Non-randomly sampled networks: Biases and Corrections

Chih-Sheng Hsieh†
Chinese University of Hong Kong

Stanley I. M. Ko‡
University of Macau

Jaromír Kovářík§
University of the Basque Country & CERGE-EI

Trevon Logan¶
Ohio State University & NBER

February 13, 2018

Abstract

Empirical network studies typically analyze partial samples of the population of interest. However, sampled network data bias systematically the properties of observed networks and suffer from non-classical measurement-error problem if applied as regressors. This paper analyzes statistical issues arising from examining networks from non-representative samples of the population. We first characterize the biases in both network statistics and the estimates of network effects under non-random sampling theoretically and numerically. Apart from the sampling rate and the elicitation procedure, these biases depend in a non-trivial way on which subpopulations are missing with higher probability. We then propose a methodology adapting post-stratification weighting approaches to networked contexts, which enables researchers to recover several network-level statistics and reduce the biases in estimates of network effects. The proposed methodology nests the corrections based on the missing-at-random assumption and–as opposed to existing approach–does not rely on any particular sampling procedure. Last, we apply our approach to two commonly used data sets and show that abstracting from non-representativeness of the sample changes dramatically the results of regression analysis.

1 Motivation

There is a growing interest in understanding the role of networks in Economics. Different “micro” and “macro” features of network architecture shape diffusion, learning, behavior, and other phenomena in a great variety of contexts of socio-economic interest. The empirical network research is probably the most dynamic part of this literature at this moment mainly due to the combination of increasing availability of large network data sets and increasing computational power at our disposal. Nevertheless, empirical network analysis inherently generates new econometric challenges (Manski (1993); De Paula (2016); Fortin and Boucher (2015); see Jackson et al. (2017) for a recent summary). This paper particularly analyzes issues arising with non-randomly sampled network data.

Large majority of empirical network studies only analyze partial samples and the sampling rates are typically low. Even though the literature across several disciplines has for long noticed that using sampled data may lead to considerable biases and other statistical issues (see Stork and Richards (1992); Kossinets (2006); or Handcock and Gile (2010) for early references), the typical approach in the literature is to treat the data “as if” complete. An important reason behind the common use of partial

---

*Jaromír Kovářík acknowledges financial support from from the Basque Government (IT-783-13), Ministerio de Economía y Competitividad and Fondo Europeo de Desarrollo Regional (ECO2015-64467-R MINECO/FEDER and ECO 2015-66027-P), and the Grant Agency of the Czech Republic (14-22044S).

†Department of Economics, The Chinese University of Hong Kong, Room 911 Esther Lee Building, CUHK, Shatin, Hong Kong (cshsieh@cuhk.edu.hk)

‡Department of Finance and Business Economics, Faculty of Business Administration, University of Macau (StanleyKo@umac.mo).

§Dpto. Fundamentos del Análisis Económico I & Bridge, University of the Basque Country, Bilbao, Spain and CERGE-EI, a joint workplace of the Charles University & the Economics Institute of the Academy of Sciences of the Czech Republic, Prague, Czech Republic (jaromir.kovarik@ehu.eus).

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

1See Vega-Redondo (2007), Jackson (2010b), and Goyal (2012) for reviews.

2Jackson (2010a) and Jackson et al. (2017) provide a survey of economic applications.

3Chandrasekhar and Lewis (2016) report that the median sampling rate in applied work is 25% and more than 66% of network studies have a sampling rate lower than 51%.

4As exceptions to this rule, several recent studies provide either robustness tests or a discussion regarding missing network data in their applications (Conley and Udry (2010); Conti et al. (2013); De Giorgi et al. (2010)). Alatas et al. (2016) apply the corrections of Chandrasekhar and Lewis (2016) assuming random missing.
samples is that network elicitation is typically more costly than collection of e.g. basic individual characteristics (see Breza et al. (2017) for a discussion). In a recent contribution, Chandrasekhar and Lewis (2016) show formally that, even in absence of other econometric issues and even if the nodes are selected randomly, sampled networks differ systematically from the population and that inference on sampled network data leads to non-classical measurement error and inconsistency problems while identifying network effects. The estimates from using sampled networks may suffer from attenuation, but also expansion or even sign switching. Chandrasekhar and Lewis (2016) propose an integral methodology of how to deal with randomly sampled networks in applied research. Their methodology consists of two alternative strategies. As the first solution, they propose analytical corrections for some standard network characteristics. This approach only works at the network level and is only tractable for a set of structural network properties though. The second approach is a two-step estimation procedure based on graphical reconstruction. Despite generating consistent estimates, this second approach does not necessarily enable to recover the network properties. Most importantly for the present work, both approaches are based on the missing-at-random assumption.

Nevertheless, samples in network studies are often non-representative. Apart from problems inherent in sampling, such as survey non-response, the existence of hard-to-reach populations, etc., it is common knowledge that non-representativeness of network data may be caused by the sampling strategy itself (Kolaczky (2009)). Snowball sampling, one prominent sampling method in applied work, is prone to finding more likely nodes with higher connectivity than with small number of network neighbors. The reason is that people are found through network links. Having more connections thus increases the probability of being sampled. Hence, more connected nodes can be overrepresented, whereas low-degree nodes underrepresented under such sampling scheme. Other issues arise with “truncated” methods due to either the specification of network boundary or fixed-choice design (e.g. nominate the best five friends). Last, several data sets exploit stratified network samples to better approximate the being sampled. The reason is that people are found through network links. Having more connections thus increases the probability of being sampled. Hence, more connected nodes can be overrepresented, whereas low-degree nodes underrepresented under such sampling scheme. Other issues arise with “truncated” methods due to either the specification of network boundary or fixed-choice design (e.g. nominate the best five friends). Last, several data sets exploit stratified network samples to better approximate the

In this paper, we argue that the existing approaches based on the missing-at-random assumption do not eliminate the statistical issues arising from sampling in the existing data set. We particularly show below that two carefully elicited data sets–National Longitudinal Survey of Adolescent Health (Add Health, hereafter) collected with the truncated fixed-choice design and a recent stratified data on microfinance take-up in a number Indian villages (Banerjee et al. (2013))–are indeed non-random samples of the population under scrutiny. To explain the issues arising from non-random sampling, let us informally decompose the problem into two effects, which we term respectively scaling and non-representativeness. Scaling refers to observing less people and relationship than there exist in the population, independently of the (non-)representativeness of the sample. In contrast, non-representativeness corresponds to non-randomness of the sample. If nodes are missing at random, only scaling matters. As an example of the effect of random missing, consider the average degree of a network. If the links between the sampled and non-sampled individuals are not observed, then the sample average degree is biased downwards by construction. In addition, imagine that the population average degree is correlated with the diffusion properties of the network. Applying the observed average degree in a regression therefore inflates the estimated impact of the average degree on diffusion under random missing. This illustrates the severity of the issue, since it represents an intuitive example of expansion and thus non-classical measurement error. However, if nodes are not missing at random, whether the observed average degree and the estimates will be inflated or attenuated will depend on who is missing. If e.g. less connected nodes are missing with higher probability (a problem inherent e.g. in snowball sampling), 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 homophily index and the clustering coefficient can be unbiased in representative samples. Nevertheless, in samples in which different types of nodes are missing with differing probabilities, homophily will be biased by definition. Since clustering is typically associated with connectivity in social networks (Jackson and Rogers (2007)), clustering is also likely to be mismeasured. The direction of these biases and those in the estimates of effects of homophily and clustering in regressions depend crucially and non-trivially on who is missing.

This paper provides a systematic analysis of problems arising from examining non-randomly sampled network data and proposes a solution that (i) allows researchers to recover the true population network and (ii) improve inferences in regressions testing the effects of large-scale network structure on both individual and group-level behaviors and outcomes. We first
characterize analytically and numerically the extent of biases both in the structural properties of the observed networks and in the estimates from regression analysis resulting from raw sampled networks as well as from corrections based on the missing-at-random assumptions. Second, we propose a set of analytical corrections for several network characteristics widely used in applications, namely average degree, clustering coefficient, graph span, epidemic threshold, and homophily. These network features represent fundamental aspects of a network architecture commonly employed in theoretical and empirical research and provide intuitive insights regarding the way social organization shapes individual and group-level outcomes. To that aim, we adapt standard (i.e. network-free) post-stratification weighting approaches to networked contexts. There is a general agreement that when population information is available, post-stratification weighting can correct sampling biases due to varying (non-)response rates among different demographic or socioeconomic categories and thus improve the precision of sample estimates for objective variables of interest.\footnote{See Holt and Smith (1979), Little (1993) and Valliant (1993) for statistical properties of poststratification weighting.}

In line with this literature, we assume that the population can be divided into $T \in \mathbb{N}$ disjoint types or groups and that sampling (or conversely missing) rates differ across types.\footnote{These types or groups are thought to represent, say, men and women, races, ethnicities, locations, different age categories, socio-economic classes, etc. or their combinations such as “white women of an age between 20 and 30 with an yearly income below $50,000” or “men of other race of an age over 70 with an income over $100,000.” The variable resulting from combining different types is termed \textit{rake} in the post-stratification literature and throughout our paper.} The main difference between the standard post-stratification and our approach is to weight on network links, rather than on individuals.\footnote{In the standard case, for each type $t$, there are $n_t$ individuals in the population of a total of $n$ individuals and $m_t$ observations in the sample of size $m$. The poststratification weight assigned to a sampled individual $i$ depends on which categories she belongs to. Formally, $p_i^t = \frac{n_t/m}{m_t/m}$. Whenever the sampled ratio in category $t$ is smaller (larger) than that in the population, the weight is larger (smaller) than one. In other words, it raises (or decreases) the weights for types of individuals who are underrepresented (or overrepresented) compared to the population. Below, we introduce a similar approach but applied to links.} Since the proposed corrections are asymptotically unbiased and by properties of conditional expectations, regressing economic outcomes on the corrected network measures delivers consistent estimates. Third, we test the ability of our approach to mitigate these biases \textit{vis-à-vis} both the population and the random corrections. Last, the proposed methodology is applied to the two data sets particularly suited for our approach, the Add Health data and the village networks from Banerjee et al. (2013). We focus on these data sets because they contain a relatively large number of networks, they are widely employed in empirical work,\footnote{See e.g. Moody (2001); Echenique and Fryer Jr (2007); Bramoulle et al. (2009); Currarini et al. (2009, 2010); Calvó-Armengol et al. (2009) among many others for the friendship networks and Chandrasekhar and Lewis (2016); Jackson et al. (2012); Banerjee et al. (2013, 2014), for the latter.} and a node and a link have very different meaning in each of them, illustrating the broad applicability of our approach.

Our theoretical and numerical analysis characterizes the biases in structural network properties under partial sampling and in corrections assuming representativeness. Their existence and magnitudes depend on typical factors, such as the missing rates, the sampling method, the network property of interest etc. Most importantly though, the directions and magnitudes of the biases depend non-trivially on who is (not) sampled with higher probability. Both the raw data and the random corrections can bias the true population statistics downwards or upwards, and the estimates using these corrections as regressors can be attenuated but also expanded and switch signs in function of who is missing. Moreover, the magnitudes of the biases are often economically significant in our numerical analysis. Therefore, abstracting from non-representativeness of the sample leads to mis-measured networks and biased network effects and one cannot rely on the classic measurement-error problem. The results show that the diagnosis of the issues with a particular data set have to be performed on the case-by-case basis, demonstrating the necessity to account for potential different missing rates of different segments of the population in applied work.

As for the recovery of the true networks of interest, we take explicit account of the missing rates of different subpopulations. Our methodology nests the existing corrections designed for random sampling. Indeed, both approaches perform similarly recovering the true network statistics both theoretically and in our numerical experiments if the sample is representative. However, our approach outperforms Chandrasekhar and Lewis (2016) if the data are not missing at random. We even document situations, in which the raw data exhibit smaller biases than corrections based on representativeness.

The two presented empirical applications corroborate the above observations in that one cannot easily predict the direction and magnitude of the biases. Not accounting for non-randomness of the sample inflates the estimates in one of our applications whereas it attenuates them in the other one. Using the Indian village networks stratified on religion and geographical location, we show that people in the working age are largely underrepresented in the sample. We then test to what extent the architecture of the network in each village predicts the village-level labour market outcomes. The social structure within the village seems to be very important under the raw data and the corrections assuming representativeness, while accounting for higher missing rates of economically more active individuals reveals that the network structure plays much smaller role for labour outcomes. In case of the Add Health data, the contrary occurs. Raw data and random corrections suggest little role of the network in participation in clubs in each village, while accounting for the missing segments of the population reveals that the patterns of interactions are key determinants of participation in within-school activities.

Before we relate our work to the literature, we discuss several aspects of the proposed methodology. First, the approach allows the researcher to have sampled data on networks, but it requires her to observe some information about the whole population. adopt their terminology. Our solution—hence, the term “large-scale”—deals with the former.

\textsuperscript{10}See Holt and Smith (1979), Little (1993) and Valliant (1993) for statistical properties of poststratification weighting.

\textsuperscript{11}These types or groups are thought to represent, say, men and women, races, ethnicities, locations, different age categories, socio-economic classes, etc. or their combinations such as “white women of an age between 20 and 30 with an yearly income below $50,000” or “men of other race of an age over 70 with an income over $100,000.” The variable resulting from combining different types is termed \textit{rake} in the post-stratification literature and throughout our paper.

\textsuperscript{12}In the standard case, for each type $t$, there are $n_t$ individuals in the population of a total of $n$ individuals and $m_t$ observations in the sample of size $m$. The poststratification weight assigned to a sampled individual $i$ depends on which categories she belongs to. Formally, $p_i^t = \frac{n_t/m}{m_t/m}$. Whenever the sampled ratio in category $t$ is smaller (larger) than that in the population, the weight is larger (smaller) than one. In other words, it raises (or decreases) the weights for types of individuals who are underrepresented (or overrepresented) compared to the population. Below, we introduce a similar approach but applied to links.
In particular, the researcher has to know the population sizes of each type and has to observe some individual heterogeneity (e.g. demographic or socio-economic characteristics) to be able to classify the sampled subjects into types and quantify who is missing. Such data structure are common in many applications and typically already considered in the sampling process. Second, the methodology indirectly assumes that individual heterogeneity provides information about the network. There is an extensive evidence that network positioning is associated to a large variety of characteristics and that who is connected with whom depend on types\textsuperscript{14,15}. We share these two aspects with other network-recovery methodologies (see e.g. Hoff et al. (2002); Breza et al. (2017)).\textsuperscript{16} Last, even if the weighting approach generalizes the correction-based strategy of Chandrasekhar and Lewis (2016) to non-representative samples, we share with them the main advantages and limitations: the corrections are computationally simple and easy to implement, but limited to particular aggregate network statistics.

The present paper is related to several streams of literature. First, we contribute to the multidisciplinary literature on missing network data. An early study of Stork and Richards (1992) analyze the issues of non-responses in network data and proposes a simple method called Reconstruction, which assumes that individual i’s description of the relationship between i and j is the same as that of individual j. Kossinets (2006) uses simulation methods to examine impacts of missing links due to different sampling methods\textsuperscript{17} and finds that biases of missing links on estimated network statistics due to the network boundary specification and the fixed choice design are dramatic; he proposes no solution though. Handcock and Gile (2010) discusses the problem of partially observed network data, i.e., missing individuals or missing links between individuals due to network sampling design. Different network sampling designs, such as Ego-centric design, One-wave link-tracking design, Multi-wave link-tracking design imply different probabilities of sampling elements of the sociomatrix. They discuss two inferential frameworks: the design-based framework and the model-based framework. In the design-based framework, in order to identify design-unbiased estimators for quantities of interest measured on the complete network, one needs the dyadic sampling probability, which is however, not observable for all sampling designs. The model-based framework (likelihood-based inference), which considers a parametric model for the random behavior of the network depending on a parameter vector, $\delta$, can provide more a flexible way to conduct network inference. Handcock and Gile (2010) point out using the MCMC to sample missing links by the conditional distribution of $W_{mis}$ given $W_{obs}$, which produce multiple imputations of the full data. In their example, individuals who did not finish the in-School survey of the Add Health data contribute to unobserved network links. In Huisman (2009), survey non-responses can be distinguished to unit (respondent) non-response and item non-response. In the longitudinal data, there is a further wave non-response. Huisman use the imputation approach instead of likelihood-based estimation. He finds that the reconstruction approach by Stork and Richards (1992) on dealing with non-response is not totally useless when the non-response rate is low. Robins et al. (2004) also questions the usefulness of the imputation method such as reconstruction.

Our methodology is closely related with respondent driven-sampling (Heckathorn (1997)), a methodology that combines snowball sampling with a model that weights the sample to compensate for the non-representativeness of the sample. Our weighting approach and RDS have the same goal but differ substantially in the underlying assumptions and the applicability. In contrast to RDS, our approach can be applied in cases in which networks are not elicited via designed sampling, such as Add Health data set or the stratified sample in Banerjee et al. (2013).

The closest to our paper is Chandrasekhar and Lewis (2016) who provide a general treatment of sample network problems under random missing of nodes. Similarly to this paper, they show that estimations with sampled networks suffer from non-classical measurement error induced by sampling and proposed an integral methodology to ensure consistent estimates. The methodology consists of two independent strategies. First, they provide formal corrections for average degree, clustering coefficient, and graph span. Our approach generalizes this first strategy.

2 Framework

2.1 Notation

Since we view our work as complementary to the previous work, we mostly adapt the existing notation in Kolaczyk (2009) and Chandrasekhar and Lewis (2016) (CL, hereafter). A population network (graph) is a pair $G = (V,E)$ of a set of nodes V and edges E. Denote $n = |V|$. The network is represented with an adjacency $n \times n$ matrix $W(G)$. We follow the theoretical and empirical literature and focus on undirected and unweighted networks. Therefore, $W_{ij} = 1(0)$ if i and j are (not) connected and if

\textsuperscript{14}Providing a survey of this literature would be well beyond the scope of this paper. Note though that our simulations show that our approach is still useful even if one applies variables uncorrelated to networks. However, its performance improves with variables correlated to networks.

\textsuperscript{15}The explicit account of types and the relationships among them inherent in our approach is particularly useful in presence of network homophily or heterophily, typical features of real-life social and economic networks (Jackson (2010b)). Since there are no types in Chandrasekhar and Lewis (2016), all links are treated equally in their approach.

\textsuperscript{16}In a similar vein, the estimation of peer effects jointly with the endogeneous network structure commonly assumes that individual characteristics determine who links up with whom (see e.g. Hsieh and Lee (2016)). Similarly, Chandrasekhar and Lewis (2016) propose to to apply individual heterogeneity to estimate the network formation model behind the data to recover the true network of interest.

\textsuperscript{17}Kossinets (2006) focuses on the boundary specification problems, which is closely related to our induced network sampling, survey non-response and fixed choice design, which are targeted in our empirical applications.
$W_{ij} = W_{ji}$ for each $i, j \in V$. Following the tradition, we set $W_{ii} = 0$. Most of the analysis extends for undirected and unweighted graphs though. We assume that the population can be classified into $T$ disjoint types with a generic type $t \in \{1, 2, ..., T\}$. Let $V_t$ be the set of nodes of type $t$, $n_t = |V_t|$ is the size of subpopulation $t$ and $\sum_t n_t = n$. We write $t_i = t$ if individual $i$ is of type $t$. Then, $t_i = t_j$ ($t_i \neq t_j$) indicates that $i$ and $j$ are (not) of the same type.

Rather than the whole graph, the researchers only observe a sample of the population network. Let $S$ be the set of sampled nodes of size $m = |S| = \psi|n_t|$, being $\psi = \frac{m}{n}$ the sampling rate. Analogously, $m_t = \psi_t n_t$ is the sampled number of individuals of type $t$ and $\psi_t$ is type $t$’s sampling rate.

There are two types of sampled networks. The first one is the induced subgraph denoted $G^s$. The induced graph restricts the network links among the $m$ sampled nodes. The second sample scheme considered is the star subgraph, denoted by $G^s$. The star network samples $S$ but allows for a link of the $m$ sampled nodes to anybody in $V$. That is, $G^s = (S, E^s)$ and $G^* = (V, E^*)$ where $E^s$ is set of edges between the sampled nodes and $E^S$ is set of all edges such that at least one of nodes is in $S$.

This work will concentrate on several network statistics. We denote a generic network statistic by $w(G)$. It can represent a scalar, vector, or even the whole adjacency matrix (i.e. $w(G) = W(G)$). The dimension of $w(G)$ will depend on the application and will be clarified in each context. Let $w(G) \in \{G^s, G^l, G\}$ be the corresponding network statistics using the sample network $\bar{G}$ and $\bar{w}(G)$ the corrected network statistic in question proposed to mitigate the sample biases with respect to the population. For example, $w(G) = \frac{1}{m} \sum_{i \in V} \sum_{j \in V} W_{ij}$ is the average degree of a graph, which we denote $d(G)$ below. Hence, $\bar{d}(G)$ is the average degree of the sample network and $\bar{d}(G)$ the proposed correction of the sample average degree to mitigate biases with respect to the true $d(G)$.

We assume that the sampling is non-random in the following sense. The analyst observes $m_t \leq n_t$ individuals of type $t$, with $m_t = \psi_t n_t$ and $\sum_t m_t = m$. If $\psi_t = \psi$ for all $t \in T$, the sampling is random as in CL. Our framework allows for $\psi_t \neq \psi$, for any $t, s \in T$, but nests the random sampling case of CL.

Last, let $\Omega_m$ be the set of all samples of size $m$ from a population of size $n$. Denote $C_m^n = \binom{n}{m}$ the number of all possible samples of size $m$ from a set of $n$ nodes. Last, let the indicator function $I(x) = 1$ if condition $x$ is satisfied; $I(x) = 0$ otherwise. Condition $x$ can naturally be a composition of several subconditions.

In applications, the researcher may observe $R$ different networks with a generic term $r \in \{1, 2, ..., R\}$. If a measure refers to a specific network, we use a subscript $r$ to specify it. That is, $G_r$ is the graph of population $r$, $\bar{G}_r \in \{G_r^s, G_r^l, G_r\}$ the corresponding sampled graphs of network $r$, and accordingly for the other variables. Therefore, $n_{r,t}$ and $m_{r,t}$ are the number of nodes of type $t$ in the population network $r$ and their corresponding sampled number. Once again, $\psi_{r,t} = \frac{m_{r,t}}{n_{r,t}}$ is the sampling rate of type $t$ in network $r$.

### 2.2 Econometric Modeling

Apart from the reconstruction of network properties of interest, we also analyze regression analysis with non-randomly sampled networks. Our approach is suitable for models where–apart from other variables–one or more network-wide characteristics are regressors. Throughout the analysis, we focus on regressions, in which the researchers is interested in understanding whether and how the global properties of a network (e.g. average degree or graph span) influence a certain socio-economic outcome of the whole network population. Formally,

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

(1)

where $y_r$ is the outcome variable of population or network $r$, $x_r$ is the set of network-level controls, and $w(G_r)$ is the true network property (or properties) of interest. The researcher is interested in estimating the parameters $\alpha$, $\beta$, and $\gamma$. We mostly focus on the estimates of $\beta$ though. Examples of applications of (1) in the literature include Alatas et al. (2016) who regress the ability of villages 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, Currarini et al. (2009, 2010) and Golub and Jackson (2012a,b) who relate homophily with school-level statistics using Add-Health data, and Toomet et al. (2013) who link regional wage differences between ethnicities with region-level homophily, or the innovation literature that model the ability of different regions to generate knowledge depending on the structure of research networks (e.g. Fleming et al., 2007). Such regressions are also of interest for the theoretical literature that proposes other hypotheses. For example, the overall clustering of a network may explain the magnitude and efficiency of risk-sharing within a society (Bloch et al., 2008), the stability of behavior in a society may be related to the minimal eigenvalue of the adjacency matrix (Bramoulle et al., 2014) among others. Nevertheless, the proposed approach also applies to models studying whether and how the overall properties of a network affect the behavior at the individual level as follows: $y_{ir} = \alpha + w(G_r) \beta + \gamma x_{ir} + \lambda_r + \epsilon_{ir}$, with $y_{ir}$ is the behavior of an individual $i$ in network $r$, $x_{ir}$ is her heterogeneity (that can include the heterogeneity of $i$’s neighborhood), and $\lambda_r$ are either fixed or random effects. For instance, the decision of an individual to adopt a product (e.g. microfinance as in Banerjee et al., 2013) or participate in an activity (e.g. recreational activity as in Bramoulle et al., 2009) can depend on the overall structure of the network. In the same vein, the innovation literature studies how the structure of regional networks shapes innovative performance of individual innovators (Schilling and
Phelps, 2007; Whittington et al., 2009). 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 and Kranton, 2007; or Bramoullé et al., 2014).

With sampled data on the network, the researchers observe $G \in \{G^3, G^{15}\}$ which is mismeasured, rather than $G$. Therefore, the scholars typically estimate

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

leading to a measurement error in the regressors. The classic measurement error and the resulting attenuation bias are based on several assumptions not generally satisfied in the discussed cases (see e.g. Wooldridge, 2015, for a textbook treatment or Hyslop and Imbens, 2001). CL show formally and via simulations that the biases are generally not tractable and can lead to expansion or sign switching even in the simplest cases and under purely random sampling. The issues becomes even more problematic if the missing-at-random assumption is violated. Moreover, if multiple regressors are included into the model as in (2), the estimates of independent variables measured without error also become biased when another one is mismeasured. That is, even if network properties only serve as controls, they may bias the estimates of the main variables of interest. This is just another application of the proposed methodology.

In the following section, we show how the biases depend on who is missing in the sample and Section 5 illustrates that non-representativeness issues are present in some widely used network data.19

3 Analytic corrections for sample network measures

This section shows formally the biases in some commonly used network characteristics arising from sampling and proposes how to correct them using post-stratification approaches. Since each network $r \in \{1, \ldots, R\}$ is corrected separately, the main text illustrates our approach for a generic network $r$.

3.1 Average Degree

The degree of a node is the number of her network connections. The average degree of population graph $G_r$ is simply the average number of network links per person in the network, defined as $d(G_r) = \frac{1}{n_r} \sum_{i \in V_r} \sum_{j \in V_r} W_{ij,r}$. The degree is a basic measure of local node’s importance or centrality. It has been applied as a regressor in numerous studies and contexts (e.g. Kremer and Miguel, 2007; Branas-Garza et al., 2010; Banerjee et al., 2013; Alatas et al., 2016, among many others).

As an introductory example, imagine that there are only two types of nodes, men and women denoted $t = 1, 2$, respectively, with differing sampling rates and different linking probabilities within or across each gender.20 The generalization for any $T \geq 2$ can be found in Appendix B. We use the two-type case here, because it allows us to illustrate the intuition behind our approach and enables an easy comparison with the approach of CL.

For the induced subgraph with $T = 2$, we can show the following:21

$$E(d(G^{15}_r)) = E \left( \frac{1}{n_r} \sum_{i,j \in S_r} W_{ij,r} \left| G_r \right. \right) = \frac{1}{n_r} \left( \frac{\psi_{r,1} \psi_{r,2}}{\psi_r} \sum_{i,j \in V_r, t_i \neq t_j} W_{ij,r} \right) + \frac{1}{n_r} \left( \frac{\psi_{r,1}}{\psi_r} (\psi_{r,1} + o(1)) \sum_{i,j \in V_r, t_i = t_j = 1} W_{ij,r} \right) + \frac{1}{n_r} \left( \frac{\psi_{r,2}}{\psi_r} (\psi_{r,2} + o(1)) \sum_{i,j \in V_r, t_i = t_j = 2} W_{ij,r} \right).$$

The intuition behind this conditional expectation is the following: There are $\sum_{i,j \in V_r} W_{ij,r}$ edges in the true network of interest, out of which $\sum_{i,j \in V_r, t_i \neq t_j} W_{ij,r}$ edges connect two individuals of different types, $\sum_{i,j \in V_r, t_i = t_j = 1} W_{ij,r}$ edges link two individuals of type 1, and $\sum_{i,j \in V_r, t_i = t_j = 2} W_{ij,r}$ edges link two type 2 nodes. Given the sampling rate of each type, we only observe for example $\frac{\psi_{r,1} \psi_{r,2}}{\psi_r} \sum_{i,j \in V_r, t_i \neq t_j} W_{ij,r}$ of the cross-type edges in expectation. As long as $\psi_r \neq 1$, the conditional expectation of

---

18 We are well aware that there exist other challenges typically present in network regressions, such that endogeneity and/or omitted variable problems. Our arguments are that sampling presents an issue even in the absence of these problems.

19 The matters become even more complex if multiple regressors are included into the model, since even variables measured without error can be biased if another one is mismeasured (CL p.12). That is, if network properties only serve as controls, they bias the estimates of the main variables of interest. This is just another application of the proposed methodology.

20 Gender homophily is one of the main features of social networks (McPherson et al., 2001).

21 The derivation of all corrections is relegated to Appendix B.
Several papers have used clustering as a regressor (e.g., Fleming et al., 2007; Alatas et al., 2016). sharing (Bloch et al., 2008), trust building (Karlan et al., 2009), cooperation enhancing (Granovetter, 1985), to name a few.

Hence, every triangle is also a triple but the converse is not true.

A triangle refers to a complete subnetwork of three individuals, while a triple is a three-node subnetwork, in which at least two edges are present. Hence, every triangle is also a triple but the converse is not true.
For the induced network, we get that

$$E(\rho(G_r^s)|G_r) = E \left( 3 \sum_{i \in S_r} \sum_{k > i} \sum_{j \neq i, k} W_{i,j,r}^s W_{j,k,r}^s W_{k,i,r}^s \mid G_r \right)$$

$$= \sum_{t=1}^{T} \left( 3 \sum_{i \in V_r} \sum_{k > i} \sum_{j \neq i, k} W_{i,j,r} W_{j,k,r} W_{k,i,r} \left( \psi_{r,t}^3 + o(1) \right) \right) + \sum_{t=1}^{T} \sum_{\ell \neq t} \left( 3 \sum_{i \in V_r} \sum_{k > i} \sum_{j \neq i, k} W_{i,j,r} W_{j,k,r} W_{k,i,r} \left( \psi_{r,\ell}^2 \psi_{r,t} + o(1) \right) \right)$$

$$+ \sum_{t=1}^{T} \sum_{\ell \neq t} \sum_{h \neq t, \ell} \left( 3 \sum_{i \in V_r} \sum_{k > i} \sum_{j \neq i, k} W_{i,j,r} W_{j,k,r} W_{k,i,r} \left( \psi_{r,\ell} \psi_{r,t} \psi_{r,h} \right) \right).$$

Analogously,

$$E(\tau(G_r^s)|G_r) = E \left( \sum_{i \in S_r} \sum_{k > i} \sum_{j \neq i, k} W_{i,j,r}^s W_{j,k,r}^s \mid G_r \right)$$

$$= \sum_{t=1}^{T} \left( \sum_{i \in V_r} \sum_{k > i} \sum_{j \neq i, k} W_{i,j,r} W_{j,k,r} \left( \psi_{r,t}^3 + o(1) \right) \right) + \sum_{t=1}^{T} \sum_{\ell \neq t} \left( \sum_{i \in V_r} \sum_{k > i} \sum_{j \neq i, k} W_{i,j,r} W_{j,k,r} \left( \psi_{r,\ell}^2 \psi_{r,t} + o(1) \right) \right)$$

$$+ \sum_{t=1}^{T} \sum_{\ell \neq t} \sum_{h \neq t, \ell} \left( \sum_{i \in V_r} \sum_{k > i} \sum_{j \neq i, k} W_{i,j,r} W_{j,k,r} \left( \psi_{r,\ell} \psi_{r,t} \psi_{r,h} \right) \right).$$

The idea behind these expressions is similar to average degree but applied to triangles and triples (rather than single edges). Based on the expressions, we propose to multiply by $\psi_{r,t}^{-3}$ all triangles and triples composed of three individuals of type $t$, by $\psi_{r,\ell}^{-2} \psi_{r,t}^{-1}$ those with two individuals of type $t$ and one of type $\ell \neq t$, and finally by $(\psi_{r,\ell} \psi_{r,t} \psi_{r,h})^{-1}$ the triangles and triples containing three individuals of three different types.

The matters are again somewhat more complex for the star subgraph.

$$E(\rho(G_r^s)|G_r) = E \left( 3 \sum_{i \in S_r} \sum_{k > i} \sum_{j \neq i, k} W_{i,j,r}^s W_{j,k,r}^s W_{k,i,r}^s \mid G_r \right)$$

$$= \sum_{t=1}^{T} \left( 3 \sum_{i \in V_r} \sum_{k > i} \sum_{j \neq i, k} W_{i,j,r} W_{j,k,r} W_{k,i,r} \left( \psi_{r,t}^3 + 3 \psi_{r,t}^2 (1 - \psi_{r,t}) + o(1) \right) \right)$$

$$+ \sum_{t=1}^{T} \sum_{\ell \neq t} \left( 3 \sum_{i \in V_r} \sum_{k > i} \sum_{j \neq i, k} W_{i,j,r} W_{j,k,r} W_{k,i,r} \left( \psi_{r,\ell}^2 \psi_{r,t} + 2 (1 - \psi_{r,t}) \psi_{r,t} \psi_{r,\ell} + \psi_{r,\ell}^2 (1 - \psi_{r,\ell}) + o(1) \right) \right)$$

$$+ \sum_{t=1}^{T} \sum_{\ell \neq t} \sum_{h \neq t, \ell} \left( 3 \sum_{i \in V_r} \sum_{k > i} \sum_{j \neq i, k} W_{i,j,r} W_{j,k,r} W_{k,i,r} \left( \psi_{r,t} \psi_{r,\ell} \psi_{r,h} + (1 - \psi_{r,t}) \psi_{r,\ell} \psi_{r,h} + \psi_{r,t} (1 - \psi_{r,\ell}) \psi_{r,h} + \psi_{r,t} \psi_{r,\ell} (1 - \psi_{r,h}) \right) \right).$$

---

23As opposed to the average degree, this and all the subsequent expressions hold for any $T \geq 2$. 
E(\tau(G_r^*) | G_r) = E \left( \sum_{i \in S_r} \sum_{k > j, k \neq i, k} W_{ij,r}^s W_{jk,r}^s G_r \right)

= \frac{1}{n_r} \sum_{t=1}^{T} \left( \sum_{\{i \in V_r, k > j, k \neq i, k \}} \sum_{i' = t}^{i} \sum_{j' = t}^{j} W_{ij,r} W_{jk,r} (\psi_{r,\ell}^2 + 3\psi_{r,\ell}^2(1 - \psi_{r,\ell}) + \psi_{r,\ell}(1 - \psi_{r,\ell})^2 + o(1)) \right)

+ \frac{1}{n_r} \sum_{t=1}^{T} \left( \sum_{\{i \in V_r, k > j, k \neq i, k \}} \sum_{i' = t}^{i} \sum_{j' = t}^{j} W_{ij,r} W_{jk,r} (\psi_{r,\ell}^2 + 2(1 - \psi_{r,\ell})\psi_{r,\ell}\psi_{r,\ell} + \psi_{r,\ell}^2(1 - \psi_{r,\ell}) + o(1)) \right)

+ \frac{1}{n_r} \sum_{t=1}^{T} \left( \sum_{\{i \in V_r, k > j, k \neq i, k \}} \sum_{i' = t}^{i} \sum_{j' = t}^{j} W_{ij,r} W_{jk,r} (\psi_{r,\ell}^2 + \psi_{r,\ell}^2(1 - \psi_{r,\ell}) + (1 - \psi_{r,\ell})^2 \psi_{r,\ell} + o(1)) \right)

+ \frac{1}{n_r} \sum_{t=1}^{T} \left( \sum_{\{i \in V_r, k > j, k \neq i, k \}} \sum_{i' = t}^{i} \sum_{j' = t}^{j} W_{ij,r} W_{jk,r} (\psi_{r,\ell}^2 + (1 - \psi_{r,\ell})\psi_{r,\ell}\psi_{r,\ell} + \psi_{r,\ell}^2(1 - \psi_{r,\ell}) + (1 - \psi_{r,\ell})^2 \psi_{r,\ell} + o(1)) \right)

Again, CL’s corrections are special cases of these expressions. CL show that \( E[\rho(G_r^*) | G_r] = (3\psi_{r,\ell}^2 - 1 - \psi_{r,\ell}) + o(1)\rho(G_r) \) and \( E[\tau(G_r^*) | G_r] = (\psi_{r,\ell}^2(1 - \psi_{r,\ell}) + \psi_{r,\ell}^3 + o(1))\tau(G_r) \). Therefore, the analytically corrected estimator for \( c(G_r) \) based on \( G_r^* \) is \( \hat{c}(G_r^*) = (\psi_{r,\ell}^2(1 - \psi_{r,\ell}) + \psi_{r,\ell}^3 + o(1))^{-1} c(G_r) \). For the induced subgraph, since \( E[\rho(G_r^*) | G_r] = (\psi_{r,\ell}^2 + o(1))\rho(G_r) \) and \( E[\tau(G_r^*) | G_r] = (\psi_{r,\ell}^3 + o(1))\tau(G_r) \), the analytically corrected estimator of \( c(G_r) \) based on \( G_r^* \) is just \( c(G_r^*) \).

### 3.3 Epidemic Threshold

There is an increasing interest in understanding the diffusion properties of networks. In Economics, the application range from diffusion of innovation (e.g. Valente, 1995; Cowan and Jonard, 2004), product adoption (Banerjee et al., 2013; Hu et al., 2014), spread of information (Alatas et al., 2016) to spread of behaviors (Centola, 2010; Jackson and Yariv, 2007). The epidemic threshold is one way of how to quantify how easy it is for a disease, information, idea, or behavior to propagate through a network. Traditionally, the lower the threshold the easier is the propagation. There is a large variety of epidemic thresholds, depending on the diffusion conditions and network properties (see e.g. Vega-Redondo, 2007, or Jackson, 2010b). We focus on the following simple version, based on mean-field approximations:

\[
Thrld_r = \frac{1}{\frac{1}{n_r} \sum_{i \in V_r} \sum_{j \in V_r} \sum_{(i, j) \neq (i', j')} W_{ij,r}^s}.
\]

The threshold is simply the ratio between the average degree, \( d(G_r) \), and the average squared degree, denoted by \( d^2(G_r) \). The corrections of \( d(G_r) \) are derived in Appendix B.1. Here we discuss the correction of \( ds(G_r) \).

Note that for induced subgraph,

\[
E(\tau(G_r^*) | G_r) = E \left( \frac{1}{m_r} \sum_{i \in S_r} \left( \sum_{j \in S_r} W_{ij,r}^s \right)^2 \right)
\]

\[
= \frac{1}{m_r} \sum_{i, j \in S_r} W_{ij,r}^s G_r + E \left( \frac{1}{m_r} \sum_{i \in S_r} \sum_{j \in S_r, k \neq j} W_{ij,r}^s W_{jk,r}^s G_r \right)
\]

\[
= E(d(G_r^*) | G_r) + E \left( \sum_{i \in S_r} \sum_{j \in S_r, k \neq j} W_{ij,r}^s W_{jk,r}^s G_r \right).
\]
Thus we only need to derived the second term in the above equation. For the general case in induced subgraph,

\[
E\left(\frac{1}{m_r} \sum_{i \in S_r} \sum_{j \in S_r, k \neq j} W_{ij,r}^{|s} W_{ik,r}^{|s} \mid G_r\right)
\]

\[
= \frac{1}{n_r} \sum_{t=1}^{T} \left( \frac{\psi^3_{r,t}}{\psi_r} + o(1) \sum_{i \in V_r, j \in V_r, k \neq j} \sum_{t_i = t_j = t_k = t} W_{ij,r} W_{ik,r} \right) + \frac{1}{n_r} \sum_{t=1}^{T} \sum_{t' \neq t} \left( \frac{\psi^2_{r,t} \psi_{r,t'} + o(1)}{\psi_r} \sum_{i \in V_r, j \in V_r, k \neq j} \sum_{t_i = t_j = t_k = t} W_{ij,r} W_{ik,r} \right)
\]

\[
+ \frac{1}{n_r} \sum_{t=1}^{T} \sum_{t' \neq t} \sum_{h \neq t} \left( \frac{\psi_{r,t} \psi_{r,t'} \psi_{r,h}}{\psi_r} \sum_{i \in V_r, j \in V_r, k \neq j} \sum_{t_i = t_j = t_k = h} W_{ij,r} W_{ik,r} \right).
\]

Hence, in the general case, we propose to multiply \((\frac{\psi^3_{r,t}}{\psi_r})^{-1}\) to the triple \((i, j, k)\) in which three individuals are of the same type \(t\); multiply \((\frac{\psi^2_{r,t} \psi_{r,t'}}{\psi_r})^{-1}\) to the triple \((i, j, k)\) in which two individuals are of the same type \(t\) and the other is of type \(t'\); multiply \((\frac{\psi_{r,t} \psi_{r,t'} \psi_{r,h}}{\psi_r})^{-1}\) to the triple \((i, j, k)\) in which three individuals are of different types, \(t, t',\) and \(h\), to correct the second term of \(E(ds(G_r^{|s}))(G_r)\). The first term follows the correction of average degree.

Similar to the induced subgraph, we consider the following correction for the second term of \(E(ds(G_r^{|s}))(G_r)\) in the case of star subgraph

\[
E\left(\frac{1}{n_r} \sum_{i \in S_r} \sum_{j \in S_r, k \neq j} W_{ij,r}^{|s} W_{ik,r}^{|s} \mid G_r\right)
\]

\[
= \sum_{t=1}^{T} \left( (\psi_{r,t}^3 + 3 \psi_{r,t}^2 (1 - \psi_{r,t}) + \psi_{r,t} (1 - \psi_{r,t})^2 + o(1)) \sum_{i \in V_r, j \in V_r, k \neq j} \sum_{t_i = t_j = t_k = t} W_{ij,r} W_{jk,r} \right)
\]

\[
+ \sum_{t=1}^{T} \sum_{t' \neq t} \left( \psi_{r,t}^2 \psi_{r,t'} (1 - \psi_{r,t}) + \psi_{r,t} \psi_{r,t'} (1 - \psi_{r,t}) + o(1) \right) \sum_{i \in V_r, j \in V_r, k \neq j} \sum_{t_i = t_j = t_k = t} W_{ij,r} W_{jk,r}
\]

\[
+ \sum_{t=1}^{T} \sum_{t' \neq t} \left( \psi_{r,t}^2 \psi_{r,t'} (1 - \psi_{r,t}) + \psi_{r,t} \psi_{r,t'} (1 - \psi_{r,t}) + o(1) \right) \sum_{i \in V_r, j \in V_r, k \neq j} \sum_{t_i = t_k = t = t_j} W_{ij,r} W_{jk,r}
\]

\[
+ \sum_{t=1}^{T} \sum_{t' \neq t} \left( \psi_{r,t}^2 \psi_{r,t'} (1 - \psi_{r,t}) + \psi_{r,t} \psi_{r,t'} (1 - \psi_{r,t}) + o(1) \right) \sum_{i \in V_r, j \in V_r, k \neq j} \sum_{t_i = t_j = t_k = t} W_{ij,r} W_{jk,r}
\]

\[
+ \sum_{t=1}^{T} \sum_{t' \neq t} \sum_{h \neq t} \left( \psi_{r,t} \psi_{r,t'} \psi_{r,h} + (1 - \psi_{r,t}) \psi_{r,t'} \psi_{r,h} + \psi_{r,t} (1 - \psi_{r,t}) \psi_{r,h} + \psi_{r,t} \psi_{r,t'} (1 - \psi_{r,h}) + \psi_{r,t} (1 - \psi_{r,t}) (1 - \psi_{r,h}) \right) \sum_{i \in V_r, j \in V_r, k \neq j} \sum_{t_i = t_j = t_k = h} W_{ij,r} W_{jk,r}
\]

3.4 Graph span

Another important network measure is the distance between nodes. The path length between \(i\) and \(j\) is the minimum number of edges between them. The average path length is simply the average path length over all the finite paths. Naturally, the
shorter the distance between nodes the easier for them is to communicate, transmit information, or influence each other. Shorter average distances consequently allow for easier transmission in the whole population. Therefore, distances play important role in diffusion (similarly to the epidemic threshold), but also in risk-sharing or flow of capital among others. Several recent papers have analyzed distances in several applications. Examples include Kinnan and Townsend (2012), Leider et al. (2009), Goeree et al. (2010), Banerjee et al. (2013), and Alatas et al. (2016).

Despite useful in many applications, path lengths are complex objects and their analytical forms are only available for specific network architectures. Therefore, we follow CL and focus on graph span, a measure that approximates the average path length in many networks (Watts and Strogatz, 1998; Jackson, 2008). Graph span is defined as

\[
\ell(G_r) = \frac{\log n - \log d(G_r)}{\log d_2(G_r) - \log d(G_r)} + 1,
\]

where \(d_2(G_r) = \frac{1}{n} \sum_{s=1}^{n} \sum_{j>i} \sum_{k \neq i,j} W_{ij,r} W_{jk,r} \) is the average number of second-order neighbors (that is, nodes at distance two or simply neighbors of neighbors). Chandrasekhar and Lewis (2016) show that, for the star subgraph, 

\[
E[d_2(G_r^{s})|G_R] = (k(\psi) + o(1))d_2(G_r),
\]

where \(k(\psi) = \psi + \psi^2 - \psi^3\), while for the induced subgraph, 

\[
E(d_2(G_r^{s})|G_r) = (\psi^2 + o(1))d_2(G_r).
\]

Therefore, let \(\hat{d}_2(G_r^{s}) = d_2(G_r^{s})/k(\psi)\) and \(\hat{d}_2(G_r^{s}) = d_2(G_r^{s})/\psi^2\), the analytically corrected estimators for \(\ell(G_r)\) based on \(G_r^{s}\) and \(G_r^{I}\) are

\[
\hat{\ell}(G_r^{s}) = \frac{\log n - \log \hat{d}(G_r^{s})}{\log d_2(G_r^{s}) - \log d(G_r^{s})} + 1 \quad \text{and} \quad \hat{\ell}(G_r^{I}) = \frac{\log(\psi^{-1}m) - \log \hat{d}(G_r^{I})}{\log d_2(G_r^{I}) - \log d(G_r^{I})} + 1.
\]

For the general case of induced subgraph,

\[
E(d_2(G_r^{s})|G_r) = E \left( \frac{1}{n_r} \sum_{i \in S_r} \sum_{j>i} \sum_{k \neq i,j} W_{ij,r} W_{jk,r} \mid G_r \right)
\]

\[
= \frac{1}{n_r} \sum_{i=1}^{T} \left( \psi_{r,t}^3 + o(1) \sum_{i \in V_r, j>i} \sum_{k \neq i,j} \sum_{t_i=t_j=t_k=t_t} W_{ij,r} W_{jk,r} \right) + \frac{1}{n_r} \sum_{i=1}^{T} \sum_{t \neq t} \left( \psi_{r,t}^2 \psi_{r,t} + o(1) \sum_{i \in V_r, j>i} \sum_{k \neq i,j} \sum_{t_i=t_j=t_k=t_t} W_{ij,r} W_{jk,r} \right)
\]

\[
+ \frac{1}{n_r} \sum_{i=1}^{T} \sum_{t \neq t} \left( \psi_{r,t}^2 \psi_{r,t} + o(1) \sum_{i \in V_r, j>i} \sum_{k \neq i,j} \sum_{t_i=t_j=t_k=t_t} W_{ij,r} W_{jk,r} \right) \]

\[
+ \frac{1}{n_r} \sum_{i=1}^{T} \sum_{t \neq t} \sum_{h \neq t} \left( \psi_{r,t}^2 \psi_{r,t} \psi_{r,h} + o(1) \sum_{i \in V_r, j>i} \sum_{k \neq i,j} \sum_{t_i=t_j=t_k=t_t} W_{ij,r} W_{jk,r} \right).
\]

Hence, in the general case, we propose to multiply \(\psi_{r,t}^3 \psi_{r,t}^{-1}\) to the triple \((i, j, k)\) in which three individuals are of the same type \(t\); multiply \(\psi_{r,t}^2 \psi_{r,t}^{-1}\) to the triple \((i, j, k)\) in which two individuals are of the same type \(t\) and the other is of type \(\ell\); multiply \(\psi_{r,t} \psi_{r,t} \psi_{r,h}^{-1}\) to the triple \((i, j, k)\) in which three individuals are of different types, \(t, \ell, \) and \(h\), to correct the second term of \(E(ds(G_r^{s})|G_r)\).
For star subgraph, the general case correction is
\[
E(d_2(G_r)) | G_r = E \left( \frac{1}{n_r} \sum_{i \in S_r} \sum_{j > i} \sum_{k \neq i,j} W_{ij,r} W_{jk,r} | G_r \right)
\]
\[
= \sum_{t=1}^{T} \left( \psi_{r,t}^3 + 3 \psi_{r,t}^2 (1 - \psi_{r,t}) + \psi_{r,t} (1 - \psi_{r,t})^2 + o(1) \right) \sum_{i \in V_r} \sum_{j > i} \sum_{k \neq i,j} W_{ij,r} W_{jk,r}
\]
\[
+ \sum_{t=1}^{T} \sum_{\ell \neq t} \left( \psi_{r,t}^2 \psi_{r,\ell} + 2 (1 - \psi_{r,t}) \psi_{r,t} \psi_{r,\ell} + \psi_{r,\ell}^2 (1 - \psi_{r,\ell}) + o(1) \right) \sum_{i \in V_r} \sum_{j > i} \sum_{k \neq i,j} W_{ij,r} W_{jk,r}
\]
\[
+ \sum_{t=1}^{T} \sum_{\ell \neq t} \left( \psi_{r,t}^2 \psi_{r,\ell} + 2 (1 - \psi_{r,t}) \psi_{r,t} \psi_{r,\ell} + \psi_{r,\ell}^2 (1 - \psi_{r,\ell}) + o(1) \right) \sum_{i \in V_r} \sum_{j > i} \sum_{k \neq i,j} W_{ij,r} W_{jk,r}
\]
\[
+ \sum_{t=1}^{T} \sum_{\ell \neq t} \left( \psi_{r,t} \psi_{r,\ell} \psi_{r,h} + (1 - \psi_{r,t}) \psi_{r,t} \psi_{r,h} + \psi_{r,t} (1 - \psi_{r,t}) \psi_{r,h} + \psi_{r,t} \psi_{r,\ell} (1 - \psi_{r,\ell}) + (1 - \psi_{r,\ell}) \psi_{r,\ell} (1 - \psi_{r,h}) \right) \sum_{i \in V_r} \sum_{j > i} \sum_{k \neq i,j} W_{ij,r}
\]

3.5 Homophily index

Many social and economic networks exhibit a feature called homophily, a tendency to bond with similar individuals. In social and professional networks, who links up with whom is typically correlated with gender, age, race, and social and economic status and many 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 deal with types. Homophily is an important measure of across-type segregation and affects many economically relevant phenomena such as diffusion or learning and their speed (Golub and Jackson, 2012a,b), labour market outcomes (Calvo-Armengol and Jackson, 2004; Toomet et al., 2013), or individual and firm-level success (McPherson and Smith-Lovin, 1987; Ibarra, 1992) to name a few.

We adapt the homophily index from Currarini et al. (2009). For \( G_r \), the homophily index within group \( t \) is defined as
\[ H_t(G_r) = \frac{s_{r,t}}{s_{r,t} + d_{r,t}} \]
where \( s_{r,t}(G_r) \) denotes the average number of friendships that agents of type \( t \) have with agents of the same type and \( d_{r,t}(G_r) \) denotes the average number of friendships that type \( t \) form with agents of type different than \( t \). The homophily index of a network is simply the average homophily across all nodes. We use \( t \) to represent different demographic characteristics, such as gender, race, age, or their combinations (i.e. rake). Specifically, let \( V_{r,t} \) denotes a set of notes with type \( t \).

\[ s_{r,t} = \frac{1}{n_{r,t}} \sum_{i,j \in V_{r,t}} W_{ij,r}, \quad d_{r,t} = \frac{1}{n_{r,t}} \sum_{i \in V_{r,t}} \sum_{j \notin V_{r,t}} W_{ij,r}. \]

In the general case of induced subgraph, fixing type \( t \), we have
\[
E(s_{r,t}(G_r)) = E \left( \frac{1}{m_{r,t}} \sum_{i \in S_r} \sum_{j \in V_{r,t}} W_{ij,r}^{\|n} | G_r \right)
\]
\[
= \frac{1}{\prod_{q=1}^{T} (m_{r,q})^{n_{r,q}}} m_{r,t} \sum_{i,j \in V_{r,t}} W_{ij,r} \left( m_{r,t} - 2 \right) \prod_{q \neq t} (m_{r,q})^{n_{r,q}}
\]
\[
= (\psi_{r,t} + o(1)) \frac{1}{n_{r,t}} \sum_{i,j \in V_{r,t}} W_{ij,r},
\]
and

\[ E(d_{r,t}^s | G_r) = \sum_{t \neq t} \mathbb{E} \left( \frac{1}{m_{r,t}} \sum_{i,j \in S_r, \, t_i = t, t_j = t} W_{ij,r}^s | G_r \right) \]

\[ = \sum_{t \neq t} \left[ \frac{1}{\prod_{q=1}^{T} (n_{r,q}) \frac{1}{m_{r,t}}} \sum_{i,j \in V_r, \, t_i = t, t_j = t} W_{ij,r} \left( \frac{n_{r,t} - 1}{m_{r,t} - 1} \right) \left( \frac{n_{r,k} - 1}{m_{r,k} - 1} \right) \prod_{q \neq t , \ell} (n_{r,q}) \right] \]

\[ = \frac{1}{n_{r,t}} \sum_{t \neq t} \left( \psi_{r,t} \sum_{i,j \in V_r} W_{ij,r} \right). \]

Therefore, we propose multiply \( \psi_{r,t}^{-1} \) on each link for the calculation of \( s_{r,t}^s \) and \( \psi_{r,t}^{-1} \) for \( d_{r,t}^s \).

In a general case of star subgraph, fixing type \( t \), we have

\[ E(s_{r,t}^s | G_r) = E \left( \frac{1}{n_{r,t}} \sum_{i,j \in S_r} W_{ij,r}^s | G_r \right) \]

\[ = \frac{1}{\prod_{q=1}^{T} (n_{r,q}) \frac{1}{m_{r,t}}} \sum_{i,j \in V_r} W_{ij,r} \left( \frac{n_{r,t} - 2}{m_{r,t} - 2} + 2 \left( \frac{n_{r,t} - 2}{m_{r,t} - 1} \right) \right) \prod_{q \neq t} (n_{r,q}) \]

\[ = \left( \psi_{r,t}^2 + 2 \psi_{r,t} (1 - \psi_{r,t}) + o(1) \right) \frac{1}{n_{r,t}} \sum_{i,j \in V_r} W_{ij,r} \]

and

\[ E(d_{r,t}^s | G_r) = \sum_{t \neq t} \mathbb{E} \left( \frac{1}{n_{r,t}} \sum_{i,j \in S_r, \, t_i = t, t_j = t} W_{ij,r}^s | G_r \right) \]

\[ = \sum_{t \neq t} \left[ \frac{1}{\prod_{q=1}^{T} (n_{r,q}) \frac{1}{m_{r,t}}} \sum_{i,j \in V_r} W_{ij,r} \left( \frac{n_{r,t} - 1}{m_{r,t} - 1} \right) \left( \frac{n_{r,k} - 1}{m_{r,k} - 1} \right) + \left( \frac{n_{r,t} - 1}{m_{r,t} - 1} \right) \left( \frac{n_{r,k} - 1}{m_{r,k} - 1} \right) \prod_{q \neq t , \ell} (n_{r,q}) \right] \]

\[ = \frac{1}{n_{r,t}} \sum_{t \neq t} \left( \psi_{r,t} \psi_{r,\ell} + \psi_{r,t} (1 - \psi_{r,t}) + \psi_{r,\ell} (1 - \psi_{r,\ell}) + o(1) \right) \sum_{i,j \in V_r} W_{ij,r} \right). \]

Therefore, we propose multiply \( \left( \psi_{r,t}^2 + 2 \psi_{r,t} (1 - \psi_{r,t}) \right) \) on each link for the calculation of \( s_{r,t}^s \) and \( \left( \psi_{r,t} \psi_{r,\ell} + \psi_{r,t} (1 - \psi_{r,t}) + \psi_{r,\ell} (1 - \psi_{r,\ell}) \right)^{-1} \) for \( d_{r,t}^s \).

### 4 Monte Carlo Simulations

This section evaluates numerically the extent of biases in the network measures and the estimates using these measures as regressors, depending on the sampling procedure (induced vs. star graph), sampling rate and the (non-)randomness of the sample. We quantify the biases in raw sampled data and in corrections based on the missing-at-random assumption, and compare their performance vis-à-vis our poststratification weighting. In this section, we concentrate on a scenario that mimics our modeling assumptions. This provides a natural testing ground of our approach.
There are several common features of the following two subsections. First, our artificial population data are generated on basis of the widely employed Add Health data. By adopting the Add Health school sample as a phototype, we can preserve certain real-life relationships between individuals’ characteristics and the network. For example, white students have on average more network connections than black students and the latter are on average more connected than other races in the data, or the patterns of homophily depend systematically on the race composition of each school (Currranini et al., 2009, 2010).

For each population network, we adjust the size of the artificial population to 1,500 so that the numbers of whites, blacks, and other races are all equal to 500 in each population. We use three demographic characteristics from the phototype to define individual’s type: seniority dummy (C1), gender dummy (C2), and race (C3). Seniority takes value of one if an individual is older than the network average and zero otherwise. For the gender dummy, one stands for males and zero for females. As for race, one denotes White, two Black, and three stands for other races. We also combine these three characteristics to form $2 \times 2 \times 3 = 12$ cross-characteristic, denoted rake throughout. 24 Seniority and gender are largely uncorrelated with individual network connectivity and they provide little information about one’s positioning and overall social organization in a school. In contrast, race is strongly correlated with network degree in the data. The average degree of white students is 9.60, black students have an average of 7.38 friends, whereas the average connectivity is 4.39 for other races.

From the above described artificial population, we vary the removal strategy in three dimensions to generate our sampled networks. First, we apply two sampling procedures, induced and star networks. In the induced subgraphs, we remove a fraction of nodes and all their links (including their connections to the non-removed individuals); for the star subgraphs, we remove a fraction of nodes and only their links to other removed individuals. Second, we consider three sampling rates, $\psi = 40\%$, $60\%$, and $80\%$ (or alternatively three missing rates, $1 - \psi = 60\%$, $40\%$, and $20\%$, respectively). Second, we employ four removal strategies with respect to the representativeness of the artificially sampled subgraphs. We either remove people randomly (Scenario R) or on basis of their connectivity. In the latter case, we employ three scenarios: (i) removal of high-degree nodes with higher probability (Scenario H), (ii) removal of intermediate-degree nodes with higher probability (Scenario M), and (iii) removal of low-degree nodes with higher probability (Scenario L). 25 When we generate our artificial samples and would like to get non-random (disproportional) missing, we use the characteristic race for such purposes. More precisely, if we want to remove highly connected nodes with higher probability, we remove white students with higher probability and so on. We perform 100 repetitions of each constellation for both induced and star subgraphs. We focus on the five network statistics treated in Section 3. In case of homophily, we mostly focus on rake homophily but homophily on other variables is sometimes treated for illustrative purposes.

### 4.1 Network characteristics

As a first step, we quantify the biases in network measures in sample networks using raw data, corrections assuming representativeness of the sample and our poststratification weighting approach, depending on the sampling rate and the (non-)randomness of the sample.

Figures 1 and 2 (and Tables A1 - A6 in Appendix A) report the first set of results from our Monte Carlo simulations. Figure 1 (and Tables A1 - A3) reports the results for the induced subgraph; Figure 2 (and Tables A4 - A6) document those for the star-network sampling. In Figures 1 and 2, the $y$-axes reflect the absolute-value of the biases with respect to the population values. The $x$-axes list the five network characteristics under scrutiny in the following order: average degree, total clustering, graph span, epidemic threshold, and homophily on rake. 26 The blue plots represent the raw sample data and the red graphs reflect the random corrections of CL. The remaining cases are variations of our methodology. The green graphs weight on the network-unrelated C1. The last two plots represent, respectively, the weighting on C3 only (violet) and rake (i.e. the combination of C1 – 3; cyanic). The rows and columns represent the four different removal strategies, Scenario R, H, M, L, and the three sampling rates, $\psi = 80\%$, $60\%$, and $40\%$, in these orders.

**Biases in the raw data.** We first discuss the biases that arise in the considered measures if raw sampled data are used to compute them. The exercise reveals that treating the data “as if” complete leads to large differences between the population and sample networks under virtually all removal strategies and in case of all network characteristics. Not surprisingly, the biases are larger in the induced subnetwork (as less information is available about the network, conditional on the sampling rate) and decrease with the sampling rate (increase with the missing rate). The most biased characteristics are variables that do not represent percentages; that is, average degree, graph span, and the epidemic threshold. The raw sample data consistently make the network look as less connected, as exhibiting longer average distances, and as less epidemic-prone than it actually is. All these findings are direct consequences of observing less links than there actually exist in the population.

To provide some—despite rough—quantification of the biases in the sample average degrees, the biases are respectively in the range of 20\%, 40\%, and 50\% in the induced subgraphs and 3\%, 15\%, and 20\% in the star subgraphs for $\psi = 80\%$, 60\%, and 40\%. The extent of biases in the average degree is clearly associated with who is removed and in the expected direction. Under

---

24Rake combines the previous three variables such that the types are for example “senior black female”, “junior male of other race” and so on.

25We prefer the notation R(andom), H(igh), M(edium), and L(ow) to e.g. Scenario 1 - 4 to make sure that readers associate the letter with the removal strategy.

26The results are qualitatively similar if we focus on gender or race homophily instead. We thus omit them here.
Figure 1: Induced subgraph: Biases (%) of network measures and their corrected versions with respect to the population network for $\psi = 0.8$ (left), 0.6 (center), 0.4 (right) and four different removal strategies.
Figure 2: Star subgraph: Biases (%) of network measures and their corrected versions with respect to the population network for $\psi = 0.8$ (left), $0.6$ (center), $0.4$ (right) and four different removal strategies.
non-random missing and conditional on $\psi$, we always detect the largest biases when there is a tendency not to observe relatively connected individuals, followed by the removal of nodes with intermediate connectivity in the second place, and with the removal of low-degree nodes exhibiting the lowest biases in average degrees. Random removal mostly leads to biases comparable to the removal of nodes with intermediate degrees. As for the graph span, the sample networks always exhibit longer average distances, compared to the population graphs, and the biases follow the same patterns as the average degree: they rise rise with the missing rate and the tendency to remove more connected nodes. The biases are respectively $8\%-20\%$, $21\%-69\%$, and $38\%-386\%$ for the induced subgraph and $1\%-4\%$, $5\%-16\%$, and $11\%-52\%$ for the star subgraph. Last, observing only a sample of nodes increases the epidemic threshold. Once again, larger missing rate is associated with larger biases and with the tendency to remove more connected individuals, with the exception of the induced network and $\psi = 80\%$ where the largest bias occurs if intermediate nodes are missing with higher probability. The epidemic threshold is biased respectively almost up to $36\%$, $77\%$, and $163\%$ in the induced graphs, and $5\%$, $19\%$, and $44\%$ in the star subnetworks.

The biases are quantitatively much lower and exhibit more complex patterns in case of the clustering coefficient and the homophily index, two network characteristics that represent shares. They also increase with the sampling rate, but they are not necessarily lower in the star graph and their magnitudes and signs are highly sensitive on the removal strategy.

As for the clustering coefficient in the induced subgraph, it is only mildly biased (less than $2\%$) under random removal and this removal scheme always deviates the clustering coefficient the least from its true population value. Removing more connected nodes with higher probability (Scenario H) always drives the clustering coefficient down, while Scenarios M and L inflate it. Quantitatively speaking, the sampled clustering coefficients are biased (downwards) by $-5.1\%$, $-10.7\%$, and $-16.8\%$ for $\psi = 80\%$, $60\%$, and $40\%$ when more higher-degree nodes are not observed; the corresponding figures are positive and relatively similar for Scenarios M and L ($3.9\%$, $9.3\%$, and $16.7\%$ vs. $4.8\%$, $10.9\%$, and $23.4\%$).

In contrast to the induced graph, the biases in the clustering coefficient are always negative and even the random removal can bias the coefficient downwards considerably under the star-network sampling scheme. Hence, the network always looks less clustered than it actually is in the star subnetworks. Under random missing, the observed biases are $-3.4\%$, $-12.6\%$, and almost $-30\%$ as the missing rate increases. Compared to the missing-at-random situation, the biases are always larger if high-degree nodes are removed with higher probability (Scenario H: $6.1\%$, $22.0\%$, and $45.6\%$ for $\psi = 80\%$, $60\%$, and $40\%$) and lower in the other two cases (Scenario M: $-2.6\%$, $-8.3\%$, and $-13.35\%$; Scenario L: $-1.6\%$, $-6.1\%$, and $-12.42\%$).

Regarding homophily on rake, it exhibits no biases under random missing (less than $0.15\%$ and even $0\%$ for the induced and star subgraphs, respectively). For non-random missing, the biases again decrease with the sampling rate. The degree of homophily is mostly biased downwards for both induced and star graphs and any removal strategy. The only regular exception occurs for Scenario IV when induced networks are elicited; one non-regular exception is the case of higher missing of high-degree nodes if $\psi = 40\%$. Virtually no biases are detected in the star graph and they are very low for the induced graph if $\psi = 80\%$. For $\psi = 60\%$, the biases in absolute value do not exceed $7\%$ and $4\%$ in the induced and star graph, respectively; the biases start to be more severe ($10\%$ or more) if the sampling rate is low. Overall, sampling makes the networks look less homophilous that they are, except when relatively non-connected nodes are not sampled with higher probability (this happens e.g. under snow-ball sampling).

**Biases in the corrections.** The second objective of this section is to compare the raw sample statistics with the random corrections and three variations of weights that apply our methodology. As mentioned above, seniority (C1) is largely uncorrelated with network positions, while race (C3) is associated with connectivity and the relationships are homophilous in C3. Therefore, we first discuss the case of C1. One would expect that C1 does not correct much of the non-representativeness of the sample but still accounts for the scaling effect. This exercise is of interest to illustrate what happens when a researcher weights on a variable that does not contain information about the network. Can it hurt, rather than help? If so, one should be very careful selecting the variables. Second, we weight on C3, which should correct the scaling problem and provide additional improvement, since this variable correlates with one’s position. Last, we weight on rake (that is, the combination of C1, C2, and C3). We hypothesize these corrections to outperform all the previous cases, since they employ the most information available.

The general findings are the following. With very few exceptions, all correction strategies outperform the raw data. As expected, the corrections assuming randomness work well under random sampling and their performance is similar to our weighting methodology. Nevertheless, the random corrections and the weighting approach diverge once the missing-at-random assumption is not respected in the sampling strategy. As hypothesized, the random corrections and weighting on the network-irrelevant C1 perform very similarly overall. This is an important result, since it shows that weighting on irrelevant variables still mitigates the biases and does not hurt, compared to the raw data and the random corrections. However, these corrections are still biased and these biases increase with the missing rate and are lower in the star-subnetwork sampling scheme.

The biases are minimal if we weight either on C3 or rake and both approaches virtually always outperform all other methods. In fact, both weighting schemes almost eliminate any biases arising from sample network data. The maximum biases are $1.1\%$, $1.6\%$, and $6.8\%$ for $\psi = 80\%$, $60\%$, and $\psi = 60\%$ for the induced subgraphs; the figures are even lower for the star subnetwork.

Observe that the corrections proposed in this paper are unbiased on average, but they are still subject to an error in finite samples. It is even possible that a correction works well on average but the corrected network statistics may still inflate the
estimated network effects. Since Figures 1 and 2 (and Tables A1 - A6 in Appendix A) only report the average performance, the panels in Figures A9 - A16 in the Appendix complement them. The latter figures plot the true population network statistics on the y-axes and, respectively, their raw sample network variants, the corrections based on randomness, and the corrections weighted on C1, C2, C3, and rake on the horizontal axes for the same network statistics, removal scenarios, and sampling rates as above. Each circle correspond to one network. The reported graphs particularly enable to evaluate two features, (a) the slope of the corrections, in comparison with the 45° line, and (b) their dispersion around the 45° degree line. If the raw data or any of the corrections are free of any error, all observations in the panel would lie right on the 45° line.

The plots reveal that the raw data are systematically biased, independently of the parameter and removal strategy. In line with the above, the corrections based on the missing-at-random assumption and our poststratification approach perform similarly under random missing in that they are distributed around the 45° line, but we can already appreciate a reduced dispersion in the corrections that take into account the non-randomness of the sample. Hence, both corrections eliminate the expansion problem and both will only suffer form attenuation, but the attenuation tend to be more severe with random corrections compared to our approach.

Once we depart from the random-missing assumption though, weighting on C3 and rake (that includes C3) clearly outperform any other method both in that the corrections exhibits slopes equal to the 45° degree line and in that they reduce the dispersion vis-à-vis raw-data statistics as well as the random corrections. Therefore, our methodology will improve inference on sampled networks and will deliver the least biased estimates overally and even if only attenuation takes place.

### 4.2 Network effects

We now turn the attention to the performance of the poststratification weighting in a regression framework, aiming at estimating population network effects on economic outcomes. Since our weighting approach is designed for correcting global network measures, the independent variables (network measures) in the regression are measured at the network-wide level. We also limit the present analysis to network-level dependent variables, but—as discussed above—the results can be extended to models, in which individual behaviors or outcomes are regressed on global network properties.

To generate the population data, we again take the Add Health school sample adjusted to the size of 1,500 as the phototype. We duplicate 200 artificial networks with node characteristics adopted from this phototype sample. Then, based on the average connectivity, clustering coefficient, and homophily index corresponding to different types of individuals in this phototype sample, we simulate network links in these 200 artificial networks. That is, we generated 200 population networks that have the same size, same node characteristics (i.e., C1, C2, and C3), but different network configurations. Particularly, simulated links exhibit uneven connectivity across types, where white nodes on average have the highest degree, followed by blacks and nodes of other races. There are also features of clustering and homophily in different types. To simulate the population dependent variable in each network, we use a simple linear regression model:

\[
y_r = \alpha + \beta w(G_r) + \varepsilon_r, \quad \varepsilon \sim \text{N}(0,0.5)
\]

For each network, we simulate network links in these 200 artificial networks. That is, we generated 200 population networks that have the same connectivity, clustering coefficient, and homophily index corresponding to different types of individuals in this phototype sample, we simulate network links in these 200 artificial networks. That is, we generated 200 population networks that have the same size, same node characteristics (i.e., C1, C2, and C3), but different network configurations. Particularly, simulated links exhibit uneven connectivity across types, where white nodes on average have the highest degree, followed by blacks and nodes of other races. There are also features of clustering and homophily in different types. To simulate the population dependent variable in each network, we use a simple linear regression model:

\[
y_r = \alpha + \beta w(G_r) + \varepsilon_r, \quad \varepsilon \sim \text{N}(0,0.5)
\]

The main results regarding the biases mimic those of network characteristics in most respects. The biases in the estimates—both from the raw data and after applying the corrections—increase in the missing rate, are higher for the induced networks compared to star subgraphs, where less data is missing (conditional on the sampling rate), and depend on who is missing. The raw network data exhibit the largest biases in the estimates for all network statistics. The performance of the corrections based on randomness and the weighting approach is similar if the data are missing at random, but they generally diverge otherwise. Both types of corrections mitigate the biases with respect to the raw data, but—some exceptions—accounting for the non-representativeness of the sample reduces the biases with respect to corrections based on randomness. This difference is not as large as in the star graph case and important biases exist even though we correct well the network statistics. The reason is that the corrected statistics still exhibit dispersion compared to the true population statistics as shown in the Appendix.

In the induced case, the biases in the estimates using raw data are typically large in percentage terms. The largest biases are observed in case of average degree, epidemic threshold, and graph span. These biases are reduced considerably when corrections are applied. As for the comparison between random corrections and post-stratification weighting, the former already outperform the latter under random removal. Remember that this due to the lower dispersion of our corrections discussed above, and the

---

27 Imagine a statistic well corrected on average, but inflated for below-average values while reduced for above-average values. This is, for instance, the case of the clustering coefficient using the raw data or corrections based on randomness in several instance in Figures A9 - A16. Regressing a variable of interest on such incorrectly corrected statistic in a linear regression will naturally lead to expansion and the possibility to accept a network effect that do not exist.

28 In contrast to the previous subsection, we omit the results corresponding to weighting on seniority, gender, or race only. The biases behave similarly to the biases detected in the previous subsection.
Figure 3: Induced subgraph: Biases (%) in network effects and their corrected versions with respect to the population network for $\psi = 0.8$ (left), 0.6 (center), and 0.4 (right) and four different removal strategies.
Figure 4: Star subgraph: Biases (%) in network effects and their corrected versions with respect to the population network for $\psi = 0.8$ (left), 0.6 (center), and 0.4 (right) and four different removal strategies.
estimates under our approach are thus less attenuated. Once random removal is abandoned, the differences between both correction types become larger even though there are cases, in which both methodologies perform similarly. In quantitative terms and with some exceptions, our corrections exhibit biases below 5%, 10%, and 20% as the missing rate increases in Figure 3, while these numbers are in the order of 10%, 40%, and 50-60% under random corrections. Figure 3 only reflects the size of the biases but hide their direction. With raw network data, we observe expansion of network effects on average in 22% instances (out of the $8 \times 12 = 96$ parameters) in Figure 3. The effect of average degree is always overestimated while the remaining cases of expansion is due to the mismeasured homophily. Other variables exhibit attenuation. When the corrected statistics are applied as regressors, 12.5% (12 cases) of estimated parameters are expanded under random corrections; this number lowers to 11.5% (11 regressions) if non-randomness taken into account. If we solely count the number of times such expansion biases the estimated in more than 1%, assuming randomness generates expansion in 8 regressions while correcting for non-randomness lowers it further to 5.

The star-sampling statistics exhibit lower biases than the induced graphs and this also holds for estimates when applied as regressors. The biases in the estimates using the raw data are mostly below 5%, 20%, and 50% for $\psi = 80\%$, $60\%$, and $40\%$ in all but Scenario H, in which these figures are higher. Random corrections mostly reduce these biases, but following the missing-at-random assumption particularly fails to recover the effect of the clustering coefficient. In fact, random corrections do not outperform raw data in reducing the biases if total clustering is applied as regressor. The post-stratification weighting clearly outperforms these approaches. Observe that there remain only minimal biases when one accounts for non-representativeness of the sample. The remaining biases are typically below 1%. Hence, our approach is remarkably successful reducing the biases under star-network sampling.

5 Empirical applications

In this section, we apply the proposed methodology to two data sets with the objective to illustrate how statistical inference can be affected if one accounts for both the scaling effect and non-representativeness. In the first subsection, we use network data from rural India. In the second part of this section, we use adolescent friendship network from U.S. middle and high schools. Both data sets contain certain non-network information on the whole population, but only sampled data on networks.

5.1 Village Networks in Rural India

We start applying our approach on data from Banerjee et al. (2013) who elicit a large variety of characteristics including network data from 75 villages in Karnattaka, India. The authors initially collected certain data on all households in all villages and later detailed follow-up survey with a subsample of the population of each village. In the latter, they also elicit the network of relationships within the village.

Two features of these data make them particularly interesting for our purpose. First, as in most studies, the network data only represent a sample of each village. Their sampling rate is on average 46%. The crucial aspect of the sampling design is the stratification by religion and geographic sublocation, generating a representative sample with respect to these two variables. Nevertheless, Table 1 reveals that the data are not representative in terms of age, gender, and household size. Below, we test how the differences between the sample and population rates affect network effects.
Table 1: Average per-village population and sample percentages of different age categories, gender ratios, and household sizes in the data from Banerjee et al. (2013).

<table>
<thead>
<tr>
<th></th>
<th>Population</th>
<th>Sample</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt; 30</td>
<td>38.71%</td>
<td>30.97%</td>
<td>7.74%</td>
</tr>
<tr>
<td>30-50</td>
<td>39.60%</td>
<td>54.11%</td>
<td>-14.51%</td>
</tr>
<tr>
<td>&gt; 50</td>
<td>21.69%</td>
<td>14.92%</td>
<td>6.77%</td>
</tr>
<tr>
<td>Sex ratio</td>
<td>50.34%</td>
<td>44.57%</td>
<td>5.77%</td>
</tr>
<tr>
<td>Household size</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt; 3</td>
<td>17.26%</td>
<td>15.49%</td>
<td>1.77%</td>
</tr>
<tr>
<td>3-8</td>
<td>71.57%</td>
<td>73.48%</td>
<td>-1.91%</td>
</tr>
<tr>
<td>&gt; 8</td>
<td>11.17%</td>
<td>11.03%</td>
<td>0.14%</td>
</tr>
<tr>
<td># of Villages</td>
<td>75</td>
<td>75</td>
<td></td>
</tr>
<tr>
<td>Total Obs.</td>
<td>48,646</td>
<td>16,995</td>
<td></td>
</tr>
</tbody>
</table>

Second, the data contain several variables regarding the labour market outcomes of the participants. Since the important role of social networks in labour markets is widely acknowledged in the literature and documented in the data, this application is interesting in its own right. The theoretical literature argues that the degree distribution (Calvo-Armengol and Jackson, 2004) or the average clustering coefficient (Espinosa et al., 2018) might affect the employment prospects directly, while network distances and the epidemic threshold might influence the flow of labour-market information and thus the labour outcomes indirectly. However, little empirical evidence exists regarding the impact of global features of networks on labour market outcomes, probably due to the lack of suitable data to test this issue. We thus ask how the fraction of people working outside the village and the village unemployment rate are determined by the global features of the underlying network of relationship within the village. Most importantly for the present study, we ask how the estimated network effects change if we account for missing network data and the non-randomness of the sample. We particularly hypothesize that the underrepresentation of people aged 30-50 in Table 1, those who typically actively participate in labour markets, might bias the estimated network effects if their underrepresentation is not accounted for.

Table 2 reports the estimated network effects in a series of estimations differing in (i) the dependent variable (left: fraction of population working outside the village; right: unemployment rate), (ii) whether raw or corrected networks are used (columns) and (iii) different network characteristics (rows). Once again, to separate the effect of scaling from the effect on non-randomness of the sample, we use the raw network data, corrections based on randomness, and our approach in which we weight on a rake variable (incorporating the information on age, gender, and household size). Each cell report the estimated network effects (and below the standard error robust to heteroscedasticity in parentheses) from a regression of one dependent variable on the corresponding network statistic and village size.

quantitatively speaking, the biases with respect to our corrections are very high. They range from 10% to 700% in case of the raw data and from 4% to 6000% under random corrections. The last figure is an outlier in Table 2, but there are still several instances of biases above 700% if only scaling is considered. Therefore, the biases under corrections based on missing-at-random assumption can be as high as under raw data. This illustrates the severity of inferences using non-representative samples even if one accounts for potential scaling issues.

5.2 Adolescent friendship networks in U.S. high schools

As a second application, this subsection applies our approach to the AdHealth data set. These data contain extensive information on friendship networks in selected group U.S. high/middle schools and detailed data on individual heterogeneity. The schools are representative. Even though the data collectors interviewed all the students present during the questionnaire day, the

<table>
<thead>
<tr>
<th>Dep. Variables</th>
<th>Share working outside the village</th>
<th>Share employed</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>raw CL Rake</td>
<td>raw CL Rake</td>
</tr>
<tr>
<td>Degree</td>
<td>-0.0461*** -0.0166*** -0.0165**</td>
<td>0.0357*** 0.0121*** 0.0119***</td>
</tr>
<tr>
<td></td>
<td>(0.0162) (0.0061) (0.0070)</td>
<td>(0.0101) (0.0037) (0.0041)</td>
</tr>
<tr>
<td>Cluster</td>
<td>-0.3436*** -0.3436*** -0.2995***</td>
<td>0.1717*** 0.1717*** 0.1696***</td>
</tr>
<tr>
<td></td>
<td>(0.1276) (0.1276) (0.0992)</td>
<td>(0.0642) (0.0642) (0.0640)</td>
</tr>
<tr>
<td>Span</td>
<td>0.0573*** 0.1759*** 0.1740***</td>
<td>-0.0259** -0.0560 -0.0511</td>
</tr>
<tr>
<td></td>
<td>(0.0174) (0.0597) (0.0576)</td>
<td>(0.0099) (0.0349) (0.0323)</td>
</tr>
<tr>
<td>Epid Thld</td>
<td>1.8841*** 4.1914*** 4.0526***</td>
<td>-1.4125*** -2.8273*** -2.5861***</td>
</tr>
<tr>
<td></td>
<td>(0.5658) (1.3554) (1.3836)</td>
<td>(0.3822) (0.8907) (0.8764)</td>
</tr>
<tr>
<td>HI-sex</td>
<td>-0.1675 -0.1675 -0.0892</td>
<td>0.2398** 0.2398** 0.1494</td>
</tr>
<tr>
<td></td>
<td>(0.2171) (0.2171) (0.2333)</td>
<td>(0.1046) (0.1046) (0.0920)</td>
</tr>
<tr>
<td>HI-age</td>
<td>-0.3321 -0.3321 0.7660**</td>
<td>0.5433** 0.5433** -0.2299</td>
</tr>
<tr>
<td></td>
<td>(0.4815) (0.4815) (0.3342)</td>
<td>(0.2507) (0.2507) (0.2261)</td>
</tr>
<tr>
<td>HI-households</td>
<td>0.4489*** 0.4489*** 0.1844</td>
<td>0.0409 0.0409 0.0028</td>
</tr>
<tr>
<td></td>
<td>(0.1422) (0.1422) (0.1709)</td>
<td>(0.0964) (0.0964) (0.0942)</td>
</tr>
<tr>
<td>HI-pool</td>
<td>0.6645* 0.6645* 0.9235***</td>
<td>0.3657*** 0.3657*** 0.0723</td>
</tr>
<tr>
<td></td>
<td>(0.3563) (0.3563) (0.3375)</td>
<td>(0.1741) (0.1741) (0.1348)</td>
</tr>
</tbody>
</table>

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

If we did not account for the non-representativeness of the sample, the general picture from Table 2 would be that the social organization in the village plays a major role in determining labour market performance. However, once we do account for non-random missing, considerably less network features affect the dependent variables significantly. That is, many estimates suffer from expansion using raw data as well as the corrections based on randomness. Only the epidemic threshold and rake homophily predict systematically that robustly relates to unemployment rates. Higher epidemic threshold suggest less integration in the village, hindering among other things the flow of job-related information from more distant network neighborhoods. Similarly, higher homophily reveals less connections and thus lower information flow across different subgroups in the village population. Both features disallow an efficient transmission of information about vacancies within a village, making people to travel for work outside the village more, compared to more integrated villages. As for the clustering coefficient, it is one important measure of whether people take care of each other within a village in adverse situations, such as e.g. unemployment (Coleman, 1988). Indeed, the estimated effect is positive, suggesting that more clustered villages exhibit higher employment rates.

Regarding the main purpose of our exercise, the estimates in Table 2 confirm that using raw network data or the corrections based on missing-at-random assumptions bias them in no specific direction. The direction and extent of biases depend on the network variable of interest and who is missing. We document instances of attenuation and no difference in the estimates between the three specifications, but expansion and sign switching are commonplace. Expanded estimates are observed in 50% of regressions with raw data and 62.5% under random corrections. Compared to accounting for non-representativeness of the sample, the signs are switched in other 25% and 31.25% regressions, respectively. In a similar vein, the network effects are significant or not under the three specifications in some instances, but we report significant associations using the raw data and random corrections but not if we account for non-random missing in cases of the average degree and some homophily indices. The effect of epidemic threshold is particularly sensitive to the network specification: correcting for scaling but not non-representativeness always suggests larger effects than raw data but correcting additionally for non-representativeness lowers it, compared to random corrections, in both cases; the effect on unemployment even becomes nonsignificant at 5% if our corrections are applied.

Table 2: Estimated network effects on the share of population working outside the village and employment probability in rural India – based on All-1 network
average sampling rate is 63.6% of the school census. Since the participation was determined non-randomly, we can expect non-representativeness of the sample. We combine these sampled network data with certain information on the roster in each school. Table XX shows the sample and population statistics regarding the variables available at the population level. Observe that ....  

Table 3: Average per-school population and sample percentages of different race categories and year grades in the Add Health data.

<table>
<thead>
<tr>
<th>Race</th>
<th>Population</th>
<th>Sample</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>White</td>
<td>60.43%</td>
<td>53.32%</td>
<td>7.11%</td>
</tr>
<tr>
<td>Non-White</td>
<td>39.57%</td>
<td>46.68%</td>
<td>-7.11%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Year grade</th>
<th>Population</th>
<th>Sample</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>9th grade</td>
<td>32.22%</td>
<td>30.00%</td>
<td>2.22%</td>
</tr>
<tr>
<td>10th grade</td>
<td>25.63%</td>
<td>26.53%</td>
<td>-0.90%</td>
</tr>
<tr>
<td>11th grade</td>
<td>22.06%</td>
<td>23.00%</td>
<td>-0.94%</td>
</tr>
<tr>
<td>12th grade</td>
<td>20.09%</td>
<td>20.47%</td>
<td>-0.38%</td>
</tr>
</tbody>
</table>

The particular application presented here is inspired by Bramoullé et al. (2009) who find large peer effects in club participation using these data. We thus expect at least some network properties to predict the average club participation at the level of schools. To this aim, we regress the average club participation in each school on the same network characteristics as in Section 5.1. Table 4 reports the estimates using again raw network data, networks corrected for scaling and those corrected by our poststratification weighting.

Table 4: Estimated network effects on club participation in U.S. high schools.

<table>
<thead>
<tr>
<th>Dep. variable</th>
<th>Average number of clubs attended</th>
<th>Average frequency of exercise</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Raw CL Rake</td>
<td>Raw CL Rake</td>
</tr>
<tr>
<td>Degree</td>
<td>0.0650* 0.0668** 0.0663**</td>
<td>0.0105 0.0183 0.0164</td>
</tr>
<tr>
<td></td>
<td>(0.0335) (0.0250) (0.0248)</td>
<td>(0.0369) (0.0162) (0.0166)</td>
</tr>
<tr>
<td>Cluster</td>
<td>-0.0300 -0.0300 -0.1016</td>
<td>1.2212*** 1.2212*** 1.2497***</td>
</tr>
<tr>
<td></td>
<td>(0.3703) (0.3703) (0.3528)</td>
<td>(0.4606) (0.4606) (0.4288)</td>
</tr>
<tr>
<td>Span</td>
<td>-0.0007 -0.2967** -0.3536***</td>
<td>0.0047*** -0.1605* -0.1804*</td>
</tr>
<tr>
<td></td>
<td>(0.0012) (0.1129) (0.1242)</td>
<td>(0.0010) (0.0890) (0.0996)</td>
</tr>
<tr>
<td>Epid Thrlid</td>
<td>-2.2472 -11.1305*** -12.9482***</td>
<td>1.3209 -4.5529 -4.7221</td>
</tr>
<tr>
<td></td>
<td>(1.3909) (3.7545) (4.0066)</td>
<td>(1.5398) (2.7815) (3.2972)</td>
</tr>
<tr>
<td>HI-grade</td>
<td>2.4308** 2.4308** 2.4729**</td>
<td>1.2698 1.2698 0.9994</td>
</tr>
<tr>
<td></td>
<td>(1.0010) (1.0010) (1.0447)</td>
<td>(1.2276) (1.2276) (1.2453)</td>
</tr>
<tr>
<td>HI-race</td>
<td>-0.8426** -0.8426** 0.0113</td>
<td>-1.2315* -1.2315* -0.7831</td>
</tr>
<tr>
<td></td>
<td>(0.4130) (0.4130) (0.5666)</td>
<td>(0.6568) (0.6568) (0.6944)</td>
</tr>
<tr>
<td>HI-pool</td>
<td>0.1294 0.1294 1.1938*</td>
<td>-0.7704 -0.7704 -0.2443</td>
</tr>
<tr>
<td></td>
<td>(0.7868) (0.7868) (0.6921)</td>
<td>(0.7865) (0.7865) (0.7300)</td>
</tr>
</tbody>
</table>

Note: Regression is based on 48 schools. Each regression controls for school size; *, **, *** stand for significance at 10%, 5%, and 1% respectively.

In contrast to the previous section, this application reveals that the overall network structure matters more for club participation than suggested by the raw data. With the exception of race homophily that is rejected as a relevant predictor under our corrections but accepted under random ones, networks corrected for non-representativeness deliver either similar or higher estimates than if only scaling is taken into account. The results suggest that participation in clubs is enhanced by higher average connectivity, lower distances, lower epidemic thresholds, and higher homophily. All these features describe higher network integration and more social capital. Nevertheless, clustering—considered one of the main features of high social cohesion—does not explain our dependent variable.
Once again, there is no general tendency in the biases and in how significance levels are affected. We observe attenuation, expansion and sign switching, and the (non-)significance can persist, appear or disappear under our corrections. If the raw sampled networks are used, the biases range from roughly 2% for average degree and homophily on grades to 7500% in case of race homophily; they lie between 50-100% in most regressions. The sign of the effect switches in one case. Concerning the significance, network distances (measured by graph span) and the epidemic threshold show no effects on club participation using the raw networks but reveal significant associations in the intuitive direction if we correct for sampling biases.

As for the comparison between our approach and the missing-at-random assumption, the biases are considerably smaller but still economically significant. The effect of homophily and clustering is the same under raw data and random corrections because these statistics are theoretically unbiased in representative samples. Therefore, the biases are still 70% for clustering and between 2% and 7500% for the homophily indices. Regarding the significance, controlling for non-representativeness of the sample may again confirm or not the estimates based on randomness.

References


