Non-Representative Sampled Networks: Estimation of Network Structural Properties by Weighting*

Chih-Sheng Hsieh† National Taiwan University

Yu-Chin Hsu‡ Academia Sinica

Stanley I. M. Ko§ Tohoku University

Jaromír Kovářík¶ University of the Basque Country & University of West Bohemia

Trevon D. Logan‖ The Ohio State University & NBER

January 20, 2024

Abstract

This paper analyzes statistical issues arising from non-representative samples of a network. Sampled network data could systematically bias the network properties and generate non-classical measurement error problems. Apart from the sampling rate and the elicitation procedure, the biases on network structural measures depend non-trivially on which subpopulations of nodes are missing with higher probability. We propose a methodology, adapting weighted estimators to networked contexts, which enables researchers to recover several network-level statistics and reduce the biases in the estimated network effects. The proposed weighted estimators are consistent and asymptotically normally distributed and have good performance in finite samples. Notably, our approach does not require users to assume any network formation model and is straightforward to implement.

*We are grateful to Isaiah Andrews, Aureo de Paula, Marco van der Leij, and participants at numerous seminars for comments and suggestions. Hsieh acknowledges financial support from the National Science and Technology Council of Taiwan (NSTC110-2410-H-002-195). Hsieh and Hsu gratefully acknowledge the research support from the Center for Research in Econometric Theory and Applications of National Taiwan University (Grant no. 112L8601). Hsu gratefully acknowledges research support from the National Science and Technology Council of Taiwan (NSTC112-2628-H-001-001), and the Academia Sinica Investigator Award of Academia Sinica, Taiwan (AS-IA-110-H01). Kovářík acknowledges financial support from *Ministerio de Economía y Competividad* and *Fondo Europeo de Desarrollo Regional* (PID2019-108718GB-I00, PID2019-106146GB-I00), the Basque Government (IT1461-22), and the Grant Agency of the Czech Republic (21-22796S).

†Department of Economics, National Taiwan University, Taipei, Taiwan (cshsieh@ntu.edu.tw)

‡Institute of Economics, Academia Sinica, Taipei, Taiwan & Department of Finance, National Central University, Taoyuan City, Taiwan & Department of Economics, National Chengchi University, Taipei, Taiwan (ychsu@econ.sinica.edu.tw)

§Graduate School of Economics and Management, Tohoku University, Japan (stanleyko@tohoku.ac.jp).

¶Dpto. del Análisis Económico, University of the Basque Country UPV-EHU, Bilbao, Spain; Faculty of Arts & Faculty of Economics, University of West Bohemia, Pilsen, Czech Republic (jaromir.kovarik@ehu.eus).

‖Department of Economics, The Ohio State University, 410 Arps Hall, 1945 N. High Street, Columbus, OH, 43210 (logan.155@osu.edu).

1 Motivation

There is growing interest in understanding the role of networks in Economics (Vega-Redondo, 2007; Jackson, 2010). Different "micro" and "macro" features of network architecture shape diffusion, learning, behavior, and other substantive phenomena in a variety of contexts.Due to the increasing availability of large network data sets and increasing computational power, empirical network research is now a dynamic and growing part of this literature. At the same time, empirical network analysis generates new econometric challenges (Fortin and Boucher, 2015; De Paula, 2017; Jackson et al., 2017). This paper tackles the challenges that arise when network data come from non-representative samples of the population, which is the most commonly encountered scenario in practical applications.

The vast majority of empirical network studies analyze sampled data, and the sampling rates are typically low.[1] Even though the literature across several disciplines has noted that using sampled data may lead to considerable biases and other statistical issues (see below for references), the typical approach is to treat the sampled data "as if" it were complete. Chandrasekhar and Lewis (2016) show formally that, even if the nodes are selected representatively through simple random sampling (SRS, henceforth), the statistics of the sampled networks differ significantly from those of the population network. This disparity results in measurement errors and inconsistency problems when we estimate network effects through regressions. The estimates from sampled networks may suffer from attenuation, expansion, or even sign-switching. As a result, one cannot rely on solutions to classical measurement-error problems to correct these issues, even if the sample is representative.

Furthermore, nodes observed in network samples are typically non-representative. First, non-representativeness may be caused by the sampling design itself (Frank, 1981; Kolaczyk, 2009; Handcock and Gile, 2010). For instance, the star subgraph sampling design analyzed in this paper is prone to including nodes with higher connectivity than nodes with a small number

[1]The reasons behind the common use of network samples are that the impracticality of analyzing the entire population and the higher costs associated with network elicitation compared to collecting basic individual characteristics (Aral, 2016; Breza et al., 2020). Chandrasekhar and Lewis (2016) report that the median sampling rate in applied work in Economics is 25% and more than 66% of network studies have a sampling rate lower than 51%. Similar rates are found in other fields.

of network neighbors. The reason is that star subgraphs encompass not only the initially sampled nodes but also their network neighbors even if the latter were not initially sampled. Having more connections thus increases the probability of a node being included. This is an example of a design that generates samples in which the inclusion probabilities of nodes are endogenous to the underlying population network structure. Non-representativeness may also arise when the inclusion (or missing) probabilities are orthogonal to the population network architecture. For instance, when network samples are collected with specified boundaries such as within schools, or within villages, etc., it is not guaranteed that samples within boundaries are representative of the entire population. Such boundary-induced network samples are equivalent to the induced subgraph sampling design analyzed in this paper. Other common sources of non-representativeness in sampling studies are non-responses or disproportionate stratified sampling. Many studies exploit stratified samples to improve precision and sampling efficiency. Unfortunately, it is difficult and costly to stratify for all relevant characteristics.

To intuitively explain the issues arising from sampled networks, we decompose the problem into two sources, *scaling* and *non-representativeness*. *Scaling* refers to observing fewer nodes and edges than there exist in the whole network, independently of the (non-) representativeness of the sample. In contrast, *non-representativeness* arises when different nodes have unequal probabilities of being included in the sample. If nodes appear in the sample with equal probability, only scaling matters. As an example of the effect of scaling, let's consider the average degree of a network. When the links between the sampled and non-sampled nodes are not observed, the sample average degree is biased downwards by construction. Furthermore, suppose the average degree is correlated with the network's diffusion properties. As a result, using the sample average degree in a regression analysis leads to an overestimation of the average degree's impact on diffusion, even when samples are representative. This is an example of the expansion of the estimated effect and thus, non-classical measurement error. However, if nodes appear in the sample with unequal probabilities, whether the observed average degree and the estimates are inflated or attenuated will depend on who is missing. For example, if less connected nodes are missing with higher probability, scaling and non-representativeness can

bias the average degree and the estimates in opposite directions, and one cannot easily predict which force will dominate. In contrast to the average degree, the global clustering coefficient and the homophily index can be unbiased in representative samples. In samples in which different types of nodes are missing with different probabilities, homophily will be biased by definition. Since clustering is typically associated with connectivity in social networks (Jackson and Rogers, 2007), it is also likely to be mismeasured. The magnitude and direction of the biases in these characteristics and their estimated effects in regressions again depend crucially and non-trivially on who is missing.

In this study, we systematically analyze the problems arising from sampled network data elicited via two widely employed sampling methods, and proposes a solution enabling to recover the true structural features of a network (e.g., average degree) and mitigate biases in regressions which study the impact of these network features on either individual or group-level behaviors and outcomes.[2] We first derive analytically weighted estimators for a set of network structural properties from sampled networks assuming that nodes appear in the sample with unequal probabilities according to their types. Secondly, we study the asymptotic properties of the proposed weighted estimators and evaluate their finite-sample performance numerically. Lastly, the proposed methodology is applied to a widely employed stratified data set on Indian villages (Banerjee et al., 2013).[3] This data set is suited for our approach because it contains a relatively large number of networks, and we document that the network data have been collected from a non-representative sample of the population under scrutiny.

This study shows that relying on the assumption of representativeness to adjust network samples, which is rarely satisfied in real-world applications, can be as biased as using raw network samples without any adjustments. Since the direction and magnitude of the biases depend on who is missing, we demonstrate the necessity of accounting for potentially different missing rates of different types of nodes in applied work. This is particularly important in

[2]Our study also improves inferences in network-formation applications studying contextual determinants of the network architecture (i.e., applying network properties as regressands). Since network formation represents a key topic in the network literature (see Jackson (2005) and De Paula (2020) for reviews), it enlarges the applicability of the proposed methodology. However, this study focuses on regressions including network properties as regressors.

[3]See, e.g., Jackson et al. (2012); Banerjee et al. (2013, 2014); Chandrasekhar and Lewis (2016) and De Paula et al. (2018), among others.

network data where population and distributional parameters are of primary interest.

As the main contribution, we propose weighted estimators for a selected set of network characteristics that are widely used in applications: average degree, global clustering coefficient, epidemic threshold, and homophily index. These network features represent fundamental aspects of network architecture employed in theoretical and empirical research and provide intuitive insights regarding the way social organization shapes individual and group-level phenomena (Jackson et al., 2017).[4] To that aim, we assume that network members can be divided into a finite number of disjoint types, and that sampling rates differ across these types. Taking explicit account of the differing sampling rates across types, we adapt standard (network-free) Horvitz-Thompson (H-T) estimators to networked contexts and propose (post-) stratification as a viable approach to correct sampling biases caused either by the sampling procedure or due to varying non-response rates among different demographic or socioeconomic categories (or both) in order to improve the precision of sample estimates for objective variables of interest (Smith, 1991; Little, 1993). The main difference between the standard H-T estimators and our approach is to weight on network objects, such as links, triples, or triangles, rather than on nodes.[5] We prove that, in sparse networks, the proposed weighted estimators are consistent and asymptotically normally distributed.[6] We also provide sufficient conditions so that we can ignore the estimation effects when the regression analysis includes the proposed weighted network measures as covariates. Our numerical analysis shows that our methodology performs well in finite samples and substantially outperforms both the naive (uncorrected) statistics from the raw data and corrections designed for representative samples.

Our empirical application shows that the Indian village network data stratified on religion and geography are non-representative in terms of age and gender. We then show that not accounting for unequal missing rates for nodes of different types affects the estimated network effects substantially and one cannot easily predict the direction and magnitude of the biases. Given the differences, applied researchers should carefully consider to what extent their results

[4]Section 6 discusses the extension of our approach to other network characteristics.

[5]Such network objects are referred to as subgraphs, subnetworks, or network motifs in different fields.

[6]Since virtually all real-life social and economic networks are sparse, our asymptotic results are broadly applicable to empirical research.

might be driven by the non-representativeness of their samples.

The present paper connects to three pieces of literature. First, our methodology complements emerging econometric literature on imperfectly measured network data and the estimation of network effects. Chandrasekhar and Lewis (2016) show that estimations with network data coming from representative samples suffer from non-classical measurement errors and propose a method to ensure consistent estimates. Their methodology consists of two alternative approaches. First, they provide formal corrections for several network measures. Our approach generalizes this first strategy. As a second approach, they propose a graphical reconstruction technique that delivers consistent estimates in both network-level and individual-level regressions. The procedure is first to estimate a network formation model and then employ the estimated model to interpolate over missing parts of the network. The network reconstruction approach requires a correct model specification and certain assumptions to ensure the consistency of the network effects. However, this second approach does not necessarily recover the structural properties of the population network, which is the primary objective of our study. Most importantly, from the perspective of the present work, both approaches are restricted to the case that the sample is representative. Chandrasekhar and Jackson (2016) propose a network formation model similar in spirit to our methodology in that it is also based on subgraphs in the function of types of nodes. However, none of these approaches can effectively recover the true network formation process from non-representative samples because, when the network-formation model is fitted on non-representative network samples, the estimated parameters in the first stage will likely be biased and potentially inconsistent even if the assumed model is correct. Thirkettle (2019) proposes a network formation model enabling the estimation of bounds on network statistics from partially observed networks. The advantage of our approach, as opposed to the graphical reconstruction techniques, is that our methodology does not rely on any assumed network formation model. Our work complements and expands the above studies by providing the first step toward the statistical treatment of network data coming from non-representative samples of the population, which is the most common type of network data available.[7]

[7] Boucher and Houndetoungan (2020) study the estimation of peer effects when the researchers only observe

Second, we contribute to the statistical sampling theory that has developed procedures for recovering the true network structural parameters from samples if the only source of non-representativeness comes from the network sampling design (see Kolaczyk (2009) for a survey). Our methodology nests these procedures as a special case (e.g., Frank, 1981, Kolaczyk, 2009, Chandrasekhar and Lewis, 2016). Our weighting method shares the same goals with these approaches but differs substantially in the underlying assumptions and applicability. Unlike these approaches which are typically suitable for specific sampling designs, our method can be applied or adapted to various sampling procedures.[8] Furthermore, our approach remains effective even in cases where non-representativeness is caused by factors unrelated to the sampling design, such as non-response or the presence of hard-to-reach subpopulations. Most importantly, existing approaches assume certain forms of representativeness in the sampling process *ex-ante*, while our proposed methodology targets both *ex-ante* and *ex-post* non-representativeness of the sample. If the sampling rates are set by the researcher before the data collection (as in the approaches discussed above and in standard stratification), they are treated as known parameters. If the sampling rates are learned after the data collection, our methodology corresponds to post-stratification by exploiting the non-representativeness of the sample and treats the sampling rates as unknown parameters to be estimated.

Last, we contribute to better practices for empirically evaluating the effects of global network features in socio-economic environments. Our study shows that, despite other econometric issues, mismeasured network features with non-representative samples might lead to a serious misunderstanding of network effects. However, our methodology mitigates this issue and provides an additional argument for the employment of sampling in empirical network work. With the increasing use of network data and corresponding empirical techniques, our proposed approach can improve the design of network sampling strategies and the inference we draw from network studies more generally. Moreover, it can serve as a standard robustness check of empirical results.

consistent estimates of aggregate network statistics. Hence, our methodology and their approach naturally complement each other in non-representative samples since our methodology delivers such consistent estimates from non-representative samples.

[8]For the sake of brevity, we concentrate on two sampling designs commonly used in economics. However, our methodology can be applied or adapted to other sampling designs. See Section 6.

2 Framework

2.1 Notation

A graph or network is defined by $G_n = (V, E)$, where V is the set of vertices (nodes) with $n = |V|$ denoting the cardinality of V, and E is the set of edges (links). The network can be represented by an $n \times n$ adjacency matrix W_n. We focus on unweighted and undirected networks; i.e., $W_{ij,n} = 1(0)$ if i and j are (not) connected and $W_{ij,n} = W_{ji,n}$ for each $i, j \in V$. Following the convention, we exclude self-loops by setting $W_{ii,n} = 0$. We assume that the nodes can be classified into T disjoint types with a generic type $t \in \mathcal{T} = \{1, 2, ..., T\}$. One can view this classification as stratification, which can be carried out either before or after sample collection. When conducted after the collection, this process is commonly referred to as post-stratification. We write $t_i = t$ if node 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. Let V_t be the set of nodes of type t, $n_t = |V_t|$ is the size of this set, and $\sum_{t=1}^{T} n_t = n$.

Rather than the whole network G_n, researchers only observe the sampled network, which is also referred to as a subgraph of G_n. Let $V^* \subseteq V$ be the set of sampled nodes of size $m = |V^*|$ and let ψ denote the sampling rate. Analogously, V_t^* denotes the set of nodes of type t in the sample and $m_t = |V_t^*|$ is the number of sampled nodes of type t and $\sum_{t=1}^{T} m_t = m$. We use ψ_t to denote type t's sampling rate. We assume that $\psi_t > \tau$ for some $\tau > 0$ for each t and is independent of n. Crucially, we assume that, within each type, individual nodes have an equal probability of being selected into the sample. Our framework primarily focuses on non-representative samples, i.e., $\psi_t \neq \psi_s$ for at least one $t, s \in \mathcal{T}$, while also encompassing the representative sample, i.e., $\psi_t = \psi$ for all $t \in \mathcal{T}$, as a special case.

In the context of (*ex-ante*) stratification, the true value of ψ_t is a known quantity specified by the researcher. However, when it comes to post-stratification, the true value of ψ_t is treated as unknown and can be estimated by $\hat{\psi}_t = \frac{m_t}{n_t}$, the ratio of the number of nodes of type t included in the sample to the population number of nodes of type t. We denote $\varphi_i = \sum_{t=1}^{T} \psi_t \mathbf{1}(t_i = t)$ the sampling probability of node i, conditional on her type, and

$\hat{\varphi}_i = \sum_{t=1}^{T} \hat{\psi}_t \mathbf{1}(t_i = t)$ the corresponding estimator based on $\hat{\psi}_t$.

Given sampled nodes, this paper focuses on two designs for eliciting network edges. The first is the *induced subgraph*, in which the sampled network is denoted by $G_n^{\mathrm{I}} = (V^*, E^{\mathrm{I}})$. In G_n^{I}, the set V^* involves m sampled nodes, and the set $E^{\mathrm{I}} \subseteq E$ involves network links among these m sampled nodes. W_n^{I} is the $m \times m$ adjacency matrix corresponding to G_n^{I}. The second is the *star subgraph*, in which the sampled network is denoted by $G_n^{\mathrm{S}} = (V^*, E^{\mathrm{S}})$.[9] In G_n^{S}, there are m initially sampled nodes in the set V_0^*. However, researchers observe not only the network links among these m sampled nodes, but also the links of the m sampled nodes to unsampled nodes in V. Hence, we use E^{S} to denote the set of edges such that at least one node of the corresponding dyad is in V_0^*. The set V_0^* is enlarged to V^* by including all the vertices $i \in V \setminus V_0^*$ that are connected through the observed links to at least one sampled node from V_0^*. The size of this enlarged vertex set is denoted by $m' = |V^*|$, and the corresponding sampling rate is denoted by ψ'. Let W_n^{S} be the $m' \times m'$ adjacency matrix corresponding to the graph G_n^{S}. In both the induced and star subgraphs, we assume that edges are reported without errors.

We study several network structural properties (measures) and refer to a generic population network measure as Λ. Let $\Lambda(G_n)$ denote the estimated network measure for Λ based on the whole network data, and let $\Lambda(\overline{G}_n)$, $\overline{G}_n \in \{G_n^{\mathrm{I}}, G_n^{\mathrm{S}}\}$, represent the corresponding estimated network measure based on the sampled network $\overline{G}_n$. We call $\Lambda(\overline{G}_n)$ the *naive* estimator of network property. Additionally, let $\widetilde{\Lambda}(\overline{G}_n)$ denote the weighted network measure proposed to mitigate sample biases with respect to the whole network. For example, $\Lambda(G_n) = \frac{1}{n}\sum_{i\in V}\sum_{j\in V} W_{ij,n}$ is the average degree of a graph, which we denote $d(G_n)$ below. Hence, $d(\overline{G}_n)$ is the average degree of the sampled network, and $\widetilde{d}(\overline{G}_n)$ is the proposed weighted estimator to mitigate biases of $d(\overline{G}_n)$.

In applications, researchers may observe multiple networks. We use a generic subscript $r \in \mathcal{R} = \{1, 2, ..., R\}$ when a measure refers to network r. That is, G_{r,n_r} denotes the graph r, and $\overline{G}_{r,n_r} \in \{G_{r,n_r}^{\mathrm{I}}, G_{r,n_r}^{\mathrm{S}}\}$ denotes the corresponding sampled network. Therefore, $n_{r,t}$ and

[9] G_n^{S} is referred to as the *labeled* star subgraph in Kolaczyk (2009) because the unsampled nodes which connect to sampled nodes are identified and labeled.

$m_{r,t}$ are the number of nodes of type t in the whole network r and its corresponding number in the sample.

2.2 Regression with Network Measures

In addition to the reconstruction of network properties of interest, we also consider regression analysis with network measures. Throughout the analysis, we focus on regressions in which researchers are interested in understanding whether and how the global measures of network properties influence a particular outcome. Formally,

$$y_r = \alpha + \beta\Lambda_r + x_r\gamma + \varepsilon_r, \tag{1}$$

where y_r is the outcome variable of network (or community) r, x_r is the set of network-level controls, and Λ_r is the population network property of interest for r-th network population. The researchers are interested in estimating the parameters α, β, and γ. Examples of the applications of (1) in the literature include Alatas et al. (2016) which regress the ability of villagers to aggregate information on a set of network characteristics in Indonesian villages, Banerjee et al. (2013) who model microfinance take-up rate in rural India in function of the average centrality of the initial seeds, Currarini et al. (2009) and Golub and Jackson (2012) who relate homophily with school-level statistics using Add Health data, or Fleming et al. (2007) who model the ability of different regions to generate knowledge depending on the structure of regional research networks. Such regressions are also of interest theoretically. For example, the overall clustering of a network may explain the magnitude and efficiency of risk-sharing within a society (Bloch et al., 2008), and the stability of behavior in a society may be related to the minimal eigenvalue of the adjacency matrix (Bramoullé et al., 2014).

The proposed approach also applies to models investigating the influence of a network's global measure on individual-level outcomes: $y_{ir} = \alpha + \beta\Lambda_r + x_{ir}\gamma + \mu_r + \varepsilon_{ir}$, where y_{ir} is the outcome of an individual i in network r, x_{ir} captures individual heterogeneity (that can also include the heterogeneity of i's neighborhood), and μ_r is a network random effect. For instance, the decision of an individual to adopt a product (e.g., microfinance as in Banerjee et al., 2013), participate in an activity (e.g., recreational activity as in Bramoullé et al., 2009),

or behave in a particular way (Centola, 2010) can depend on the overall structure of the network. In the same vein, the innovation literature studies how the structure of regional networks shapes the innovative performance of individual innovators (Schilling and Phelps, 2007). There also exist theories arguing that the overall structure of a network may determine the behavior at the individual level (see, e.g., Ballester et al., 2006; Bramoullé et al., 2014).

With sampled data, researchers observe $\overline{G}_{r,n_r} \in \{G^{\mathrm{I}}_{r,n_r}, G^{\mathrm{S}}_{r,n_r}\}$, and the naive estimator $\Lambda(\overline{G}_{r,n_r})$ is not a consistent estimator for Λ_r. Therefore, when researchers estimate

$$y_r = \alpha + \beta\Lambda(\overline{G}_{r,n_r}) + x_r\gamma + u_r, \tag{2}$$

it leads to a measurement error in the regressor. The classic measurement error and the resulting attenuation bias are based on several assumptions that are generally not satisfied in the case of network measures.[10] Chandrasekhar and Lewis (2016) show analytically and via simulations that the biases are generally not tractable and can lead to expansion or sign switching under representativeness. The issues become even more problematic if the representativeness assumption is violated. On the other hand, when researchers estimate

$$y_r = \alpha + \beta\widetilde{\Lambda}(\overline{G}_{r,n_r}) + x_r\gamma + u_r, \tag{3}$$

it leads to consistent estimation of the parameters. In this regard, we later provide sufficient conditions such that we can ignore the estimated effect of $\widetilde{\Lambda}(\overline{G}_{r,n_r})$ in the OLS regression.

3 Weighted Estimators for Sampled Network Measures

This section proposes weighted estimators for commonly used network measures when sampled data are used. We also address the biases present in both the naive (unweighted) estimators and weighted estimators that solely account for scaling effects. One key assumption made throughout this section is that the network measures (statistics) under consideration are well-defined. For example, when the sampling rate is extremely low, the global clustering coefficient could be zero if no closed triplets are observed in the sampled network. In such scenarios,

[10] Although network regressions often face additional challenges such as endogeneity and omitted variable problems, we contend that the sampling issue persists even in the absence of these problems.

both the naive estimator and our proposed weighted estimator are null and it is not possible to recover the true value of the coefficient. Hence, we stress that our corrections for network statistics are applicable under sampling rates in which well-defined naive estimators exist. To maintain notational simplicity, we will omit the network index r in the subscripts throughout Sections 3.1 and 3.2. We reintroduce it when discussing the asymptotics of regressions with network measures in Section 3.3.

3.1 Average Degree

The degree is the number of connections of a node, which is a basic measure of a node's importance or local centrality. The average degree of the graph G_n is simply the average number of network links per node in the network, defined as $d(G_n) = \frac{1}{n}\sum_{i\in V}\sum_{j\in V} W_{ij,n}$. It has been applied as a regressor in numerous empirical studies of different contexts (see, e.g., Branas-Garza et al., 2010; Banerjee et al., 2013; Alatas et al., 2016, among many others).

For an induced subgraph, the naive estimator of the average degree is computed as

$$d(G_n^{\mathrm{I}}) = \frac{1}{m}\sum_{i\in V^*}\sum_{j\in V^*} W_{ij,n}^{\mathrm{I}} = \frac{1}{m}\sum_{i\in V}\sum_{j\in V} W_{ij,n} D_i D_j, \tag{4}$$

where D_i is a binary variable that takes the value 1 if $i \in V^*$, and 0 otherwise. To correct the biases from both scaling and non-representativeness in (4), we propose the weighted sample average degree by multiplying each observed sample edge $W_{ij,n}^{\mathrm{I}}$ with the weight, $\left(\hat{\varphi}_i\hat{\varphi}_j\right)^{-1}$, which is the inverse of the estimated inclusion probability. Thus, the weighted sample average degree is given by

$$\widetilde{d}(G_n^{\mathrm{I}}) = \frac{1}{n}\sum_{i\in V^*}\sum_{j\in V^*} W_{ij,n}^{\mathrm{I}}\left(\hat{\varphi}_i\hat{\varphi}_j\right)^{-1} = \frac{1}{n}\sum_{i\in V}\sum_{j\in V} W_{ij,n}\frac{D_i}{\hat{\varphi}_i}\frac{D_j}{\hat{\varphi}_j}. \tag{5}$$

As the true value of the inclusion probability $(\varphi_i\varphi_j)$ is typically unknown and needs to be estimated from the sample, we refer to the weighted estimator in (5) as a post-stratification estimator. However, when the true value of $(\varphi_i\varphi_j)$ is known and applied in (5), the estimator follows the general principle of the H-T estimator (Horvitz and Thompson, 1952). To show why the proposed weighted estimator (5) removes the bias in (4), assume known $(\varphi_i\varphi_j)$. Then,

$$\mathrm{E}\left(\frac{1}{n}\sum_{i\in V}\sum_{j\in V}W_{ij,n}\frac{D_i}{\varphi_i}\frac{D_j}{\varphi_j}\middle| G_n\right) = \frac{1}{n}\sum_{i\in V}\sum_{j\in V}W_{ij,n}\left(\frac{\mathrm{E}(D_iD_j|G_n)}{\varphi_i\varphi_j}\right)$$
$$= \frac{1}{n}\sum_{i\in V}\sum_{j\in V}W_{ij,n}\left(\frac{\varphi_i\varphi_j}{\varphi_i\varphi_j}\right) = \frac{1}{n}\sum_{i\in V}\sum_{j\in V}W_{ij,n}.$$

That is, the expected value of our weighted estimator is the true average degree of the population network. The intuition behind (5) is as follows. There are $\sum_{i\in V}\sum_{j\in V}W_{ij,n}$ edges to account for in G_n. However, due to variations in the inclusion probabilities of sample edges, we only observe $\sum_{i\in V}\sum_{j\in V}W_{ij,n}(\varphi_i\varphi_j)$ edges in an induced subgraph in expectation. Even if samples are representative (i.e., $\varphi_i = \varphi_j = \psi$), as long as $\psi < 1$, a bias emerges due to scaling. Moreover, as φ_i and φ_j are not necessarily the same, we have the second source of bias, non-representativeness, and the issues become more complicated.

For a star subgraph, the naive (sample) average degree is defined as

$$d(G_n^{\mathrm{S}}) = \frac{1}{m'}\sum_{i\in V^*}\sum_{i\in V^*}W_{ij,n}^{\mathrm{S}} = \frac{1}{m'}\sum_{i\in V}\sum_{j\in V}W_{ij,n}\Big(1-(1-D_i)(1-D_j)\Big), \tag{6}$$

where D_i is a binary variable that takes the value 1 if $i \in V_0^*$, and 0 otherwise. To correct the bias, we propose the following weighted sample average degree,

$$\widetilde{d}(G_n^{\mathrm{S}}) = \frac{1}{n}\sum_{i\in V^*}\sum_{j\in V^*}W_{ij,n}^{\mathrm{S}}\left(1-(1-\hat{\varphi}_i)(1-\hat{\varphi}_j)\right)^{-1} = \frac{1}{n}\sum_{i\in V}\sum_{j\in V}W_{ij,n}\frac{1-(1-D_i)(1-D_j)}{1-(1-\hat{\varphi}_i)(1-\hat{\varphi}_j)}. \tag{7}$$

Once again, assuming that φ_i's are known, we can demonstrate the following:

$$\mathrm{E}\left(\frac{1}{n}\sum_{i\in V}\sum_{j\in V}W_{ij,n}\frac{1-(1-D_i)(1-D_j)}{1-(1-\varphi_i)(1-\varphi_j)}\middle| G_n\right)$$
$$= \frac{1}{n}\sum_{i\in V}\sum_{j\in V}W_{ij,n}\left(\frac{\mathrm{E}(1-(1-D_i)(1-D_j)|G_n)}{1-(1-\varphi_i)(1-\varphi_j)}\right)$$
$$= \frac{1}{n}\sum_{i\in V}\sum_{j\in V}W_{ij,n}\left(\frac{1-(1-\varphi_i)(1-\varphi_j)}{1-(1-\varphi_i)(1-\varphi_j)}\right) = \frac{1}{n}\sum_{i\in V}\sum_{j\in V}W_{ij,n}.$$

This result justifies why the weighted sample average in (7) mitigates the bias problem.

The weighted estimators proposed in (5) and (7) account for two phenomena. Firstly, they account for the differing inclusion probabilities of the links in the function of the types of the involved nodes. Secondly, they respect the correlations in who is connected to whom in the observed part of the network (i.e., they respect the network homophily). If one applies the corrections assuming representativeness of the sample, (5) and (7) will change to

$$\widetilde{d}(G_n^{\mathrm{I}}) = \frac{1}{n} \sum_{i \in V^*} \sum_{j \in V^*} W_{ij,n}^{\mathrm{I}} \left(\hat{\psi}^2 \right)^{-1} \tag{8}$$

and

$$\widetilde{d}(G_n^{\mathrm{S}}) = \frac{1}{n} \sum_{i \in V^*} \sum_{j \in V^*} W_{ij,n}^{\mathrm{S}} \left(1 - (1 - \hat{\psi})^2 \right)^{-1}, \tag{9}$$

respectively. These corrections are exactly the same as shown by Chandrasekhar and Lewis (2016). However, biases would still emerge in (8) and (9) if the sample is not truly representative. Importantly, there is no reason for these biases to be smaller than in the raw (uncorrected) data as their size depends on who is missing.

One can perceive the proposed weighted estimators (5) and (7) as a design-based approach. However, the remainder of this subsection characterizes the asymptotic properties of the estimators. To this aim, we envision that the underlying finite-population network (G_n) expands progressively toward a hypothetical superpopulation network. This prompts a natural transition to a model-based approach, targeting the unknown (model) parameters that characterize this hypothetical superpopulation for asymptotic statistical inference.[11] Consequently, we advocate a synthesis of design-based and model-based approaches (Binder and Roberts, 2003; Sterba, 2009). We expand upon the framework introduced by Bickel et al. (2011) to account for non-representativeness of nodes under the assumption that the network is sparse. Sparse networks refer to networks where the number of observed links is considerably lower than the maximum number of possible links, a common feature of real-life social networks. Formally, sparseness is defined as the property of an infinite sequence of graphs where the (average)

[11] An alternative possibility is asymptotic analysis with finite population sampling (Prášková and Sen, 2009; Li and Ding, 2017). However, the methodology cannot currently handle sparse networks in finite populations. Consequently, we adopt the approach in Bickel et al. (2011) to investigate the asymptotics of a superpopulation and defer the analysis of a finite population for future research.

degree is bounded as $n \to \infty$ (Bickel and Chen, 2009; Lovász, 2012).[12] In addition, similar to Bickel et al. (2011), we assume that the adjacency matrix of the whole network W_n is exchangeable.[13] As a result, according to the Aldous-Hoover theorem (Aldous, 1981; Hoover, 1979), the adjacency matrix can be represented by

$$W_{ij,n} \overset{\text{D}}{=} g_n(\xi_i, \xi_j, \epsilon_{ij}, t_i, t_j), \tag{10}$$

where $\overset{\text{D}}{=}$ denotes equality in distribution, and g_n is a measurable function symmetric in its first two and last two arguments. In (10), ξ_i and ϵ_{ij} are i.i.d. uniform random variables on $[0,1]$, $\epsilon_{ij} = \epsilon_{ji}$, and $\{t_i\}_{i=1}^n$ are independent of $\{\xi_i\}_{i=1}^n$ and $\{\epsilon_{ij}\}_{i,j=1}^n$. Note that this implies $W_{ij,n} = W_{ji,n}$.

Since the function $g_n(.)$ in (10) cannot be uniquely identified (Bickel and Chen, 2009), it would be advisable to explore an alternative parameterization, $h_{ts,n}(u,v) \equiv \text{P}[W_{ij,n} = 1 | \xi_i = u, \xi_j = v, t_i = t, t_j = s]$ for $t, s \in \mathcal{T}$, which refers to the unique canonical $h_{ts,\text{can}}$ such that $\int_0^1 h_{ts,\text{can}}(u,v)dv$ is monotone non-decreasing in u. Also, let $p_t = \text{P}(t_i = t)$ and assume for all $t \in \mathcal{T}$, $p_t \geq \tau$ for some $\tau > 0$, and is independent of n. Under these assumptions, we have $h_{ts,n}(u,v) = h_{st,n}(u,v)$ and $h_n(u,v) = \text{P}[W_{ij,n} = 1 | \xi_i = u, \xi_j = v] = \sum_{t=1}^T \sum_{s=1}^T h_{ts,n}(u,v) p_t p_s$. Let

$$\rho_n = \int_0^1 \int_0^1 h_n(u,v)\, du\, dv \tag{11}$$

be the probability of an edge in the network (i.e., network density). We can then write $w_{ts,n}(u,v) = \rho_n^{-1} h_{ts,n}(u,v)$, which represents the conditional density of (ξ_i, ξ_j) given that there

[12] The sparse networks that we consider here are restricted to a particular class of networks with $o(n^2)$ edges, or equivalently $o(n)$ network degrees, and do not contain dense spots. This may thus preclude some real-life social networks that exhibit the power-law degree distribution (Borgs et al., 2019).

[13] To be precise, a network is *relatively exchangeable* with respect to the type variable t if

$$[W_{\sigma_t(i)\sigma_t(j),n}] \overset{\text{D}}{=} [W_{ij,n}]$$

for all n and all permutations σ_t satisfying $[t_{\sigma_t(i)}]_{i \in n} = [t_i]_{i \in n}$ (Crane and Towsner, 2018). Exchangeability implies a particular dependence structure across the elements of $W_{ij,n}$. In particular, $W_{ij,n}$ and $W_{i'j',n}$ are dependent if $i = i'$ or $j = j'$. This type of dependence is implied by many statistical and econometric network formation models, such as stochastic blockmodels (Holland et al., 1983), latent position model (Hoff et al., 2002), and other conditional edge independence models (Chandrasekhar, 2016). More details of this framework are available in the handbook chapter by Graham (2020).

is an edge between i and j. The expression $w_{ts,n}$ decouples the network density from the inhomogeneity structure. For the asymptotics, we will assume that $w_{ts,n}(u,v) = w_{ts}(u,v)$, where $w_{ts}(u,v)$ is independent of n. Let $w_t(u,v) = \sum_{s=1}^{T} w_{ts}(u,v)p_s$ and $w(u,v) = \sum_{t,s=1}^{T} w_{ts}(u,v)p_t p_s = \sum_{t=1}^{T} w_t(u,v)p_t$. We will control the rate of the expected degree $\lambda_n = (n-1)\rho_n > 0$ as $n \to \infty$.[14]

The asymptotics of the average degree for the whole network G_n, $d(G_n)$, and our proposed weighted estimators $\widetilde{d}(G_n^{\mathrm{I}})$ and $\widetilde{d}(G_n^{\mathrm{S}})$ can be summarized in the following theorem.[15]

Theorem 1. *Suppose that* $\int_0^1 \int_0^1 w^2(u,v)dvdu < \infty$ *and* $\lim_{n\to\infty} \lambda_n = \lambda < \infty$. *Then,*

(a) *under the population (non-sampled) network,*

$$d(G_n) \xrightarrow{p} \lambda, \quad \sqrt{n}(d(G_n) - \lambda) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2_{d(G)}\right)$$

for some $\sigma^2_{d(G)} > 0$*;*

(b) *under the induced subgraph,*

$$\widetilde{d}(G_n^{\mathrm{I}}) \xrightarrow{p} \lambda, \quad \sqrt{n}(\widetilde{d}(G_n^{\mathrm{I}}) - \lambda) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2_{\widetilde{d}(G^{\mathrm{I}})}\right)$$

for some $\sigma^2_{\widetilde{d}(G^{\mathrm{I}})} > 0$*;*

(c) *under the star subgraph,*

$$\widetilde{d}(G_n^{\mathrm{S}}) \xrightarrow{p} \lambda, \quad \sqrt{n}(\widetilde{d}(G_n^{\mathrm{S}}) - \lambda) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2_{\widetilde{d}(G^{\mathrm{S}})}\right)$$

for some $\sigma^2_{\widetilde{d}(G^{\mathrm{S}})} > 0$.

Theorem 1 establishes that, if the network is sparse, our weighted estimators $\widetilde{d}(G_n^{\mathrm{I}})$ and $\widetilde{d}(G_n^{\mathrm{S}})$ are consistent and asymptotically normally distributed with finite variance.[16] This is the case independently of whether the sampling rates are treated as estimators or not. Section

[14] Specifically, we require $\rho_n = \Theta(1/n)$, i.e., ρ_n grows as fast as $1/n$, so that λ_n converges to a non-zero constant when n goes to infinity.

[15] The proofs of Theorem 1 and Lemma 2 are relegated into the Supplementary Appendix B.

[16] The analytical expressions for the asymptotic variances of $\widetilde{d}(G_n^{\mathrm{I}})$ and $\widetilde{d}(G_n^{\mathrm{S}})$ are complex due to intricate network patterns (Bickel et al., 2011; Bhattacharyya and Bickel, 2015; Graham, 2020) and we therefore leave their derivations for future research. Nevertheless, Bickel et al. (2011) propose subsampling bootstrap methods for approximation and conjecture–although do not prove–that these methods might work properly in sparse networks; see p.2291-2292 of their paper.

4 complements the asymptotic analysis (in Theorem 1 and Supplementary Appendix B) with numerical analysis assessing to what extent the corrections proposed in this section differ from their true values in finite samples.

3.2 Other Network Measures

In addition to the average degree, we also study three other fundamental network measures: the global clustering coefficient, epidemic threshold, and homophily index. We will now provide a brief introduction to these three network measures.

Global Clustering Coefficient. The global clustering coefficient is defined as the ratio between the number of closed triplets $T_c(G_n)$ and the number of connected triples $N_c(G_n)$ in the network (Watts and Strogatz, 1998),[17] calculated as

$$c(G_n) = \frac{T_c(G_n)}{N_c(G_n)}, \tag{12}$$

where

$$T_c(G_n) = \frac{1}{2}\sum_{i\in V}\sum_{\substack{j\in V\\ i\neq j}}\sum_{\substack{k\in V\\ i\neq j\neq k}} W_{ij,n}W_{jk,n}W_{ki,n} \quad \text{and} \quad N_c(G_n) = \frac{1}{2}\sum_{i\in V}\sum_{\substack{j\in V\\ i\neq j}}\sum_{\substack{k\in V\\ i\neq j\neq k}} W_{ij,n}W_{jk,n}.$$

The global clustering coefficient has traditionally been considered a measure of social capital. For example, it plays an important role in risk-sharing (Bloch et al., 2008), trust building (Karlan et al., 2009), job search (Ruiz-Palazuelos et al., 2023), and enhancing cooperation (Granovetter, 1985). Several empirical studies have used the global clustering coefficient as a regressor or a dependent variable (e.g., Fleming et al., 2007; Alatas et al., 2016).

Supplementary Appendix A.1 shows that the naive estimators of the global clustering coefficients (12) under the induced and star subgraphs display biases. We propose their corrections, which differ from those for the average degree: rather than edges connecting dyads

[17] The number of closed triplets also equals three times the number of triangles. A triangle refers to a complete subnetwork of three individuals, which consists of three closed triplets, one centered on each node. A connected triple is a three-node subnetwork in which at least two edges are present. Hence, every triangle is a connected triple, but the reverse is not necessarily true.

(pairs of individuals), we adjust "relationships" involving three individuals, taking into account their interconnections as closed triplets or connected triples, and accounting for the associated sampling probabilities. Nevertheless, research has demonstrated that the global clustering coefficient in (12) approaches zero in sparse networks as $n \to \infty$ (see Supplementary Appendix B.2 for a formal proof; see also Bhattacharyya and Bickel (2015) and Graham (2020) for further discussion). Therefore, the asymptotic analysis of $c(G_n)$ is uninformative. To overcome this issue, we follow the literature, employing a normalized global clustering coefficient which converges to a non-zero value asymptotically and is robust to network size, network density, and degree heterogeneity. In particular, we focus on the normalized coefficient proposed in Li et al. (2019), calculated as

$$c_{norm}(G_n) = \frac{\frac{T_c(G_n)}{3\binom{n}{3}}\left(\frac{nd(G_n)}{2\binom{n}{2}}\right)^3}{\left(\frac{N_c(G_n)}{3\binom{n}{3}}\right)^3} = \frac{(n-2)^2 tr(W_n^3)(\mathbf{1}'W_n\mathbf{1})^3}{n(n-1)(\mathbf{1}'W_n^2\mathbf{1} - tr(W_n^2))^3}, \tag{13}$$

where $\mathbf{1}$ is n-dimensional vector of 1's. It is straightforward to see that the normalization in (13) balances the exponents regarding the network size n and the adjacency matrix W_n in the numerator and denominator. Therefore, as $n \to \infty$, the numerator and the denominator will converge at the same rate.[18] After rearranging, (13) can be expressed as follows:

$$c_{norm}(G_n) = \zeta_n \frac{T_c(G_n)d(G_n)^3}{\left(\frac{N_c(G_n)}{n}\right)^3}, \tag{14}$$

with $\zeta_n = \frac{(n-2)^2}{4n(n-1)}$. Supplementary Appendices A.1 and B.2 analyze the biases in the naive estimators of the normalized global clustering coefficient in (14), provide the corresponding corrections, and show that the weighted estimators are consistent and asymptotically normally distributed as $n \to \infty$.

Epidemic Threshold. There is an increasing interest in understanding the diffusion properties of networks. The epidemic threshold is one way to quantify how easy it is for a disease,

[18] In addition to (13), there is an alternative normalized global clustering coefficient proposed in Bhattacharyya and Bickel (2015), which we discuss in further detail in Supplementary Appendix B.2. We focus on (13) in the main text for the sake of brevity.

information, idea, or behavior to propagate through a network. The applications range from product adoption (Banerjee et al., 2013), spread of information (Alatas et al., 2016) to spread of behaviors (Centola, 2010). There is a large variety of epidemic thresholds, depending on the diffusion conditions and network properties (see, e.g., Vega-Redondo, 2007, and Jackson, 2010). We focus on the following widely used version, based on the mean-field approximation (Pastor-Satorras and Vespignani, 2002):

$$\delta(G_n) = \frac{\frac{1}{n}\sum_{i\in V}\sum_{j\in V} W_{ij,n}}{\frac{1}{n}\sum_{i\in V}(\sum_{j\in V} W_{ij,n})^2}.$$

Supplementary Appendices A.2 and B.3 show that the naive estimators are biased and our proposed weighted estimators are consistent and normally distributed.

Homophily Index. Social and economic networks exhibit a feature called homophily, a tendency to bond with similar individuals. In social and economic networks, who links with whom is typically correlated with characteristics such as gender, age, race, and social and economic status, among others (see McPherson et al. (2001) for a survey). This phenomenon of "birds of a feather flock together" gains particular relevance in our approach because we explicitly consider the types of nodes in the network. Homophily is an important measure of cross-type segregation and affects many economically relevant phenomena such as diffusion or learning and their speeds (Golub and Jackson, 2012), labor market outcomes (Calvo-Armengol and Jackson, 2004), or individual and firm-level success (McPherson and Smith-Lovin, 1987).

We adopt the homophily index from Currarini et al. (2009). The index for type t is defined as $H_t(G_n) = \frac{d_{tt}(G_n)}{d_t(G_n)}$, where $d_{tt}(G_n)$ denotes the average number of friendships that agents of type t have *within* the same type and $d_t(G_n)$ denotes the average number of friendships that type t form regardless of others' types. Supplementary Appendix A.3 contains detailed derivations of the weighted estimator for $H_t(G_n)$ under induced and star subgraphs. Supplementary Appendix B.4 again proves that our weighted estimators are consistent and asymptotically normally distributed.

3.3 Asymptotics of Regressions with Estimated Network Measures

This section discusses the asymptotic properties of OLS regressions in (3), in which our weighted estimators are employed as regressors. Suppose we have R networks. Let n_r denote the number of nodes in the r^{th} network for $r = 1, \cdots, R$. Let $n_r = a_r \cdot n$ and assume that $0 < \varsigma_\ell \leq a_r \leq \varsigma_u < \infty$ for all r with ς_ℓ and ς_u being constants that do not depend on r. That is, we assume that the number of nodes in each network is of the same order. It is well-known that in OLS regressions with covariates being estimated, if the estimating error of the regressor is independent of the regression error ϵ_r and $\max_{r=1,\ldots,R}\{|\widetilde{\Lambda}(\overline{G}_{r,n_r}) - \Lambda_r|\} = o_p(1)$, the estimation effect can be ignored asymptotically. To be specific, let

$$
\begin{aligned}
(\hat{\alpha}_{in}, \hat{\beta}_{in}, \hat{\gamma}_{in}')' &= \arg\min_{(\alpha,\beta,\gamma')} \frac{1}{R}\sum_{r=1}^{R}\Big(y_r - \alpha - \beta\Lambda_r - x_r\gamma\Big)^2, \\
(\hat{\alpha}, \hat{\beta}, \hat{\gamma}')' &= \arg\min_{(\alpha,\beta,\gamma')} \frac{1}{R}\sum_{r=1}^{R}\Big(y_r - \alpha - \beta\widetilde{\Lambda}(\overline{G}_{r,n_r}) - x_r\gamma\Big)^2,
\end{aligned} \tag{15}
$$

where $(\hat{\alpha}_{in}, \hat{\beta}_{in}, \hat{\gamma}_{in}')'$ denotes the infeasible OLS estimator because the true Λ_r's are not observable and $(\hat{\alpha}, \hat{\beta}, \hat{\gamma}')'$ denotes the OLS estimator when the true Λ_r's are replaced with their estimates. If $\max_{r=1,\ldots,R}\{|\widetilde{\Lambda}(\overline{G}_{r,n_r}) - \Lambda_r|\} = o_p(1)$, then

$$
\sqrt{R}\Big((\hat{\alpha}, \hat{\beta}, \hat{\gamma}')' - (\hat{\alpha}_{in}, \hat{\beta}_{in}, \hat{\gamma}_{in}')'\Big) = o_p(1),
$$

i.e., the infeasible OLS estimator and the OLS estimator based on estimated $\widetilde{\Lambda}(\overline{G}_{r,n_r})$'s are asymptotically equivalent. In other words, we can treat $\widetilde{\Lambda}(\overline{G}_{r,n_r})$'s as the true Λ_r's in the regression without the need to correct for the estimation effect of $\widetilde{\Lambda}(\overline{G}_{r,n_r})$'s. The following lemma provides sufficient conditions for $\max_{r=1,\ldots,R}\{|\widetilde{\Lambda}(\overline{G}_{r,n_r}) - \Lambda_r|\} = o_p(1)$.

Lemma 2. *Assume that the variance of $\sqrt{n}(\widetilde{\Lambda}(\overline{G}_{r,n_r}) - \Lambda_r)$ is uniformly bounded above by a finite constant M, for all r and $n \geq N$ for some finite large number N, and $R/n \to 0$. Then,* $\max_{r=1,\ldots,R}\{|\widetilde{\Lambda}(\overline{G}_{r,n_r}) - \Lambda_r|\} = o_p(1)$.

4 Monte Carlo Simulations

This section complements the previous one in assessing the performance of our approach in finite samples. In particular, we evaluate numerically the estimation biases in the network measures under study (in Sections 4.1), as well as the network effects when using these measures as regressors in regression analysis (in Section 4.2). The evaluation considers various factors such as the sampling design (induced vs. star subgraph), the sampling rate, and whether SRS (representativeness) is assumed when applying the weighted estimators. We quantify the biases present in the naive estimators and the corrections made under the SRS assumption and compare their performances *vis-à-vis* our post-stratification estimators. For ease of interpretation, we concentrate on the scenarios that mimic our modeling assumptions.

In this simulation exercise, we demonstrate the effectiveness of our post-stratification approach by analyzing the network measures discussed in Section 3. These measures include the average degree, global clustering coefficient, normalized global clustering coefficient, epidemic threshold, and homophily index.[19] The network data in our simulation study are adopted from the Add Health Wave-I In-school data.[20] In particular, we adopt one school as a prototype.[21] By adopting the real-life friendship network in that school, we can preserve certain relationships between students' characteristics and the friendship network.[22] For example, white students tend to have more friends than black students, while black students have more connections than other racial groups. Additionally, the prevalence of homophily patterns varies systematically based on the racial makeup of each school (Currarini et al., 2009).

[19]We include both the standard global clustering coefficient and its normalized variant. Our corrections of the latter are asymptotically well-behaved and we would like to assess its performance in finite samples. However, the (non-normalized) coefficient is widely employed in the literature. Hence, although we know it converges to zero in sparse networks asymptotically (see Section 3.2), we analyze its performance in finite samples.

[20]This is a program project designed by J. Richard Udry, Peter S. Bearman, and Kathleen Mullan Harris, and funded by a grant P01-HD31921 from the National Institute of Child Health and Human Development, with cooperative funding from 17 other agencies. Special acknowledgment is due Ronald R. Rindfuss and Barbara Entwisle for assistance in the original design. Persons interested in obtaining data files from Add Health should contact Add Health, Carolina Population Center, 123 W. Franklin Street, Chapel Hill, NC 27516-2524 (addhealth@unc.edu). No direct support was received from grant P01-HD31921 for this analysis.

[21]This adopted school is a public suburban school with 1606 students from grades 9 to 12. The school is located in the southern U.S.

[22]In Add Health In-School survey, each participant can nominate up to five male and five female friends, and we use this nomination information to build their friendship network. These friendship links are treated as undirected, i.e., there is a link between i and j as long as i nominates j or j nominates i.

For ease of interpretation, we prune this prototype to the size of 1,500 so that the numbers of whites, blacks, and other races are all equal to 500. We use three demographic characteristics from the prototype to define an individual's type: seniority (C1), gender (C2), and race (C3). Seniority takes the value of one if an individual is older than the school average and zero otherwise. For gender, one stands for male, and zero for female. For race, one represents individuals who identify as white, two represents those who identify as black, and three represents individuals from other racial backgrounds. We also combine these three characteristics to form $2 \times 2 \times 3 = 12$ cross-characteristics, such as "senior black female," "junior male of other race," and so on, denoted by *Cross* throughout. Seniority and gender are largely uncorrelated with individual network connectivity. In contrast, race is strongly correlated with network degree in the data. The average network degree is 9.60 for white students, 7.38 for black students and 4.39 for students of other races. Also note that the homophily index can be calculated for various types of characteristics, including cross-characteristics. For illustration, we will only focus on the homophily index of blacks in this simulation study.

4.1 Network Measures

As a first step, we quantify the biases in the naive estimators (such as (4) and (6) for the average degree), corrections assuming representativeness (e.g., (8) and (9)), and our proposed weighted estimators (e.g., (5) and (7)).

From the "population" network data described above, we generate 1,000 sampled networks using different removal schemes, which vary in three dimensions. First, we mimic two network sampling designs, i.e., induced and star subgraph sampling. For the induced subgraphs, we remove a fraction of nodes and all of their links. For the star subgraphs, we remove a fraction of nodes and only their links to other removed nodes. Second, we consider three removal rates $(1 - \psi)$, which are 20%, 40%, and 60%, corresponding to sampling rates (ψ) of 80%, 60%, and 40%, respectively. Third, we employ four removal scenarios to reflect representative or non-representative sampled networks. The first scenario (scenario R) removes nodes completely at random, which mimics SRS and produces representative samples. The other

scenarios incorporate disproportional removals to produce non-representative samples, where the probability of node removal is linked to the node's connectivity. Specifically, scenario H considers the removal of high-degree nodes with a higher probability, scenario M removes intermediate-degree nodes with a higher probability, and scenario L removes low-degree nodes with a higher probability. Due to a strong correlation between race and network degree in the prototype data, we implement disproportional removals based on race. For instance, to remove highly connected nodes with a higher probability, we remove white students with a higher probability, and so forth.[23]

We perform 1,000 Monte Carlo repetitions. Figures 1 and 2 summarize the directions and average magnitudes of the biases (in percentage terms) out of 1,000 simulation repetitions with respect to the whole network values for the induced and the star subgraph, respectively. In these two figures, the x-axes list the five network measures under scrutiny in the following order: average degree, global clustering coefficient, normalized global clustering coefficient, epidemic threshold, and homophily index of blacks. The blue bars in Figures 1 and 2 represent the *Raw* unweighted sample statistics, and the red bars, denoted SRS, reflect the corrections based on the representativeness assumption. The remaining three other colored bars are our weighted estimators. The green bars weight on the network-unrelated characteristic $C1$, whereas the last two bars represent, respectively, the weighting on $C3$ (dark red) and the weighting on *Cross*, i.e., the combination of C1 to C3 (gray).[24] The rows and columns represent, respectively, the four different removal scenarios, R, H, M, L, and the three removal rates, 20%, 40%, and 60%, in these orders.

Biases in Raw (Unweighted) Sample Statistics. We first discuss the biases that emerge in the network measures when computing raw sample statistics and emphasize the impact of scaling and non-representativeness. This exercise reveals that treating the data "as if" complete leads to considerable differences between the whole and sampled networks under almost all removal scenarios. The biases are more prominent in the induced subgraph due to

[23]Specifically, the amounts of removal for (white, black, and other races) are $(\frac{1-\psi}{2}, \frac{1-\psi}{3}, \frac{1-\psi}{6}) \times 1500$ in scenario H, $(\frac{1-\psi}{4}, \frac{1-\psi}{2}, \frac{1-\psi}{4}) \times 1500$ in scenario M, and $(\frac{1-\psi}{6}, \frac{1-\psi}{3}, \frac{1-\psi}{2}) \times 1500$ in scenario L.

[24]Weighting on C2 performs very similarly to weighting on C1 and is thus omitted.

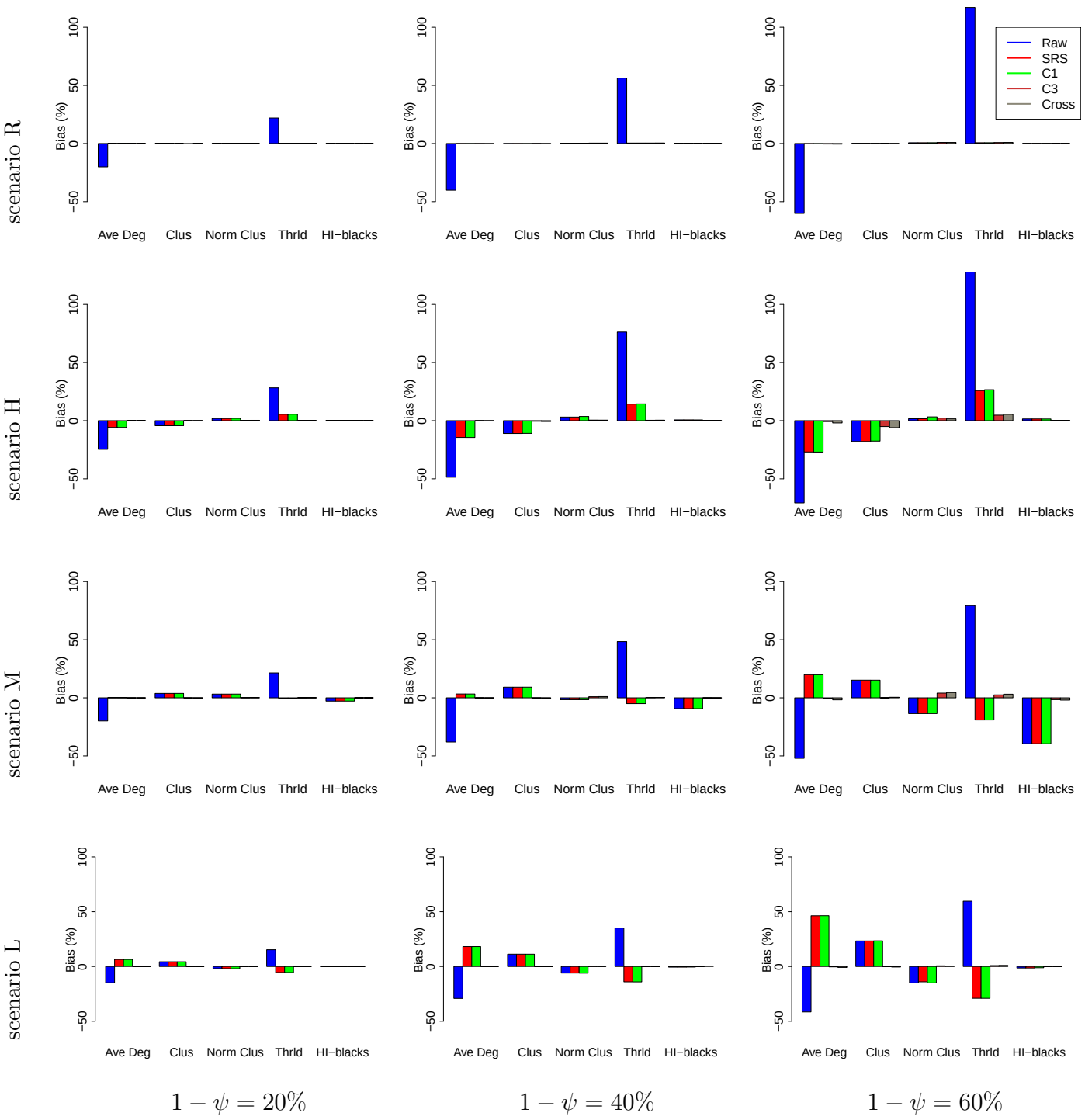

Figure 1: Induced subgraph. Biases (%) in five estimated network measures (average degree, global clustering coefficient, normalized global clustering coefficient, epidemic threshold, and homophily index of blacks) from the raw sample and their corrected versions by weighting for three different removal rates 20% (left), 40% (center), 60% (right) and four different removal scenarios (R, H, M, and L). The bias is computed by subtracting the whole network value from the average across 1,000 simulation repetitions. Raw indicates an unweighted sample statistic, SRS signifies the correction based on the representativeness assumption, and C1, C3, and Cross denote the weighting on the respective characteristic variables.

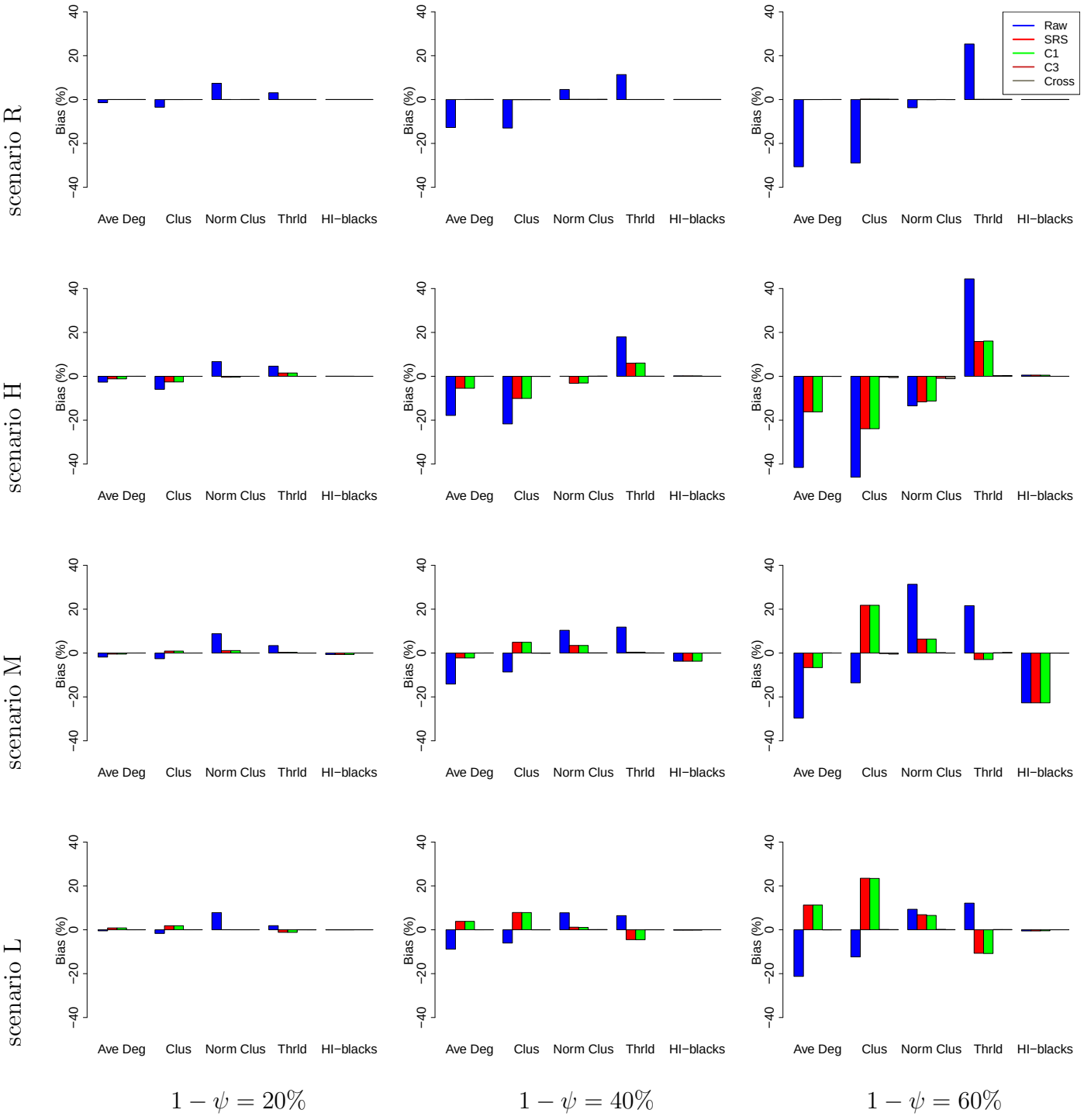

Figure 2: Star subgraph. Biases (%) in five estimated network measures (average degree, global clustering coefficient, normalized global clustering coefficient, epidemic threshold, and homophily index of blacks) from the raw sample and their corrected versions by weighting for three different removal rates 20% (left), 40% (center), 60% (right) and four different removal scenarios (R, H, M, and L). The bias is computed by subtracting the whole network value from the average across 1,000 simulation repetitions. Raw indicates an unweighted sample statistic, SRS signifies the correction based on the representativeness assumption, and C1, C3, and Cross denote the weighting on the respective characteristic variables.

less information available about the network conditional on the sampling rate, and decrease (increase) with the sampling rate (removal rate).

The average degree and the epidemic threshold are two measures that exhibit particularly significant biases that increase dramatically with the removal rate. Sampled networks, which have fewer observed links than the whole network, consistently appear less connected and less prone to epidemic spread than they actually are. The magnitude of the biases varies depending on which nodes are removed. When highly connected nodes are not observed, the largest biases are detected. Biases are smaller but still significant when more nodes with intermediate connectivity are removed, followed by the removal of more low-degree nodes. SRS leads to biases that are comparable to the removal of nodes with intermediate degrees.

In case of the global clustering coefficient and homophily index, the biases exhibit more intricate patterns. The biases tend to increase with higher removal rates. However, they are not necessarily lower in the star subgraph, and their magnitudes and signs are highly dependent on the specific removal scenarios employed. This is consistent with what we noted earlier, namely that the direction of bias is difficult to predict a priori. In the induced subgraphs, the global clustering coefficient is barely biased (less than 1%) under SRS. Under disproportional removals, the direction of bias is contingent upon which individuals are more likely to be removed. Removing highly connected nodes with higher probability (scenario H) drives the global clustering coefficient down, while scenarios M and L inflate it. Unlike induced subgraphs, biases in the global clustering coefficient under star subgraphs are consistently negative (even in representative samples). To explain this observation, consider that nodes are sampled representatively under SRS. Under the induced subgraphs, as all links of all nodes in the sample are observed, there is no systematic bias. However, in a star subgraph, when two or more neighbors of a sampled node are not observed, their links to the sampled node will be included in the sampled network but their mutual connections will not. This situation systematically decreases the global clustering coefficient of the sampled network, and this effect becomes more pronounced as the sampling rate decreases. These biases are always the largest in absolute terms in scenario H where high-degree nodes are removed with higher

probability and lower in scenarios M and L. Additionally, note that non-representative samples may generate lower biases in the global clustering coefficient than representative samples.

The homophily index of black students exhibits no biases whatsoever (less than 0.20%) in scenario R. This explains why the literature has ignored the effects of sampling on homophily. However, disproportional removals of different types lead to mismeasured values. Figures 1 and 2 report downward biases in scenario M, where blacks are more likely to be removed, and these biases increase with the removal rate. Scenarios H and L lead to more moderate biases.

Biases in Weighted Estimators. The second objective of this section is to evaluate the effectiveness of two weighted estimators in correcting biases. The first is based on the assumption of SRS, while the second is our post-stratification estimator, which involves three variations of weights based on three auxiliary variables. It should be noted that seniority (C1) is generally uncorrelated with connectivity, while race (C3) is strongly associated. We first examine the weight based on C1, expecting it to correct, at least partially, the bias resulting from the scaling effect. Next, we consider the weight based on C3, which is expected to correct the bias arising from the problems of scaling and non-representativeness. This is due to C3 serving as the variable used in designing the removal scenarios and correlates with one's network position. Finally, we apply weights based on *Cross* (i.e., the combination of C1, C2, and C3). We hypothesize that weighting on *Cross* should outperform the previous two cases as it utilizes all available information.

With only a few exceptions (none involving our preferred strategy), all weighted estimators yield either smaller or equal biases when compared to the raw sample statistics. As expected, the weighted estimator designed for SRS performs well in scenario R and our post-stratification estimator yields similar results. However, we observe divergent performances of the two weighted estimators under scenarios H, M, and L. First, the weighting based on SRS and the weighting on the network-irrelevant variable C1 show similar overall performance. This is an important finding as it indicates that, regardless of which variable is used for weighting, our post-stratification estimator mitigates potential biases and does not produce worse results than raw sample statistics and the weighted estimator based on SRS. However,

certain biases remain and they increase with the missing rate. Most important, the biases are virtually eliminated once we weight on either C3 or *Cross*.

Root Mean Square Errors. Figures D.1 and D.2 in Supplementary Appendix D provide the corresponding normalized root mean squared errors (RMSEs). These RMSEs reflect both the average biases and variances, allowing us to assess whether there is a bias-variance trade-off when employing the proposed weighted estimators. The RMSEs support the conclusions drawn based on the average biases. Notably, our preferred strategy, which weights on the variable *Cross*, outperforms all other strategies in terms RMSEs with two exceptions involving the clustering coefficient. Specifically, scenarios H and M under the induced subgraph and $\psi = 40\%$ generate larger RMSEs of the global clustering coefficient weighted on *Cross* compared to raw network statistics and the corrections based on the SRS assumption. Although our corrections lead in these two cases to essentially unbiased estimators on average, their variances are larger for high missing rates. Therefore, we conclude that our estimators do not come at the cost of larger variances overall. Nevertheless, scholars should be careful while recovering the global clustering coefficient from non-representative samples under high missing rates. In these cases, our methodology eliminates the biases, but a certain bias-variance trade-off exists.

Drawing from our simulation results, the proposed approach generally improves inference on sampled networks by providing estimates of network effects that are less biased and more stable compared to using raw sample statistics or the weighted estimator that only targets the scaling effect. Under our approach, we still expect certain attenuation when our weighted estimators are employed as regressors and the biases should be more pronounced in the induced subgraphs. The next section analyzes these conjectures.

4.2 Network Effects

We now focus on the performance of our weighting approach in a regression framework, aiming at estimating the impact of global network features on economic outcomes. Our weighting approach is specifically designed for global network measures, so the dependent variables in our regressions are measured at the network-wide level, such as the mean, median, or other

statistics computed from individuals' outcomes in the network. However, as discussed in Section 2, our method can also be extended to models where individual behaviors or outcomes are directly regressed on global network properties, or in network formation applications where the network properties are used as regressands.

To generate the population network data, we start with the same prototype network from the pruned Add Health school sample with 1,500 students. We create 200 artificial networks, each with 1,500 nodes, using the node characteristics adopted from the prototype network. Based on the average connectivity, global clustering coefficient, and homophily index of different types of individuals in the prototype network, we simulate network links in these 200 artificial networks. This results in 200 networks that have the same size and node characteristics (i.e., C1, C2, and C3), but different network configurations. Notably, the simulated links exhibit uneven connectivity across types, with white nodes having the highest average degree, followed by blacks and other races. There are also clustering and homophily features among different types. To simulate the dependent variable y_r in each network, we follow a simple linear regression model: $y_r = \alpha + \beta\Lambda(G_{r,n_r}) + \varepsilon_r$, $r = 1, \cdots, 200$, with ε_r being an i.i.d. random error from $N(0, 1)$. We generate the data with designed parameters $\alpha = 1$ and different β's corresponding to different network measure $\Lambda(G_{r,n_r})$: $\beta = 0.5$, 2, -2, and -0.5 for the average degree, global clustering coefficient, epidemic threshold, and homophily index of blacks, respectively. Finally, for each of the whole networks, we generate 1,000 artificial samples following the same removal strategies as described in the previous section. This generates the raw sample data, on which we apply corrections based on the assumption of SRS and our post-stratification approach. We estimate the regression model based on five cases, which are the raw sample statistics, corrections based on SRS, and post-stratification weighting on C1, C3, and *Cross*, and compare them to the estimates based on the whole network.

Figures 3 and 4 display the average biases in the estimates of β for the induced and star subgraphs, respectively. The y-axes represent the biases in percentage terms, while the x-axes list the five network measures under study. Again, the blue bars represent the raw sample statistics, the red bars represent the corrections based on the assumption of SRS, and the

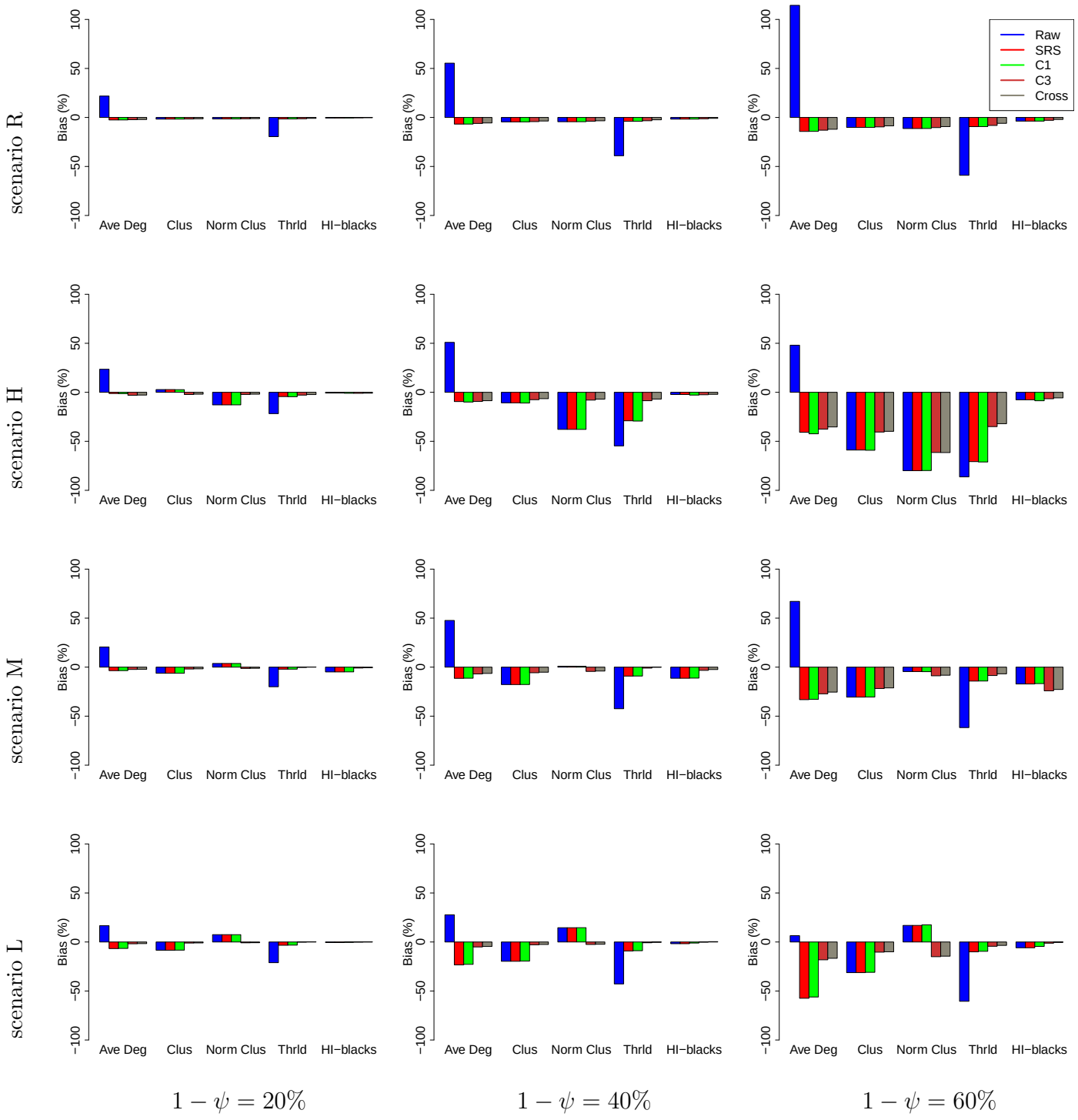

Figure 3: Induced subgraph. Biases (%) in five estimated network effects (average degree, global clustering coefficient, normalized global clustering coefficient, epidemic threshold, and homophily index of blacks) from the raw sample statistics and their corrected versions by weighting for three different removal rates 20% (left), 40% (center), 60% (right) and four different removal scenarios (R, H, M, and L). The bias is computed by subtracting the whole network value from the average across 1,000 simulation repetitions. Raw indicates an unweighted sample statistic, SRS signifies the correction based on the representativeness assumption, and C1, C3, and Cross denote the weighting on the respective characteristic variables.

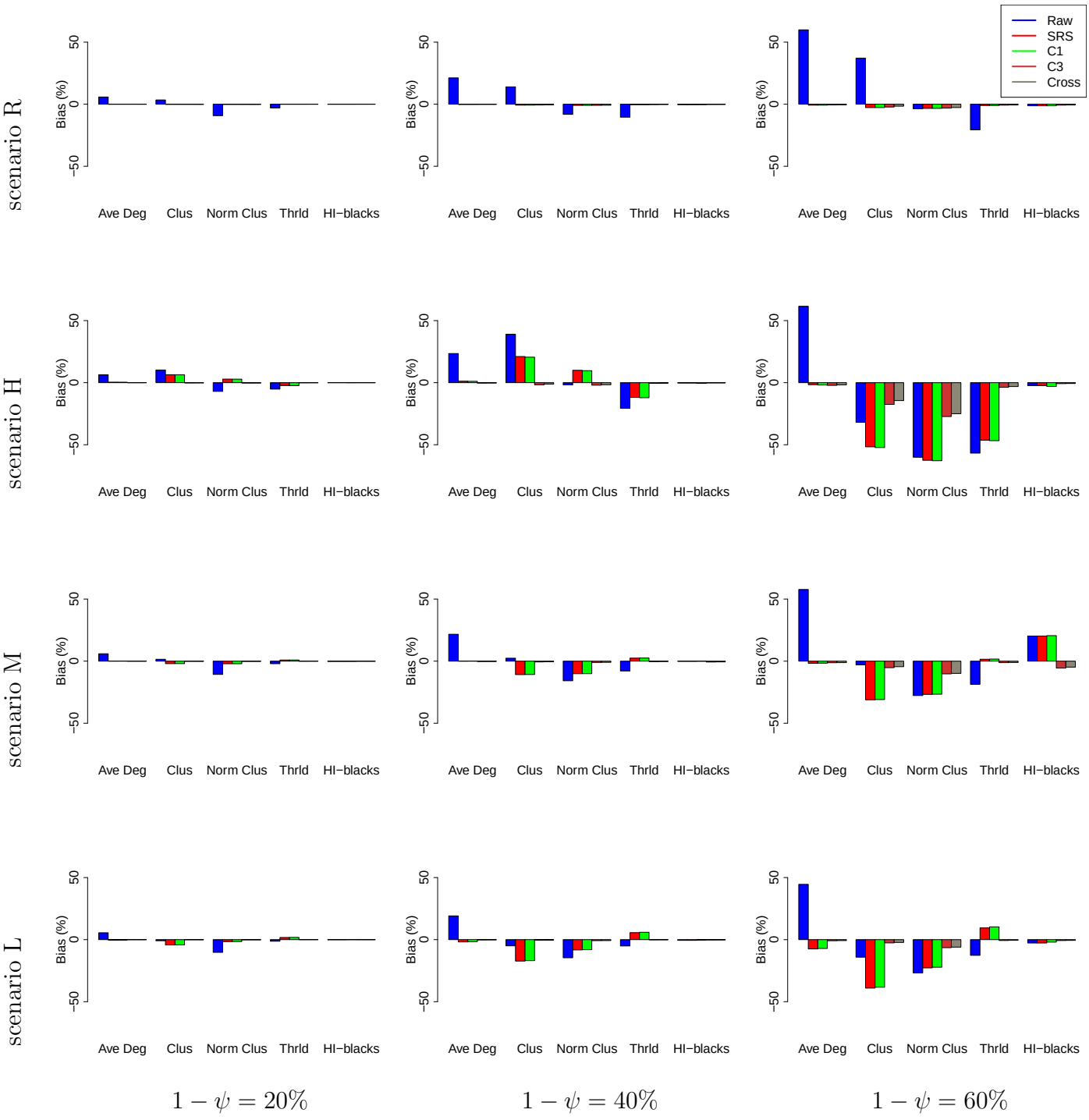

Figure 4: Star subgraph. Biases (%) in five estimated network effects (average degree, global clustering coefficient, normalized global clustering coefficient, epidemic threshold, and homophily index of blacks) from the raw sample statistics and their corrected versions by weighting for three different removal rates 20% (left), 40% (center), 60% (right) and four different removal scenarios (R, H, M, and L). The bias is computed by subtracting the whole network value from the average across 1,000 simulation repetitions. Raw indicates an unweighted sample statistic, SRS signifies the correction based on the representativeness assumption, and C1, C3, and Cross denote the weighting on the respective characteristic variables.

green, dark red, and gray bars represent the weighting on C1, C3, and *Cross*, respectively. Figures D.3 and D.4 in Supplementary Appendix D plot the corresponding normalized RMSEs.

The findings regarding biases in network effects are consistent with those observed in network measures in Section 4.1. The estimates, whether based on raw samples or corrections, exhibit increased biases as the missing rate increases. The biases are more pronounced in induced than star subgraphs, and the magnitudes are influenced by the pattern of missingness.

When raw sample statistics are employed, the estimates of all network effects display serious biases. Specifically, the impact of the average degree is consistently overestimated and the biases are dramatic. The bias in the estimated influence of the global clustering coefficient depends on the configuration. In most cases, the effects are attenuated and the magnitude of the biases is considerable. The biases in the estimated effects of the epidemic threshold and the homophily index are always attenuated; the attenuation is often substantial in the case of the former and generally moderate for homophily.

As for the corrections based on weighted estimators, we first discuss the induced subgraph sampling scheme. Almost without exception, all the corrections eliminate the expansion problem, but some negative biases persist. Both the corrections assuming representativeness and our post-stratification approach mitigate the biases in the uncorrected estimators. More importantly, our methodology, which takes into account the non-representativeness of the sample, generally leads to lower biases than the corrections that solely target the scaling effect. The RMSEs reported in Supplementary Appendix Figure D.3 corroborate all these findings.

The network statistics derived from star subgraphs consistently display lower biases compared to those obtained from induced subgraphs, and this pattern also holds when the estimated network statistics are applied as regressors. Based on the raw sample statistics, the effect of average degree tends to be overestimated, the effect of global clustering coefficients can be either overestimated or underestimated depending on the configuration, and the effects of other network measures are generally attenuated. With few exceptions, the biases in the estimates of network effects using the uncorrected estimators are mostly below 5%, 20%, and 60% for $\psi = 80\%$, 60%, and 40%. Both types of corrections reduce the biases drastically.

Our preferred strategy consistently outperforms corrections that assume representativeness, especially in recovering the true effect of the global clustering coefficient. Overall, our approach is remarkably successful in recovering the true network effects under the star subgraph sampling, a conclusion that is even reinforced if we look at the RMSEs in Supplementary Appendix Figure D.4.

5 Empirical Application

In this section, we illustrate how statistical inference with field data can be affected if one does not account for non-representativeness of network samples. To that aim, we apply the proposed methodology to the data from Banerjee et al. (2013) who elicit a large variety of characteristics, including social networks, from 75 villages in southern Karnattaka, India. The authors initially collected the census information for each household in all villages. Subsequently, they conducted a comprehensive follow-up survey with a subset of each village, wherein they also recorded the networks of relationships among surveyed individuals. As is common in most studies, the survey respondents only represent a sample of each village, and their reported network is an induced subgraph of the whole network. The average sampling rate across villages is 35%. The crucial aspect of the sampling design in Banerjee et al. (2013) is the stratification by religion and geographic sub-location, generating a representative sample with respect to these two variables. This is a common approach in many applications. Despite the stratification based on religion and geography, Table 1 reveals that the data are not representative in terms of age, gender, and–to a lesser extent–household size. Below, we show to what extent the differences between the village population and sample shares of these categories affect the estimation of network effects in regressions discussed in Section 2.2.

Table 1: Population and sample shares of different characteristics and labor market outcomes in the Indian rural village data from Banerjee et al. (2013).

	Population	Sample	Diff. (p-value)
Age			
< 30	38.71%	30.97%	7.74% (0.000)
30 - 50	39.60%	54.11%	-14.51% (0.000)
> 50	21.69%	14.92%	6.77% (0.000)
Male	50.34%	44.57%	5.77% (0.000)
Household size			
< 3	17.26%	15.49%	1.77% (0.038)
3 - 8	71.57%	73.48%	-1.91% (0.039)
> 8	11.17%	11.03%	0.14% (0.879)
Labor market outcome			
employed		62.49%	
work outside village		21.21%	
Number of villages	75	75	
Observations	48,646	16,995	

The data contain several variables regarding the labor market outcomes of the participants, such as their employment status, whether they work outside the village, and their occupation. Since the important role of social networks in labor markets is widely acknowledged (Granovetter, 1985; Calvo-Armengol and Jackson, 2004; Cingano and Rosolia, 2012), we ask how the village employment rate and the fraction of people working outside the village correlate with the global features of the underlying network of relationships within the village.[25] Theoretical literature suggests that both connectivity and the global clustering coefficient can have a direct impact on employment prospects (Calvo-Armengol and Jackson, 2004; Ruiz-Palazuelos et al., 2023). Additionally, the epidemic threshold can indirectly influence labor outcomes by affecting the flow of labor-market information (Calvo-Armengol and Jackson, 2004). Similarly, the degree of segregation can determine which individuals have access to job information and those who do not. Most importantly, for the present study, we ask how the estimated network effects change if we account for non-representativeness of the network sample. We hypothesize

[25]To maintain simplicity and align better with the assumptions of our analysis, we focus on a simpler application compared to Banerjee et al. (2013), who propose a more intricate estimation strategy.

that the over-representation of individuals aged 30-50 and the under-representation of men in the sample (as evident in Table 1), who are typically more active participants in labor markets in a country like India, could bias the estimated network effects if this misrepresentation is not taken into account.

Table 2 reports the estimated network effects in a series of regressions differing in (i) the dependent variable (employment rate or fraction of working outside the village), (ii) whether raw sample statistics or corrections are used and (iii) different network measures. As for (ii), to separate the effect of scaling from the effect on non-representativeness of the sample, we use the naive estimators (denoted *Raw* in Table 2), corrections assuming SRS (denoted *SRS*), and our approach in which we weight on cross-characteristics (incorporating the information on age, gender, and household size; denoted *Cross*). Table 1 illustrates the distributions of these three variables, from which we compute the $\hat{\psi}_t$ for the $3 \times 2 \times 3 = 18$ types according to the variable *Cross*. Each row reports the estimated network effect (and the standard error robust to heteroskedasticity in parentheses) from a separate regression of one dependent variable on the corresponding network statistic and village size, mimicking the structure of the regressions in Section 2.2. We also apply the post-stratification weighting on the dependent variables (i.e., employment and working outside villages) at the village level to correct measurement errors.[26] Consequently, the columns *Cross* provide a typical example of standard post-stratification with a reasonable number of stratification groups, where the sampling rates are estimated from the differences between the sample and population shares of auxiliary variables. Since we show that our approach delivers consistent estimates, we believe that applied researchers should report estimates such as those in the columns *Cross* as their main result while estimating the effect of network measures on outcomes in non-representative samples or, at least, as a robustness check of their main analysis.

[26] We use the network constructed by the union of all relationships reported by survey respondents (e.g., borrowing, lending, seeking advices, going to temple together, visiting home, etc.). We find similar results if we only focus on friendships (see Table C.1 in the Supplementary Appendix).

Table 2: Estimated network effects on the labor market outcomes of villagers in rural India villages

Dependent Variable	(I) Employed (%)			(II) Work Outside Village (%)		
	Raw	SRS	Cross	Raw	SRS	Cross
Average 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)
Global Clustering	0.4989**	0.4989**	0.4240**	-0.6410**	-0.6410**	-0.4996***
	(0.1930)	(0.1930)	(0.1830)	(0.2666)	(0.2666)	(0.1879)
Norm. Global Clustering	-0.0047	-0.0047	-0.0022	0.0038	0.0038	0.0037
	(0.0061)	(0.0061)	(0.0051)	(0.0048)	(0.0048)	(0.0054)
Epidemic Threshold	-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-male	0.1939*	0.1939*	0.1445	-0.1374	-0.1374	-0.0463
	(0.1028)	(0.1028)	(0.0939)	(0.1490)	(0.1490)	(0.1621)
HI-middle age	0.2848	0.2848	-0.2386	-0.5150**	-0.5150**	0.0010
	(0.1956)	(0.1956)	(0.2119)	(0.2033)	(0.2033)	(0.2428)
HI-small household size	0.0930	0.0930	0.1866**	-0.2677**	-0.2677**	-0.0818
	(0.0991)	(0.0991)	(0.0856)	(0.0992)	(0.0992)	(0.0920)

Note: Regressions are based on 75 villages. Standard errors robust to heteroskedasticity are reported in parentheses. *, **, *** stand for significance at 10%, 5%, and 1% respectively. Each row represents a separate regression with a different network measure, and the village size is included in every regression as a default control. Raw indicates an unweighted sample statistic, SRS signifies the correction based on the representativeness assumption, and Cross denotes the weighting on the *Cross* characteristic variable.

As for the influence of village networks on labor market outcomes, our findings support existing literature, highlighting the significant role played by the structure of social networks in shaping labor markets. By accounting for the non-representativeness of the sample (as indicated by the *Cross* columns in Table 2), certain features of the social networks have a meaningful impact on average labor outcomes within the village. Moreover, the effects of these various network characteristics largely exhibit consistency with one another.

Regarding the main purpose of this exercise, Table 2 shows the sensitivity of the results with respect to (non-)representativeness of network samples. In contrast to Section 4, we

do not know the true impact of the different network measures. However, since the data and the performed regressions match the assumptions behind our approach, all the previous analysis suggests that the results using our methodology are consistent, less biased, and more stable than either the naive estimators or corrections assuming representativeness. As a result, the following discussion provides an informal assessment of the disparities among the results obtained from naive estimators, the corrections assuming SRS, and our approach.

Table 2 documents that the estimates using raw data or corrections based on the representativeness assumption are mostly expanded compared to the corrections that account for both scaling and the non-representativeness of the network data. However, we also observe instances of attenuation and even sign-switching. There are three cases in which we observe a network effect when employing the naive estimators or corrections under SRS, but this effect does not show up using our weighting approach. In one other case, the network effect is absent with the naive estimators and corrections for scaling, but this effect becomes significant in the *Cross* column. All these four cases are associated with the impact of homophily. Quantitatively speaking, the effect of the average degree, when based on raw data, is overestimated in Table 2 by over 130% compared to the effect observed through our weighting approach. Likewise, the effect of the global clustering coefficient is overestimated by more than 17%, while the effect of the epidemic threshold is underestimated by at least 45%. Hence, some of these differences are economically significant. The corrections assuming representativeness either alleviate or maintain the biases when compared to the results obtained through our approach. These corrections effectively reduce the biases with respect to the *Cross* column to below 10% for the average degree and epidemic threshold. However, the biases remain economically significant for network measures that are unbiased in representative samples but generally biased in non-representative samples, such as the clustering coefficients and the homophily indices.

In sum, significant differences are present between the estimates obtained using our approach and those from the naive estimators as well as the corrections assuming representativeness. These findings suggest that false positives (or negatives), expansion of network effects, and sign switching might be common phenomena resulting from non-representativeness of net-

work samples. Given that most network data share the underlying properties of this data, the results here imply that applied researchers should consider the effect of weighting on the sign, size, and magnitude of network effects. More importantly, the direction and the magnitude of the biases depend non-trivially on the particular network statistics, the dependent variable under study, and who is missing. Hence, this exercise corroborates that researchers cannot easily predict the direction of the biases and consequently, they should not rely on classical measurement-error solutions, even in the simplest cases analyzed here.

6 Discussion

This section discusses potential extensions and limitations of our methodology and provide several recommendations concerning the selection of auxiliary variables for weighting.

Alternative Network Sampling Designs. Although this paper focuses on the induced and star subgraphs, the proposed methodology can be adapted to other sampling schemes as long as the researcher knows the strategy employed for the elicitation of the sample and possesses some information about the whole population. We present several examples illustrating how the proposed approach can be applied to different sampling strategies and discuss cases where our methodology cannot be directly applied, or requires modification.

As a first example, consider the issue known as the boundary specification problem. Researchers sometimes set a boundary to determine the whole network of interest. Imagine a researcher who collects a network sample from a few classes within a school, excluding individuals from other classes and any connections between the classes under investigation and individuals outside the class. Although the sampled network may provide a comprehensive representation of the analyzed classes, it remains incomplete in capturing the entirety of the true social network within the school. If one would like to study the school network, and individual characteristics are available for the whole school, one can mitigate the boundary specification problem by applying our method directly because setting a boundary is mathematically equivalent to the induced subgraph sampling.

As a second example, consider snowball sampling, a sampling procedure commonly applied

in Sociology, Marketing, and Epidemiology (see, e.g., Berg (2004); Browne (2005)). In snowball sampling, a researcher begins by randomly selecting seed nodes. These seeds serve as the starting point for the first wave, during which the researcher collects information on all the contacts of the initially selected nodes. In subsequent waves, the researcher expands the sample by eliciting the contacts of the nodes identified in the previous wave, and this process continues iteratively. Note that conducting a one-wave snowball sampling is essentially equivalent to the star subgraph sampling approach discussed earlier, thus making our methodology directly applicable. The literature has suggested corrections for one-wave snowball sampling (Frank, 1977; Kolaczyk, 2009), but these corrections only align with our approach when the initial seeds are representative samples of the population. We argue this is rarely the case even in very carefully and systematically collected data sets. Although the computation becomes increasingly complex as more waves are performed, one can adapt our approach to multiple waves of snowball sampling taking into account the missing frequencies of each type and the information about the within-type and across-type connectivity from the observed part of the network using combinatorial arguments. In fact, our methodology has certain parallelism with Respondent Driven Sampling (Heckathorn, 1997), a weighting approach on snowball samples to compensate analytically for the non-randomness of snowball-sampling procedures. In contrast to this approach that corrects for the non-representativeness *ex-ante*, our approach adjusts for these issues *ex-post* by mitigating the discrepancy between the sampled and population networks and treating the sampling rates as estimators.

Unsurprisingly, our corrections cannot be applied to some alternative sampling designs or should be tailored to the specific sampling strategy employed in the corresponding study. Consider, for example, random selection of links (also known as random edge sampling) where an individual i is included in the sample if at least one of her edges is sampled. Such sampling is commonplace in communication data, where only random samples of phone calls or e-mails are selected. We do not target this procedure in this study as additional assumptions would be necessary, but see, e.g., Kolaczyk (2009) for a potential direction. Relatedly, our approach assumes that, conditionally on observing a particular sample of nodes and the

sampling design, the links are observed perfectly. That is, this study specifically analyzes issues arising from imperfect observation of network members but cannot solve issues arising from mismeasured links (see, e.g., Hardy et al. (2019)). A notable example of this issue is the truncated fixed-choice survey design, where respondents are constrained to nominate a certain number of friends (e.g., up to ten friends). Our approach mitigates the biases due to the non-representativeness but not those due to the truncation. However, both issues might be targeted simultaneously by combining our post-stratification weighting with the approach proposed by Griffith (2022), which is specifically designed to mitigate the issues due to the truncation. Similarly, additional applications of our approach might result from combining our approach with methods designed for other purposes. The extension of our approach to these other more specialized sampling procedures is left for future research.

Other Network Measures. Due to their theoretical and empirical relevance, this study focuses on four fundamental network measures commonly seen in the empirical literature. Nevertheless, one can adapt the methodology to other measures that solely require the knowledge of nodes' local information.[27] The first set of examples allows for a direct application of our methodology, which includes the assortativity coefficient and the average size of the second-order neighborhood. Assortativity plays a crucial role in the process of diffusion, as it can either impede or facilitate the transmission of diseases, behaviors, and social norms (Newman, 2002; Jackson et al., 2017). The average size of the second-order neighborhood enables us to assess how fast diffusion spreads, and it is important in labor markets (Calvo-Armengol and Jackson, 2004). Since the computation of both the assortativity coefficient and the second-order neighborhood only requires the knowledge of an individual's degree and the degrees of their neighbors, their weighted corrections follow directly from Section 3.

Other measures do not follow directly from Section 3, but our approach can still be applied. For instance, Eagle et al. (2010) apply the concept of entropy to capture the diversity of connections of an individual to different types in the network. Since their measure only relies on the neighborhood of each node, the corrected variation of this measure for sampled

[27] In this paper, local information always refers to the first- and second-order neighborhoods of each node. One can go further and incorporate more distant neighbors probably at the cost of lower precision of the proposed corrections.

networks is straightforward. Similarly, cycles of length four have recently received certain attention in sociology (Opsahl, 2013) and economics (Ruiz-Palazuelos et al., 2023). One can recover it following our approach using the combinatorial logic. Since these characteristics are extensions of the ideas of homophily and the global clustering coefficient, respectively, we focus on the more common variations and do not propose the corrections of these two in this study.

The proposed methodology cannot recover global network measures computed based on the entire network architecture. This includes spectral properties, average betweenness or eigenvalue centrality, and network distances. However, there is a rich literature proposing approximations, bounds, or "plug-in" estimators computed on the basis of nodes' local information (e.g., Van Mieghem, 2010; Comellas and Gago, 2007). Hence, one can correct these bounds and approximations using our approach either directly or by plugging some of our corrections into more general expressions. Future research shall establish the finite-sample as well as asymptotic properties of such bounds, approximations, and plug-in estimators. The proposed approach cannot recover the network characteristics at the individual node level.

Selection of (Auxiliary) Weighting Variables. A natural question arising from the proposed methodology is the choice of the (auxiliary) weighting variables for post-stratification. The evidence points out that different characteristics matter in different contexts and situations. For instance, Morelli et al. (2017) report that positive emotions explain positioning in network reflecting time sharing, while empathy plays a role in intimate networks of the same people describing trust and support. Similarly, firms may form ties differently if searching for providers (or buyers) compared to innovation collaborations. Hence, one has to know the particular application under study to assess which node-level characteristic might provide valuable information about the network and we prefer to refrain from making general recommendations regarding the application of particular variables. For this reason, we would generally encourage applied researchers to first analyze the degree of non-representativeness and then use that information to inform the variables chosen for the correction.

Practically speaking, most data sets are limited to a relatively small set of variables that

encompass census information. Since our results show that the performance improves with more information and applying variables that provide no information about the network does *not* affect the performance negatively, we recommend employing all the available information in such cases. In contrast, when many variables are available for weighting, a problem would be to have too few observations in each stratified cell. This can lead to an increase in variance, resulting in reduced efficiency of the weighting estimates for the characteristic being studied. One straightforward solution is to apply the principal component analysis to filter the relevant independent information from a large number of potentially correlated variables and construct the weights using the discretized components. Another solution can be a simple two-step algorithm, outlined in Supplementary Appendix E, that we propose for the selection of the "right" variables.

We remain agnostic about the specific approach a researcher would take for a particular project. However, that researchers should be aware of the inferential problem addressed here and the general limits of treating the network as if it were complete or assuming representativeness of the network sample. Given that sensitivity, a variety of weights should be used to discover if the results are sensitive to accounting for non-representativeness. Such analysis should serve as a standard robustness check of empirical network results, giving scholars confidence that the results reflect network effects and are not a figment of the sampling strategy.

References

Alatas, Vivi, Abhijit Banerjee, Arun G Chandrasekhar, Rema Hanna, and Benjamin A Olken (2016) "Network structure and the aggregation of information: Theory and evidence from Indonesia," *American Economic Review*, 106 (7), 1663–1704.

Aldous, David J (1981) "Representations for partially exchangeable arrays of random variables," *Journal of Multivariate Analysis*, 11 (4), 581–598.

Aral, Sinan (2016) "Networked experiments," *The Oxford Handbook of The Economics of Networks*, 376–411.

Ballester, Coralio, Antoni Calvó-Armengol, and Yves Zenou (2006) "Who's who in networks. Wanted: The key player," *Econometrica*, 74 (5), 1403–1417.

Banerjee, Abhijit, Arun G Chandrasekhar, Esther Duflo, and Matthew O Jackson (2013) "The diffusion of microfinance," *Science*, 341 (6144), 1236498.

——— (2014) "Gossip: Identifying central individuals in a social network," *No. w20422 NBER Working paper*.

Berg, Sven (2004) "Snowball sampling—I," *Encyclopedia of Statistical Sciences*, 12.

Bhattacharyya, Sharmodeep and Peter J Bickel (2015) "Subsampling Bootstrap of Count Features of Networks," *Annals of Statistics*, 43 (6).

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

Bickel, Peter J, Aiyou Chen, and Elizaveta Levina (2011) "The method of moments and degree distributions for network models," *The Annals of Statistics*, 39 (5), 2280–2301.

Binder, David A and Georgia R Roberts (2003) "Design-based and model-based methods for estimating model parameters," *Analysis of survey data*, 29, 33–54.

Bloch, Francis, Garance Genicot, and Debraj Ray (2008) "Informal insurance in social networks," *Journal of Economic Theory*, 143 (1), 36–58.

Borgs, Christian, Jennifer Chayes, Henry Cohn, and Yufei Zhao (2019) "An L^p theory of sparse graph convergence I: Limits, sparse random graph models, and power law distributions," *Transactions of the American Mathematical Society*, 372 (5), 3019–3062.

Boucher, Vincent and Aristide Houndetoungan (2020) "Estimating peer effects using partial network data," *Working paper*.

Bramoullé, Yann, Habiba Djebbari, and Bernard Fortin (2009) "Identification of peer effects through social networks," *Journal of Econometrics*, 150 (1), 41–55.

Bramoullé, Yann, Rachel Kranton, and Martin D'amours (2014) "Strategic interaction and networks," *American Economic Review*, 104 (3), 898–930.

Branas-Garza, Pablo, Ramón Cobo-Reyes, María Paz Espinosa, Natalia Jiménez, Jaromír Kovářík, and Giovanni Ponti (2010) "Altruism and social integration," *Games and Economic Behavior*, 69 (2), 249–257.

Breza, Emily, Arun G Chandrasekhar, Tyler H McCormick, and Mengjie Pan (2020) "Using aggregated relational data to feasibly identify network structure without network data," *American Economic Review*, 110 (8), 2454–84.

Browne, Kath (2005) "Snowball sampling: using social networks to research non-heterosexual women," *International Journal of Social Research Methodology*, 8 (1), 47–60.

Calvo-Armengol, Antoni and Matthew O Jackson (2004) "The effects of social networks on employment and inequality," *American Economic Review*, 94 (3), 426–454.

Centola, Damon (2010) "The spread of behavior in an online social network experiment," *Science*, 329 (5996), 1194–1197.

Chandrasekhar, Arun (2016) "Econometrics of network formation," *The Oxford Handbook of the Economics of Networks*, 303–357.

Chandrasekhar, Arun G and Matthew O Jackson (2016) "A network formation model based on subgraphs," *Working paper*.

Chandrasekhar, Arun and Randall Lewis (2016) "Econometrics of sampled networks," *Available at SSRN: https://ssrn.com/abstract=2660381 or http://dx.doi.org/10.2139/ssrn.2660381*.

Cingano, Federico and Alfonso Rosolia (2012) "People I know: job search and social networks," *Journal of Labor Economics*, 30 (2), 291–332.

Comellas, F and S Gago (2007) "Spectral bounds for the betweenness of a graph," *Linear Algebra and Its Applications*, 423 (1), 74–80.

Crane, Harry and Henry Towsner (2018) "Relatively exchangeable structures," *The Journal of Symbolic Logic*, 83 (2), 416–442.

Currarini, Sergio, Matthew O Jackson, and Paolo Pin (2009) "An economic model of friendship: Homophily, minorities, and segregation," *Econometrica*, 77 (4), 1003–1045.

De Paula, Aureo (2017) "Econometrics of Network Models," *In B. Honoré, A. Pakes, M. Piazzesi, & L. Samuelson (Eds.), Advances in Economics and Econometrics: Eleventh World Congress (Econometric Society Monographs, pp. 268-323). Cambridge: Cambridge University Press.*

De Paula, Áureo (2020) "Econometric models of network formation," *Annual Review of Economics*, 12, 775–799.

De Paula, Áureo, Imran Rasul, and Pedro Souza (2018) "Recovering social networks from panel data: Identification, simulations and an application," *Working paper*.

Eagle, Nathan, Michael Macy, and Rob Claxton (2010) "Network diversity and economic development," *Science*, 328 (5981), 1029–1031.

Fleming, Lee, Charles King III, and Adam I Juda (2007) "Small worlds and regional innovation," *Organization Science*, 18 (6), 938–954.

Fortin, Bernard and Vincent Boucher (2015) "Some Challenges in the Empirics of the Effects of Networks," in *The Oxford Handbook of the Economics of Networks*.

Frank, Ove (1977) "Survey sampling in graphs," *Journal of Statistical Planning and Inference*, 1 (3), 235–264.

——— (1981) "A survey of statistical methods for graph analysis," *Sociological Methodology*, 12, 110–155.

Golub, Benjamin and Matthew O Jackson (2012) "How homophily affects the speed of learning and best-response dynamics," *Quarterly Journal of Economics*, 127 (3), 1287–1338.

Graham, Bryan S (2020) "Network data," in *Handbook of Econometrics*, 7, 111–218: Elsevier.

Granovetter, Mark (1985) "Economic action and social structure: The problem of embeddedness," *American Journal of Sociology*, 91 (3), 481–510.

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

Handcock, Mark S and Krista J Gile (2010) "Modeling social networks from sampled data," *Annals of Applied Statistics*, 4 (1), 5.

Hardy, Morgan, Rachel M Heath, Wesley Lee, and Tyler H McCormick (2019) "Estimating spillovers using imprecisely measured networks," *arXiv preprint arXiv:1904.00136*.

Heckathorn, Douglas D (1997) "Respondent-driven sampling: a new approach to the study of hidden populations," *Social Problems*, 44 (2), 174–199.

Hoff, Peter D, Adrian E Raftery, and Mark S Handcock (2002) "Latent space approaches to social network analysis," *Journal of the American Statistical Association*, 97 (460), 1090–1098.

Holland, Paul W, Kathryn Blackmond Laskey, and Samuel Leinhardt (1983) "Stochastic blockmodels: First steps," *Social Networks*, 5 (2), 109–137.

Hoover, Douglas N (1979) "Relations on probability spaces and arrays of random variables," *Preprint, Institute for Advanced Study, Princeton, NJ*, 2, 275.

Horvitz, Daniel G and Donovan J Thompson (1952) "A generalization of sampling without replacement from a finite universe," *Journal of the American Statistical Association*, 47 (260), 663–685.

Jackson, Matthew O (2005) "A survey of network formation models: stability and efficiency," *Group Formation in Economics: Networks, Clubs, and Coalitions*, 11–49.

——— (2010) *Social and Economic Networks*: Princeton university press.

Jackson, Matthew O, Tomas Rodriguez-Barraquer, and Xu Tan (2012) "Social capital and social quilts: Network patterns of favor exchange," *American Economic Review*, 102 (5), 1857–1897.

Jackson, Matthew O and Brian W Rogers (2007) "Meeting strangers and friends of friends: How random are social networks?" *American Economic Review*, 97 (3), 890–915.

Jackson, Matthew O, Brian W Rogers, and Yves Zenou (2017) "The economic consequences of social-network structure," *Journal of Economic Literature*, 55 (1), 49–95.

Karlan, Dean, Markus Mobius, Tanya Rosenblat, and Adam Szeidl (2009) "Trust and social collateral," *Quarterly Journal of Economics*, 124 (3), 1307–1361.

Kolaczyk, Eric D (2009) *Statistical Analysis of Network Data: Methods and Models*: Springer Science & Business Media.

Li, Ting, Xianshi Yu, and Bing-Yi Jing (2019) "Measuring the clustering strength of a network via the normalized clustering coefficient," *arXiv preprint arXiv:1908.00523*.

Li, Xinran and Peng Ding (2017) "General forms of finite population central limit theorems with applications to causal inference," *Journal of the American Statistical Association*, 112 (520), 1759–1769.

Little, Roderick JA (1993) "Post-stratification: a modeler's perspective," *Journal of the American Statistical Association*, 88 (423), 1001–1012.

Lovász, László (2012) *Large Networks and Graph Limits*, 60: American Mathematical Society.

McPherson, J Miller and Lynn Smith-Lovin (1987) "Homophily in voluntary organizations: Status distance and the composition of face-to-face groups," *American Sociological Review*, 370–379.

McPherson, Miller, Lynn Smith-Lovin, and James M Cook (2001) "Birds of a feather: Homophily in social networks," *Annual Review of Sociology*, 27 (1), 415–444.

Morelli, Sylvia A, Desmond C Ong, Rucha Makati, Matthew O Jackson, and Jamil Zaki (2017) "Empathy and well-being correlate with centrality in different social networks," *Proceedings of the National Academy of Sciences*, 114 (37), 9843–9847.

Newman, Mark EJ (2002) "Assortative mixing in networks," *Physical Review Letters*, 89 (20), 208701.

Opsahl, Tore (2013) "Triadic closure in two-mode networks: Redefining the global and local clustering coefficients," *Social Networks*, 35 (2), 159–167.

Pastor-Satorras, Romualdo and Alessandro Vespignani (2002) "Immunization of complex networks," *Physical Review E*, 65 (3), 036104.

Prášková, Zuzana and Pranab Kumar Sen (2009) "Asymptotics in finite population sampling," *Handbook of Statistics*, 29, 489–522.

Ruiz-Palazuelos, Sofía, María Paz Espinosa, and Jaromír Kovářík (2023) "The weakness of common job contacts," *European Economic Review*, 160, 104594.

Schilling, Melissa A and Corey C Phelps (2007) "Interfirm collaboration networks: The impact of large-scale network structure on firm innovation," *Management Science*, 53 (7), 1113–1126.

Smith, Terence MF (1991) "Post-stratification," *Journal of the Royal Statistical Society Series D: The Statistician*, 40 (3), 315–323.

Sterba, Sonya K (2009) "Alternative model-based and design-based frameworks for inference from samples to populations: From polarization to integration," *Multivariate Behavioral Research*, 44 (6), 711–740.

Thirkettle, Matthew (2019) "Identification and Estimation of Network Statistics with Missing Link Data," *Working paper*.

Van Mieghem, Piet (2010) *Graph Spectra for Complex Networks*: Cambridge University Press.

Vega-Redondo, Fernando (2007) *Complex Social Networks* (44): Cambridge University Press.

Watts, Duncan J and Steven H Strogatz (1998) "Collective dynamics of "small-world" networks," *Nature*, 393 (6684), 440–442.