Electronic Journal of Statistics
Vol. 14 (2020) 1577–1610
ISSN: 1935-7524
https://doi.org/10.1214/20-EJS1698

Asymptotic seed bias in respondent-driven sampling

Yuling Yan

Department of Operations Research and Financial Engineering, Princeton University
Sherrerd Hall, Charlton Street, Princeton, NJ 08544, USA
e-mail: yulingy@princeton.com

Bret Hanlon

Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison
610 Walnut Street, Madison, WI 53726, USA
e-mail: bret.hanlon@wisc.edu

Sebastien Roch

Department of Mathematics, University of Wisconsin-Madison
480 Lincoln Drive, Madison, WI 53792, USA
e-mail: roch@math.wisc.edu

and

Karl Rohe

Department of Statistics, University of Wisconsin-Madison
1300 University Avenue, Madison, WI 53706, USA
e-mail: karlrohe@stat.wisc.edu

Abstract: Respondent-driven sampling (RDS) collects a sample of individuals in a networked population by incentivizing the sampled individuals to refer their contacts into the sample. This iterative process is initialized from some seed node(s). Sometimes, this selection creates a large amount of seed bias. Other times, the seed bias is small. This paper gains a deeper understanding of this bias by characterizing its effect on the limiting distribution of various RDS estimators. Using classical tools and results from multi-type branching processes [12], we show that the seed bias is negligible for the Generalized Least Squares (GLS) estimator and non-negligible for both the inverse probability weighted and Volz-Heckathorn (VH) estimators. In particular, we show that (i) above a critical threshold, VH converge to a non-trivial mixture distribution, where the mixture component depends on the seed node, and the mixture distribution is possibly multi-modal. Moreover, (ii) GLS converges to a Gaussian distribution independent of the seed node, under a certain condition on the Markov process. Numerical experiments with both simulated data and empirical social networks suggest that these results appear to hold beyond the Markov conditions of the theorems.

MSC 2010 subject classifications: Primary 62D05; secondary 60J20.
Keywords and phrases: Limit distribution, Galton-Watson process, Volz-Heckathorn estimator.

Received August 2019.

Contents

1. Introduction

Network sampling techniques, including web crawling, snowball sampling, and respondent-driven sampling (RDS), contact individuals in hard-to-reach populations by following edges in a social network. This paper uses RDS as a motivating example [9]. It is used by the Centers for Disease Control (CDC) and the Joint United Nations Programme on HIV/AIDS (UN-AIDS) to sample populations most at risk for HIV (injection drug users, sex workers, and men who have sex

with men) [4, 11]. In the most recent survey of the literature [20], RDS had been applied in over 460 different studies, in 69 different countries.

An RDS sample is initialized with one or more "seed individuals" selected by convenience from the population. These individuals participate in the survey and are incentivized to refer additional participants (often up to 3 or 5 participants) into the sample. This process iterates until reaching the target sample size or there are no referrals. All participants are incentivized to take a survey and an HIV test. With this sample, we wish to estimate the proportion of individuals in the population that are HIV+.

The Markov model for the RDS process has provided fundamental insight into RDS sampling [17, 6, 16]. For example, nodes with more connections are more likely to be sampled [13]. This creates bias and there are ways to adjust for it [17, 18]. While the inverse probability weighted (IPW) estimator requires a normalizing constant that is unknown in practice, the Volz-Heckathorn (VH) estimator provides a way to estimate this normalizing constant [18]. More recently, [16] studied the variability of the IPW estimators and showed that there are two regimes (low variance and high variance). This regime is determined by two parameters of the Markov process that is described in Section 2.1. In brief, let λ_2 be the second eigenvalue of the Markov transition matrix on the social network and let m be the average number of referrals provided by each node. When $m < \lambda_2^{-2}$, the variance of the IPW estimator decays at rate n^{-1}, where n is the sample size. However, when $m > \lambda_2^{-2}$, the variance of IPW decays at a slower rate. Later, [14] showed that the VH and IPW estimators are asymptotically normal under the Markov model in the low variance regime. More recently, [15] proposed a generalized least squares (GLS) estimator for the high variance regime and showed that the variance of this estimator is $O(n^{-1})$, even when $m > \lambda_2^{-2}$. These previous results are summarized in Table 1.

This paper studies the limit distribution of (i) the GLS estimator and (ii) the IPW estimator in the high variance regime. These results also allow for the Volz-Heckathorn adjustment. For technical reasons, our analysis of the GLS estimator is restricted to a special case of the Markov model that was first used to study RDS in [6].

These technical results make many unrealistic assumptions which we discuss below. In particular, the Markov model allows for resampling of individuals. The results are asymptotic in the sample size, while the population size is fixed. This

TABLE 1

Summary of properties of IPW and GLS estimators. In the columns, m refers to the number of participants that the typical participant refers into the study and λ_2 is the second eigenvalue of the Markov transition matrix.

Result	Estimator	Low variance, i.e. $m < \lambda_2^{-2}$	High variance, i.e. $m > \lambda_2^{-2}$
Variance	IPW	$O(n^{-1})$ [16]	$O(n^{2\log_m \lambda_2})$ [16]
	GLS	$O(n^{-1})$ [15]	
Distribution	IPW&VH	Asymptotically normal [14]	Non-trivial mixture [Current Paper]
	GLS	Asymptotically normal [Current Paper]	

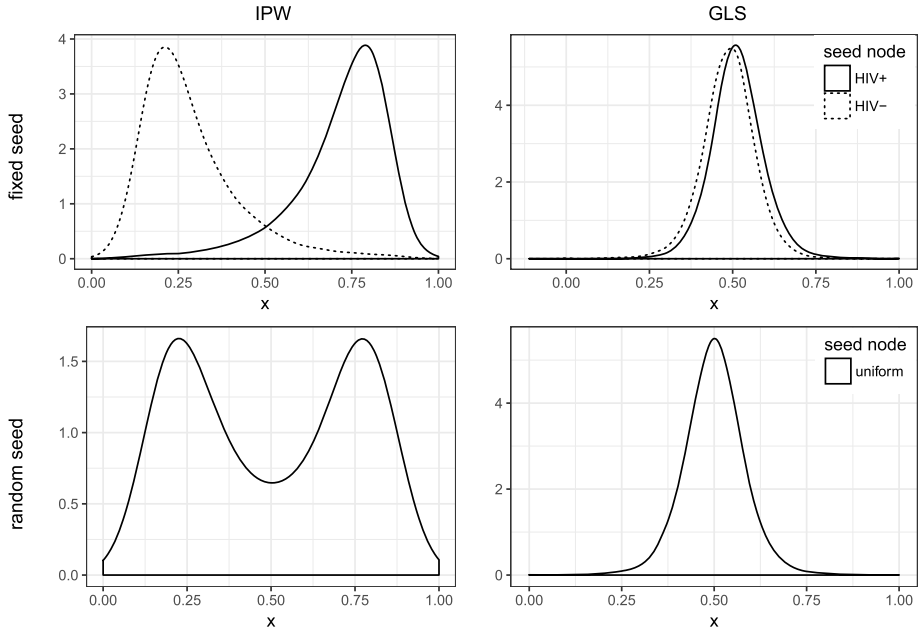

FIG 1. *The model for this simulation is described in Section 1.1. The two left panels show the distribution of sample proportion (i.e. the IPW estimator in this model). The two right panels show the distribution of GLS estimator. Each panel in the top row has two curves corresponding to whether or not the seed node is HIV+. The solid line gives the distribution of the estimator when the process is initialized with an HIV+ node. The dashed line is initialized with an HIV- node. In the bottom row, the seed participant is selected uniformly at random. This figure demonstrates how the limit distribution of the IPW estimator can have two modes which correspond to whether the seed is HIV+ or HIV-. Moreover, the figure suggests that the GLS estimator is asymptotically normal and the dependence on the seed node is negligible.*

creates extensive resampling. Nevertheless, this model provides fundamental insights into the properties of the estimators and these properties continue to hold under more realistic simulation models in Sections 4 and 5.

1.1. A simple motivating example

Here we consider a model studied in [6], which we refer to as the Blockmodel with 2 blocks. In this example, the population that we wish to sample is equally divided into two groups: HIV+ and HIV-. The seed participant is selected from one of the two groups with equal probability. Each participant refers an iid number of offspring, generated from some offspring distribution. With probability p, the referred participant matches the HIV status of the participant that referred them. With probability $1-p$, their statuses differ. Each referral is independent, conditional on the status of the referring participant. Using a sample generated in this way, we wish to estimate the proportion of the population that is HIV+ (in this case, the true proportion is 0.5).

Figure 1 displays a motivating simulation from this Blockmodel with 2 blocks. Each sample size is 1000 individuals, sampled from the Blockmodel with $p = .95$ and offspring distribution $1 + \text{Binomial}(2, 0.5)$. For each sample of 1000, we construct both sample proportion (equivalent to the IPW estimator, see Section 2.4) and GLS estimator. This process is repeated 10000 times. Figure 1 displays a kernel density estimate of the resulting distribution.

1.2. Main contributions

Many RDS papers discuss the "bias from seed selection". Section 3.1 shows that the IPW and VH estimators have a limit distribution and this limit distribution depends on where the process is initialized (i.e. the "seed" node). If the seed node is randomized, then in simulations, the limit distribution of the IPW and VH estimators can have multiple modes, where each mode corresponds to a different set of initial conditions. The limit results for the IPW and VH estimators highlight how, conditioned on the seed node, the bias of these estimators decays at the same rate as the variance. So, unconditional on the seed node, this can create multiple modes in the limit distributions of the IPW and VH estimators. Similarly to classical results in multitype branching process theory [12], the exact limit distribution does not appear to have a concise and easily interpretable closed form.

While the IPW and VH estimators are not asymptotically normal in the high variance regime, Section 3.2 shows that the GLS estimator is asymptotically normal in this regime and this limit distribution does not depend on where the process is initialized. This pair of results provides additional insight into the notions of "bias" and "variance" for network sampling. In particular, the GLS estimator is the linear estimator with the smallest variance and that measure of variance includes the variability that comes from selecting the seed node (i.e. from the stationary distribution of the Markov process). Hence, it adjusts for the seed selection. Another way of saying this is that the GLS estimator reduces "the bias from seed selection". This blurring of the divide between "variance" and "bias from seed selection" highlights one potential problem of conditioning on the seed node in a bootstrap resampling procedure [2]; in the high variance regime, conditioning on the seed node removes a large source of variability in the VH estimator.

2. Background and notation

This section (i) defines the Markov model, (ii) illustrates how this model is particularly tractable when the underlying network is a Blockmodel [19], and (iii) defines the IPW, VH, and GLS estimators.

2.1. Markov model

The Markov model consists of (1) a social network represented as a graph, (2) a Markov transition matrix on the nodes of the graph, (3) a referral tree to

index the Markov process on the graph, and finally, (4) a node feature defined for each node in the graph. Each of these are defined below.

The results in this paper allow for an undirected, weighted graph. Let $G = (V, E)$ be a graph with vertex set $V = \{1, \ldots, N\}$ containing the people and edge set $E = \{(i, j) : i, j \in V \text{ are connected}\}$ containing the friendships. Let w_{ij} be the weight of the edge $(i, j) \in E$. For notational convenience, define $w_{ij} = 0$ if $(i, j) \notin E$. If the graph is unweighted, define $w_{ij} = 1$ for all $(i, j) \in E$. Throughout this paper, the graph is undirected (i.e. $w_{ij} = w_{ji}$ for all pairs (i, j)). Define the degree of node i as $\deg(i) = \sum_j w_{ij}$ and the volume of the graph as $\text{vol}(G) = \sum_i \deg(i)$. For simplicity, $i \in G$ is used synonymously with $i \in V$. Define the Markov transition matrix $\boldsymbol{P} \in \mathbb{R}^{N \times N}$ as

$$P_{ij} = \frac{w_{ij}}{\deg(i)}. \tag{2.1}$$

Since G is undirected, $\boldsymbol{P}$ is a reversible Markov transition matrix with a stationary distribution $\boldsymbol{\pi} : G \to \mathbb{R}$ with $\pi(i) = \deg(i)/\text{vol}(G)$.

The referral tree is a rooted tree, i.e. a connected graph with n nodes, no cycles, and a vertex 0. This tree, $\mathbb{T}$, can be random (a Galton-Watson tree with expected offspring number m) or nonrandom (an m-tree, where each node has exactly m offspring). If $\mathbb{T}$ is randomly generated, then the Markov process is conditioned on the tree. For simplicity, $\sigma \in \mathbb{T}$ is used synonymously with σ belonging to the vertex set of $\mathbb{T}$. The seed participant is the root vertex 0 in $\mathbb{T}$. For each non-root node $\sigma \in \mathbb{T}$, denote $p(\sigma) \in \mathbb{T}$ as the parent of σ (i.e. the node one step closer to the root).

Assume that the nodes are sampled with a Markov process that is indexed by $\mathbb{T}$: each node $\sigma \in \mathbb{T}$ corresponds to an individual X_σ sampled from the population G, and an edge (σ, τ) of $\mathbb{T}$ denotes that the sampled individual X_σ referred the individual X_τ into the sample. Mathematically, let $\{X_\sigma^{(\cdot)} : \sigma \in \mathbb{T}\}$ be a tree-indexed Markov process on the individuals from the social network G:

$$\mathbb{P}(X_\sigma^{(\cdot)} = j \mid X_{p(\sigma)}^{(\cdot)} = i, X_\tau^{(\cdot)} : \tau \in \mathscr{D}(\sigma)^c) = \mathbb{P}(X_\sigma^{(\cdot)} = j \mid X_{p(\sigma)} = i) = P_{ij},$$

where $\mathscr{D}(\sigma) \subset \mathbb{T}$ denotes the set of σ and all its descendants in $\mathbb{T}$. The superscript $(\cdot)$ indicates the initial condition: if the superscript is some $i \in G$, X_0 is initialized from i; if the superscript is some distribution $\boldsymbol{\nu} : G \to \mathbb{R}$ (e.g. the stationary distribution $\boldsymbol{\pi}$ of $\boldsymbol{P}$), X_0 is initialized from $\boldsymbol{\nu}$. When the initial state does not matter, we leave off the superscript. Following [3], we call this process a $(\mathbb{T}, \boldsymbol{P})$-walk on G.

In a special case, $\mathbb{T}$ can be the chain graph $(0-1-2-3-\ldots)$; this results in the model being a Markov chain. Just as a chain graph indexes a Markov chain, the graph $\mathbb{T}$ provides the indexing in this model. For simplicity, $\sigma \in \mathbb{T}$ is used synonymously with σ belonging to the vertex set of $\mathbb{T}$. The seed participant is root vertex 0 in $\mathbb{T}$. For each non-root node $\sigma \in \mathbb{T}$, denote $p(\sigma) \in \mathbb{T}$ as the parent of σ (i.e. the node one step closer to the root). Assume that the nodes are sampled with a Markov process that is indexed by $\mathbb{T}$.

For each node $i \in G$, let $y(i)$ denote some characteristic of this node, for example whether i is HIV+ or HIV-. Sometimes we regard $\boldsymbol{y}$ as a vector in

$\mathbb{R}^N$, where N is the number of nodes in G. We want to estimate the population average $\mu_{\mathsf{true}} = \sum_{i \in G} y(i)/N$ by the RDS sample $\{y(X_\sigma) : \sigma \in \mathbb{T}\}$.

2.2. A special case: Blockmodel

Consider G as coming from a Blockmodel with k blocks [19]. That is, each node $i \in G$ is assigned to a block with $b(i) \in \{1, \ldots, k\}$, where each block j contains N/k nodes. If $b(i) = b(j)$, then $w_{i\ell} = w_{j\ell}$ for all $\ell \in \{1, \ldots N\}$. Further suppose that if $b(i) = b(j)$, then $y(i) = y(j)$. The Stochastic Blockmodel [10] is derived from this model.

The idea behind a Blockmodel with k blocks is clear: people in the same block share the same feature and the same friendship patterns. [6] studied RDS with this model. The motivating example in Section 1 also uses a Blockmodel with 2 blocks.

Let $\mathcal{W} \in \mathbb{R}^{k \times k}$ denote the weight matrix between blocks, where $\mathcal{W}_{b(i),b(j)} = w_{ij}$. Define the corresponding Markov transition matrix between blocks $\mathcal{P} \in \mathbb{R}^{k \times k}$ from $\mathcal{W}$ similarly to (2.1). Since $\mathcal{W}$ is symmetric, $\mathcal{P}$ is reversible.

Let $\{B_\sigma^{(\cdot)} : \sigma \in \mathbb{T}\}$ denote a Markov process indexed by $\mathbb{T}$, where the state space is the block labels $\{1, \ldots, k\}$ and the transition matrix is $\mathcal{P}$. The superscript of $B_\sigma^{(\cdot)}$ indicates the initial state of B_0 and is in correspondence with the initial state X_0 of the Markov process over G: if X_0 is initialized at $i \in G$, B_0 is initialized at $b(i)$ and the superscript is $b(i)$; if X_0 is initialized from any distribution $\boldsymbol{\nu} : G \to \mathbb{R}$, B_0 is initialized from the distribution $\boldsymbol{\mu} : \{1, \ldots, k\} \to \mathbb{R}$ with $\mu_j = \sum_{i \in G : b(i) = j} \nu_i$. For any $\{\sigma_{i_1}, \ldots, \sigma_{i_s}\} \subset \mathbb{T}$ and $b_{i_1}, \ldots, b_{i_s} \in \{1, \ldots, k\}$,

$$\mathbb{P}(B_{\sigma_{i_1}}^{(\boldsymbol{\mu})} = b_{i_1}, \ldots, B_{\sigma_{i_s}}^{(\boldsymbol{\mu})} = b_{i_s}) = \mathbb{P}(b(X_{\sigma_{i_1}}^{(\boldsymbol{\nu})}) = b_{i_1}, \ldots, b(X_{\sigma_{i_s}}^{(\boldsymbol{\nu})}) = b_{i_s}). \tag{2.2}$$

The proof of (2.2) is in Appendix A. Therefore $\{B_\sigma^{(\cdot)} : \sigma \in \mathbb{T}\}$ is equal in distribution to $\{b(X_\sigma^{(\cdot)}) : \sigma \in \mathbb{T}\}$ with suitable initializations. Instead of studying the Markov process $\{X_\sigma^{(\cdot)} : \sigma \in \mathbb{T}\}$ in Section 2.1, we study the Markov process $\{B_\sigma^{(\cdot)} : \sigma \in \mathbb{T}\}$. Intuitively, the original process $\{X_\sigma^{(\cdot)} : \sigma \in \mathbb{T}\}$ keeps track of the individuals while $\{B_\sigma^{(\cdot)} : \sigma \in \mathbb{T}\}$ keeps track of some feature of the individuals. This time the node feature $\boldsymbol{y} \in \mathbb{R}^N$ is replaced by the block feature $\boldsymbol{b} \in \mathbb{R}^k$ and the Markov transition matrix is replaced by the Markov transition matrix between blocks $\mathcal{P} \in \mathbb{R}^{k \times k}$.

The Blockmodel is a special case of the Markov model in Section 2.1. In this paper, Theorem 3.1, Corollary 3.1 and 3.2 apply to the Markov model. Theorem 3.2 and Corollary 3.3 only apply to the Blockmodel with 2 blocks.

2.3. Estimators

Denote $\mathbb{E}_{\boldsymbol{\pi}}(y) = \sum_i \pi(i) y(i)$. The theoretical results in this paper study two estimators defined in this section. They are unbiased estimators of $\mathbb{E}_{\boldsymbol{\pi}}(y)$. When applying inverse probability weighting (in Section 2.4), these estimators become

unbiased estimators of μ_{true} instead. Further, the VH adjustment provides a way to estimate the inverse probability weights.

Sample average The RDS sample average is

$$\hat{\mu}^{(\cdot)} = \frac{1}{n}\sum_{\sigma\in\mathbb{T}} y(X_\sigma^{(\cdot)}). \tag{2.3}$$

When X_0 is initialized from $\boldsymbol{\pi}$, $\hat{\mu}^{(\boldsymbol{\pi})}$ is an unbiased estimator of $\mathbb{E}_{\boldsymbol{\pi}}(y)$. When X_0 is initialized from $i \in G$, $\hat{\mu}^{(i)}$ is an asymptotically unbiased estimator of $\mathbb{E}_{\boldsymbol{\pi}}(y)$ (see Claim C.1).

GLS estimator [15] proposed generalize least squares (GLS) in RDS to reduce the variance, particularly in the high variance regime. The GLS estimator is the weighted average

$$\hat{\mu}_{\mathsf{GLS}}^{(\cdot)} = \sum_{\sigma\in\mathbb{T}} w_\sigma^* y(X_\sigma^{(\cdot)}) \tag{2.4}$$

where $\boldsymbol{w}^*$ minimizes the variance of the weighted average initialized from $\boldsymbol{\pi}$

$$\boldsymbol{w}^* = \arg\min_{\boldsymbol{w}} \mathsf{Var}\left(\sum_{\sigma\in\mathbb{T}} w_\sigma y(X_\sigma^{(\boldsymbol{\pi})})\right) \quad s.t. \quad \sum_{\sigma\in\mathbb{T}} w_\sigma = 1. \tag{2.5}$$

When X_0 is initialized from $\boldsymbol{\pi}$, $\hat{\mu}_{\mathsf{GLS}}^{(\boldsymbol{\pi})}$ is an unbiased estimator of $\mathbb{E}_{\boldsymbol{\pi}}(y)$. When X_0 is initialized from $i \in G$, $\hat{\mu}_{\mathsf{GLS}}^{(i)}$ is an *asymptotically* unbiased estimator of $\mathbb{E}_{\boldsymbol{\pi}}(y)$ (see Theorem 3.2).

2.4. Inverse probability weighting

In general $\mu_{\mathsf{true}} \neq \mathbb{E}_{\boldsymbol{\pi}}(y)$. So $\hat{\mu}$ and $\hat{\mu}_{\mathsf{GLS}}$ are biased estimators for μ_{true}. Inverse probability weighting can adjust for this bias. Define $y^{\boldsymbol{\pi}}(i) = y(i)/(N\pi(i))$. The IPW estimator and GLS estimator with IPW adjustment are the sample average and the GLS estimator of $y^{\boldsymbol{\pi}}(X_\sigma)$'s:

$$\hat{\mu}_{\mathsf{IPW}} = \frac{1}{n}\sum_{\sigma\in\mathbb{T}} y^{\boldsymbol{\pi}}(X_\sigma) = \frac{1}{n}\frac{\mathrm{vol}(G)}{N}\sum_{\sigma\in\mathbb{T}}\frac{y(X_\sigma)}{\deg(X_\sigma)}, \text{ and}$$

$$\hat{\mu}_{\mathsf{IPW,GLS}} = \sum_{\sigma\in\mathbb{T}} w_\sigma^{\boldsymbol{\pi}} y^{\boldsymbol{\pi}}(X_\sigma) = \frac{\mathrm{vol}(G)}{N}\sum_{\sigma\in\mathbb{T}} w_\sigma^{\boldsymbol{\pi}}\frac{y(X_\sigma)}{\deg(X_\sigma)}.$$

When X_0 is initialized from the stationary distribution $\boldsymbol{\pi}$, they are unbiased estimates of μ_{true}. However, computing these two estimators requires the average node degree $\mathrm{vol}(G)/N$, which is typically not available in practice.

The popular VH estimator replaces $\mathrm{vol}(G)/N$ in the IPW estimator with the harmonic mean of the degrees of the RDS samples [18]. Define

$$H^{-1} = \frac{1}{n}\sum_{\sigma\in\mathbb{T}}\frac{1}{\deg(X_\sigma)}, \qquad \hat{\pi}(i) = H^{-1}\deg(i), \qquad y^{\hat{\boldsymbol{\pi}}}(i) = \frac{y(i)}{\hat{\pi}(i)}.$$

The VH estimator is the sample average of $y^{\hat{\boldsymbol{\pi}}}(X_\sigma)$'s. The GLS estimator with VH adjustment uses a similar reweighting, but replaces $\mathrm{vol}(G)/N$ with a GLS estimate of $\mathbb{E}_{\boldsymbol{\pi}}(1/\deg(i))$ [15].

The VH estimator and GLS estimator with VH adjustment are two asymptotically unbiased estimators of μ_{true} under the $(\mathbb{T}, \boldsymbol{P})$-walk on G. Theorem 3.1 and 3.2 study the limit distribution of the sample average and GLS estimator. By a simple transformation (defining a new node function $y^{\boldsymbol{\pi}}(i) = y(i)/(N\pi(i))$), these results can also be applied to the IPW estimator and the GLS estimator with IPW adjustment. Corollary 3.2 and 3.3 extend these results to the VH estimator and GLS estimator with VH adjustment.

2.5. Additional notation

For two sequences a_n and b_n, define the following notation: (i) $a_n = O(b_n)$ if and only if $|a_n|$ is bounded above by b_n (up to constant factor) asymptotically, i.e. $\exists k > 0,\ \exists n_0,\ \forall n > n_0, |a_n| \leq k b_n$. (ii) $a_n = \Theta(b_n)$ if and only if a_n is bounded both above and below by b_n (up to constant factors) asymptotically, i.e. $\exists k_1 > 0,\ \exists k_2 > 0,\ \exists n_0,\ \forall n > n_0,\ k_1 b_n \leq a_n \leq k_2 b_n$.

3. Main results

This section shows that, after proper scaling, the GLS estimator and the sample average both have a limit distribution. For GLS, the limit distribution is a normal distribution. For the sample average, on the other hand, the limit distribution is a non-trivial mixture distribution, where the mixture component is determined by the seed node. This mixture distribution can be multi-modal as illustrated in Figure 1. These results can be further extended to the GLS estimator with VH adjustment and to the VH estimator respectively.

We will need the following standard lemma (e.g. [13, Lemma 12.2]) which provides the eigendecomposition of the Markov transition matrix $\boldsymbol{P}$.

Lemma 3.1. *Let $\boldsymbol{P}$ be a reversible Markov transition matrix on the nodes in G with respect to the stationary distribution $\boldsymbol{\pi}$. The eigenvectors of $\boldsymbol{P}$, denoted as $\boldsymbol{f}_1, \ldots, \boldsymbol{f}_N$, are real valued functions of the nodes $i \in G$ and orthonormal with respect to the inner product*

$$\langle \boldsymbol{f}_a, \boldsymbol{f}_b \rangle_{\boldsymbol{\pi}} = \sum_{i \in G} f_a(i) f_b(i) \pi(i). \tag{3.1}$$

If λ is an eigenvalue of $\boldsymbol{P}$, then $|\lambda| \leq 1$. The eigenfunction $\boldsymbol{f}_1$ corresponding to the eigenvalue 1 can be taken to be the constant vector $\mathbf{1}$.

Assume that the eigenvalues of $\boldsymbol{P}$ are

$$|\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_N|.$$

Since it is a Markov transition matrix, its largest eigenvalue is $\lambda_1 = 1$. Let $\boldsymbol{f}_i$ be the eigenvector corresponding to λ_i, normalized as in Lemma 3.1. The eigenvector $\boldsymbol{f}_1$ corresponding to λ_1 is taken to be the constant vector $\mathbf{1}$. Expanding the node feature $y \in \mathbb{R}^N$ in the eigenbasis yields

$$\boldsymbol{y} = \sum_{j=1}^{N} \langle \boldsymbol{y}, \boldsymbol{f}_j \rangle_{\boldsymbol{\pi}} \boldsymbol{f}_j. \tag{3.2}$$

3.1. Results for the sample average and the IPW and VH estimators

This section shows that the sample average, IPW and VH estimators have a limit distribution and that this limit distribution in fact depends on where the process is initialized (i.e. the "seed" node).

For each node $\sigma \in \mathbb{T}$, let $|\sigma|$ be the distance of σ from the root 0. Define $\{X_\sigma : \sigma \in \mathbb{T}, |\sigma| = t\}$ as the individuals in the t-th generation of the sample. Denote the sample average up to generation t as $\hat{\mu}_t$. Superscripts on $\hat{\mu}$ will denote how X_0 is initialized.

Theorem 3.1 studies the limit distribution of the sample average $\hat{\mu}_t^{(i)}$. Recall that the sample average of RDS samples is $\hat{\mu}^{(\cdot)} = n^{-1} \sum_{\sigma \in \mathbb{T}} y(X_\sigma^{(\cdot)})$.

Theorem 3.1. *Assume the eigenvalues of the transition matrix $\boldsymbol{P}$ are*

$$1 = \lambda_1 > \lambda_2 > |\lambda_3| \geq \cdots \geq |\lambda_N|. \tag{3.3}$$

Assume $\mathbb{T}$ is an m-tree. When $m > \lambda_2^{-2}$, there exist a random variable $X^{(i)} \in L^2$ such that

$$\lambda_2^{-t} \left[\hat{\mu}_t^{(i)} - \mathbb{E}_{\boldsymbol{\pi}}(y) \right] \to X^{(i)} \tag{3.4}$$

almost surely and in L^2 as $t \to \infty$, and

$$\mathbb{E} X^{(i)} = \frac{(m-1)\lambda_2}{m\lambda_2 - 1} \langle \boldsymbol{y}, \boldsymbol{f}_2 \rangle_{\boldsymbol{\pi}} f_2(i). \tag{3.5}$$

Moreover, if $\langle \boldsymbol{y}, \boldsymbol{f}_2 \rangle_{\boldsymbol{\pi}} \neq 0$, then $\mathsf{Var}(X^{(i)}) > 0$ for any $i = 1, \ldots, N$.

Note that the result is based on the technicial condition that $\mathbb{T}$ is an m-tree. The simulations in Section 4 suggest that the result still holds when $\mathbb{T}$ is a Galton-Watson tree. Condition (3.3) in Theorem 3.1 can be weakened to

$$1 = \lambda_1 > \lambda_2 = \cdots = \lambda_k > |\lambda_{k+1}| \geq \cdots \geq |\lambda_N|,$$

but the statement of the conclusion becomes more involved. See Remark 6.1 for a complete statement.

Using the above result, we can study how the bias and variance of the sample average decays, conditioned on the seed node.

Corollary 3.1. *Assume the conditions of Theorem 3.1 hold.*

1. *When* $\langle \boldsymbol{y}, \boldsymbol{f}_2 \rangle_{\boldsymbol{\pi}} \neq 0$ *and* $f_2(i) \neq 0$*, the bias of* $\hat{\mu}_t^{(i)}$ *decays like*

$$\left[\mathbb{E}(\hat{\mu}_t^{(i)}) - \mathbb{E}_{\boldsymbol{\pi}}(y)\right]^2 = \Theta(\lambda_2^{2t}). \tag{3.6}$$

2. *When* $\langle \boldsymbol{y}, \boldsymbol{f}_2 \rangle_{\boldsymbol{\pi}} \neq 0$*, the variance of* $\hat{\mu}_t^{(i)}$ *decays like*

$$\mathsf{Var}(\hat{\mu}_t^{(i)}) = \Theta(\lambda_2^{2t}). \tag{3.7}$$

When X_0 is initialized from $\boldsymbol{\pi}$, $\hat{\mu}_t^{(\boldsymbol{\pi})}$ is an unbiased estimator of μ_{true}. By (3.5), for i, j such that $f_2(i) \neq f_2(j)$, the limit distributions of $\lambda_2^{-t}\hat{\mu}_t^{(i)}$ and $\lambda_2^{-t}\hat{\mu}_t^{(j)}$ are different because $X^{(i)}$ and $X^{(j)}$ have different expectations. Thus the limit distribution of $\lambda_2^{-t}\hat{\mu}_t^{(\boldsymbol{\pi})}$ is a non-trivial mixture. The motivating example in the introduction illustrates this mixture. It is further explored with the simulation in Section 4.

Theorem 3.1 studies the limit distribution of the sample average. Using the transformation discussed in Section 2.4, the result also applies to the IPW estimator. Denote the VH estimator up to generation t as $\hat{\mu}_{\mathsf{VH},t}$. The following corollary extends the result to the VH estimator.

Corollary 3.2. *Under the conditions of Theorem 3.1, there exists a random variable* $\tilde{X}^{(i)} \in L^2$ *such that*

$$\lambda_2^{-t}\left[\hat{\mu}_{\mathsf{VH},t}^{(i)} - \mu_{\mathsf{true}}\right] \to \tilde{X}^{(i)}$$

almost surely, and

$$\mathbb{E}\tilde{X}^{(i)} = \mathbb{E}_{\boldsymbol{\pi}}(y')^{-1}\frac{(m-1)\lambda_2}{m\lambda_2 - 1}\,\langle \boldsymbol{y}'', \boldsymbol{f}_2 \rangle_{\boldsymbol{\pi}}\, f_2(i),$$

where $y'(j) = \deg(j)^{-1}$ *and* $y''(j) = y(j)/\deg(j)$*. Moreover, if* $\langle \boldsymbol{y}'', \boldsymbol{f}_2 \rangle_{\boldsymbol{\pi}} \neq 0$*, then* $\mathsf{Var}(\tilde{X}^{(i)}) > 0$ *for any* $i = 1, \ldots, N$*.*

Similarly, when X_0 is initialized from $\boldsymbol{\pi}$, the limit distirbution of $\hat{\mu}_{\mathsf{VH},t}^{(\boldsymbol{\pi})}$ is a non-trivial mixture of the limit distributions of $\hat{\mu}_{\mathsf{VH},t}^{(i)}$ for all $i \in G$.

3.2. Results for the GLS estimator

For the GLS estimator, the two right panels of Figure 1 suggest that the estimator is not sensitive to the initial distribuiton of X_0. This section shows that the GLS estimator is asymptotically normal with parameters that do not depend on the initial distribution of X_0.

Given the referral tree $\mathbb{T}$, define the covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{n \times n}$ as

$$\boldsymbol{\Sigma}_{\sigma,\tau} = \mathsf{Cov}(y(X_\sigma), y(X_\tau))$$

for any $\sigma, \tau \in \mathbb{T}$, where n is the number of nodes in $\mathbb{T}$. According to [15], $\boldsymbol{w}^*$ in (2.5) is given by

$$\boldsymbol{w}^* = (\boldsymbol{x}^\top \mathbf{1})^{-1} \boldsymbol{x}^\top, \quad \text{where} \quad \boldsymbol{\Sigma} \boldsymbol{x} = \mathbf{1}. \tag{3.8}$$

Here $\boldsymbol{x}$ is the vectorization of the RDS sample $\{X_\sigma^{(\cdot)} : \sigma \in \mathbb{T}\}$. For the Blockmodel with 2 blocks, the GLS estimator admits a closed-form expression:

$$\hat{\mu}_{\mathsf{GLS}} = \sum_{\sigma \in \mathbb{T}} \frac{1 - \lambda_2(\deg(\sigma) - 1)}{n(1 - \lambda_2(1 - \frac{2}{n}))} y(X_\sigma), \tag{3.9}$$

where λ_2 is the second eigenvalue of the Markov transition matrix between blocks and $\deg(\sigma)$ is the degree of $\sigma \in \mathbb{T}$.

Let $\hat{\mu}_{\mathsf{GLS},t}$ be the GLS estimator of the RDS samples up to generation t. Based on (3.9), the following theorem establishes the asymptotic normality of the GLS estimator.

Theorem 3.2. *Consider the Blockmodel with 2 blocks on an m-tree $\mathbb{T}$. Assume $|\lambda_2| < 1$. Then, for any initial distribution $\boldsymbol{\nu}$ of X_0,*

$$\sqrt{n_t} \left[\hat{\mu}_{\mathsf{GLS},t}^{(\boldsymbol{\nu})} - \mathbb{E}_{\boldsymbol{\pi}}(y) \right] \to \mathcal{N}\left(0, \frac{1 + \lambda_2}{1 - \lambda_2} \mathsf{Var}_{\boldsymbol{\pi}}(y)\right). \tag{3.10}$$

in distribution as $t \to \infty$, where $\mathsf{Var}_{\boldsymbol{\pi}}(y) = \mathbb{E}_{\boldsymbol{\pi}}(y^2) - (\mathbb{E}_{\boldsymbol{\pi}}(y))^2$ and $n_t = 1 + m + \cdots + m^t$ is the number of RDS samples up to generation t.

Theorem 3.2 shows that the GLS estimator is asymptotically normal *both in the low variance and high variance regimes*. Note that the result is based on (3.9) and the technical condition that $\mathbb{T}$ is an m-tree. The simulations in Section 4 suggest that the asymptotic normality of the GLS estimator still holds when $\mathbb{T}$ is a Galton-Watson tree, or the model is no longer a Blockmodel with 2 blocks.

Theorem 3.2 studies the limit distribution of the GLS estimator. Using the transformation discussed in Section 2.4, the result also applies to the GLS estimator with IPW adjustment. Denote the GLS estimator with VH adjustment of RDS samples up to generation t as $\hat{\mu}_{\mathsf{GLS},\mathsf{VH},t}^{(\cdot)}$. The following corollary extends the result to the GLS estimator with VH adjustment.

Corollary 3.3. *Under the conditions in Theorem 3.2, for any initial distribution $\boldsymbol{\nu}$ of X_0,*

$$\sqrt{n_t} \left[\hat{\mu}_{\mathsf{GLS},\mathsf{VH},t}^{(\boldsymbol{\nu})} - \mu_{\mathsf{true}} \right] \xrightarrow{d} \mathcal{N}\left(0, \frac{1 + \lambda_2}{1 - \lambda_2} \mathbb{E}_{\boldsymbol{\pi}}(y')^{-2} \mathsf{Var}_{\boldsymbol{\pi}}(y'')\right).$$

where $y'(i) = \deg(i)^{-1}$ and $y''(i) = y(i)/\deg(i)$.

4. Simulation studies

In this section, data are simulated from a Blockmodel with 2 or 3 blocks. As stated in Section 2.2, a Blockmodel with k blocks consists of a reversible transition matrix $\mathcal{P} \in \mathbb{R}^{k \times k}$ between blocks, block feature $\boldsymbol{y} \in \mathbb{R}^k$, and a referral

tree $\mathbb{T}$. In this specification, the block feature $\boldsymbol{y}$ is assumed to be centralized, so that $\mathbb{E}_{\boldsymbol{\pi}}(y) = 0$. For a Blockmodel with 2 blocks, let

$$\mathcal{P} = \begin{pmatrix} p & 1-p \\ 1-q & q \end{pmatrix}.$$

denote the transition matrix between 2 blocks. The second eigenvalue of $\mathcal{P}$ is $\lambda_2 = p + q - 1$.

In the simulation settings below, the block feature is given prior to centralization. In fact, all of the 2-Blockmodels use $\boldsymbol{y} = (1, 0)^\top$ and the 3-Blockmodels use $\boldsymbol{y} = (0, 1, 2)^\top$. All of the experiments are based on 5000 simulated datasets.

4.1. Sample average

Here we consider the behavior of the sample average $\hat{\mu}_t$ in the high variance regime $m > \lambda_2^{-2}$. In this setting, the asymptotic distribution of $\lambda_2^{-t}\hat{\mu}_t^{(\boldsymbol{\pi})}$ is no longer normal, unlike the low variance regime. Instead, its asymptotic distribution is a mixture of the distributions of $\lambda_2^{-t}\hat{\mu}_t^{(i)}$ for all $i \in G$.

The simulation is performed on two different Blockmodels with 2 blocks. We consider a balanced model with $p = q = .95$ and an unbalanced model with $p = 0.95$ and $q = 0.85$. For both models, $\mathbb{T}$ is a Galton-Watson tree with offspring distribution $1 + \text{Binomial}(2, 1/2)$. Under these settings, $m > \lambda_2^{-2}$ for both models. Figure 2 displays the results of the experiment with $t = 50$.

4.2. GLS estimator

Here we consider the behavior of the GLS estimator in both the low and high variance regimes. The first experiment corroborates the result of Theorem 3.2, namely that the GLS estimator is asymptotically normal in both variance regimes. The simulation is performed on two different Blockmodels with 2 blocks. In the first model $(p, q) = (0.95, 0.85)$; in the second model $(p, q) = (0.8, 0.7)$. For both models, $\mathbb{T}$ is a 2-tree. Under these settings, $m > \lambda_2^{-2}$ for the first model and $m < \lambda_2^{-2}$ for the second model. The two quantile-quantile plots in Figure 3 correspond to the two models. It appears that the distribution of the GLS estimator gets closer to the normal distribution as the sample size increases.

The second experiment suggests that the asymptotic normality of GLS estimator extends beyond the conditions in Theorem 3.2. We consider a two-block model with $(p, q) = (0.8, 0.7)$ and a three-block model, where the transition matrix between the blocks is

$$\mathcal{P} = \begin{pmatrix} 0.8 & 0.1 & 0.1 \\ 0.2 & 0.6 & 0.2 \\ 0.2 & 0.2 & 0.6 \end{pmatrix}.$$

For both models, $\mathbb{T}$ is a Galton-Watson tree with offspring distribution $1 + \text{Binomial}(2, 1/2)$. Results for this experiment are displayed in Figure 4.

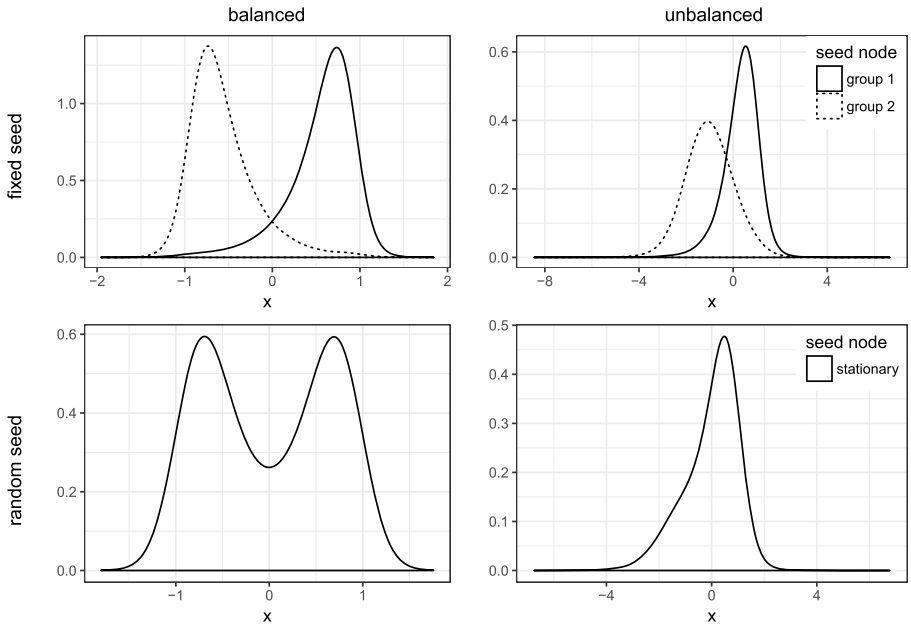

FIG 2. *Kernel density estimates of* $\lambda_2^{-t}\hat{\mu}_t$ *for balanced (the left panels) and unbalanced (the right panels) Blockmodel with* 2 *blocks over* 5000 *replicates. For each scenario, the top panel corresponds to the case when* X_0 *is initialized from group 1 (the solid curve) and group 2 (the dashed curve), the lower panel corresponds to the case when* X_0 *is initialized from the stationary distribution.*

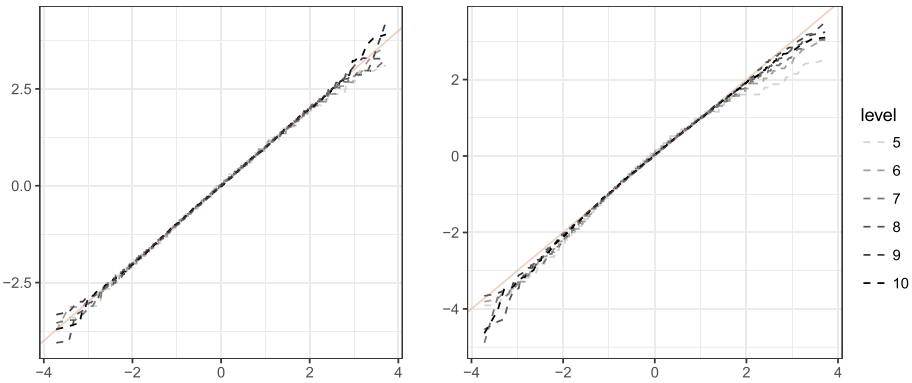

FIG 3. *Q-Q plot of* $\mu_{t,GLS}$ *for the Blockmodels with* 2 *blocks, with* $m > \lambda_2^{-2}$ *(left panel) and* $m < \lambda_2^{-2}$ *(right panel).* $\mathbb{T}$ *is a* 2*-tree. For each scenario, the Q-Q plot is created over* 5000 *replicates. The six dashed Q-Q lines with different colors correspond to* $\mathbb{T}$ *with 5, 6, 7, 8, 9 or 10 levels. The red solid line is* $y = x$*..*

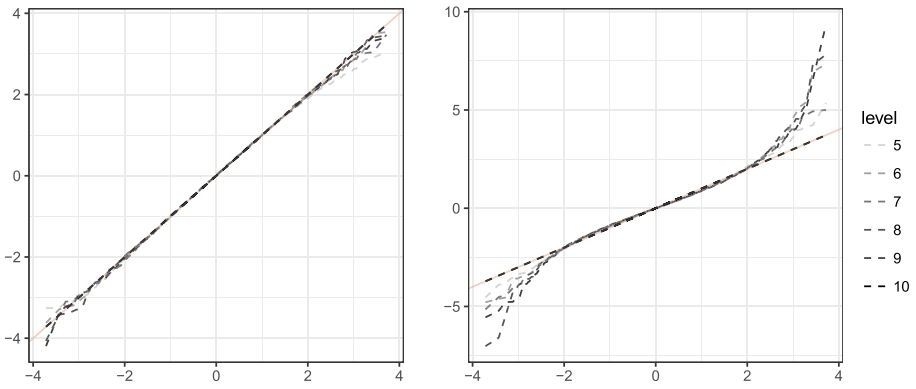

FIG 4. *Q-Q plot of $\hat{\mu}_{t,GLS}$ for the Blockmodels with 2 blocks (left panel) and 3 blocks (right panel), where $\mathbb{T}$ is a Galton-Watson tree. For each scenario, the Q-Q plot is created over 5000 replicates. The six dashed Q-Q lines with different colors correspond to $\mathbb{T}$ with 5, 6, 7, 8, 9 or 10 levels. The red solid line is $y = x$.*

5. Analysis of Adolescent Health Data

In this section, we consider numerical experiments where the RDS samples are simulated without replacement from empirically derived social networks. Specifically, we use social networks collected in the National Longitudinal Study of Adolescent Health (Add Health). In the 1994–95 school year, the Add Health study collected a nationally representative sample of adolescents in grades seven through twelve. The sample covers 84 pairs of middle and high schools in which students nominated up to five male and five female friends in their middle or high school network [7].

In this analysis, we consider 25 networks with at least 1000 nodes. All contacts are symmetrized and all graphs are restricted to the largest connected component. The RDS sampling process is initialized from a seed node which is selected with probability proportional to node degree (i.e. the stationary distribution). Then, each participant recruits $\xi \sim 1 + \text{Binomial}(2, 1/2)$ participants uniformly at random from their contacts whom have not yet been recruited. If the participant has fewer than ξ contacts eligible to recruit, then the participant recruits all of their eligible contacts. The RDS process stops when there are 500 participants. If the process terminates before collecting 500 participants, then the process is restarted. For each network, we collect 500 different RDS samples. We generate 2000 such simulated data sets.

We use school-status as the binary node feature and focus on estimating the proportion of the population in high school. We construct a sample average, a GLS estimator and a SBM-fGLS estimator for the proportion of students in high school. The GLS estimator requires an estimate of the covariance matrix Σ, which can be calculated from the Markov transition matrix of the network (typically not available in practice) and equation (6) in [16]. The SBM-fGLS estimator proposed in [15] estimates Σ using the RDS samples.

TABLE 2

Network characteristics for the 25 networks in the Add Health study used in the numerical experiments in Section 5. ID gives the network ID (school ID) from the study listed in increasing order by $\tilde{\lambda}$, a measure of the strength of bottleneck in the network, see (5.1).

ID	$\tilde{\lambda}$	ID	$\tilde{\lambda}$	ID	$\tilde{\lambda}$	ID	$\tilde{\lambda}$	ID	$\tilde{\lambda}$
17	0.739	39	0.842	45	0.869	54	0.879	60	0.911
75	0.744	40	0.844	48	0.869	59	0.881	58	0.917
42	0.771	41	0.847	36	0.874	73	0.886	84	0.923
15	0.818	50	0.867	43	0.874	44	0.889	57	0.925
28	0.839	34	0.868	61	0.878	68	0.897	49	0.944

Consider a measure of the network bottleneck. Let $\boldsymbol{A} \in \mathbb{R}^{N\times N}$ denote the adjacency matrix of the network. Define the diagonal matrix $\boldsymbol{D} \in \mathbb{R}^{N\times N}$ and the matrix $\boldsymbol{L} \in \mathbb{R}^{N\times N}$ so that

$$D_{ii} = \sum_{k=1}^{N} W_{ik}, \quad \boldsymbol{L} = \boldsymbol{D}^{-1/2}\boldsymbol{A}\boldsymbol{D}^{-1/2}.$$

Then $\tilde{\lambda}$ is defined as

$$\tilde{\lambda} = \tilde{\boldsymbol{y}}^{\top}\boldsymbol{L}\tilde{\boldsymbol{y}}, \tag{5.1}$$

where $\tilde{\boldsymbol{y}}$ is the standardized form of the node feature $\boldsymbol{y}$, so that $\sum_{i=1}^{N}\tilde{y}_i = 0$ and $\|\tilde{\boldsymbol{y}}\|_2 = 1$. $\tilde{\lambda}$ provides a measure of the network bottleneck; as long as the second eigenvalue λ_2 is not too close to 1, then this quantity will not be close to 1. Table 2 displays the $\tilde{\lambda}$ of the 25 networks.

In Figure 5, the 25 subplots show the kernel density estimation of VH estimator corresponding to the 25 networks. In Figure 6 and 7, the 25 subplots show the kernel density estimation and quantile-quantile plots of GLS and SBM-fGLS estimator with VH adjustment corresponding to the 25 networks. We plot these results over 2000 replicates. The 25 subplots are in order of descending $\tilde{\lambda}$. It is clear that the VH estimator has two modes, so these networks are all beyond the critical threshold. Except for networks with extremely strong bottleneck (i.e. with large $\tilde{\lambda}$), the GLS estimators with VH adjustment are approximately normally distributed. The distribution of SBM-fGLS estimator with VH adjustment are not enough close to the normal distribution for some networks, which means that our results for the GLS estimator might not always hold for the SBM-fGLS estimator. It is possible for the GLS estimator to exceed one. In practice, one would provide a modified estimate capped at one.

6. Proof outlines for the main results

This section outlines the proofs for Theorems 3.1 and 3.2. Well-established theory for multi-type branching processes and martingale limit theorems play an important role. For each proof, the main idea is to extract the underlying martingale structure for the estimator; it is this structure that determines the asymptotic behavior. The proofs of Corollary 3.1, 3.2 and 3.3 are relegated to Appendices B and C.

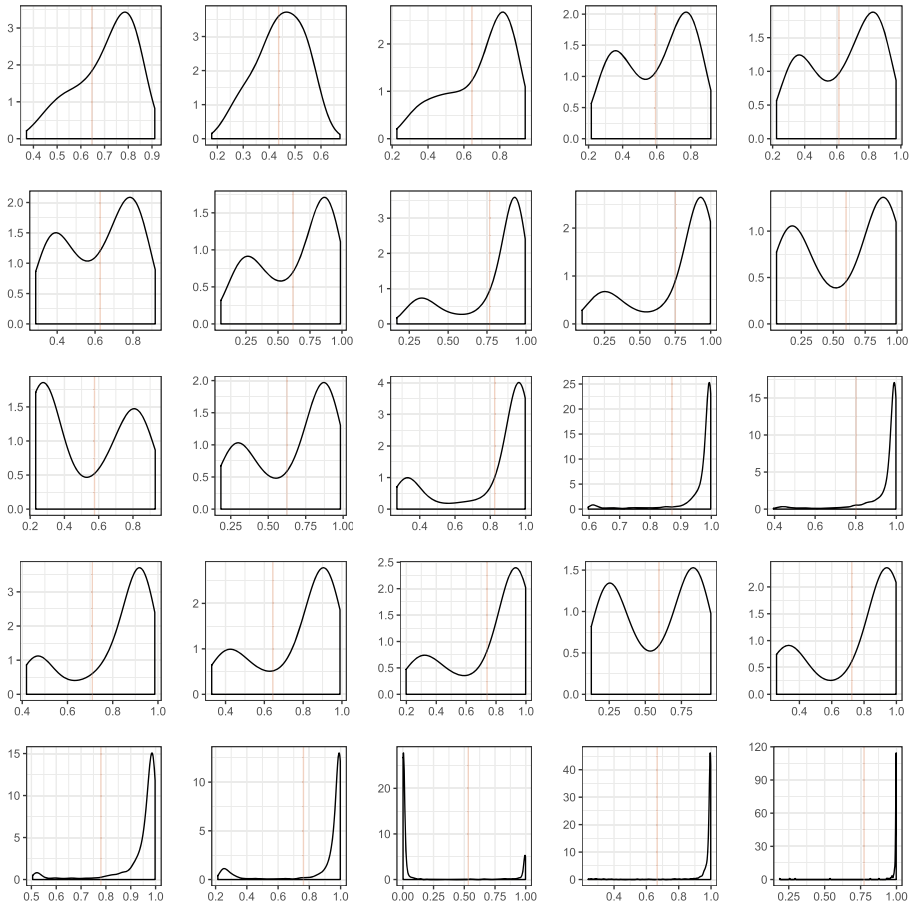

FIG 5. *Simulation results based on the Add Health Study described in Section 5. The figures display kernel density estimates of the sample average. The 25 subplots correspond to the Comm 17, 75, 42, 15, 28, 39, 40, 41, 50, 34, 45, 48, 36, 43, 61, 54, 59, 73, 44, 68, 60, 58, 84, 57, 49 networks. The red solid line is* $x = \mu_{true}$*. This figure suggests that VH estimator has multiple modes.*

6.1. Analysis of the sample average

Denote $\boldsymbol{Z}_{t,j}$ as the number of $j \in G$ in the t-th generation and define $\boldsymbol{Z}_t = (Z_{t,1}, \ldots, Z_{t,N})$. When $\mathbb{T}$ is an m-tree and X_0 is initialized from $i \in G$,

$$\boldsymbol{Z}_t^{(i)} = (Z_{t,1}^{(i)}, \ldots, Z_{t,N}^{(i)})$$

is a multitype Galton-Watson process [8, 1]. The next lemma can be derived from a standard result in the literature of multitype Galton-Watson processes.

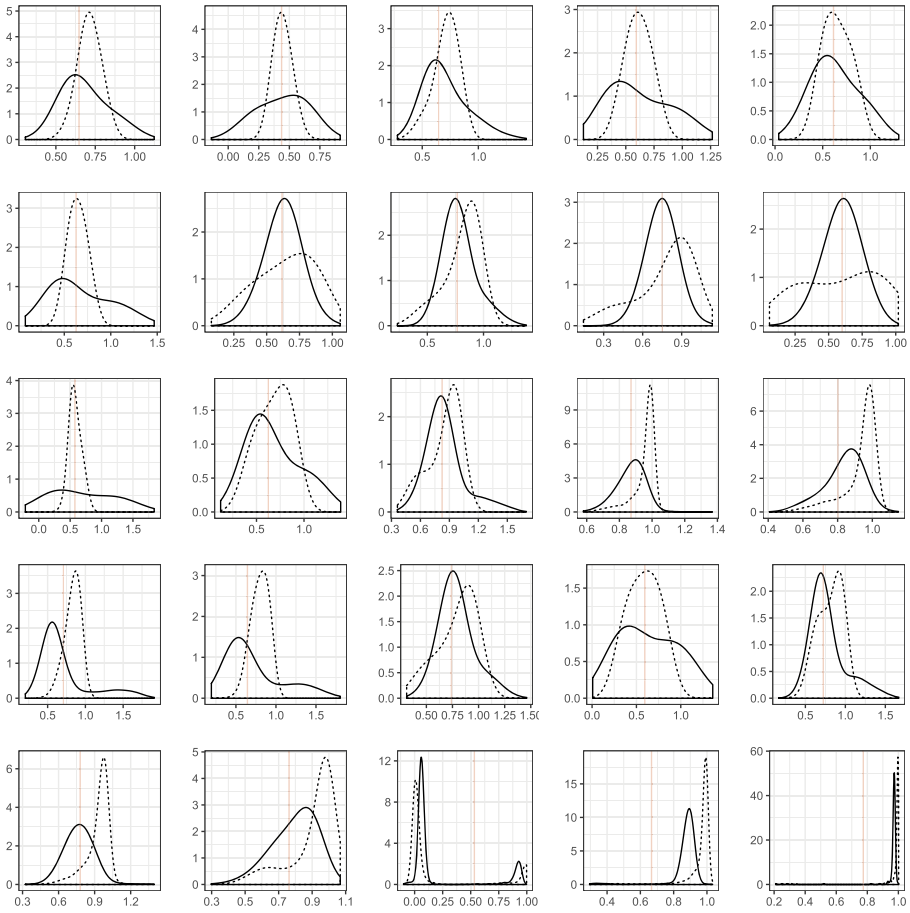

FIG 6. *Simulation results based on the Add Health Study described in Section 5. The figures display kernel density estimates of the GLS estimator (solid line) and the SBM-fGLS estimator (dashed line). The 25 subplots correspond to the Comm 17, 75, 42, 15, 28, 39, 40, 41, 50, 34, 45, 48, 36, 43, 61, 54, 59, 73, 44, 68, 60, 58, 84, 57, 49 networks. The red solid line is $x = \mu_{true}$. This figure shows that when the bottleneck of the network is not too strong, both estimators have only one mode.*

Lemma 6.1. *Assume the conditions of Theorem 3.1. For any $j = 1, \ldots, N$,*

$$Y_{t,j}^{(i)} = (m\lambda_j)^{-t} \langle \boldsymbol{Z}_t^{(i)}, \boldsymbol{f}_j \rangle$$

is a real-valued martingale adapted to $\mathcal{F}_t = \sigma\{\boldsymbol{Z}_l^{(i)} : 1 \le l \le t\}$.

Proof. See Appendix B.1. □

Let W_t denote the summation of the t-th generation RDS samples,

$$W_t = \sum_{\sigma \in \mathbb{T}:|\sigma|=t} y(X_\sigma),$$

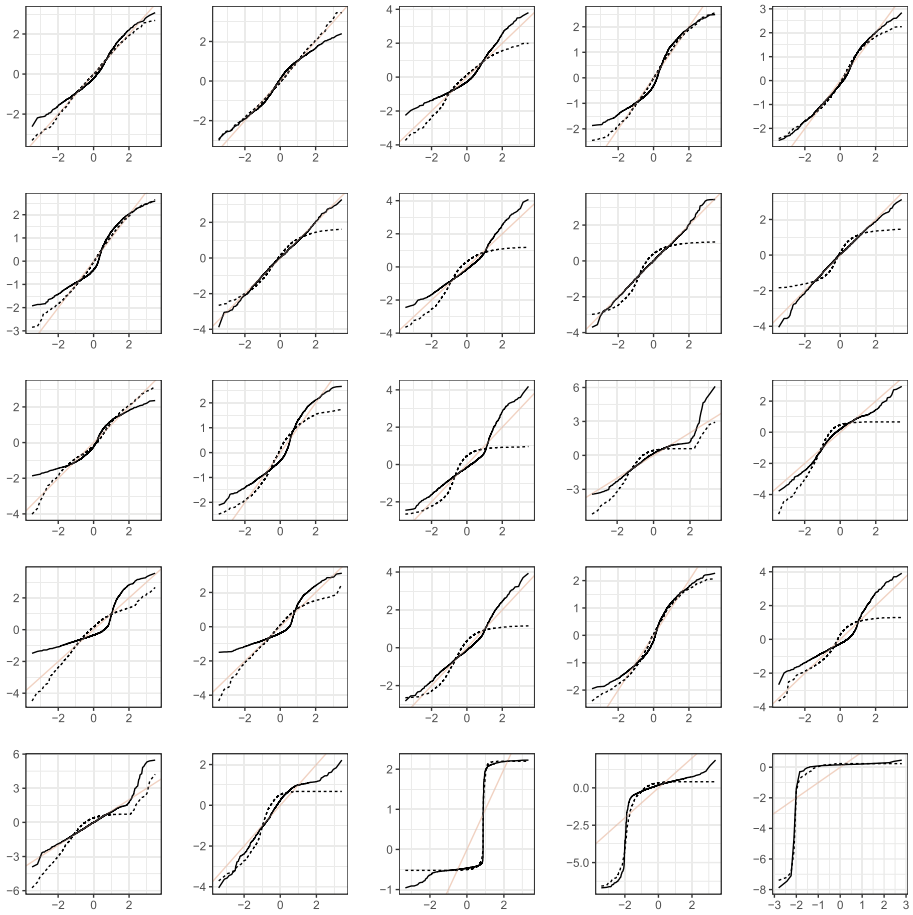

FIG 7. *Simulation results based on the Add Health Study described in Section 5. The figures display Q-Q plots of the GLS estimator (solid line) and the SBM-fGLS estimator (dashed line). The 25 subplots correspond to the Comm 17, 75, 42, 15, 28, 39, 40, 41, 50, 34, 45, 48, 36, 43, 61, 54, 59, 73, 44, 68, 60, 58, 84, 57, 49 networks. This figure illustrates that when the bottleneck of the network is not too strong, both estimators appear approximately normal (even under without replacement sampling).*

and let $S_t = \sum_{j=0}^{t} W_j$ denote the summation up to generation t. Recall that n_t is the number of nodes in $\mathbb{T}$ between the root 0 and generation t (inclusive), i.e. $n_t = |\{\sigma \in \mathbb{T}, |\sigma| \leq t\}|$. Thus the sample average up to generation t is $\hat{\mu}_t = S_t/n_t$. Superscripts on $\boldsymbol{Z}, S$ and W will denote how X_0 is initialized if necessary.

Recall from (3.2) and $\boldsymbol{f}_1 = \mathbf{1}$ that

$$\boldsymbol{y} = \sum_{j=1}^{N} \langle \boldsymbol{y}, \boldsymbol{f}_j \rangle_{\boldsymbol{\pi}} \boldsymbol{f}_j = \mathbb{E}_{\boldsymbol{\pi}}(y)\mathbf{1} + \sum_{j=2}^{N} \langle \boldsymbol{y}, \boldsymbol{f}_j \rangle_{\boldsymbol{\pi}} \boldsymbol{f}_j.$$

As a result, one obtains the following decomposition of $W_t^{(i)}$

$$W_t^{(i)} = \sum_{\sigma\in\mathbb{T},|\sigma|=t} y(X_\sigma^{(i)}) = \boldsymbol{y}^\top \boldsymbol{Z}_t^{(i)} = m^t \mathbb{E}_{\boldsymbol{\pi}}(y) + \sum_{j=2}^{N} \langle \boldsymbol{y}, \boldsymbol{f}_j\rangle_{\boldsymbol{\pi}} (m\lambda_j)^t Y_{t,j}^{(i)}. \quad (6.1)$$

The last step utilizes the simple fact that $\mathbf{1}^\top \boldsymbol{Z}_t^{(i)} = m^t$. This motivates us to study the limit distribution of $Y_{t,j}^{(i)}$.

Lemma 6.2. *Assume the conditions of Theorem 3.1. Then there exists a random variable $Y_2^{(i)}$ such that*

$$Y_{t,2}^{(i)} \to Y_2^{(i)}$$

almost surely and in L^2. For $j \geq 3$,

$$(\lambda_2^{-1}\lambda_j)^t\, Y_{t,j}^{(i)} \to 0$$

almost surely and in L^2.

Proof. See Appendix B.2. □

Lemma 6.2 informally reveals that, under proper scaling, the asymptotic distributional characterization of $W_t^{(i)}$ is determined by $Y_{t,2}^{(i)}$. The next lemma derives the first and second moments of $Y_2^{(i)}$.

Lemma 6.3. *Assume the conditions of Theorem 3.1. Then $\mathbb{E}Y_2^{(i)} = f_2(i)$, and $\mathsf{Var}(Y_2^{(i)}) > 0$ for any $i = 1, \ldots, N$ if we further assume that $\langle \boldsymbol{y}, \boldsymbol{f}_2\rangle_{\boldsymbol{\pi}} \neq 0$ holds.*

Proof. See Appendix B.3. □

6.2. Proof of Theorem 3.1

Apply (6.1) and Lemma 6.2 collectively to obtain

$$\frac{W_t^{(i)} - m^t \mathbb{E}_{\boldsymbol{\pi}}(y)}{(m\lambda_2)^t} = \sum_{j=2}^{N} \langle \boldsymbol{y}, \boldsymbol{f}_j\rangle_{\boldsymbol{\pi}} (\lambda_2^{-1}\lambda_j)^t Y_{t,j}^{(i)} \to \langle \boldsymbol{y}, \boldsymbol{f}_2\rangle_{\boldsymbol{\pi}} Y_2^{(i)} \quad (6.2)$$

almost surely and in L^2. Recalling that $S_t = \sum_{j=0}^{t} W_j$, $\hat{\mu}_t = S_t/n_t$, and the number of samples between 0 and generation t is $n_t = \sum_{l=0}^{t} m^l$, one arrives at

$$\lambda_2^{-t}\left[\hat{\mu}_t^{(i)} - \mathbb{E}_{\boldsymbol{\pi}}(y)\right] = \frac{S_t^{(i)} - n_t \mathbb{E}_{\boldsymbol{\pi}}(y)}{n_t \lambda_2^t} = \frac{m^t}{n_t} \sum_{l=0}^{t} (m\lambda_2)^{l-t} \frac{W_l^{(i)} - m^l \mathbb{E}_{\boldsymbol{\pi}}(y)}{(m\lambda_2)^l}. \quad (6.3)$$

Since $\lim_{t\to\infty} m^t/n_t = (m-1)/m$, from (6.2)

$$\begin{aligned} \lambda_2^{-t}\left[\hat{\mu}_t^{(i)} - \mathbb{E}_{\boldsymbol{\pi}}(y)\right] &\to \frac{m-1}{m} \sum_{r=0}^{\infty} (m\lambda_2)^{-r} \langle \boldsymbol{y}, \boldsymbol{f}_2\rangle_{\boldsymbol{\pi}} Y_2^{(i)} \\ &= \frac{(m-1)\lambda_2}{m\lambda_2 - 1} \langle \boldsymbol{y}, \boldsymbol{f}_2\rangle_{\boldsymbol{\pi}} Y_2^{(i)} \triangleq X^{(i)} \end{aligned} \quad (6.4)$$

almost surely as $t \to \infty$.

To prove L^2 convergence, recall that if a sequence of random variables $X_n \to X$ in probability, and $\|X_n\|_{L^2} \to \|X\|_{L^2}$, then $X_n \to X$ in L^2. Observe, similarly to the above limits, that

$$\begin{aligned}
\left\|\lambda_2^{-t}(\hat{\mu}_t^{(i)} - \mathbb{E}_{\boldsymbol{\pi}}(y))\right\|_{L^2}^2 &= \frac{m^{2t}}{n_t^2}\mathbb{E}\left[\left(\sum_{l=0}^{t}\frac{W_l^{(i)} - m^l\mathbb{E}_{\boldsymbol{\pi}}(y)}{(m\lambda_2)^t}\right)^2\right] \\
&= \frac{m^{2t}}{n_t^2}\sum_{k=1}^{t}\sum_{l=1}^{t}\mathbb{E}\left\{(m\lambda_2)^{-2t}\left[W_k^{(i)} - m^k\mathbb{E}_{\boldsymbol{\pi}}(y)\right]\left[W_l^{(i)} - m^l\mathbb{E}_{\boldsymbol{\pi}}(y)\right]\right\} \\
&\xrightarrow{t\to\infty} \lim_{r\to\infty}\left(\frac{m-1}{m}\right)^2\mathbb{E}\left[\left(\langle \boldsymbol{y}, \boldsymbol{f}_2\rangle_{\boldsymbol{\pi}} Y_2^{(i)}\right)^2\right]\sum_{k=1}^{r}\sum_{l=1}^{r}(m\lambda_2)^{k+l-2r} \\
&= \mathbb{E}\left[\left(\frac{(m-1)\lambda_2}{m\lambda_2 - 1}\langle \boldsymbol{y}, \boldsymbol{f}_2\rangle_{\boldsymbol{\pi}} Y_2^{(i)}\right)^2\right].
\end{aligned}$$

So the convergence in (6.4) is also in L_2.

Finally, in view of Lemma 6.3 and the definition of $X^{(i)}$ in (6.4), it is straightforward to check that (3.5) holds. Moreover, if $\langle \boldsymbol{y}, \boldsymbol{f}_2\rangle_{\boldsymbol{\pi}} \neq 0$, then $\mathsf{Var}(X^{(i)}) > 0$ for any $i = 1, \ldots, N$.

Remark 6.1. If condition (3.3) in Theorem 3.1 is weakened to

$$1 = \lambda_1 > \lambda_2 = \cdots = \lambda_k > |\lambda_{k+1}| \geq \cdots \geq |\lambda_N|,$$

then (3.4) becomes

$$\lambda_2^{-t}(\hat{\mu}_t^{(i)} - \mathbb{E}_{\boldsymbol{\pi}}(y)) \to \frac{(m-1)\lambda_2}{m\lambda_2 - 1}\sum_{j=2}^{k}\langle \boldsymbol{y}, \boldsymbol{f}_j\rangle_{\boldsymbol{\pi}} Y_j^{(i)}.$$

6.3. Analysis of the GLS estimator

In previous sections, the subscript of the estimators is t or l, which denotes the generation. This section requires us to study each node in a generation. Accordingly we order the nodes of the m-tree $\mathbb{T}$ by scanning each level from the root down. For example, for a 2-tree, the root node is 1, its offsprings are 2 and 3, the offsprings of 2 are 4 and 5, the offsprings of 3 are 6 and 7, etc. In a change of notation from the previous sections, when the subscript is n, $\hat{\mu}_n^{(\cdot)}$ now denotes the sample mean up to node n, i.e. $\hat{\mu}_n = n^{-1}\sum_{k=1}^{n} y(X_k)$.

Assume the Markov transition matrix between blocks is

$$\mathcal{P} = \begin{pmatrix} p & 1-p \\ 1-q & q \end{pmatrix}.$$

The second eigenvalue is $\lambda_2 = p + q - 1$ and the stationary distribution is $\boldsymbol{\pi} = (\frac{1-q}{1-\lambda_2}, \frac{1-p}{1-\lambda_2})$. For $k \geq 1$, define

$$M_n = \sum_{k=1}^{n} \left[y(X_k^{(\boldsymbol{\nu})}) - \lambda_2 y(X_{p(k)}^{(\boldsymbol{\nu})}) \right] - n(1-\lambda_2)\mathbb{E}_{\boldsymbol{\pi}}(y), \tag{6.5}$$

where $p(k)$ is the parent node of k in the ordering defined above. In view of (3.9), the relation between M_n and $\hat{\mu}_{\mathsf{GLS,t}}$ is

$$\sqrt{n_t}\left[\hat{\mu}_{\mathsf{GLS},t}^{(\boldsymbol{\nu})} - \mathbb{E}_{\boldsymbol{\pi}}(y)\right] = \frac{M_{n_t}}{\sqrt{n_t}(1-\lambda_2)} + O_{\mathbb{P}}\left(\frac{1}{\sqrt{n_t}}\right). \tag{6.6}$$

Thus it suffices to study the asymptotic behavior of M_n. One can check that M_n is a martingale adapted to the filtration $\mathcal{F}_n = \sigma(X_k^{(\nu)} : k \leq n)$:

$$\mathbb{E}\left(M_n - M_{n-1} \mid \mathcal{F}_{n-1}\right) = \begin{cases} py_1 + (1-p)y_2 - \lambda_2 y_1 - (1-\lambda_2)\mathbb{E}_{\boldsymbol{\pi}}(y) & \text{if } X_{p(n)}^{(\boldsymbol{\nu})} = 1 \\ (1-q)y_1 + qy_2 - \lambda_2 y_2 - (1-\lambda_2)\mathbb{E}_{\boldsymbol{\pi}}(y) & \text{if } X_{p(n)}^{(\boldsymbol{\nu})} = 2 \end{cases},$$

both of which are 0, as can be seen from $(1-\lambda_2)\mathbb{E}_{\boldsymbol{\pi}}(y) = (1-q)y_1 + (1-p)y_2$ and the expression for λ_2 above. It is necessary to introduce a martingale central limit theorem (see e.g. [5, Fifth Edition, Theorem 8.2.8]).

Theorem 6.1 (Martingale CLT)**.** *Let a martingale M_n satisfy $\mathbb{E}(M_n) = 0$, and*

1. *$\frac{1}{n}\sum_{k=1}^{n} \mathbb{E}((M_k - M_{k-1})^2 \mid M_1, \ldots, M_{k-1}) \to \sigma^2 > 0$ in probability as $n \to \infty$, and*
2. *for every $\epsilon > 0$, $\frac{1}{n}\sum_{k=1}^{n} \mathbb{E}((M_k - M_{k-1})^2; |M_k - M_{k-1}| > \epsilon\sqrt{n}) \to 0$ as $n \to \infty$,*

then $M_n/\sqrt{n} \to \mathcal{N}(0, \sigma^2)$ in distribution as $n \to \infty$.

It then boils down to showing that M_n defined in (6.5) satisfies the conditions in the above theorem. The detailed proof is provided in the next section. We will need a technical lemma which states that, although the limit distribution of the sample average $\hat{\mu}_n^{(i)}$ differs in the high and low variance regimes under appropriate scalings, $\hat{\mu}_n^{(i)}$ itself always converges to $\mathbb{E}_{\boldsymbol{\pi}}(y)$ in L^2.

Lemma 6.4. *Assume the conditions of Theorem 3.2. Then for any initial distribution $\boldsymbol{\nu}$ of X_0, $\hat{\mu}_n^{(\boldsymbol{\nu})} \to \mathbb{E}_{\boldsymbol{\pi}}(y)$ in L^2.*

Proof. See Appendix C.1 □

6.4. Proof of Theorem 3.2

Without loss of generality, assume $\mathbb{E}_{\boldsymbol{\pi}}(y) = 0$ and $\mathsf{Var}_{\boldsymbol{\pi}}(y) = 1$, which can be equivalently viewed as applying the same linear transformation to each entry of $\boldsymbol{y}$ as well as $\hat{\mu}_{\mathsf{GLS},t}$.

We begin by showing that M_n defined in (6.5) satisfies the first condition in Theorem 6.1. By invoking the martingale property $\mathbb{E}(M_k - M_{k-1} \mid \mathcal{F}_{k-1}) = 0$, one obtains

$$\mathbb{E}\left[(M_k - M_{k-1})^2 \mid \mathcal{F}_{k-1}\right] = \mathsf{Var}\left(M_k - M_{k-1} \mid \mathcal{F}_{k-1}\right) = \mathsf{Var}\big(y(X_k^{(\boldsymbol{\nu})}) \mid X_{p(k)}^{(\boldsymbol{\nu})}\big).$$

Notice that $\mathbb{E}_{\boldsymbol{\pi}}(y) = 0$ and $\mathsf{Var}_{\boldsymbol{\pi}}(y) = 1$ imply $(1-q)y_1^2 + (1-p)y_2^2 = 2 - p - q$, which yields

$$\begin{aligned}\mathsf{Var}\big(y(X_k^{(\boldsymbol{\nu})}) \mid X_{p(k)}^{(\boldsymbol{\nu})}\big) &= \begin{cases} py_1^2 + (1-p)y_2^2 - \lambda_2^2 y_1^2 & \text{if } X_{p(n)}^{(\boldsymbol{\nu})} = 1 \\ (1-q)y_1^2 + qy_2^2 - \lambda_2^2 y_2^2 & \text{if } X_{p(n)}^{(\boldsymbol{\nu})} = 2 \end{cases} \\ &= (1-\lambda_2)(1 + \lambda_2 y(X_{p(k)}^{(\boldsymbol{\nu})})^2).\end{aligned}$$

Thus $\mathbb{E}[(M_k - M_{k-1})^2 \mid \mathcal{F}_{k-1}] = (1-\lambda_2)(1 + \lambda_2 y(X_{p(k)}^{(\boldsymbol{\nu})})^2)$. For notational simplicity, denote

$$V_n = \frac{1}{n}\sum_{k=1}^{n} \mathbb{E}\left[(M_k - M_{k-1})^2 \mid \mathcal{F}_{k-1}\right] = \frac{1-\lambda_2}{n}\sum_{k=1}^{n}(1 + \lambda_2 y(X_{p(k)}^{(\boldsymbol{\nu})})^2).$$

When $\mathbb{T}$ is an m-tree, each node from level 0 to $t-1$ is counted m times as a parent. Define a new node feature $\boldsymbol{y}' = (y(1)^2, y(2)^2)^\top$. Let $\hat{\gamma}_n^{(\boldsymbol{\nu})}$ be the sample average of $y'(X_\sigma^{(\boldsymbol{\nu})})$'s up to node n. By Lemma 6.4 applied to $\hat{\gamma}_n^{(\boldsymbol{\nu})}$,

$$\begin{aligned} V_n &= 1 - \lambda_2 + \lambda_2(1-\lambda_2)\frac{m\sum_{k=1}^{p(n)} y'(X_k^{(\nu)}) + O(1)}{n} \\ &= 1 - \lambda_2 + \lambda_2(1-\lambda_2)\hat{\gamma}_{p(n)}^{(\boldsymbol{\nu})} + O(1/n) \\ &\xrightarrow{L^2} 1 - \lambda_2 + \lambda_2(1-\lambda_2)\mathbb{E}_{\boldsymbol{\pi}}(y') = 1 - \lambda_2^2 \end{aligned}$$

as $n \to \infty$. Here $O(1)$ in the first line comes from the fact that $y'(X_{p(n)}^{(\boldsymbol{\nu})})$ might be counted less than m times, which results in a remainder term bounded by $m\|\boldsymbol{y}'\|_\infty$, and the second line uses $\mathbb{E}_{\boldsymbol{\pi}}(y') = \mathsf{Var}_{\boldsymbol{\pi}}(y) = 1$. Since L^2 convergence implies convergence in probability, the first condition is verified.

We now move on to the second condition. Notice that

$$|M_k - M_{k-1}| = \left|y(X_k^{(\boldsymbol{\nu})}) - \lambda_2 y(X_{p(k)}^{(\boldsymbol{\nu})})\right| \le (1+\lambda_2)\|\boldsymbol{y}\|_\infty.$$

So $\mathbb{P}(|M_k - M_{k-1}| > \epsilon\sqrt{n}) = 0$ when $n > \epsilon^{-2}(1+\lambda_2)^2\|\boldsymbol{y}\|_\infty^2$. This gives

$$\frac{1}{n}\sum_{k=1}^{n}\mathbb{E}((M_k - M_{k-1})^2; |M_k - M_{k-1}| > \epsilon\sqrt{n}) = 0$$

for sufficiently large n. Thus the second condition is verified.

As a result, we obtain from Theorem 6.1 that $M_n/\sqrt{n} \to \mathcal{N}(0, 1-\lambda_2^2)$ in distribution. Combined with (6.6) and Slutsky's theorem, one finally arrives at

$$\sqrt{n_t}\left[\hat{\mu}_{\mathsf{GLS},t}^{(\boldsymbol{\nu})} - \mathbb{E}_{\boldsymbol{\pi}}(y)\right] \xrightarrow{d} \mathcal{N}\left(0, \frac{1+\lambda_2}{1-\lambda_2}\mathsf{Var}_{\boldsymbol{\pi}}(y)\right).$$

7. Discussion

We prove the existence of a limit distribution for the IPW estimator under the Markov model of respondent-driven sampling and show that this limit distribution depends on the seed node—thus the limit distribution is a non-trivial mixture distribution when the seed is randomized. This result also shows that the "seed bias" of IPW is non-negligible. We also establish the asymptotic normality of the GLS estimator under certain conditions and show that this limiting normal does not depend on the seed node. This implies that the "seed bias" of GLS is negligible. Both results allow for the VH adjustment. Our empirical study on social networks as well as on simulated data illustrate that these theoretical results appear to hold beyond the technical conditions given in the theorems.

Acknowledgements

Y. Yan was partially supported by the elite undergraduate training program of School of Mathematical Sciences in Peking University. S. Roch is supported by the NSF grants DMS-1614242 and CCF-1740707 (TRIPODS) and DMS-1916378, and by a Simons Fellowship. K. Rohe is supported by the NSF grants DMS-1612456 and DMS-1916378, and by the ARO grant W911NF-15-1-0423.

Appendix A: Proof of Equation (2.2)

First we use mathematical induction to show that, for every $n \in \mathbb{Z}^+$, the following statement $P(n)$ holds:

For any given referral tree $\mathbb{T}$ with n vertices $\{\sigma_1, \ldots, \sigma_n\}$, for any initial distribution $\boldsymbol{\nu}$ of X_0 and $b_1, \ldots, b_n \in \{1, \ldots, k\}$, the following holds

$$\mathbb{P}(B_{\sigma_1}^{(\boldsymbol{\mu})} = b_1, \ldots, B_{\sigma_n}^{(\boldsymbol{\mu})} = b_n) = \mathbb{P}(b(X_{\sigma_1}^{(\boldsymbol{\nu})}) = b_1, \ldots, b(X_{\sigma_n}^{(\boldsymbol{\nu})}) = b_n) \tag{A.1}$$

with $\mu_j = \sum_{i \in G: b(i)=j} \nu_i$ for $j = 1, \ldots, k$.

Base case: We prove that $P(1)$ holds. Since $\mathbb{T}$ only contains the seed vertex 0, it suffices to show that $\mathbb{P}(B_0^{(\boldsymbol{\mu})} = b_0) = \mathbb{P}(b(X_0^{(\boldsymbol{\nu})}) = b_0)$ for any $b_0 \in \{1, \ldots, k\}$. However,

$$\mathbb{P}(b(X_0^{(\boldsymbol{\nu})}) = b_0) = \sum_{i \in G: b(i)=b_0} \nu_i = \mu_{b_0} = \mathbb{P}(B_0^{(\boldsymbol{\mu})} = b_0).$$

So $P(0)$ is true.

Inductive step: We prove that if $P(n-1)$ holds for some unspecified value of $n \geq 2$, then $P(n)$ also holds. Assume σ_n is a leaf node (i.e. σ_n has no descendant) and σ_{n-1} is the parent of σ_n. Then $\mathbb{T} \setminus \{\sigma_n\}$ is a referral tree with $n-1$ vertex.

By the Markov property,

$$\begin{aligned}
&\mathbb{P}\left(b(X_{\sigma_n}^{(\boldsymbol{\nu})}) = b_n \mid b(X_{\sigma_1}^{(\boldsymbol{\nu})}) = b_1, \ldots, b(X_{\sigma_{n-1}}^{(\boldsymbol{\nu})}) = b_{n-1}\right) \\
&= \Bigg[\sum_{i \in G: b(i) = b_{n-1}} \mathbb{P}\left(b(X_{\sigma_n}^{(\boldsymbol{\nu})}) = b_n \mid B_{\sigma_{n-1}}^{(\boldsymbol{\nu})} = i\right) \\
&\quad \times \mathbb{P}\Big(X_{\sigma_{n-1}}^{(\boldsymbol{\nu})} = i \mid b(X_{\sigma_1}^{(\boldsymbol{\nu})}) = b_1, \ldots, b(X_{\sigma_{n-2}}^{(\boldsymbol{\nu})}) = b_{n-2}\Big)\Bigg] / \\
&\quad \mathbb{P}\left(b(X_{\sigma_{n-1}}^{(\boldsymbol{\nu})}) = b_{n-1} \mid b(X_{\sigma_1}^{(\boldsymbol{\nu})}) = b_1, \ldots, b(X_{\sigma_{n-2}}^{(\boldsymbol{\nu})}) = b_{n-2}\right) \\
&= \frac{\sum_{i \in G: b(i) = b_{n-1}} \mathcal{P}_{b_{n-1} b_n} \mathbb{P}\left(X_{\sigma_{n-1}}^{(\boldsymbol{\nu})} = i \mid b(X_{\sigma_1}^{(\boldsymbol{\nu})}) = b_1, \ldots, b(X_{\sigma_{n-2}}^{(\boldsymbol{\nu})}) = b_{n-2}\right)}{\mathbb{P}\left(b(X_{\sigma_{n-1}}^{(\boldsymbol{\nu})}) = b_{n-1} \mid b(X_{\sigma_1}^{(\boldsymbol{\nu})}) = b_1, \ldots, b(X_{\sigma_{n-2}}^{(\boldsymbol{\nu})}) = b_{n-2}\right)} \\
&\quad = \mathcal{P}_{b_{n-1} b_n} = \mathbb{P}\left(B_{\sigma_n}^{(\boldsymbol{\mu})} = b_n \mid B_{\sigma_1}^{(\boldsymbol{\mu})} = b_1, \ldots, B_{\sigma_{n-1}}^{(\boldsymbol{\mu})} = b_{n-1}\right).
\end{aligned}$$

Additionally, the induction hypothesis that $P(n-1)$ holds gives

$$\mathbb{P}(B_{\sigma_1}^{(\boldsymbol{\mu})} = b_1, \ldots, B_{\sigma_{n-1}}^{(\boldsymbol{\mu})} = b_{n-1}) = \mathbb{P}(b(X_{\sigma_1}^{(\boldsymbol{\nu})}) = b_1, \ldots, b(X_{\sigma_{n-1}}^{(\boldsymbol{\nu})}) = b_{n-1}).$$

The above two equations give (A.1), thereby showing $P(n)$ is true.

Since both the base case and the inductive step have been performed, by mathematical induction the statement $P(n)$ holds for all $n \in \mathbb{Z}^+$.

Finally we prove (2.2) based on the above result. Assume $\mathbb{T}$ has n vertices. For any $\{\sigma_{i_1}, \ldots, \sigma_{i_s}\} \subset \mathbb{T}$ and $b_{i_1}, \ldots, b_{i_s} \in \{1, \ldots, k\}$, let $\{\sigma_{j_1}, \ldots, \sigma_{j_{n-s}}\} = \mathbb{T} \setminus \{\sigma_{i_1}, \ldots, \sigma_{i_s}\}$. Then

$$\begin{aligned}
&\mathbb{P}(B_{\sigma_{i_1}}^{(\boldsymbol{\mu})} = b_{i_1}, \ldots, B_{\sigma_{i_s}}^{(\boldsymbol{\mu})} = b_{i_s}) \\
&\quad = \sum_{b_{j_1}=1}^{k} \cdots \sum_{b_{j_{n-s}}=1}^{k} \mathbb{P}(B_{\sigma_1}^{(\boldsymbol{\mu})} = b_1, \ldots, B_{\sigma_n}^{(\boldsymbol{\mu})} = b_n) \\
&\quad = \sum_{b_{j_1}=1}^{k} \cdots \sum_{b_{j_{n-s}}=1}^{k} \mathbb{P}(b(X_{\sigma_1}^{(\boldsymbol{\nu})}) = b_1, \ldots, b(X_{\sigma_n}^{(\boldsymbol{\nu})}) = b_n) \\
&\quad = \mathbb{P}(b(X_{\sigma_{i_1}}^{(\boldsymbol{\nu})}) = b_{i_1}, \ldots, b(X_{\sigma_{i_s}}^{(\boldsymbol{\nu})}) = b_{i_s})
\end{aligned}$$

as claimed.

Appendix B: Proofs: Sample average

Define the mean matrix $\boldsymbol{M} \in \mathbb{R}^{N \times N}$ as

$$\boldsymbol{M} = \{\mathbb{E} Z_{1,j}^{(i)} : i, j = 1, \ldots, N\}.$$

Let $\boldsymbol{V}_i$ denote the variance-covariance matrix of $\boldsymbol{Z}_1^{(i)}$, and define

$$\boldsymbol{C}_t^{(i)} = \{\mathbb{E}Z_{t,j}^{(i)} Z_{t,k}^{(i)} : j,k = 1,\ldots,N\}.$$

All components of $\boldsymbol{M}$ and $\boldsymbol{C}_t^{(i)}$ are finite. The following lemma is a standard result of multitype Galton-Watson process, see e.g. [8] or [1].

Lemma B.1. *The expectation of $\boldsymbol{Z}_t^{(i)}$ and $\boldsymbol{C}_t^{(i)}$ can be calculated from*

$$\mathbb{E}(\boldsymbol{Z}_t^{(i)}) = \boldsymbol{Z}_0^{(i)} \boldsymbol{M}^t, \quad \text{and} \tag{B.1}$$

$$\boldsymbol{C}_t^{(i)} = (\boldsymbol{M}^\top)^t \boldsymbol{C}_0^{(i)} \boldsymbol{M}^t + \sum_{l=1}^{t} (\boldsymbol{M}^\top)^{t-l} \left(\sum_{k=1}^{N} \boldsymbol{V}_k \mathbb{E} Z_{l-1,k}^{(i)} \right) \boldsymbol{M}^{t-l}. \tag{B.2}$$

B.1. Proof of Lemma 6.1

For the Markov model, $\boldsymbol{M} = m\boldsymbol{P}$ so $\boldsymbol{f}_j$ is the eigenvector of $\boldsymbol{M}$ corresponding to the eigenvalue $m\lambda_j$. The following lemma comes from the well-established theory of multitype Galton-Watson process.

Lemma B.2. *Let $\boldsymbol{\xi}$ be a right eigenvector of $\boldsymbol{M}$ and λ be the corresponding eigenvalue. Then*

$$\lambda^{-t} \langle \boldsymbol{Z}_t^{(i)}, \boldsymbol{\xi} \rangle$$

is a (complex-valued) martingale adapted to $\mathcal{F}_t = \sigma(\boldsymbol{Z}_l^{(i)} : 1 \le l \le t)$.

Proof. See Theorem 4' on Page 196 of [1]. □

According to Lemma 3.1, all λ_j and $\boldsymbol{f}_j$ are real. One then applies Lemma B.2 to the Markov model with $\lambda = m\lambda_j$ and $\boldsymbol{\xi} = \boldsymbol{f}_j$ to complete the proof.

B.2. Proof of Lemma 6.2

The next theorem is the martingale L^p convergence theorem (see e.g. [5]).

Theorem B.1. *If X_n is a martingale with $\sup_n \mathbb{E}\,|X_n|^p < \infty$ where $p > 1$, then $X_n \to X$ almost surely and in L^p.*

It is essential to derive the variance of $\langle \boldsymbol{Z}_t^{(i)}, f_j \rangle$ before applying Theorem B.1 to the martingales $Y_{t,j}^{(i)}, j \ge 2$. We conclude the result in the following claim and defer the proof to the end of this section.

Claim B.1. *The variance of $\langle \boldsymbol{Z}_t^{(i)}, \boldsymbol{f}_j \rangle$ is*

$$\mathsf{Var}(\langle \boldsymbol{Z}_t^{(i)}, \boldsymbol{f}_j \rangle) = \begin{cases} O((m\lambda_j)^{2t}) & \text{if } m\lambda_j^2 > 1, \\ O(t(m\lambda_j)^{2t}) & \text{if } m\lambda_j^2 = 1, \\ O(m^t) & \text{if } m\lambda_j^2 < 1. \end{cases} \tag{B.3}$$

We begin with $Y_2^{(i)}$. By Theorem B.1, we only need to show $\sup_t \mathbb{E}(Y_{t,2}^{(i)})^2 < \infty$. However,

$$\mathbb{E}(Y_{t,2}^{(i)})^2 = \mathsf{Var}(Y_{t,2}^{(i)}) + (\mathbb{E}Y_{t,2}^{(i)})^2 = (m\lambda_2)^{-2t}\,\mathsf{Var}(\langle \boldsymbol{Z}_t^{(i)}, \boldsymbol{f}_2\rangle) + (Y_{0,2}^{(i)})^2.$$

Since $m > \lambda_2^{-2}$, by Claim B.1, $\mathsf{Var}(\langle \boldsymbol{Z}_t^{(i)}, \boldsymbol{f}_2\rangle) = O((m\lambda_2)^{2t})$. This gives $\sup_t \mathbb{E}(Y_{t,2}^{(i)})^2 < \infty$.

We move on to $Y_j^{(i)}$ for $j \geq 3$. By Theorem B.1,

$$\begin{aligned}\mathbb{E}[(\lambda_2^{-1}\lambda_j)^t Y_{t,j}^{(i)}]^2 &= (\lambda_2^{-1}\lambda_j)^{2t}[\mathsf{Var}(Y_{t,2}^{(i)}) + (\mathbb{E}Y_{t,2}^{(i)})^2] \\ &= (m\lambda_2)^{-2t}\,\mathsf{Var}(\langle \boldsymbol{Z}_t^{(i)}, \boldsymbol{f}_j\rangle) + (\lambda_2^{-1}\lambda_j)^{2t}(Y_{0,2}^{(i)})^2.\end{aligned}$$

Since $\lambda_2^{-1}\lambda_j < 1$, $(\lambda_2^{-1}\lambda_j)^{2t}(Y_{0,2}^{(i)})^2 \to 0$. Additionally, for $j \geq 3$,

$$(m\lambda_2)^{-2t}\,\mathsf{Var}(\langle \boldsymbol{Z}_t^{(i)}, \boldsymbol{f}_j\rangle) = \begin{cases} O((\lambda_2^{-1}\lambda_j)^{2t}) & \text{if } m\lambda_j^2 > 1 \\ O(t(\lambda_2^{-1}\lambda_j)^{2t}) & \text{if } m\lambda_j^2 = 1 \\ O((m\lambda_2^2)^{-t}) & \text{if } m\lambda_j^2 < 1 \end{cases}$$

which converges to 0 in all cases. Thus $\mathbb{E}[(\lambda_2^{-1}\lambda_j)^t Y_{t,j}^{(i)}]^2 \to 0$, which leads to $(\lambda_2^{-1}\lambda_j)^t Y_{t,j}^{(i)} \to 0$ in L^2. To prove almost sure convergence, let $\delta = \max\{(\lambda_2^{-1}\lambda_j)^2, m\lambda_2^{-2}\} \in (0,1)$. There exists $C > 0$ such that

$$\mathbb{E}[(\lambda_2^{-1}\lambda_j)^t Y_{t,j}^{(i)}]^2 \leq Ct\delta^t$$

always holds. Then $\forall \epsilon > 0$,

$$\mathbb{P}(|(\lambda_2^{-1}\lambda_j)^t Y_{t,j}^{(i)}| > \epsilon) \leq \epsilon^{-2}\mathbb{E}[(\lambda_2^{-1}\lambda_j)^t Y_{t,j}^{(i)}]^2 \leq \epsilon^{-2}Ct\delta^t.$$

So

$$\sum_{t=1}^{\infty} \mathbb{P}(|(\lambda_2^{-1}\lambda_j)^t Y_{t,j}^{(i)}| > \epsilon) \leq \epsilon^{-2}C\sum_{t=1}^{\infty} t\delta^t < \infty.$$

By the Borel-Cantelli lemma, $(\lambda_2^{-1}\lambda_j)^t Y_{t,j}^{(i)} \to 0$ almost surely.

Proof of Claim B.1. From Lemma B.1 and the fact that $\boldsymbol{C}_0^{(i)} = (\boldsymbol{Z}_0^{(i)})^\top \boldsymbol{Z}_0^{(i)}$,

$$\begin{aligned}\mathsf{Var}(\boldsymbol{Z}_t^{(i)}) &= \boldsymbol{C}_t^{(i)} - (\boldsymbol{M}^\top)^t (\boldsymbol{Z}_0^{(i)})^\top \boldsymbol{Z}_0^{(i)} \boldsymbol{M}^t \\ &= \sum_{l=1}^{t} (\boldsymbol{M}^\top)^{t-l} \left(\sum_{k=1}^{N} \boldsymbol{V}_k \mathbb{E} Z_{l-1,k}^{(i)}\right) \boldsymbol{M}^{t-l}.\end{aligned}$$

As a result,

$$\mathsf{Var}(\langle \boldsymbol{Z}_t^{(i)}, \boldsymbol{f}_j\rangle) = \sum_{l=1}^{t} \boldsymbol{f}_j^\top (\boldsymbol{M}^\top)^{t-l} \left(\sum_{k=1}^{N} \boldsymbol{V}_k \mathbb{E} Z_{l-1,k}^{(i)}\right) \boldsymbol{M}^{t-l} \boldsymbol{f}_j.$$

Since $\boldsymbol{f}_j$ is the eigenvector of $\boldsymbol{M}$ corresponding to the eigenvalue $m\lambda_j$, for every $n \in \mathbb{Z}^+$, $\boldsymbol{M}^n \boldsymbol{f}_j = (m\lambda_j)^n \boldsymbol{f}_j$. This yields

$$\mathsf{Var}(\langle \boldsymbol{Z}_t^{(i)}, \boldsymbol{f}_j \rangle) = \sum_{l=1}^{t} (m\lambda_j)^{2t-2l} \sum_{k=1}^{N} (\boldsymbol{f}_j^\top \boldsymbol{V}_k \boldsymbol{f}_j) \mathbb{E} Z_{l-1,k}^{(i)}. \tag{B.4}$$

Notice that $\sum_{k=1}^{N} \mathbb{E} Z_{l-1,k} = m^{l-1}$,

$$\mathsf{Var}(\langle \boldsymbol{Z}_t^{(i)}, \boldsymbol{f}_j \rangle) \le c \sum_{l=1}^{t} (m\lambda_j)^{2t-2k} m^k = c(m\lambda_j)^{2t} \sum_{l=1}^{t} (m\lambda_j^2)^{-l},$$

where $c = \max\{\boldsymbol{f}_j^\top \boldsymbol{V}_k \boldsymbol{f}_j : 1 \le j, k \le N\}$. This gives (B.3). □

B.3. Proof of Lemma 6.3

By Lemma 6.1 and 6.2, $\mathbb{E} Y_2^{(i)} = Y_{0,2}^{(i)} = f_2(i)$ and

$$\mathsf{Var}(Y_2^{(i)}) = \lim_{t\to\infty} \mathsf{Var}(Y_{t,2}^{(i)}) = \lim_{t\to\infty} (m\lambda_2)^{-2t} \mathsf{Var}(\langle \boldsymbol{Z}_t^{(i)}, \boldsymbol{f}_2 \rangle).$$

By (B.4)

$$(m\lambda_2)^{-2t} \mathsf{Var}(\langle \boldsymbol{Z}_t^{(i)}, \boldsymbol{f}_2 \rangle) = \sum_{l=1}^{t} (m\lambda_2)^{-2l} \sum_{k=1}^{N} (\boldsymbol{f}_2^\top \boldsymbol{V}_k \boldsymbol{f}_2) \mathbb{E} Z_{l-1,k}^{(i)}.$$

Notice that $\boldsymbol{V}_k = m\left(\mathsf{diag}\{P_{k1}, \ldots, P_{kN}\} - \boldsymbol{P}_k \boldsymbol{P}_k^T\right)$, where $\boldsymbol{P}_k^\top = (P_{k1}, \ldots, P_{kN})$ is the k-th row of $\boldsymbol{P}$. Notice that $\sum_{j=1}^{N} P_{kj} = 1$, by the Jensen's inequality

$$\boldsymbol{f}_2^\top \boldsymbol{V}_k \boldsymbol{f}_2 = m \sum_{j=1}^{N} P_{kj} f_2(j)^2 - m \left(\sum_{j=1}^{N} P_{kj} f_2(j) \right)^2 \ge 0 \tag{B.5}$$

for any $k = 1, \ldots, N$. The assumptions $\langle \boldsymbol{f}_1, \boldsymbol{f}_2 \rangle_\pi = 0$ and $\boldsymbol{f}_1 = \boldsymbol{1}$ imply that f_2 is not a constant vector, thus the equality in (B.5) does not hold. Similar to the proof of Theorem B.1,

$$(m\lambda_2)^{-2t} \mathsf{Var}(\langle \boldsymbol{Z}_t^{(i)}, \boldsymbol{f}_2 \rangle) \ge c \sum_{l=1}^{t} (m\lambda_2)^{-2l} m^k = c \sum_{l=1}^{t} (m\lambda_2^2)^{-l}$$

where $c = \min\{\boldsymbol{f}_2^\top \boldsymbol{V}_k \boldsymbol{f}_2 : 1 \le k \le N\} > 0$. Since $m\lambda_2^2 > 1$, this yields $\mathsf{Var}(Y_2^{(i)}) > 0$ for any $i = 1, \ldots, N$.

B.4. Proof of Corollary 3.1

The L^2 convergence in Theorem 3.1 implies L^1 convergence. If a sequence of random variables $X_n \xrightarrow{L^1} X$, then $|\mathbb{E}(X_n - X)| \le \mathbb{E}\,|X_n - X|$ implies $\mathbb{E}X_n \to \mathbb{E}X$. So

$$\lim_{t\to\infty} \mathbb{E}\left(\lambda_2^{-t}[\hat{\mu}_t^{(i)} - \mathbb{E}_{\boldsymbol{\pi}}(y)]\right) = \mathbb{E}X^{(i)} = \frac{(m-1)\lambda_2}{m\lambda_2 - 1}\langle \boldsymbol{y}, \boldsymbol{f}_2\rangle_{\boldsymbol{\pi}} f_2(i).$$

Since $\langle \boldsymbol{y}, \boldsymbol{f}_2\rangle_{\boldsymbol{\pi}} f_2(i) \neq 0$, the bias term decays like

$$\left(\mathbb{E}(\hat{\mu}_t^{(i)}) - \mathbb{E}_{\boldsymbol{\pi}}(y)\right)^2 = \Theta(\lambda_2^{2t}).$$

Additionally, the L^2 convergence in Theorem 3.1 also yields

$$\lambda_2^{-2t}\,\mathsf{Var}(\hat{\mu}_t^{(i)}) = \mathsf{Var}(\lambda_2^{-t}[\hat{\mu}_t^{(i)} - \mathbb{E}_{\boldsymbol{\pi}}(y)]) \xrightarrow{t\to\infty} \mathsf{Var}(X^{(i)}) > 0.$$

So the variance term decays like $\mathsf{Var}(\hat{\mu}_t^{(i)}) = \Theta(\lambda_2^{2t})$.

B.5. Proof of Corollary 3.2

By the definition of the VH estimator in Section 2.4,

$$\hat{\mu}_{\mathsf{VH},t}^{(i)} = H_t \cdot \frac{1}{n_t} \sum_{\sigma\in\mathbb{T},|\sigma|\le t} \frac{y(X_\sigma^{(i)})}{\deg(X_\sigma^{(i)})}, \quad \text{where } H_t^{-1} = \frac{1}{n_t}\sum_{\sigma\in\mathbb{T}} \frac{1}{\deg(X_\sigma^{(i)})}.$$

H_t^{-1} is the sample average of $y'(X_\sigma^{(i)})$'s up to generation t, where $y'(j) = \deg(j)^{-1}$. In view of Theorem 3.1, H_t^{-1} converges to $\mathbb{E}_{\boldsymbol{\pi}}(y') > 0$ almost surely. Additionally,

$$\hat{\mu}_t'' = \frac{1}{n_t} \sum_{\sigma\in\mathbb{T},|\sigma|\le t} \frac{y(X_\sigma^{(i)})}{\deg(X_\sigma^{(i)})}$$

is the sample average of $y''(X_\sigma^{(i)})$'s up to generation t, where $y''(j) = y(j)/\deg(j)$. By Theorem 3.1, there exists some random variable $\bar{X}^{(i)} \in L^2$ such that $\lambda_2^{-t}[\hat{\mu}_t'' - \mathbb{E}_{\boldsymbol{\pi}}(y'')] \to \bar{X}^{(i)}$ almost surely and in L^2. So

$$\lambda_2^{-t}\left[\hat{\mu}_{\mathsf{VH},t}^{(i)} - \frac{\mathbb{E}_{\boldsymbol{\pi}}(y'')}{\mathbb{E}_{\boldsymbol{\pi}}(y')}\right] \to \mathbb{E}_{\boldsymbol{\pi}}(y')^{-1}\bar{X}^{(i)} \triangleq \tilde{X}^{(i)}$$

almost surely. Notice that $\mathbb{E}_{\boldsymbol{\pi}}(y') = N/\mathrm{vol}(G)$ and $\mathbb{E}_{\boldsymbol{\pi}}(y'') = \sum_i y(i)/\mathrm{vol}(G)$, this gives the result that

$$\lambda_2^{-t}\left[\hat{\mu}_{\mathsf{VH},t}^{(i)} - \mu_{\mathsf{true}}\right] \to \tilde{X}^{(i)}$$

almost surely. The mean and variance of $\tilde{X}^{(i)}$ comes directly from Theorem 3.1.

Appendix C: Proofs: GLS estimator

This section contains the proof of Lemma 6.4 and Corollary 3.3.

C.1. Proof of Lemma 6.4

For a given n, there exists t such that $n_{t-1} \le n < n_t$. Throughout this proof, t is determined by the corresponding n in this way.

We consider two cases. First, when $n < n_{t-1} + m^{t-1}$, in base m, $n_t - n$ is represented as

$$n_t - n = a_{t-1}m^{t-1} + \cdots + a_1 m + a_0, \tag{C.1}$$

where $a_i \in \{0, 1, \ldots, m-1\}$ for $0 \le i \le t-1$, $a_{t-1} \ge 1$. And $\hat{\mu}_n^{(\nu)}$ can be represented as

$$\hat{\mu}_n^{(\boldsymbol{\nu})} = \frac{n_t \hat{\mu}_t^{(\boldsymbol{\nu})} - \sum_{k=n+1}^{n_t} y(X_k^{(\boldsymbol{\nu})})}{n}. \tag{C.2}$$

Note that $\{X_k^{(\boldsymbol{\nu})} : n_t - m^{t-1} + 1 \le k \le n_t\}$ form the $(t-1)$-st generation of a subtree of $\mathbb{T}$ (rooted at a child of the root $\mathbb{T}$) and let $W_{t-1}^1 = \sum_{k=n_t-m^{t-1}+1}^{n_t} y(X_k^{(\nu)})$. Similarly we can determine a_{t-1} such subtrees by scanning the nodes from right to left in the t-th generation of $\mathbb{T}$ and define accordingly $W_{t-1}^2, \ldots, W_{t-1}^{a_{t-1}}$. Next we can determine a subtree of $\mathbb{T}$ where the next m^{t-2} nodes in the t-th generation of $\mathbb{T}$ form its $(t-2)$-nd generation. We can determine a_{t-2} such subtrees by continuing to scan the nodes from right to left in the t-th generation of $\mathbb{T}$ and define $W_{t-2}^1, \ldots, W_{t-2}^{a_{t-2}}$. And so on. By (C.1),

$$\sum_{k=n+1}^{n_t} y(X_k^{(\boldsymbol{\nu})}) = \sum_{k=1}^{a_{t-1}} W_{t-1}^k + \sum_{k=1}^{a_{t-2}} W_{t-2}^k + \cdots + \sum_{k=1}^{a_0} W_0^k.$$

To proceed, we need the following concentration bounds for $m^{-t}W_t^{(\boldsymbol{\nu})}$ and $\hat{\mu}_t^{(\boldsymbol{\nu})}$ (proof below).

Claim C.1. *For any initial distribution $\boldsymbol{\nu}$ of X_0, $m^{-t}W_t^{(\boldsymbol{\nu})} \to \mathbb{E}_{\boldsymbol{\pi}}(y)$ and $\hat{\mu}_t^{(\boldsymbol{\nu})} \to \mathbb{E}_{\boldsymbol{\pi}}(y)$ in L^2. For any $0 < \delta < 1$, there exists $C > 0$ such that*

$$\mathbb{E}[(m^{-t}W_t^{(\boldsymbol{\nu})} - \mathbb{E}_{\boldsymbol{\pi}}(y))^2] \le Cm^{-(1-\delta)t}, \quad \mathbb{E}[(\hat{\mu}_t^{(\boldsymbol{\nu})} - \mathbb{E}_{\boldsymbol{\pi}}(y))^2] \le Ctm^{-(1-\delta)t}. \tag{C.3}$$

The constant C does not depend on the initial distribution ν.

Then by Claim C.1, the triangle inequality, $a_{t-1} \ge 1$ and $a_l \le m-1$ for $0 \le l \le t-1$, one has

$$\left\| \frac{\sum_{k=n+1}^{n_t} y(X_k^{(\boldsymbol{\nu})})}{n_t - n} - \mathbb{E}_{\boldsymbol{\pi}}(y) \right\|_{L^2} = \left\| \frac{\sum_{l=0}^{t-1} m^l \sum_{k=1}^{a_l} m^{-l}(W_l^k - \mathbb{E}_{\boldsymbol{\pi}}(y))}{a_{t-1}m^{t-1} + \cdots + a_1 m + a_0} \right\|_{L^2}$$
$$\le \sum_{l=0}^{t-1} \frac{a_l m^l}{a_{t-1}m^{t-1} + \cdots + a_1 m + a_0} Cm^{-\frac{1-\delta}{2}l}$$

$$\leq C \sum_{l=0}^{t-1} \frac{(m-1)m^l}{m^{t-1}} m^{-\frac{1-\delta}{2}l} = O(m^{-\frac{1-\delta}{2}t}). \tag{C.4}$$

For any subsequence such that $n < n_{t-1} + m^{t-1}$, $n \to \infty$ implies $t \to \infty$. As a result,

$$\frac{\sum_{k=n+1}^{n_t} y(X_k^{(\boldsymbol{\nu})})}{n_t - n} \xrightarrow{L^2} \mathbb{E}_{\boldsymbol{\pi}}(y).$$

From Claim C.1, $\hat{\mu}_t^{(\boldsymbol{\nu})} \xrightarrow{L^2} \mathbb{E}_{\boldsymbol{\pi}}(y)$. By (C.2), the triangle inequality, the fact that $n_t/n = O(1)$ and $(n_t - n)/n = O(1)$,

$$\begin{aligned}\left\|\hat{\mu}_n^{(\nu)} - \mathbb{E}_{\boldsymbol{\pi}}(y)\right\|_{L^2} &\leq \frac{n_t}{n}\left\|\hat{\mu}_t^{(\nu)} - \mathbb{E}_{\boldsymbol{\pi}}(y)\right\|_{L^2} \\ &\quad + \frac{n_t - n}{n}\left\|\frac{\sum_{k=n+1}^{n_t} y(X_k^{(\nu)})}{n_t - n} - \mathbb{E}_{\boldsymbol{\pi}}(y)\right\|_{L^2} \\ &\xrightarrow{n\to\infty} 0.\end{aligned}$$

In the second case, when $n \geq n_{t-1} + m^{t-1}$, in base m, $n - n_{t-1}$ is represented as

$$n - n_{t-1} = a_{t-1}m^{t-1} + \cdots + a_1 m + a_0, \tag{C.5}$$

where $a_i \in \{0, 1, \ldots, m-1\}$ for $0 \leq i \leq t-1$, $a_{t-1} \geq 1$. And $\hat{\mu}_n^{(\boldsymbol{\nu})}$ can be represented as

$$\hat{\mu}_n^{(\boldsymbol{\nu})} = \frac{n_{t-1}\hat{\mu}_{t-1}^{(\boldsymbol{\nu})} + \sum_{k=n_{t-1}+1}^{n} y(X_k^{(\boldsymbol{\nu})})}{n}. \tag{C.6}$$

Arguing as above, we can write

$$\sum_{k=n_{t-1}+1}^{n} y(X_k^{(\boldsymbol{\nu})}) = \sum_{k=1}^{a_{t-1}} W_{t-1}^k + \sum_{k=1}^{a_{t-2}} W_{t-2}^k + \cdots + \sum_{k=1}^{a_0} W_0^k.$$

Similarly to the previous case, we can prove that for any subsequence such that $n \geq n_{t-1} + m^{t-1}$, when $n \to \infty$,

$$\frac{\sum_{k=n_{t-1}+1}^{n} y(X_k^{(\boldsymbol{\nu})})}{n - n_{t-1}} \xrightarrow{L^2} \mathbb{E}_{\boldsymbol{\pi}}(y),$$

and

$$\begin{aligned}&\left\|\hat{\mu}_n^{(\boldsymbol{\nu})} - \mathbb{E}_{\boldsymbol{\pi}}(y)\right\|_{L^2} \\ &\quad\leq \frac{n_{t-1}}{n}\left\|\hat{\mu}_{t-1}^{(\boldsymbol{\nu})} - \mathbb{E}_{\boldsymbol{\pi}}(y)\right\|_{L^2} + \frac{n - n_{t-1}}{n}\left\|\frac{\sum_{k=n_{t-1}+1}^{n} y(X_k^{(\nu)})}{n - n_{t-1}} - \mathbb{E}_{\boldsymbol{\pi}}(y)\right\|_{L^2} \\ &\quad\xrightarrow{n\to\infty} 0.\end{aligned}$$

Since $\hat{\mu}_n^{(\boldsymbol{\nu})} \xrightarrow{L^2} \mathbb{E}_{\boldsymbol{\pi}}(y)$ holds for both $n < n_{t-1} + m^{t-1}$ and $n \geq n_{t-1} + m^{t-1}$ as $n \to \infty$, one finally arrives at $\hat{\mu}_n^{(\boldsymbol{\nu})} \xrightarrow{L^2} \mathbb{E}_{\boldsymbol{\pi}}(y)$, which completes the proof.

Proof of Claim C.1. The proof is similar to the proof of Theorem 3.1. First,

$$\mathbb{E}[(\lambda_2^t Y_{t,2}^{(i)})^2] = \mathsf{Var}(\lambda_2^t Y_{t,2}^{(i)}) + (\lambda_2^t \mathbb{E} Y_{t,2}^{(i)})^2 = m^{-2t}\mathsf{Var}(\langle \boldsymbol{Z}_t^{(i)}, \boldsymbol{f}_2 \rangle) + (\lambda_2^t Y_{0,2}^{(i)})^2.$$

From (B.3), for any $0 < \delta < 1$, $\mathsf{Var}(\langle \boldsymbol{Z}_t^{(i)}, \boldsymbol{f}_2 \rangle) = O(m^{(1+\delta)t})$ holds for all $i \in G$. As a result,

$$\mathbb{E}[(\lambda_2^t Y_{t,2}^{(i)})^2] = O(m^{-(1-\delta)t}). \tag{C.7}$$

From (6.2), one has $m^{-t} W_t^{(i)} - \mathbb{E}_{\boldsymbol{\pi}}(y) = \langle \boldsymbol{y}, \boldsymbol{f}_2 \rangle_{\boldsymbol{\pi}} \lambda_2^t Y_{t,2}^{(i)}$ and as a result,

$$\mathbb{E}[(m^{-t} W_t^{(i)} - \mathbb{E}_{\boldsymbol{\pi}}(y))^2] \leq \langle \boldsymbol{y}, \boldsymbol{f}_2 \rangle_{\boldsymbol{\pi}}^2 \mathbb{E}[(\lambda_2^t Y_{t,2}^{(i)})^2] = O(m^{-(1-\delta)t}). \tag{C.8}$$

Recall from (6.3) that

$$\hat{\mu}_t^{(i)} - \mathbb{E}_{\boldsymbol{\pi}}(y) = \frac{m^t}{n_t} \sum_{l=0}^{t} \frac{W_l^{(i)} - m^l \mathbb{E}_{\boldsymbol{\pi}}(y)}{m^t}.$$

By the Cauchy-Schwarz inequality,

$$\begin{aligned} \mathbb{E}\left[(\hat{\mu}_t^{(i)} - \mathbb{E}_{\boldsymbol{\pi}}(y))^2\right] &\leq \left(\frac{m^t}{n_t}\right)^2 (t+1) \sum_{l=0}^{t} m^{2(l-t)} \mathbb{E}[(m^{-l} W_t^{(i)} - \mathbb{E}_{\boldsymbol{\pi}}(y))^2] \\ &= \left(\frac{m^t}{n_t}\right)^2 (t+1) \sum_{l=0}^{t} O(m^{-2t+(1+\delta)l}) \\ &= O(t m^{-(1-\delta)t}). \end{aligned} \tag{C.9}$$

So there exists $C > 0$ such that for all $i \in G$,

$$\mathbb{E}[(m^{-t} W_t^{(i)} - \mathbb{E}_{\boldsymbol{\pi}}(y))^2] \leq C m^{-(1-\delta)t}, \quad \mathbb{E}[(\hat{\mu}_t^{(i)} - \mathbb{E}_{\boldsymbol{\pi}}(y))^2] \leq C t m^{-(1-\delta)t}.$$

So for any initial distribution $\boldsymbol{\nu}$ of X_0, since $\sum_{i \in G} \nu_i = 1$,

$$\begin{aligned} \mathbb{E}[(\hat{\mu}_t^{(\boldsymbol{\nu})} - \mathbb{E}_{\boldsymbol{\pi}}(y))^2] &= \sum_{i \in G} \nu_i \mathbb{E}[(\hat{\mu}_t^{(i)} - \mathbb{E}_{\boldsymbol{\pi}}(y))^2] \leq C t m^{(1-\delta)t}, \\ \mathbb{E}[(m^{-t} W_t^{(\boldsymbol{\nu})} - \mathbb{E}_{\boldsymbol{\pi}}(y))^2] &= \sum_{i \in G} \nu_i \mathbb{E}[(m^{-t} W_t^{(i)} - \mathbb{E}_{\boldsymbol{\pi}}(y))^2] \leq C m^{-(1-\delta)t}. \end{aligned}$$

So $m^{-t} W_t^{(\boldsymbol{\nu})} \to \mathbb{E}_{\boldsymbol{\pi}}(y)$ and $\hat{\mu}_t^{(\boldsymbol{\nu})} \to \mathbb{E}_{\boldsymbol{\pi}}(y)$ in L^2. □

C.2. Proof of Corollary 3.3

By the definition of the GLS estimator with VH adjustment in Section 2.4,

$$\hat{\mu}^{(\boldsymbol{\nu})}_{\mathsf{GLS},\mathsf{VH},t} = H_t \cdot \sum_{\sigma\in\mathbb{T},|\sigma|\le t} w^*_{\sigma,t} \frac{y(X^{(\nu)}_\sigma)}{\deg(X^{(\nu)}_\sigma)}.$$

H_t^{-1} is the GLS estimator of $\mathbb{E}_{\boldsymbol{\pi}}(y')$ where $y'(i) = \deg(i)^{-1}$. So H_t^{-1} converges to $\mathbb{E}_{\boldsymbol{\pi}}(y')$ in distribution (thus in probability). Additionally,

$$\hat{\mu}''_{\mathsf{GLS},t} = \sum_{\sigma\in\mathbb{T},|\sigma|\le t} w^*_{\sigma,t} \frac{y(X^{(\boldsymbol{\nu})}_\sigma)}{\deg(X^{(\boldsymbol{\nu})}_\sigma)}$$

is the GLS estimator of $\mathbb{E}_{\boldsymbol{\pi}}(y'')$ where $y''(i) = y(i)/\deg(i)$. Then

$$\sqrt{n_t}\left[\hat{\mu}''_{\mathsf{GLS},t} - \mathbb{E}_{\boldsymbol{\pi}}(y'')\right] \xrightarrow{d} \mathcal{N}\left(0, \frac{1+\lambda_2}{1-\lambda_2}\mathsf{Var}_{\boldsymbol{\pi}}(y'')\right).$$

By Slutsky's theorem,

$$\sqrt{n_t}\left[\hat{\mu}^{(\boldsymbol{\nu})}_{\mathsf{GLS},\mathsf{VH},t} - \frac{\mathbb{E}_{\boldsymbol{\pi}}(y'')}{\mathbb{E}_{\boldsymbol{\pi}}(y')}\right] \xrightarrow{d} \mathcal{N}\left(0, \frac{1+\lambda_2}{1-\lambda_2}\mathbb{E}_{\boldsymbol{\pi}}(y')^{-2}\mathsf{Var}_{\boldsymbol{\pi}}(y'')\right).$$

Notice that $\mathbb{E}_{\boldsymbol{\pi}}(y') = N/\mathrm{vol}(G)$ and $\mathbb{E}_{\boldsymbol{\pi}}(y'') = \sum_i y(i)/\mathrm{vol}(G)$, this gives the result

$$\sqrt{n_t}\left[\hat{\mu}^{(\boldsymbol{\nu})}_{\mathsf{GLS},\mathsf{VH},t} - \mu_{\mathsf{true}}\right] \xrightarrow{d} \mathcal{N}\left(0, \frac{1+\lambda_2}{1-\lambda_2}\mathbb{E}_{\boldsymbol{\pi}}(y')^{-2}\mathsf{Var}_{\boldsymbol{\pi}}(y'')\right).$$

References

[1] Athreya, K. B. and Ney, P. E. (2004). *Branching processes*. Courier Corporation. MR0373040

[2] Baraff, A. J., McCormick, T. H. and Raftery, A. E. (2016). Estimating uncertainty in respondent-driven sampling using a tree bootstrap method. *Proceedings of the National Academy of Sciences* 201617258.

[3] Benjamini, I. and Peres, Y. (1994). Markov chains indexed by trees. *The Annals of Probability* 219–243. MR1258875

[4] CDC (2017). National HIV Behavioral Surveillance (NHBS). *Division of HIV/AIDS Prevention*.

[5] Durrett, R. (2019). *Probability: theory and examples* **49**. Cambridge university press. MR3930614

[6] Goel, S. and Salganik, M. J. (2009). Respondent-driven sampling as Markov chain Monte Carlo. *Statistics in medicine* **28** 2202–2229. MR2751515

[7] Harris, K. M. (2011). The national longitudinal study of adolescent health: Research design. http://www.cpc.unc.edu/projects/addhealth/design.
[8] Harris, T. E. (2002). *The theory of branching processes*. Courier Corporation. MR1991122
[9] Heckathorn, D. D. (1997). Respondent-driven sampling: a new approach to the study of hidden populations. *Social problems* **44** 174–199.
[10] Holland, P. W. and Laskey, K. B. (1983). Stochastic blockmodels: First steps. *Social Networks* **5** 109–137. MR0718088
[11] Johnston, L. (2013). Introduction to HIV/AIDS and sexually transmitted infection surveillance: Module 4: Introduction to Respondent Driven Sampling. *World Health Organization*.
[12] Kesten, H. and Stigum, B. P. (1966). Additional limit theorems for indecomposable multidimensional Galton-Watson processes. *The Annals of Mathematical Statistics* **37** 1463–1481. MR0200979
[13] Levin, D. A., Peres, Y. and Wilmer, E. L. (2009). *Markov chains and mixing times*. American Mathematical Soc. MR2466937
[14] Li, X. and Rohe, K. (2017). Central limit theorems for network driven sampling. *Electronic Journal of Statistics* **11** 4871–4895. MR3733297
[15] Roch, S. and Rohe, K. (2018). Generalized least squares can overcome the critical threshold in respondent-driven sampling. *Proceedings of the National Academy of Sciences* **115** 10299–10304. MR3865545
[16] Rohe, K. (2019). A critical threshold for design effects in network sampling. *The Annals of Statistics* **47** 556–582. MR3909942
[17] Salganik, M. J. and Heckathorn, D. D. (2004). Sampling and estimation in hidden populations using respondent-driven sampling. *Sociological methodology* **34** 193–240.
[18] Volz, E. and Heckathorn, D. D. (2008). Probability based estimation theory for respondent driven sampling. *Journal of official statistics* **24** 79.
[19] White, H. C., Boorman, S. A. and Breiger, R. L. (1976). Social Structure from Multiple Networks. I. Blockmodels of Roles and Positions. *American Journal of Sociology* **81** 730–780.
[20] White, R. G., Hakim, A. J., Salganik, M. J., Spiller, M. W., Johnston, L. G., Kerr, L., Kendall, C., Drake, A., Wilson, D., Orroth, K. et al. (2015). Strengthening the reporting of observational studies in epidemiology for respondent-driven sampling studies:"STROBE-RDS" statement. *Journal of clinical epidemiology* **68** 1463–1471.