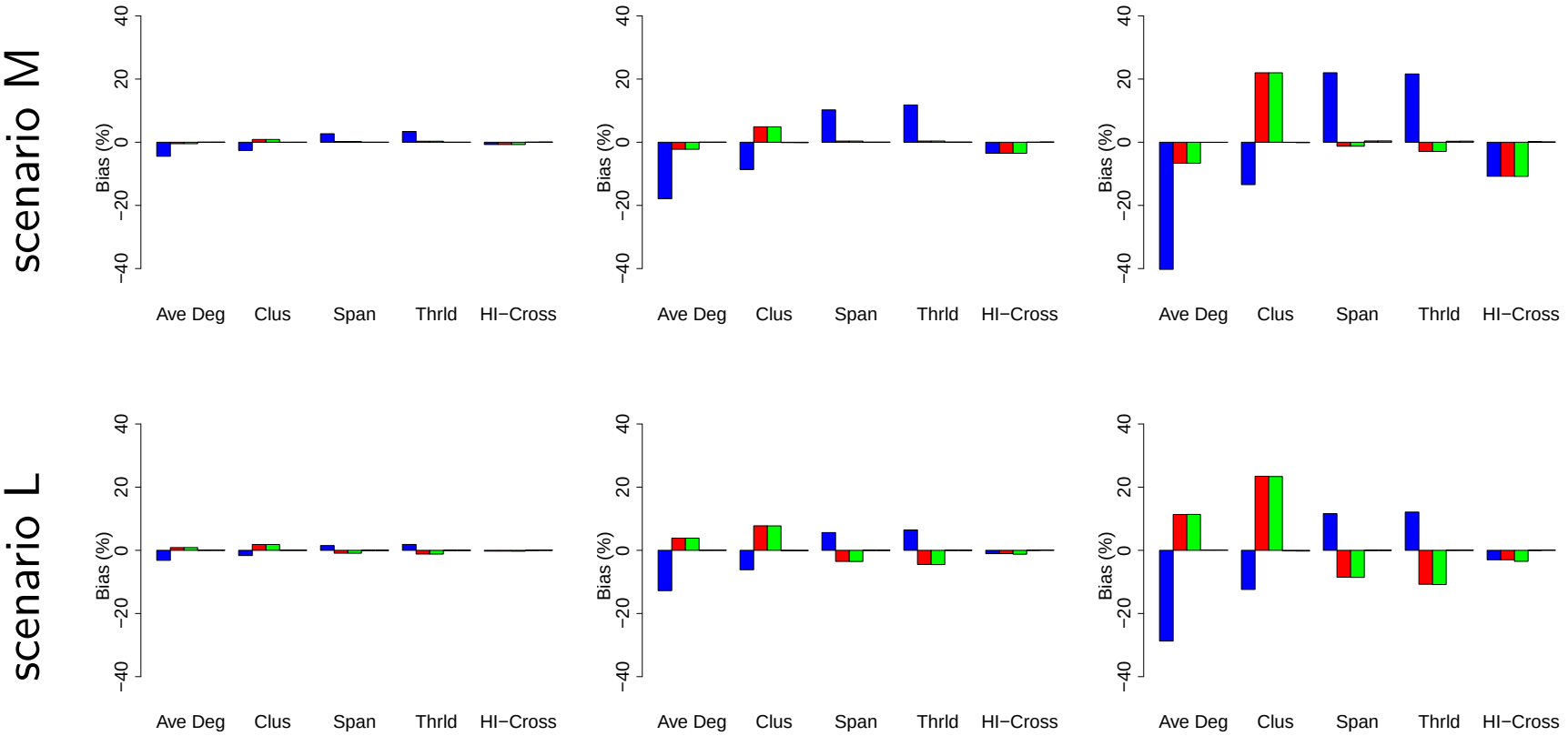
Statistical Correction Methods for Sampled Networks



(a) 20% (b) 40% (c) 60%

Figure: Scenarios R and H on star subgraph

Statistical Correction Methods for Sampled Networks



(a) 20% (b) 40% (c) 60%

Figure: Scenarios M and L on star subgraph

Statistical Correction Methods for Sampled Networks

- We have shown that our correction approach can remedy the biases in network statistics when network sampling is non-random.
- Next we look at the bias of network effects in regression

$$y_r = \alpha + \Lambda(G_r)\beta + \gamma x_r + \epsilon_r,$$

where y_r is the outcome variable of population or network r, x_r is the set of network-level controls, and $\Lambda(G_r)$ is the *true* network property (or properties) of interest.

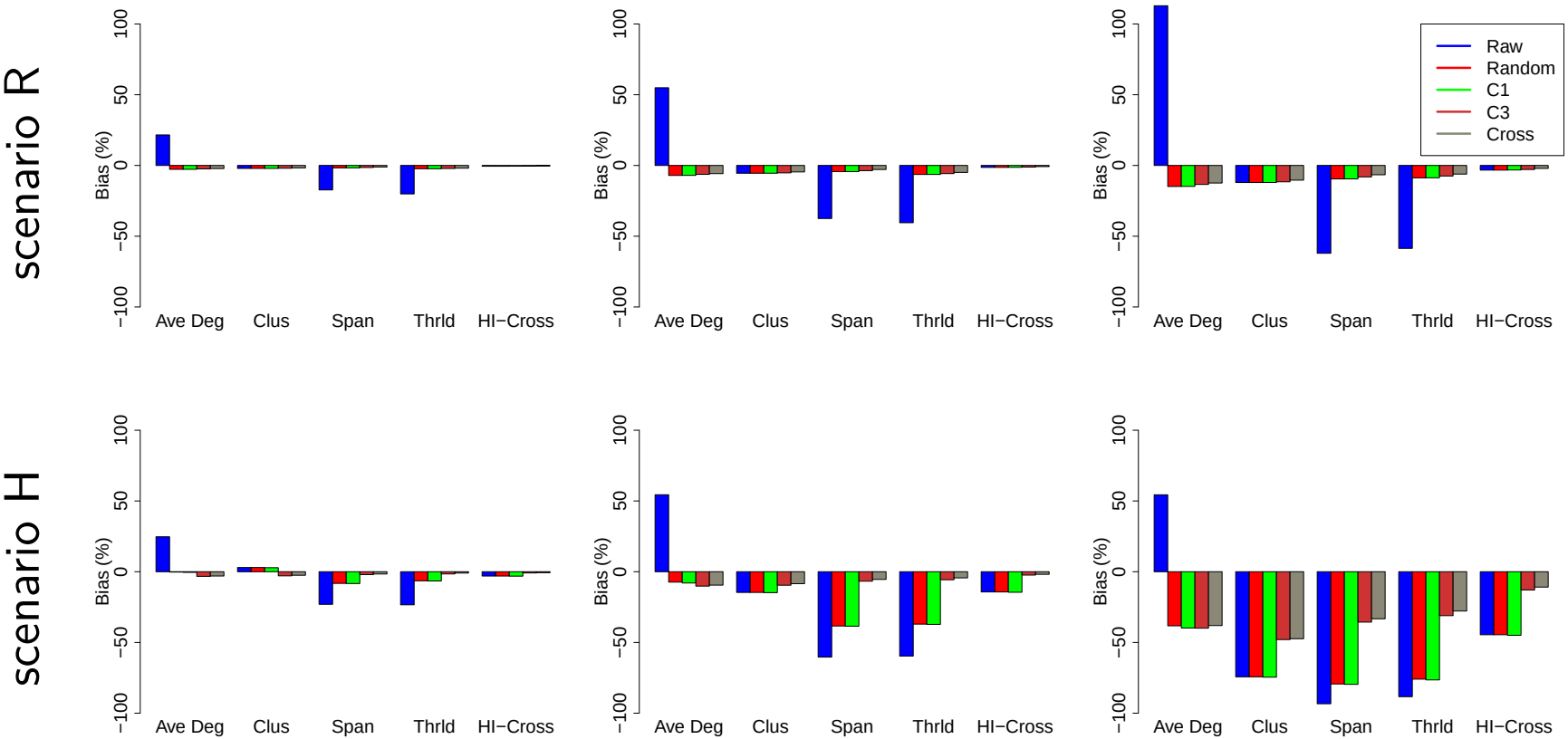
- With sampled data on the network, the researchers observe $\overline{G}_r \in \{G_r^S, G_r^{|S}\}$. Therefore, the scholars typically estimate

$$y_r = \alpha + \Lambda(\overline{G}_r)\beta + \gamma x_r + \epsilon_r,$$

which leads to the bias on $\hat{\beta}$.

- We also conduct a simulation study to show the bias on $\hat{\beta}$ can be alleviated by our correction approach.

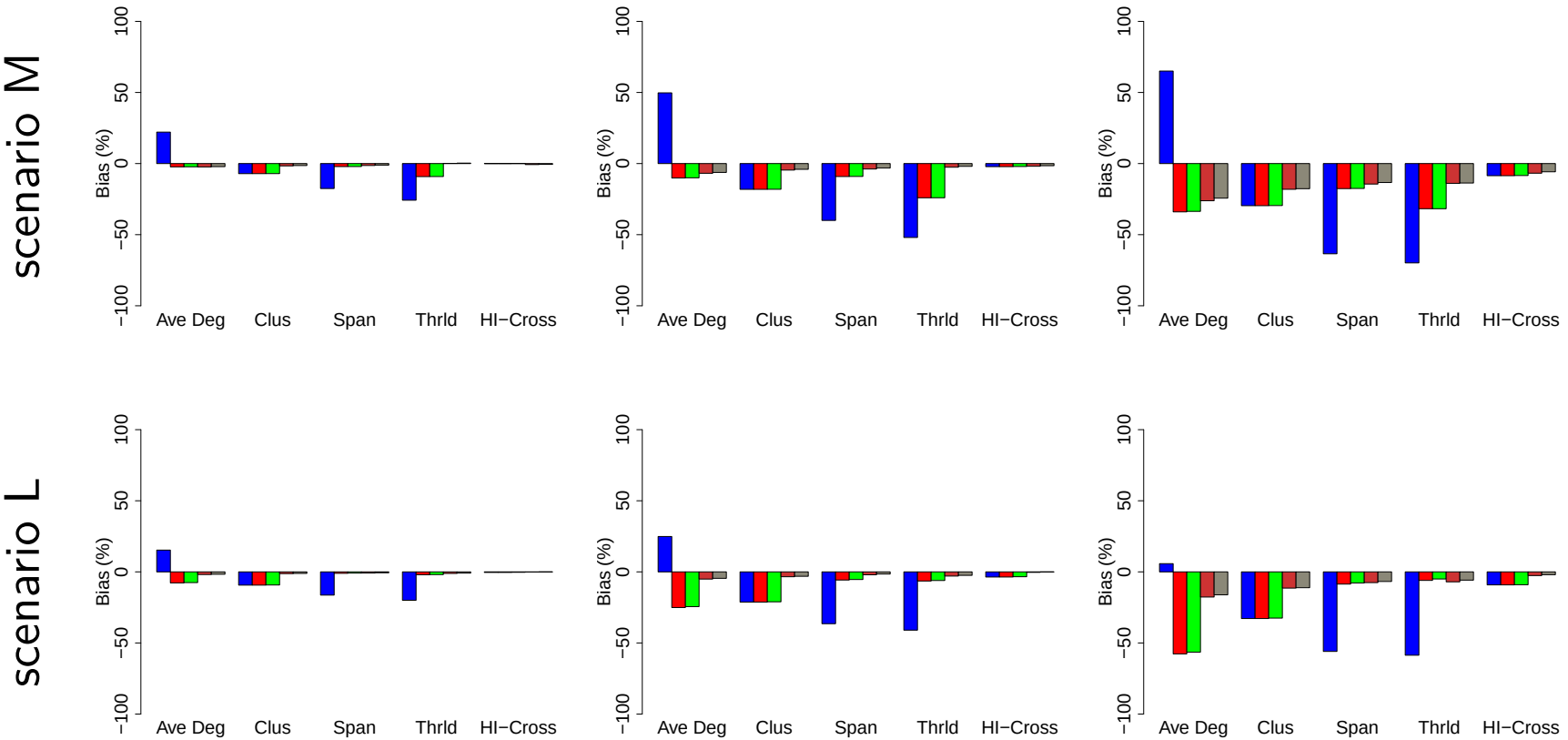
Statistical Correction Methods for Sampled Networks



(a) 20% (b) 40% (c) 60%

Figure: Scenarios R and H on induced subgraph

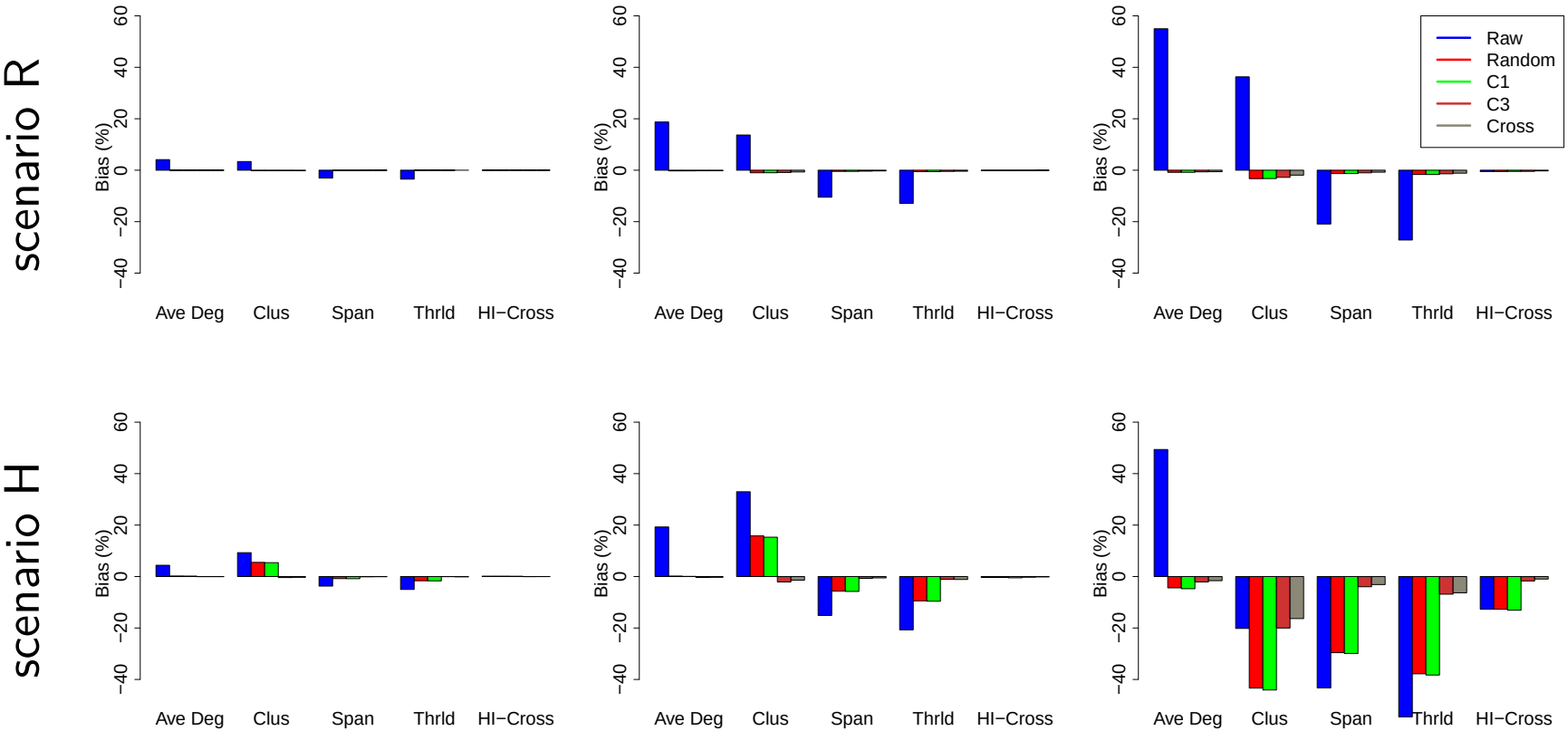
Statistical Correction Methods for Sampled Networks



(a) 20% (b) 40% (c) 60%

Figure: Scenarios M and L on induced subgraph

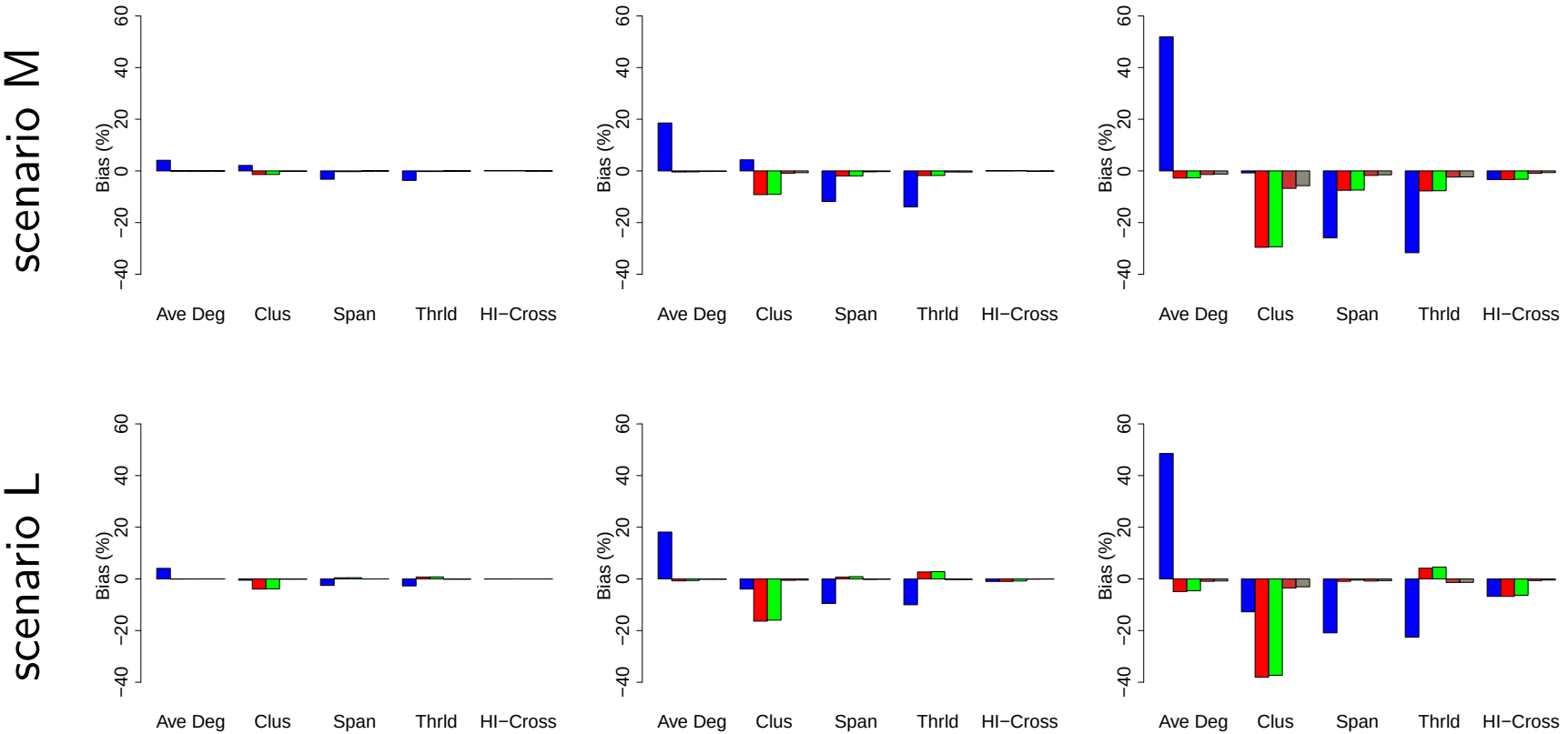
Statistical Correction Methods for Sampled Networks



(a) 20% (b) 40% (c) 60%

Figure: Scenarios R and H on star subgraph

Statistical Correction Methods for Sampled Networks



(a) 20% (b) 40% (c) 60%

Figure: Scenarios M and L on star subgraph

Statistical Correction Methods for Sampled Networks

- We study an empirical application on Indian rural village network sample.
- Banerjee et al. (2013) elicit a large variety of household characteristics including network data from 75 villages in southern Karnattaka, India.
- They first conducted a census to collect basic information (age, gender) of each villagers in all villages and later collected detailed follow-up survey with a subsample of the population of each village.
- The survey respondents only represent a sample of each village and thus their reported network is an induced subgraph of the population network.
- The average sampling rate across villages is 35%.
- The crucial aspect of their sampling design is the stratification by religion and geographic sub-location, generating a representative sample with respect to these two variables.

Statistical Correction Methods for Sampled Networks

Table: Average per-village population and sample percentages in the data from Banerjee et al. (2013).

	Population	Sample	Difference
Age			
< 30	38.71%	30.97%	7.74%
30-50	39.60%	54.11%	-14.51%
> 50	21.69%	14.92%	6.77%
Sex ratio	50.34%	44.57%	5.77%
Houshold size			
< 3	17.26%	15.49%	1.77%
3-8	71.57%	73.48%	-1.91%
> 8	11.17%	11.03%	0.14%
# of Villages	75	75	
Total Obs.	48,646	16,995	

The table reveals that the data are not representative in terms of age, gender, and to a lesser extent household size.

Statistical Correction Methods for Sampled Networks

Table: Estimated network effects on the share of population in rural India village that (I) employed and (II) work outside the village.

Dependent Variable	(I) Employed (%)			(II) Work Outside Village (%)		
	Raw	Random	Rake	Raw	Random	Rake
Degree	0.0269***	0.0091**	0.0088**	-0.0235*	-0.0093**	-0.0101*
	(0.0095)	(0.0035)	(0.0039)	(0.0120)	(0.0044)	(0.0051)
Cluster	0.1663**	0.1663**	0.1413**	-0.2137**	-0.2137**	-0.1665***
	(0.0643)	(0.0643)	(0.0610)	(0.0889)	(0.0889)	(0.0626)
Span	-0.0248**	-0.0746**	-0.0650*	0.0335***	0.1253***	0.1255***
	(0.0096)	(0.0349)	(0.0345)	(0.0117)	(0.0427)	(0.0435)
Epid. Thrld	-1.1530***	-2.3017***	-2.0965**	0.9357**	2.3498**	2.3924**
	(0.3589)	(0.8442)	(0.8341)	(0.4148)	(0.9967)	(1.0430)
HI-sex	0.1622	0.1622	0.0974	-0.0689	-0.0689	-0.0368
	(0.1017)	(0.1017)	(0.0929)	(0.1344)	(0.1344)	(0.1473)
HI-age	0.4015*	0.4015*	-0.1271	-0.1253	-0.1253	0.4875**
	(0.2356)	(0.2356)	(0.1932)	(0.3026)	(0.3026)	0.1953
HI-householdsize	0.0176	0.0176	-0.0127	0.2465**	0.2465**	0.0950
	(0.1036)	(0.1036)	(0.1003)	(0.0968)	(0.0968)	(0.1163)
HI-cross	0.2484	0.2484	0.0443	0.4158*	0.4158*	0.5957***
	(0.1790)	(0.1790)	(0.1628)	(0.2098)	(0.2098)	(0.2135)

Note: Regression is based on 75 villages. Standard errors robust to heteroscedasticity are reported in parentheses. *, **, *** stand for significance at 10%, 5%, and 1% respectively. Each row corresponds to one regression and the village size is included as a default control.

Reference

Banerjee, Abhijit, Arun G Chandrasekhar, Esther Duflo, and Matthew O Jackson (2013) "The diffusion of microfinance," *Science*, Vol. 341, p. 1236498.

Chandrasekhar, Arun and Randall Lewis (2011) "Econometrics of sampled networks," *Unpublished manuscript, MIT.[422]*.

Chen, Xinlei, Yuxin Chen, and Ping Xiao (2013) "The impact of sampling and network topology on the estimation of social intercorrelations," *Journal of Marketing Research*, Vol. 50, pp. 95–110.

Coleman, James (1958) "Relational analysis: the study of social organizations with survey methods," *Human Organization*, Vol. 17, pp. 28–36.

Frenzen, Jonathan K and Harry L Davis (1990) "Purchasing behavior in embedded markets," *Journal of Consumer Research*, Vol. 17, pp. 1–12.

Gjoka, Minas, Maciej Kurant, Carter T Butts, and Athina Markopoulou

(2010) "Walking in facebook: A case study of unbiased sampling of osns," in *Infocom, 2010 Proceedings IEEE*, pp. 1–9, Ieee.

Goel, Sharad and Matthew J Salganik (2010) "Assessing respondent-driven sampling," *Proceedings of the National Academy of Sciences*, Vol. 107, pp. 6743–6747.

Goodman, Leo A (1961) "Snowball sampling," *The Annals of Mathematical Statistics*, pp. 148–170.

Handcock, Mark S and Krista J Gile (2010) "Modeling social networks from sampled data," *The Annals of Applied Statistics*, Vol. 4, p. 5.

Heckathorn, Douglas D (1997) "Respondent-driven sampling: a new approach to the study of hidden populations," *Social Problems*, Vol. 44, pp. 174–199.

Henry, Paul C (2005) "Social class, market situation, and consumers' metaphors of (dis) empowerment," *Journal of Consumer Research*, Vol. 31, pp. 766–778.

Holland, Paul W and Samuel Leinhardt (1973) "The structural implications of measurement error in sociometry," *Journal of Mathematical Sociology*, Vol. 3, pp. 85–111.

Horvitz, Daniel G and Donovan J Thompson (1952) "A generalization of sampling without replacement from a finite universe," *Journal of the American Statistical Association*, Vol. 47, pp. 663–685.

Hsieh, Chih-Sheng, I. M. Stanley Ko, Jaromír Kovářík, and Trevon Logan (2018a) "Non-randomly sampled networks: Biases and Corrections," *working paper*.

Hsieh, Chih-Sheng, Stanley IM Ko, Jaromír Kovářík, and Trevon Logan (2018b) "Non-Randomly Sampled Networks: Biases and Corrections,"Technical report, National Bureau of Economic Research.

Hu, Pili and Wing Cheong Lau (2013) "A survey and taxonomy of graph sampling," *arXiv preprint arXiv:1308.5865*.

Huisman, Mark and Christian Steglich (2008) "Treatment of

non-response in longitudinal network studies," *Social Networks*, Vol. 30, pp. 297–308.

Klovdahl, Alden S (1989) "Urban social networks: Some methodological problems and possibilities," *The Small World*, pp. 176–210.

Koskinen, Johan H, Garry L Robins, and Philippa E Pattison (2010) "Analysing exponential random graph (p-star) models with missing data using Bayesian data augmentation," *Statistical Methodology*, Vol. 7, pp. 366–384.

Kossinets, Gueorgi (2006) "Effects of missing data in social networks," *Social networks*, Vol. 28, pp. 247–268.

Laumann, Edward O, Peter V Marsden, and David Prensky (1989) "The boundary specification problem in network analysis," *Research Methods in Social Network Analysis*, Vol. 61, p. 87.

Leskovec, Jure and Christos Faloutsos (2006) "Sampling from large graphs," in *Proceedings of the 12th ACM SIGKDD international*

conference on Knowledge discovery and data mining, pp. 631–636, ACM.

Lovász, László (1993) "Random walks on graphs," *Combinatorics, Paul erdos is eighty*, Vol. 2, p. 4.

Robins, Garry, Philippa Pattison, and Jodie Woolcock (2004) "Missing data in networks: exponential random graph (p∗) models for networks with non-respondents," *Social Networks*, Vol. 26, pp. 257–283.

Rumsey, Deborah J (1993) "Nonresponse models for social network: stochastic processes," Ph.D. dissertation, The Ohio State University.

Salganik, Matthew J and Douglas D Heckathorn (2004) "Sampling and estimation in hidden populations using respondent-driven sampling," *Sociological Methodology*, Vol. 34, pp. 193–240.

Stork, Diana and William D Richards (1992) "Nonrespondents in communication network studies: Problems and possibilities," *Group & Organization Management*, Vol. 17, pp. 193–209.

Stumpf, Michael PH, Carsten Wiuf, and Robert 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*, Vol. 102, pp. 4221–4224.

Tepper, Kelly (1994) "The role of labeling processes in elderly consumers' responses to age segmentation cues," *Journal of Consumer Research*, Vol. 20, pp. 503–519.