NBER WORKING PAPER SERIES

NON-RANDOMLY SAMPLED NETWORKS: BIASES AND CORRECTIONS

Chih-Sheng Hsieh
Stanley I. M. Ko
Jaromír Kovář
Trevon Logan

Working Paper 25270
http://www.nber.org/papers/w25270

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
November 2018, Revised July 2022

We are grateful to Isaiah Andrews, Aureo de Paula, Marco van der Leij, and participants at numerous seminars for comments and suggestions. Jaromír Kovář acknowledges financial support from Ministerio de Economía y Competividad and Fondo Europeo de Desarrollo Regional (PID2019-108718GB-I00, PID2019-106146GB-I00), the Basque Government (IT1461-22), and the Grant Agency of the Czech Republic (21-22796S). The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Non-Randomly Sampled Networks: Biases and Corrections
Chih-Sheng Hsieh, Stanley I. M. Ko, Jaromír Ková ík, and Trevon Logan
NBER Working Paper No. 25270
November 2018, Revised July 2022
JEL No. C4,D85,L14,Z13

ABSTRACT

This paper analyzes statistical issues arising from networks based on non-representative samples of the population. We first characterize the biases in both network statistics and estimates of network effects under non-random sampling analytically and numerically. Sampled network data systematically bias the properties of population networks and suffer from non-classical measurement-error problems when applied as regressors. Apart from the sampling rate and the elicitation procedure, these biases depend in a nontrivial way on which subpopulations are missing with higher probability. We 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 the estimated network effects. The advantages of the proposed methodology are that it can be applied to network data collected via both designed and non-designed sampling procedures, does not require one to assume any network formation model, and is straightforward to implement. We apply our approach to two widely used network data sets and show that accounting for the non-representativeness of the sample dramatically changes the results of regression analysis.

Chih-Sheng Hsieh
Chinese University of Hong Kong
Department of Economics
Room 911 Esther Lee Building
Hong Kong
cshsieh@cuhk.edu.hk

Jaromír Ková ík
Departamento Fundamentos
Análisis Económico I
Universidad del País Vasco
Av. Lehendakari Aguirre 83
Spain
jaromir.kovarik@ehu.eus

Stanley I. M. Ko
Department of Finance and
Business Economics
Faculty of Business Administration
Room 4003, E22
Macau
stanleyko@umac.mo

Trevon Logan
The Ohio State University
410 Arps Hall
1945 N. High Street
Columbus, OH 43210
and NBER
logan.155@osu.edu

1 Motivation

There is growing interest in understanding the role of networks in Economics.[1] Different "micro" and "macro" features of network architecture shape diffusion, learning, behavior, and other substantive phenomena in a variety of contexts.[2] Due to the increasing availability of large network data sets and increasing computational power, empirical network research is now a dynamic part of this literature. At the same time, empirical network analysis generates new econometric challenges (Fortin and Boucher, 2015; De Paula, 2017; Jackson et al., 2017). This paper investigates issues arising with *non-randomly sampled* network data, the most common network data used.

The vast majority of empirical network studies analyze sampled data, and the sampling rates are typically low.[3] Even though the literature across several disciplines has noted that using sampled data may lead to considerable biases and other statistical issues (see below for references), the typical approach is to treat the sampled data "as if" it were complete.[4] Chandrasekhar and Lewis (2016) show formally that, even in absence of other econometric issues and even if the nodes are selected randomly, the statistics of sampled networks differ systematically from the population network features and the inference based on sampled network data leads to measurement errors and inconsistency problems when estimating network effects. The estimates from sampled networks may suffer from attenuation, but also expansion or even sign switching. As a result, one cannot rely on solutions to classical measurement-error problems to correct these issues even if the sample is representative.

Furthermore, samples in network studies are typically non-representative. Apart from problems inherent in sampling, non-representativeness of network data may be caused by the sampling strategy itself (Frank, 1981; Kossinets, 2006; Kolaczyk, 2009; Handcock and Gile, 2010). For instance, snowball sampling, one prominent sampling method in applied work, is prone to finding nodes with higher connectivity than nodes with a 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. Other issues arise with "truncated" network data due to either the specification of network boundary or fixed-choice design (e.g., fixed-number friendship nomination). Last, several studies exploit stratified samples to better approximate the missing-at-random assumption. Unfortunately, it is difficult and costly to stratify on all relevant characteristics. Many issues with collection of network data might generate, and often do, samples in which the observed network structure is endogenous to the missing data mechanism.

To intuitively explain the issues arising from non-random sampling, we informally decompose the problem into two effects, *scaling* and *non-representativeness*. *Scaling* refers to

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

[2] Jackson (2010a) provides a survey of economic applications. Jackson et al. (2017) develop a taxonomy of macro, aggregate, or global network properties as opposed to micro, local, or individual ones, which we adopt here. They consequently review the influence of these characteristics at both levels in different socio-economic settings.

[3] The main reasons behind the common use of partial samples are that sampling the whole population is typically infeasible and network elicitation is more costly than collection of basic individual characteristics (Aral, 2016; Breza et al., 2020). Chandrasekhar and Lewis (2016) report that the median sampling rate in applied work in Economics is 25% and more than 66% of network studies have a sampling rate lower than 51%. Similar rates are found in other fields.

[4] Early exceptions include Ammermueller and Pischke (2009), Conley and Udry (2010), and Conti et al. (2013).

observing fewer nodes and edges 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 sample average degree in a regression therefore inflates the estimated impact of average degree on diffusion under random missing. This is an example of expansion of the estimated effect and thus non-classical measurement error. However, if nodes are not missing at random, whether the observed average degree and the estimates are inflated or attenuated will depend on who is missing. For example, if less connected nodes are missing with higher probability (a problem inherent in edge-based or 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 average degree, the homophily index and clustering coefficient can be unbiased in representative samples. Nevertheless, in samples in which different types of nodes are missing with different probabilities, homophily will be biased by definition. Since clustering is typically associated with connectivity in social networks (Jackson and Rogers, 2007), it is also likely to be mismeasured. The magnitude and direction of the biases in these characteristics and in their estimated effects in regressions again depend crucially and non-trivially on who is missing.

This study provides a systematic analysis of the problems arising from examining non-randomly sampled network data[5] elicited via two widely employed sampling methods and proposes a solution that (i) allows researchers to recover the true population network features (such as the average degree, global clustering, or homophily of a network) and (ii) mitigates biases in regressions testing the impact of these network features on either individual or group-level behaviors and outcomes.[6] We first characterize analytically and numerically the extent of biases under non-random missing both in the structural properties of observed networks and in estimates from a regression analysis, employing raw sampled networks as well as the existing corrections based on the missing-at-random assumption. Second, we propose a set of analytical corrections which allow us to recover the true values of several network characteristics widely used in applications. Third, we test the ability of our approach to mitigate these biases *vis-à-vis* uncorrected sample estimates, and random corrections. Last, the proposed methodology is applied to the National Longitudinal Study of Adolescent to Adult Health (hereafter Add Health) and a more recent stratified data set on microfinance take-up in a number of Indian villages (Banerjee et al., 2013). These data sets are particularly suited for our approach because they contain a relatively large number of networks, are widely employed in empirical work,[7] and a node and link have different meanings in each setting, illustrating the broad applicability of our approach. We document that these two datasets are indeed non-random

[5]Formal definition of (non-)randomness can be found in Section 2.1.

[6]Our study also improves inferences in network-formation applications studying contextual determinants of the network architecture (i.e. applying different network properties as regressands). Since network formation represents a key topic in the network literature (see Jackson (2005) for a review), it enlarges the applicability of the proposed methodology. However, since mismeasured dependent variables are less problematic as they only affect the estimated errors, this study focuses on regressions including network properties as regressors.

[7]See e.g., Moody (2001); Echenique and Fryer (2007); Bramoullé 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) and De Paula et al. (2018) for the latter.

samples of the population under scrutiny.[8] We argue that the existing approaches to sampled network data do not eliminate the statistical problems arising from non-random sampling and show how failure to account for non-representativeness of the sample causes problems with inference about the size and even the direction of network effects.

This study shows that relying on the missing-at-random assumption to adjust the raw network sample, which is rarely satisfied empirically, may be as serious as applying raw network data. Since the direction and magnitude of the biases depend on who is missing, we demonstrate the necessity of accounting for potential different missing rates of different segments of the population in applied work. This is particularly important in network data as population and distributional parameters are of main interest.

As the main contribution of this paper, we propose analytical corrections for a set of network characteristics widely used in applications: average degree, degree distribution, clustering coefficient, graph span, epidemic threshold, bounds on the maximal eigenvalue of the adjacency matrix of a network, and homophily. These network features represent fundamental aspects of network architecture commonly employed in theoretical and empirical research and provide intuitive insights regarding the way social organization shapes individual and group-level phenomena (Jackson et al., 2017). To that aim, we take explicit account of the missing rates of different sub-populations and 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 response rates among different demographic or socioeconomic categories and thus improve the precision of sample estimates for objective variables of interest (Holt and Smith, 1979; Little, 1993; Valliant, 1993). Intuitively, we assume that the population can be divided into a finite number of disjoint types or groups and that sampling (or conversely missing) rates differ across types.[9] The main difference between the standard post-stratification and our approach is to weight on network links, triples, or triangles (or subgraphs in the terminology of Chandrasekhar and Jackson, 2016), rather than on individuals. Since the proposed corrections are asymptotically unbiased, regressing economic outcomes on the corrected network measures, or using the corrected network features as dependent variables, delivers consistent estimates under standard assumptions. Moreover, our methodology nests the existing corrections designed for random sampling as a special case (e.g., Frank, 1980, 1981, Kolaczyk, 2009, Zhang et al., 2015, Chandrasekhar and Lewis, 2016). Our methodology outperforms approaches based on the missing-at-random assumption in terms of both the average biases of the proposed corrections and their mean squared errors if the sample is non-representative, making our methodology more broadly applicable.

Our applications corroborate that one cannot easily predict the direction and magnitude of these biases, and that they are substantively significant. The Indian village networks stratified on religion and geographical location are non-representative in terms of age and gender, while senior and non-white students are overrepresented in the Add Health data. We report that not accounting for unequal missing rates of different segments of the population affects the estimated network effects significantly in either data set. In a battery of regressions, we

[8]The Add Health networks were collected with the truncated fixed-choice design, introducing further biases (Griffith, 2022). In this paper, we focus on the issues generated by the non-representativeness.

[9]These types or groups are thought to represent, say, gender, race, ethnicity, location, age, education, etc. or their combinations (say, "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 *cross* throughout this paper.

frequently observe attenuation and false negative findings, but expansion, sign-switching, and false positives are also commonplace. Moreover, the biases are economically important; in many instances, the network effects are over/underestimated by roughly 100% or more.

The present paper connects to three literatures. First, we contribute to the literature on missing social network data. Numerous studies across fields drew attention to the issues arising with sampled networks and how particular sampling methods affect observed networks, some of which provide partial solutions to different issues (Frank, 1980, 1981; Stork and Richards, 1992; Stumpf et al., 2005; Kossinets, 2006; Huisman, 2009; Handcock and Gile, 2010). Within this literature, Zhang et al. (2015) build on Frank (1980, 1981) and propose an estimation procedure to recover the true degree distribution from sampled data if all randomness comes from the sampling method itself. We generalize their methodology to other potential sources of non-randomness. Our approach has certain parallelism with respondent driven-sampling, a methodology that combines snowball sampling with a model that weights the sample to compensate for the non-representativeness of the sample (Heckathorn, 1997), as well as statistical sampling theory that has developed procedures for how to recover true population networks if the only source of randomness is the sampling design (see Kolaczyk (2009) for a survey). Last, Thirkettle (2019) propose a procedure for the estimation of bounds on network statistics from sampled networks. Our weighting method has the same goals but differs substantially from these approaches in the underlying assumptions and in applicability: the former approaches are only applicable under specific sampling designs, whereas our approach can be applied both under designed sampling that depends on the network structure (such as snowball sampling) and also in cases in which networks are not elicited via designed sampling and the non-representativeness is caused by reasons orthogonal to the network structure, such as non-response or the existence of hard-to-reach subpopulations (as the case in e.g., the stratified sample in Banerjee et al. (2013)). Most importantly, existing approaches assume certain forms of representativeness in the sampling process *ex ante*, while our proposed methodology exploits the *ex-post* non-representativeness of the sample.

Second, our methodology complements emerging econometric literature on imperfectly measured network data and the estimation of network effects. Chandrasekhar and Lewis (2016) propose an integral methodology of dealing with *randomly* sampled networks. Similar to this paper, they show that estimations with sampled networks suffer from non-classical measurement error and propose a method to ensure consistent estimates. Their methodology consists of two alternative approaches. First, they provide formal corrections for the average degree, clustering coefficient, and graph span under an assumption of random sampling. Our approach generalizes this first strategy. As a second approach, they propose a graphical reconstruction technique that delivers consistent estimates in both network-level and individual-level regressions.[10] The procedure is first to estimate a network formation model, and then employ the estimated model to interpolate over missing parts of the network. As mentioned in their paper, the network reconstruction approach requires a correct model specification and certain assumptions to ensure consistency of the network statistics. This second approach does not necessarily recover the network properties, however. Most importantly for the present work, both approaches are based on a missing-at-random assumption. Breza et al. (2020) propose a two-stage strategy for network elicitation using responses to questions such as "How many of your social connections have trait k?" that permits the estimation of both node- or graph-

[10]This includes the estimation of network effects using instrumentation techniques proposed by Bramoullé et al. (2009) and De Giorgi et al. (2010). Liu (2013) shows that the solution in Chandrasekhar and Lewis (2016) may still suffer from weak-instrument problems.

level network properties. Chandrasekhar and Jackson (2016) propose a network formation model similar in the spirit to our recovery methodology in that it is also based on subgraphs in function of types of the nodes. The advantage of our approach, as opposed to the graphical reconstruction in Chandrasekhar and Lewis (2016) and the approaches in Breza et al. (2020) and Chandrasekhar and Jackson (2016), is that our methodology does not rely on any particularly assumed network formation model. This is crucial because when the model is fitted on non-representative network samples, the estimated network-formation parameters in the first stage will likely be biased and potentially inconsistent even if the assumed model is correct. As a result, none of these approaches can effectively recover the true network formation process from non-representative samples. Our methodology overcomes this issue as well as additional statistical problems arising from assuming an incorrect network formation model and introducing additional uncertainty via the two-stage procedures in Chandrasekhar and Lewis (2016), Breza et al. (2020), and Chandrasekhar and Jackson (2016). More recently, De Paula et al. (2018) propose a methodology allowing estimates of the entire network structure from panel data simultaneously with peer effects if either no or partial information on networks is available. The latter is restricted to panel data containing a large enough number of time series observations, a very strong requirement in many applications. We are interested in recovering the true network of interest from partial network data and how global network properties shape individual and group-level behaviors and outcomes, a type of network effects on which their approach does not apply. Boucher and Houndetoungan (2020) study the estimation of peer effects when the researchers only observe a consistent estimates of aggregate network statistics. Hence, our methodology and their approach naturally complement each other in non-representative samples since our methodology allows one to estimate unbiased statistics in non-representative samples. Our work complements and expands the above studies by providing the first step toward the statistical treatment of network data coming from non-representative samples of the population, which is the most common type of network data available.

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

2 Framework

2.1 Notation

A population network (graph) is a pair $G = (V, E)$ which consists of the sets of nodes V and edges E. Denote $n = |V|$ the cardinality of V. The network is represented with an $n \times n$ adjacency matrix $W(G)$. We follow the theoretical and empirical literature and focus on undirected and unweighted networks, i.e., $W_{ij} = 1(0)$ if i and j are (not) connected and $W_{ij} = W_{ji}$ for each $i, j \in V$. However, most of our analysis extends for directed and weighted

graphs. Following convention, we set $W_{ii} = 0$. 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 $\sum_{t=1}^{T} n_t = n$. 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 population network, researchers observe only the sample network. Let $S \subset V$ be the set of sampled nodes of size $m = |S| = \psi n$, where $\psi = \frac{m}{n}$ is the sampling rate. Analogously, S_t denotes the set of nodes of type t in the sample and $m_t = \psi_t n_t$ is the number of sampled individuals of type t and ψ_t is type t's sampling rate. We assume that $\psi_t > 0$ for each t throughout.

This paper focuses on two types of network sampling methods.[11] The first is the *induced subgraph*, denoted by $G^{|s}$. In the induced subgraph, the network links are only observable among the m sampled nodes. The second is the *star subgraph*, denoted by G^s. In the star subgraph, we observe not only the network links among m sampled nodes but also the links of the m sampled nodes to unsampled nodes in V. That is, $G^{|s} = (S, E^{|S})$ and $G^s = (V, E^S)$ 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.

We concentrate on several network properties, and denote a generic network property as $w(G)$. It can represent a scalar, vector, or even the whole adjacency matrix itself (i.e., $w(G) = W(G)$). The dimension of $w(G)$ will depend on the application and will be defined in each context. Let $w(\overline{G})$, $\overline{G} \in \{G^s, G^{|s}\}$, be the corresponding network statistic using the sample network $\overline{G}$ and $\widetilde{w}(\overline{G})$ the corrected network statistic in question proposed to mitigate the sample biases with respect to the population. For example, $w(G) = \frac{1}{n}\sum_{i \in V}\sum_{j \in V} W_{ij}$ is the average degree of a graph, which we denote $d(G)$ below. Hence, $d(\overline{G})$ is the average degree of the sample network and $\widetilde{d}(\overline{G})$ 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 sample is representative. Our framework allows for $\psi_t \neq \psi_s$ for any $t, s \in T$, which nests the random sampling case.

In applications, researchers may observe multiple networks. If a measure refers to a specific network, we a generic term $r \in \{1, 2, ..., R\}$. That is, G_r is the graph of population r, $\overline{G}_r \in \{G_r^s, G_r^{|s}\}$ 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 its corresponding number in the sample. Once again, $\psi_r = \frac{m_r}{n_r}$ and $\psi_{r,t} = \frac{m_{r,t}}{n_{r,t}}$.

2.2 Econometric Modeling

In addition to the reconstruction of network properties of interest, we also consider 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 researchers are interested in understanding whether and how the global properties of a network (e.g., average degree or the average distance in the network) influence a particular outcome. Formally,

$$y_r = \alpha + w(G_r)\beta + x_r\gamma + \epsilon_r, \tag{1}$$

[11] Section 6 discusses other sampling schemes and their relation with the proposed methodology.

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 α, β, and γ. Examples of the 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, 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 regional research networks (e.g., Fleming et al., 2007). Such regressions are also of interest theoretically. For example, the overall clustering of a network may explain the magnitude and efficiency of risk-sharing within a society (Bloch et al., 2008), and the stability of behavior in a society may be related to the minimal eigenvalue of the adjacency matrix (Bramoullé et al., 2014).

The proposed approach also applies to models studying whether and how the overall properties of a network affect outcomes at the individual level: $y_{ir} = \alpha + w(G_r)\beta + x_{ir}\gamma + \lambda_r + \epsilon_{ir}$, with y_{ir} is the outcome of an individual i in network r, x_{ir} captures her individual heterogeneity (that can include the heterogeneity of i's neighborhood), and λ_r are network random effects. For instance, the decision of an individual to adopt a product (e.g., microfinance as in Banerjee et al., 2013), participate in an activity (e.g., recreational activity as in Bramoullé et al., 2009), or behave in a particular way (Centola, 2010) can depend on the overall structure of the network. In the same vein, the innovation literature studies how the structure of regional networks shapes innovative performance of individual innovators (Schilling and Phelps, 2007; Whittington et al., 2009; Galaso and Kovářík, 2021). 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; Bramoullé et al., 2014; Ruiz-Palazuelos, 2021).

With sampled data on the network, the researchers observe $\overline{G} \in \{G^S, G^{|S}\}$ which is mismeasured, different from G. Therefore, scholars typically estimate

$$y_r = \alpha + w(\overline{G}_r)\beta + x_r\gamma + \epsilon_r, \tag{2}$$

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 case of network measures (see e.g., Wooldridge, 2015, and Hyslop and Imbens, 2001).[12] Chandrasekhar and Lewis (2016) 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 become even more problematic if the missing-at-random assumption is violated. Moreover, if multiple regressors are included in the model as in (2), the estimates of independent variables measured without error also become biased when others are mismeasured. That is, even if network properties only serve as controls, they may bias the estimates of the main variables of interest.

In Sections 3 and 4, we show analytically and numerically how the biases depend on who is missing in the sample. Section 5 further illustrates the applications of our correction methods on two widely used network data sets that come from non-representative samples of

[12] While there are other challenges typically present in network regressions, such as endogeneity and/or omitted variable problems, we argue that the sampling issue presents even in the absence of these problems.

the population under study.

3 Analytic corrections for sample network measures

This section shows formally the biases arising from sampling in case of some commonly used network measures and proposes how to correct them using the post-stratification weighting approach.[13]

Our approach naturally shares the assumptions and data requirements of standard (non-network) post-stratification weighting and raises the same issues.[14] Since issues such as the optimal construction of post-stratification bins or raking strategies have been largely discussed in the post-stratification literature (Holt and Smith, 1979; Little, 1993; Valliant, 1993), we do not analyze them here. In Section 6 we further provide an algorithm designed to select the most effective weighting variables. Another issue is the existence of the statistics discussed in this section. For example, the graph span might not exist if the sampling rate is so low that second-order neighbors of the sampled nodes are not observed. In such a case, neither the (uncorrect) raw sample statistics, nor the corrections based on the random missing assumption, nor our proposed corrections exist. Hence, we stress that our corrections for network statistics are applicable under sampling rates in which the raw sample statistics exist.

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 a node's importance or centrality. It has been applied as a regressor in numerous empirical studies and contexts (see, e.g., Kremer and Miguel, 2007; Branas-Garza et al., 2010; Kovářík et al., 2012; Banerjee et al., 2013; Alatas et al., 2016, among many others).

For induced subgraphs, the sample average degree is defined as $d(G^{|s}) = \frac{1}{m_r}\sum_{i\in S_r}\sum_{j\in S_r} W^{|s}_{ij,r}$. In Supplementary Appendix A.1, we show that

$$\mathrm{E}(d(G_r^{|s})|G_r) = \frac{1}{n_r}\sum_{t=1}^{T}\left(\sum_{i\in V_{r,t}}\sum_{j\in V_{r,t}} W_{ij,r}\frac{\psi_{r,t}^2}{\psi_r}\right)+\frac{1}{n_r}\sum_{t=1}^{T}\sum_{\ell\neq t}^{T}\left(\sum_{i\in V_{r,t}}\sum_{j\in V_{r,\ell}} W_{ij,r}\left(\frac{\psi_{r,t}\psi_{r,\ell}}{\psi_r}\right)\right)+o(1). \tag{3}$$

[13] In standard post-stratification weighting, 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 post-stratification weight assigned to a sampled individual i depends on which categories she belongs to. Formally, $p_i^t = \frac{n_t/n}{m_t/m}$. 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 different types of relationships (or subgraphs in the terminology of Chandrasekhar and Jackson, 2016).

[14] First, we assume that individual heterogeneity provides valuable information about the positioning and who is connected with whom. There exist large evidence that it is the case and we share this assumption with the network-formation estimation techniques discussed in Section 1. Second, post-stratification weighting–including our approach–requires the analyst to know the exact joint distributions of non-network characteristics such as sex, age, and race in the population and the sample, a data structure common in many applications (see e.g. the Add Health data, or the data in Banerjee et al. (2013), and Conley and Udry (2010) among others). If the joint distribution is not available but the marginal distributions are, one can still use the raking, logistic regression, or calibration to construct the weights.

The intuition behind (3) is the following: there are $\sum_{i\in V_r}\sum_{j\in V_r} W_{ij,r}$ edges in the population network G_r of interest, out of which $\sum_{i\in V_{r,t}}\sum_{j\in V_{r,t}} W_{ij,r}$ edges link two individuals of the same type $t \in \{1\cdots,T\}$, and $\sum_{i\in V_{r,t}}\sum_{j\in V_{r,\ell}} W_{ij,r}$ edges connect two individuals of different types, $t \neq \ell$. Given the sampling rate of each type, we only observe, for example, $\frac{\psi_{r,t}\psi_{r,\ell}}{\psi_r}\sum_{i\in V_{r,t}}\sum_{j\in V_{r,\ell}} W_{ij,r}$ of the cross-type edges in expectation.

Most importantly, as long as $\psi_r < 1$ for some r, the conditional expectation of $d(G_r^{|s})$ is not equal to $d(G_r)$ and the bias emerges due to scaling, even if the sample is representative. Moreover, as $\psi_{r,t}$ and $\psi_{r,\ell}$ are not necessarily the same as ψ_r, we have the second source of bias, non-representativeness, and the issue becomes more complicated.

To correct the biases from both scaling and non-representativeness, we follow the principle of Horvitz-Thompson (H-T, hereafter) estimator (Horvitz and Thompson, 1952) to propose the weighted sample average degree:

$$\widetilde{d}(G_r^{|s}) = \frac{1}{m_r}\sum_{t=1}^{T}\left(\sum_{i\in S_{r,t}}\sum_{j\in S_{r,t}} W_{ij,r}^{|s}\left(\frac{\psi_{r,t}^2}{\psi_r}\right)^{-1}\right) + \frac{1}{m_r}\sum_{t=1}^{T}\sum_{\ell\neq t}^{T}\left(\sum_{i\in S_{r,t}}\sum_{j\in S_{r,\ell}} W_{ij,r}^{|s}\left(\frac{\psi_{r,t}\psi_{r,\ell}}{\psi_r}\right)^{-1}\right). \tag{4}$$

In (4), we multiply each observed same-type sample link by a post-stratification weight $\left(\psi_{r,t}^2\right)^{-1}$ and each cross-type sample link by a weight $\left(\psi_{r,t}\psi_{r,\ell}\right)^{-1}$ and then all sample links by a weight ψ_r to further adjust the sample-population size ratio in calculating the average. These weights account for the missing rates of the links in the sample, respecting the number of such links in the observed part of the network. That is, the corrections respect the observed correlations in who is connected to whom (i.e. they respect network homophily).

Note the presence of little $o(1)$ term in (3). The proposed corrected $\widetilde{d}(G_r^{|s})$ is only unbiased if a network grows large; otherwise it is subject to an error due to the randomness in the sampling process. Under standard assumptions regarding the asymptotic variance of the estimator, we could additionally show that the correction (4) is not only asymptotically unbiased but also consistent. However, rather than making additional assumptions, and since samples are finite in applications, Section 4 complements this section with numerical analysis using standard population and sample sizes, assessing to which extent all the corrections proposed in this section and the estimates using them as regressors differ from their true values.

For star subgraphs, the sample average degree is defined as $d(G_r^s) = \frac{1}{m_r}\sum_{i\in S_r}\sum_{j\in S_r} W_{ij,r}^s$. In Supplementary Appendix A.1, we show that

$$\begin{aligned}\mathrm{E}(d(G_r^s)|G_r) = {} & \frac{1}{n_r}\sum_{t=1}^{T}\left(\sum_{i\in V_{r,t}}\sum_{j\in V_{r,t}} W_{ij,r}\left(\frac{\psi_{r,t}^2 + 2\psi_{r,t}(1-\psi_{r,t})}{\psi_r}\right)\right) \\ & + \frac{1}{n_r}\sum_{t=1}^{T}\sum_{\ell\neq t}^{T}\left(\sum_{i\in V_{r,t}}\sum_{j\in V_{r,\ell}} W_{ij,r}\left(\frac{\psi_{r,t}\psi_{r,\ell} + \psi_{r,\ell}(1-\psi_{r,t}) + \psi_{r,t}(1-\psi_{r,\ell})}{\psi_r}\right)\right) + o(1).\end{aligned} \tag{5}$$

Again, as long as $\psi_r \neq 1$ and the sampling rate across types are not the same, i.e., $\psi_{r,t} \neq \psi_{r,\ell}$, the conditional expectation of $d(G_r^s)$ in (5) does not equal to $d(G_r)$. To correct the bias, our

proposed weighted sample average degree takes the following form:

$$\widetilde{d}(G_r^s) = \frac{1}{m_r}\sum_{t=1}^{T}\left(\sum_{i\in S_{r,t}}\sum_{j\in S_{r,t}} W_{ij,r}^s \left(\frac{\left(\psi_{r,t}^2 + 2\psi_{r,t}(1-\psi_{r,t})\right)}{\psi_r}\right)^{-1}\right)$$
$$+ \frac{1}{m_r}\sum_{t=1}^{T}\sum_{\ell\neq t}^{T}\left(\sum_{i\in S_{r,t}}\sum_{j\in S_{r,\ell}} W_{ij,r}^s \left(\frac{\left(\psi_{r,t}\psi_{r,\ell} + \psi_{r,\ell}(1-\psi_{r,t}) + \psi_{r,t}(1-\psi_{r,\ell})\right)}{\psi_r}\right)^{-1}\right). \quad (6)$$

Note that if $\psi_{r,t} = \psi_{r,\ell} = \psi_r$ (i.e. under random sampling), (3) collapses to $\mathrm{E}[d(G_r^{|s})|G_r] = (\psi_r + o(1))d(G_r)$ and (5) collapses to $\mathrm{E}[d(G_r^s)|G_r] = [1-(1-\psi_r)^2 + o(1)]\, d(G_r)$, which would be exactly the same as shown by Chandrasekhar and Lewis (2016). The key difference between the correction approach proposed in Chandrasekhar and Lewis (2016) and the weighted sample average degrees in (4) or (6) is that the corrections based on the missing-at-random assumption are only a rescaling of the corresponding sample average degree, while matters become more complex if sampling differs across types and the network is homophilous or heterophilous. In such a case, one has to rescale on links and different links have to be weighted differently to yield an asymptotically unbiased correction. If one applies the random corrections to non-representative data, biases emerge and there is no reason for these biases to be smaller than in the raw (uncorrected) data.

3.2 Clustering coefficient

The global clustering coefficient of a graph is the ratio between the number of triangles and the number of connected triples in the network,[15] calculated as $c(G_r) = \frac{\rho(G_r)}{\tau(G_r)}$, where

$$\rho(G_r) = \sum_{\substack{i\in V_r}}\sum_{\substack{j\in V_r\\ i\neq j}}\sum_{\substack{k\in V_r\\ i\neq j\neq k}} W_{ij,r}W_{jk,r}W_{ki,r} \quad \text{and} \quad \tau(G_r) = \sum_{\substack{i\in V_r}}\sum_{\substack{j\in V_r\\ i\neq j}}\sum_{\substack{k\in V_r\\ i\neq j\neq k}} W_{ij,r}W_{jk,r}.$$

The clustering coefficient has traditionally been considered a measure of social capital. For example, it plays an important role in risk-sharing (Bloch et al., 2008), trust building (Karlan et al., 2009), enhancing cooperation (Granovetter, 1985), and labor markets (Espinosa et al., 2021). Several empirical studies have used the clustering coefficient as a regressor or a dependent variable (e.g., Fleming et al., 2007; Kovářík and Van der Leij, 2014; Alatas et al., 2016; Kovářík et al., 2017).

Similarly to average degree, we show in Supplementary Appendix A.2 that the sample clustering coefficients $c(G_r^{\overline{s}}) = \frac{\rho(G_r^{\overline{s}})}{\tau(G_r^{\overline{s}})}$; $\overline{s} \in \{|s, s\}$ are biased in non-representative samples and that the biases can be corrected via the H-T estimator with the weights summarized in Table 1. However, since the clustering coefficient is measured using triangles and connected triples, the corrections in Table 1 differ from the average degree in that, instead of dyads, we correct "relationships" of three individuals depending on how they are interconnected (triangle or connected triple) and their type composition.

More precisely, for induced subgraphs, we propose to multiply each triangle and connected triple composed of three individuals of the same type t by the weight $\psi_{r,t}^{-3}$, those with two individuals of type t and one of type $\ell \neq t$ by $(\psi_{r,t}^2\psi_{r,\ell})^{-1}$, and those with three individuals of three

[15] 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.

different types by $(\psi_{r,t}\psi_{r,\ell}\psi_{r,h})^{-1}$. Similarly, we can obtain an asymptoticly unbiased estimator of the clustering coefficient in star subgraphs by multiplying all triangles and connected triples by different weights from Table 1 depending on their type composition. Then, the proposed corrections are $\widetilde{c}(G_r^{\overline{s}}) = \widetilde{\rho}(G_r^{\overline{s}})/\widetilde{\tau}(G_r^{\overline{s}})$ with $\overline{s} \in \{|s, s\}$. See Supplementary Appendix A.2 for the exact expressions and other details.

3.3 Epidemic threshold

There has been an increasing interest in understanding the diffusion properties of networks. The applications range from diffusion of innovation (e.g., Valente, 1996; 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 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 the propagation. There are 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 version which is widely used and based on the mean-field approximation:

$$Thrld_r = \frac{\frac{1}{n_r}\sum_{i\in V_r}\sum_{j\in V_r} W_{ij,r}}{\frac{1}{n_r}\sum_{i\in V_r}(\sum_{j\in V_r} W_{ij,r})^2}.$$

The threshold is simply the ratio between the average degree, $d(G_r)$, and the average squared degree, denoted by $ds(G_r)$. The corrections of $d(G_r)$ has been treated in Section 3.1. In Supplementary Appendix A.3, we show that

$$ds(G_r) = d(G_r) + \frac{1}{n_r}\sum_{i\in V_r}\sum_{j\in V_r}\sum_{\substack{k\in V_r \\ k\neq j}} W_{ij,r}W_{ik,r}, \tag{7}$$

where the second term at the right hand side of (7) involves calculation of the number of two-stars stemming from i.

Again, the correcting weights for the sample number of two-stars are summarized in Table 1 for both elicitation strategies.[16] The remaining steps of the correction procedure are analogous to previous sections and outlined in detail in Supplementary Appendix A.3.

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 finite paths. Naturally, the shorter the distance between nodes, the easier it is for them to communicate, transmit information, or influence each other. Shorter average distances consequently allow for easier transmission throughout the whole population. Therefore, distances play important role in diffusion (similarly to the epidemic threshold and spectral properties), but also in risk-sharing or flow of capital among others. Several

[16] Although triangles, triples, and two-stars are somewhat different objects, their correcting weights are identical under induced-subgraph elicitation. The reason is that the named people are sampled by construction. In contrast, the weights differ across the three network objects in star subgraphs, where some nodes can be named by others as friends without being sampled.

Table 1: The correction weights for three-node subnetworks (triangles, triples, and two-stars) for induced (top) and star (bottom) subgraphs and in function of the type composition of the subnetwork.

Subnetwork	Type composition of individuals i, j, k	Correction weights
Induced subgraph ($G_r^{\vert s}$)		
$W_{ij,r}^{\vert s} W_{jk,r}^{\vert s} W_{ki,r}^{\vert s}$	$t_i = t_j = t_k = t$	$\psi_{r,t}^{-3}$
$W_{ij,r}^{\vert s} W_{jk,r}^{\vert s}$	$t_i = t_j = t, t_k = \ell$ $t_j = t_k = t, t_i = \ell$ $t_i = t_k = t, t_j = \ell$	$\left(\psi_{r,t}^2 \psi_{r,\ell}\right)^{-1}$
$W_{ij,r}^{\vert s} W_{ik,r}^{\vert s}$	$t_i = t, t_j = \ell, t_k = h$	$\left(\psi_{r,t} \psi_{r,\ell} \psi_{r,h}\right)^{-1}$
Star subgraph (G_r^s)		
$W_{ij,r}^s W_{jk,r}^s W_{ki,r}^s$	$t_i = t_j = t_k = t$	$\left(\psi_{r,t}^3 + 3\psi_{r,t}^2(1-\psi_{r,t})\right)^{-1}$
	$t_i = t_j = t, t_k = \ell$ $t_j = t_k = t, t_i = \ell$ $t_i = t_k = t, t_j = \ell$	$\left(\psi_{r,t}^2 \psi_{r,\ell} + 2(1-\psi_{r,t})\psi_{r,t}\psi_{r,\ell} + \psi_{r,t}^2(1-\psi_{r,\ell})\right)^{-1}$
	$t_i = t, t_j = \ell, t_k = h$	$\left(\psi_{r,t}\psi_{r,\ell}\psi_{r,h} + (1-\psi_{r,t})\psi_{r,\ell}\psi_{r,h} + \psi_{r,t}(1-\psi_{r,\ell})\psi_{r,h} + \psi_{r,t}\psi_{r,\ell}(1-\psi_{r,h})\right)^{-1}$
$W_{ij,r}^s W_{jk,r}^s$	$t_i = t_j = t_k = t$	$\left(\psi_{r,t}^3 + 3\psi_{r,t}^2(1-\psi_{r,t}) + \psi_{r,t}(1-\psi_{r,t})^2\right)^{-1}$
	$t_i = t_j = t, t_k = \ell$ $t_j = t_k = t, t_i = \ell$	$\left(\psi_{r,t}^2 \psi_{r,\ell} + 2(1-\psi_{r,t})\psi_{r,t}\psi_{r,\ell} + \psi_{r,t}^2(1-\psi_{r,\ell})\right)^{-1}$
	$t_i = t_k = t, t_j = \ell$	$\left(\psi_{r,t}^2 \psi_{r,\ell} + 2(1-\psi_{r,t})\psi_{r,t}\psi_{r,\ell} + \psi_{r,t}^2(1-\psi_{r,\ell}) + (1-\psi_{r,t})^2\psi_{r,\ell}\right)^{-1}$
	$t_i = t, t_j = \ell, t_k = h$	$\Big(\psi_{r,t}\psi_{r,\ell}\psi_{r,h} + (1-\psi_{r,t})\psi_{r,\ell}\psi_{r,h} + \psi_{r,t}(1-\psi_{r,\ell})\psi_{r,h} + \psi_{r,t}\psi_{r,\ell}(1-\psi_{r,h}) + (1-\psi_{r,t})\psi_{r,\ell}(1-\psi_{r,h})\Big)^{-1}$
$W_{ij,r}^s W_{ik,r}^s$	$t_i = t_j = t_k = t$	$\left(\psi_{r,t}^3 + 3\psi_{r,t}^2(1-\psi_{r,t}) + \psi_{r,t}(1-\psi_{r,t})^2\right)^{-1}$
	$t_i = t_j = t, t_k = \ell$ $t_i = t_k = t, t_j = \ell$	$\left(\psi_{r,t}^2 \psi_{r,\ell} + 2(1-\psi_{r,t})\psi_{r,t}\psi_{r,\ell} + \psi_{r,t}^2(1-\psi_{r,\ell})\right)^{-1}$
	$t_j = t_k = t, t_i = \ell$	$\left(\psi_{r,t}^2 \psi_{r,\ell} + 2(1-\psi_{r,t})\psi_{r,t}\psi_{r,\ell} + \psi_{r,t}^2(1-\psi_{r,\ell}) + (1-\psi_{r,t})^2\psi_{r,\ell}\right)^{-1}$
	$t_i = t, t_j = \ell, t_k = h$	$\Big(\psi_{r,t}\psi_{r,\ell}\psi_{r,h} + (1-\psi_{r,t})\psi_{r,\ell}\psi_{r,h} + \psi_{r,t}(1-\psi_{r,\ell})\psi_{r,h} + \psi_{r,t}\psi_{r,\ell}(1-\psi_{r,h}) + \psi_{r,t}(1-\psi_{r,\ell})(1-\psi_{r,h})\Big)^{-1}$

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

$$\ell(G_r) = \frac{\log n_r - \log d(G_r)}{\log d_2(G_r) - \log d(G_r)} + 1,$$

where $d_2(G_r) = \frac{1}{n_r}\sum_{i\in V_r}\sum_{\substack{j\in V_r \\ i\neq j\neq k}}\sum_{k\in V_r} W_{ij,r}W_{jk,r}$ is the average number of second-order neighbors (that is, nodes at distance two or simply neighbors of neighbors).

In order to correct $d_2(G_r^{|s})$ and $d_2(G_r^s)$, Table 1 again shows how the sampled connected triples should be multiplied by different correction weights in function of the type composition and the network elicitation procedure. Finally, as illustrated in Supplementary Appendix A.4, the weights are additionally multiplied by ψ_r to adjust the sample-population size ratio in the averages. The final expressions for $\tilde{d}_2(G_r^{|s})$ and $\tilde{d}_2(G_r^s)$ can be found in Supplementary Appendix A.4.

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 with whom is typically correlated with characteristics such as gender, age, race, social and economic status, among others (see McPherson et al. (2001) for a survey). This phenomenon of "birds of a feather flock together" gains particular relevance in our approach, because we explicitly consider types in the population and network, respectively. Homophily is an important measure of cross-type segregation and affects many economically relevant phenomena such as diffusion or learning and their speeds (Golub and Jackson, 2012a,b), labor market outcomes (Calvo-Armengol and Jackson, 2004; Toomet et al., 2013), or individual and firm-level success (McPherson and Smith-Lovin, 1987; Ibarra, 1992).

We adapt the homophily index from Currarini et al. (2009). The homophily index within type 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 types different than t. Specifically,

$$s_{r,t}(G_r) = \frac{1}{n_{r,t}}\sum_{i\in V_{r,t}}\sum_{j\in V_{r,t}} W_{ij,r}, \quad d_{r,t}(G_r) = \frac{1}{n_{r,t}}\sum_{i\in V_{r,t}}\sum_{j\notin V_{r,t}} W_{ij,r}.$$

In the case of induced subgraphs, fixing type t, we propose the following weighted estimators

$$\tilde{s}_{r,t}^{|s} = \frac{1}{m_{r,t}}\sum_{i\in S_{r,t}}\sum_{j\in S_{r,t}} W_{ij,r}^{|s}\psi_{r,t}^{-1} \quad \text{and} \quad \tilde{d}_{r,t}^{|s} = \frac{1}{m_{r,t}}\sum_{\ell\neq t}^{T}\left(\sum_{i\in S_{r,t}}\sum_{j\in S_{r,\ell}} W_{ij,r}^{|s}\psi_{r,\ell}^{-1}\right).$$

Therefore, we propose to multiply $\psi_{r,t}^{-1}$ on each link for the calculation of $\tilde{s}_{r,t}^{|s}$ and $\psi_{r,\ell}^{-1}$ for $\tilde{d}_{r,t}^{|s}$.

In the case of star subgraphs, fixing type t, we propose the following weighted estimators

$$\tilde{s}^s_{r,t} = \frac{1}{m_{r,t}} \sum_{i \in S_{r,t}} \sum_{j \in S_{r,t}} W^s_{ij,r} \Big(2 - \psi_{r,t}\Big)^{-1} \tag{8}$$

and

$$\tilde{d}^s_{r,t} = \frac{1}{m_{r,t}} \sum_{\ell \neq t}^{T} \left(\sum_{i \in S_{r,t}} \sum_{j \in S_{r,\ell}} W^s_{ij,r} \Big(\psi_{r,\ell}/\psi_{r,t} + (1 - \psi_{r,\ell})\Big)^{-1} \right). \tag{9}$$

Therefore, we propose to multiply $(2 - \psi_{r,t})^{-1}$ to each link for the calculation of $\tilde{s}^s_{r,t}$ and multiply $(\psi_{r,t}/\psi_{r,\ell} + (1 - \psi_{r,\ell}))^{-1}$ to the links between type t and type ℓ in $\tilde{d}^s_{r,t}$.

The corrected homophily indexes for both elicitation procedures then are $\widetilde{H}_t(G^{\bar{s}}_r) = \frac{\tilde{s}^{\bar{s}}_{r,t}}{\tilde{s}^{\bar{s}}_{r,t} + \tilde{d}^{\bar{s}}_{r,t}}$. The derivations can be found in Supplementary Appendix A.5.

3.6 Degree distribution

Average degree is the first moment of the degree distribution of a graph. Nevertheless, other higher order moments and, in fact, the whole degree distribution are fundamental descriptors of a network structure with important consequences (Vega-Redondo, 2007; Jackson and Rogers, 2007; Jackson, 2010b). Both the first and the second moments of the degree distribution, as well as the tails of the distribution, are key for the understanding of diffusion properties of a graph (see e.g., Acemoglu et al. (2012) for an application in Economics; see Sections 3.3 and 3.7 for further discussion of diffusion properties of a network). The degree distribution also affects behavior in network games, as in Jackson and Yariv (2007) and Galeotti et al. (2010). Moreover, different moments may serve for the computation of bounds on spectral properties of a graph as illustrated in Section 3.7.

We show formally in Supplementary Appendix A.6 that our methodology allows to recover the entire degree distribution from non-randomly sampled network data. To that aim, we generalize the approaches of Frank (1980, 1981) and Zhang et al. (2015). However, since the derivations of the corrections and the description of the whole procedure are complex, they require several pages of text. We thus relegate all the details to Supplementary Appendix A.6.

One may wonder why we still propose the correction for the average degree in Section 3.1 if one can compute it from the corrected degree counts in this section. First, the recovery of the degree distribution relies on degree counts and estimating the degree distribution may be computationally costly with a large number of types, since researchers have to estimate the degree counts separately for different types or even type pairs. Such a concern is particularly relevant for the estimation strategy proposed by Zhang et al. (2015) outlined in Supplementary Appendix A.6. Second, the estimators differ in their accuracy. Estimating the average degree improves with the size of the sample while the estimates of counts depend on the sampling rate (e.g., Zhang et al., 2015). Hence, we still propose corrections for the average degree (Section 3.1) and the average degree squared (as part of the epidemic threshold in Section 3.3). Most importantly though, the fact that our methodology can also recover the entire degree distribution illustrates how powerful our approach is.

3.7 Largest eigenvalue of a network

Spectral properties of the adjacency matrix provide rich information about many topological properties of a network, including features of the degree distribution, component and community structure, and network distances, to name a few examples (see e.g., Faloutsos et al., 1999; Van Mieghem, 2010). They are particularly useful for modeling dynamic phenomena that take place on networks. For instance, the epidemic threshold from Section 3.3 is a good measure of how a network diffuses viruses, diseases, or behaviors if there are no degree correlations. If a network exhibits such correlations (and real-life networks typically do), the epidemic threshold is equal to the inverse of the largest eigenvalue of the adjacency matrix, $\lambda(G_r)$ (e.g., Boguñá et al., 2003). The largest eigenvalue also plays a role in network games (e.g., Ballester et al., 2006), public good provision (Elliott and Golub, 2019), or propagation of shocks in interbank or financial networks (Bardoscia et al., 2017). The remaining eigenvalues have also been shown to matter for dynamics, speed, and stability of learning and behavior (Golub and Jackson, 2010, 2012b; Bramoullé et al., 2014).

Since spectral properties depend on the whole network architecture, our approach that relies on nodes' local information cannot recover their exact values. Indeed, no existing approach can work without assumptions on the whole network. Nevertheless, to illustrate another application of our approach, we propose corrections for the bounds on $\lambda(G_r)$ from a sampled network. Lovász (2007) and Van Mieghem (2010) show that $d(G_r) \leq \sqrt{ds(G_r)} \leq \lambda(G_r) \leq U_r$, where $U_r = (2|E_r|(n_r - 1)/n_r)^{1/2}$. That is, the largest eigenvalue is bounded below by the average degree and the square root of the average squared degree and bounded above by an expression that depends on the number of nodes and edges in the graph. Notice that, even though we cannot recover the eigenvalue precisely, the bounds of the eigenvalue can be estimated using our approach.

We only focus on $\lambda(G_r)$ here, but large literature across disciplines has also proposed bounds for other eigenvalues (e.g., Das and Kumar, 2004; Walker, 2011) or the average betweenness centrality of a graph (Comellas and Gago, 2007). Hence, one could potentially employ the poststratification weighting to propose population-level bounds for other features of the network that cannot be recovered exactly based only on local information. Such a strategy enlarges the applicability of the methodology proposed in this study.

4 Monte Carlo Simulations

This section evaluates numerically the estimation biases in the network measures (in Section 4.1) and the network effects using these measures as regressors in regression (in Section 4.2), depending on the sample network type (induced vs. star subgraph), sampling rate, and (non-)randomness of the sample. We quantify the biases in raw sample data and in corrections based on the missing-at-random assumption, and compare their performance *vis-à-vis* our poststratification weighting. We concentrate on the scenarios that mimic our modeling assumptions. This provides a natural testing ground of our approach.

To illustrate the usefulness of our poststratification weighting approach, we focus on six network measures from Section 3 in this simulation exercise: average degree, global clustering coefficient, graph span, epidemic threshold, homopily, and bounds on the maximal eigenvalue. We do not include the degree distribution because it relies on a complex constrained penalized weighted least square approach from Zhang et al. (2015) and it is beyond the scope of this paper to demonstrate that detailed procedure. Moreover, one objective of this study is to

analyze how the proposed estimates perform in regression analysis. It is thus not clear how to incorporate the degree distribution into a regression.[17]

The population network data in our simulation study are adopted from the Add Health Wave-I In-school data.[18] In particular, we adopt one school as a prototype.[19] By adopting a real-life friendship network in school, we can preserve certain relationships between students' characteristics and the friendship network.[20] For example, white students have on average more network (friendship) connections than black students and the latter are on average more connected than other races in the data, and the patterns of homophily depend systematically on the race composition of each school (Currarini et al., 2009, 2010).

For ease of interpretation, we prune this prototype to the size of 1,500 so that the numbers of whites, blacks, and other races are all equal to 500 in each race group. We use three demographic characteristics from the prototype to define individual's type: seniority (C1), gender (C2), and race (C3). Seniority takes value of one if an individual is older than the population average and zero otherwise. For gender, one stands for male and zero for female. For race, one denotes White, two denotes Black, and three stands for other races. We also combine these three characteristics to form $2 \times 2 \times 3 = 12$ cross-characteristics, denoted by *Cross* throughout.[21] Seniority and gender are largely uncorrelated with individual network connectivity. In contrast, race is strongly correlated with network degree in the data. The average network degree for white students is 9.60, black students have an average network degree of 7.38, whereas the average connectivity is 4.39 for other races. Naturally, the three variables can partially predict who is connected to whom. In the following sections, homophily mostly refers to homophily with respect to variable *Cross*. The results are very similar in case of homophily on other variables.

4.1 Network measures

As a first step, we quantify the biases in network measures in sample networks using uncorrected estimators, corrections assuming representativeness, and our postsratification weighting approach. From the population described above, we generate 1,000 artificial sampled networks using different removal schemes, which vary in three dimensions. First, we apply *two* sampled network types, induced and star subgraphs. for 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 indi-

[17] The same issue applies to the bounds on the maximal eigenvalue. Should we use the lower or the upper bound, or the middle point between these two? We therefore skip the bounds on the maximal eigenvalue in the regression analysis.

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

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

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

[21] *Cross* combines the previous three variables such that the types are for example "senior black female," "junior male of other race," and so on.

viduals. Second, we consider *three* sampling rates, $\psi = 80\%$, 60%, and 40% (or alternatively three removal rates, $1 - \psi = 20\%$, 40%, and 60%, respectively). Third, we employ *four* removal scenarios with respect to the representativeness of the artificially sampled subgraphs. We either remove individuals 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). Since race is strongly correlated with network degree in the data, when we generate our artificial samples and would like non-random (disproportional) missing, we use 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.[22]

We perform 1,000 Monte Carlo repetitions. Figures 1 and 2 summarize the first set of results from our simulations for the induced subgraph and the star subgraph sampling, respectively. In Figures 1 and 2, the y-axes reflect the average magnitudes and signs of the biases (out of 1,000 simulation repetitions) in percentage terms with respect to the population values. The x-axes list the five network characteristics under scrutiny in the following order: average degree, global clustering, graph span, epidemic threshold, and homophily index of cross-characteristics. To save on space, we do not include the results for the bounds on the maximal eigenvalue in the main text; they can be found in Figures C.5 and C.6 in Supplementary Appendix C. The blue bars in Figures 1 and 2 represent the raw sample data and the red bars, denoted Random, reflect the corrections based on the missing-at-random assumption. The remaining three other colored bars are variations of our methodology. The green bars weight on the network-unrelated characteristic C1, whereas the last two bars represent, respectively, the weighting on C3 only (dark red) and the weighting on *cross* (i.e., the combination of C1 to C3; gray).[23] The rows and columns represent respectively the four different removal scenarios, scenarios R, H, M, L, and the three removal rates, $1 - \psi = 20\%$, 40%, and 60%, in these orders.

Biases in the uncorrected estimators. We first discuss the biases that arise in the considered measures if raw sampled data are used to compute the network statistics and stress the effect of non-representativeness. This exercise reveals that treating the data "as if" complete leads to large differences between the population and sample networks under virtually all removal scenarios. Not surprisingly, the biases are larger in the induced subgraph (as less information is available about the network, conditional on the sampling rate) and increase with the removal rate (decrease with the sampling rate).

The most biased characteristics are average degree, graph span, and the epidemic threshold under both sampling methods. The raw sample data consistently make the network appear less connected, exhibiting longer average distances, and as less epidemic-prone than it actually is. All these findings are direct consequences of observing fewer links than there actually exist in the population. The biases in the sample average degree are in the range of 15-20%, 30-49%, and 42-71% in the induced subgraphs and 4-5%, 12-20%, and 28-46% in the star subgraphs for $\psi = 80\%$, 60%, and 40%, respectively; the biases in the sample graph span are are respectively 8-20%, 20-70%, and 38-390% for the induced subgraph and 1-4%, 5-16%, and 12-52% for the star subgraph; and the epidemic threshold is biased respectively almost up to 28%, 77%,

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

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

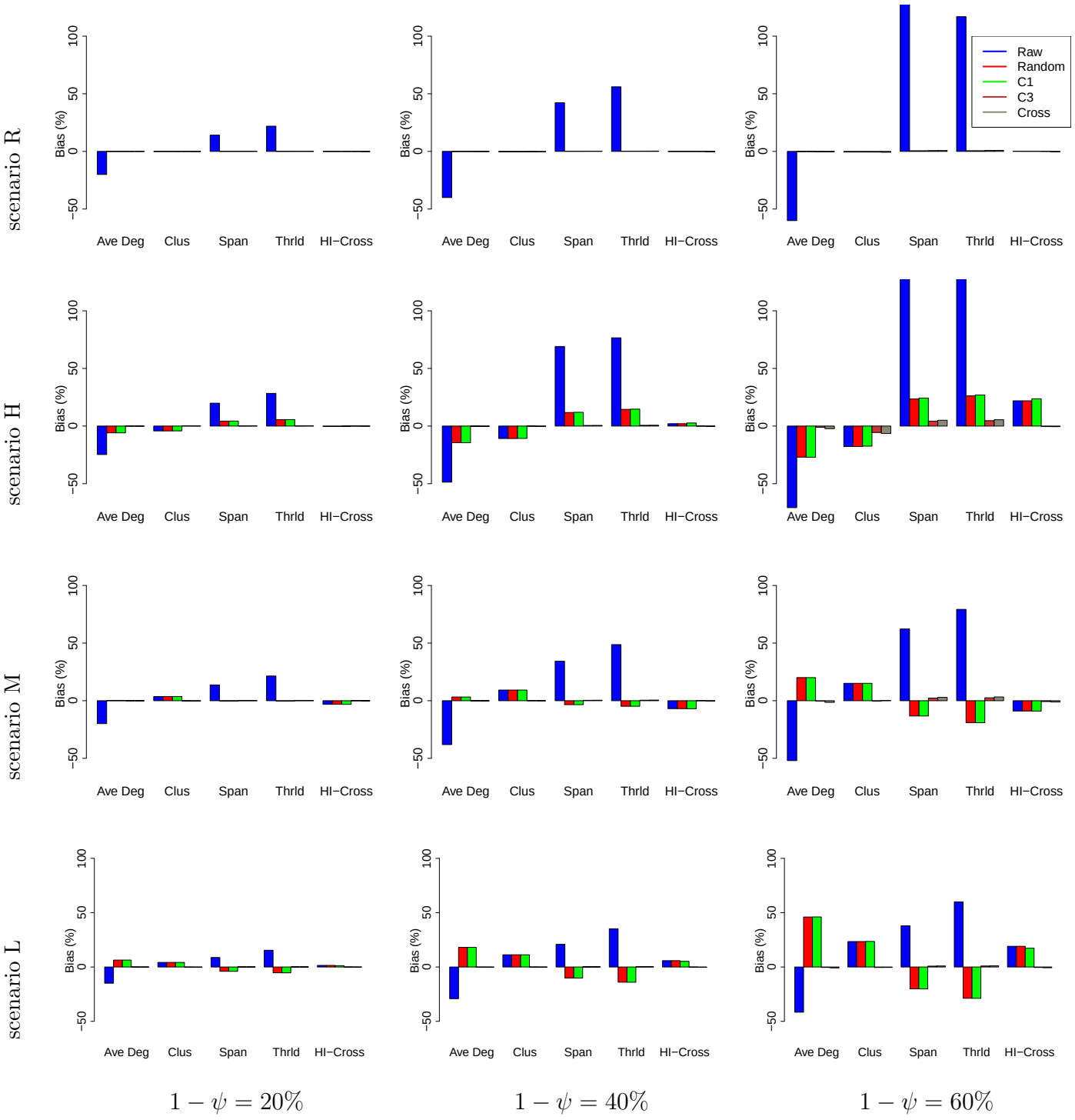

Figure 1: Induced subgraph. Biases (%) in five estimated network measures (average degree, clustering coefficient, graph span, epidemic threshold, and homophily index of cross-characteristics) from the raw sample and their corrected versions by weighting for three different removal rates 20% (left), 60% (center), 40% (right) and four different removal scenarios, R, H, M, and L. The bias is computed by subtracting the population parameter value from the average across 1000 simulation repetitions.

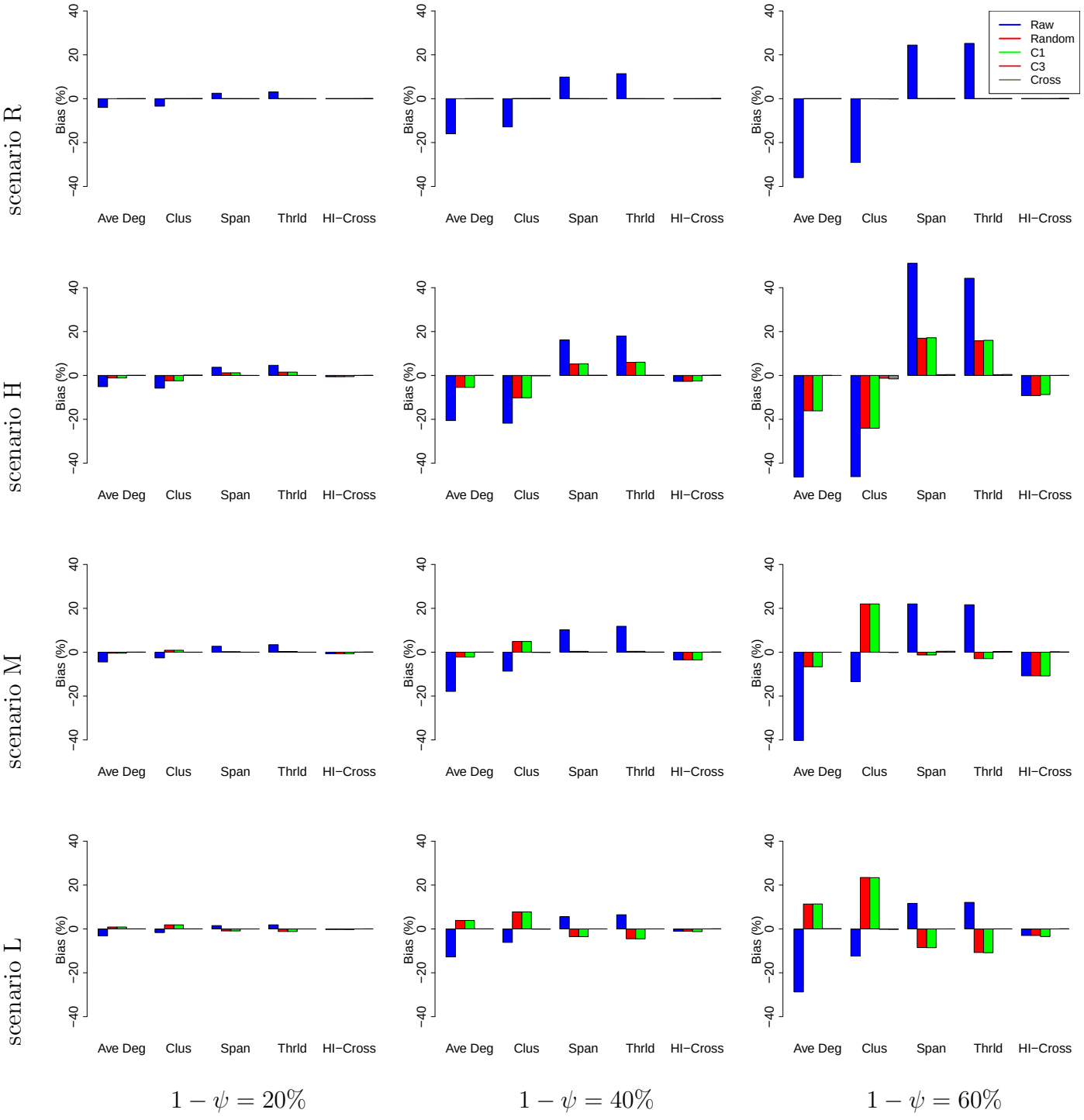

Figure 2: Star subgraph. Biases (%) in five estimated network measures (average degree, clustering coefficient, graph span, epidemic threshold, and homophily index of cross-characteristics) from the raw sample and their corrected versions by weighting for three different removal rates 20% (left), 60% (center), 40% (right) and four different removal scenarios, R, H, M, and L. The bias is computed by subtracting the population parameter value from the average across 1000 simulation repetitions.

and 163% in the induced subgraphs, and 5%, 18%, and 44% in the star subgraphs. In the case of the three measures, the extent of biases is clearly associated with who is removed and in the expected direction. Under non-random missing and conditional on ψ, we detect the largest biases when there is a tendency not to observe relatively connected individuals, followed by the removal of nodes with intermediate connectivity, and then with the removal of low-degree nodes exhibiting the lowest biases in average degrees. Random removal leads to biases comparable to the removal of nodes with intermediate degrees.

The biases exhibit more complex patterns in case of the global clustering coefficient and the homophily index, two network measures that represent shares and are thus restricted to values between zero and one. The biases also increase with the removal rate, but they are not necessarily lower in the star subgraph and their magnitudes and signs are highly sensitive to the removal scenarios.

In the induced subgraphs, the clustering coefficient is barely biased under random removal (less than 1%) and this removal scheme always deviates the clustering coefficient the least from its true population value. However, the sign depends on who is removed with higher probability. In particular, 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 clustering coefficients are biased (downwards) by -4.2%, -10.8%, and -17.8% for $\psi = 80\%$, 60%, and 40% when more higher-degree nodes are not sampled; the corresponding figures are positive and relatively similar for scenarios M and L (3.7%, 9.3%, and 15.1% vs. 4.2%, 11.3%, and 23.4%).

In contrast to the induced subgraph, the biases in the clustering coefficient under the star subgraphs are always negative, they rival those in the average degree, graph span, and the epidemic threshold, and even the random removal can bias the coefficient downwards considerably.[24] Under random removal, the observed biases are -3.4%, -12.9%, and almost -30% as the missing rate increases and they are similar in magnitude to the non-random missing scenarios. The biases are always the largest in absolute terms if high-degree nodes are removed with higher probability (scenario H: -5.8%, -21.8%, and -46.1% for $\psi = 80\%$, 60%, and 40%) and lower in the other two cases (scenario M: -2.6%, -8.6%, and -12.4; scenario L: -1.7%, -6.1%, and -12.4%). This is an example showing that non-random missing may generate lower biases than representative samples for the clustering coefficient if the "right" nodes are removed.

Regarding homophily on cross-characteristics, it exhibits no biases whatsoever under random missing (less than 0.20%). This is probably the reason why the literature has ignored the effects of sampling on homophily. However, non-random missing leads to mismeasured values because different types of nodes play crucial role in the homophily index and they are now missing in different proportions. Under disproportional missing of different types, the biases again increase with the missing rate but they are similar under both sampling methods. In addition, the effect of who is missing is non-monotonic: when either high-degree or low-degree nodes are missing with higher probabilities, induced subgraphs look more homophilous than

[24] In order to explain this result, consider random removal. Since the inclusion probability of the nodes does not depend on their clustering, no bias exists under induced subgraph. However, if two or more neighbors of a sampled node are not observed in a star subgraph, their links to the sampled individuals are included in the sampled network but their mutual connections are not, driving the sample clustering coefficient down. Since the number of such cases increases as we decrease the sampling rate, the network always looks as if it had lower clustering coefficient than it actually has in the star subgraphs and this negative bias increases with $1 - \psi$.

they actually are while they look less homophilous if nodes with median degrees are more likely to be non-sampled. In star subgraphs, sampling uniformly makes the networks look less homophilous. Quantitatively, the biases are minor for high sampling rate, but they are economically relevant for $\psi = 40\%$ (raising above 20%).

Biases in the corrections. The second objective of this section is to compare the raw network sample statistics with the random corrections and our weighting approach with three variations of weights. Remember that seniority (C1) is generally uncorrelated with connectivity while race (C3) is strongly associated, but both variables determine who links with whom. Therefore, we first discuss the case of C1. One would expect that weighting on C1 corrects at least the scaling effect. This exercise is of interest to illustrate what happens when a researcher weights on a variable that contains little information about the network. Can performing our methodology based on such a variable 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 network position. Last, we weight on *Cross* (that is, the combination of C1, C2, and C3). We hypothesize that weighting on *Cross* would outperform the previous cases, since this correction employs most of the available information.

With a few exceptions (that never involve our preferred strategy), all correction strategies lead to smaller or equal biases than the results based on the raw sample. As expected, the corrections assuming randomness work well under random sampling and our weighting approach has similar performance. Nevertheless, the random corrections and our weighting approach diverge once the missing-at-random assumption is violated. As hypothesized, the random corrections and weighting on the network-irrelevant variable C1 perform very similarly overall. This is an important result, since it shows that weighting on any variable still mitigates the biases and does not hurt, compared to the raw sample data and the random corrections. However, these corrections are still biased and these biases increase with the missing rate and are lower in star subgraphs. Most importantly, the biases are minimal if we weight either on C3 or *Cross*. Both approaches always outperform all other methods under non-proportional missing and in most cases by an order of magnitude. In fact, both weighting schemes almost eliminate any biases arising from sample network data. The maximum biases are below 0.2%, 0.8%, and 5.5% for $\psi = 80\%$, 60%, and $\psi = 40\%$ for the induced subgraphs; the figures are even lower for the star subnetwork. Our approach perform even better in case of the bounds on the maximal eigenvalue. See Supplementary Appendix C.

Root mean square errors. Since Figures 1 and 2 only reflect the average biases without reflecting the variability of the estimators across the 1000 simulated repetitions, Figures C.1 and C.2 in Supplementary Appendix C complement Figures 1 and 2 by reporting the normalized root mean squared errors (RMSEs) that reflect both the average biases and variances. The RMSEs allow us to assess whether one faces a bias-variance trade-off while employing our corrections.

The RMSEs corroborate the conclusions drawn based on the average biases. Most importantly, our preferred strategy weighting on the variable *Cross* not only outperforms all the other strategies in terms of the average biases but also in terms of RMSEs. There are two exceptions though, both involving the clustering coefficient.[25] We thus conclude that our

[25]Under the induced subgraph and $\psi = 40\%$, scenarios H and M generate larger RMSEs of the clustering coefficient weighted on *Cross* than random corrections and the raw network measures. In these cases, although our corrections lead to essentially unbiased estimators on average, their variances are slightly larger for high missing rates in general.

estimators do not come at a cost of larger variances overall, but we note that scholars should be careful while recovering the clustering coefficient from non-representative samples if the missing rate is high. In these cases, our methodology eliminates the biases on average but a certain bias-variance trade-off exists.

As a result, the proposed approach should improve inference on sampled networks by delivering the least biased and more stable estimates of network effects compared to raw data and random corrections in most cases. Under our approach, we should at most expect certain attenuation and if some biases remain they would be more pronounced in the induced subgraphs. The next section analyzes these conjectures.

4.2 Network effects

We now turn the attention to the performance of our weighting approach in a regression framework, aiming at estimating global network effects on economic outcomes. Since our weighting approach is designed for global network measures, the independent variables of network measures in the regressions are measured at the network-wide level. We also limit the present analysis to network-level dependent variables, such as the mean, median, or other statistics computed from individuals' outcome in the network. However, as discussed in Section 2, the method can be extended either to models in which we directly regress individual behaviors or outcomes on global network properties, or in network formation applications in which the network properties are the regressands.

To generate the population data, we again take the manipulated Add Health school sample of the size 1,500 students as the phototype. We create 200 artificial populations of the size 1,500 with the 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 populations. 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 then nodes of other races. There are also features of clustering and homophily in different types. To simulate the population dependent variable y in each network, we follow a simple linear regression model: $y_r = \alpha + \beta w(G_r) + \varepsilon_r$, with ε_r being an i.i.d. random error from $N(0, 1)$. We generate the data with designed parameters $\alpha = 1$ and different β's corresponding to different network measure $w(G_r)$: $\beta = 0.5$ for average degree; $\beta = 5$ for the clustering coefficient; $\beta = -0.5$ for the epidemic threshold; $\beta = -0.5$ for graph span; and $\beta = -0.5$ for each of the homophily indices. Lastly, for each of the population networks, we generate 1000 artificial samples following the same removal strategies from the previous section. This generates the artificial raw sample data, on which we apply the corrections based on the missing-at-random assumption and our poststratification weighting approach. We estimate the model based on five cases, which are the uncorrected estimators, random corrections, and poststratification weighting on C1, C3, and *Cross*, and compare them to the estimates based on the artificial population networks.

Figures 3 and 4 report the average biases in the estimated β's (out of 1,000 Monte Carlo repetitions) with respect to the population for the induced and star subgraphs, respectively. Again, y-axes reflect the biases in percentage terms and their signs and x-axes the five network statistics.The blue bars correspond to the raw sample, the red ones correspond to the corrections based on random missing, and the green, dark red, and gray bars correspond to

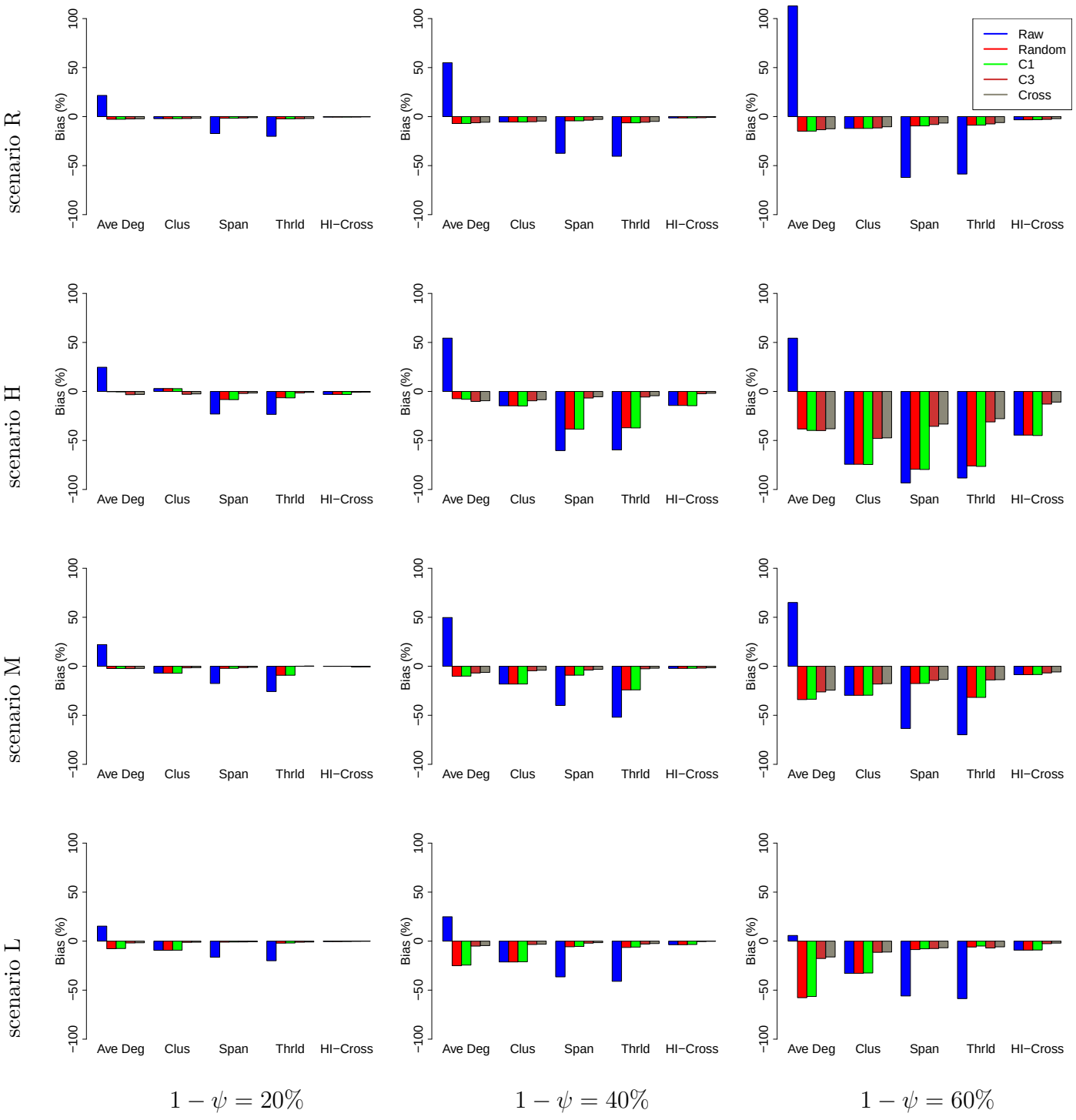

Figure 3: Induced subgraph. Biases (%) in five estimated network effects (average degree, clustering coefficient, graph span, epidemic threshold, and homophily index of cross-characteristics) from the raw sample and their corrected versions by weighting for three different removal rates 20% (left), 60% (center), 40% (right) and four different removal scenarios, R, H, M, and L. The bias is computed by subtracting the population parameter value from the average across 1,000 simulation repetitions.

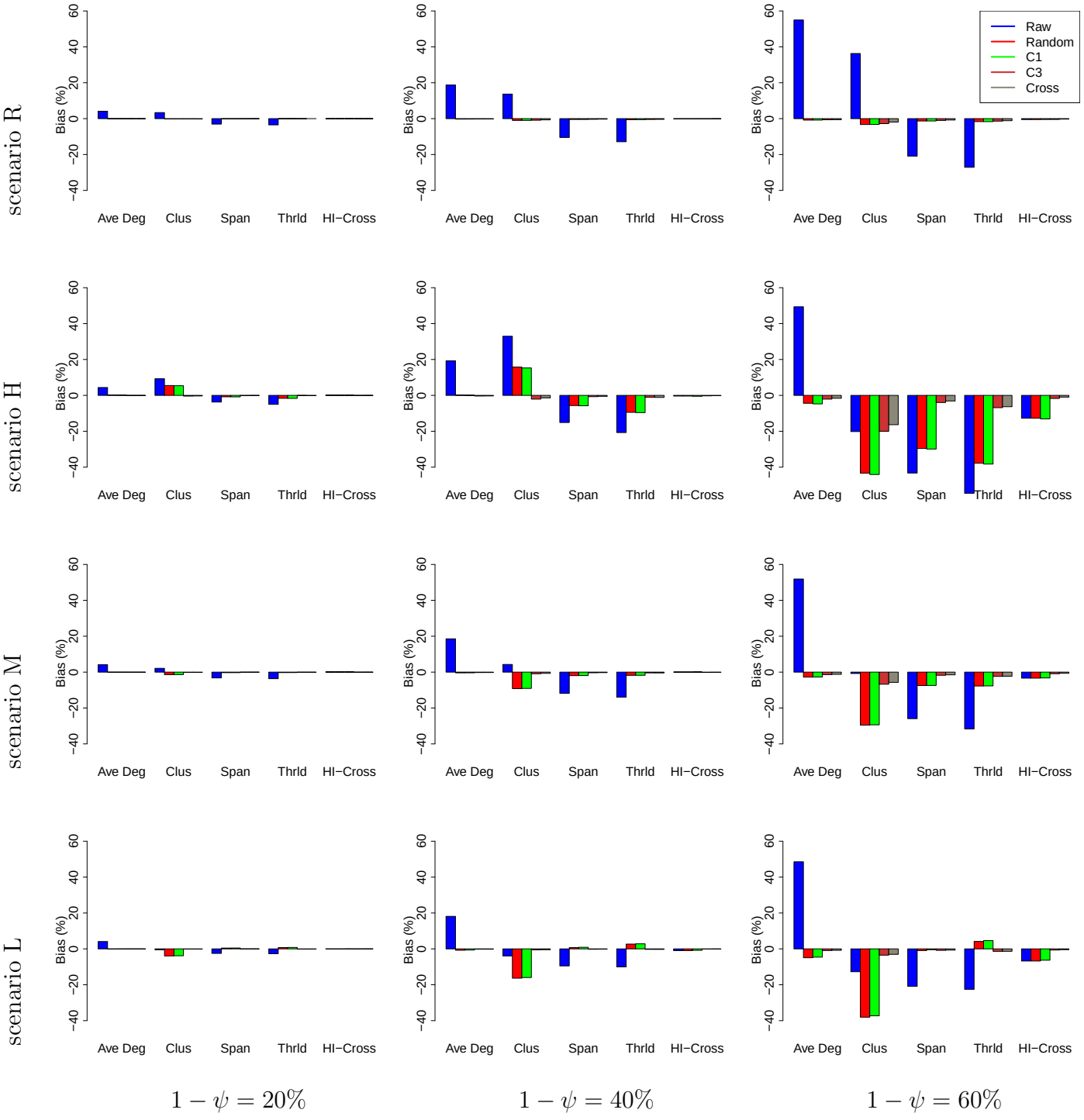

Figure 4: Star subgraph. Biases (%) in five estimated network effects (average degree, clustering coefficient, graph span, epidemic threshold, and homophily index of cross-characteristics) from the raw sample and their corrected versions by weighting for three different removal rates 20% (left), 60% (center), 40% (right) and four different removal scenarios, R, H, M, and L. The bias is computed by subtracting the population parameter value from the average across 1,000 simulation repetitions.

weighting on C1, C3, and *Cross*, respectively. Again, Figures C.3 and C.4 in Supplementary Appendix C plot the corresponding normalized RMSEs.

The general results regarding the biases mimic those of network characteristic measures in most respects: the biases in the estimates–both from the uncorrected estimators and after applying the corrections–increase with the missing rate, are higher for the induced subgraphs compared to the star subgraphs in which less data is missing (conditional on the sampling rate), depend on who is missing, and, under non-random sampling, increase as we move from scenario L through scenario M to scenario H.

The raw network data exhibit the largest biases in the estimates for all network statistics. The estimated effects of average degrees are always inflated; those of the clustering are generally attenuated under induced subgraphs, but the sign of the biases depends on the parameter constellation under star-subgraph sampling; the biases in the estimated effects of the remaining three network characteristics are attenuated.

We first discuss the induced-subgraph sampling scheme. The biases in the estimates using the uncorrected estimators are sizable in percentage terms. They are specifically large for the average degree, the epidemic threshold, and the graph span. The raw data always inflate the network effects of the average degree; the effects are inflated by roughly 20%, 50%, and above 50% for $1-\psi = 20, 40, 60\%$, with the exception of Scenario L under which the figures are lower but still important. Hence, expansion and non-classical measurement error problem present a serious issue in case of the impact of connectivity on economic outcomes. The graph span and the epidemic threshold computed from the raw data induce slightly smaller biases than average degree but with reverse signs.

The magnitudes of the biases in case of the clustering coefficient are largely sensitive to the removal strategy and sampling rate. As mentioned above, the estimated effects of clustering are biased downwards (with the exception of $\psi = 0.8$ in scenario H, in which the bias is positive and almost 3%). The biases are relatively low (-2,1%, -5.5%, and -12.1% for $\psi = 80\%$, 60%, and 40%) under random sampling, However, under non-random sampling, the values raise to below -10% if $\psi = 20\%$, around 20% if $\psi = 40\%$, and well above 30% if $\psi = 40\%$. The biases in the estimates of homophily on *Cross* are negative and quantitatively small.

With virtually no exception, all the corrections eliminate the expansion problem but negative biases remain. Both the corrections based on randomness and the weighting approach mitigate the biases with respect to the uncorrected estimators, but accounting for the non-representativeness of the sample using our methodology generally reduces the biases as compared to the corrections based on randomness. This is already the case in representative samples. Once the sampling is not random, our approach outperforms random corrections. In quantitative terms and independently of the network measure, our corrections based on *Cross* exhibit negative biases of at most 3% and 10% if $\psi = 80\%$ and 60%, respectively. These figures are lower that the biases under random corrections slightly if $\psi = 80\%$ and in many cases considerably if $\psi = 60\%$. Under $\psi = 40\%$, the estimated effects of all the corrected measures are biased considerably independently of the correction, but our methodology always mitigates the biases *vis-à-vis* the random corrections and the improvement is often dramatic. The RMSEs reported in Figure C.3 corroborate all these conclusions.

The star-subgraph statistics exhibit lower biases than the induced subgraphs and this also holds when the estimated network statistics are applied as regressors. Using the raw network data, the effect of average degree is overestimated, that of the clustering coefficient can be over- and underestimated depending on the parameters, and the impacts or the other network measures are attenuated. With few exceptions, the biases in the estimates of network effects

using the uncorrected estimators are mostly below 5%, 20%, and 60% for $\psi = 80\%$, 60%, and 40%. Both types of corrections reduce the biases drastically. Random corrections never outperform our preferred weighting strategy and they particularly fail to recover the true effect of the clustering coefficient. If $\psi = 40\%$, the corrections under the missing-at-random assumption also fail to recover the true effects of the graph span and the epidemic threshold if this assumption is violated. Under the star sampling procedure, our weighting approach eliminates virtually all the serious biases in the estimates. Overall, our approach is remarkably successful recovering the true network effects under the star-subgraph sampling, a conclusion that is even reinforced if we look at the RMSEs in Figure C.4.

Hence, our methodology is generally the most successful in mitigating the biases in the network effects and preventing false positives in finite samples.

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 does not account for non-representativeness of network samples. In the first subsection, we use village network data from rural India. In the second, we employ adolescent friendship networks from U.S. high schools. Both data sets contain information on (non-network) characteristics for the whole population, but only sampled data on networks.

5.1 Village Networks in Rural India

Banerjee et al. (2013) elicit a large variety of characteristics including network data from 75 villages in southern Karnattaka, India, corresponding to 75 different networks. The authors initially collected the census information for each household (on age and gender of household members) in all villages and later conducted detailed follow-up survey with a subsample of the population of each village. In the latter, they also elicit the network of relationships among individuals. Two features of this data make them particularly interesting for our purpose. First, as in most studies, the survey respondents only represent a sample of each village and their reported network is an induced subgraph of the population network. The average sampling rate across villages is 35%. The crucial aspect of their sampling design is the stratification by religion and geographic sub-location, generating a representative sample with respect to these two variables. This is a common approach in many applications. The stratification on religion and geography notwithstanding, Table 2 reveals that the data are not representative in terms of age, gender, and–to a lesser extent–household size. Below, we test to what extent the differences between the sample and population shares of these categories affect the estimation of network effects in regressions like (1).

Table 2: Population and sample shares of different characteristics categories and labor outcomes in the Indian rural village data from Banerjee et al. (2013).

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

Second, the data contain several variables regarding the labor market outcomes of the participants, such as the employment status, whether they work outside the village, and their occupation. Since the important role of social networks in labor markets is widely acknowledged in the literature and documented in the data (Granovetter, 1985; Montgomery, 1991; Calvo-Armengol and Jackson, 2004; Granovetter, 2005) this application is interesting in its own right. The theoretical literature argues that the degree distribution (Calvo-Armengol and Jackson, 2004) and the average clustering coefficient (Espinosa et al., 2021) might affect the employment prospects directly, while average network distances and the epidemic threshold might influence the flow of labor-market information and thus the labor outcomes indirectly (see Calvo-Armengol and Jackson (2004) for examples). Similarly, the extent of segregation may determine who hears about jobs and who does not. However, little empirical evidence exists regarding the impact of different macro features of networks on labor markets, due primarily to the lack of suitable data containing enough networks. We thus ask how the village employment rate and the fraction of people working outside the village correlate with the global features of the underlying network of relationships within the village. Most importantly, for the present study, we ask how the estimated network effects change if we account for missing network data and non-randomness of the sample. Specifically, we hypothesize that the over-representation of people aged 30-50 and under-representation of men in the sample (see Table 2), those who typically actively participate in labor markets in a country like India, might bias the estimated network effects if their misrepresentation is not accounted for.

Table 3 reports the estimated network effects in a series of estimations differing in (i) the dependent variable (I. employment rate; II. fraction of population working outside the village), (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 *cross* variable (incorporating the information on age, gender, and household size). Each row reports the estimated network effect (and the standard error robust to heteroscedasticity in parentheses) from a regression of one dependent variable on the corresponding network statistic and village size. There are two important things to note. First, we also apply the post-stratification cross weighting on the dependent variables, i.e., employment and working outside villages, at the village level to correct measurement errors due to non-randomness of the sample.[26] Second, we use the network constructed by the union of all relationships reported by survey respondents, e.g., borrowing, lending, seeking advices, going to temple together, visiting home, and others following Banerjee et al. (2013). We also study the network effects which are only based on reported friendships and find similar results (see Supplementary Appendix Table B.1).

Table 3: Estimated network effects on the share of population in rural India village that (I) employed and (II) work outside the village.

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

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

Regarding the main purpose of our exercise, Table 3 shows the sensitivity of results with respect to non-representative sampling. First, the estimates using raw data or corrections based on the missing-at-random assumption might be attenuated, expanded, and can even switch signs, compared to the corrections that account for both scaling and the non-representativeness of the network data. There are examples of false positive findings: two regressions in Table 3 detect a significant network effect using both the raw network statistics and those correcting for scaling, but this effect disappears if we correct for non-randomness of the network sample. There exist one case, in which we detect no effect using the uncorrected estimators and random corrections but we do using our weighting approach. The sign switching occurs three

[26] See also footnote 13.

times in case of the uncorrected estimators and three times in case of the random corrections; it always concerns the effect of homophily. Second, overestimation with respect to the effect of the weighted network statistics – disregarding the (non-)significance – is commonplace.

In the $2 \times 8 = 16$ regressions employing the raw network data, the biases in the estimates are always higher than 10% and 10 of the regressions are overestimated. Under overestimation, the estimates are inflated between 17% in case of clustering and 2,226% in case of the average degree; if attenuated, the biases range from 30% in case of homophily on *cross* to 415% in case of the age homophily. The remaining biases are spread somehow in between these extremes and are overally economically important.

In case of the random corrections, 11 (out of the 16) estimates reported in Table 3 are biased more than 10% with respect to the weighting approach. The biases are below 10% in case of the average degree and the epidemic threshold in both applications and in case of graph span when regressed on the fraction of people working outside the village. If biased more than 10%, the estimates are inflated in 7 regressions (43.8%). These biases lie between 14.8% in case of the clustering coefficient and 460.7% for cross homophily. Underestimation of the estimates by more 10% is observed in 4 instances (25%) and the biases range from 30% in case of cross homophily to 415.9% in case of age homophily. Again, the remaining cases are distributed in between.

In sum, there are important biases in the estimates under both the raw data as well as corrections assuming representativeness. Crucially, false positives, expansion of network effects, and sign switching are common. More important, the direction and the magnitude of the biases depend non-trivially on the particular network statistics, the dependent variable under study, and who is missing. Observe that the largest biases are detected in case of age-related statistics, the category that differs the most between the population and the sample. Hence, researchers cannot easily predict the direction of the biases and thus cannot rely on classical measurement-error solutions, even in the simplest cases analyzed here.

As for how the village networks shape labor outcomes, we corroborate the literature in that the architecture of social structures plays a key role in labor markets. Accounting for the non-representativeness of the sample (that is, columns denoted *Cross* in Table 3) shows that virtually all the features of social organizations under study matter for the average labor outcomes in the village and the effects of the different characteristics are to a large extent consistent with each other. More precisely, both higher average degrees and more dense social circles stimulate employment and prevent people from having to travel for work outside their village. Higher average degree favors the flow of information about jobs throughout the village, naturally making people more likely to find a job and to find it within the village. As for the clustering coefficient, it is an important measure of whether people take care of each other within a village in adverse situations, such as unemployment (Coleman, 1988). Indeed, the estimated effect is positive, suggesting that more clustered villages exhibit higher employment rates and less need to search for jobs outside the village. Relatedly, shorter distances and higher epidemic thresholds lead to lower employment rates and more traveling for work outside the community. This is in line with the idea that higher values of both variables correspond to less integration, hindering–among other things–the flow of job-related information from more distant network neighborhoods. Last, homophily does not seem to effect the labor outcomes systematically. The only exception is age homophily. The estimates reveal that higher segregation across different age categories makes people more likely to work outside the village. Recall that higher homophily corresponds to less connections and thus lower information flow across different subgroups in the village population. This effect is thus

in line with the effect of the graph span.

5.2 Adolescent friendship networks in U.S. high schools

As a second application, this subsection applies our approach to the Add Health data set. These data contain extensive information on friendship networks in selected U.S. high/middle schools and detailed data on individual heterogeneity. The schools are representative.[27] Even though the survey conductors collected information from all the students present during the questionnaire day, the average sampling rate is 63.6% of the school census. We combine the sampled network data with certain information on the roster in each school. Since the participation was determined non-randomly, we can expect non-representativeness of the sample. As Table 4 indeed shows, the sample and the population differ in terms of race and grade compositions. In particular, white students and students of the ninth grade are underrepresented in the sample. We already know that white students are more connected. Thus, higher missing of white students will directly affect the observed network connectedness and the degree of race segregation. As for the grade composition, younger students might for instance be less integrated in the school networks, compared to their older schoolmates. Hence, there are good reasons to believe that the observed school networks are mismeasured and that this may affect inferences on these networks.

Table 4: Population and sample statistics of different characteristics categories and school activities in the Add Health data.

	Population	Sample	Difference (p-value)
White Ratio	60.43%	53.32%	7.11% (0.214)
Year grade			
9th	32.22%	30.00%	2.22% (0.044)
10th	25.63%	26.53%	-0.90% (0.121)
11th	22.06%	23.00%	-0.94% (0.139)
12th	20.09%	20.47%	-0.38% (0.566)
Activity			
club		2.03	
exercise		4.35	
Num. of Schools	48	48	
Observations	66,025	40,898	

We study two particular school activities here. The first activity is inspired by Bramoullé et al. (2009) who find large peer effects in club participation using the Add Health data. This raises the question of whether the global features of friendship networks predict the average club participation at high schools. The second activity is inspired by the experiments of Centola (2010, 2011) regarding the effect of friendship networks on the spread of health-related

[27] We use 48 high schools (out of 80) in Add Health data which consist of year 9 to year 12 and student of different races. These 48 schools have a complete registration record (i.e., population record) on race and year compositions of their students.

behaviors. He reports that health related behaviors, including exercising, are heavily influenced by the clustering coefficient, network distances (Centola, 2010), and homophily (Centola, 2011). Therefore, we regress two dependent variables, the average number of clubs attended and average exercise frequency in the schools, on the same network characteristics as in Section 5.1 and the school size. Table 5 reports the estimates using again raw network data, networks corrected for scaling, and those corrected by our poststratification weighting. The table has the same structure as Table 3.

Table 5: Estimated network effects on club participation and frequency of exercise in U.S. high schools.

Dependent Variable	Number of clubs attended			Frequency of exercise		
	Raw	Random	Cross	Raw	Random	Cross
Degree	0.0681**	0.0646**	0.0660**	0.0970**	0.0526***	0.0498***
	(0.0329)	(0.0257)	(0.0255)	(0.0421)	(0.0186)	(0.0194)
Cluster	-0.0116	-0.0116	-0.0583	0.5104	0.5104	0.4591
	(0.3729)	(0.3729)	(0.3571)	(0.6480)	(0.6480)	(0.6289)
Span	-0.0009	-0.2886**	-0.3561***	-0.0079***	-0.1940*	-0.2399**
	(0.0012)	(0.1117)	(0.1240)	(0.0014)	(0.1049)	(0.1192)
Epid Thrld	-2.3991*	-10.7449***	-12.9819**	-4.8215***	-6.6086*	-8.3189**
	(1.3756)	(3.7717)	(4.0284)	(1.5809)	(3.3198)	(3.9157)
HI-grade	2.4168**	2.4168**	2.6725**	1.3443	1.3443	0.6779
	(1.0237)	(1.0237)	(1.0310)	(1.1274)	(1.1274)	(1.3677)
HI-race	-0.8281*	-0.8281*	0.0248	-1.9841***	-1.9841***	-0.8637
	(0.4314)	(0.4314)	(0.5886)	(0.6582)	(0.6582)	(0.7602)
HI-cross	0.1263	0.1263	1.2347*	-1.6500*	-1.6500*	-0.4103
	(0.7949)	(0.7949)	(0.6998)	(0.9354)	(0.9354)	(0.8941)

Note: Regression is based on 48 schools. *, **, *** stand for significance at 10%, 5%, and 1% respectively. Each row corresponds to one regression with the school size included as a default control.

Again, we observe that some network effects appear and some disappear, and there are two instances of sign switching once we account for non-randomness. Similarly, the (non-) significance can persist, appear, or disappear. There are three cases in the table, in which both the uncorrected estimates and the random corrections generate significant impact on the dependent variables, while they become insignificant with our weighting approach; in one case, the opposite occurs. Contrary to Section 5.1, whether the effects are overall attenuated or inflated depends crucially on the dependent variable.

In case of the club participation, the only case in which we detect expansion corresponds to the effect of average degree using the uncorrected estimators and the bias is small (3.2%). All the remaining effects are attenuated with respect to our weighting approach. The attenuation is small for the effect of the average degree corrected for random samples (-2.1%) and homophily on grade (-9.6%). Nevertheless, the estimated effects are biased downwards considerably for the other variables. The estimated effects are more biased for the raw data. The effects of clustering are biased downwards by 80.1%, whereas those of the graph span and

epidemic threshold by more than 80% and roughly 18% for the uncorrected estimators and random corrections, respectively. The more dramatic biases appear–similarly to Section 5.1–for the variables affected the most by the differing sampling probabilities of different types. The estimated effect of homophily on race, the variable closely related to network positioning in the data, is biased downwards by more than 3,000% if we do not correct for the higher missing rates of races other than white. Race homophily seems to affect negatively club participation if we disregard the non-representativeness of the network data, but once we correct for it the effect switches the sign and becomes statistically insignificant.

As for the exercise frequency, the effects of the average degree, the clustering coefficient, and homophily are always overestimated, while those of the epidemic threshold and graph span are attenuated using both the raw data and corrections assuming representativeness. The biases in the effects of average degree are 94.8% with uncorrected estimators but only 5.6% if corrected for scaling. The effect of the clustering coefficient is overestimated by 11.2% in both cases. The influence of homophily is always inflated by more than 98% independently of the homophily type. The estimated effects of graph span and epidemic threshold are underestimated by around 20% if we employ random corrections; the figures are considerably larger with the raw network data.

Regarding the effects of the network if corrected for both scaling and non-representativeness, the average connectivity of the network stimulates both club participation and exercise frequency, and longer distances and higher epidemic threshold hinder them. All these features reflect different aspects of higher network integration and the signs align with the intuition. Homophily positively affects the club attendance but does not influence how many people exercise. The former finding confirms Centola (2011) in that higher homophily stimulates diffusion, but the latter contradicts his findings that higher homophily stimulates the adoption of health behaviors. In contrast to Centola (2010), the clustering does not explain any dependent variable. Since the experiment of Centola (2010) cannot disentangle the effect of clustering from that of distances by design, we re-estimate the models from Table 5 including both the clustering coefficient and graph span in one model (see Supplementary Appendix Table B.2) but the results do not change. Hence, the (treatment) effect detected by Centola (2010) seems to be driven by network distances, whereas the effect of clustering might play a minor role in promoting health behaviors in his data. This conclusion is reinforced by the fact that it holds for both dependent variables.

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. Most importantly, these applications illustrate that both the magnitudes and the directions of the biases depend non-trivially on the dependent variable under study.

6 Discussion

In this section we briefly comment on potential extensions and limitations of our methodology regarding other sampling methods and network measures. We also provide several recommendations concerning the selection of the weighting variables.

Alternative sampling strategies. Although this paper focuses on two standard sampling methods, the induced- and star-subgraph elicitation, the proposed methodology can

be adopted to other sampling schemes as long as the researcher knows the strategy employed for the elicitation of the sample and possesses some information on the whole population. In the following, we provide several examples illustrating the applicability of the proposed approach to other sampling strategies.

As a first example, consider the issue known as the boundary specification problem. Regardless of whether a researcher elicits a complete or sampled network, she must set a boundary to determine the population of interest. Imagine for simplicity that the researcher elicits network sample from one class in a school, disregarding any individual from other classes and the ties from the class under scrutiny to people outside the class. It is very likely that there exist connections between the members of the class and other people who do not belong to the class. Hence, even if there exists a clearly defined boundary and the class network sample is complete, the true social network of the studied population is most likely incomplete. If one would like to study the complete network sample, say, at the school level and individual characteristics are available for the whole school, one can mitigate the boundary specification problem by applying directly our method because setting a boundary is mathematically equivalent to our induced subgraph sampling.

As a second example, consider snow-ball sampling, a sampling procedure commonly applied in Sociology, Marketing and Epidemiology (Berg, 2004; Browne, 2005; Chen et al., 2013). Under the snow-ball sampling, a researcher initially selects a randomly selected subset of nodes. Then, she performs the first wave by eliciting all the contacts of the initially selected nodes. In the second wave, she elicits all the contacts of the nodes found in the first wave, and so on. Observe that the star-subgraph sampling treated above is formally equivalent to one-wave snow-ball sampling and our methodology directly applies. Nevertheless, there is a difference between our approach and the corrections proposed in the literature for one-wave snow-ball sampling: they only coincide if each type is missing with the same probability or if the weighting variables provide no information about the network under study (see Kolaczyk (2009); Zhang et al. (2015)). We argue this is rarely the case even in very carefully and systematically collected data sets. Although the computation becomes increasingly complex as more and more waves are preformed, one can adapt our approach to multiple waves of snowball sampling taking into account the missing frequencies of each type and the information about the within-type and across-type connectivity from the observed part of the network using combinatorial arguments (see e.g., Frank (1977); Snijders (1992)). It is also possible to take into account that all connections are observed for one part of the sample and apply the corrections to nodes for which some information might be missing. This notwithstanding, the complexity of the problem increases with the number of waves as mentioned above and the literature almost exclusively focuses on the one-wave variant even if all the randomness solely comes from the sampling process (Frank, 1977). Similar considerations apply for forest-fire sampling, a generalization of snow-ball sampling. It again starts with a set of randomly chosen individuals and operates in waves. In the first wave, the analyst elicits the links from the initially chosen individuals, but each link is only followed with a probability $p \leq 1$, and similarly in the following waves. Naturally, if $p = 1$ forest-fire sampling is equivalent to snow-ball sampling. Again, one can incorporate the probability p to the variations of our corrections adapted to snow-ball sampling.

Unsurprisingly, the corrections proposed in Section 2 cannot be employed directly under other sampling designs, or should be adapted to the employed sampling strategy in the corresponding study. Consider, for example, random selection of links (also known as incident edge sampling) such that an individual i belongs to the sample if and only if at least one

of her edges is sampled. Such sampling is commonplace in communication data, where only a random sample of phone calls or e-mails is selected. The main problem of this sampling design is to compute the theoretical probability with which a particular individual belongs to the sample, but this probability is observed in the type of data we target by this study. Additionally, since the probability of being sampled depends on nodes' degrees, the differing sampling rates across types already provide information regarding the different connectivity of each type. On the other hand, the probability of each dyad belonging to the sample is the same for each link. Hence, combining this information with the observed connectivity across and within types, one can compute, for instance, the expected average degree of the network a là Section 2 but such computation might be more involved. A similar approach would apply for other network measures of interest.

In fact, Zhang et al. (2015) show for representative samples that the corrections for the degree distribution under the incidence edge sampling are very similar to those under the induced subgraph sampling; see Lee et al. (2006) for other network measures. Hence, one can propose the corresponding estimates following our approach is links (rather than nodes) are selected.

More generally, parting from the Horvitz-Thompson estimator (Horvitz and Thompson, 1952) the statistical sampling theory has developed a number of estimators of different network measures under a variety of sampling schemes (see Frank (2005) or Kolaczyk (2009) for textbook treatments) and our approach combines their developments with the ideas of post-stratification weighting. Whenever a correction exists under random sampling, it can be generalized to non-representative samples.

These conclusions notwithstanding, there exist sampling methods that cannot be targeted using our approach. For example, our approach cannot be applied to networks collected using the truncated fixed-choice design (that is, nominate up to, say, five friends) if one does not have a theory regarding who is (not) named among the five friends. For example, the Add Health data is a non-representative sample, as shown in Section 5, collected via a truncated fixed-choice design. Our approach mitigates the biases due to the non-representativeness but not those due to the truncation. However, both issues might be targeted simultaneously by combining our post-stratification weighting with the approach proposed by Griffith (2022) designed to mitigate the issues due to the truncation. Similarly, additional applications of our approach might result from combining our approach with methods designed for other purposes.

Other network measures. For their theoretical and empirical relevance, this study focuses on seven basic network properties: average degree, degree distribution, global clustering, graph span, epidemic threshold, the maximal eigenvalue, and homophily index (Jackson et al., 2017). Nevertheless, one can easily adapt the methodology to any other network measure that solely requires the knowledge of nodes' local information.[28] The examples include the assortativity coefficient, the average size of the second-order neighborhood, network entropy, or the average number of cycles of size four within nodes' neighborhoods. Assortativity is a common feature of network architectures and measures a correlation between the degrees of connected nodes. Assortativity plays a crucial rule in diffusion: it can slow down or enhance the disease trans-

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

mission but also that of behaviors and social norms (Newman, 2002; Jackson et al., 2017). Social networks typically exhibit positive assortativity, the tendency of more connected individuals to be connected to more connected individual. In contrast, many technological or biological networks exhibit negative correlations between the degrees of connected individuals. Lee et al. (2006) show how to recover assortativity from samples obtained through several sampling schemes. The average size of the second-order neighborhood (compared to that of direct neighborhood) enables to assess how fast diffusion spreads and it has also shown to be important in labour markets (Calvo-Armengol and Jackson, 2004; Espinosa et al., 2021). Since the computation of both the assortativity coefficient and the second-order neighborhood only requires the knowledge of one's degree and the degree of her neighbors, their weighted corrections follow directly from Section 2.

Other measures do not follow directly from Section 2, but our approach can still be applied. For instance, Eagle et al. (2010) apply the concept of network entropy that reflects the diversity of connections of an individual to different groups (or types in the terminology of the current paper) in the population. They report large economic advantages from belonging to communities with geographically diversified distributions of contacts. Since their measure only relies on the distribution of different types in the neighborhood of each node, the corrected variation of this measure for sampled networks is straightforward. Similarly, cycles of size four have recently received certain attention in Sociology (Opsahl, 2013), Physics (Yin et al., 2018), and Economics (Espinosa et al., 2021). One can recover it following our approach using the combinatorial logic. Since these characteristics are extensions of the ideas of homophily and the clustering coefficient, respectively, we focus on the more common variations and do not propose the corrections of these two in this study.

These examples notwithstanding, the proposed methodology cannot be applied in other cases. First, it does not allow to recover exactly global network measures computed on basis of the whole network architecture such as the spectral properties, the average betweenness or eigenvalue centrality, or network distances. However, as illustrated in Section 3, one can overcome this problem by using approximations and bounds computed on basis of local information. There is a rich literature proposing approximations of average and maximal distances and bounds on the leading eigenvalue, and an emerging literature for other eigenvalues and measures (see Sections 3.3 and 3.7 for references). Clearly, the proposed approach cannot recover the network characteristics at the individual level. Similarly to the degree distribution, one can propose recovery strategies for the distributions of, say, the clustering coefficient or the homophily but not their values for one particular node. Whether and how the logic of our methodology can be applied in these cases is left for future research.

Selection of (auxiliary) weighting variables. A natural question arising from the proposed methodology is the choice of the (auxiliary) weighting variables. Our approach is based on the idea that individual heterogeneity and the observed part of the network provide valuable information about the missing part. The evidence supports this assumption and reveals that two types of correlations are of particular relevance: the positioning in a network correlates with a series of individual characteristics and similar people (or dissimilar as in, say, romantic networks) are more likely to be connected. However, the evidence also points out that different characteristics matter in different contexts and situations. For instance, Morelli et al. (2017) report that positive emotions explain positioning in network reflecting time sharing, while empathy plays a role in intimate networks of the same people describing trust and support. Similarly, firms will probably form ties differently if searching for providers (or buyers),

compared to innovation collaborations. Hence, one has to know the particular application under study to assess which node-level characteristic might provide valuable information about the network and we prefer to refrain from making general recommendations regarding the application of particular variables.

Practically speaking, most data sets we are aware of are restricted to a relatively small number of variables at the population level. Since our results show that the performance increases with more information and that applying variables that provide no information about the network does *not* affect the performance negatively, we would recommend to employ all the available information in such cases.

If, in contrast, too many variables are available for weighting, one would like to avoid to having too many types and probably disregard some of them. In such a situation, the main problem would be to have too few observations in each stratified cell. It may increase the variance–and thus the efficiency–of the proposed weighting estimates of the characteristic under study and thus decrease the performance of the corrections. One straightforward solution is to apply the principal component analysis to filter the relevant independent information from a large number of potentially correlated variables. Another solution can be a simple two-step algorithm. Consider a set Z of population-level variables with $z = |Z|$. Denote $\widetilde{w}_{-i}(\overline{G})$ the correction of a network statistics using all the available variables but i. Last, let $\theta \geq 0$ be a (small) threshold chosen by the researcher. Then, we propose the following algorithm to eliminate some of the variables available at the population level:

Step 1: use all the available z variables and generate the correction $\widetilde{w}(\overline{G})$ of the network characteristic of interest;

Step 2: for each $i = 1, 2, ..., z$, compute $\widetilde{w}_{-i}(\overline{G})$. If $|\widetilde{w}(\overline{G}) - \widetilde{w}_{-i}(\overline{G})| \leq \theta$, remove variable i.

Despite its simplicity, the proposed algorithm targets several key features of the selection of the "right" variables. The main contribution of the proposed algorithm lies in eliminating the variables that either provide too little information for the network statistic of interest (where too little is determined by the researcher with choice of θ) or provide the same information as some other considered variable (thus providing too little *additional* information). Moreover, the algorithm selects endogenously the most suitable variables for each network statistic. As a result, different network statistics might be corrected with different variables.[29]

As for θ, it represents a threshold to be decided by the researcher. It can be set to zero if the researcher would like to maintain all the information; alternatively, it can be set equal to any arbitrary (most likely) small number or computed on basis of a desirable percentage improvement if the research would like to filter out only the most relevant variables. We remain agnostic about the specific approach a researcher would take for a particular project. We stress, however, that researchers should be aware of the inferential problem addressed here and the general limits of assumed randomness of the network sampled. Given that sensitivity, a variety of weights should be used to discover if results are sensitive to the random network assumption. Such analysis should serve as a standard robustness check of empirical network results, giving scholars confidence that the results reflect network effects and are not a figment of the sampling strategy.

[29]This is an important issue as different features of network positioning are commonly explained by different individual characteristics. For instance, in the context of friendships, Branas-Garza et al. (2010) report that social-norm adherence explains one's centrality but not the clustering coefficient whereas Kovářík and Van der Leij (2014) document that risk attitudes predict the clustering coefficient but not the centrality.

References

Acemoglu, Daron, Vasco M Carvalho, Asuman Ozdaglar, and Alireza Tahbaz-Salehi (2012) "The network origins of aggregate fluctuations," *Econometrica*, 80 (5), 1977–2016.

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

Ammermueller, Andreas and Jörn-Steffen Pischke (2009) "Peer effects in European primary schools: Evidence from the progress in international reading literacy study," *Journal of Labor Economics*, 27 (3), 315–348.

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

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

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

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

Bardoscia, Marco, Stefano Battiston, Fabio Caccioli, and Guido Caldarelli (2017) "Pathways towards instability in financial networks," *Nature Communications*, 8, 14416.

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

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

Boguñá, Marián, Romualdo Pastor-Satorras, and Alessandro Vespignani (2003) "Epidemic spreading in complex networks with degree correlations," in *Statistical Mechanics of Complex Networks*, 127–147: Springer.

Boucher, Vincent and Aristide Houndetoungan (2020) *Estimating peer effects using partial network data*: Centre de recherche sur les risques les enjeux économiques et les politiques.

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

Bramoullé, Yann and Rachel Kranton (2007) "Public goods in networks," *Journal of Economic Theory*, 135 (1), 478–494.

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

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

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

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

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

Calvó-Armengol, Antoni, Eleonora Patacchini, and Yves Zenou (2009) "Peer effects and social networks in education," *Review of Economic Studies*, 76 (4), 1239–1267.

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

——— (2011) "An experimental study of homophily in the adoption of health behavior," *Science*, 334 (6060), 1269–1272.

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

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

Chen, Xinlei, Yuxin Chen, and Ping Xiao (2013) "The impact of sampling and network topology on the estimation of social intercorrelations," *Journal of Marketing Research*, 50 (1), 95–110.

Coleman, James S (1988) "Social capital in the creation of human capital," *American Journal of Sociology*, 94, S95–S120.

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

Conley, Timothy G and Christopher R Udry (2010) "Learning about a new technology: Pineapple in Ghana," *American Economic Review*, 100 (1), 35–69.

Conti, Gabriella, Andrea Galeotti, Gerrit Mueller, and Stephen Pudney (2013) "Popularity," *Journal of Human Resources*, 48 (4), 1072–1094.

Cowan, Robin and Nicolas Jonard (2004) "Network structure and the diffusion of knowledge," *Journal of Economic Dynamics and Control*, 28 (8), 1557–1575.

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

——— (2010) "Identifying the roles of race-based choice and chance in high school friendship network formation," *Proceedings of the National Academy of Sciences*, 107 (11), 4857–4861.

Das, Kinkar Ch and Pawan Kumar (2004) "Some new bounds on the spectral radius of graphs," *Discrete Mathematics*, 281 (1-3), 149–161.

De Giorgi, Giacomo, Michele Pellizzari, and Silvia Redaelli (2010) "Identification of social interactions through partially overlapping peer groups," *American Economic Journal: Applied Economics*, 2 (2), 241–275.

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

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

Dong, Jianping and Jeffrey S Simonoff (1994) "The construction and properties of boundary kernels for smoothing sparse multinomials," *Journal of Computational and Graphical Statistics*, 3 (1), 57–66.

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

Echenique, Federico and Roland G Fryer (2007) "A measure of segregation based on social interactions," *Quarterly Journal of Economics*, 122 (2), 441–485.

Elliott, Matthew and Benjamin Golub (2019) "A network approach to public goods," *Journal of Political Economy*, 127 (2), 730–776.

Espinosa, M.P., J. Kovářík, and S. Ruiz-Palazuelos (2021) "The impact of network cycles on employment and inequality," *mimeo.*

Faloutsos, Michalis, Petros Faloutsos, and Christos Faloutsos (1999) "On power-law relationships of the internet topology," in *ACM SIGCOMM Computer Communication Review*, 29, 251–262, ACM.

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

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

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

——— (1980) "Estimation of the number of vertices of different degrees in a graph," *Journal of Statistical Planning and Inference*, 4 (1), 45–50.

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

——— (2005) "Network sampling and model fitting," *Models and Methods in Social Network Analysis*, 31–56.

Galaso, Pablo and Jaromír Kovářík (2021) "Collaboration networks, geography and innovation: Local and national embeddedness," *Papers in Regional Science*, 100 (2), 349–377.

Galeotti, Andrea, Sanjeev Goyal, Matthew O Jackson, Fernando Vega-Redondo, and Leeat Yariv (2010) "Network games," *Review of Economic Studies*, 77 (1), 218–244.

Goeree, Jacob K, Margaret A McConnell, Tiffany Mitchell, Tracey Tromp, and Leeat Yariv (2010) "The 1/d law of giving," *American Economic Journal: Microeconomics*, 2 (1), 183–203.

Golub, Benjamin and Matthew O Jackson (2010) "Naive learning in social networks and the wisdom of crowds," *American Economic Journal: Microeconomics*, 2 (1), 112–49.

——— (2012a) "Does homophily predict consensus times? Testing a model of network structure via a dynamic process," *Review of Network Economics*, 11 (3).

——— (2012b) "How homophily affects the speed of learning and best-response dynamics," *Quarterly Journal of Economics*, 127 (3), 1287–1338.

Goyal, Sanjeev (2012) *Connections: An Introduction to the Economics of Networks*: Princeton University Press.

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

——— (2005) "The impact of social structure on economic outcomes," *Journal of Economic Perspectives*, 19 (1), 33–50.

Griffith, Alan (2022) "Name your friends, but only five? The importance of censoring in peer effects estimates using social network data," *Journal of Labor Economics*, forthcoming.

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

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

Holt, D and TM Fred Smith (1979) "Post stratification," *Journal of the Royal Statistical Society. Series A (General)*, 33–46.

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

Hu, Mandy, Chih-Sheng Hsieh, and Jianmin Jamie Jia (2014) "Predicting Social Influence Based on Dynamic Network Structures," *mimeo*.

Huisman, Mark (2009) "Imputation of missing network data: Some simple procedures," *Journal of Social Structure*, 10 (1), 1–29.

Hyslop, Dean R and Guido W Imbens (2001) "Bias from classical and other forms of measurement error," *Journal of Business & Economic Statistics*, 19 (4), 475–481.

Ibarra, Herminia (1992) "Homophily and differential returns: Sex differences in network structure and access in an advertising firm," *Administrative Science Quarterly*, 422–447.

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

——— (2008) "Average distance, diameter, and clustering in social networks with homophily," in *International Workshop on Internet and Network Economics*, 4–11, Springer.

——— (2010a) "An overview of social networks and economic applications," *Handbook of Social Economics*, 1, 511–85.

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

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

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

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

Jackson, Matthew O and Leeat Yariv (2007) "Diffusion of behavior and equilibrium properties in network games," *American Economic Review*, 97 (2), 92–98.

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

Kinnan, Cynthia and Robert Townsend (2012) "Kinship and financial networks, formal financial access, and risk reduction," *American Economic Review*, 102 (3), 289–293.

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

Kossinets, Gueorgi (2006) "Effects of missing data in social networks," *Social Networks*, 28 (3), 247–268.

Kovářík, Jaromír, Pablo Brañas-Garza, Ramón Cobo-Reyes, María Paz Espinosa, Natalia Jiménez, and Giovanni Ponti (2012) "Prosocial norms and degree heterogeneity in social networks," *Physica A Statistical Mechanics and its Applications*, 391, 849–853.

Kovářík, Jaromír, Pablo Brañas-Garza, Michael W Davidson, Dotan A Haim, Shannon Carcelli, and James H Fowler (2017) "Digit ratio (2D: 4D) and social integration: an effect of prenatal sex hormones," *Network Science*, 5 (4), 476–489.

Kovářík, Jaromír and Marco J Van der Leij (2014) "Risk aversion and social networks," *Review of Network Economics*, 13 (2), 121–155.

Kremer, Michael and Edward Miguel (2007) "The illusion of sustainability," *Quarterly Journal of Economics*, 122 (3), 1007–1065.

Lee, Sang Hoon, Pan-Jun Kim, and Hawoong Jeong (2006) "Statistical properties of sampled networks," *Physical Review E*, 73 (1), 016102.

Leider, Stephen, Markus M Möbius, Tanya Rosenblat, and Quoc-Anh Do (2009) "Directed altruism and enforced reciprocity in social networks," *Quarterly Journal of Economics*, 124 (4), 1815–1851.

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

Liu, Xiaodong (2013) "Estimation of a local-aggregate network model with sampled networks," *Economics Letters*, 118 (1), 243–246.

Lovász, László (2007) *Combinatorial Problems and Exercises*, 361: American Mathematical Soc.

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

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

Montgomery, James D (1991) "Social networks and labor-market outcomes: Toward an economic analysis," *American Economic Review*, 81 (5), 1408–1418.

Moody, James (2001) "Race, school integration, and friendship segregation in America," *American Journal of Sociology*, 107 (3), 679–716.

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

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

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

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

Ruiz-Palazuelos, Sofía (2021) "Clustering in network games," *Economics Letters*, 109922.

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

Snijders, Tom AB (1992) "Estimation on the basis of snowball samples: how to weight?" *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique*, 36 (1), 59–70.

Stork, Diana and William D Richards (1992) "Nonrespondents in communication network studies: Problems and possibilities," *Group & Organization Management*, 17 (2), 193–209.

Stumpf, Michael PH, Carsten Wiuf, and Robert M May (2005) "Subnets of scale-free networks are not scale-free: sampling properties of networks," *Proceedings of the National Academy of Sciences*, 102 (12), 4221–4224.

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

Toomet, Ott, Marco Van Der Leij, and Meredith Rolfe (2013) "Social networks and labor market inequality between ethnicities and races," *Network Science*, 1 (3), 321–352.

Valente, Thomas W (1996) "Network models of the diffusion of innovations," *Computational & Mathematical Organization Theory*, 2 (2), 163–164.

Valliant, Richard (1993) "Poststratification and conditional variance estimation," *Journal of the American Statistical Association*, 88 (421), 89–96.

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

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

Walker, Stephen G (2011) "Bounds for the second largest eigenvalue of a transition matrix," *Linear and Multilinear Algebra*, 59 (7), 755–760.

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

Whittington, Kjersten Bunker, Jason Owen-Smith, and Walter W Powell (2009) "Networks, propinquity, and innovation in knowledge-intensive industries," *Administrative Science Quarterly*, 54 (1), 90–122.

Wooldridge, Jeffrey M (2015) *Introductory Econometrics: A Modern Approach*: Nelson Education.

Yin, Hao, Austin R Benson, and Jure Leskovec (2018) "Higher-order clustering in networks," *Physical Review E*, 97 (5), 052306.

Zhang, Yaonan, Eric D Kolaczyk, Bruce D Spencer et al. (2015) "Estimating network degree distributions under sampling: An inverse problem, with applications to monitoring social media networks," *The Annals of Applied Statistics*, 9 (1), 166–199.

Supplementary Appendix

A Derivation of Corrections

In the following derivations, we denote $\binom{n}{m}$ the number of all possible samples of size m from a set of n nodes. We use $I(A)$ to denote an indicator function which equals to 1 if the condition A satisfies and 0 otherwise. We also use the standard "little-o" notation $o(1)$ to denote a term which converges to zero when the sample sizes of different types and the whole population go to infinity, holding fixed ψ_r and $\psi_{r,t}$ for each t.

A.1 Average degree

For induced subgraphs, the conditional expectation of the sample average degree, $d(G_r^{|s})$, is

$$
\begin{aligned}
&\mathrm{E}(d(G_r^{|s})|G_r) = \mathrm{E}\left(\frac{1}{m_r}\sum_{i\in S_r}\sum_{j\in S_r} W_{ij,r}^{|s}\middle| G_r\right)\\
&= \sum_{t=1}^{T}\left(\frac{1}{m_r}\sum_{i\in V_{r,t}}\sum_{j\in V_{r,t}} W_{ij,r}\mathrm{E}\left(I\left(i\in S_{r,t}, j\in S_{r,t}\right)\middle| G_r\right)\right) + \sum_{t=1}^{T}\sum_{\ell\neq t}^{T}\left(\frac{1}{m_r}\sum_{i\in V_{r,t}}\sum_{j\in V_{r,\ell}} W_{ij,r}\mathrm{E}\left(I\left(i\in S_{r,t}, j\in S_{r,\ell}\right)\middle| G_r\right)\right)\\
&= \sum_{t=1}^{T}\left[\frac{1}{m_r}\sum_{i\in V_{r,t}}\sum_{j\in V_{r,t}} W_{ij,r}\left(\frac{\binom{n_{r,t}-2}{m_{r,t}-2}\prod_{q\neq t}\binom{n_{r,q}}{m_{r,q}}}{\prod_{q=1}^{T}\binom{n_{r,q}}{m_{r,q}}}\right)\right] + \sum_{t=1}^{T}\sum_{\ell\neq t}^{T}\left[\frac{1}{m_r}\sum_{i\in V_{r,t}}\sum_{j\in V_{r,\ell}} W_{ij,r}\left(\frac{\binom{n_{r,t}-1}{m_{r,t}-1}\binom{n_{r,\ell}-1}{m_{r,\ell}-1}\prod_{q\neq t,\ell}\binom{n_{r,q}}{m_{r,q}}}{\prod_{q=1}^{T}\binom{n_{r,q}}{m_{r,q}}}\right)\right]\\
&= \frac{1}{n_r}\sum_{t=1}^{T}\left(\sum_{i\in V_{r,t}}\sum_{j\in V_{r,t}} W_{ij,r}\left(\frac{\psi_{r,t}\left(\psi_{r,t}+o(1)\right)}{\psi_r}\right)\right) + \frac{1}{n_r}\sum_{t=1}^{T}\sum_{\ell\neq t}^{T}\left(\sum_{i\in V_{r,t}}\sum_{j\in V_{r,\ell}} W_{ij,r}\left(\frac{\psi_{r,t}\psi_{r,\ell}}{\psi_r}\right)\right),
\end{aligned}
$$

where $o(1) = \frac{m_{r,t}-1}{n_{r,t}-1} - \frac{m_{r,t}}{n_{r,t}} = \frac{n_t-m_t}{n_t(n_t-1)}$.[1]

The conditional expectation of average degree for the star subgraph, $d(G_r^s)$, is

$$
\begin{aligned}
&\mathrm{E}(d(G_r^s)|G_r) = \mathrm{E}\left(\frac{1}{m_r}\sum_{i\in S_r}\sum_{j\in S_r} W_{ij,r}^{s}\middle| G_r\right)\\
&= \sum_{t=1}^{T}\left(\frac{1}{m_r}\sum_{i\in V_{r,t}}\sum_{j\in V_{r,t}} W_{ij,r}\mathrm{E}\left(I\left(\begin{smallmatrix} i\in S_{r,t} \text{ and} \\ j\in S_{r,t}\end{smallmatrix} \vee \begin{smallmatrix} i\in S_{r,t} \text{ or} \\ j\in S_{r,t}\end{smallmatrix}\right)\middle| G_r\right)\right) + \sum_{t=1}^{T}\sum_{\ell\neq t}^{T}\left(\frac{1}{m_r}\sum_{i\in V_{r,t}}\sum_{j\in V_{r,\ell}} W_{ij,r}\mathrm{E}\left(I\left(\begin{smallmatrix} i\in S_{r,t} \text{ and} \\ j\in S_{r,\ell}\end{smallmatrix} \vee \begin{smallmatrix} i\in S_{r,t} \text{ or} \\ j\in S_{r,\ell}\end{smallmatrix}\right)\middle| G_r\right)\right)\\
&= \sum_{t=1}^{T}\left[\frac{1}{m_r}\sum_{i\in V_{r,t}}\sum_{j\in V_{r,t}} W_{ij,r}\left(\frac{\left(\binom{n_{r,t}-2}{m_{r,t}-2} + 2\binom{n_{r,t}-2}{m_{r,t}-1}\right)\prod_{q\neq t}\binom{n_{r,q}}{m_{r,q}}}{\prod_{q=1}^{T}\binom{n_{r,q}}{m_{r,q}}}\right)\right]\\
&\quad + \sum_{t=1}^{T}\sum_{\ell\neq t}^{T}\left[\frac{1}{m_r}\sum_{i\in V_{r,t}}\sum_{j\in V_{r,\ell}} W_{ij,r}\left(\frac{\left(\binom{n_{r,t}-1}{m_{r,t}-1}\binom{n_{r,\ell}-1}{m_{r,\ell}-1} + \binom{n_{r,t}-1}{m_{r,t}}\binom{n_{r,\ell}-1}{m_{r,\ell}-1} + \binom{n_{r,t}-1}{m_{r,t}-1}\binom{n_{r,\ell}-1}{m_{r,\ell}}\right)\prod_{q\neq t,\ell}\binom{n_{r,q}}{m_{r,q}}}{\prod_{q=1}^{T}\binom{n_{r,q}}{m_{r,q}}}\right)\right]\\
&= \frac{1}{n_r}\sum_{t=1}^{T}\left(\sum_{i\in V_{r,t}}\sum_{j\in V_{r,t}} W_{ij,r}\left(\frac{\psi_{r,t}^2 + 2\psi_{r,t}(1-\psi_{r,t}) + o(1)}{\psi_r}\right)\right)\\
&\quad + \frac{1}{n_r}\sum_{t=1}^{T}\sum_{\ell\neq t}^{T}\left(\sum_{i\in V_{r,t}}\sum_{j\in V_{r,\ell}} W_{ij,r}\left(\frac{\psi_{r,t}\psi_{r,\ell} + \psi_{r,\ell}(1-\psi_{r,t}) + \psi_{r,t}(1-\psi_{r,\ell}) + o(1)}{\psi_r}\right)\right).
\end{aligned}
$$

[1] Since the form of $o(1)$ is very similar in the other cases discussed below, we only specify its exact form in this case.

A.2 Clustering coefficient

The clustering coefficient of a graph is a ratio of the number of triangles to the number of connected triples, calculated as $c(G_r) = \frac{\rho(G_r)}{\tau(G_r)}$, where

$$\rho(G_r) = \sum_{\substack{i\in V_r \\ i\neq j\neq k}} \sum_{\substack{j\in V_r \\ i\neq j\neq k}} \sum_{\substack{k\in V_r \\ i\neq j\neq k}} W_{ij,r}W_{jk,r}W_{ki,r} \quad \text{and} \quad \tau(G_r) = \sum_{\substack{i\in V_r \\ i\neq j\neq k}} \sum_{\substack{j\in V_r \\ i\neq j\neq k}} \sum_{\substack{k\in V_r \\ i\neq j\neq k}} W_{ij,r}W_{jk,r}.$$

For induced subgraphs, the conditional expectation of $\rho(G_r^{|s})$ can be decomposed as follows,

$$\begin{aligned}
&\mathrm{E}(\rho(G_r^{|s})|G_r) = \mathrm{E}\left(\sum_{\substack{i\in S_r \\ i\neq j\neq k}} \sum_{\substack{j\in S_r \\ i\neq j\neq k}} \sum_{\substack{k\in S_r \\ i\neq j\neq k}} W_{ij,r}^{|s}W_{jk,r}^{|s}W_{ki,r}^{|s} \middle| G_r \right) \\
&= \sum_{t=1}^{T} \left(\sum_{\substack{i\in V_{r,t} \\ i\neq j\neq k}} \sum_{\substack{j\in V_{r,t} \\ i\neq j\neq k}} \sum_{\substack{k\in V_{r,t} \\ i\neq j\neq k}} W_{ij,r}W_{jk,r}W_{ki,r}\mathrm{E}\left(I\left(i,j,k\in S_{r,t}\right)|\, G_r\right) \right) \\
&+ \sum_{t=1}^{T}\sum_{\ell\neq t}^{T} \left(\sum_{\substack{i\in V_{r,t} \\ i\neq j\neq k}} \sum_{\substack{j\in V_{r,t} \\ i\neq j\neq k}} \sum_{\substack{k\in V_{r,\ell} \\ i\neq j\neq k}} W_{ij,r}W_{jk,r}W_{ki,r}\mathrm{E}\left(I\left(i,j\in S_{r,t},k\in S_{r,\ell}\right)|\, G_r\right) \right) \\
&+ \sum_{t=1}^{T}\sum_{\ell\neq t}^{T} \left(\sum_{\substack{i\in V_{r,\ell} \\ i\neq j\neq k}} \sum_{\substack{j\in V_{r,t} \\ i\neq j\neq k}} \sum_{\substack{k\in V_{r,t} \\ i\neq j\neq k}} W_{ij,r}W_{jk,r}W_{ki,r}\mathrm{E}\left(I\left(i\in S_{r,\ell},j,k\in S_{r,t}\right)|\, G_r\right) \right) \\
&+ \sum_{t=1}^{T}\sum_{\ell\neq t}^{T} \left(\sum_{\substack{i\in V_{r,t} \\ i\neq j\neq k}} \sum_{\substack{j\in V_{r,\ell} \\ i\neq j\neq k}} \sum_{\substack{k\in V_{r,t} \\ i\neq j\neq k}} W_{ij,r}W_{jk,r}W_{ki,r}\mathrm{E}\left(I\left(i,k\in S_{r,t},j\in S_{r,\ell}\right)|\, G_r\right) \right) \\
&+ \sum_{t=1}^{T}\sum_{\ell\neq t}^{T}\sum_{h\neq t,\ell}^{T} \left(\sum_{\substack{i\in V_{r,t} \\ i\neq j\neq k}} \sum_{\substack{j\in V_{r,\ell} \\ i\neq j\neq k}} \sum_{\substack{k\in V_{r,h} \\ i\neq j\neq k}} W_{ij,r}W_{jk,r}W_{ki,r}\mathrm{E}\left(I\left(i\in S_{r,t},j\in S_{r,\ell},k\in S_{r,h}\right)|\, G_r\right) \right).
\end{aligned}$$

When three individuals are of the same type, i.e., $t_i = t_j = t_k = t$, we have

$$\mathrm{E}\left(I\left(i,j,k\in S_{r,t}\right)|\, G_r\right) = \left(\frac{\binom{n_{r,t}-3}{m_{r,t}-3}\prod_{q\neq t}\binom{n_{r,q}}{m_{r,q}}}{\prod_{q=1}^{T}\binom{n_{r,q}}{m_{r,q}}} \right).$$

When two individuals are of the same type and the third is of a different type, including $t_i = t_j = t$ and $t_k = \ell$; $t_j = t_k = t$ and $t_i = \ell$; and $t_i = t_k = t$ and $t_j = \ell$,

$$\begin{aligned}
&\mathrm{E}\left(I\left(i,j\in S_{r,t},k\in S_{r,\ell}\right)|\, G_r\right) = \mathrm{E}\left(I\left(i\in S_{r,\ell},j,k\in S_{r,t}\right)|\, G_r\right) = \mathrm{E}\left(I\left(i,k\in S_{r,t},j\in S_{r,\ell}\right)|\, G_r\right) \\
&= \left(\frac{\binom{n_{r,t}-2}{m_{r,t}-2}\binom{n_{r,\ell}-1}{m_{r,\ell}-1}\prod_{q\neq t,\ell}\binom{n_{r,q}}{m_{r,q}}}{\prod_{q=1}^{T}\binom{n_{r,q}}{m_{r,q}}} \right).
\end{aligned}$$

When three individuals are of different types, i.e., $t_i = t, t_j = \ell, t_k = h$,

$$\mathrm{E}\left(I\left(i\in S_{r,t},j\in S_{r,\ell},k\in S_{r,h}\right)|\, G_r\right) = \left(\frac{\binom{n_{r,t}-1}{m_{r,t}-1}\binom{n_{r,\ell}-1}{m_{r,\ell}-1}\binom{n_{r,h}-1}{m_{r,h}-1}\prod_{q\neq t,\ell,h}\binom{n_{r,q}}{m_{r,q}}}{\prod_{q=1}^{T}\binom{n_{r,q}}{m_{r,q}}} \right).$$

Therefore, we can further obtain

$$
\begin{aligned}
&\mathrm{E}(\rho(G_r^{|s})|G_r) = \\
&= \sum_{t=1}^{T}\left(\sum_{i\in V_{r,t}}\sum_{\substack{j\in V_{r,t}\\ i\neq j\neq k}}\sum_{k\in V_{r,t}} W_{ij,r}W_{jk,r}W_{ki,r}\left(\psi_{r,t}^3+o(1)\right)\right)+\sum_{t=1}^{T}\sum_{\ell\neq t}^{T}\left(\sum_{i\in V_{r,t}}\sum_{\substack{j\in V_{r,t}\\ i\neq j\neq k}}\sum_{k\in V_{r,\ell}} W_{ij,r}W_{jk,r}W_{ki,r}\left(\psi_{r,t}^2\psi_{r,\ell}+o(1)\right)\right)\\
&+\sum_{t=1}^{T}\sum_{\ell\neq t}^{T}\left(\sum_{i\in V_{r,\ell}}\sum_{\substack{j\in V_{r,t}\\ i\neq j\neq k}}\sum_{k\in V_{r,t}} W_{ij,r}W_{jk,r}W_{ki,r}\left(\psi_{r,t}^2\psi_{r,\ell}+o(1)\right)\right)+\sum_{t=1}^{T}\sum_{\ell\neq t}^{T}\left(\sum_{i\in V_{r,t}}\sum_{\substack{j\in V_{r,\ell}\\ i\neq j\neq k}}\sum_{k\in V_{r,t}} W_{ij,r}W_{jk,r}W_{ki,r}\left(\psi_{r,t}^2\psi_{r,\ell}+o(1)\right)\right)\\
&+\sum_{t=1}^{T}\sum_{\ell\neq t}^{T}\sum_{h\neq t,\ell}^{T}\left(\sum_{i\in V_{r,t}}\sum_{\substack{j\in V_{r,\ell}\\ i\neq j\neq k}}\sum_{k\in V_{r,h}} W_{ij,r}W_{jk,r}W_{ki,r}\left(\psi_{r,t}\psi_{r,\ell}\psi_{r,h}\right)\right).
\end{aligned}
$$

As a result, we propose the following weighted estimator for the number of triangles,

$$
\begin{aligned}
&\widetilde{\rho}(G_r^{|s}) = \\
&\sum_{t=1}^{T}\left(\sum_{i\in S_{r,t}}\sum_{\substack{j\in S_{r,t}\\ i\neq j\neq k}}\sum_{k\in S_{r,t}} W_{ij,r}^{|s}W_{jk,r}^{|s}W_{ki,r}^{|s}\left(\psi_{r,t}^3\right)^{-1}\right)+\sum_{t=1}^{T}\sum_{\ell\neq t}^{T}\left(\sum_{i\in S_{r,t}}\sum_{\substack{j\in S_{r,t}\\ i\neq j\neq k}}\sum_{k\in S_{r,\ell}} W_{ij,r}^{|s}W_{jk,r}^{|s}W_{ki,r}^{|s}\left(\psi_{r,t}^2\psi_{r,\ell}\right)^{-1}\right)\\
&+\sum_{t=1}^{T}\sum_{\ell\neq t}^{T}\left(\sum_{i\in S_{r,\ell}}\sum_{\substack{j\in S_{r,t}\\ i\neq j\neq k}}\sum_{k\in S_{r,t}} W_{ij,r}^{|s}W_{jk,r}^{|s}W_{ki,r}^{|s}\left(\psi_{r,t}^2\psi_{r,\ell}\right)^{-1}\right)\\
&+\sum_{t=1}^{T}\sum_{\ell\neq t}^{T}\left(\sum_{i\in S_{r,t}}\sum_{\substack{j\in S_{r,\ell}\\ i\neq j\neq k}}\sum_{k\in S_{r,t}} W_{ij,r}^{|s}W_{jk,r}^{|s}W_{ki,r}^{|s}\left(\psi_{r,t}^2\psi_{r,\ell}\right)^{-1}\right)\\
&+\sum_{t=1}^{T}\sum_{\ell\neq t}^{T}\sum_{h\neq t,\ell}^{T}\left(\sum_{i\in S_{r,t}}\sum_{\substack{j\in S_{r,\ell}\\ i\neq j\neq k}}\sum_{k\in S_{r,h}} W_{ij,r}^{|s}W_{jk,r}^{|s}W_{ki,r}^{|s}\left(\psi_{r,t}\psi_{r,\ell}\psi_{r,h}\right)^{-1}\right).
\end{aligned}
$$

The conditional expectation of the denominator $\tau(G_r^{|s})$ has the following form:

$$
\begin{aligned}
\mathrm{E}(\tau(G_r^{|s})|G_r) &= \mathrm{E}\left(\sum_{\substack{i\in S_r \\ i\neq j\neq k}} \sum_{j\in S_r} \sum_{k\in S_r} W_{ij,r}^{|s} W_{jk,r}^{|s} \,\middle|\, G_r \right) \\
&= \sum_{t=1}^{T} \left(\sum_{\substack{i\in V_{r,t} \\ i\neq j\neq k}} \sum_{j\in V_{r,t}} \sum_{k\in V_{r,t}} W_{ij,r} W_{jk,r} \mathrm{E}\left(I\left(i,j,k\in S_{r,t}\right) | \, G_r\right) \right) \\
&+ \sum_{t=1}^{T}\sum_{\ell\neq t}^{T} \left(\sum_{\substack{i\in V_{r,t} \\ i\neq j\neq k}} \sum_{j\in V_{r,t}} \sum_{k\in V_{r,\ell}} W_{ij,r} W_{jk,r} \mathrm{E}\left(I\left(i,j\in S_{r,t},k\in S_{r,\ell}\right) | \, G_r\right) \right) \\
&+ \sum_{t=1}^{T}\sum_{\ell\neq t}^{T} \left(\sum_{\substack{i\in V_{r,\ell} \\ i\neq j\neq k}} \sum_{j\in V_{r,t}} \sum_{k\in V_{r,t}} W_{ij,r} W_{jk,r} \mathrm{E}\left(I\left(i\in S_{r,\ell},j,k\in S_{r,t}\right) | \, G_r\right) \right) \\
&+ \sum_{t=1}^{T}\sum_{\ell\neq t}^{T} \left(\sum_{\substack{i\in V_{r,t} \\ i\neq j\neq k}} \sum_{j\in V_{r,\ell}} \sum_{k\in V_{r,t}} W_{ij,r} W_{jk,r} \mathrm{E}\left(I\left(i,k\in S_{r,t},j\in S_{r,\ell}\right) | \, G_r\right) \right) \\
&+ \sum_{t=1}^{T}\sum_{\ell\neq t}^{T}\sum_{h\neq t,\ell}^{T} \left(\sum_{\substack{i\in V_{r,t} \\ i\neq j\neq k}} \sum_{j\in V_{r,\ell}} \sum_{k\in V_{r,h}} W_{ij,r} W_{jk,r} \mathrm{E}\left(I\left(i\in S_{r,t},j\in S_{r,\ell},k\in S_{r,h}\right) | \, G_r\right) \right).
\end{aligned}
$$

When three individuals are of the same type, i.e., $t_i = t_j = t_k = t$, we have

$$
\mathrm{E}\left(I\left(i,j,k\in S_{r,t}\right) | \, G_r\right) = \left(\frac{\binom{n_{r,t}-3}{m_{r,t}-3} \prod_{q\neq t} \binom{n_{r,q}}{m_{r,q}}}{\prod_{q=1}^{T} \binom{n_{r,q}}{m_{r,q}}} \right).
$$

When two individuals are of the same type and the third is of a different type, including $t_i = t_j = t$ and $t_k = \ell$; $t_j = t_k = t$ and $t_i = \ell$; and $t_i = t_k = t$ and $t_j = \ell$,

$$
\begin{aligned}
&\mathrm{E}\left(I\left(i,j\in S_{r,t},k\in S_{r,\ell}\right) | \, G_r\right) = \mathrm{E}\left(I\left(i\in S_{r,\ell},j,k\in S_{r,t}\right) | \, G_r\right) = \mathrm{E}\left(I\left(i,k\in S_{r,t},j\in S_{r,\ell}\right) | \, G_r\right) \\
&= \left(\frac{\binom{n_{r,t}-2}{m_{r,t}-2}\binom{n_{r,\ell}-1}{m_{r,\ell}-1} \prod_{q\neq t,\ell} \binom{n_{r,q}}{m_{r,q}}}{\prod_{q=1}^{T} \binom{n_{r,q}}{m_{r,q}}} \right).
\end{aligned}
$$

When three individuals are of different types, i.e., $t_i = t, t_j = \ell, t_k = h$,

$$
\mathrm{E}\left(I\left(i\in S_{r,t},j\in S_{r,\ell},k\in S_{r,h}\right) | \, G_r\right) = \left(\frac{\binom{n_{r,t}-1}{m_{r,t}-1}\binom{n_{r,\ell}-1}{m_{r,\ell}-1}\binom{n_{r,h}-1}{m_{r,h}-1} \prod_{q\neq t,\ell,h} \binom{n_{r,q}}{m_{r,q}}}{\prod_{q=1}^{T} \binom{n_{r,q}}{m_{r,q}}} \right).
$$

Therefore,

$$
\begin{aligned}
&\mathrm{E}(\tau(G_r^{|s})|G_r) = \\
&\sum_{t=1}^{T}\left(\sum_{i\in V_{r,t}}\sum_{\substack{j\in V_{r,t}\\ i\neq j\neq k}}\sum_{k\in V_{r,t}} W_{ij,r}W_{jk,r}\left(\psi_{r,t}^3+o(1)\right)\right)+\sum_{t=1}^{T}\sum_{\ell\neq t}^{T}\left(\sum_{i\in V_{r,t}}\sum_{\substack{j\in V_{r,t}\\ i\neq j\neq k}}\sum_{k\in V_{r,\ell}} W_{ij,r}W_{jk,r}\left(\psi_{r,t}^2\psi_{r,\ell}+o(1)\right)\right)\\
&+\sum_{t=1}^{T}\sum_{\ell\neq t}^{T}\left(\sum_{i\in V_{r,\ell}}\sum_{\substack{j\in V_{r,t}\\ i\neq j\neq k}}\sum_{k\in V_{r,t}} W_{ij,r}W_{jk,r}\left(\psi_{r,t}^2\psi_{r,\ell}+o(1)\right)\right)+\sum_{t=1}^{T}\sum_{\ell\neq t}^{T}\left(\sum_{i\in V_{r,t}}\sum_{\substack{j\in V_{r,\ell}\\ i\neq j\neq k}}\sum_{k\in V_{r,t}} W_{ij,r}W_{jk,r}\left(\psi_{r,t}^2\psi_{r,\ell}+o(1)\right)\right)\\
&+\sum_{t=1}^{T}\sum_{\ell\neq t}^{T}\sum_{h\neq t,\ell}^{T}\left(\sum_{i\in V_{r,t}}\sum_{\substack{j\in V_{r,\ell}\\ i\neq j\neq k}}\sum_{k\in V_{r,h}} W_{ij,r}W_{jk,r}\left(\psi_{r,t}\psi_{r,\ell}\psi_{r,h}\right)\right).
\end{aligned}
$$

The resulting conditional expectation of $c(G_r^{|s})$ is $\mathrm{E}(c(G_r^{|s}|G_r)) = \mathrm{E}(\rho(G_r^{|s})|G_r)/\mathrm{E}(\tau(G_r^{|s})|G_r)$. If $\psi_{r,t} = \psi_{r,\ell} = \psi_{r,h}$, the above result will collapse to $\mathrm{E}(\rho(G_r^{|s}|G_r)) = (\psi_r^3 + o(1))\rho(G_r)$ and $\mathrm{E}(\tau(G_r^{|s}|G_r)) = (\psi_r^3 + o(1))\tau(G_r)$ and $\mathrm{E}(c(G_r^{|s})|G_r) = c(G_r)$, the results derived in Chandrasekhar and Lewis (2016). However, $c(G_r^{|s})$ would be a biased estimator of $c(G_r)$ in non-representative samples.

We propose the following weighted estimator for the number of connected triples,

$$
\begin{aligned}
&\widetilde{\tau}(G_r^{|s}) = \\
&\sum_{t=1}^{T}\left(\sum_{i\in S_{r,t}}\sum_{\substack{j\in S_{r,t}\\ i\neq j\neq k}}\sum_{k\in S_{r,t}} W_{ij,r}^{|s}W_{jk,r}^{|s}\left(\psi_{r,t}^3\right)^{-1}\right)+\sum_{t=1}^{T}\sum_{\ell\neq t}^{T}\left(\sum_{i\in S_{r,t}}\sum_{\substack{j\in S_{r,t}\\ i\neq j\neq k}}\sum_{k\in S_{r,\ell}} W_{ij,r}^{|s}W_{jk,r}^{|s}\left(\psi_{r,t}^2\psi_{r,\ell}\right)^{-1}\right)\\
&+\sum_{t=1}^{T}\sum_{\ell\neq t}^{T}\left(\sum_{i\in S_{r,\ell}}\sum_{\substack{j\in S_{r,t}\\ i\neq j\neq k}}\sum_{k\in S_{r,t}} W_{ij,r}^{|s}W_{jk,r}^{|s}\left(\psi_{r,t}^2\psi_{r,\ell}\right)^{-1}\right)+\sum_{t=1}^{T}\sum_{\ell\neq t}^{T}\left(\sum_{i\in S_{r,t}}\sum_{\substack{j\in S_{r,\ell}\\ i\neq j\neq k}}\sum_{k\in S_{r,t}} W_{ij,r}^{|s}W_{jk,r}^{|s}\left(\psi_{r,t}^2\psi_{r,\ell}\right)^{-1}\right)\\
&+\sum_{t=1}^{T}\sum_{\ell\neq t}^{T}\sum_{h\neq t,\ell}^{T}\left(\sum_{i\in S_{r,t}}\sum_{\substack{j\in S_{r,\ell}\\ i\neq j\neq k}}\sum_{k\in S_{r,h}} W_{ij,r}^{|s}W_{jk,r}^{|s}\left(\psi_{r,t}\psi_{r,\ell}\psi_{r,h}\right)^{-1}\right).
\end{aligned}
$$

For star subgraphs, the conditional expectation of the sample number of triangles is

$$
\begin{aligned}
\mathrm{E}(\rho(G_r^s)|G_r) &= \mathrm{E}\left(\sum_{\substack{i\in S_r \\ i\neq j\neq k}} \sum_{j\in S_r} \sum_{k\in S_r} W_{ij,r}^s W_{jk,r}^s W_{ki,r}^s \middle| G_r \right) \\
&= \sum_{t=1}^{T} \left(\sum_{\substack{i\in V_{r,t} \\ i\neq j\neq k}} \sum_{j\in V_{r,t}} \sum_{k\in V_{r,t}} W_{ij,r}W_{jk,r}W_{ki,r} \mathrm{E}\left(I\left(i,j,k \in S_{r,t} \vee {}^{\text{any two of } i,j,k\in S_r}_{t_i=t_j=t_k=t} \right) \middle| G_r \right) \right) \\
&\quad + \sum_{t=1}^{T}\sum_{\ell\neq t}^{T} \left(\sum_{\substack{i\in V_{r,t} \\ i\neq j\neq k}} \sum_{j\in V_{r,t}} \sum_{k\in V_{r,\ell}} W_{ij,r}W_{jk,r}W_{ki,r} \mathrm{E}\left(I\left({}^{i,j\in S_{r,t}}_{k\in S_{r,\ell}} \vee {}^{\text{any two of } i,j,k\in S_r}_{t_i=t_j=t,t_k=\ell} \right) \middle| G_r \right) \right) \\
&\quad + \sum_{t=1}^{T}\sum_{\ell\neq t}^{T} \left(\sum_{\substack{i\in V_{r,\ell} \\ i\neq j\neq k}} \sum_{j\in V_{r,t}} \sum_{k\in V_{r,t}} W_{ij,r}W_{jk,r}W_{ki,r} \mathrm{E}\left(I\left({}^{i\in S_{r,\ell}}_{j,k\in S_{r,t}} \vee {}^{\text{any two of } i,j,k\in S_r}_{t_j=t_k=t,t_i=\ell} \right) \middle| G_r \right) \right) \\
&\quad + \sum_{t=1}^{T}\sum_{\ell\neq t}^{T} \left(\sum_{\substack{i\in V_{r,t} \\ i\neq j\neq k}} \sum_{j\in V_{r,\ell}} \sum_{k\in V_{r,t}} W_{ij,r}W_{jk,r}W_{ki,r} \mathrm{E}\left(I\left({}^{i,k\in S_{r,t}}_{j\in S_{r,\ell}} \vee {}^{\text{any two of } i,j,k\in S_r}_{t_i=t_k=t,t_j=\ell} \right) \middle| G_r \right) \right) \\
&\quad + \sum_{t=1}^{T}\sum_{\ell\neq t}^{T}\sum_{h\neq t,\ell}^{T} \left(\sum_{\substack{i\in V_{r,t} \\ i\neq j\neq k}} \sum_{j\in V_{r,\ell}} \sum_{k\in V_{r,h}} W_{ij,r}W_{jk,r}W_{ki,r} \mathrm{E}\left(I\left(i\in S_{r,t}, j\in S_{r,\ell}, k\in S_{r,h} \vee {}^{\text{any two of } i,j,k\in S_r}_{t_i=t,t_j=\ell,t_k=h} \right) \middle| G_r \right) \right).
\end{aligned}
$$

When three individuals are of the same type, i.e., $t_i = t_j = t_k = t$, we have

$$
\mathrm{E}\left(I\left(i,j,k \in S_{r,t} \vee {}^{\text{any two of } i,j,k\in S_r}_{t_i=t_j=t_k=t} \right) \middle| G_r \right) = \left(\frac{\left(\binom{n_{r,t}-3}{m_{r,t}-3} + 3\binom{n_{r,t}-3}{m_{r,t}-2} \right) \prod_{q\neq t} \binom{n_{r,q}}{m_{r,q}}}{\prod_{q=1}^{T} \binom{n_{r,q}}{m_{r,q}}} \right).
$$

When two individuals are of the same type and the third is of a different type, including $t_i = t_j = t$ and $t_k = \ell$; $t_j = t_k = t$ and $t_i = \ell$; and $t_i = t_k = t$ and $t_j = \ell$,

$$
\begin{aligned}
&\mathrm{E}\left(I\left({}^{i,j\in S_{r,t}}_{k\in S_{r,\ell}} \vee {}^{\text{any two of } i,j,k\in S_r}_{t_i=t_j=t,t_k=\ell} \right) \middle| G_r \right) = \mathrm{E}\left(I\left({}^{i\in S_{r,\ell}}_{j,k\in S_{r,t}} \vee {}^{\text{any two of } i,j,k\in S_r}_{t_j=t_k=t,t_i=\ell} \right) \middle| G_r \right) \\
=&\mathrm{E}\left(I\left({}^{i,k\in S_{r,t}}_{j\in S_{r,\ell}} \vee {}^{\text{any two of } i,j,k\in S_r}_{t_i=t_k=t,t_j=\ell} \right) \middle| G_r \right) = \left(\frac{\left(\binom{n_{r,t}-2}{m_{r,t}-2}\binom{n_{r,\ell}-1}{m_{r,\ell}-1} + 2\binom{n_{r,t}-2}{m_{r,t}-1}\binom{n_{r,\ell}-1}{m_{r,\ell}-1} + \binom{n_{r,t}-2}{m_{r,t}-2}\binom{n_{r,\ell}-1}{m_{r,\ell}} \right) \prod_{q\neq t,\ell} \binom{n_{r,q}}{m_{r,q}}}{\prod_{q=1}^{T} \binom{n_{r,q}}{m_{r,q}}} \right).
\end{aligned}
$$

When three individuals are of different types, i.e., $t_i = t, t_j = \ell, t_k = h$,

$$
\begin{aligned}
&\mathrm{E}\left(I\left(i\in S_{r,t}, j\in S_{r,\ell}, k\in S_{r,h} \vee {}^{\text{any two of } i,j,k\in S_r}_{t_i=t,t_j=\ell,t_k=h} \right) \middle| G_r \right) \\
&= \left(\prod_{q=1}^{T} \binom{n_{r,q}}{m_{r,q}} \right)^{-1} \left[\left(\binom{n_{r,t}-1}{m_{r,t}-1}\binom{n_{r,\ell}-1}{m_{r,\ell}-1}\binom{n_{r,h}-1}{m_{r,h}-1} + \binom{n_{r,t}-1}{m_{r,t}-1}\binom{n_{r,\ell}-1}{m_{r,\ell}}\binom{n_{r,h}-1}{m_{r,h}-1} \right.\right. \\
&\quad \left.\left. + \binom{n_{r,t}-1}{m_{r,t}}\binom{n_{r,\ell}-1}{m_{r,\ell}-1}\binom{n_{r,h}-1}{m_{r,h}-1} + \binom{n_{r,t}-1}{m_{r,t}-1}\binom{n_{r,\ell}-1}{m_{r,\ell}-1}\binom{n_{r,h}-1}{m_{r,h}} \right) \prod_{q\neq t,\ell,h} \binom{n_{r,q}}{m_{r,q}} \right]
\end{aligned}
$$

Therefore, we can further obtain

$$
\begin{aligned}
\mathrm{E}(\rho(G_r^s)|G_r) = &\sum_{t=1}^{T}\left(\sum_{i\in V_{r,t}}\sum_{\substack{j\in V_{r,t}\\ i\neq j\neq k}}\sum_{k\in V_{r,t}} W_{ij,r}W_{jk,r}W_{ki,r}\left(\psi_{r,t}^3+3\psi_{r,t}^2(1-\psi_{r,t})+o(1)\right)\right)\\
&+\sum_{t=1}^{T}\sum_{\ell\neq t}^{T}\left(\sum_{i\in V_{r,t}}\sum_{\substack{j\in V_{r,t}\\ i\neq j\neq k}}\sum_{k\in V_{r,\ell}} W_{ij,r}W_{jk,r}W_{ki,r}\left(\psi_{r,t}^2\psi_{r,\ell}+2(1-\psi_{r,t})\psi_{r,t}\psi_{r,\ell}+\psi_{r,t}^2(1-\psi_{r,\ell})+o(1)\right)\right)\\
&+\sum_{t=1}^{T}\sum_{\ell\neq t}^{T}\left(\sum_{i\in V_{r,\ell}}\sum_{\substack{j\in V_{r,t}\\ i\neq j\neq k}}\sum_{k\in V_{r,t}} W_{ij,r}W_{jk,r}W_{ki,r}\left(\psi_{r,t}^2\psi_{r,\ell}+2(1-\psi_{r,t})\psi_{r,t}\psi_{r,\ell}+\psi_{r,t}^2(1-\psi_{r,\ell})+o(1)\right)\right)\\
&+\sum_{t=1}^{T}\sum_{\ell\neq t}^{T}\left(\sum_{i\in V_{r,t}}\sum_{\substack{j\in V_{r,\ell}\\ i\neq j\neq k}}\sum_{k\in V_{r,t}} W_{ij,r}W_{jk,r}W_{ki,r}\left(\psi_{r,t}^2\psi_{r,\ell}+2(1-\psi_{r,t})\psi_{r,t}\psi_{r,\ell}+\psi_{r,t}^2(1-\psi_{r,\ell})+o(1)\right)\right)\\
&+\sum_{t=1}^{T}\sum_{\ell\neq t}^{T}\sum_{h\neq t,\ell}^{T}\left(\sum_{i\in V_{r,t}}\sum_{\substack{j\in V_{r,\ell}\\ i\neq j\neq k}}\sum_{k\in V_{r,h}} W_{ij,r}W_{jk,r}W_{ki,r}\left(\psi_{r,t}\psi_{r,\ell}\psi_{r,h}+(1-\psi_{r,t})\psi_{r,\ell}\psi_{r,h}+\psi_{r,t}(1-\psi_{r,\ell})\psi_{r,h}+\psi_{r,t}\psi_{r,\ell}(1-\psi_{r,h})\right)\right).
\end{aligned}
$$

We propose the following weighted estimator for the number of triangles for the star subgraph:

$$
\begin{aligned}
\tilde{\rho}(G_r^s) = &\sum_{t=1}^{T}\left(\sum_{i\in S_{r,t}}\sum_{\substack{j\in S_{r,t}\\ i\neq j\neq k}}\sum_{k\in S_{r,t}} W_{ij,r}^sW_{jk,r}^sW_{ki,r}^s\left(\psi_{r,t}^3+3\psi_{r,t}^2(1-\psi_{r,t})\right)^{-1}\right)\\
&+\sum_{t=1}^{T}\sum_{\ell\neq t}^{T}\left(\sum_{i\in S_{r,t}}\sum_{\substack{j\in S_{r,t}\\ i\neq j\neq k}}\sum_{k\in S_{r,\ell}} W_{ij,r}^sW_{jk,r}^sW_{ki,r}^s\left(\psi_{r,t}^2\psi_{r,\ell}+2(1-\psi_{r,t})\psi_{r,t}\psi_{r,\ell}+\psi_{r,t}^2(1-\psi_{r,\ell})\right)^{-1}\right)\\
&+\sum_{t=1}^{T}\sum_{\ell\neq t}^{T}\left(\sum_{i\in S_{r,\ell}}\sum_{\substack{j\in S_{r,t}\\ i\neq j\neq k}}\sum_{k\in S_{r,t}} W_{ij,r}^sW_{jk,r}^sW_{ki,r}^s\left(\psi_{r,t}^2\psi_{r,\ell}+2(1-\psi_{r,t})\psi_{r,t}\psi_{r,\ell}+\psi_{r,t}^2(1-\psi_{r,\ell})\right)^{-1}\right)\\
&+\sum_{t=1}^{T}\sum_{\ell\neq t}^{T}\left(\sum_{i\in S_{r,t}}\sum_{\substack{j\in S_{r,\ell}\\ i\neq j\neq k}}\sum_{k\in S_{r,t}} W_{ij,r}^sW_{jk,r}^sW_{ki,r}^s\left(\psi_{r,t}^2\psi_{r,\ell}+2(1-\psi_{r,t})\psi_{r,t}\psi_{r,\ell}+\psi_{r,t}^2(1-\psi_{r,\ell})\right)^{-1}\right)\\
&+\sum_{t=1}^{T}\sum_{\ell\neq t}^{T}\sum_{h\neq t,\ell}^{T}\left(\sum_{i\in S_{r,t}}\sum_{\substack{j\in S_{r,\ell}\\ i\neq j\neq k}}\sum_{k\in S_{r,h}} W_{ij,r}^sW_{jk,r}^sW_{ki,r}^s\Big(\psi_{r,t}\psi_{r,\ell}\psi_{r,h}+(1-\psi_{r,t})\psi_{r,\ell}\psi_{r,h}+\psi_{r,t}(1-\psi_{r,\ell})\psi_{r,h}\right.\\
&\qquad\left.+\psi_{r,t}\psi_{r,\ell}(1-\psi_{r,h})\Big)^{-1}\right).
\end{aligned}
$$

The conditional expectation of $\tau(G_r^s)$ is:

$$
\begin{aligned}
\mathrm{E}(\tau(G_r^s)|G_r) &= \mathrm{E}\left(\sum_{\substack{i\in S_r \\ i\neq j\neq k}} \sum_{j\in S_r} \sum_{k\in S_r} W_{ij,r}^s W_{jk,r}^s \middle| G_r \right) \\
&= \sum_{t=1}^{T} \left(\sum_{\substack{i\in V_{r,t} \\ i\neq j\neq k}} \sum_{j\in V_{r,t}} \sum_{k\in V_{r,t}} W_{ij,r} W_{jk,r} \mathrm{E}\left(I\left(i,j,k \in S_{r,t} \vee {}^{\text{any two of } i,j,k\in S_r}_{t_i=t_j=t_k=t} \vee {}^{\text{only } j\in S_r}_{t_i=t_j=t_k=t}\right) \middle| G_r \right) \right) \\
&\quad + \sum_{t=1}^{T}\sum_{\ell\neq t}^{T} \left(\sum_{\substack{i\in V_{r,t} \\ i\neq j\neq k}} \sum_{j\in V_{r,t}} \sum_{k\in V_{r,\ell}} W_{ij,r} W_{jk,r} \mathrm{E}\left(I\left({}^{i,j\in S_{r,t}}_{k\in S_{r,\ell}} \vee {}^{\text{any two of } i,j,k\in S_r}_{t_i=t_j=t,t_k=\ell} \vee {}^{\text{only } j\in S_r}_{t_i=t_j=t,t_k=\ell}\right) \middle| G_r \right) \right) \\
&\quad + \sum_{t=1}^{T}\sum_{\ell\neq t}^{T} \left(\sum_{\substack{i\in V_{r,\ell} \\ i\neq j\neq k}} \sum_{j\in V_{r,t}} \sum_{k\in V_{r,t}} W_{ij,r} W_{jk,r} \mathrm{E}\left(I\left({}^{i\in S_{r,\ell}}_{j,k\in S_{r,t}} \vee {}^{\text{any two of } i,j,k\in S_r}_{t_j=t_k=t,t_i=\ell} \vee {}^{\text{only } j\in S_r}_{t_j=t_k=t,t_i=\ell}\right) \middle| G_r \right) \right) \\
&\quad + \sum_{t=1}^{T}\sum_{\ell\neq t}^{T} \left(\sum_{\substack{i\in V_{r,t} \\ i\neq j\neq k}} \sum_{j\in V_{r,\ell}} \sum_{k\in V_{r,t}} W_{ij,r} W_{jk,r} \mathrm{E}\left(I\left({}^{i,k\in S_{r,t}}_{j\in S_{r,\ell}} \vee {}^{\text{any two of } i,j,k\in S_r}_{t_i=t_k=t,t_j=\ell} \vee {}^{\text{only } j\in S_r}_{t_i=t_k=t,t_j=\ell}\right) \middle| G_r \right) \right) \\
&\quad + \sum_{t=1}^{T}\sum_{\ell\neq t}^{T}\sum_{h\neq t,\ell}^{T} \left(\sum_{\substack{i\in V_{r,t} \\ i\neq j\neq k}} \sum_{j\in V_{r,\ell}} \sum_{k\in V_{r,h}} W_{ij,r} W_{jk,r} \mathrm{E}\left(I\left({}^{i\in S_{r,t},j\in S_{r,\ell},k\in S_{r,h}}_{t_i=t,t_j=\ell,t_k=h} \vee {}^{\text{only } j\in S_r}_{t_i=t,t_j=\ell,t_k=h}\right) \middle| G_r \right) \right).
\end{aligned}
$$

When the three individuals involved are of the same type, i.e., $t_i = t_j = t_k = t$, we have

$$
\mathrm{E}\left(I\left(i,j,k \in S_{r,t} \vee {}^{\text{any two of } i,j,k\in S_r}_{t_i=t_j=t_k=t} \vee {}^{\text{only } j\in S_r}_{t_i=t_j=t_k=t}\right) \middle| G_r \right) = \left(\frac{\left(\binom{n_{r,t}-3}{m_{r,t}-3} + 3\binom{n_{r,t}-3}{m_{r,t}-2} + \binom{n_{r,t}-3}{m_{r,t}-1} \right) \prod_{q\neq t} \binom{n_{r,q}}{m_{r,q}}}{\prod_{q=1}^{T} \binom{n_{r,q}}{m_{r,q}}} \right).
$$

When two individuals are of the same type and the third is of a different type that $t_i = t_j = t$ and $t_k = \ell$ or $t_j = t_k = t$ and $t_i = \ell$

$$
\begin{aligned}
&\mathrm{E}\left(I\left({}^{i,j\in S_{r,t}}_{k\in S_{r,\ell}} \vee {}^{\text{any two of } i,j,k\in S_r}_{t_i=t_j=t,t_k=\ell} \vee {}^{\text{only } j\in S_r}_{t_i=t_j=t,t_k=\ell}\right) \middle| G_r \right) \\
&= \mathrm{E}\left(I\left({}^{i\in S_{r,\ell}}_{j,k\in S_{r,t}} \vee {}^{\text{any two of } i,j,k\in S_r}_{t_j=t_k=t,t_i=\ell} \vee {}^{\text{only } j\in S_r}_{t_j=t_k=t,t_i=\ell}\right) \middle| G_r \right) \\
&= \left(\frac{\left(\binom{n_{r,t}-2}{m_{r,t}-2}\binom{n_{r,\ell}-1}{m_{r,\ell}-1} + 2\binom{n_{r,t}-2}{m_{r,t}-1}\binom{n_{r,\ell}-1}{m_{r,\ell}-1} + \binom{n_{r,t}-2}{m_{r,t}-2}\binom{n_{r,\ell}-1}{m_{r,\ell}} + \binom{n_{r,t}-2}{m_{r,t}-1}\binom{n_{r,\ell}-1}{m_{r,\ell}} \right) \prod_{t\neq t,\ell} \binom{n_{r,q}}{m_{r,q}}}{\prod_{q=1}^{T} \binom{n_{r,q}}{m_{r,q}}} \right);
\end{aligned}
$$

and for $t_i = t_k = t$ and $t_j = \ell$,

$$
\begin{aligned}
&\mathrm{E}\left(I\left({}^{i,k\in S_{r,t}}_{j\in Sr,\ell} \vee {}^{\text{any two of } i,j,k\in S_r}_{t_i=t_k=t,t_j=\ell} \vee {}^{\text{only } j\in S_r}_{t_i=t_k=t,t_j=\ell}\right) \middle| G_r \right) \\
&= \left(\frac{\left(\binom{n_{r,t}-2}{m_{r,t}-2}\binom{n_{r,\ell}-1}{m_{r,\ell}-1} + 2\binom{n_{r,t}-2}{m_{r,t}-1}\binom{n_{r,\ell}-1}{m_{r,\ell}-1} + \binom{n_{r,t}-2}{m_{r,t}-2}\binom{n_{r,\ell}-1}{m_{r,\ell}} + \binom{n_{r,t}-2}{m_{r,t}}\binom{n_{r,\ell}-1}{m_{r,\ell}-1} \right) \prod_{t\neq t,\ell} \binom{n_{r,q}}{m_{r,q}}}{\prod_{q=1}^{T} \binom{n_{r,q}}{m_{r,q}}} \right)
\end{aligned}
$$

Last, when the three individuals are of different types, i.e., $t_i = t, t_j = \ell, t_k = h$,

$$
\begin{aligned}
&\mathrm{E}\left(I\left(i\in S_{r,t}, j\in S_{r,\ell}, k\in S_{r,h} \vee \underset{t_i=t,t_j=\ell,t_k=h}{\overset{\text{any two of } i,j,k\in S_r}{\vee}} \vee \underset{t_i=t,t_j=\ell,t_k=h}{\overset{\text{only } j\in S_r}{\vee}}\right) \middle| G_r\right) \\
&\quad = \left(\prod_{q=1}^{T}\binom{n_{r,q}}{m_{r,q}}\right)^{-1}\Bigg[\left(\binom{n_{r,t}-1}{m_{r,t}-1}\binom{n_{r,\ell}-1}{m_{r,\ell}-1}\binom{n_{r,h}-1}{m_{r,h}-1} + \binom{n_{r,t}-1}{m_{r,t}-1}\binom{n_{r,\ell}-1}{m_{r,\ell}}\binom{n_{r,h}-1}{m_{r,h}-1}\right. \\
&\qquad\qquad + \binom{n_{r,t}-1}{m_{r,t}}\binom{n_{r,\ell}-1}{m_{r,\ell}-1}\binom{n_{r,h}-1}{m_{r,h}-1} + \binom{n_{r,t}-1}{m_{r,t}-1}\binom{n_{r,\ell}-1}{m_{r,\ell}-1}\binom{n_{r,h}-1}{m_{r,h}} \\
&\qquad\qquad \left. + \binom{n_{r,t}-1}{m_{r,t}}\binom{n_{r,\ell}-1}{m_{r,\ell}-1}\binom{n_{r,h}-1}{m_{r,h}}\right)\prod_{q\neq t,\ell,h}\binom{n_{r,q}}{m_{r,q}}\Bigg].
\end{aligned}
$$

Therefore, we obtain the following:

$$
\begin{aligned}
\mathrm{E}(\tau(G_r^s)|G_r) &= \sum_{t=1}^{T}\left(\sum_{i\in V_{r,t}}\sum_{\substack{j\in V_{r,t}\\ i\neq j\neq k}}\sum_{k\in V_{r,t}} W_{ij,r}W_{jk,r}\left(\psi_{r,t}^3 + 3\psi_{r,t}^2(1-\psi_{r,t}) + \psi_{r,t}(1-\psi_{r,t})^2 + o(1)\right)\right) \\
&\quad + \sum_{t=1}^{T}\sum_{\ell\neq t}^{T}\left(\sum_{i\in V_{r,t}}\sum_{\substack{j\in V_{r,t}\\ i\neq j\neq k}}\sum_{k\in V_{r,\ell}} W_{ij,r}W_{jk,r}\left(\psi_{r,t}^2\psi_{r,\ell} + 2(1-\psi_{r,t})\psi_{r,t}\psi_{r,\ell} + \psi_{r,t}^2(1-\psi_{r,\ell}) + o(1)\right)\right) \\
&\quad + \sum_{t=1}^{T}\sum_{\ell\neq t}^{T}\left(\sum_{i\in V_{r,\ell}}\sum_{\substack{j\in V_{r,t}\\ i\neq j\neq k}}\sum_{k\in V_{r,t}} W_{ij,r}W_{jk,r}\left(\psi_{r,t}^2\psi_{r,\ell} + 2(1-\psi_{r,t})\psi_{r,t}\psi_{r,\ell} + \psi_{r,t}^2(1-\psi_{r,\ell}) + o(1)\right)\right) \\
&\quad + \sum_{t=1}^{T}\sum_{\ell\neq t}^{T}\left(\sum_{i\in V_{r,t}}\sum_{\substack{j\in V_{r,\ell}\\ i\neq j\neq k}}\sum_{k\in V_{r,t}} W_{ij,r}W_{jk,r}\left(\psi_{r,t}^2\psi_{r,\ell} + 2(1-\psi_{r,t})\psi_{r,t}\psi_{r,\ell} + \psi_{r,t}^2(1-\psi_{r,\ell}) + (1-\psi_{r,t})^2\psi_{r,\ell} + o(1)\right)\right) \\
&\quad + \sum_{t=1}^{T}\sum_{\ell\neq t}^{T}\sum_{h\neq t,\ell}^{T}\left(\sum_{i\in V_{r,t}}\sum_{\substack{j\in V_{r,\ell}\\ i\neq j\neq k}}\sum_{k\in V_{r,h}} W_{ij,r}W_{jk,r}\Big(\psi_{r,t}\psi_{r,\ell}\psi_{r,h} + (1-\psi_{r,t})\psi_{r,\ell}\psi_{r,h} + \psi_{r,t}(1-\psi_{r,\ell})\psi_{r,h}\right. \\
&\qquad\qquad \left.\left. + \psi_{r,t}\psi_{r,\ell}(1-\psi_{r,h}) + (1-\psi_{r,t})\psi_{r,\ell}(1-\psi_{r,h})\Big)\right).
\end{aligned}
$$

The weighted estimator for the number of connected triples has the following form:

$$\tilde{\tau}(G_r^s) = \sum_{t=1}^{T}\left(\sum_{\substack{i\in S_{r,t}\\ i\neq j\neq k}}\sum_{j\in S_{r,t}}\sum_{k\in S_{r,t}} W_{ij,r}^s W_{jk,r}^s \left(\psi_{r,t}^3 + 3\psi_{r,t}^2(1-\psi_{r,t}) + \psi_{r,t}(1-\psi_{r,t})^2\right)^{-1}\right)$$

$$+\sum_{t=1}^{T}\sum_{\ell\neq t}^{T}\left(\sum_{\substack{i\in S_{r,t}\\ i\neq j\neq k}}\sum_{j\in S_{r,t}}\sum_{k\in S_{r,\ell}} W_{ij,r}^s W_{jk,r}^s \left(\psi_{r,t}^2\psi_{r,\ell} + 2(1-\psi_{r,t})\psi_{r,t}\psi_{r,\ell} + \psi_{r,t}^2(1-\psi_{r,\ell})\right)^{-1}\right)$$

$$+\sum_{t=1}^{T}\sum_{\ell\neq t}^{T}\left(\sum_{\substack{i\in S_{r,\ell}\\ i\neq j\neq k}}\sum_{j\in S_{r,t}}\sum_{k\in S_{r,t}} W_{ij,r}^s W_{jk,r}^s \left(\psi_{r,t}^2\psi_{r,\ell} + 2(1-\psi_{r,t})\psi_{r,t}\psi_{r,\ell} + \psi_{r,t}^2(1-\psi_{r,\ell})\right)^{-1}\right)$$

$$+\sum_{t=1}^{T}\sum_{\ell\neq t}^{T}\left(\sum_{\substack{i\in S_{r,t}\\ i\neq j\neq k}}\sum_{j\in S_{r,\ell}}\sum_{k\in S_{r,t}} W_{ij,r}^s W_{jk,r}^s \left(\psi_{r,t}^2\psi_{r,\ell} + 2(1-\psi_{r,t})\psi_{r,t}\psi_{r,\ell} + \psi_{r,t}^2(1-\psi_{r,\ell}) + (1-\psi_{r,t})^2\psi_{r,\ell}\right)^{-1}\right)$$

$$+\sum_{t=1}^{T}\sum_{\ell\neq t}^{T}\sum_{h\neq t,\ell}^{T}\left(\sum_{\substack{i\in S_{r,t}\\ i\neq j\neq k}}\sum_{j\in S_{r,\ell}}\sum_{k\in S_{r,h}} W_{ij,r}^s W_{jk,r}^s \Big(\psi_{r,t}\psi_{r,\ell}\psi_{r,h} + (1-\psi_{r,t})\psi_{r,\ell}\psi_{r,h} + \psi_{r,t}(1-\psi_{r,\ell})\psi_{r,h}\right.$$

$$\left.\left. + \psi_{r,t}\psi_{r,\ell}(1-\psi_{r,h}) + (1-\psi_{r,t})\psi_{r,\ell}(1-\psi_{r,h})\Big)^{-1}\right).$$

A.3 Epidemic threshold

There is one version of the epidemic threshold derived based on the mean-field approximation (Pastor-Satorras and Vespignani, 2002), which is stated as

$$Thrld_r = \frac{\frac{1}{n_r}\sum_{i\in V_r}\sum_{j\in V_r} W_{ij,r}}{\frac{1}{n_r}\sum_{i\in V_r}(\sum_{j\in V_r} W_{ij,r})^2}.$$

The numerator is the average degree denoted by $d(G_r)$ and the denominator is average of degree square denoted by $ds(G_r)$. The corrections of $d(G_r)$ are derived in section A.1. Here we discuss the correction of $ds(G_r)$. Since

$$ds(G_r) = \frac{1}{n_r}\sum_{i\in V_r}\left(\sum_{j\in V_r} W_{ij,r}\right)^2 = \frac{1}{n_r}\sum_{i\in V_r}\sum_{j\in V_r} W_{ij,r} + \frac{1}{n_r}\sum_{i\in V_r}\sum_{j\in V_r}\sum_{\substack{k\in V_r\\ k\neq j}} W_{ij,r}W_{ik,r}$$

$$= d(G_r) + \frac{1}{n_r}\sum_{i\in V_r}\sum_{j\in V_r}\sum_{\substack{k\in V_r\\ k\neq j}} W_{ij,r}W_{ik,r},$$

we only need to consider the second term in the above equation. For induced subgraphs,

$$
\begin{aligned}
&\mathrm{E}\left(\left.\frac{1}{m_r}\sum_{i\in S_r}\sum_{j\in S_r}\sum_{\substack{k\in S_r\\k\neq j}} W^{|s}_{ij,r}W^{|s}_{ik,r}\right| G_r\right)\\
&=\sum_{t=1}^{T}\left(\frac{1}{m_r}\sum_{i\in V_{r,t}}\sum_{j\in V_{r,t}}\sum_{\substack{k\in V_{r,t}\\k\neq j}} W_{ij,r}W_{ik,r}\mathrm{E}\left(I\left(i,j,k\in S_{r,t}\right)|\,G_r\right)\right)\\
&+\sum_{t=1}^{T}\sum_{\ell\neq t}^{T}\left(\frac{1}{m_r}\sum_{i\in V_{r,t}}\sum_{j\in V_{r,t}}\sum_{\substack{k\in V_{r,\ell}\\k\neq j}} W_{ij,r}W_{ik,r}\mathrm{E}\left(I\left(i,j\in S_{r,t},k\in S_{r,\ell}\right)|\,G_r\right)\right)\\
&+\sum_{t=1}^{T}\sum_{\ell\neq t}^{T}\left(\frac{1}{m_r}\sum_{i\in V_{r,\ell}}\sum_{j\in V_{r,t}}\sum_{\substack{k\in V_{r,t}\\k\neq j}} W_{ij,r}W_{ik,r}\mathrm{E}\left(I\left(i\in S_{r,t},j,k\in S_{r,\ell}\right)|\,G_r\right)\right)\\
&+\sum_{t=1}^{T}\sum_{\ell\neq t}^{T}\left(\frac{1}{m_r}\sum_{i\in V_{r,t}}\sum_{j\in V_{r,\ell}}\sum_{\substack{k\in V_{r,t}\\k\neq j}} W_{ij,r}W_{ik,r}\mathrm{E}\left(I\left(i,k\in S_{r,t},j\in S_{r,\ell}\right)|\,G_r\right)\right)\\
&+\sum_{t=1}^{T}\sum_{\ell\neq t}^{T}\sum_{h\neq t,\ell}^{T}\frac{1}{m_r}\left(\sum_{i\in V_{r,t}}\sum_{j\in V_{r,\ell}}\sum_{\substack{k\in V_{r,k}\\k\neq j}} W_{ij,r}W_{ik,r}\mathrm{E}\left(I\left(i\in S_{r,t},j\in S_{r,\ell},k\in S_{r,h}\right)|\,G_r\right)\right).
\end{aligned}
$$

When three individuals are of the same type, i.e., $t_i=t_j=t_k=t$, we have

$$
\mathrm{E}\left(I\left(i,j,k\in S_{r,t}\right)|\,G_r\right)=\left(\frac{\binom{n_{r,t}-3}{m_{r,t}-3}\prod_{q\neq t}\binom{n_{r,q}}{m_{r,q}}}{\prod_{q=1}^{T}\binom{n_{r,q}}{m_{r,q}}}\right).
$$

When two individuals are of the same type and the third is of a different type, including $t_i=t_j=t$ and $t_k=\ell$; $t_j=t_k=t$ and $t_i=\ell$; and $t_i=t_k=t$ and $t_j=\ell$,

$$
\begin{aligned}
&\mathrm{E}\left(I\left(i,j\in S_{r,t},k\in S_{r,\ell}\right)|\,G_r\right)=\mathrm{E}\left(I\left(i\in S_{r,\ell},j,k\in S_{r,t}\right)|\,G_r\right)=\mathrm{E}\left(I\left(i,k\in S_{r,t},j\in S_{r,\ell}\right)|\,G_r\right)\\
&=\left(\frac{\binom{n_{r,t}-2}{m_{r,t}-2}\binom{n_{r,\ell}-1}{m_{r,\ell}-1}\prod_{q\neq t,\ell}\binom{n_{r,q}}{m_{r,q}}}{\prod_{q=1}^{T}\binom{n_{r,q}}{m_{r,q}}}\right).
\end{aligned}
$$

When three individuals are of different types, i.e., $t_i = t, t_j = \ell, t_k = h$, $\mathrm{E}\left(\left(i \in S_{r,t}, j \in S_{r,\ell}, k \in S_{r,h}\right)\middle|\, G_r\right) = \left(\frac{\binom{n_{r,t}-1}{m_{r,t}-1}\binom{n_{r,\ell}-1}{m_{r,\ell}-1}\binom{n_{r,h}-1}{m_{r,h}-1}\prod_{q\neq t,\ell,h}\binom{n_{r,q}}{m_{r,q}}}{\prod_{q=1}^{T}\binom{n_{r,q}}{m_{r,q}}}\right)$. Therefore,

$$
\begin{aligned}
&\mathrm{E}\left(\frac{1}{m_r}\sum_{i\in S_r}\sum_{j\in S_r}\sum_{\substack{k\in S_r\\k\neq j}} W^{|s}_{ij,r}W^{|s}_{ik,r}\,\middle|\, G_r\right)\\
&= \frac{1}{n_r}\sum_{t=1}^{T}\left(\sum_{i\in V_{r,t}}\sum_{j\in V_{r,t}}\sum_{\substack{k\in V_{r,t}\\k\neq j}} W_{ij,r}W_{ik,r}\left(\frac{\psi^3_{r,t}+o(1)}{\psi_r}\right)\right)\\
&\quad+\frac{1}{n_r}\sum_{t=1}^{T}\sum_{\ell\neq t}^{T}\left(\sum_{i\in V_{r,t}}\sum_{j\in V_{r,t}}\sum_{\substack{k\in V_{r,\ell}\\k\neq j}} W_{ij,r}W_{ik,r}\left(\frac{\psi^2_{r,t}\psi_{r,\ell}+o(1)}{\psi_r}\right)\right)\\
&\quad+\frac{1}{n_r}\sum_{t=1}^{T}\sum_{\ell\neq t}^{T}\left(\sum_{i\in V_{r,\ell}}\sum_{j\in V_{r,t}}\sum_{\substack{k\in V_{r,t}\\k\neq j}} W_{ij,r}W_{ik,r}\left(\frac{\psi^2_{r,t}\psi_{r,\ell}+o(1)}{\psi_r}\right)\right)\\
&\quad+\frac{1}{n_r}\sum_{t=1}^{T}\sum_{\ell\neq t}^{T}\left(\sum_{i\in V_{r,t}}\sum_{j\in V_{r,\ell}}\sum_{\substack{k\in V_{r,t}\\k\neq j}} W_{ij,r}W_{ik,r}\left(\frac{\psi^2_{r,t}\psi_{r,\ell}+o(1)}{\psi_r}\right)\right)\\
&\quad+\frac{1}{n_r}\sum_{t=1}^{T}\sum_{\ell\neq t}^{T}\sum_{h\neq t,\ell}^{T}\left(\sum_{i\in V_{r,t}}\sum_{j\in V_{r,\ell}}\sum_{\substack{k\in V_{r,k}\\k\neq j}} W_{ij,r}W_{ik,r}\left(\frac{\psi_{r,t}\psi_{r,\ell}\psi_{r,h}}{\psi_r}\right)\right).
\end{aligned}
$$

For the case of induced subgraphs, we propose the following weighted estimator for $ds(G^{|s}_r)$,

$$
\begin{aligned}
\widetilde{ds}(G^{|s}_r) &= \widetilde{d}(G^{|s}_r) + \frac{1}{m_r}\sum_{t=1}^{T}\left(\sum_{i\in S_{r,t}}\sum_{j\in S_{r,t}}\sum_{\substack{k\in S_{r,t}\\k\neq j}} W^{|s}_{ij,r}W^{|s}_{ik,r}\left(\frac{\psi^3_{r,t}}{\psi_r}\right)^{-1}\right)\\
&+\frac{1}{m_r}\sum_{t=1}^{T}\sum_{\ell\neq t}^{T}\left(\sum_{i\in S_{r,t}}\sum_{j\in S_{r,t}}\sum_{\substack{k\in S_{r,\ell}\\k\neq j}} W^{|s}_{ij,r}W^{|s}_{ik,r}\left(\frac{\psi^2_{r,t}\psi_{r,\ell}}{\psi_r}\right)^{-1}\right)\\
&+\frac{1}{m_r}\sum_{t=1}^{T}\sum_{\ell\neq t}^{T}\left(\sum_{i\in S_{r,\ell}}\sum_{j\in S_{r,t}}\sum_{\substack{k\in S_{r,t}\\k\neq j}} W^{|s}_{ij,r}W^{|s}_{ik,r}\left(\frac{\psi^2_{r,t}\psi_{r,\ell}}{\psi_r}\right)^{-1}\right)\\
&+\frac{1}{m_r}\sum_{t=1}^{T}\sum_{\ell\neq t}^{T}\left(\sum_{i\in S_{r,t}}\sum_{j\in S_{r,\ell}}\sum_{\substack{k\in S_{r,t}\\k\neq j}} W^{|s}_{ij,r}W^{|s}_{ik,r}\left(\frac{\psi^2_{r,t}\psi_{r,\ell}}{\psi_r}\right)^{-1}\right)\\
&+\frac{1}{m_r}\sum_{t=1}^{T}\sum_{\ell\neq t}^{T}\sum_{h\neq t,\ell}^{T}\left(\sum_{i\in S_{r,t}}\sum_{j\in S_{r,\ell}}\sum_{\substack{k\in S_{r,h}\\k\neq j}} W^{|s}_{ij,r}W^{|s}_{ik,r}\left(\frac{\psi_{r,t}\psi_{r,\ell}\psi_{r,h}}{\psi_r}\right)^{-1}\right).
\end{aligned}
$$

For star subgraphs, the conditional expectation is

$$
\begin{aligned}
&\mathrm{E}\left(\frac{1}{m_r}\sum_{i\in S_r}\sum_{j\in S_r}\sum_{\substack{k\in S_r\\k\neq j}} W^s_{ij,r}W^s_{ik,r}\,\middle|\, G_r\right)\\
&=\sum_{t=1}^{T}\left(\frac{1}{m_r}\sum_{i\in V_{r,t}}\sum_{j\in V_{r,t}}\sum_{\substack{k\in V_{r,t}\\k\neq j}} W_{ij,r}W_{ik,r}\mathrm{E}\left(I\left(i,j,k\in S_{r,t}\vee{}^{\text{any two of } i,j,k\in S_r}_{t_i=t_j=t_k=t}\vee{}^{\text{only } i\in S_r}_{t_i=t_j=t_k=t}\right)\middle|\, G_r\right)\right)\\
&\quad+\sum_{t=1}^{T}\sum_{\ell\neq t}^{T}\left(\frac{1}{m_r}\sum_{i\in V_{r,t}}\sum_{j\in V_{r,t}}\sum_{\substack{k\in V_{r,\ell}\\k\neq j}} W_{ij,r}W_{ik,r}\mathrm{E}\left(I\left({}^{i,j\in S_{r,t}}_{k\in S_{r,\ell}}\vee{}^{\text{any two of } i,j,k\in S_r}_{t_i=t_j=t,t_k=\ell}\vee{}^{\text{only } i\in S_r}_{t_i=t_j=t,t_k=\ell}\right)\middle|\, G_r\right)\right)\\
&\quad+\sum_{t=1}^{T}\sum_{\ell\neq t}^{T}\left(\frac{1}{m_r}\sum_{i\in V_{r,t}}\sum_{j\in V_{r,\ell}}\sum_{\substack{k\in V_{r,t}\\k\neq j}} W_{ij,r}W_{ik,r}\mathrm{E}\left(I\left({}^{i,k\in S_{r,t}}_{j\in S_{r,\ell}}\vee{}^{\text{any two of } i,j,k\in S_r}_{t_i=t_k=t,t_j=\ell}\vee{}^{\text{only } i\in S_r}_{t_i=t_k=t,t_j=\ell}\right)\middle|\, G_r\right)\right)\\
&\quad+\sum_{t=1}^{T}\sum_{\ell\neq t}^{T}\left(\frac{1}{m_r}\sum_{i\in V_{r,\ell}}\sum_{j\in V_{r,t}}\sum_{\substack{k\in V_{r,t}\\k\neq j}} W_{ij,r}W_{ik,r}\mathrm{E}\left(I\left({}^{j,k\in S_{r,t}}_{i\in S_{r,\ell}}\vee{}^{\text{any two of } i,j,k\in S_r}_{t_j=t_k=t,t_i=\ell}\vee{}^{\text{only } i\in S_r}_{t_j=t_k=t,t_i=\ell}\right)\middle|\, G_r\right)\right)\\
&\quad+\sum_{t=1}^{T}\sum_{\ell\neq t}^{T}\sum_{h\neq t,\ell}^{T}\left(\frac{1}{m_r}\sum_{i\in V_{r,t}}\sum_{j\in V_{r,\ell}}\sum_{\substack{k\in V_{r,h}\\k\neq j}} W_{ij,r}W_{ik,r}\mathrm{E}\left(I\left(i\in S_{r,t},j\in S_{r,\ell},k\in S_{r,h}\vee{}^{\text{any two of } i,j,k\in S_r}_{t_i=t,t_j=\ell,t_k=h}\vee{}^{\text{only } i\in S_r}_{t_i=t,t_j=\ell,t_k=h}\right)\middle|\, G_r\right)\right).
\end{aligned}
$$

When three individuals are of the same type, i.e., $t_i=t_j=t_k=t$, we have

$$
\mathrm{E}\left(I\left(i,j,k\in S_{r,t}\vee{}^{\text{any two of } i,j,k\in S_r}_{t_i=t_j=t_k=t}\vee{}^{\text{only } i\in S_r}_{t_i=t_j=t_k=t}\right)\middle|\, G_r\right)=\left(\frac{\left(\binom{n_{r,t}-3}{m_{r,t}-3}+3\binom{n_{r,t}-3}{m_{r,t}-2}+\binom{n_{r,t}-3}{m_{r,t}-1}\right)\prod_{q\neq t}\binom{n_{r,q}}{m_{r,q}}}{\prod_{q=1}^{T}\binom{n_{r,q}}{m_{r,q}}}\right).
$$

When two individuals are of the same type and the third is of a different type that $t_i=t_j=t$ and $t_k=\ell$ or $t_i=t_k=t$ and $t_j=\ell$

$$
\begin{aligned}
&\mathrm{E}\left(I\left({}^{i,j\in S_{r,t}}_{k\in S_{r,\ell}}\vee{}^{\text{any two of } i,j,k\in S_r}_{t_i=t_j=t,t_k=\ell}\vee{}^{\text{only } i\in S_r}_{t_i=t_j=t,t_k=\ell}\right)\middle|\, G_r\right)\\
&=\mathrm{E}\left(I\left({}^{i,k\in S_{r,t}}_{j\in S_{r,\ell}}\vee{}^{\text{any two of } i,j,k\in S_r}_{t_i=t_k=t,t_j=\ell}\vee{}^{\text{only } i\in S_r}_{t_i=t_k=t,t_j=\ell}\right)\middle|\, G_r\right)\\
&=\left(\frac{\left(\binom{n_{r,t}-2}{m_{r,t}-2}\binom{n_{r,\ell}-1}{m_{r,\ell}-1}+2\binom{n_{r,t}-2}{m_{r,t}-1}\binom{n_{r,\ell}-1}{m_{r,\ell}-1}+\binom{n_{r,t}-2}{m_{r,t}-2}\binom{n_{r,\ell}-1}{m_{r,\ell}}+\binom{n_{r,t}-2}{m_{r,t}-1}\binom{n_{r,\ell}-1}{m_{r,\ell}}\right)\prod_{t\neq t,\ell}\binom{n_{r,q}}{m_{r,q}}}{\prod_{q=1}^{T}\binom{n_{r,q}}{m_{r,q}}}\right);
\end{aligned}
$$

and for $t_j=t_k=t$ and $t_i=\ell$,

$$
\begin{aligned}
&\mathrm{E}\left(I\left({}^{i\in S_{r,\ell}}_{j,k\in S_{r,t}}\vee{}^{\text{any two of } i,j,k\in S_r}_{t_j=t_k=t,t_i=\ell}\vee{}^{\text{only } i\in S_r}_{t_j=t_k=t,t_i=\ell}\right)\middle|\, G_r\right)\\
&=\left(\frac{\left(\binom{n_{r,t}-2}{m_{r,t}-2}\binom{n_{r,\ell}-1}{m_{r,\ell}-1}+2\binom{n_{r,t}-2}{m_{r,t}-1}\binom{n_{r,\ell}-1}{m_{r,\ell}-1}+\binom{n_{r,t}-2}{m_{r,t}-2}\binom{n_{r,\ell}-1}{m_{r,\ell}}+\binom{n_{r,t}-2}{m_{r,t}}\binom{n_{r,\ell}-1}{m_{r,\ell}-1}\right)\prod_{t\neq t,\ell}\binom{n_{r,q}}{m_{r,q}}}{\prod_{q=1}^{T}\binom{n_{r,q}}{m_{r,q}}}\right)
\end{aligned}
$$

When three individuals are of different types, i.e., $t_i = t, t_j = \ell, t_k = h$,

$$
\begin{aligned}
&\mathrm{E}\left(I\left(i \in S_{r,t}, j \in S_{r,\ell}, k \in S_{r,h} \vee {}^{\text{any two of } i,j,k \in S_r}_{t_i=t,t_j=\ell,t_k=h} \vee {}^{\text{only } i \in S_r}_{t_i=t,t_j=\ell,t_k=h}\right) \middle| G_r \right) \\
&\quad = \left(\prod_{q=1}^{T} \binom{n_{r,q}}{m_{r,q}}\right)^{-1} \Bigg[\left(\binom{n_{r,t}-1}{m_{r,t}-1}\binom{n_{r,\ell}-1}{m_{r,\ell}-1}\binom{n_{r,h}-1}{m_{r,h}-1} + \binom{n_{r,t}-1}{m_{r,t}-1}\binom{n_{r,\ell}-1}{m_{r,\ell}}\binom{n_{r,h}-1}{m_{r,h}-1} \right. \\
&\qquad\qquad + \binom{n_{r,t}-1}{m_{r,t}}\binom{n_{r,\ell}-1}{m_{r,\ell}-1}\binom{n_{r,h}-1}{m_{r,h}-1} + \binom{n_{r,t}-1}{m_{r,t}-1}\binom{n_{r,\ell}-1}{m_{r,\ell}-1}\binom{n_{r,h}-1}{m_{r,h}} \\
&\qquad\qquad \left. + \binom{n_{r,t}-1}{m_{r,t}-1}\binom{n_{r,\ell}-1}{m_{r,\ell}}\binom{n_{r,h}-1}{m_{r,h}} \right) \prod_{q \neq t,\ell,h} \binom{n_{r,q}}{m_{r,q}} \Bigg].
\end{aligned}
$$

Therefore, we can further obtain

$$
\begin{aligned}
&\mathrm{E}\left(\frac{1}{m_r} \sum_{i \in S_r} \sum_{j \in S_r} \sum_{\substack{k \in S_r \\ k \neq j}} W^s_{ij,r} W^s_{ik,r} \middle| G_r \right) \\
&= \frac{1}{n_r} \sum_{t=1}^{T} \left(\sum_{i \in V_{r,t}} \sum_{j \in V_{r,t}} \sum_{\substack{k \in V_{r,t} \\ k \neq j}} W_{ij,r} W_{ik,r} \left(\frac{\psi^3_{r,t} + 3\psi^2_{r,t}(1-\psi_{r,t}) + \psi_{r,t}(1-\psi_{r,t})^2 + o(1)}{\psi_r} \right) \right) \\
&+ \frac{1}{n_r} \sum_{t=1}^{T} \sum_{\ell \neq t}^{T} \left(\sum_{i \in V_{r,t}} \sum_{j \in V_{r,t}} \sum_{\substack{k \in V_{r,\ell} \\ k \neq j}} W_{ij,r} W_{ik,r} \left(\frac{\psi^2_{r,t}\psi_{r,\ell} + 2(1-\psi_{r,t})\psi_{r,t}\psi_{r,\ell} + \psi^2_{r,t}(1-\psi_{r,\ell}) + o(1)}{\psi_r} \right) \right) \\
&+ \frac{1}{n_r} \sum_{t=1}^{T} \sum_{\ell \neq t}^{T} \left(\sum_{i \in V_{r,t}} \sum_{j \in V_{r,\ell}} \sum_{\substack{k \in V_{r,t} \\ k \neq j}} W_{ij,r} W_{ik,r} \left(\frac{\psi^2_{r,t}\psi_{r,\ell} + 2(1-\psi_{r,t})\psi_{r,t}\psi_{r,\ell} + \psi^2_{r,t}(1-\psi_{r,\ell}) + o(1)}{\psi_r} \right) \right) \\
&+ \frac{1}{n_r} \sum_{t=1}^{T} \sum_{\ell \neq t}^{T} \left(\sum_{i \in V_{r,\ell}} \sum_{j \in V_{r,t}} \sum_{\substack{k \in V_{r,t} \\ k \neq j}} W_{ij,r} W_{ik,r} \left(\frac{\psi^2_{r,t}\psi_{r,\ell} + 2(1-\psi_{r,t})\psi_{r,t}\psi_{r,\ell} + \psi^2_{r,t}(1-\psi_{r,\ell}) + (1-\psi_{r,t})^2\psi_{r,\ell} + o(1)}{\psi_r} \right) \right) \\
&+ \frac{1}{n_r} \sum_{t=1}^{T} \sum_{\ell \neq t}^{T} \sum_{h \neq t,\ell}^{T} \left(\sum_{i \in V_{r,t}} \sum_{j \in V_{r,\ell}} \sum_{\substack{k \in V_{r,h} \\ k \neq j}} W_{ij,r} W_{ik,r} \Big(\psi_r^{-1} \Big(\psi_{r,t}\psi_{r,\ell}\psi_{r,h} + (1-\psi_{r,t})\psi_{r,\ell}\psi_{r,h} + \psi_{r,t}(1-\psi_{r,\ell})\psi_{r,h} \right. \right. \\
&\qquad\qquad \left. \left. + \psi_{r,t}\psi_{r,\ell}(1-\psi_{r,h}) + \psi_{r,t}(1-\psi_{r,\ell})(1-\psi_{r,h}) \Big) \Big) \right).
\end{aligned}
$$

For star subgraphs, we consider the following weighted estimator for $ds(G_r^s)$,

$$\widetilde{ds}(G_r^s) = \widetilde{d}(G_r^s) + \frac{1}{m_r}\sum_{t=1}^{T}\left(\sum_{i\in S_{r,t}}\sum_{j\in S_{r,t}}\sum_{\substack{k\in S_{r,t}\\k\neq j}} W_{ij,r}^s W_{ik,r}^s \left(\frac{\psi_{r,t}^3 + 3\psi_{r,t}^2(1-\psi_{r,t}) + \psi_{r,t}(1-\psi_{r,t})^2}{\psi_r}\right)^{-1}\right)$$

$$+ \frac{1}{m_r}\sum_{t=1}^{T}\sum_{\ell\neq t}^{T}\left(\sum_{i\in S_{r,t}}\sum_{j\in S_{r,t}}\sum_{\substack{k\in S_{r,\ell}\\k\neq j}} W_{ij,r}^s W_{ik,r}^s \left(\frac{\psi_{r,t}^2\psi_{r,\ell} + 2(1-\psi_{r,t})\psi_{r,t}\psi_{r,\ell} + \psi_{r,t}^2(1-\psi_{r,\ell})}{\psi_r}\right)^{-1}\right)$$

$$+ \frac{1}{m_r}\sum_{t=1}^{T}\sum_{\ell\neq t}^{T}\left(\sum_{i\in S_{r,t}}\sum_{j\in S_{r,\ell}}\sum_{\substack{k\in S_{r,t}\\k\neq j}} W_{ij,r}^s W_{ik,r}^s \left(\frac{\psi_{r,t}^2\psi_{r,\ell} + 2(1-\psi_{r,t})\psi_{r,t}\psi_{r,\ell} + \psi_{r,t}^2(1-\psi_{r,\ell})}{\psi_r}\right)^{-1}\right)$$

$$+ \frac{1}{m_r}\sum_{t=1}^{T}\sum_{\ell\neq t}^{T}\left(\sum_{i\in S_{r,\ell}}\sum_{j\in S_{r,t}}\sum_{\substack{k\in S_{r,t}\\k\neq j}} W_{ij,r}^s W_{ik,r}^s \left(\frac{\psi_{r,t}^2\psi_{r,\ell} + 2(1-\psi_{r,t})\psi_{r,t}\psi_{r,\ell} + \psi_{r,t}^2(1-\psi_{r,\ell}) + (1-\psi_{r,t})^2\psi_{r,\ell}}{\psi_r}\right)^{-1}\right)$$

$$+ \frac{1}{m_r}\sum_{t=1}^{T}\sum_{\ell\neq t}^{T}\sum_{h\neq t,\ell}^{T}\left(\sum_{i\in S_{r,t}}\sum_{j\in S_{r,\ell}}\sum_{\substack{k\in S_{r,h}\\k\neq j}} W_{ij,r}^s W_{ik,r}^s \Big(\psi_r^{-1}\Big(\psi_{r,t}\psi_{r,\ell}\psi_{r,h} + (1-\psi_{r,t})\psi_{r,\ell}\psi_{r,h} + \psi_{r,t}(1-\psi_{r,\ell})\psi_{r,h}\right.\right.$$
$$\left.\left. + \psi_{r,t}\psi_{r,\ell}(1-\psi_{r,h}) + \psi_{r,t}(1-\psi_{r,\ell})(1-\psi_{r,h})\Big)\Big)^{-1}\right).$$

A.4 Graph span

The graph span is defined as

$$\ell(G_r) = \frac{\log n_r - \log d(G_r)}{\log d_2(G_r) - \log d(G_r)} + 1,$$

where $d_2(G_r) = \frac{1}{n_r}\sum_{i\in V_r}\sum_{j>i}\sum_{k\neq i,j} W_{ij,r}W_{jk,r}$ is the average number of second 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 $\tilde{d}_2(G_r^s) = d_2(G_r^2)/k(\psi)$ and $\tilde{d}_2(G_r^{|s}) = d_2(G_r^{|s})/\psi^2$, the analytically corrected estimators for $\ell(G_r)$ based on G_r^s and $G_r^{|s}$ are

$$\tilde{\ell}(G_r^s) = \frac{\log n - \log \tilde{d}(G_r^s)}{\log \tilde{d}_2(G_r^s) - \log \tilde{d}(G_r^s)} + 1 \quad \text{and} \quad \tilde{\ell}(G_r^{|s}) = \frac{\log(\psi^{-1}m) - \log \tilde{d}(G_r^{|s})}{\log \tilde{d}_2(G_r^{|s}) - \log \tilde{d}(G_r^{|s})} + 1.$$

For the case of induced subgraph,

$$
\begin{aligned}
\mathrm{E}(d_2(G_r^{|s})|G_r) &= \mathrm{E}\left(\frac{1}{m_r} \sum_{i\in S_r} \sum_{\substack{j\in S_r \\ i\neq j}} \sum_{\substack{k\in S_r \\ i\neq j\neq k}} W_{ij,r}^{|s} W_{jk,r}^{|s} \middle| G_r \right) \\
&= \sum_{t=1}^{T} \left(\frac{1}{m_r} \sum_{i\in V_{r,t}} \sum_{\substack{j\in V_{r,t} \\ i\neq j}} \sum_{\substack{k\in V_{r,t} \\ i\neq j\neq k}} W_{ij,r} W_{jk,r} \mathrm{E}\left(I\left(i,j,k\in S_{r,t}\right) | \, G_r \right) \right) \\
&+ \sum_{t=1}^{T} \sum_{\ell\neq t}^{T} \left(\frac{1}{m_r} \sum_{i\in V_{r,t}} \sum_{\substack{j\in V_{r,t} \\ i\neq j}} \sum_{\substack{k\in V_{r,\ell} \\ i\neq j\neq k}} W_{ij,r} W_{jk,r} \mathrm{E}\left(I\left(i,j\in S_{r,t},k\in S_{r,\ell}\right) | \, G_r \right) \right) \\
&+ \sum_{t=1}^{T} \sum_{\ell\neq t}^{T} \left(\frac{1}{m_r} \sum_{i\in V_{r,\ell}} \sum_{\substack{j\in V_{r,t} \\ i\neq j}} \sum_{\substack{k\in V_{r,t} \\ i\neq j\neq k}} W_{ij,r} W_{jk,r} \mathrm{E}\left(I\left(j,k\in S_{r,t},i\in S_{r,\ell}\right) | \, G_r \right) \right) \\
&+ \sum_{t=1}^{T} \sum_{\ell\neq t}^{T} \left(\frac{1}{m_r} \sum_{i\in V_{r,t}} \sum_{\substack{j\in V_{r,\ell} \\ i\neq j}} \sum_{\substack{k\in V_{r,t} \\ i\neq j\neq k}} W_{ij,r} W_{jk,r} \mathrm{E}\left(I\left(i,k\in S_{r,t},j\in S_{r,\ell}\right) | \, G_r \right) \right) \\
&+ \sum_{t=1}^{T} \sum_{\ell\neq t}^{T} \sum_{h\neq t,\ell}^{T} \left(\frac{1}{m_r} \sum_{i\in V_{r,t}} \sum_{\substack{j\in V_{r,\ell} \\ i\neq j}} \sum_{\substack{k\in V_{r,h} \\ i\neq j\neq k}} W_{ij,r} W_{jk,r} \mathrm{E}\left(I\left(i\in S_{r,t},j\in S_{r,\ell},k\in S_{t,h}\right) | \, G_r \right) \right).
\end{aligned}
$$

When three individuals are of the same type, i.e., $t_i = t_j = t_k = t$, we have

$$
\mathrm{E}\left(I\left(i,j,k\in S_{r,t}\right) | \, G_r \right) = \left(\frac{\binom{n_{r,t}-3}{m_{r,t}-3} \prod_{q\neq t} \binom{n_{r,q}}{m_{r,q}}}{\prod_{q=1}^{T} \binom{n_{r,q}}{m_{r,q}}} \right).
$$

When two individuals are of the same type and the third is of a different type, including $t_i = t_j = t$ and $t_k = \ell$; $t_j = t_k = t$ and $t_i = \ell$; and $t_i = t_k = t$ and $t_j = \ell$,

$$
\begin{aligned}
&\mathrm{E}\left(I\left(i,j\in S_{r,t},k\in S_{r,\ell}\right) | \, G_r \right) = \mathrm{E}\left(I\left(i\in S_{r,\ell},j,k\in S_{r,t}\right) | \, G_r \right) = \mathrm{E}\left(I\left(i,k\in S_{r,t},j\in S_{r,\ell}\right) | \, G_r \right) \\
&= \left(\frac{\binom{n_{r,t}-2}{m_{r,t}-2} \binom{n_{r,\ell}-1}{m_{r,\ell}-1} \prod_{q\neq t,\ell} \binom{n_{r,q}}{m_{r,q}}}{\prod_{q=1}^{T} \binom{n_{r,q}}{m_{r,q}}} \right).
\end{aligned}
$$

When three individuals are of different types, i.e., $t_i = t, t_j = \ell, t_k = h$,

$$
\mathrm{E}\left(I\left(i\in S_{r,t},j\in S_{r,\ell},k\in S_{r,h}\right) | \, G_r \right) = \left(\frac{\binom{n_{r,t}-1}{m_{r,t}-1} \binom{n_{r,\ell}-1}{m_{r,\ell}-1} \binom{n_{r,h}-1}{m_{r,h}-1} \prod_{q\neq t,\ell,h} \binom{n_{r,q}}{m_{r,q}}}{\prod_{q=1}^{T} \binom{n_{r,q}}{m_{r,q}}} \right).
$$

Therefore,

$$
\begin{aligned}
&\mathrm{E}(d_2(G_r^{|s})|G_r) = \\
&\frac{1}{n_r}\sum_{t=1}^{T}\left(\sum_{i\in V_{r,t}}\sum_{\substack{j\in V_{r,t}\\ i\neq j\neq k}}\sum_{k\in V_{r,t}} W_{ij,r}W_{jk,r}\left(\frac{\psi_{r,t}^3+o(1)}{\psi_r}\right)\right)+\frac{1}{n_r}\sum_{t=1}^{T}\sum_{\ell\neq t}^{T}\left(\sum_{i\in V_{r,t}}\sum_{\substack{j\in V_{r,t}\\ i\neq j\neq k}}\sum_{k\in V_{r,\ell}} W_{ij,r}W_{jk,r}\left(\frac{\psi_{r,t}^2\psi_{r,\ell}+o(1)}{\psi_r}\right)\right)\\
&+\frac{1}{n_r}\sum_{t=1}^{T}\sum_{\ell\neq t}^{T}\left(\sum_{i\in V_{r,\ell}}\sum_{\substack{j\in V_{r,t}\\ i\neq j\neq k}}\sum_{k\in V_{r,t}} W_{ij,r}W_{jk,r}\left(\frac{\psi_{r,t}^2\psi_{r,\ell}+o(1)}{\psi_r}\right)\right)+\frac{1}{n_r}\sum_{t=1}^{T}\sum_{\ell\neq t}^{T}\left(\sum_{i\in V_{r,t}}\sum_{\substack{j\in V_{r,\ell}\\ i\neq j\neq k}}\sum_{k\in V_{r,t}} W_{ij,r}W_{jk,r}\left(\frac{\psi_{r,t}^2\psi_{r,\ell}+o(1)}{\psi_r}\right)\right)\\
&+\frac{1}{n_r}\sum_{t=1}^{T}\sum_{\ell\neq t}^{T}\sum_{h\neq t,\ell}^{T}\left(\sum_{i\in V_{r,t}}\sum_{\substack{j\in V_{r,\ell}\\ i\neq j\neq k}}\sum_{k\in V_{r,h}} W_{ij,r}W_{jk,r}\left(\frac{\psi_{r,t}\psi_{r,\ell}\psi_{r,h}}{\psi_r}\right)\right).
\end{aligned}
$$

Hence, in the general case, we propose to multiply $\left(\frac{\psi_{r,t}^3}{\psi_r}\right)^{-1}$ to the triple (i,j,k) in which three individuals are of the same type t; multiply $\left(\frac{\psi_{r,t}^2\psi_{r,\ell}}{\psi_r}\right)^{-1}$ to the triple (i,j,k) in which two individuals are of the same type t and the other is of type ℓ; multiply $\left(\frac{\psi_{r,t}\psi_{r,\ell}\psi_{r,h}}{\psi_t}\right)^{-1}$ to the triple (i,j,k) in which three individuals are of different types, t, ℓ, and h, to correct the second term of $\mathrm{E}(ds(G_r^{|s})|G_r)$. The the resulting estimator for $d_2(G_r^{|s})$ thus is

$$
\begin{aligned}
\tilde{d}_2(G_r^{|s}) = &\frac{1}{m_r}\sum_{t=1}^{T}\left(\sum_{\substack{i\in S_r\\ t_i=t_j=t_k=t}}\sum_{j>i}\sum_{k\neq i,j} W_{ij,r}^{|s}W_{jk,r}^{|s}\left(\frac{\psi_{r,t}^3}{\psi_r}\right)^{-1}\right)\\
&+\frac{1}{m_r}\sum_{t=1}^{T}\sum_{\ell\neq t}^{T}\left(\sum_{\substack{i\in S_r\\ t_i=t_j=t,t_k=\ell}}\sum_{j>i}\sum_{k\neq i,j} W_{ij,r}^{|s}W_{jk,r}^{|s}\left(\frac{\psi_{r,t}^2\psi_{r,\ell}}{\psi_r}\right)^{-1}\right)\\
&+\frac{1}{m_r}\sum_{t=1}^{T}\sum_{\ell\neq t}^{T}\left(\sum_{\substack{i\in S_r\\ t_j=t_k=t,t_i=\ell}}\sum_{j>i}\sum_{k\neq i,j} W_{ij,r}^{|s}W_{jk,r}^{|s}\left(\frac{\psi_{r,t}^2\psi_{r,\ell}}{\psi_r}\right)^{-1}\right)\\
&+\frac{1}{m_r}\sum_{t=1}^{T}\sum_{\ell\neq t}^{T}\left(\sum_{\substack{i\in S_r\\ t_i=t_k=t,t_j=\ell}}\sum_{j>i}\sum_{k\neq i,j} W_{ij,r}^{|s}W_{jk,r}^{|s}\left(\frac{\psi_{r,t}^2\psi_{r,\ell}}{\psi_r}\right)^{-1}\right)\\
&+\frac{1}{m_r}\sum_{t=1}^{T}\sum_{\ell\neq t}^{T}\sum_{h\neq t,\ell}^{T}\left(\sum_{\substack{i\in S_r\\ t_i=t,t_j=\ell,t_k=h}}\sum_{j>i}\sum_{k\neq i,j} W_{ij,r}^{|s}W_{jk,r}^{|s}\left(\frac{\psi_{r,t}\psi_{r,\ell}\psi_{r,h}}{\psi_r}\right)^{-1}\right). \qquad (1)
\end{aligned}
$$

For star subgraph, the conditional expectation is

$$
\begin{aligned}
\mathrm{E}(d_2(G_r^s)|G_r) &= \mathrm{E}\left(\frac{1}{m_r} \sum_{i\in S_r} \sum_{\substack{j\in S_r \\ i\neq j}} \sum_{\substack{k\in S_r \\ i\neq j\neq k}} W_{jk,r}^s \middle| G_r \right) \\
&= \sum_{t=1}^{T} \left(\frac{1}{m_r} \sum_{i\in V_{r,t}} \sum_{\substack{j\in V_{r,t} \\ i\neq j}} \sum_{\substack{k\in V_{r,t} \\ i\neq j\neq k}} W_{ij,r}W_{jk,r}\mathrm{E}\left(I\left(i,j,k \in S_{r,t} \vee {}^{\text{any two of } i,j,k\in S_r}_{t_i=t_j=t_k=t} \vee {}^{\text{only } j\in S_r}_{t_i=t_j=t_k=t}\right)\middle| G_r\right)\right) \\
&\quad + \sum_{t=1}^{T}\sum_{\ell\neq t}^{T} \left(\frac{1}{m_r} \sum_{i\in V_{r,t}} \sum_{\substack{j\in V_{r,t} \\ i\neq j}} \sum_{\substack{k\in V_{r,\ell} \\ i\neq j\neq k}} W_{ij,r}W_{jk,r}\mathrm{E}\left(I\left({}^{i,j\in S_{r,t}}_{k\in S_{r,\ell}} \vee {}^{\text{any two of } i,j,k\in S_r}_{t_i=t_j=t,t_k=\ell} \vee {}^{\text{only } j\in S_r}_{t_i=t_j=t,t_k=\ell}\right)\middle| G_r\right)\right) \\
&\quad + \sum_{t=1}^{T}\sum_{\ell\neq t}^{T} \left(\frac{1}{m_r} \sum_{i\in V_{r,\ell}} \sum_{\substack{j\in V_{r,t} \\ i\neq j}} \sum_{\substack{k\in V_{r,t} \\ i\neq j\neq k}} W_{ij,r}W_{jk,r}\mathrm{E}\left(I\left({}^{i\in S_{r,\ell}}_{j,k\in S_{r,t}} \vee {}^{\text{any two of } i,j,k\in S_r}_{t_j=t_k=t,t_i=\ell} \vee {}^{\text{only } j\in S_r}_{t_j=t_k=t,t_i=\ell}\right)\middle| G_r\right)\right) \\
&\quad + \sum_{t=1}^{T}\sum_{\ell\neq t}^{T} \left(\frac{1}{m_r} \sum_{i\in V_{r,t}} \sum_{\substack{j\in V_{r,\ell} \\ i\neq j}} \sum_{\substack{k\in V_{r,t} \\ i\neq j\neq k}} W_{ij,r}W_{jk,r}\mathrm{E}\left(I\left({}^{i,k\in S_{r,t}}_{j\in S_{r,\ell}} \vee {}^{\text{any two of } i,j,k\in S_r}_{t_i=t_k=t,t_j=\ell} \vee {}^{\text{only } j\in S_r}_{t_i=t_k=t,t_j=\ell}\right)\middle| G_r\right)\right) \\
&\quad + \sum_{t=1}^{T}\sum_{\ell\neq t}^{T}\sum_{h\neq t,\ell}^{T} \left(\frac{1}{m_r} \sum_{i\in V_{r,t}} \sum_{\substack{j\in V_{r,\ell} \\ i\neq j}} \sum_{\substack{k\in V_{r,h} \\ i\neq j\neq k}} W_{ij,r}W_{jk,r}\mathrm{E}\left(I\left(i\in S_{r,t}, j\in S_{r,\ell}, k\in S_{r,h} \vee {}^{\text{any two of } i,j,k\in S_r}_{t_i=t,t_j=\ell,t_k=h} \vee {}^{\text{only } j\in S_r}_{t_i=t,t_j=\ell,t_k=h}\right)\middle| G_r\right)\right).
\end{aligned}
$$

When three individuals are of the same type, i.e., $t_i = t_j = t_k = t$, we have

$$
\mathrm{E}\left(I\left(i,j,k \in S_{r,t} \vee {}^{\text{any two of } i,j,k\in S_r}_{t_i=t_j=t_k=t} \vee {}^{\text{only } j\in S_r}_{t_i=t_j=t_k=t}\right)\middle| G_r\right) = \left(\frac{\left(\binom{n_{r,t}-3}{m_{r,t}-3} + 3\binom{n_{r,t}-3}{m_{r,t}-2} + \binom{n_{r,t}-3}{m_{r,t}-1}\right)}{\prod_{q\neq t}\binom{n_{r,q}}{m_{r,q}}\prod_{q=1}^{T}\binom{n_{r,q}}{m_{r,q}}} \right).
$$

When two individuals are of the same type and the third is of a different type that $t_i = t_j = t$ and $t_k = \ell$ or $t_j = t_k = t$ and $t_i = \ell$

$$
\begin{aligned}
&\mathrm{E}\left(I\left({}^{i,j\in S_{r,t}}_{k\in S_{r,\ell}} \vee {}^{\text{any two of } i,j,k\in S_r}_{t_i=t_j=t,t_k=\ell} \vee {}^{\text{only } j\in S_r}_{t_i=t_j=t,t_k=\ell}\right)\middle| G_r\right) \\
=&\mathrm{E}\left(I\left({}^{i\in S_{r,\ell}}_{j,k\in S_{r,t}} \vee {}^{\text{any two of } i,j,k\in S_r}_{t_j=t_k=t,t_i=\ell} \vee {}^{\text{only } j\in S_r}_{t_j=t_k=t,t_i=\ell}\right)\middle| G_r\right) \\
=&\left(\frac{\left(\binom{n_{r,t}-2}{m_{r,t}-2}\binom{n_{r,\ell}-1}{m_{r,\ell}-1} + 2\binom{n_{r,t}-2}{m_{r,t}-1}\binom{n_{r,\ell}-1}{m_{r,\ell}-1} + \binom{n_{r,t}-2}{m_{r,t}-2}\binom{n_{r,\ell}-1}{m_{r,\ell}} + \binom{n_{r,t}-2}{m_{r,t}-1}\binom{n_{r,\ell}-1}{m_{r,\ell}}\right)\prod_{t\neq t,\ell}\binom{n_{r,q}}{m_{r,q}}}{\prod_{q=1}^{T}\binom{n_{r,q}}{m_{r,q}}} \right);
\end{aligned}
$$

and for $t_i = t_k = t$ and $t_j = \ell$,

$$
\begin{aligned}
&\mathrm{E}\left(I\left({}^{i,k\in S_{r,t}}_{j\in S_{r,\ell}} \vee {}^{\text{any two of } i,j,k\in S_r}_{t_i=t_k=t,t_j=\ell} \vee {}^{\text{only } j\in S_r}_{t_i=t_k=t,t_j=\ell}\right)\middle| G_r\right) \\
&= \left(\frac{\left(\binom{n_{r,t}-2}{m_{r,t}-2}\binom{n_{r,\ell}-1}{m_{r,\ell}-1} + 2\binom{n_{r,t}-2}{m_{r,t}-1}\binom{n_{r,\ell}-1}{m_{r,\ell}-1} + \binom{n_{r,t}-2}{m_{r,t}-2}\binom{n_{r,\ell}-1}{m_{r,\ell}} + \binom{n_{r,t}-2}{m_{r,t}}\binom{n_{r,\ell}-1}{m_{r,\ell}-1}\right)\prod_{t\neq t,\ell}\binom{n_{r,q}}{m_{r,q}}}{\prod_{q=1}^{T}\binom{n_{r,q}}{m_{r,q}}} \right)
\end{aligned}
$$

When three individuals are of different types, i.e., $t_i = t, t_j = \ell, t_k = h$,

$$
\begin{aligned}
&\mathrm{E}\left(I\left(i \in S_{r,t}, j \in S_{r,\ell}, k \in S_{r,h} \vee {}^{\text{any two of } i,j,k \in S_r}_{t_i=t,t_j=\ell,t_k=h} \vee {}^{\text{only } j \in S_r}_{t_i=t,t_j=\ell,t_k=h}\right) \middle| G_r \right) \\
&\quad = \left(\prod_{q=1}^{T} \binom{n_{r,q}}{m_{r,q}}\right)^{-1} \Bigg[\left(\binom{n_{r,t}-1}{m_{r,t}-1}\binom{n_{r,\ell}-1}{m_{r,\ell}-1}\binom{n_{r,h}-1}{m_{r,h}-1} + \binom{n_{r,t}-1}{m_{r,t}-1}\binom{n_{r,\ell}-1}{m_{r,\ell}}\binom{n_{r,h}-1}{m_{r,h}-1} \right. \\
&\qquad\qquad + \binom{n_{r,t}-1}{m_{r,t}}\binom{n_{r,\ell}-1}{m_{r,\ell}-1}\binom{n_{r,h}-1}{m_{r,h}-1} + \binom{n_{r,t}-1}{m_{r,t}-1}\binom{n_{r,\ell}-1}{m_{r,\ell}-1}\binom{n_{r,h}-1}{m_{r,h}} \\
&\qquad\qquad \left. + \binom{n_{r,t}-1}{m_{r,t}}\binom{n_{r,\ell}-1}{m_{r,\ell}-1}\binom{n_{r,h}-1}{m_{r,h}} \right) \prod_{q \neq t,\ell,h} \binom{n_{r,q}}{m_{r,q}} \Bigg].
\end{aligned}
$$

Therefore, we can further obtain

$$
\begin{aligned}
\mathrm{E}(d_2(G_r^s)|G_r) = {} & \frac{1}{n_r}\sum_{t=1}^{T}\left(\sum_{\substack{i,j,k \in V_{r,t} \\ i\neq j\neq k}} W_{ij,r}W_{jk,r}\left(\frac{\psi_{r,t}^3 + 3\psi_{r,t}^2(1-\psi_{r,t}) + \psi_{r,t}(1-\psi_{r,t})^2 + o(1)}{\psi_r}\right)\right) \\
& + \frac{1}{n_r}\sum_{t=1}^{T}\sum_{\ell\neq t}^{T}\left(\sum_{i\in V_{r,t}}\sum_{\substack{j\in V_{r,t} \\ i\neq j\neq k}}\sum_{k\in V_{r,\ell}} W_{ij,r}W_{jk,r}\left(\frac{\psi_{r,t}^2\psi_{r,\ell} + 2(1-\psi_{r,t})\psi_{r,t}\psi_{r,\ell} + \psi_{r,t}^2(1-\psi_{r,\ell}) + o(1)}{\psi_r}\right)\right) \\
& + \frac{1}{n_r}\sum_{t=1}^{T}\sum_{\ell\neq t}^{T}\left(\sum_{i\in V_{r,\ell}}\sum_{\substack{j\in V_{r,t} \\ i\neq j\neq k}}\sum_{k\in V_{r,t}} W_{ij,r}W_{jk,r}\left(\frac{\psi_{r,t}^2\psi_{r,\ell} + 2(1-\psi_{r,t})\psi_{r,t}\psi_{r,\ell} + \psi_{r,t}^2(1-\psi_{r,\ell}) + o(1)}{\psi_r}\right)\right) \\
& + \frac{1}{n_r}\sum_{t=1}^{T}\sum_{\ell\neq t}^{T}\left(\sum_{i\in V_{r,t}}\sum_{\substack{j\in V_{r,\ell} \\ i\neq j\neq k}}\sum_{k\in V_{r,t}} W_{ij,r}W_{jk,r}\left(\frac{\psi_{r,t}^2\psi_{r,\ell} + 2(1-\psi_{r,t})\psi_{r,t}\psi_{r,\ell} + \psi_{r,t}^2(1-\psi_{r,\ell}) + (1-\psi_{r,t})^2\psi_{r,\ell} + o(1)}{\psi_r}\right)\right) \\
& + \frac{1}{n_r}\sum_{t=1}^{T}\sum_{\ell\neq t}^{T}\sum_{h\neq t,\ell}^{T}\left(\sum_{i\in V_{r,t}}\sum_{\substack{j\in V_{r,\ell} \\ i\neq j\neq k}}\sum_{k\in V_{r,h}} W_{ij,r}W_{jk,r}\Big(\psi_r^{-1}\big(\psi_{r,t}\psi_{r,\ell}\psi_{r,h} + (1-\psi_{r,t})\psi_{r,\ell}\psi_{r,h} + \psi_{r,t}(1-\psi_{r,\ell})\psi_{r,h} \right.\right. \\
& \qquad \left.\left. + \psi_{r,t}\psi_{r,\ell}(1-\psi_{r,h}) + (1-\psi_{r,t})\psi_{r,\ell}(1-\psi_{r,h})\big)\Big)\right).
\end{aligned}
$$

The resulting estimator of $d_2(G_r^s)$ is

$$\tilde{d}_2(G_r^s) = \frac{1}{m_r}\sum_{t=1}^{T}\left(\sum_{\substack{i\in S_r\\ t_i=t_j=t_k=t}}\sum_{j>i}\sum_{k\neq i,j} W_{ij,r}^s W_{jk,r}^s \left(\frac{\psi_{r,t}^3 + 3\psi_{r,t}^2(1-\psi_{r,t}) + \psi_{r,t}(1-\psi_{r,t})^2}{\psi_r}\right)^{-1}\right)$$

$$+ \frac{1}{m_r}\sum_{t=1}^{T}\sum_{\ell\neq t}^{T}\left(\sum_{\substack{i\in S_r\\ t_i=t_j=t,t_k=\ell}}\sum_{j>i}\sum_{k\neq i,j} W_{ij,r}^s W_{jk,r}^s \left(\frac{\psi_{r,t}^2\psi_{r,\ell} + 2(1-\psi_{r,t})\psi_{r,t}\psi_{r,\ell} + \psi_{r,t}^2(1-\psi_{r,\ell})}{\psi_r}\right)^{-1}\right)$$

$$+ \frac{1}{m_r}\sum_{t=1}^{T}\sum_{\ell\neq t}^{T}\left(\sum_{\substack{i\in S_r\\ t_j=t_k=t,t_i=\ell}}\sum_{j>i}\sum_{k\neq i,j} W_{ij,r}^s W_{jk,r}^s \left(\frac{\psi_{r,t}^2\psi_{r,\ell} + 2(1-\psi_{r,t})\psi_{r,t}\psi_{r,\ell} + \psi_{r,t}^2(1-\psi_{r,\ell})}{\psi_r}\right)^{-1}\right)$$

$$+ \frac{1}{m_r}\sum_{t=1}^{T}\sum_{\ell\neq t}^{T}\left(\sum_{\substack{i\in S_r\\ t_i=t_k=t,t_j=\ell}}\sum_{j>i}\sum_{k\neq i,j} W_{ij,r}^s W_{jk,r}^s \left(\frac{\psi_{r,t}^2\psi_{r,\ell} + 2(1-\psi_{r,t})\psi_{r,t}\psi_{r,\ell} + \psi_{r,t}^2(1-\psi_{r,\ell}) + (1-\psi_{r,t})^2\psi_{r,\ell}}{\psi_r}\right)^{-1}\right)$$

$$+ \frac{1}{m_r}\sum_{t=1}^{T}\sum_{\ell\neq t}^{T}\sum_{h\neq t,\ell}^{T}\left(\sum_{\substack{i\in S_r\\ t_i=t,t_j=\ell,t_k=h}}\sum_{j>i}\sum_{k\neq i,j} W_{ij,r}^s W_{jk,r}^s \Big(\psi_r^{-1}\big(\psi_{r,t}\psi_{r,\ell}\psi_{r,h} + (1-\psi_{r,t})\psi_{r,\ell}\psi_{r,h} + \psi_{r,t}(1-\psi_{r,\ell})\psi_{r,h}\right.\right.$$

$$\left.\left. + \psi_{r,t}\psi_{r,\ell}(1-\psi_{r,h}) + (1-\psi_{r,t})\psi_{r,\ell}(1-\psi_{r,h})\big)\Big)^{-1}\right). \tag{2}$$

A.5 Homophily index

Currarini et al. (2009) define the homophily index of graph G_r as $H_{r,t} = \frac{s_{r,t}}{s_{r,t}+d_{r,t}}$, where $s_{r,t}$ denotes the average number of friendships that agents of type t have with agents of the same type and $d_{r,t}$ denotes the average number of friendships that type t form with agents of type different than t. Here we may use type t to represent different demographic characteristics, e.g., gender, race, and age, etc. Specifically, let $V_{r,t}$ denotes the set of notes of type t,

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

In the case of induced subgraphs, fixing type t, we have

$$\begin{aligned}
\mathrm{E}(s_{r,t}^{|s}|G_r) &= \mathrm{E}\left(\frac{1}{m_{r,t}}\sum_{i\in S_{r,t}}\sum_{j\in S_{r,t}} W_{ij,r}^{|s}\middle| G_r\right) \\
&= \frac{1}{m_{r,t}}\sum_{i\in V_{r,t}}\sum_{j\in V_{r,t}} W_{ij,r}\mathrm{E}\left(I\left(i,j\in S_{r,t}\right)|\, G_r\right) \\
&= \frac{1}{m_{r,t}}\sum_{i\in V_{r,t}}\sum_{j\in V_{r,t}} W_{ij,r}\left(\frac{\binom{n_{r,t}-2}{m_{r,t}-2}\prod_{q\neq t}\binom{n_{r,q}}{m_{r,q}}}{\prod_{q=1}^{T}\binom{n_{r,q}}{m_{r,q}}}\right) \\
&= \frac{1}{n_{r,t}}\sum_{i\in V_{r,t}}\sum_{j\in V_{r,t}} W_{ij,r}\left(\psi_{r,t}+o(1)\right),
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{E}(d_{r,t}^{|s}|G_r) &= \sum_{\ell\neq t}^{T}\mathrm{E}\left(\frac{1}{m_{r,t}}\sum_{i\in S_{r,t}}\sum_{j\in S_{r,\ell}} W_{ij,r}^{|s}\middle| G_r\right) \\
&= \sum_{\ell\neq t}^{T}\left(\frac{1}{m_{r,t}}\sum_{i\in V_{r,t}}\sum_{j\in V_{r,\ell}} W_{ij,r}\mathrm{E}\left(I\left(i\in S_{r,t}, j\in S_{r,\ell}\right)|\, G_r\right)\right) \\
&= \sum_{\ell\neq t}^{T}\left[\frac{1}{m_{r,t}}\sum_{i\in V_{r,t}}\sum_{j\in V_{r,\ell}} W_{ij,r}\left(\frac{\binom{n_{r,t}-1}{m_{r,t}-1}\binom{n_{r,\ell}-1}{m_{r,\ell}-1}\prod_{q\neq t,\ell}\binom{n_{r,q}}{m_{r,q}}}{\prod_{q=1}^{T}\binom{n_{r,q}}{m_{r,q}}}\right)\right] \\
&= \frac{1}{n_{r,t}}\sum_{\ell\neq t}^{T}\left(\sum_{i\in V_{r,t}}\sum_{j\in V_{r,\ell}} W_{ij,r}\psi_{r,\ell}\right).
\end{aligned}$$

Therefore, we propose to multiply each link by $\psi_{r,t}^{-1}$ for the calculation of $s_{r,t}^{|s}$ and each link between type t and type ℓ by $\psi_{r,\ell}^{-1}$ for $d_{r,t}^{|s}$.

In the case of star subgraphs, fixing type t, we have

$$\begin{aligned}
\mathrm{E}(s_{r,t}^{s}|G_r) &= \mathrm{E}\left(\frac{1}{m_{r,t}}\sum_{i\in S_{r,t}}\sum_{j\in S_{r,t}} W_{ij,r}^{s}\middle| G_r\right) \\
&= \frac{1}{m_{r,t}}\sum_{i\in V_{r,t}}\sum_{j\in V_{r,t}} W_{ij,r}\mathrm{E}\left(I\left((i\in S_{r,t}, j\in S_{r,t})\vee(i\in S_{r,t}\text{ or }j\in S_{r,t})\right)|\, G_r\right) \\
&= \frac{1}{m_{r,t}}\sum_{i\in V_{r,t}}\sum_{j\in V_{r,t}} W_{ij,r}\left(\frac{\left(\binom{n_{r,t}-2}{m_{r,t}-2}+2\binom{n_{r,t}-2}{m_{r,t}-1}\right)\prod_{q\neq t}\binom{n_{r,q}}{m_{r,q}}}{\prod_{q=1}^{T}\binom{n_{r,q}}{m_{r,q}}}\right) \\
&= \frac{1}{n_{r,t}}\sum_{i\in V_{r,t}}\sum_{j\in V_{r,t}} W_{ij,r}\left(2-\psi_{r,t}+o(1)\right),
\end{aligned}$$

and

$$
\begin{aligned}
&\mathrm{E}(d^s_{r,t}|G_r) \\
&= \sum_{\ell\neq t}^{T} \mathrm{E}\left(\frac{1}{m_{r,t}} \sum_{i\in S_{r,t}} \sum_{j\in S_{r,\ell}} W^s_{ij,r} \middle| G_r \right) \\
&= \sum_{\ell\neq t}^{T} \left(\frac{1}{m_{r,t}} \sum_{i\in V_{r,t}} \sum_{j\in V_{r,\ell}} W_{ij,r} \mathrm{E}\left(I\left((i\in S_{r,t}, j\in S_{r,\ell}) \vee (i\in S_{r,t} \text{ or } j\in S_{r,\ell}) \right) \middle| G_r \right) \right) \\
&= \sum_{\ell\neq t}^{T} \left(\frac{1}{m_{r,t}} \sum_{i\in V_{r,t}} \sum_{j\in V_{r,\ell}} W_{ij,r} \frac{\left(\binom{n_{r,t}-1}{m_{r,t}-1}\binom{n_{r,\ell}-1}{m_{r,\ell}-1} + \binom{n_{r,t}-1}{m_{r,t}}\binom{n_{r,\ell}-1}{m_{r,\ell}-1} + \binom{n_{r,t}-1}{m_{r,t}-1}\binom{n_{r,\ell}-1}{m_{r,\ell}} \right) \prod_{q\neq t,\ell} \binom{n_{r,q}}{m_{r,q}}}{\prod_{q=1}^{T} \binom{n_{r,q}}{m_{r,q}}} \right) \\
&= \frac{1}{n_{r,t}} \sum_{\ell\neq t}^{T} \left(\sum_{i\in V_{r,t}} \sum_{j\in V_{r,\ell}} W_{ij,r} \left(\psi_{r,\ell}/\psi_{r,t} + (1-\psi_{r,\ell}) + o(1) \right) \right).
\end{aligned}
$$

Therefore, we propose to multiply each link by $(2-\psi_{r,t})^{-1}$ for the calculation of $s^s_{r,t}$ and each link between type t and type ℓ by $(\psi_{r,\ell}/\psi_{r,t} + (1-\psi_{r,\ell}))^{-1}$ for $d^s_{r,t}$.

A.6 Degree distribution

In this section, we show how one can estimate the entire degree distribution from non-randomly sampled network data. To that aim, we generalize the approaches of Frank (1980, 1981) and Zhang et al. (2015). Since we observe the sizes of the (sub)population(s), our strategy targets the degree counts (rather then percentages). Let $N^{t\ell}_d(G_r)$ denote the number of nodes of type t in G_r who have d connections to nodes of type ℓ. Analogously, $N^t_d(G_r)$ is the number of nodes of type t who have degree d and $N_d(G_r) = \sum_{t\in T} N^t_d(G_r)$ is the number of nodes with degree d. $N^{t\ell}(G_r)$, $N^t(G_r)$, and $N(G_r)$ are the corresponding vectors of the numbers of nodes with different degrees in the network. Last, $\upsilon^{t\ell}(G_r)$, $\upsilon^t(G_r)$, and $\upsilon(G_r)$ stand for the largest number of links between two individuals of types t and ℓ, the maximal degree of type-t nodes, and the maximal degree in network G_r, respectively. Hence, the vectors $N^{t\ell}(G_r)$, $N^t(G_r)$, and $N(G_r)$ are of sizes $\upsilon^{t\ell}(G_r)$, $\upsilon^t(G_r)$, and $\upsilon(G_r)$.

First, under the induced subgraph sampling, the probability that a node of type t with d links to individuals of type $\ell \neq t$ in the population network G_r is selected and observed to have $d' \leq d$ links to type-ℓ nodes in $G^{|s}_r$ corresponds to the joint probability that (i) the node of type t is in the sample and (ii) d' out of her d neighbors are in the sample. Formally,

$$
P^{|s}_{r,t\ell}(d',d) = \frac{m_{r,t}}{n_{r,t}} \frac{\binom{d}{d'}\binom{n_{r,\ell}-d}{m_{r,\ell}-d'}}{\binom{m_{r,\ell}}{n_{r,\ell}}}
$$

If the network is large enough and the sampling rates are low enough,

$$
P^{|s}_{r,t\ell}(d',d) \simeq \binom{d}{d'} \frac{m_{r,t}}{n_{r,t}} \frac{m^{d'}_{r,\ell}(n_{r,\ell}-m_{r,\ell})^{d-d'}}{n^d_{r,\ell}} = \binom{d}{d'} \psi_{r,t}\psi^{d'}_{r,\ell}(1-\psi_{r,\ell})^{d-d'}
$$

For $\ell = t$,

$$P^{|s}_{r,t\ell}(d',d) = \frac{m_{r,t}}{n_{r,t}} \frac{\binom{d}{d'}\binom{n_{r,t}-1-d}{m_{r,t}-1-d'}}{\binom{m_{r,t}-1}{n_{r,t}-1}} \simeq \binom{d}{d'} \frac{m_{r,t}}{n_{r,t}} \frac{(m_{r,t}-1)^{d'}(n_{r,t}-m_{r,t})^{d-d'}}{(n_{r,t}-1)^{d}}$$
$$= \binom{d}{d'} \psi_{r,t}[\psi_{r,t}+o(1)]^{d'}[1-(\psi_{r,t}+o(1))]^{d-d'}$$
$$\approx \binom{d}{d'} \psi_{r,t} \left(\psi_{r,\ell}\right)^{d'} \left(1-\psi_{r,\ell}\right)^{d-d'}, \quad \text{for } 0 \leq d' \leq d \leq \upsilon^{t\ell}(G_r).$$

Naturally, $P^{|s}_{r,t\ell}(d',d) = 0$ for $d' > d$.[2] Let $P^{|s}_{r,t\ell}$ denote the $\upsilon^{t\ell}(G_r) \times \upsilon^{t\ell}(G_r)$ matrix of all such probabilities for any $d, d' \leq \upsilon^{t\ell}(G_r)$. Note that there are T^2 such matrices, one for each (t,ℓ) pair (including pairs of the same type). The conditional expectation for the number of sampled nodes of type t have degree d to type ℓ in $G^{|s}_r$ is:

$$E[N^{t\ell}_d(G^{|s}_r)|G_r] = \sum_{j=d}^{\upsilon^{t\ell}(G_r)} \binom{j}{d} \psi_{r,t} \left(\psi_{r,\ell}\right)^{d} \left(1-\psi_{r,\ell}\right)^{j-d} N^{t\ell}_j(G_r).$$

Also, $E[N_d(G^{|s}_r)|G_r] = \sum_{t\in T}\sum_{\ell\in T} E[N^{t\ell}_d(G^{|s}_r)|G_r]$. The *naive* estimators for $N^{t\ell}(G_r)$ and $N(G_r)$ then are $\widetilde{N}^{t\ell}(G^{|s}_r) = (P^{|s}_{r,t\ell})^{-1} N^{t\ell}(G^{|s}_r)$ and $\widetilde{N}(G^{|s}_r) = \sum_{t\in T}\sum_{\ell\in T} \widetilde{N}^{t\ell}(G^{|s}_r)$.[3]

The matter is simpler under the star subgraph sampling because the true degree of each sampled node is observed without error. Hence, the probability that a type-t node has degree d in the true population network G_r have degree d' in G^s_r corresponds to the probability that she is sampled. That is, $P^s_{r,t}(d',d) = \psi_{r,t}$ if $d = d'$ and $P^s_{r,t}(d',d) = 0$ otherwise. Consequently, $P^s_{r,t} = \psi_{r,t}\mathbb{1}_{\upsilon^t(G_r)}$, is a $\upsilon^t(G_r) \times \upsilon^t(G_r)$ diagonal matrix with the type-specific sampling rate $\psi_{r,t}$ on the main diagonal that summarizes all these probabilities. In this case, we only need T such $P^s_{r,t}$ matrices, one for each type. As a result, $E[N_d(G^s_r)|G_r] = E[\sum_t N^t_d(G^s_r)|G_r] = \sum_t \psi_{r,t} N^t_d(G_r)$ and the estimators of $N^t_d(G_r)$ and $N_d(G_r)$ are $\widetilde{N}^t_d(G^s_r) = \psi^{-1}_{r,t} N^t_d(G^s_r)$ and $\widetilde{N}_d(G^s_r) = \sum_t \widetilde{N}^t_d(G^s_r)$. In matrix terminology, $\widetilde{N}(G^s_r) = \sum_t (P^s_{r,t})^{-1} N^t(G^s_r)$.

Observe that, if $\psi_{r,t} = \psi_{r,\ell} = \psi_r$ for each $t,\ell \in T$, $E[N(\overline{G}_r)|G_r] = \overline{P}_r N(G_r)$ where $\overline{G}_r \in \{G^{|s}_r, G^s_r\}$ and $\overline{P}_r \in \{P^{|s}_r, P^s_r\}$ and the naive estimators for $N(G_r)$ collapse to $\widetilde{N}(\overline{G}_r) = \overline{P}_r^{-1} N(\overline{G}_r)$, the estimator proposed by Frank (1980, 1981).

The proposed corrections notwithstanding, Zhang et al. (2015) show that the estimators of Frank (1980, 1981) are ill-posed in two respects. The operators $\overline{P}_r$'s are not necessarily invertible in general and, even if they are, the elements of $\widetilde{N}(\overline{G}_r)$ may be non-negative. The particular matrices $P^{|s}_{r,t\ell}$ and $P^s_{r,t}$ are invertible because the former is upper triangular and the latter is diagonal. However, they can still generate negative estimates of the number of nodes of certain degree. In fact, our simulations corroborate this concern for induced subgraphs. To

[2] For example, the probability that a sampled node of type t with two friends of type ℓ in G_r has degree 1 in $G^{|s}_r$ (that is, the element (1,2) of matrix $P^{|s}_{r,t\ell}$ defined below) is $P^{|s}_{r,t\ell}(1,2) = \binom{2}{1}\psi_{r,t}\left(\psi_{r,\ell}\right)\left(1-\psi_{r,\ell}\right)$. This probability takes into account that the node in question is sampled and has two neighbors of type ℓ and that one of these neighbors is sampled while the other is not.

[3] Zhang et al. (2015) show that the estimator of Frank (1980, 1981) is ill-posed. They thus label this estimator as *naive*. We follow this terminology.

overcome these issues, Zhang et al. (2015) propose a constrained, penalized weighted least-squares estimator which avoids the inversion of $\overline{P}_r$'s and for which the estimator $\hat{N}(\bar{G}_r) \geq 0$ by construction.

Since a naive inversion of the matrices a là Frank (1980, 1981) is problematic, Zhang et al. (2015) propose a constrained, penalized weighted least-squares estimation framework, which we extend as follows. $\hat{N}(\overline{G}_r)$, $\overline{G}_r \in \{G_r^{|s}, G_r^s\}$, is the solution resulted from the following minimization problem:

$$\begin{aligned} \min_{N} \quad & (PN - \overline{N})^T \overline{C}^{-1} (PN - \overline{N}) + \lambda \cdot \text{pen}(N) \\ \text{s.t.} \quad & N_i \geq 0, i = 0, 1, ..., \upsilon(G_r), \\ & \sum_{i=0}^{\upsilon^t(G_r)} N_i^t = n_t, t \in \{1, ..., T\}. \end{aligned}$$

where N is the estimated vector of degree counts with an element N_i.[4] P and $\overline{N}$ are respectively the operator of the problem and the degree counts in the sampled graph (both defined below) and $\overline{C}$ is (the approximation of) the covariance matrix of $\overline{N}$. λ is a smoothing parameter and pen(N) reflects the penalty on the complexity of N.

In case of the induced subgraphs, consider $N = [N^{t\ell}(G_r)]_{t\ell \in T}$ and $\overline{N} = [N^{t\ell}(G_r^{|s})]_{t\ell \in T}$. The operator P is constructed as follows:

$$P = \begin{bmatrix} P_{11}^{|s} & 0 & \cdots & 0 \\ 0 & P_{12}^{|s} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & P_{T\times T}^{|s} \end{bmatrix}$$

Last, Zhang et al. (2015) show that, even though $C^{|s} = \text{cov}(\overline{N})$ has non-zero off-diagonal elements under the induced subgraph sampling, it is well approximated by a diagonal matrix[5]

$$C^{|s} = \text{diag}(\overline{N}_{smooth}) + \delta \mathbb{1}. \tag{3}$$

The term $\text{diag}(\overline{N}_{smooth})$ in (3) is a diagonal matrix with the diagonal elements being equal to the smoothed version of the observed degree counts $\overline{N}$, for which they propose to employ the smoothing method of Dong and Simonoff (1994). The second part of (3) ensures that the approximation of $C^{|s}$ is positive definite. Zhang et al. (2015) discuss the choice of δ in detail.

As for the star subgraph, $N = [N^t(G_r)]_{t\in T}$, $\overline{N} = [N^t(G_r^s)]_{t\in T}$, and

$$\overline{P} = \begin{bmatrix} P_1^s & 0 & \cdots & 0 \\ 0 & P_2^s & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & P_T^s \end{bmatrix}.$$

Since the covariance matrix is diagonal under the star sampling scheme, no approximation is necessary and

[4] Note that $\sum_{i=0}^{\upsilon^t(G_r)} N_i^t = n_t$ for each t implies that $\sum_i N_i = n$.

[5] $\mathbb{1}$ is an identity matrix of a size corresponding to the dimension of the problem.

$$C^s = \begin{bmatrix} \psi_{r,1}(1-\psi_{r,1})\text{diag}[N^1(G_r^s)] & 0 & \cdots & 0 \\ 0 & \psi_{r,2}(1-\psi_{r,2})\text{diag}[N^2(G_r^s)] & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \psi_{r,T}(1-\psi_{r,T})\text{diag}[N^T(G_r^s)] \end{bmatrix}.$$

For both sampling schemes, we refer to Zhang et al. (2015) for an exhaustive discussion and analysis of the selection of both the smoothing parameter λ and the penalty function $\text{pen}(N)$ in the above optimization problem. Under a convex penalty, the above minimization problem belongs to the class of convex optimization problems and standard software can be applied. Since Zhang et al. (2015) recommend the use of an ℓ_2 norm, the optimization becomes a quadratic programming exercise. The estimated degree counts $\hat{N}(\overline{G}_r)$, $\overline{G}_r \in \{G_r^{|s}, G_r^s\}$, are obtained by adding up the estimated $\hat{N}_{t\ell}(G_r^{|s})$ for the induced subgraph and $\hat{N}_t(G_r^s)$ for the star subgraph case.

B Additional Empirical Results

Table B.1: Estimated network effects on the share of population in rural India village that (I) employed and (II) work outside the village – based on friendship network

Dependent Variable	Employed(%)			Work Outside Village(%)		
	Raw	Random	Cross	Raw	Random	Cross
Degree	0.0425**	0.0161**	0.0120	-0.0216	-0.0103	-0.0118
	(0.0197)	(0.0075)	(0.0076)	(0.0213)	(0.0083)	(0.0090)
Cluster	0.2490***	0.2490***	0.1272**	0.0194	0.0194	0.0464
	(0.0714)	(0.0714)	(0.0555)	(0.0935)	(0.0935)	(0.0629)
Span	0.0001	-0.0102	-0.0033	0.0001	0.0145	0.0022
	(0.0002)	(0.0097)	(0.0077)	(0.0001)	(0.0103)	(0.0083)
Epid Thrld	-0.4304**	-0.7882*	-0.4498	0.2627	0.6378	0.4709
	(0.2005)	(0.3997)	(0.3608)	(0.1893)	(0.3876)	(0.3552)
HI-sex	0.3504***	0.3504***	0.2745**	-0.0008	-0.0008	0.0926
	(0.1107)	(0.1107)	(0.1106)	(0.1401)	(0.1401)	(0.1416)
HI-age	0.4039***	0.4039***	0.0603	-0.1427	-0.1427	0.4020**
	(0.1437)	(0.1437)	(0.1478)	(0.2374)	(0.2374)	(0.2086)
HI-householdsize	0.0377	0.0377	-0.0382	0.1901**	0.1901**	0.1004
	(0.0842)	(0.0842)	(0.0832)	(0.0820)	(0.0820)	(0.0968)
HI-cross	0.1905	0.1905	0.0181	0.2488*	0.2488*	0.4177***
	(0.1221)	(0.1221)	(0.1166)	(0.1344)	(0.1344)	(0.1362)

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

Table B.2: Effects of clustering and graph span on club participation and frequency of exercise in U.S. high schools.

Dependent Variable	Number of clubs attended			Frequency of exercise		
	Raw	Random	Cross	Raw	Random	Cross
Cluster	0.0519	-0.3535	-0.3891	1.1620**	0.3139	-0.3891
	(0.4294)	(0.3895)	(0.3718)	(0.5264)	(0.6989)	(0.3718)
Span	-0.0010	-0.3079**	-0.3779***	-0.0103***	-0.1769*	-0.3779***
	(0.0014)	(0.1160)	(0.1237)	(0.0019)	(0.1152)	(0.1237)

Note: Regression is based on 48 schools. *, **, *** stand for significance at 10%, 5%, and 1% respectively. The model includes both the clustering coefficient and graph span as network regressors, and has the school size included as a default control.

C Additional Figures

In this section, we report two types of figures: (i) the normalized root mean-square error (RMSE), and (ii) the simulation results on the bounds of the maximal eigenvalues of a network

First, in order to assess the bias-variance trade-off in the random corrections and our post-stratification weighting approach, Figures C.1 - C.4 plot the RMSEs corresponding to Figures 1 - 4 in the main text with the same layout.

Second, Figures C.5 and C.6 for induced and star subgraphs, respectively, present the results of our simulation exercise for the bounds of the maximal eigenvalue in terms of box plots with estimates from 100 repetitions. Again, we compare the population values (lines), raw data, as well as the two types of corrections (box plots). We only consider the sampling rates $\psi = 80\%$ and $\psi = 40\%$ for simplicity. The results for the intermediate sampling rate are in line with those reported here. While the two lower bounds ($d(G)$ and $\sqrt{d_s(G)}$) and the upper bound (U) (from Section 3.7) are all positive, they are quite different in magnitudes, so for a better exposition, we transform them into reciprocals. The blue lines in the figures represent the true value of the bounds and the red line represents the true maximal eigenvalue.

The results clearly show that the bounds computed from the raw sample are largely biased from their actual values upwards in the reciprocal form (i.e., downwards in their original form). As expected, the biases scale down with the sampling rate. The random corrections eliminate the biases in scenario R, but large biases remain under non-random sampling. The corrections based on the missing-at-random assumption may even change the sign of the biases compared to raw data and remain large. In contrast, our poststratification weighting is remarkably successful in eliminating the biases in the bounds.

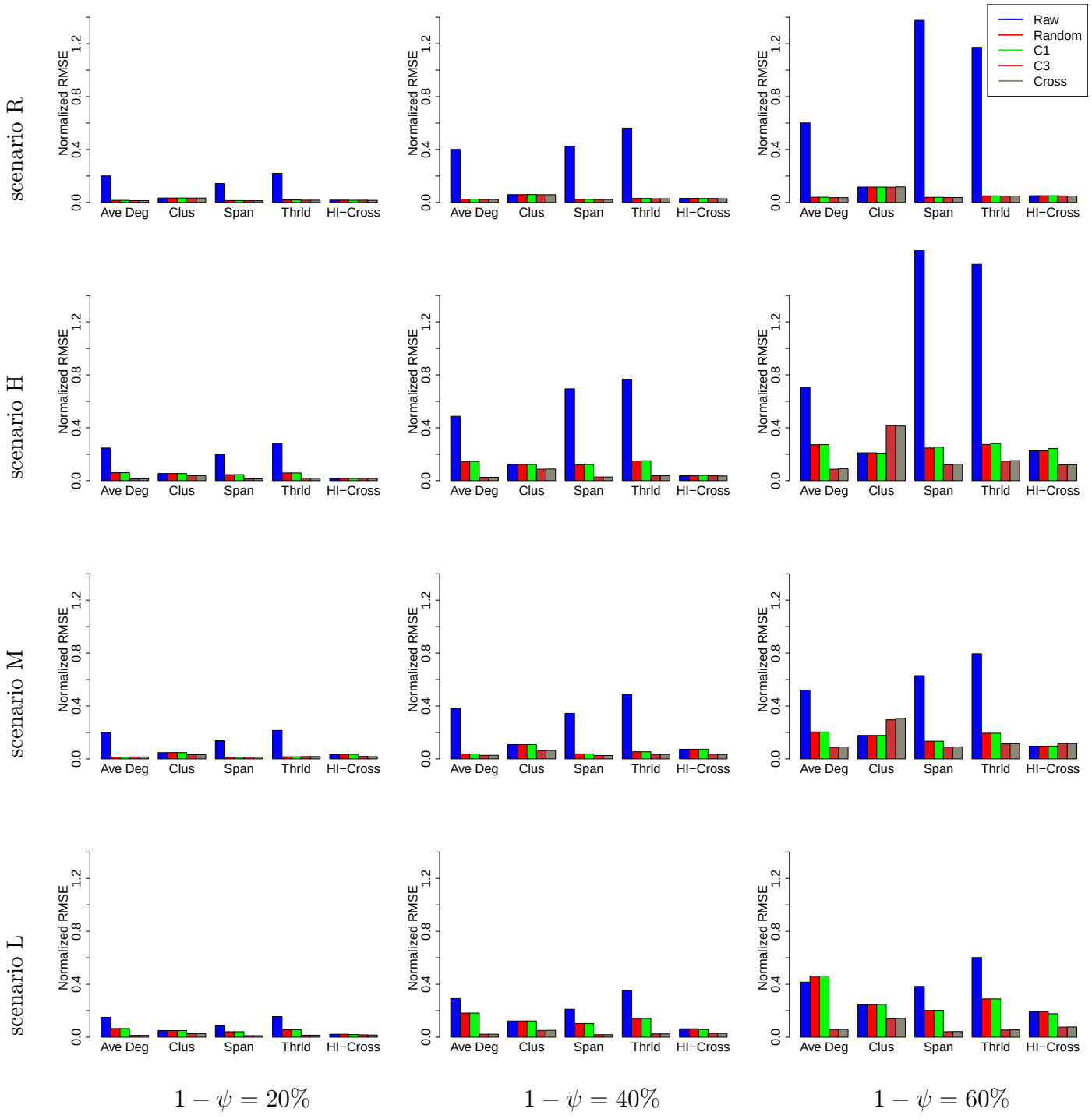

Figure C.1: Induced subgraph. Normalized root mean square error (RMSE) of five estimated network measures (average degree, clustering coefficient, graph span, epidemic threshold, and homophily index of cross-characteristics) from the raw sample and their corrected versions by weighting for three different removal rates 20% (left), 60% (center), 40% (right) and four different removal scenarios, R, H, M, and L. The normalized RMSE is computed based on 1,000 simulation repetitions.

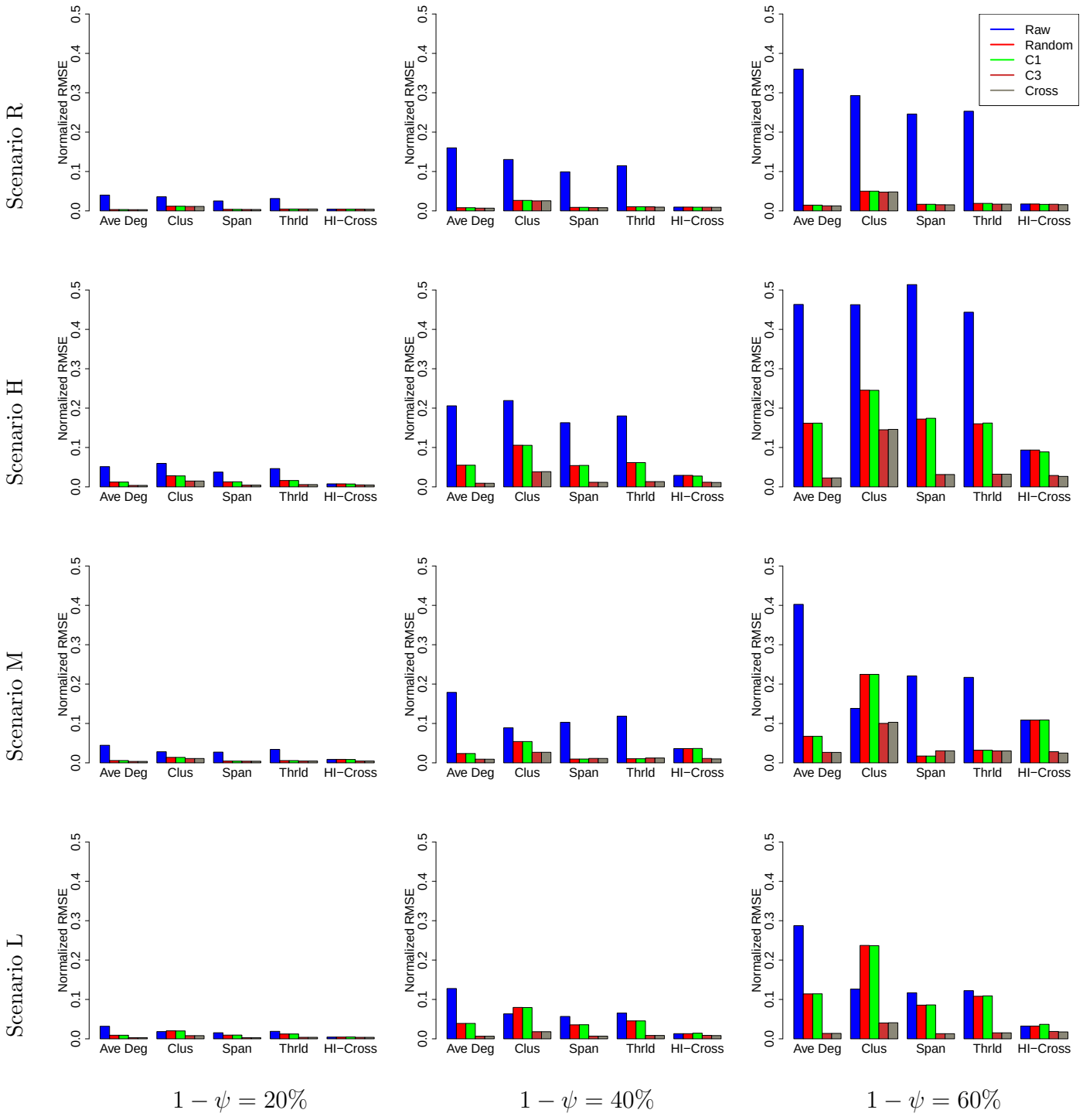

Figure C.2: Star subgraph. Normalized root mean square error (RMSE) of five estimated network measures (average degree, clustering coefficient, graph span, epidemic threshold, and homophily index of cross-characteristics) from the raw sample and their corrected versions by weighting for three different removal rates 20% (left), 60% (center), 40% (right) and four different removal scenarios, R, H, M, and L. The normalized RMSE is computed based on 1,000 simulation repetitions.

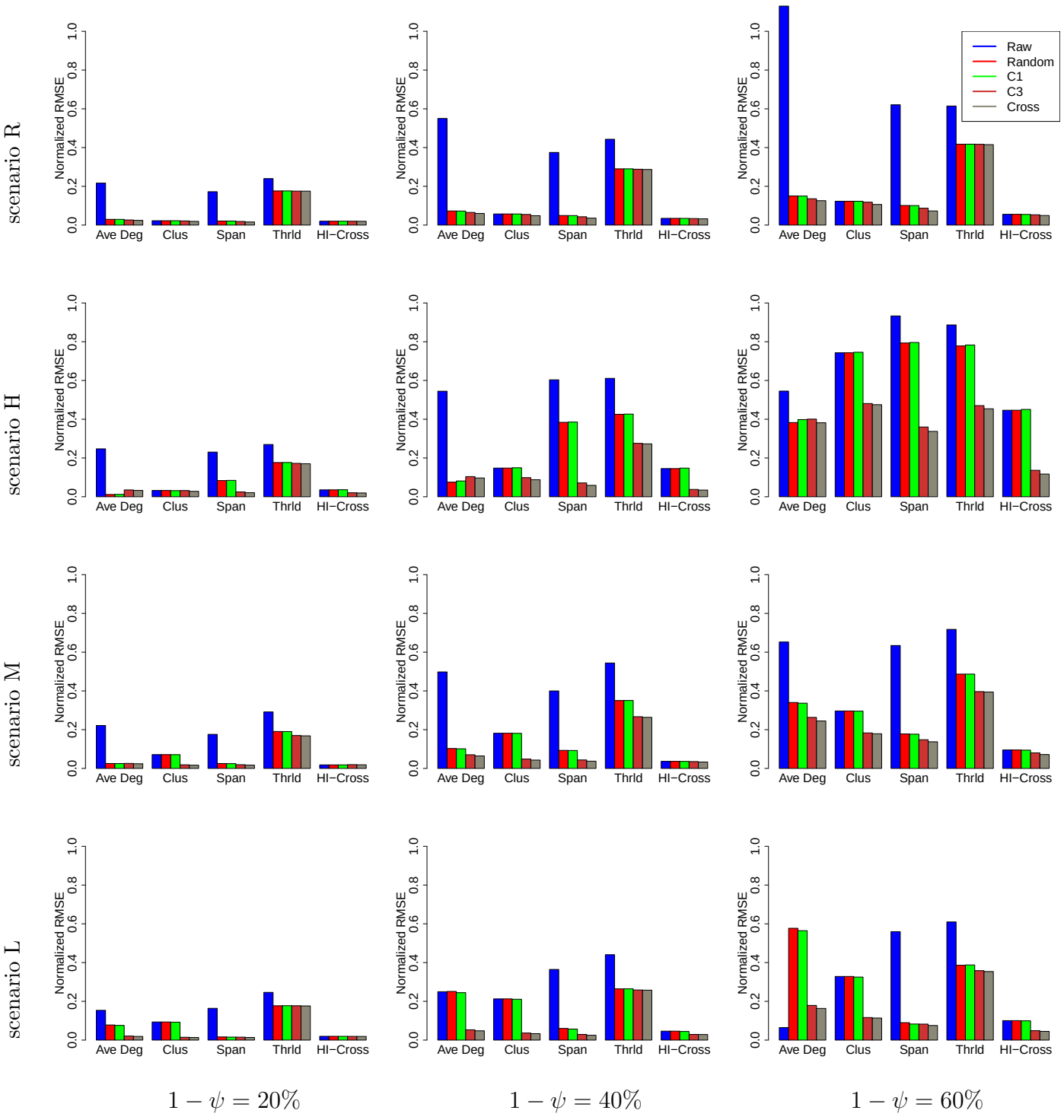

Figure C.3: Induced subgraph. Normalized root mean square error (RMSE) of five estimated network effects (average degree, clustering coefficient, graph span, epidemic threshold, and homophily index of cross-characteristics) from the raw sample and their corrected versions by weighting for three different removal rates 20% (left), 60% (center), 40% (right) and four different removal scenarios, R, H, M, and L. The normalized RMSE is computed based on 1,000 simulation repetitions.

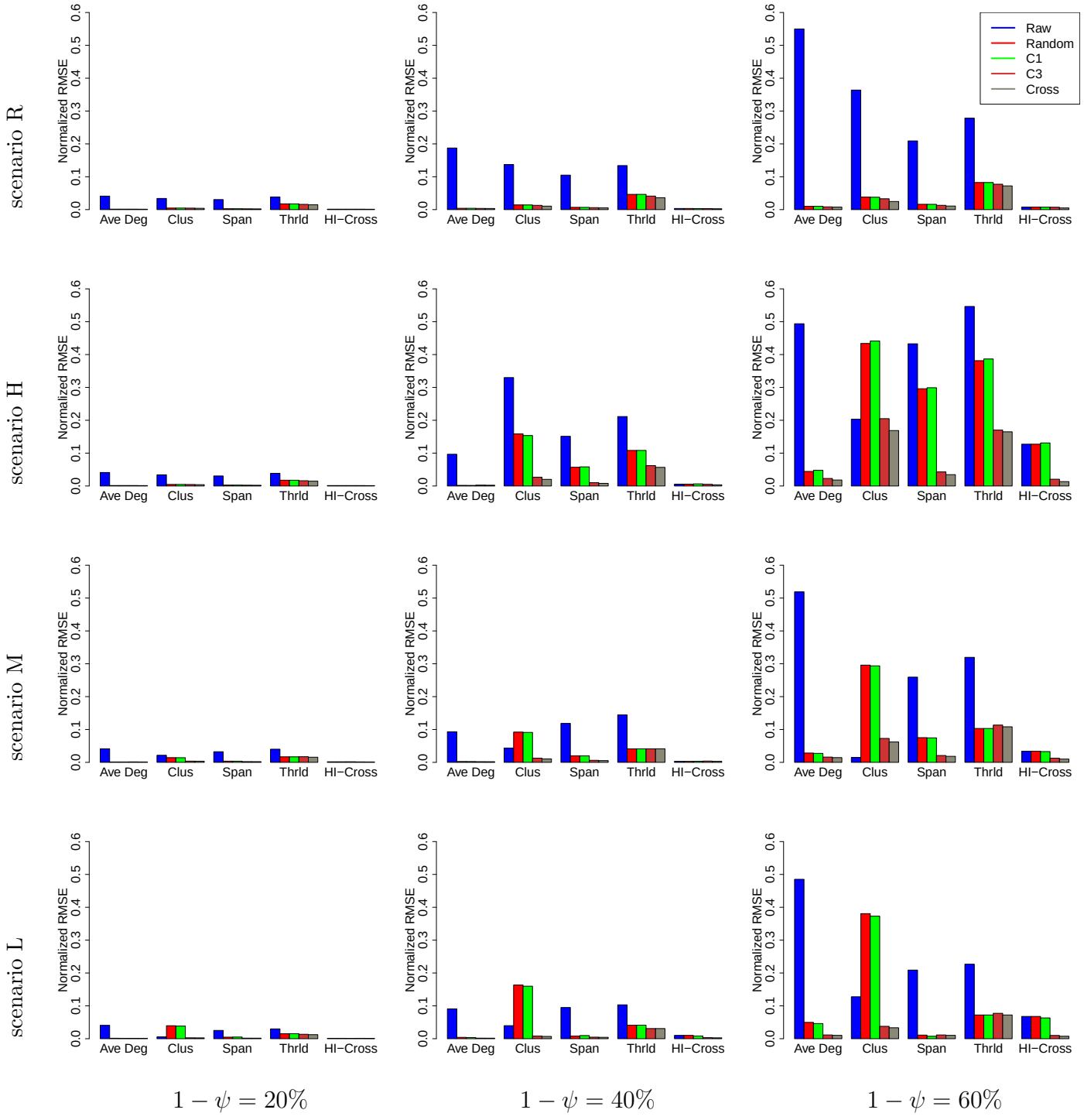

Figure C.4: Star subgraph. Normalzied root mean square error (RMSE) of five estimated network effects (average degree, clustering coefficient, graph span, epidemic threshold, and homophily index of cross-characteristics) from the raw sample and their corrected versions by weighting for three different removal rates 20% (left), 60% (center), 40% (right) and four different removal scenarios, R, H, M, and L. The normalized RMSE is computed based on 1,000 simulation repetitions.

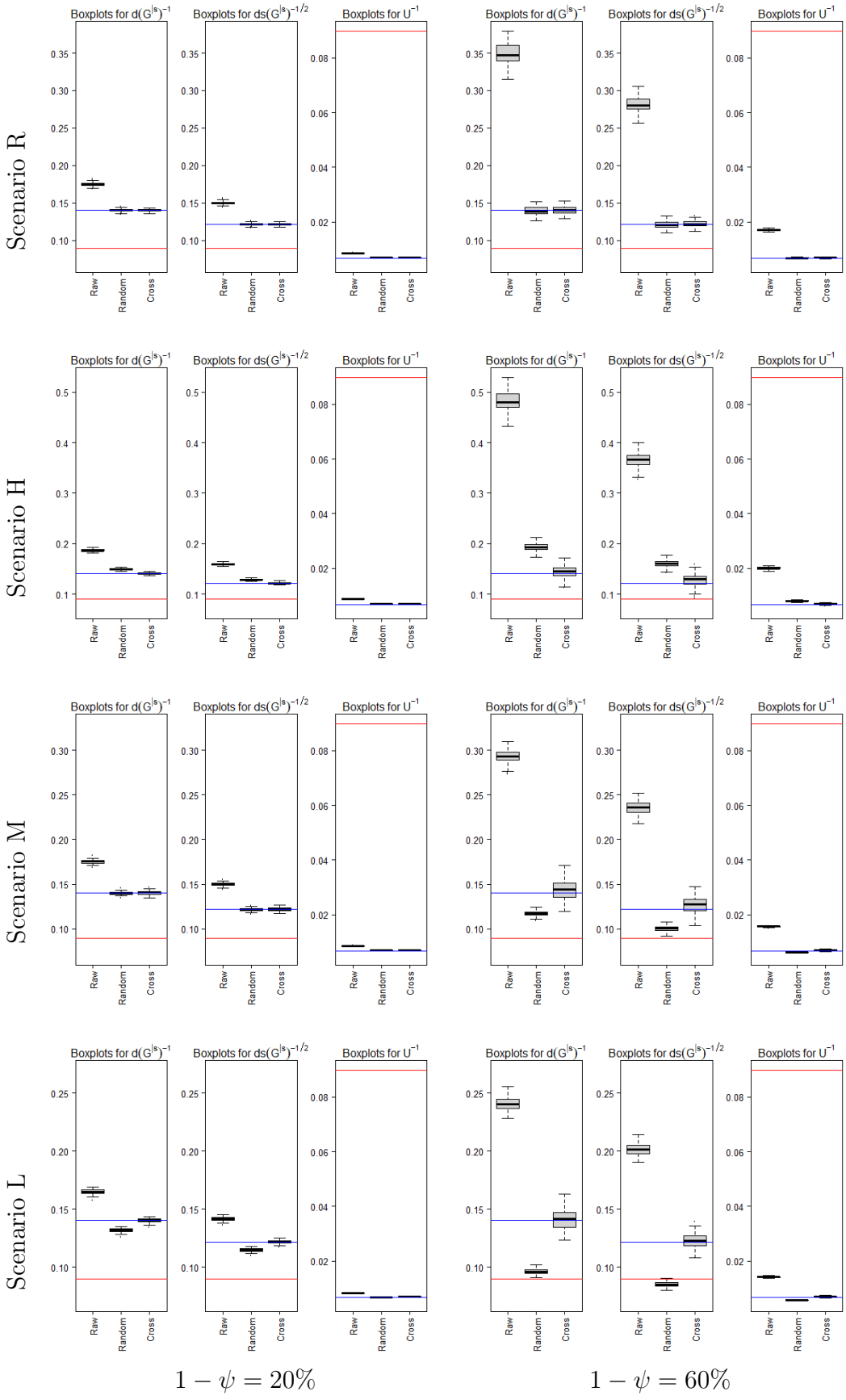

Figure C.5: Induced subgraph: Boxplots of bounds corrections for the maximal eigenvalue with respect to the population network for $\psi = 80\%$ (left), and 40% (right) and four different removal scenarios.

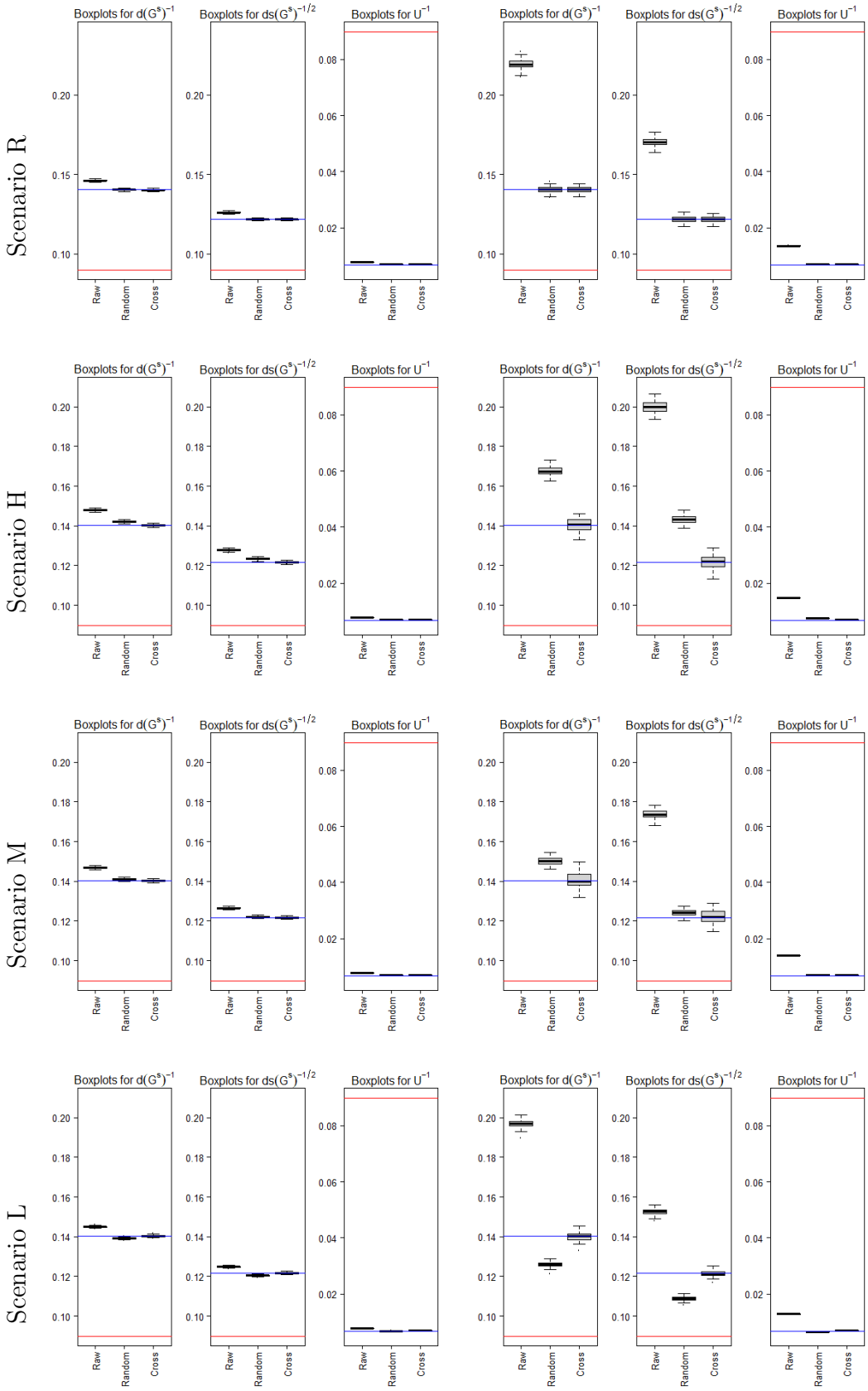

Figure C.6: Star subgraph: Boxplots of bounds corrections for the maximal eigenvalue with respect to the population network for $\psi = 80\%$ (left), and 40% (right) and four different removal scenarios.