A CRITICAL THRESHOLD FOR DESIGN EFFECTS IN NETWORK SAMPLING

BY KARL ROHE[1]

University of Wisconsin–Madison

Web crawling, snowball sampling, and respondent-driven sampling (RDS) are three types of network sampling techniques used to contact individuals in hard-to-reach populations. This paper studies these procedures as a Markov process on the social network that is indexed by a tree. Each node in this tree corresponds to an observation and each edge in the tree corresponds to a referral. Indexing with a tree (instead of a chain) allows for the sampled units to refer multiple future units into the sample.

In survey sampling, the design effect characterizes the additional variance induced by a novel sampling strategy. If the design effect is some value DE, then constructing an estimator from the novel design makes the variance of the estimator DE times greater than it would be under a simple random sample with the same sample size n. Under certain assumptions on the referral tree, the design effect of network sampling has a critical threshold that is a function of the referral rate m and the clustering structure in the social network, represented by the second eigenvalue of the Markov transition matrix, λ_2. If $m < 1/\lambda_2^2$, then the design effect is finite (i.e., the standard estimator is $\sqrt{n}$-consistent). However, if $m > 1/\lambda_2^2$, then the design effect grows with n (i.e., the standard estimator is no longer $\sqrt{n}$-consistent). Past this critical threshold, the standard error of the estimator converges at the slower rate of $n^{\log_m \lambda_2}$. The Markov model allows for nodes to be resampled; computational results show that the findings hold in without-replacement sampling. To estimate confidence intervals that adapt to the correct level of uncertainty, a novel resampling procedure is proposed. Computational experiments compare this procedure to previous techniques.

Introduction. This paper is motivated by respondent-driven sampling (RDS), a popular technique to sample marginalized and/or hard-to-reach populations [Heckathorn (1997)]. RDS has become particularly popular in HIV research because the populations most at risk for HIV (i.e., people who inject drugs, female sex workers, and men who have sex with men) cannot be sampled using conventional techniques. Several domestic and international institutions use RDS to quantify the prevalence of HIV in at risk populations, including the Centers for Disease Control and Prevention (CDC), the World Health Organization (WHO), and the
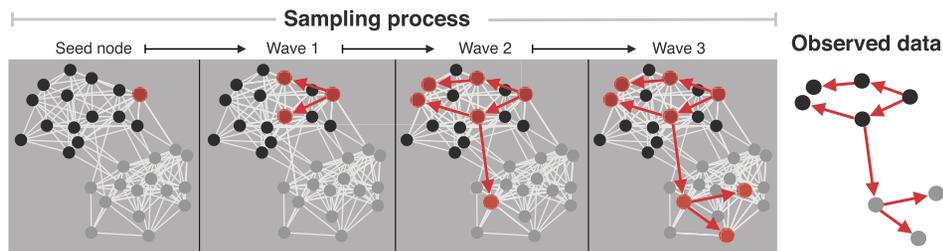
FIG. 1. *Network sampling has two graphs: the underlying social network and the referral tree. Each node in the social network has some feature (e.g., HIV status). In this diagram, the node feature is denoted by color. When we sample a node, we observe* (i) *the node's color and* (ii) *which node referred the node into the sample. In the end, we want to estimate the proportion of nodes that are grey.*

Joint United Nations Programme on HIV/AIDS (UNAIDS) [World Health Organization and UNAIDS (2013)]. It has been applied in over 460 different studies, in 69 different countries [White et al. (2015)].

The RDS process starts with a convenience sample of "seeds" from the target population. They form wave zero. These participants are incentivized to (1) participate in the study and (2) pass three (or sometimes up to five) referral coupons to their friends. The friends that return to the study site with a coupon form the first wave of the RDS. The process iterates until the procedure reaches the target sample size, or until the process dies because participants stop passing coupons. Figure 1 gives an illustration of this process.

If we presume that each participant refers a random subset of their friends,[2] then RDS is a stochastic process on the members of the social network. In the RDS literature, it is common to assume a Markov model because it is analytically tractable. The Markovian assumption is knowingly incorrect in practice. For example, it samples with-replacement; in practice, the sampling is performed without-replacement. Simulation studies suggest that, when the sample size is much smaller than the population size, the Markov model provides an approximation to more accurate simulation models [Lu et al. (2012)]. Under the Markov model, Salganik and Heckathorn (2004) and Volz and Heckathorn (2008) construct unbiased estimators. While they are unbiased, they often suffer from high variability [Goel and Salganik (2009, 2010)]. In particular, Goel and Salganik (2010)

---

[2]In current implementations of RDS, randomization is not produced by researchers. Rather, it is presumed that people refer friends randomly. The validity of such assumptions has been studied in several several ways in empirical and statistical papers. For example, Gile, Johnston and Salganik (2015) proposed statistical diagnostics to examine the convergence properties, Arayasirikul, Cai and Wilson (2015) performed qualitative follow-up interviews to ask participants about difficulties in finding referrals and McCreesh et al. (2012) compared a respondent-driven sample in Uganda with a total population survey on the same population.

shows in a wide range of computer experiments that (1) RDS often produces estimators with exceedingly large variance and (2) the popular bootstrap technique in Salganik (2006) produced nominal 95% confidence intervals with coverage probabilities between 40% and 70%. This paper aims to build on these earlier results to provide a rigorous description of the inadequacies.

This paper focuses on one particular assumption of the Markov chain model which has received insufficient scrutiny. In practice, each participant can refer between zero and three (sometimes up to five) future participants. However, in the Markov chain model, each participant refers exactly one individual. In the previous simulation study of Goel and Salganik (2010), the "chain" assumption was relaxed, while the "Markov" assumption was retained. This model has drastically different behaviors. The results below show that this "Markov tree" model remains analytically tractable.

The paper is organized as follows. Section 1 defines the Markov model, the quantity we wish to estimate, and the estimators. Section 2 provides an exact formula for the variance of an RDS estimator in Theorem 2.1. Section 3 specifies the asymptotic behavior of the design effect in Theorem 3.1. Section 4 studies the rate at which the Markov model resamples nodes in Theorem 4.1. With Theorems 2.1 and 3.1, Section 5 reinterprets the previous simulation results in Goel and Salganik (2010). Section 6 proposes a novel resampling technique A-TREE-BOOTSTRAP and compares it to previous techniques in computational experiments. Finally, Section 7 concludes the paper. All proofs are contained in the Appendix and Supplementary Material [Rohe (2019)].

1. Preliminaries. The model described below is a straightforward combination of the Markov models developed in the previous literature [e.g., Heckathorn (1997), Salganik and Heckathorn (2004), Volz and Heckathorn (2008) and Goel and Salganik (2009)]. There are four necessary mathematical pieces: a social network represented as a graph, a Markov transition matrix on the nodes of the graph, a referral tree to index the Markov process on the graph, and finally, a node feature defined for each node in the graph.

1.1. *Markov processes on a graph.* A social network $G = (V, E)$ consists of the set of people $V = \{1, \ldots, N\}$ and the set of friendships $E = \{(i, j) : i \text{ and } j \text{ are friends}\}$. V is referred to as the node set and E is referred to as the edge set. The results in this paper allow for a weighted graph. Let w_{ij} be the weight of the edge $(i, j) \in E$; if $(i, j) \notin E$, define $w_{ij} = 0$. If the graph is unweighted, then let $w_{ij} = 1$ for all $(i, j) \in E$. Throughout this paper, the graph is undirected, $w_{ij} = w_{ji}$; 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 $\text{vol}(G) = \sum_i \deg(i)$. If the graph is unweighted, $\deg(i)$ is the number of connections to node i. To simplify notation, $i \in G$ is used synonymously with $i \in V$.

1.1.1. *Markov chain on G.* Denote $X(0), X(1), X(2), \ldots \in G$ as a Markov chain on the individuals from the social network G. The transition matrix $P \in \mathbb{R}^{N \times N}$ is defined so that transition probabilities are proportional to edge weights,

$$P_{ij} = \mathbb{P}(X(t+1) = j | X(t) = i) = \frac{w_{ij}}{\deg(i)}.$$

Let $|\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_N|$ denote the eigenvalues of P. All eigenvalues of P are less than or equal to one in absolute value [see, e.g., Lemma 12.1 in Levin, Peres and Wilmer (2009)]. Because the edge weights are symmetric, $w_{ij} = w_{ji}$ for all i, j, the Markov chain is *reversible*. If $|\lambda_2| < 1$, then the stationary distribution $\pi : G \to \mathbb{R}$ is

$$\pi_j = \lim_{t \to \infty} \mathbb{P}(X(t) = j | X(0) = i) = \frac{\deg(j)}{\text{vol}(G)} \qquad \text{for all } i, j \in G.$$

1.1.2. *Markov process on G indexed by a tree.* Let $\mathbb{T}$ be a rooted tree—a connected graph with n nodes, no cycles, and a vertex 0. The seed participant is vertex 0 in $\mathbb{T}$ (cf. Figure 1). Note that the node set of G indexes the population and the node set of $\mathbb{T}$ indexes the sample. To simplify notation, $\sigma \in \mathbb{T}$ is used synonymously with σ belonging to the vertex set of $\mathbb{T}$. For any node in the tree $\sigma \in \mathbb{T}$, denote $\sigma' \in \mathbb{T}$ as the parent of σ (the node one step closer to the root). Let $\mathscr{D}(\sigma) \subset \mathbb{T}$ denote the set of σ and all its descendants in $\mathbb{T}$. Denote the height of $\mathbb{T}$ as $h(\mathbb{T})$; this is the number of rounds of sampling in the RDS, or the maximum graph distance in $\mathbb{T}$ from the root to any node.

A Markov process indexed by $\mathbb{T}$ is a set of random variables $\{X_\sigma : \sigma \in \mathbb{T}\}$ satisfying the Markov property

$$\mathbb{P}(X_\sigma | X_{\sigma'}, X_\tau : \tau \in \mathscr{D}(\sigma)^c) = \mathbb{P}(X_\sigma | X_{\sigma'}).$$

The transition matrix $P \in [0, 1]^{N \times N}$ describes these transition probabilities:

$$\mathbb{P}(X_\sigma = j | X_{\sigma'} = i) = P_{ij} \qquad \text{for } i, j \in G.$$

Benjamini and Peres (1994) called this process a $(\mathbb{T}, P)$-*walk on G*. Unless stated otherwise, it will be presumed throughout that under the $(\mathbb{T}, P)$-*walk on G*, X_0 is initialized from the stationary distribution of P.

For example, if $\mathbb{C}$ is the chain graph, then the $(\mathbb{C}, P)$-*walk* on G is a Markov chain on G, $X(0), X(1), X(2), \ldots \in G$. One key property of the Markov model is that it allows for resampling. Said another way, it "samples with-replacement" because it is possible for $X(i) = X(j)$ for $i \neq j$. The same is true in the tree model. In particular, it is possible for $X_\tau = X_\sigma$ for $\tau, \sigma \in \mathbb{T}$ with $\tau \neq \sigma$.

1.2. *Measurements and estimators.* For each node $i \in G$, let $y(i) \in \mathbb{R}$ denote some characteristic of this node. We wish to estimate the population average

$$\mu_{\text{true}} = \frac{1}{N} \sum_{i \in G} y(i).$$

In the motivating RDS example, $y(i) = 1$ denotes that $i \in G$ is HIV+, $y(i) = 0$ denotes that $i \in G$ is HIV−, and μ_{true} is the proportion of the population that is HIV+. We estimate μ_{true} with observations

$$Y_\tau = y(X_\tau) \qquad \text{for } \tau \in \mathbb{T},$$

where X_τ is a $(\mathbb{T}, P)$-*walk on* G. Denote

$$\mu = \mathbb{E}_{\text{RDS}}(Y_0) = \sum_i y(i) \pi_i,$$

where the subscript RDS denotes that the expectation is computed with the $(\mathbb{T}, P)$-*walk on* G. In general, $\mu \neq \mu_{\text{true}}$. The sample average

$$(1) \qquad \hat{\mu} = \frac{1}{n} \sum_{\tau \in \mathbb{T}} Y_\tau$$

is an unbiased estimate of μ. Theorem 2.1 below gives a closed-form expression for $\text{Var}_{\text{RDS}}(\hat{\mu})$.

With $\pi_i = \deg(i)/\text{vol}(G)$, the inverse probability weighted estimator (IPW),

$$\hat{\mu}_{\text{IPW}} = \frac{1}{n} \sum_{\tau \in \mathbb{T}} \frac{Y_\tau}{\pi_{X_\tau} N} = \frac{\text{vol}(G)}{N} \frac{1}{n} \sum_{\tau \in \mathbb{T}} \frac{Y_\tau}{\deg(X_\tau)},$$

is an unbiased estimator for μ_{true}. The results in this paper can be applied to $\hat{\mu}_{\text{IPW}}$ via a transformation that is described in the next remark. Computing the IPW estimator requires $\text{vol}(G)$ or the average node degree $\text{vol}(G)/N$. This is typically not available in practice. When the sampling weights can be identified up to a constant of proportionality [i.e., $\pi_i \propto \deg(i)$], estimating $\text{vol}(G)/N$ with the harmonic mean of the observed node degrees,

$$H = \left(n^{-1} \sum_{\tau \in \mathbb{T}} 1/\deg(X_\tau) \right)^{-1},$$

leads the Hajek or Volz–Heckathorn estimator [Volz and Heckathorn (2008)],

$$\hat{\mu}_{\text{VH}} = H \frac{1}{n} \sum_{\tau \in \mathbb{T}} \frac{Y_\tau}{\deg(X_\tau)}.$$

REMARK 1.1. Define a new node feature

$$y^\pi(i) = \frac{y(i)}{\pi_i N}$$

and new node measurements $Y_\tau^\pi = y^\pi(X_\tau)$. The sample average of the Y_τ^π's is exactly the IPW estimator using the nontransformed Y_τ's. Because of this simple transformation, the theorems below that study $\hat{\mu}$ can also study $\hat{\mu}_{\text{IPW}}$ by substituting y^π for y.

Define $W_1, \ldots, W_n \in G$ as independent random samples with $\mathbb{P}(W_i = j) = \pi_j$. Define

$$(2) \qquad \mathrm{Var}_\pi(\hat{\mu}) = \mathrm{Var}\left(\frac{1}{n} \sum_{i=1}^{n} y(W_i)\right).$$

Define the *design effect* of the $(\mathbb{T}, P)$-*walk on G* as

$$(3) \qquad \mathrm{DE}(\hat{\mu}) = \frac{\mathrm{Var}_{\mathrm{RDS}}(\hat{\mu})}{\mathrm{Var}_\pi(\hat{\mu})}.$$

The standard definition of DE contains the variance under simple random sampling (SRS) in the denominator. For simplicity, the DE in this paper contains Var_π in the denominator. The key difficulty of comparing SRS to the $(\mathbb{T}, P)$-*walk on G* is that SRS is without-replacement. Instead of SRS, the denominator in equation (3) could be replaced by the variance under uniform sampling (with-replacement) and this would only change the DE by a constant factor. This is because G and π do not change with n.

The standard O-notation is used below. In particular, $h(n) = o(g(n))$ means that $h(n)/g(n) \to 0$ as $n \to \infty$ and $h(n) = O(g(n))$ means that $h(n) \le M g(n)$ for all n, for some constant M.

2. The variance under RDS.
The key result of this section, Theorem 2.1, expresses $\mathrm{Var}_{\mathrm{RDS}}(\hat{\mu})$ as a function of the eigenproperties of P. The following lemma from Levin, Peres and Wilmer (2009) provides the eigendecomposition of the matrix P.

LEMMA 2.1 [Lemma 12.2 in Levin, Peres and Wilmer (2009)]. *Let P be a reversible Markov transition matrix on the nodes in G with respect to the stationary distribution π. The eigenvectors of P, denoted as $f_1, \ldots, f_N$, are real valued functions of the nodes $i \in G$ and orthonormal with respect to the inner product*

$$(4) \qquad \langle f_a, f_b \rangle_\pi = \sum_{i \in G} f_a(i) f_b(i) \pi_i.$$

If λ is an eigenvalue of P, then $|\lambda| \le 1$. The eigenfunction f_1 corresponding to the eigenvalue 1 can be taken to be the constant vector $\mathbf{1}$.

The structure of the tree $\mathbb{T}$ is a key component in the variance. That structure is summarized in the following definition of $\mathbb{G} : [-1, 1] \to \mathbb{R}$, which is a functional of $\mathbb{T}$. All of the statements in this section are conditional on the tree.

DEFINITION 1. Select two nodes $I, J \in \mathbb{T}$ uniformly and independently. Define $D = d(I, J)$ to be the graph distance between I and J in $\mathbb{T}$. Define $\mathbb{G}$ as the probability generating function for D,

$$\mathbb{G}(\lambda) = \mathbb{E}(\lambda^D).$$
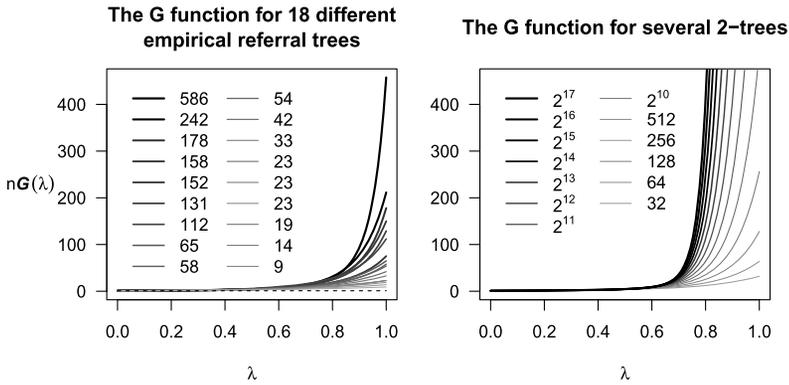
FIG. 2. *Each line corresponds to a referral tree. The vertical axis gives $n\mathbb{G}(\lambda)$. The legend gives the number of nodes in each tree. In the left panel, there are eighteen different referral trees from published RDS studies. The tree of 586 comes from a study of drug users in New York City [Abdul-Quader et al. (2006)]. The tree of 112 comes from a study of injection drug users in Connecticut [Heckathorn (1997)]. The trees of 14, 19, 23, 23, 65 and 152 come from a study of men who have sex with men in Higuey, Dominican Republic [Gile, Johnston and Salganik (2015)]. The remaining ten trees come from a study of 25 villages in rural Uganda [McCreesh et al. (2012)]. In the right panel, each line represents a 2-tree, where each node creates two nodes in the next wave. The results of the next section are foreshadowed by the critical threshold at $\lambda = 1/\sqrt{2} \approx 0.7$.*

Note that because the tree $\mathbb{T}$ is observed, the function $\mathbb{G}$ can be computed in practice. Figure 2 gives an illustration of $n\mathbb{G}(\lambda)$ for $\lambda \in [0, 1]$, where n is the number of nodes in $\mathbb{T}$.

THEOREM 2.1. *Suppose that the Markov transition matrix P is reversible with respect to π and that the second eigenvalue of P is less than one in absolute value, then*

$$(5) \qquad \mathrm{Var}_{\mathrm{RDS}}(\hat{\mu}) = \sum_{\ell=2}^{N} \langle y, f_\ell \rangle_\pi^2 \mathbb{G}(\lambda_\ell),$$

where the subscript RDS denotes that data have been collected through a $(\mathbb{T}, P)$-walk on G, $\hat{\mu}$ is defined in equation (1), $\langle \cdot, \cdot \rangle_\pi$ is defined in equation (4), $f_1, \ldots, f_N : G \to \mathbb{R}$ are the eigenvectors of P corresponding to eigenvalues $\lambda_1 > |\lambda_2| \geq \cdots \geq |\lambda_N|$ and $\mathbb{G}$ is defined in Definition 1.

In previous research, Verdery et al. (2015) and Khabbazian et al. (2017) prove this theorem for the special case that $\mathbb{T}$ is a chain. The first step to prove Theorem 2.1 is to show that if $d(\sigma, \tau) = t$, then by the reversibility of P,

$$(X_\sigma, X_\tau) \overset{d}{=} (X(0), X(t)),$$

where $X(0), \ldots, X(t) \in G$ is a Markov chain with the same transition matrix P. Then, expanding y in the eigenbasis from Lemma 2.1,

$$(6) \qquad \mathrm{Cov}_{\mathrm{RDS}}(Y_\sigma, Y_\tau) = \sum_{\ell=2}^{N} \lambda_\ell^{d(\sigma,\tau)} \langle y, f_\ell \rangle_\pi^2.$$

Averaging over σ, τ and exchanging summations yields $\mathbb{G}$ and the final result. Section A in the Appendix contains a full proof.

REMARK 2.1. Using Remark 1.1, Theorem 2.1 also gives the variance for $\hat{\mu}_{\mathrm{IPW}}$. This theorem presumes that X_0 (i.e., the seed node) is sampled from the stationary distribution. Under this assumption, $\hat{\mu}_{\mathrm{IPW}}$ is unbiased for finite samples. However, conditionally on the seed node, $\hat{\mu}_{\mathrm{IPW}}$ and $\hat{\mu}_{\mathrm{VH}}$ are biased [Gile and Handcock (2010)]. The law of total variance shows how $\mathrm{Var}_{\mathrm{RDS}}(\hat{\mu}_{\mathrm{IPW}})$ includes the seed-bias, defined as $\mathrm{bias}(\hat{\mu}_{\mathrm{IPW}}, X_0) = \mathbb{E}_{\mathrm{RDS}}(\hat{\mu}_{\mathrm{IPW}}|X_0) - \mu_{\mathrm{true}}$,

$$\mathrm{Var}_{\mathrm{RDS}}(\hat{\mu}_{\mathrm{IPW}}) = \mathbb{E}_\pi\big(\mathrm{Var}_{\mathrm{RDS}}(\hat{\mu}_{\mathrm{IPW}}|X_0)\big) + \mathbb{E}_\pi\big(\mathrm{bias}(\hat{\mu}_{\mathrm{IPW}}, X_0)\big)^2,$$

where $\mathbb{E}_\pi$ is the expectation with respect to X_0 having distribution π.

The law of total variance shows that that conditioning on the seed node decreases the variance by the squared seed-bias. As such, any technique which attempts to estimate the uncertainty (e.g., bootstrap procedures) should ensure that the seed node is allowed to be random. The bootstrap procedure proposed in Section 6.1 proposes a way to randomize the seed node(s).

3. The asymptotic behavior of the design effect. The eigenproperties of P have been extensively studied in the literature on spectral graph theory and spectral clustering [Chung (1997), von Luxburg (2007)]. Cheeger's inequality shows that if λ_2 is close to one, then there are clusters or communities in the graph. For RDS, this creates a "referral bottleneck" where the referral process has difficulty mixing between the two communities. For example, if $\lambda_2 = 1$, then the social network is disconnected; this represents an extreme bottleneck, where the referral process will never cross the divide. This section shows that the asymptotic behavior of DE depends upon the relationship between λ_2 and the growth rate of the referral tree $\mathbb{T}$.

To study how $\mathrm{Var}_{\mathrm{RDS}}$ and DE behave as the sample size increases, it is necessary to describe how the referral tree $\mathbb{T}$ grows. Theorem 3.1 grows a random Galton–Watson tree. A Galton–Watson tree is initialized with a single root node and is parameterized by its *offspring distribution*. Starting with the root node and iterating through all future generations, each node generates a random number of offspring, drawn from the offspring distribution. The number of offspring produced by each node is independent across nodes. This process is highly studied with several well-known results [e.g., Athreya and Ney (1972)].

Let ξ be a generic draw from the offspring distribution and denote $\mathbb{E}\xi = m$. To have a positive probability that the tree generates an infinite number of nodes, the results below require that $m > 1$. Denote $\mathbb{T}_h$ as the subtree of $\mathbb{T}$ that includes all nodes within distance h of the root.

THEOREM 3.1. *Suppose $\mathbb{T}$ is a random Galton–Watson tree. Let ξ be a single draw for the offspring distribution with $m = \mathbb{E}(\xi) > 1$ and $\mathbb{E}(\xi^4) < \infty$. Condition on the survival of the Galton–Watson process. Define $\mathbb{T}_h$ as the node induced subgraph of $\mathbb{T}$ that contains all nodes $\tau \in \mathbb{T}$ within distance h from the root node. Let P be a Markov transition matrix on G that is reversible with respect to its stationary distribution π. Let $\hat{\mu}_h$ be constructed with the samples from a $(\mathbb{T}_h, P)$-walk on G. If $\mathrm{Var}_\pi Y_0 > 0$, $\langle y, f_2 \rangle_\pi^2 > 0$, and $\lambda_2 > 0$, then*

$$(7) \qquad \mathrm{DE}(\hat{\mu}_h) \asymp \begin{cases} c & \text{if } m \leq \beta, \\ n^{1-\alpha} & \text{if } m > \beta, \end{cases}$$

where DE *is defined in equation* (3) *conditionally on* $\mathbb{T}$, $\asymp$ *is equality up to* $(\log n)^2$ *terms,* $\beta = \lambda_2^{-2}$ *and* $\alpha = \log_m \lambda_2^{-2}$.

The proof of this result has four pieces, divided into four subsections of Section 1 in the Supplementary Material [Rohe (2019)]. Section 1.1 shows that DE behaves asymptotically similar to $n\mathbb{G}(\lambda_2)$. Then Section 1.2 gives a lower bound for $\mathbb{G}(\lambda_2)$ that depends only on the growth rate of the tree $\mathbb{T}$. Section 1.3 gives an upper bound for $\mathbb{G}(\lambda_2)$ that requires a "balanced assumption" on $\mathbb{T}$. These three subsections do not require that $\mathbb{T}$ comes from the Galton–Watson distribution. Then, in Section 1.4 the Kesten–Stigum theorem shows that when $\mathbb{T}$ comes from the Galton–Watson distribution, it grows at rate m (satisfying the lower bounds in Section 1.2). Then Lemma 1.3 applies the L^p maximal inequality for martingales to the Galton–Watson martingale to show that Galton–Watson trees with $\mathbb{E}\xi^4 < \infty$ satisfy the "balanced assumption."

The assumption that $\mathbb{E}(\xi^4) < \infty$ is a strong assumption in the literature on the Galton–Watson process. However, there are two important points. First, in the context of RDS, the offspring distribution is typically bounded by three or five. As such, this condition is certainly satisfied. Second, the finite fourth moment is only needed for the *upper* bound; it implies the "balanced assumption" on $\mathbb{T}$. To motivate the necessity of the "balanced assumption," the beginning of Section 1.3 in the Supplementary Material [Rohe (2019)] gives a deterministic tree for which the upper bounds will not hold (i.e., RDS has much larger DE).

To see why there is a critical threshold, note that $\mathrm{Var}_{\mathrm{RDS}}(\hat{\mu})$ is the average of the covariances $\mathrm{Cov}_{\mathrm{RDS}}(Y_\sigma, Y_\tau)$. From equation (6), each covariance term decays exponentially, $O(\lambda_2^{d(\sigma,\tau)})$, where $d(\sigma, \tau)$ is the graph distance between σ and τ in $\mathbb{T}$. However, these graph distances grow logarithmically; when $m > 1$, $d(\sigma, \tau) =$

$O(\log_m n)$. For example, if $\mathbb{T}$ is a complete m-tree with n nodes, $h(\mathbb{T}) \leq \log_m n$ implies $d(\sigma, \tau) \leq 2h(\mathbb{T}) \leq 2\log_m n$. Using these bounds,

$$\lambda_2^{d(\sigma,\tau)} \geq \lambda_2^{2\log_m n} = n^{2\log_m \lambda_2}.$$

Below the critical threshold, this term is $o(n^{-1})$ and $\mathrm{Var}_{\mathrm{RDS}}(\hat{\mu})$ is controlled by the terms $\sigma = \tau$ (i.e., the diagonal terms of the covariance matrix). Above the critical threshold, the upper bound in the Supplementary Material confirms that $n^{2\log_m \lambda_2}$ is the rate that $\mathrm{Var}_{\mathrm{RDS}}(\hat{\mu})$ decays to zero. For more details, see the Supplementary Material [Rohe (2019)].

REMARK 3.1. In practice, it is possible to estimate where a given sample falls with respect to the critical threshold; first, a warning. The empirical mean of the offspring distribution for any finite tree $\mathbb{T}$ is always less than one. As such, this cannot be used as an estimator for m. Instead, use the function $\mathbb{G}(\lambda)$ for $\lambda \in [0, 1]$ which can be computed from the graph topology of $\mathbb{T}$. Figure 2 gives an illustration for several empirical and synthetic trees. Each line of Figure 2 corresponds to a single referral tree. In these figures, one can see that at some value of λ, $\mathbb{G}(\lambda)$ begins to grow at a fast rate. One must then compute $\hat{\lambda}_2$, an estimate of λ_2, the second largest eigenvalue of P (this requires statistical inference). By evaluating $\mathbb{G}(\hat{\lambda}_2)$, one can gauge where a given sample falls with respect to the critical threshold. A follow-up paper provides an approach to estimate the largest eigenvalues of the matrix P from a $(\mathbb{T}, P)$-walk on G sample [Roch and Rohe (2017)]. The code is available on github (https://github.com/karlrohe/mRDS).

4. The gap between sampling with and without-replacement.
Define the number of repeated pairs as

$$R_n = \big|\{\sigma, \tau \in \mathbb{T} | \tau \neq \sigma, X_\tau = X_\sigma\}\big|.$$

This section studies $\mathbb{E}_{\mathrm{RDS}}(R_n)$ as n and N grow in tandem. Because R_n counts *pairs* of repeats, $\mathbb{E}(R_n)$ could grow at rate n^2. Proposition 4.1 and Theorem 4.1 show that if $n = o(\sqrt{N})$ and some additional assumptions, then $\mathbb{E}_{\mathrm{RDS}}(R_n) \asymp n$. In particular, this shows that the rate of resampling does not depend on λ_2.

PROPOSITION 4.1. *Under the $(\mathbb{T}, P)$-walk on G, suppose that G is undirected and P is a simple random walk. If $\deg(i) < D$ for all nodes $i \in G$, then*

$$\mathbb{E}(R_n) \geq n/D.$$

The proof is based on the fact that if $X_\sigma = i$, the probability of transitioning back to the state of $X_{\sigma'}$ is $1/\deg(i) \geq 1/D$. The full proof is contained in Section 2 in the Supplementary Material [Rohe (2019)].

As Proposition 4.1 shows, the $(\mathbb{T}, P)$-walk on G can have several repeated samples. However, this alone does not prevent the variance from decaying at rate $1/n$;

the decay of the variance is determined by the critical threshold, $m > \lambda_2^{-2}$. The next result gives a matching upper bound for $\mathbb{E}(R_n)$. This shows that the rate of $\mathbb{E}(R_n)$ does not depend on the critical threshold.

THEOREM 4.1.   *Consider a sequence of samples $\{X_\tau : \tau \in \mathbb{T}_n\}$ that are sampled from a $(\mathbb{T}_n, P_N)$-walk on G_N, where n and N are both growing. Suppose that the sequence $\mathbb{T}_n$ satisfies the conditions of Theorem 1.1 in the Supplementary Material; that is, there is a balanced infinite tree $\mathbb{T}$ that grows at rate m and $\mathbb{T}_n$ is a sequence of subtrees that successively add one generation at a time.*

If (1) the stationary distribution is bounded, $\pi_i \leq c/N$ for all i and all N; (2) the number of eigenvalues λ_ℓ that exceed the critical threshold $1/\sqrt{m}$ is bounded by k for all N; and (3) $n = o(\sqrt{N})$, then

$$\mathbb{E}(R_n) = O\big((\log n)n\big).$$

Notice that condition (1) is implied by the bounded degree assumption in Proposition 4.1. Importantly, the rate of this upper bound does not depend on λ_2. So, under the conditions of these results, λ_2 and the critical threshold do not effect the rate of $\mathbb{E}(R_n)$.

The key to proving Theorem 4.1 is the relationship between the trace of a matrix and its eigenvalues. First, notice that

(8) $$\mathbb{E}(R_n) = \sum_{\sigma \neq \tau} \mathbb{P}(X_\sigma = X_\tau).$$

Let $\mathrm{tr}(P)$ denote the trace of P:

$$\mathbb{P}(X_\sigma = X_\tau) = \sum_{i \in G} \pi_i \mathbb{P}(X_\tau = i | X_\sigma = i) = \sum_{i \in G} \pi_i P_{ii}^{d(\sigma, \tau)} \leq cN^{-1} \mathrm{tr}\big(P^{d(\sigma, \tau)}\big)$$

$$= cN^{-1} \sum_\ell \lambda_\ell^{d(\sigma, \tau)}.$$

To bound $\mathbb{E}(R_n)$, exchange the summation over $\sigma \neq \tau$ from equation (8) with the summation over ℓ in the line above. Each term in the resulting summation can be expressed with $\mathbb{G}$ functions and bounded by Theorem 1.1 in the supplement. The full proof is contained in Section 2 of the Supplementary Material [Rohe (2019)].

4.1. *Comparison to a more realistic model with simulation.*   For mathematical tractability, the theorems above make two simplifications. First, the theorems use the $(\mathbb{T}, P)$-*walk on* G, which samples with-replacement. Second, the theorems study the IPW estimator. The simulations in this section (and in the rest of the paper) use a more realistic setting. First, the simulated samples are collected without-replacement. Second, the simulations study the Volz–Heckathorn estimator. These

With and without replacement sampling display a similar critical threshold
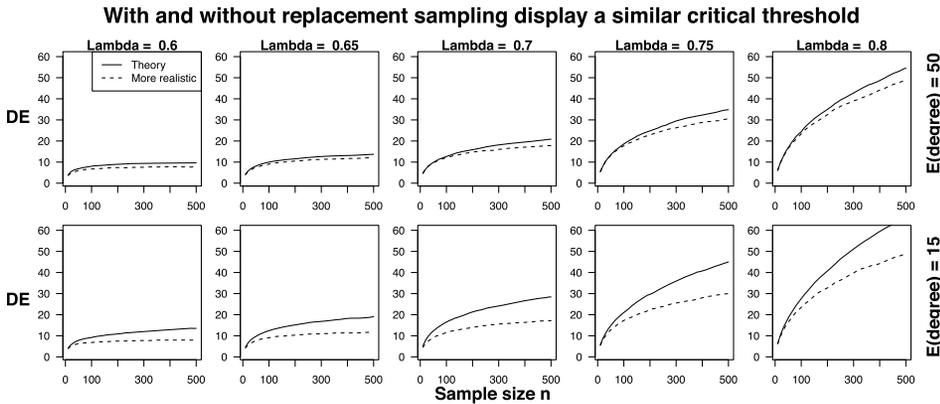
FIG. 3. *In all figures, $m = 2$. Each column of panels corresponds to a different value of λ_2, from left to right, $\lambda_2 \in (0.6, 0.65, 0.7, 0.75, 0.8)$. In the panels on the left, the lines are roughly flat. In the panels on the right, the lines are quickly increasing. This shows that the $(\mathbb{T}, P)$-walk on G and the more realistic model have a critical threshold somewhere between $\lambda_2(\mathscr{P}) = 0.6$ and $\lambda_2(\mathscr{P}) = 0.8$.*

simulation results find that the Markov model with the IPW estimator is a good approximation to the more realistic model, so long as the number of sampled nodes is much smaller than the population size, as predicted by Theorem 4.1.

The simulations are performed on networks simulated from the Stochastic Blockmodel. The ten panels in Figure 3 correspond to ten different model settings. Each of the ten models has $N = 10k$ nodes, equally balanced between group zero and group one. This simulation is repeated in Section B for unbalanced group sizes.

The probability of a connection between two nodes in different blocks is r and the probability of connection between two nodes in the same block is p. Figure 3 parameterizes this model via (1) the expected degree $(p + r)N/2$ and (2) the second eigenvalue of $\mathscr{P} = \mathbb{E}(D)^{-1}\mathbb{E}(A)$,

$$(9) \qquad \lambda_2(\mathscr{P}) = \frac{p - r}{p + r},$$

where expectations are under the Stochastic Blockmodel [cf. example on page 1893 of Rohe, Chatterjee and Yu (2011)]. In group zero, $y_i = 0$ and in group one, $y_i = 1$. The horizontal axis in each plot represents the sample size; the vertical axis represents the design effect (as estimated via simulation). The five columns of plots correspond to five different values of $\lambda_2(\mathscr{P})$.

To simulate from the more realistic model, the simulation first generates $\mathbb{T}$ as a Galton–Watson tree with offspring distribution $1 + \text{Binomial}(2, 1/2)$. A tree is grown until it reaches 2000 nodes; while only 500 samples are kept, it will become clear why $\mathbb{T}$ must be initialized to be larger than 500. This tree is seeded with a participant selected from the stationary distribution. Then each participant randomly

selects their referrals from their "viable" friend list without-replacement; a friend is viable if it has not yet appeared in the sample. One participant at a time makes all of their referrals, iterating through the tree in the fashion of a breadth first search. A difficulty arises if $\sigma \in \mathbb{T}$ should produce three referrals, but X_σ does not have that many viable friends. When this happens in the simulation, all viable friends are referred and the remaining descendants in $\mathbb{T}$ are removed; this happens infrequently in the simulation. Once this process samples 500 nodes, the remaining nodes in $\mathbb{T}$ are pruned. This pruned tree is then used to run the $(\mathbb{T}, P)$-*walk on* G. For each of the ten networks, this process is simulated 1000 times. The sample variance across these 1000 samples is divided by the variance of uniform with-replacement sampling, $(4n)^{-1}$.

Because the trees are simulated to have $m = 2$, Theorem 3.1 suggests that the design effect grows when λ_2 exceeds $1/\sqrt{2} \approx 0.7$. In the left most plots, the solid lines are roughly flat. In the right most plots, the solid lines are quickly increasing. This shows that the $(\mathbb{T}, P)$-*walk on* G has a critical threshold somewhere between 0.6 and 0.8; this is consistent with the theory. Similarly, the dashed lines are roughly flat in the left plots and quickly increasing in the right plots. Under these simulation settings, the more realistic model mimics the critical threshold behavior identified in Theorem 3.1.

In the first row of plots, each node has an expected degree of fifty. In the second row of plots, each node has an expected degree of fifteen. In the top row, the solid and dashed lines are close because there are fewer repeated samples. In the bottom row, the lines for the sparse graphs are not as close. However, both rows display the same qualitative behavior (flat when $\lambda_2 = 0.6$ and increasing when $\lambda_2 = 0.8$). This simulation is repeated in the Appendix (Section B) with a Stochastic Blockmodel that has unbalanced block sizes. Under the unbalanced blocks, Figure 6 displays the same critical threshold as Figure 3 above.

5. Reinterpreting the results of Goel and Salganik (2010) with Theorem 3.1. One of the most highly cited bootstrap procedures in the previous literature was proposed in Salganik (2006) and is often referred to as the Salganik bootstrap. Later, Goel and Salganik (2010) showed in simulation experiments that this procedure produces "misleadingly narrow" confidence intervals. This section reinterprets those simulation results using Theorems 2.1 and 3.1 above. This reinterpretation motivates an alternative bootstrap procedure which is explored in the next section.

In the simulation study, Goel and Salganik (2010) used several different graphs G that were collected in previous empirical social network research. In each of several experiments, y is a demographic measurement such as race or gender. Given G and y, Goel and Salganik (2010) simulated the respondent-driven sample with a $(\mathbb{T}, P)$-*walk on* G, where $\mathbb{T}$ is a Galton–Watson tree with $m = 1.5$. After collecting a sample of $n = 500$, Goel and Salganik (2010) constructed a bootstrapped confidence interval with the Salganik bootstrap [Salganik (2006)]. To resample the

observed individuals, the Salganik bootstrap constructs a Markov transition matrix $\hat{P}_o \in \mathbb{R}^{n \times n}$ on the observed individuals as follows:

> [D]ivide the sample members into two sets based on how they were recruited: people recruited by someone in group A (which we will call A_{rec}) and people recruited by someone in group B (which we will call B_{rec}). For example, A_{rec} could be the set of all sample members who were recruited by someone with HIV. ... [B]ased on the group membership [of the current state], we draw with-replacement from either A_{rec} or B_{rec} [Salganik (2006)].

The fundamental problem with the Salganik bootstrap is that each bootstrap sample is a $(\mathbb{C}, \hat{P}_o)$-*walk* on the observed individuals, where $\mathbb{C}$ is a chain graph. By using $\mathbb{C}$ instead of $\mathbb{T}$, *the Salganik bootstrap resampling distribution is a Markov chain, not a "Markov tree."*

In the simulation results of Goel and Salganik (2010), the bootstrap has particularly poor coverage on a subset of the features. These features are correlated with the underlying social network. In particular, if there is an eigenpair (λ_ℓ, f_ℓ) of P where $\langle y, f_\ell \rangle_\pi^2$ is large and $1.5 > 1/\lambda_\ell^2$, then the $(\mathbb{T}, P)$-*walk on G* exceeds the critical threshold, while the $(\mathbb{C}, \hat{P}_o)$-*walk* does not. If the original sample $(\mathbb{T}, P)$-*walk on G* exceeds the critical threshold, then estimates derived from this sample will be highly variable. However, because the $(\mathbb{C}, \hat{P}_o)$-*walk* resamples with a chain graph $\mathbb{C}$, it has $m = 1$. As such, the $(\mathbb{C}, \hat{P}_o)$-*walk* will never exceed the critical threshold. The confidence intervals from the Salganik bootstrap will contract at rate $O(n^{-1/2})$, while the true uncertainty is decaying at a slower rate. This leads to confidence intervals which are too narrow.

6. Bootstrap resampling with $\mathbb{T}$.

To allow for the bootstrap distribution to exceed the critical threshold, this section proposes A-TREE-BOOTSTRAP. The Salganik bootstrap will be referred to as A-CHAIN-BOOTSTRAP. The A- prefix stands for *assisted*, because they are both assisted by some node feature to create a Markov transition matrix. In the A-CHAIN-BOOTSTRAP, the construction of the matrix $\hat{P}_o$ is assisted by the outcome of interest y (via the sets A_{rec}, B_{rec}). The A-TREE-BOOTSTRAP also constructs a Markov transition matrix on the observed individuals, $\hat{P} \in \mathbb{R}^{n \times n}$, and the construction of this matrix is assisted by some node features. However, unlike A-CHAIN-BOOTSTRAP, A-TREE-BOOTSTRAP does not require that $\hat{P}$ is constructed from the same variable as the outcome of interest y. As described in the previous section, the A-CHAIN-BOOTSTRAP directly samples from the $(\mathbb{C}, \hat{P}_o)$-*walk*. Similarly, the A-TREE-BOOTSTRAP directly samples from the $(\mathbb{T}, \hat{P})$-*walk*, where the construction of $\hat{P}$ is described in the next subsection. R code for A-TREE-BOOTSTRAP is available at https://github.com/karlrohe/mRDS.

Over the course of this research, Baraff, McCormick and Raftery (2016) proposed another bootstrap procedure which also uses $\mathbb{T}$ to perform the resampling. This procedure will be referred to as U-TREE-BOOTSTRAP. The U- prefix stands for *unassisted* because it does not require any node features to construct its Markov

transition matrix. In particular, the U-TREE-BOOTSTRAP resamples in a way that the topology of the referral tree changes between bootstrap samples. As such, it cannot be expressed as a $(\mathbb{T}, P)$-*walk on* G. However, it is still a Markov process with a transition matrix $\hat{P}_u$.

> …the initial step is to resample with-replacement from the seeds of the trees. Next, from each of those seeds, we resample with-replacement from their recruits, creating the second level of the bootstrap sample trees. From each of these new recruits, we then resample with-replacement from their recruits to create a third level. This process continues iteratively until no further recruits are available [Baraff, McCormick and Raftery (2016)].

To define $\hat{P}_u$ in the notation of this paper, let $A_{\mathbb{T}} \in \{0, 1\}^{n \times n}$ be the (asymmetric) adjacency matrix of the directed graph $\mathbb{T}$. So, for $\sigma \in \mathbb{T}$ with $\sigma \neq 0$, $[A_{\mathbb{T}}]_{\sigma', \sigma} = 1$. All other elements of $A_{\mathbb{T}}$ are zero. Define $D_{\mathbb{T}}$ as a diagonal matrix containing the number of referrals from σ in element σ, σ; $[D_{\mathbb{T}}]_{\sigma, \sigma} = \sum_{\tau} A_{\sigma, \tau}$. Note that if $\sigma \in \mathbb{T}$ is a leaf node, then $[D_{\mathbb{T}}]_{\sigma, \sigma} = 0$. The Markov transition matrix is $\hat{P}_u = D_{\mathbb{T}}^{-1} A_{\mathbb{T}}$, where $0/0$ is defined to be zero and the process terminates upon reaching a leaf node. This $\hat{P}_u$ is neither irreducible nor reversible.

6.1. *The* A-TREE-BOOTSTRAP *procedure.* This subsection describes the construction of $\hat{P}$ used in the A-TREE-BOOTSTRAP. Presume that every node in G belongs to a class, $z : V \to \{1, \ldots, K\}$, and $z(i)$ is observed if node i is sampled. These variables could denote some demographic characteristics or HIV status. The variables $\{z(X_\tau) : \tau \in \mathbb{T}\}$ assist the estimation of the Markov transition matrix on the n individuals in the original sample X_τ.

All probability statements in this section are conditional on the original sample. So, to temporarily conceal the randomness of the original sample, denote the observed individuals with lower-case letters, $\{x_\tau : \tau \in \mathbb{T}\}$. Recall that for any $\sigma \in \mathbb{T}$ with $\sigma \neq 0$, the parent node of σ is denoted as node $\sigma' \in \mathbb{T}$. Denote $N(u) = \sum_\sigma \mathbf{1}\{z(x_\sigma) = u\}$ as the number of nodes in class u. Define $\hat{A} : \{1, \ldots, K\}^2 \to \mathbb{R}$ to count the number of transitions between node types; for $u, v \in \{1, \ldots, K\}$,

$$(10) \qquad \hat{A}(u, v) = \sum_{\sigma \neq 0} \mathbf{1}\{z(x_{\sigma'}) = u, z(x_\sigma) = v\}.$$

Denote $\hat{D}(u)$ as the number of samples in class u that make a referral, $\hat{D}(u) = \sum_v \hat{A}(u, v)$.

If X_0^* and X_1^* represent one step of U-TREE-BOOTSTRAP, then X_0^* and X_1^* take values in the set of originally sampled individuals $\{x_\tau : \tau \in \mathbb{T}\}$. For any x_σ and x_τ in the original sample, define $u = z(x_\sigma)$ and $v = z(x_\tau)$. Then the probability of a transition from $x(\sigma)$ to $x(\tau)$ in U-TREE-BOOTSTRAP is defined to be

$$(11) \qquad \hat{P}_{x_\sigma, x_\tau} = \mathbb{P}(X_1^* = x_\tau | X_0^* = x_\sigma) = \frac{\hat{A}(u, v)}{\hat{D}(u)} \frac{1}{N(v)}.$$

This is equivalent to first taking a Markov transition from $z(x_\sigma)$ to some other node type v and then choosing an individual uniformly from the set of $N(v)$-many individuals of this type. Using the matrix $\hat{P}$, the resampling distribution of A-TREE-BOOTSTRAP is a $(\mathbb{T}, \hat{P})$-*walk* on $\{x_\tau : \tau \in \mathbb{T}\}$. Denote a resample as $\{X_\tau^* : \tau \in \mathbb{T}\}$; using these samples, construct $\hat{\mu}^*$ using $\{y(X_\tau^*) : \tau \in \mathbb{T}\}$ and any other measured features on the originally sampled individuals $\{x_\tau : \tau \in \mathbb{T}\}$ (e.g., their degree in G).

Recall from Remark 2.1 that it is necessary to randomize the seed node in a bootstrap procedure. To sample the seed node(s), A-TREE-BOOTSTRAP first samples a "mother node" uniformly at random from the original sample. Then this mother node refers all of the seed nodes in the bootstrap sample. The mother node only makes a difference when there are multiple seeds. It simulates the fact that some group is responsible for finding the seed nodes and this group is likely to constrained in their ability to select seeds. In U-TREE-BOOTSTRAP, the seed node(s) in the bootstrap match the seed nodes in the original sample. This makes the resampling unable to realize the uncertainty that comes from seed selection. By incorporating a mother node, A-TREE-BOOTSTRAP increases the dependence of the seeds and the variability of $\hat{\mu}^*$, thus making more conservative intervals. Section C further examines the sensitivity of A-TREE-BOOTSTRAP to using a "mother node."

The key ideas of A-TREE-BOOTSTRAP can also be used to perform sample size calculations. To do this, one must guess (i) K, (ii) for each $u, v \in 1, \ldots K$, the probability that someone in class u refers someone in class v, (iii) the proportion of individuals that belong to each class, (iv) the values of y within each class and (v) the topological structure of $\mathbb{T}$. R code for this is available at https://github.com/karlrohe/mRDS.

6.2. *Simulations to compare the bootstrap procedures.*

This section investigates the coverage properties of the confidence intervals generated from A-TREE-BOOTSTRAP, U-TREE-BOOTSTRAP, A-CHAIN-BOOTSTRAP and SS-BOOTSTRAP. The successive-sampling (SS) model was first described for RDS in Gile (2011). The SS-BOOTSTRAP fits and resamples from the SS model and was introduced in the R package RDS [Handcock, Fellows and Gile (2016)]. The SS model is not Markovian and so it cannot be described as a $(\mathbb{T}, P)$-*walk*. The SS-BOOTSTRAP requires an estimate of the population size; in the simulations below, the function was provided with the true value of the population size.

6.2.1. *Simulation settings.*

In total, the figures below change three aspects of the simulations settings: (i) the sample size of the RDS, (ii) the strength of the bottleneck in G, that is, λ_2 and (iii) the strength of the relationship between the outcome y and the bottleneck in G, that is, $\rho_\pi^2(y, f_2) = \sigma^{-1}\langle y, f_2 \rangle_\pi$, where $\sigma^2 = \text{Var}_{\text{RDS}} Y_0$. While the asymptotic properties of Var_{RDS} only depend on whether $\rho_\pi^2(y, f_2)$ is zero or nonzero, the magnitude of $\rho_\pi^2(y, f_2)$ is highly relevant in a finite sample.

To collect the desired sample size, each tree is initially sampled as a Galton–Watson tree with offspring distribution $1 + W$, where $W \sim \text{Binomial}(2, 1/2)$. So, $m = 2$ and the critical threshold is when the second eigenvalue is equal to $1/\sqrt{2} \approx 0.71$. Then the RDS sample is constructed without-replacement, using the procedure described in Section 4.1.

To vary the value of λ_2, each network G is simulated from a two block Stochastic Blockmodel [Holland, Laskey and Leinhardt (1983)] with 1/3 of the nodes in block 0 with $z(i) = 0$ and 2/3 of the nodes in block 1 with $z(i) = 1$. The size of the networks is set to $N = 50{,}000$ and the probability of a connection between two nodes in different blocks is $r = 15/N$. Then λ_2 varies between 0.5 and 0.9 by varying the probability of a connection between two nodes in the same block. These within-block probabilities are parameterized as follows:

$$(12) \qquad \begin{aligned} \mathbb{P}\big((i, j) \in E \mid z(i) = z(j) = 0\big) &= 7/3\omega r, \\ \mathbb{P}\big((i, j) \in E \mid z(i) = z(j) = 1\big) &= 3/7\omega r, \end{aligned}$$

where ω ranges between 2.5 and 20. As ω increases, so does λ_2. The constants 7/3 and 3/7 ensure that, for the values of ω that we consider, the expected degree of a node in block zero [$z(i) = 0$] divided by the expected degree of a node in block one [$z(i) = 1$] is roughly constant at 2.5. This ensures that (1) the sample average is biased under the Markov model and (2) this bias does not significantly change with λ_2. As such, this parameterization isolates the effects of changing λ_2 and is not confounded by any issues of changing bias.

To control $\rho_\pi^2(y, f_2)$, the simulations examine two types of node features y, `aligned` and `correlated`. In the simulations for `aligned`, $y(i) = z(i)$ for all $i \in G$. In the `correlated` simulation, 45% of the nodes in block 0 have $y(i) = 1$ and 10% of the nodes in block 1 have $y(i) = 1$; the rest of the nodes have $y(i) = 0$. In the `aligned` simulation, $\rho_\pi^2(y, f_2)$ is close to one. In the `correlated` simulation, $\rho_\pi^2(y, f_2)$ is around .15; see Figure 2 in the Supplementary Material [Rohe (2019)] for more details.

The first step of the simulation is to generate the referral tree $\mathbb{T}$ from the Galton–Watson distribution with $n = 1000$ nodes. Then the two types of node features y are generated (`aligned` and `correlated`). The $\mathbb{T}$ and y's are fixed across all simulations. Then the following six steps create one replicate of the experiment:

1. Simulate the underlying network G from a Stochastic Blockmodel. To parameterize the Stochastic Blockmodel, edges between blocks occur with probability $r = 15/N$ and edges within each block occur with a probability that changes as a function of ω, as defined in (12).

2. Simulate a respondent-driven sample of 1000 nodes, without-replacement, using the more realistic model described in Section 4.1.

3. To examine sample sizes $n \neq 1000$, retain only the first n samples.

4. Draw 500 samples from each of the resampling distributions (A-TREE-BOOTSTRAP, U-TREE-BOOTSTRAP, A-CHAIN-BOOTSTRAP and SS-BOOTSTRAP).

Coverage probabilities for nominally 90% intervals when outcome is `aligned` with the network.
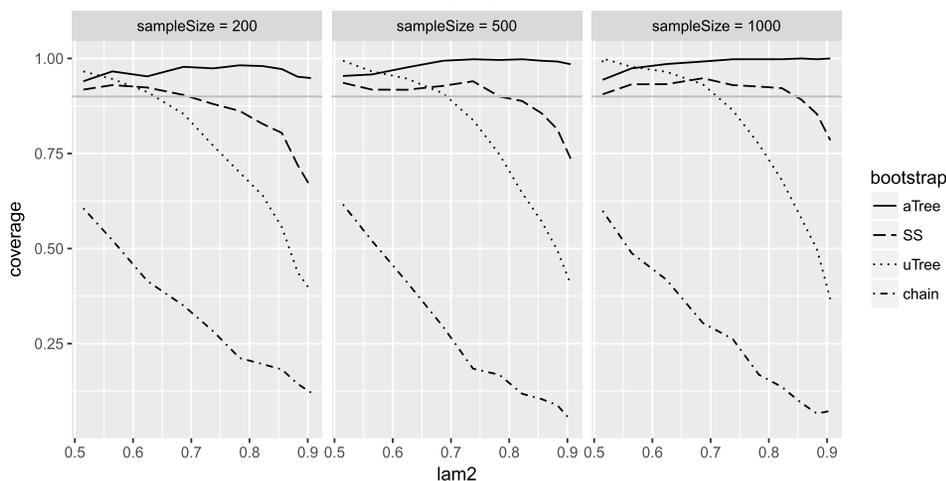
FIG. 4. *In these simulations, y is perfectly* `aligned` *with z, the referral bottleneck in the graph. Across different sample sizes and varying strengths of referral bottlenecks,* A-TREE-BOOTSTRAP *creates confidence intervals with conservative coverage probabilities.*

5. Compute $\hat{\mu}^*_{\text{VH}}$ on each of the bootstrap samples.

6. For each resampling distribution, use the 500 values of $\hat{\mu}^*_{\text{VH}}$ to compute the percentile confidence interval with the 5th to the 95th percentile of the bootstrap distribution for $\hat{\mu}^*_{\text{VH}}$.

To examine the frequentist properties of these confidence intervals, the above five steps are repeated 501 times; 501 to avoid confusion with the number of bootstrap samples in step 3.

6.3. *Simulation results*; *the coverage probabilities of the confidence intervals.* Figures 4 and 5 display the estimated coverage probabilities as a function of the bottleneck strength λ_2. The three panels correspond to different sample sizes. Figure 4 displays the results for `aligned` y. Figure 5 gives the results for `correlated` y. While all of the confidence intervals are nominally 90%, the figures show that the actual coverage probabilities can deviate substantially from 90%.

Across simulation settings, the nominally 90% confidence intervals from A-CHAIN-BOOTSTRAP have coverage probabilities ranging from 90% to 10%. These coverage probabilities are small in situations where the bottleneck is strong. This demonstrates the sensitivities of A-CHAIN-BOOTSTRAP discussed in Section 5 and in Goel and Salganik (2010). SS-BOOTSTRAP has coverage probabilities close to 90% when y is `aligned` and λ_2 is not too large. However, when y is `correlated`, the coverage probabilities for SS-BOOTSTRAP quickly diminish for moderate to large values of λ_2. The U-TREE-BOOTSTRAP confidence intervals are con-

Coverage probabilities for nominally 90% intervals when outcome is `correlated` with the network.
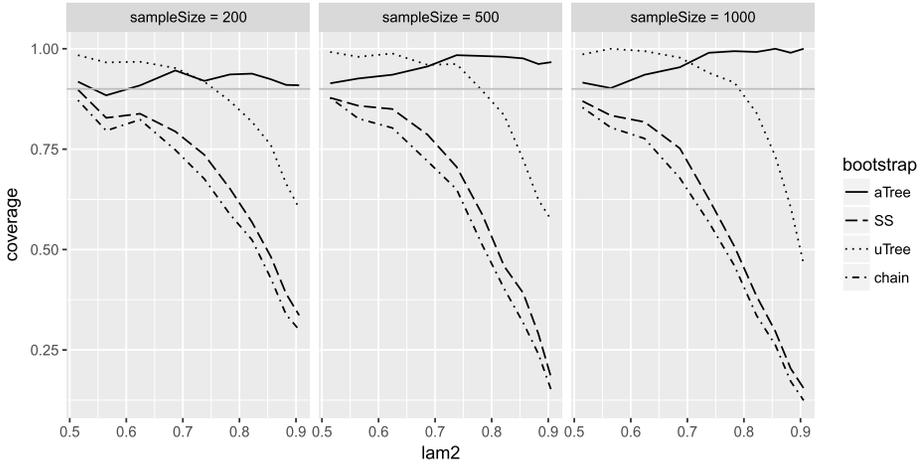
FIG. 5.   *In these simulations,* y *is* `correlated` *with* z, *the referral bottleneck in the graph. Across different sample sizes and varying strengths of referral bottlenecks,* A-TREE-BOOTSTRAP *creates confidence intervals with conservative coverage probabilities.*

servative for small values of λ_2 and anticonservative for larger values of λ_2. The intervals produced by A-TREE-BOOTSTRAP are conservative across the simulation settings.

Note that in these simulations, the intervals from A-TREE-BOOTSTRAP appropriately cover μ_{true} [i.e., not $\mathbb{E}(\hat{\mu}_{\text{VH}}|X_0)$]. Similar to Remark 2.1, these intervals account for the uncertainty due to seed selection (sometimes called seed-bias).

The results for the "studentized" confidence intervals were studied, but are not displayed. The "studentized" confidence intervals are constructed as $\hat{\mu}_{\text{VH}} \pm 1.65\hat{\sigma}$, where $\hat{\sigma}$ is the standard error of $\hat{\mu}_{\text{VH}}^*$ in the 500 bootstrap samples. In the simulations, the studentized intervals from the A-TREE-BOOTSTRAP often fail to be contained in $[0, 1]$, despite the fact that $y_i \in \{0, 1\}$ for all nodes i. Perhaps one reason for these strange results is that the accuracy of the studentized intervals depends on $\hat{\mu}_{\text{VH}}$ being asymptotically normal, while results in Li and Rohe (2017) suggest that it is not. In the limited simulations that were performed, the percentile interval was often (i) narrower and (ii) more likely to cover μ_{true} than the studentized interval. The percentile interval can simultaneously improve both of these metrics because it is not necessarily symmetric around the point estimate.

Section 3 in the Supplementary Material [Rohe (2019)] presents another simulation which studies the widths of the confidence intervals [Rohe (2019)]. Using a network with $\lambda_2 \approx 0.82$ (i.e., beyond the critical threshold), it studies how the width of the confidence interval decays as the sample size increases; see Figure 1 in the Supplementary Material [Rohe (2019)] for more details.

7. Discussion. A common concern in the RDS literature has been the design effect of network sampling techniques [Salganik (2006), Goel and Salganik (2010), Szwarcwald et al. (2011), Johnston et al. (2013), Verdery et al. (2015)]. Theorems 2.1 and 3.1 use the Markov model to give a rigorous account of the variance and design effect of RDS. In particular, if $m > \lambda_2^{-2}$, then the design effect can grow with the sample size; this is equivalent to saying the the variance of the estimator decays slower than $O(n^{-1})$. If the design effect is growing, then it should not be used for sample size or power calculations for two reasons. First, there might not be a central limit theorem to justify this approach [Li and Rohe (2017)]. Second, if DE changes with n, then many of the standard formulas are not well-defined (or they are incorrect). Instead of using DE to summarize the quality of the sample, a more reasonable summary would be the "half-life of the standard error." That is, given an RDS with sample size n, how much larger should $\tilde{n}$ be such to decrease the standard error by 50%. For example, estimators which are $\sqrt{n}$-consistent (i.e., constant DE) have a half-life of 4. Past the critical threshold in RDS, the standard error decays like n^{γ}, where $\gamma = \log_m \lambda_2 \in (-1/2, 0)$. This means that the half-life of the standard error is $(1/2)^{(1/\gamma)} > 4$.

Section 4 examines how well the $(\mathbb{T}, P)$-*walk on G* (which samples with-replacement) approximates a more accurate simulation model (which samples without-replacement). Proposition 4.1 and Theorem 4.1 give matching lower and upper bounds on the expected number of repeated pairs in a $(\mathbb{T}, P)$-*walk on G*. So long as $n = o(\sqrt{N})$, and some further technical conditions, these bounds show that λ_2 and the critical threshold do not affect the rate of $\mathbb{E}(R_n)$. As such, the critical threshold does not create additional repeated pairs. Section 4.1 presents a simulation comparing the $(\mathbb{T}, P)$-*walk on G* to a network sample taken without-replacement. Under the simulation settings, both the with-replacement and without-replacement samples displayed a similar critical threshold.

Section 6 introduces A-TREE-BOOTSTRAP, a new resampling procedure for computing confidence intervals for $\hat{\mu}_{VH}$. In a wide range of simulation settings, the intervals from A-TREE-BOOTSTRAP produced intervals with conservative coverage probabilities (i.e., the nominally 90% intervals had actual coverage that exceeded 90%). In contrast, there were simulation settings under which A-CHAIN-BOOTSTRAP, SS-BOOTSTRAP and U-TREE-BOOTSTRAP produced intervals with coverage probabilities that fall short of their nominal values. A key advantage of the U-TREE-BOOTSTRAP and SS-BOOTSTRAP is that they do not require z. In contrast, a key practical limitation of the A-CHAIN-BOOTSTRAP is that it requires a choice of z; that is, we must identify the referral bottleneck. More research is needed to (1) make U-TREE-BOOTSTRAP and SS-BOOTSTRAP less sensitive to λ_2 and (2) guide the choice of z for A-TREE-BOOTSTRAP.

APPENDIX A: PROOF OF THEOREM 2.1

The proof requires some notation and the following lemma. Throughout, let $\{X_\sigma : \sigma \in \mathbb{T}\}$ be a $(\mathbb{T}, P)$-*walk on G*. Let $\{X(i) : i \in 0, 1, \dots\}$ be a Markov chain

with the same transition matrix P that is initialized from π. Define $d(\sigma, \tau)$ as the graph distance between nodes σ and τ in $\mathbb{T}$.

LEMMA A.1. *If the transition matrix P is reversible, then for any two nodes σ and τ in the referral tree,*

$$\mathbb{P}(X_\sigma = u, X_\tau = v) = \mathbb{P}(X(0) = u, X(d(\sigma, \tau)) = v).$$

PROOF. Let $p = \sigma \wedge \tau$ be the most recent common ancestor of σ and τ. By the reversibility of the process, express $\mathbb{P}(X_\sigma = u, X_\tau = v)$ as

$$\sum_\ell \mathbb{P}(X_\sigma = u, X_p = \ell, X_\tau = v)$$

$$= \sum_\ell \pi_\ell \mathbb{P}(X_\sigma = u | X_p = \ell) \mathbb{P}(X_\tau = v | X_p = \ell)$$

$$= \sum_\ell \pi_u \mathbb{P}(X_p = \ell | X_\sigma = u) \mathbb{P}(X_\tau = v | X_p = \ell)$$

$$= \sum_\ell \pi_u \mathbb{P}(X(d(\sigma, p)) = \ell | X(0) = u) \mathbb{P}(X(d(p, \tau) + d(\sigma, p))$$

$$= v | X(d(\sigma, p)) = \ell)$$

$$= \mathbb{P}(X(0) = u, X(d(\sigma, \tau)) = v). \qquad \square$$

Also, we require a fuller version of Lemma 2.1, which comes from Levin, Peres and Wilmer (2009).

LEMMA A.2 [Lemma 12.2 in Levin, Peres and Wilmer (2009)]. *Let P be a reversible Markov transition matrix on the nodes in G with respect to the stationary distribution π. The eigenvectors of P, denoted as $f_1, \ldots, f_N$, are real valued functions of the nodes $i \in G$ and orthonormal with respect to the inner product*

$$(13) \qquad \langle f_a, f_b \rangle_\pi = \sum_{i \in G} f_a(i) f_b(i) \pi_i.$$

If λ is an eigenvalue of P, then $|\lambda| \leq 1$. The eigenfunction f_1 corresponding to the eigenvalue 1 can be taken to be the constant vector $\mathbf{1}$, in which case the probability of a transition from $i \in G$ to $j \in G$ in t steps can be written as

$$(14) \qquad \mathbb{P}(X(t) = j | X(0) = i) = P_{ij}^t = \pi_j + \pi_j \sum_{\ell=2}^N \lambda_\ell^t f_\ell(i) f_\ell(j).$$

The following is a proof of Theorem 2.1.

PROOF.

$$\text{Var}_{\text{RDS}}(\hat{\mu}) = \frac{1}{n^2} \text{Var}_{\text{RDS}} \left(\sum_{\tau \in \mathbb{T}}^{N} y(X_\tau) \right)$$

$$= \frac{1}{n^2} \sum_{\sigma, \tau \in \mathbb{T}} \text{Cov}_{\text{RDS}}(y(X_\sigma), y(X_\tau)).$$

For ease of notation, let $t = d(\sigma, \tau)$. From Lemma A.1 (and suppressing the RDS subscript),

$$\text{Cov}(y(X_\sigma), y(X_\tau)) = \mathbb{E}(y(X(0))y(X(t))) - (\mathbb{E}y(X(0)))^2.$$

Using the spectral decomposition of P (see Lemma A.2), with the fact that f_1 is a constant vector and $\lambda_1 = 1$ [Levin, Peres and Wilmer (2009)],

$$\mathbb{E}(y(X(0))y(X(t))) = \sum_{u,v \in G} y(u)y(v)\mathbb{P}(X(0) = u, X(t) = v)$$

$$= \sum_{u,v \in G} y(u)y(v)\pi_u P^t(u, v)$$

$$= \sum_{u,v \in G} y(u)y(v)\pi_u \pi_v \sum_{\ell=1}^{N} \lambda_\ell^t f_\ell(u) f_\ell(v)$$

$$= \sum_{u,v \in G} y(u)y(v)\pi_u \pi_v \left(1 + \sum_{\ell=2}^{N} \lambda_\ell^t f_\ell(u) f_\ell(v) \right)$$

$$= \left(\sum_{u \in G} y(u)\pi_u \right)^2 + \sum_{\ell=2}^{N} \lambda_\ell^t \left(\sum_{u \in G} y(u)\pi_u f_\ell(u) \right)^2$$

$$= (\mathbb{E}y(X(0)))^2 + \sum_{\ell=2}^{N} \lambda_\ell^t \langle y, f_\ell \rangle_\pi^2.$$

Terms cancel. So,

(15) $$\text{Cov}(y(X_\sigma), y(X_\tau)) = \sum_{\ell=2}^{N} \lambda_\ell^{d(\sigma,\tau)} \langle y, f_\ell \rangle_\pi^2.$$

Then,

$$\text{Var}_{\text{RDS}}(\hat{\mu}) = n^{-2} \sum_{\sigma, \tau \in \mathbb{T}} \text{Cov}(y(X_\sigma), y(X_\tau))$$

$$= n^{-2} \sum_{\sigma, \tau \in \mathbb{T}} \sum_{\ell=2}^{N} \lambda_\ell^{d(\sigma,\tau)} \langle y, f_\ell \rangle_\pi^2$$

$$= n^{-2} \sum_{\ell=2}^{N} \langle y, f_\ell \rangle_\pi^2 \sum_{\sigma, \tau \in \mathbb{T}} \lambda_\ell^{d(\sigma, \tau)}$$

$$= \sum_{\ell=2}^{N} \langle y, f_\ell \rangle_\pi^2 \mathbb{G}(\lambda_\ell). \qquad \square$$

APPENDIX B: COMPARING THE CRITICAL THRESHOLD BETWEEN SAMPLING WITH AND WITHOUT REPLACEMENT UNDER AN UNBALANCED STOCHASTIC BLOCKMODEL

Figure 6 repeats the simulation given in Section 4.1 and Figure 3. In the previous simulation, each block in the Stochastic Blockmodel accounted for 50% of the nodes in the graph. In Figure 6, 1/3 of the nodes are in block zero and the rest are in block one. To parameterize the model with the specified expected degree and λ_2, define $r = 15/N$. Using parameters α and ω to control the expected degree and λ_2 (resp.), the edge probabilities are specified as follows:

$$\mathbb{P}\big((i, j) \in E | z(i) = z(j) = 0\big) = \alpha r,$$
$$\mathbb{P}\big((i, j) \in E | z(i) = z(j) = 0\big) = \alpha 7/3 \omega r \quad \text{and}$$
$$\mathbb{P}\big((i, j) \in E | z(i) = z(j) = 1\big) = \alpha 3/7 \omega r.$$

Each column of panels in Figure 6 corresponds to a value of λ_2. These values of λ_2 are created with $\omega \in \{3.55, 4.45, 5.35, 6.78, 9\}$. Each row of panels in Figure 6 corresponds to a different expected degree (for a node chosen uniformly at random). The value of α is adjusted to create the desired expected value.

APPENDIX C: FURTHER JUSTIFICATION FOR THE MOTHER NODE IN A-TREE-BOOTSTRAP

A-TREE-BOOTSTRAP is a $(\mathbb{T}, \hat{P})$-*walk*, where $\hat{P}$ is defined in (11). As such, Theorem 2.1 and $\mathbb{G}(\lambda)$ can be used to study the variance of $\hat{\mu}^*$, providing a closed-form expression for the bootstrap variance (and standard error). This can be computed with $\mathbb{G}$ and the eigenproperties of $\hat{P}$ and does not require simulating the bootstrap. Here, we focus on the role of $\mathbb{T}$ and $\mathbb{G}$ in that formula. In particular, Figure 7 studies how the standard error of $\hat{\mu}^*$ is sensitive to whether or not $\mathbb{T}$ includes a mother node. Note that the bootstrap standard error should not be used to construct confidence intervals because $\hat{\mu}^*$ is not necessarily normally distributed; Figure 7 uses standard error as merely one measure of uncertainty.

As discussed after Remark 2.1, it is necessary to randomize the seed node(s) in the bootstrap in order to incorporate the additional uncertainty that comes from the selection of the seeds. That uncertainty is often called seed-bias. However, if we

With and without replacement sampling display a similar critical threshold (unbalanced model)
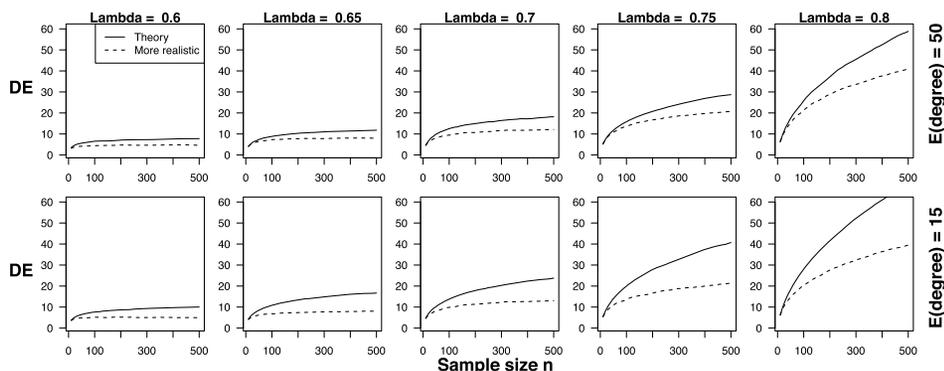
FIG. 6. *This simulation is identical to the simulation displayed in Figure 3, with one exception. In this figure, 1/3 of the nodes are in block zero (instead of 1/2, as in Figure 3). Under this deviation, the sample average becomes biased under the Markov model. The theoretical results show that the critical threshold depends on m and λ_2, not the bias of the sample average. This figure confirms that the critical threshold for the more realistic model is also insensitive to the bias of the sample average.*

randomize the seed node(s), then seed-bias becomes seed-variance. In this way, A-TREE-BOOTSTRAP includes seed-bias/variance in its estimates of uncertainty by randomizing the seed node(s).

When there are multiple seed nodes, the bootstrap should randomize all of them. However, in typical RDS practice, seed nodes are not sampled independently from the population. As such, the bootstrap should not simulate the seeds to be independent. To mimic the dependence of the seeds, while still randomly selecting them within the bootstrap, A-TREE-BOOTSTRAP creates an artificial "mother node" in

Bootstrap standard error computed with square root of G;
dependent vs independent seed selection
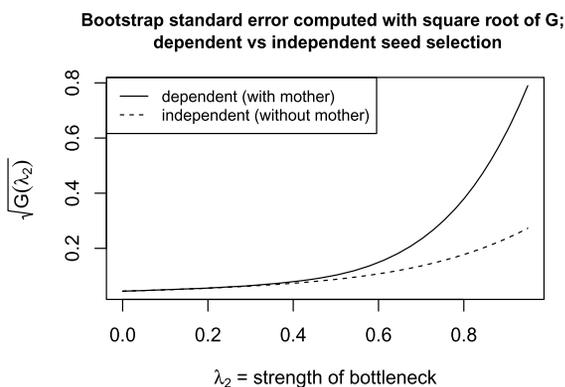
FIG. 7. *When bootstrapped seeds are sampled independently, the bootstrapped estimator $\hat{\mu}^*$ has a drastically smaller standard error when compared to the seed generation mechanism in A-TREE-BOOTSTRAP. Unless the seeds from the original sample were properly sampled from the target population, this drastic reduction in variance is a bootstrap induced illusion.*

$\mathbb{T}$ that refers all of the seed nodes via the estimated Markov process (11). In this way, A-TREE-BOOTSTRAP generates seeds in a dependent fashion.

If there are multiple seeds, then before the addition of a mother node, $\mathbb{T}$ is a forest of disconnected trees. Each tree has its own root and these roots have a one-to-one correspondence with the seeds. A-TREE-BOOTSTRAP artificially adds a mother node to $\mathbb{T}$ and connects the mother node to all of the seeds (i.e., the former root nodes). This mother node can be thought of as the research group that selects the seeds, typically from a convenience sample. Adding this node to $\mathbb{T}$ drastically changes the variability in Theorem 2.1 because the $\mathbb{G}$ function drastically changes.

The vertical axis in Figure 7 gives $\sqrt{\mathbb{G}(\lambda_2)}$; the two lines correspond to whether or not $\mathbb{T}$ contains a mother node. Under a simple construction of $\hat{P}$, $\sqrt{\mathbb{G}(\lambda_2)}$ is the bootstrapped standard error up to a constant of proportionality that can be easily computed for a specific data set and do not depend on the tree.[3] In Figure 7, the forest $\mathbb{T}$ is ten 2-trees of size fifty ($n = 500$) and is used to create the dashed line. The solid line is constructed by adding a mother node.

Without a mother node, different components of the forest are independent. By averaging over independent seed nodes, there is a drastic reduction in the seed-bias/variance; this is a bootstrap induced illusion (unless the actual data collection sampled the seeds independently from the stationary distribution π). With the mother node, A-TREE-BOOTSTRAP does not average over independent seeds. This leads to a larger standard error and more conservative bootstrapped confidence intervals.

SUPPLEMENTARY MATERIAL

Supplement: Proofs for Sections 3 and 4 (DOI: 10.1214/18-AOS1700SUPP; .pdf). Due to space constraints, this supplement contains the proofs for the results in Sections 3 and 4. Moreover, it contains an addition computational experiment to study the widths of the bootstrap confidence intervals.

---

[3]Consider an example where $\hat{P}$ in equation (11) is constructed with $z(X_\tau) = y_{X_\tau} \in \{0, 1\}$ (cf. Section 6.1). This formulation ensures that the sum in Equation (5) simplifies to a single term which is proportional to $\mathbb{G}(\lambda_2)$.

REFERENCES

ABDUL-QUADER, A. S., HECKATHORN, D. D., MCKNIGHT, C., BRAMSON, H., NEMETH, C., SABIN, K., GALLAGHER, K. and DES JARLAIS, D. C. (2006). Effectiveness of respondent-driven sampling for recruiting drug users in New York City: Findings from a pilot study. *J. Urban Health* **83** 459–476.

ARAYASIRIKUL, S., CAI, X. and WILSON, E. C. (2015). A qualitative examination of respondent-driven sampling (RDS) peer referral challenges among young transwomen in the San Francisco bay area. *JMIR Public Health Surveill.* **1** e9.

ATHREYA, K. B. and NEY, P. E. (1972). *Branching Processes. Die Grundlehren der Mathematischen Wissenschaften* **196**. Springer, New York. MR0373040

BARAFF, A. J., MCCORMICK, T. H. and RAFTERY, A. E. (2016). Estimating uncertainty in respondent-driven sampling using a tree bootstrap method. *Proc. Natl. Acad. Sci. USA* 201617258.

BENJAMINI, I. and PERES, Y. (1994). Markov chains indexed by trees. *Ann. Probab.* **22** 219–243. MR1258875

CHUNG, F. R. K. (1997). *Spectral Graph Theory. CBMS Regional Conference Series in Mathematics* **92**. Published for the Conference Board of the Mathematical Sciences, Washington, DC; by the Amer. Math. Soc., Providence, RI. MR1421568

GILE, K. J. (2011). Improved inference for respondent-driven sampling data with application to HIV prevalence estimation. *J. Amer. Statist. Assoc.* **106** 135–146. MR2816708

GILE, K. J. and HANDCOCK, M. S. (2010). Respondent-driven sampling: An assessment of current methodology. *Sociol. Method.* **40** 285–327.

GILE, K. J., JOHNSTON, L. G. and SALGANIK, M. J. (2015). Diagnostics for respondent-driven sampling. *J. Roy. Statist. Soc. Ser. A* **178** 241–269. MR3291770

GOEL, S. and SALGANIK, M. J. (2009). Respondent-driven sampling as Markov chain Monte Carlo. *Stat. Med.* **28** 2202–2229. MR2751515

GOEL, S. and SALGANIK, M. J. (2010). Assessing respondent-driven sampling. *Proc. Natl. Acad. Sci. USA* **107** 6743–6747.

HANDCOCK, M. S., FELLOWS, I. E. and GILE, K. J. (2016). RDS: Respondent-driven sampling. Los Angeles, CA, R package version 0.7-5. http://CRAN.R-project.org/package=RDS.

HECKATHORN, D. D. (1997). Respondent-driven sampling: A new approach to the study of hidden populations. *Soc. Probl.* **44** 174–199.

HOLLAND, P. W., LASKEY, K. B. and LEINHARDT, S. (1983). Stochastic blockmodels: First steps. *Soc. Netw.* **5** 109–137. MR0718088

JOHNSTON, L. G., CHEN, Y.-H., SILVA-SANTISTEBAN, A. and RAYMOND, H. F. (2013). An empirical examination of respondent driven sampling design effects among HIV risk groups from studies conducted around the world. *AIDS Behav.* **17** 2202–2210.

KHABBAZIAN, M., HANLON, B., RUSSEK, Z. and ROHE, K. (2017). Novel sampling design for respondent-driven sampling. *Electron. J. Stat.* **11** 4769–4812.

LEVIN, D. A., PERES, Y. and WILMER, E. L. (2009). *Markov Chains and Mixing Times*. Amer. Math. Soc., Providence, RI. MR2466937

LI, X. and ROHE, K. (2017). Central limit theorems for network driven sampling. *Electron. J. Stat.* **11** 4871–4895. MR3733297

LU, X., BENGTSSON, L., BRITTON, T., CAMITZ, M., KIM, B. J., THORSON, A. and LILJEROS, F. (2012). The sensitivity of respondent-driven sampling. *J. Roy. Statist. Soc. Ser. A* **175** 191–216. MR2873802

MCCREESH, N., FROST, S., SEELEY, J., KATONGOLE, J., TARSH, M. N., NDUNGUSE, R., JICHI, F., LUNEL, N. L., MAHER, D., JOHNSTON, L. G. et al. (2012). Evaluation of respondent-driven sampling. *Epidemiology* **23** 138.

ROCH, S. and ROHE, K. (2017). Generalized least squares can overcome the critical threshold in respondent-driven sampling. ArXiv Preprint ArXiv:1708.04999.

ROHE, K. (2019). Supplement to "A critical threshold for design effects in network sampling." DOI:10.1214/18-AOS1700SUPP.

ROHE, K., CHATTERJEE, S. and YU, B. (2011). Spectral clustering and the high-dimensional stochastic blockmodel. *Ann. Statist.* **39** 1878–1915. MR2893856

SALGANIK, M. J. (2006). Variance estimation, design effects, and sample size calculations for respondent-driven sampling. *J. Urban Health* **83** 98–112.

SALGANIK, M. J. and HECKATHORN, D. D. (2004). Sampling and estimation in hidden populations using respondent-driven sampling. *Sociol. Method.* **34** 193–240.

SZWARCWALD, C. L., DE SOUZA JÚNIOR, P. R. B., DAMACENA, G. N., JUNIOR, A. B. and KENDALL, C. (2011). Analysis of data collected by RDS among sex workers in 10 Brazilian cities, 2009: Estimation of the prevalence of HIV, variance, and design effect. *JAIDS J. Acquir. Immune Defic. Syndr.* **57** S129–S135.

VERDERY, A. M., MOUW, T., BAULDRY, S. and MUCHA, P. J. (2015). Network structure and biased variance estimation in respondent driven sampling. *PLoS ONE* **10** e0145296.

VOLZ, E. and HECKATHORN, D. D. (2008). Probability based estimation theory for respondent driven sampling. *J. Off. Stat.* **24** 79.

VON LUXBURG, U. (2007). A tutorial on spectral clustering. *Stat. Comput.* **17** 395–416. MR2409803

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. *J. Clin. Epidemiol.* **68** 1463–1471.

WORLD HEALTH ORGANIZATION and UNAIDS (2013). *Introduction To HIV/AIDS And Sexually Transmitted Infection Surveillance Module 4: Introduction to Respondent-drive Sampling.* World Health Organization & UNAIDS. http://applications.emro.who.int/dsaf/EMRPUB_2013_EN_1539.pdf.

STATISTICS DEPARTMENT
UNIVERSITY OF WISCONSIN–MADISON
1300 UNIVERSITY AVE
MADISON, WISCONSIN 53706
USA
E-MAIL: karlrohe@stat.wisc.edu