Network driven sampling; a critical threshold for design effects

Karl Rohe∗
University of Wisconsin - Madison

September 10, 2015

Abstract

Web crawling, snowball sampling, and respondent-driven sampling (RDS) are three types of network driven sampling techniques that are popular when it is difficult to contact individuals in the population of interest. This paper studies network driven sampling 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 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. Under certain assumptions on the referral tree, the design effect of network driven 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, $\lambda_2$. 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 the critical threshold, the estimator converges at the slower rate of $\log_m \lambda_2$. The Markov model allows for nodes to be resampled. Under certain conditions, the rate of resampling is not affected by the critical threshold, so long as $n = o(\sqrt{N})$, where $n$ is the sample size and $N$ is the population size.

Introduction

Statistical inference takes conclusions from a small set of observations and extends them to the entire population; “sampling” drastically reduces the burden of research in various disciplines. However, classical sampling techniques require a sampling frame that lists each individual in the population and a way of contacting each individual. In many settings, a sampling frame is not available. In others, it is too expensive to compile or only covers a biased subset of the population. Given the continual decline in response rates to telephone surveys, even standard sampling frames are becoming less useful; Pew Research Center reported that the typical response rate in telephone surveys declined from 36% to 9% between 1997 and 2012 [Kohut et al., 2012]. Network driven sampling has the potential to circumvent these difficulties because it does not require a sampling frame (e.g. Heckathorn [1997]).

∗Thank you Zoe Russek and Emma Krauska for recording the referral trees used in Figure 5 and helpful comments on this draft. Thank you Bret Hanlon, Mohammad Khazzabian, Mark Handcock, Sebastien Roch, Quansheng Liu, and Russell Lyons for thoughtful and helpful discussions over the course of this research.

1
In many populations of interest, the members are connected by a network; a social network, a hyperlinked network, or a protein-protein interaction network. This network provides a viable means for reaching the target population; the people, the webpages, or the proteins. Using a network to reach a target population goes by many names in various disciplines; respondent-driven sampling / snowball sampling, web crawling / link-tracing / breadth-first search, and Co-immunoprecipitation / chromatin immunoprecipitation. These disparate techniques are united in providing researchers access to otherwise hard-to-reach populations by essentially asking friends to refer friends. This paper studies random sampling in the context of a networked population.

Within public health research on HIV, respondent-driven sampling (RDS) is a popular technique to sample marginalized and/or hard-to-reach populations [Heckathorn 1997]. RDS has become widely and 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 by 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 (CDC), the World Health Organization (WHO), and the Joint United Nations Programme on HIV/AIDS (UNAIDS) [WHO 2013]. RDS serves as a motivating application in this paper. However, the results herein are applicable to network driven sampling in other domains.

This paper studies the design effect of network driven sampling. The design effect is the variance of an estimator constructed from a network driven sample divided by the variance of that estimator constructed from an independent random sample. For example, if the design effect is three, then a network driven sample increase the variance by a factor of three. Said another way, a network driven sample would require $\sqrt{3}$ times more samples to obtain an estimator with the same standard error.

The Markov model of network driven sample has been extensively used in the previous literature. Salganik and Heckathorn 2004 and others have modeled RDS as a simple random walk on the social network; such a model presumes that each person refers exactly one participant. Goel and Salganik 2009 modeled the RDS referral process as a Markov process indexed by a tree; this allows for the fact that some individuals make multiple referrals. When indexing the Markov process by a tree, the initial participant (i.e. the “seed node”) forms the root of the tree. Following the tree down to any leaf creates a standard Markov chain. Due to properties of the Markov chain, the dependence between samples decays exponentially fast along the branches of the referral tree. However, due to the branching structure, most pairs of samples are merely $O(\log n)$ apart. As such, observations in a tree are more dependent than observations in a chain.

![Sampling process and tree](image-url)

Figure 1: Network driven sampling has two graphs: the social network $G$ and the referral tree $T$. 

Diagnostics:

```
Estimation: 
```

...
Previous results in [Goel and Salganik 2009] studied a tree structure where every individual refers exactly two more individuals. These results only apply to a social network in which the adjacency matrix is rank two; this supposes that there are exactly two communities in the network and connection within and between these communities are entirely homogeneous. Theorem 2.1 extends these results by making no assumptions on the referral tree and requiring only minimal assumptions on the social network (e.g. connected, unipartite, etc.). To do this, the proof relies on the spectral decomposition of the Markov transition matrix. This allows for the comparison of (1) the logarithmically expanding distances in the tree against (2) the exponential convergence of the Markov chain. Comparing these quantities identifies the critical threshold that is discussed in the abstract.

A key function $G_n$ relates the referral tree to the Markov process and determines the asymptotic behavior of the design effect. Sample two nodes $I, J$ uniformly from the referral tree and compute their graph distance in the tree, $D = d(I, J)$; this is the number of referrals between $I$ and $J$. Evaluate the probability generating function for this random variable $D$ at $\lambda_2 < 1$, the second eigenvalue of the Markov transition matrix on the social network:

$$G_n(\lambda_2) = E\lambda_2^D,$$

where the subscript $n$ denotes the number of samples in the referral tree. As $n$ grows, the tree grows and the distribution of $D$ changes. Under certain conditions on the distribution of the outcome of interest (e.g. HIV status) in the network, the design effect of a Horvitz-Thompson (HT) based estimator [Volz and Heckathorn 2008] is asymptotically proportional to $nG_n(\lambda_2)$. The rate of $nG_n(\lambda_2)$ has two regimes. Suppose that the tree is “balanced” in a sense described in Section 4 and let $m$ be the average number of referrals. If $m < \lambda_2^{-2}$, then the design effect is bounded, $nG_n(\lambda_2) = O(1)$. However, if $m > \beta$, then the design effect grows with $n$, $nG_n(\lambda_2) = O(n^{1-\alpha})$, where $\alpha = \log m \lambda_2^{-2} \in (0, 1)$. When the design effect grows with $n$, it means that under the network driven sample, the variance of the HT estimator converges slower than the standard $1/n$ rate.

The value $\lambda_2$ is widely studied in the literature on Markov chains and spectral graph theory [Levin et al. 2009, Chung 1997] and it plays a key role in the critical threshold. It measures what the RDS literature refers to as “bottlenecks” in the referral process. Suppose there are two communities in the network, EAST and WEST. If the seed belongs to EAST and several waves of sampling fail to refer someone in WEST, then this indicates a bottleneck in the referral process and it creates additional dependence between successive samples. Such a bottleneck corresponds to a large value of $\lambda_2$ [Von Luxburg 2007]; this concept is formalized by the Cheegar cut [Chung 1997]. Because $G(z)$ is increasing with $z > 0$, a strong bottleneck creates additional variance.

The paper is organized as follows. Section 1 defines the following four separate pieces that constitute the model for RDS that is used throughout the paper: a social network, a Markov transition matrix, a referral tree, and a node feature that we wish to describe. Section 2 gives Theorem 2.1 which provides an exact formula for the variance of an RDS estimator. In standard applications, the formula for the variance can not be computed by observed data. Section 3 briefly mentions two possible paths that allow for tractable estimation of the variance. Section 4 gives upper and lower bounds on $G_n(z)$ as a function of the sample size $n$ (i.e. the number of nodes in the tree). These bounds require some conditions on the referral tree. $m$-trees, where every non-leaf node refers exactly $m$ participants, trivially satisfy these conditions. Theorem 4.3 combines the previous results to show that, under certain conditions, the bounds apply to random Galton-Watson referral trees. Section 5 studies the rate at which the Markov model resamples nodes. Section 6
illustrates the key elements of the variance formula in (1) the popular political blog network and (2) 18 different referral trees that have been published in previous RDS studies.

0.1 Previous literature

There are three strains of literature that are relevant to this paper. First, the extensive empirical and statistical literature on RDS. Second, the probability literature on Markov processes indexed by chains and trees. Finally, a highly computational and experimental data mining literature in web graph sampling. The relevant pieces of these literatures are summarized below.

RDS was first proposed in Heckathorn [1997] and it is an extension of snowball sampling, a more commonly used term and technique [Handcock and Gile, 2011]. The classical statistical results on snowball sampling in Goodman [1961] studies estimators of the social network structure. More recently, these sampling techniques have been applied to situations where the network structure is ancillary; it is used only to reach/sample participants from a hard to reach population. A key innovation of RDS is the use of dual incentives to first participate and then, to refer future participants.

The techniques of RDS have been refined in numerous publications since its introduction. Salganik and Heckathorn [2004] first discussed RDS as a random walk on a social network. Several others examined the procedure and proposed various refined estimators [Volz and Heckathorn, 2008, Goel and Salganik, 2009, 2010, Gile and Handcock, 2010]. A common concern in this literature has been the variability of the RDS estimators. Salganik [2006], Gile [2011], Gile and Handcock [2015] proposed techniques to estimate the variance and the design effects of RDS estimators. Using the bootstrap estimator proposed in Salganik [2006], Szwarcwald et al. [2011] and Johnston et al. [2013] reanalyzed a wide variety of RDS samples. In these range of studies, the estimated design effect of RDS is between four and eight. These bootstrap techniques underestimate the actual variability (e.g. Neely [2009], Goel and Salganik [2010], Verdery et al. [2013]). Using a rank two network model and a perfect binary referral tree, Goel and Salganik [2009] shows that that the standard RDS estimator can converge slower than the standard rate. The results in the current paper extend this result to a more general network model. Moreover, the results below allow for more general referral trees. These generalizations allow for the identification of the critical threshold.

The current paper relates two strings of previous literature in probability; one that studies the variability of Markov chain sampling and another that studies Markov processes indexed by a tree. For example, primarily motivated by applications to Markov Chain Monte Carlo, Geyer [1992], Aldous and Fill [2002] and Jones et al. [2004] studied the variability of Markov chain sampling. This extant literature does not address tree indexing. The literature that has studied Markov processes indexed by a tree (e.g. Benjamin and Peres [1994], Peres [1999], Yang [2003], Dembo et al. [2005]), has not addressed the questions relevant to RDS—in particular, where there is an underlying graph with node features. Both of these literatures have open source texts for reference—see Aldous and Fill [2002], and Levin et al. [2009] for results on Markov chains and mixing times and see Lyons and Peres [2005] for random processes indexed by trees or graphs.

In an entirely separate thread of literature, several data mining papers have applied network driven sampling to hyperlinked webpages. These sampling techniques have been applied for various purposes (visualization, compression, storing and retrieval) [Krishnamurthy et al., 2005]. More recently, this literature has studied the various biases induced by different sampling mechanisms (e.g. Leskovec and Faloutsos [2006], Stutzbach et al. [2009], Kurant et al. [2010], Gjoka et al. [2010], Ribeiro and Towsley [2010], Avrachenkov et al. [2010]). These data mining papers seek to find a
representative sample of the nodes and edges so that conclusions from the sample can be extended to the entire graph. Much of the evidence in this literature is highly computational, based on simulations and computer experiments rather than statistical theory.

1 Preliminaries

There are four mathematical pieces necessary to model the statistical performance of the RDS estimators; 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. The aim of network sampling is to measure the average node feature for all nodes in the graph, using only the features of the sampled nodes.

1.1 Markov processes on a graph

A node set $V = \{1, \ldots, N\}$ contains the people in the social network and an edge set $E = \{(i, j) : i$ and $j$ are friends$\}$ contains the friendships. Together, they form the social network, $G = (V, E)$. The results in this paper require that the graph is undirected; they do not require that the graph is unweighted. If the graph is unweighted, define the degree of node $i$ as the number of connections to node $i$, $\text{deg}(i) = \sum_j 1\{(i, j) \in E\}$. If the graph is weighed, define the degree as the sum of the edge weights. Throughout the text $i \in G$ is used synonymously with $i \in V$ to simplify notation.

1.1.1 Markov chain on $G$

Denote $X(0), X(1), X(2) \cdots \in G$ as a Markov chain on the individuals from the social network $G$. For example, given that $X(t) = i \in G$, choose $X(t + 1)$ uniformly from $i$’s friends, $P(X(t + 1) = j | X(t) = i) = \frac{1\{(i, j) \in E\}}{\text{deg}(i)}$; this is a simple random walk on the nodes of $G$. The transition matrix for a Markov chain $P \in \mathbb{R}^{N \times N}$ defines the transition probabilities, $P_{ij} = P(X(t + 1) = j | X(t) = i)$.

If we presume that coupons can only be transmitted along edges in the social network, then $P$ has the restriction that $(i, j) \notin E \implies P_{ij} = 0$.

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 et al. [2009]). The literature on spectral graph theory and spectral clustering has extensively studied the second largest eigenvalue of $P$ [Chung, 1997, Von Luxburg, 2007]. If $\lambda_2$ is close to one, then there are clusters or communities in the graph. In particular, if the simple random walk transition matrix has $\lambda_2 = 1$, then the social network is disconnected; this represents an extreme bottleneck.

The Markov chain is reversible if there is a function $\pi : G \rightarrow \mathbb{R}$ that satisfies

$$\pi_i P_{i,j} = \pi_j P_{j,i}, \quad \text{for all } i, j \in G.$$\(^1\)

\(^1\)If $|\lambda_2| < 1$ and the chain is reversible, then the chain is irreducible and aperiodic.
For example, if \( G \) is undirected, then the simple random walk on \( G \) is reversible. If \( P \) is reversible and \( |\lambda_2| < 1 \), the function \( \pi \) is the stationary distribution and it satisfies

\[
\lim_{t \to \infty} \mathbb{P}(X(t) = j | X(0) = i) = \pi_j, \quad \text{for all } i, j \in G.
\]

When \( G \) is undirected and \( P \) is the simple random walk,

\[
\pi_j \propto \text{deg}(j) \quad \forall \ j \in G.
\]

The popular Volz-Heckathorn (VH) estimator is a Horvitz-Thompson estimator (defined in Section 1.2) that uses these terms as inclusion probabilities. Under certain assumptions, the VH estimator is asymptotically unbiased.

The spectral properties of the Markov transition matrix enable tractable calculations. The following Lemma from Chapter 12 of Levin et al. [2009] provides the essential pieces.

**Lemma 1.1.** Let \( P \) be a reversible Markov transition matrix on the nodes in \( G \) with respect to the stationary distribution \( \pi \). 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

\[
\langle f_a, f_b \rangle_\pi = \sum_{i \in G} f_a(i) f_b(i) \pi_i.
\]

(1)

If \( \lambda \) 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

\[
\mathbb{P}(X(t) = j | X(0) = i) = \mathbb{P}_{ij}^t = \pi_j + \pi_j \sum_{\ell=2}^{N} \lambda_\ell^t f_\ell(i) f_\ell(j).
\]

(2)

The second eigenvector \( f_2 \) plays a fundamental role in the rest of the paper. Each of the leading eigenvectors represents a bottleneck in the referral process. In the previous example with two communities EAST and WEST, suppose that \( P \) corresponds to the simple random walk. If EAST and WEST correspond to the most dominant partition in the network, then \( f_2(i) \) and \( f_2(j) \) will have the same signs (i.e. +/−) if and only if \( i \) and \( j \) belong to the same community. By looking at the signs, \( f_2 \) partitions the graph into EAST and WEST. The Cheegar bound [Chung, 1997] makes this concept rigorous and provides an argument for why spectral clustering can partition a graph [Von Luxburg, 2007].

### 1.1.2 Markov process on \( G \) indexed by a tree

Another graph beyond the social network \( G \) is needed to index the network driven sample. While \( G \) contains the entire population, the second graph contains the sampled nodes. For example, under a Markov chain, this second graph contains the nodes 0, 1, 2, \ldots, \( n \) and the edges \( t - 1 \to t \) for \( t = 1, \ldots, n \). In this second graph, an edge corresponds to a referral in the sampling process. Because some participants provide multiple referrals and other participants provide zero referrals, this rest of the paper indexes the Markov process with a tree. Let \( T \) be a rooted tree—a connected graph with \( n \) nodes, no cycles, and a vertex 0. The seed participant is vertex 0 in \( T \) (cf Figure 1).
To simplify notation, \( \sigma \in \mathbb{T} \) is used synonymously with \( \sigma \) 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 \( \sigma \) (the node one step closer to the root). Let \( \mathcal{D}(\sigma) \subset \mathbb{T} \) denote the set of \( \sigma \) and all its descendants in \( \mathbb{T} \). The Markov process indexed by \( \mathbb{T} \) is the set of random variables \( \{X_\sigma : \sigma \in \mathbb{T} \} \) satisfying the Markov property

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

Just as in the standard Markov chain, the transition matrix \( P \in [0, 1]^{N \times N} \) describes the transition probabilities,

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

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 \( P \).

In the social network \( G \), an edge represents friendship. In the tree, an edge from \( \tau \in \mathbb{T} \) to \( \sigma \in \mathbb{T} \) represents that random individual \( X_\tau \in G \) refers random individual \( X_\sigma \in G \) in the \((\mathbb{T}, P)\)-walk on \( G \).

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.

### 1.2 Measurements and estimators

For each node \( i \in G \), let \( y(i) \in \mathbb{R} \) denote some characteristic of this node. For example, \( y(i) = 1 \) represents that \( i \in G \) is HIV+ and \( y(i) = 0 \) represents that \( i \in G \) is HIV-. We wish to estimate the population average

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

Under the previous example, this is the proportion of the population that is HIV+. We estimate \( \mu_{\text{true}} \) with observations

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

where \( X_\tau \) is a \((\mathbb{T}, P)\)-walk on \( G \). Denote

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

where a subscript of \( \cdot_{\text{RDS}} \) denotes that expectation are computed with the \((\mathbb{T}, P)\)-walk on \( G \). The sample average,

\[
\hat{\mu} = \frac{1}{n} \sum_{\tau \in \mathbb{T}} Y_\tau
\]

is an unbiased estimate of \( \mu \). However, \( \pi \) is not necessarily uniform. So, it is not necessarily true that \( \mu = \mu_{\text{true}} \).

If \( P \) is a simple random walk on \( G \), then the popular Volz-Heckathorn estimator [Volz and Heckathorn, 2008] provides an asymptotically unbiased estimate of \( \mu_{\text{true}} \) under the \((\mathbb{T}, P)\)-walk on \( G \). It uses generalized Horvitz-Thompson reweighting,

\[
\hat{\mu}_{\text{VH}} = \frac{\sum_{\tau \in \mathbb{T}} Y_\tau / \text{deg}(X_\tau)}{\sum_{\tau \in \mathbb{T}} 1 / \text{deg}(X_\tau)}.
\]
The results in this paper apply to an idealized form of the Volz-Heckathorn estimator, where we suppose that the normalization factor in $\pi_i \propto \text{deg}(i)$ can be properly identified. For $i \in G$, $w_i = n\pi_i$ is the expected number of times that node $i$ appears in the sample. Thus, the Horvitz-Thompson estimator is

$$\hat{\mu}_{HT} = \frac{1}{N} \sum_{\tau \in T} Y_{\tau} w_{X_{\tau}}^{-1} = \frac{1}{n} \sum_{\tau \in T} \frac{Y_{\tau}}{\pi_{X_{\tau}} N}.$$ 

**Remark 1.1.** The graph $G$, the node features $y(i)$, and the probabilities $\pi_i$ are all fixed. It is the sample $X_{\tau}$ that is random. Define a new node feature

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

and new node measurements $Y^\pi_{\tau} = y^\pi(X_{\tau})$. The sample average of the new measurements is exactly the idealized Volz-Heckathorn estimator $\hat{\mu}_{HT}$ that is an unbiased estimator of $\mu_{\text{true}}$. Because of this simple transformation, the theorems below that study $\hat{\mu}$ can also study $\hat{\mu}_{HT}$ by substituting $y^\pi$ for $y$.

This paper studies the design effect of the $(T, P)$-walk on $G$ with the estimator $\hat{\mu}$. The design effect compares the $(T, P)$-walk on $G$ to a random sample. Define $\text{Var}_{RDS} (\hat{\mu})$ as the variance of $\hat{\mu}$ under the $(T, P)$-walk on $G$. Define $W_1, \ldots, W_n \in G$ as independent random samples with $\mathbb{P}(W_i = j) = \pi_j$. Define

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

The **design effect** of the $(T, P)$-walk on $G$ is

$$DE(\hat{\mu}) = \frac{\text{Var}_{RDS} (\hat{\mu})}{\text{Var}_\pi (\hat{\mu})}. \quad (5)$$

In many of the results below, a key assumption is that $y$ has non-negligible correlation with $f_2$, the second eigenvector of $P$.

**Definition 1.** Denote the population correlation between $y$ and the second eigenvector of $P$ as

$$\rho_\pi(y, f_2) = \sigma^{-1} \langle y, f_2 \rangle_\pi, \quad (6)$$

where $\sigma^2 = \text{Var}_{RDS} Y_0$ and $\langle \cdot, \cdot \rangle_\pi$ is defined in equation (1).

If the social graph is split into two communities (e.g. EAST and WEST) and $f_2$ corresponds to this partition, then the $(T, P)$-walk on $G$ has the potential to oversample one of these communities. If $\rho_\pi(y, f_2)$ is large, then the referral bottleneck corresponding to $f_2$ induces dependence between the samples. However, if the correlation $\rho_\pi(y, f_2)$ is zero or close to zero, then the two communities have similar distributions of $y$ values (e.g. similar rate of HIV+). As such, over sampling one of the communities will not contribute excess variance to the estimator, making this bottleneck is irrelevant.
1.3 Two examples

Throughout the paper, two specific models will illustrate the results.

Example 1. Continuing the previous example where $f_2$ partitions the network into the EAST and WEST communities, suppose that $f_2(i)$ is positive when $i$ is a node in EAST and negative when $i$ is a node in WEST. Assume that for each node $i \in G$, $y(i) = \mu + \sigma f_2(i)$. This simple model implies that $y$ is larger than $\mu$ on the EAST nodes and smaller than $\mu$ on the WEST nodes.

This model will be useful below because $\rho_\tau(y, f_2) = 1$ and $\rho_\tau(y, f_\ell) = 0$ for $\ell > 2$. This leads to several simplifications.

A key limitation of the $(T, P)$-walk on $G$ is that it samples nodes “with replacement” (Section 5 examines the rate of resampling). In classical sampling, “sampling with replacement” is justified when the population $N$ is infinite. The second example model will illustrate the results for network driven sampling on a model for an infinite population graph.

Example 2. The Stochastic Blockmodel for a random network $G$ supposes that each node is assigned to one of $K$ classes with probabilities $\pi_1, \ldots, \pi_K$. Conditioning on these class assignments, nodes connect independently with probabilities that depend only on the class memberships [Holland et al., 1983]; for symmetric $\Psi \in [0, 1]^{K \times K}$,

$$\mathbb{P}(i \text{ connects to } j \in G \mid i \text{ belongs to class } u, j \text{ belongs to class } v ) = \Psi_{uv}.$$  

Let $G$ be sampled from the Stochastic Blockmodel and let $P$ be a simple random walk. Define

$$B_{uv} = \frac{\pi_v \Psi_{uv}}{\sum_{w=1}^{K} \pi_w \Psi_{uw}}, B \in [0, 1]^{K \times K}. \quad (7)$$

As the graph population $N$ increases, $B_{uv}$ contains the limiting probabilities that a node in class $u$ refers a node in class $v$. If the node features $y(i)$ are constant within the blocks, then the infinite population graph can be studied with

$$\{X_\tau \in \{1, \ldots, K\} : \tau \in T\} \sim (T, B)\text{-walk on } \{1, \ldots, K\}.$$  

Here, the random walk on the infinite population simplifies to a random walk on the $K$ different classes in the Stochastic Blockmodel.

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

2 The variance of RDS

The next theorem provides an exact formula for $\text{Var}_{RDS}(\hat{\mu})$ as a functional of $G$, defined as follows.

Definition 2. Select two nodes $I, J$ uniformly at random from the tree $T$. Define the random variable $D = d(I, J)$ to be the graph distance in $T$ between $I$ and $J$. Define $G$ as the probability generating function for $D$,

$$G(z) = \mathbb{E}(z^D).$$
In practice, $T$ is observed. So, the function $G$ can be computed (e.g. see Figure 5). In many studies there are multiple seed nodes. In these cases, $T$ is a “forest” and $G$ can be computed by setting $d(I,J) = \infty$ if $I$ and $J$ are in different connected components of $T$.

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

$$Var_{RDS}(\hat{\mu}) = \sum_{\ell=2}^{N} \langle y, f_\ell \rangle_\pi^2 G(\lambda_\ell),$$

where the subscript $RDS$ denotes that data have been collected through a $(T,P)$-walk on $G$, $\hat{\mu}$ is defined in Equation (3), $\langle \cdot, \cdot \rangle_\pi$ is defined in Equation (1), $f_1, \ldots, f_N : G \to \mathbb{R}$ are the eigenvectors of $P$ corresponding to eigenvalues $\lambda_1 > |\lambda_2| \geq \cdots \geq |\lambda_N|$, and $G$ is defined in Definition 2.

This theorem gives a closed form expression for the variance of $\hat{\mu}$ by decomposing $y$ with the eigenbasis for the Markov transition matrix $P$. In an asymptotic setting where the number of samples is growing, the coefficient $\langle y, f_\ell \rangle_\pi^2$ and the eigenvalues $\lambda_\ell$ remain unchanged; it is the function $G$ that changes with $n$. In previous research, Verdery et al. [2013] and Khabbazian et al. [2015] have proven this theorem for the special case that $T$ is a chain.

In the various forms of network driven sampling, it is likely that the tree $T$ will change from one experiment to the next. Theorem 2.1 shows that the variance of $\hat{\mu}$ changes with the tree, via the function $G$. A similar phenomenon happens in linear regression, where the design matrix controls the covariance of the estimator and the design matrix may change if the experiment were repeated. Because of this, the standard estimators of the covariance condition on the design. One argument for conditioning is formalized by the conditionality principal in Birnbaum [1962]. In an analogous fashion, Theorem 2.1 conditions on the tree $T$.

The first step of the proof shows 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, using Equation (2) in Lemma 1.1,

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

(9)

Summing over all $\sigma, \tau$ and exchanging summations yields the result. The appendix contains the full proof.

In Example 1, $y(i) = \mu + \sigma f_2(i)$ for all nodes $i \in G$. Under this model, Equation (8) has a particularly simple form,

$$Var_{RDS}(\hat{\mu}) = \sigma^2 G(\lambda_2).$$  

(10)

Motivated by this example, the next corollary simplifies Equation (8) with bounds that do not depend on the simple form $y(i) = \mu + \sigma f_2(i)$. Recall that the ordering of the eigenvalues is in absolute value, $|\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_N|$.

**Corollary 2.1.** Under the conditions of Theorem 2.1, if $\lambda_2 > 0$, then

$$\rho_2^2(y, f_2) n G_n(\lambda_2) \leq DE(\hat{\mu}) \leq n G_n(\lambda_2),$$

where $DE(\hat{\mu})$ is the design effect defined in Equation (5).
If the outcome of interest \( y \) correlates with the largest bottleneck in the network (i.e., \( \rho^2(y, f_2) > 0 \)), then the design effect is asymptotically proportional to \( n^2 \mathcal{G}(\lambda_2) \). As such, the \((T, P)\)-walk on \( G \) has a bounded design effect if and only if \( \mathcal{G}(\lambda_2) = O(n^{-1}) \). To examine this, Section 4 gives upper and lower bounds for \( \mathcal{G}_n(\lambda_2) \).

### 3 Paths towards estimating the variance

Motivated by the two example models in Section 1.3, this section highlights two paths towards estimating the variance defined in Theorem 2.1. The first estimator does not have any precedence in the literature. The second estimator provides a closed form generalization of a popular bootstrap estimator that was proposed in Salganik [2006]. Due to space limitations, the details of these estimators are not explored in the current paper.

#### 3.1 Example 1

Under the model \( y = \mu f_1 + \sigma f_k \), for \( k > 1 \), \( \text{Var}_{RDS}(\hat{\mu}) \) has a particularly simple form, \( \sigma^2 \mathcal{G}(\lambda_k) \). To form a plug-in estimator, there are standard estimators for \( \sigma \) and the function \( \mathcal{G} \) is easily computable. To estimate the only remaining quantity \( \lambda_k \) requires the model on \( y \). Using this model, the covariance between adjacent observations \( \sigma, \sigma' \in \mathcal{T} \) simplifies from Equation (9),

\[
\text{Cov}_{RDS}(Y_{\sigma}, Y_{\sigma'}) = \sum_{\ell=2}^{\ell=N} \lambda_{\ell}^2 \mathcal{G}(\lambda_{\ell}) = \lambda_k \sigma^2.
\]

So, the autocorrelation (at lag one) of the \((T, P)\)-walk on \( G \) is exactly \( \lambda_k \). The sample autocorrelation provides an estimator. The resulting plug-in estimator will be explored in future research.

#### 3.2 Example 2: model based inference

To examine the \((T, P)\)-walk on \( G \) in an infinite population model, Example 2 proposes the Stochastic Blockmodel. Suppose that each of the \( K \) blocks has a mean outcome \( y(1), \ldots, y(K) \). For each observation \( \tau \in \mathcal{T} \), we observe \( Y_\tau = y(X_\tau) + \epsilon_\tau \), where \( X_\tau \) is the block membership of sample \( \tau \in \mathcal{T} \) and \( \epsilon_\tau \) is an independent error with mean zero and variance \( \tilde{\sigma}^2 \). From Theorem 2.1, if \( X_\tau \) is a \((T, B)\)-walk on \( \{1, \ldots, K\} \),

\[
\text{Var}_{RDS}(\hat{\mu}) = \sum_{\ell=2}^{K} \langle y, f_\ell \rangle^2 \mathcal{G}(\lambda_\ell) + \tilde{\sigma}^2/n,
\]

where \( f_\ell, \lambda_\ell, \) and \( \pi \) are not defined with respect the transition matrix \( P \), but instead with respect to the transition matrix \( B \) in Equation (7) on the blocks \( 1, \ldots, K \).

Previous research has estimated \( \text{Var}_{RDS}(\hat{\mu}) \) for binary outcome \( y \) by using \( y \) to construct a two block Stochastic Blockmodel and employing a parametric bootstrap [Salganik, 2006, Szwarcwald et al., 2011, Johnston et al., 2013]. Equation (11) generalizes this approach and provides a closed form solution. In both the case here and the case in Salganik [2006], the underlying network is modeled with a Stochastic Blockmodel where observed node features (e.g., demographic characteristics, health status, etc) determine the blocks; then, \( \hat{B} \) and \( \hat{y} \) can be computed with sample
proportions. The resulting plug-in estimator is
\[
\hat{\text{Var}}_{\text{RDS}}(\hat{\mu}) = \sum_{\ell=2}^{K} \langle \hat{y}, \hat{f}_\ell \rangle^2 \hat{G}(\hat{\lambda}_\ell) + \hat{\sigma}^2 / n.
\]
This estimator will be explored in future research.

4 Asymptotic behavior of the design effect

To understand the design effect of RDS, Corollary 2.1 shows that it is necessary to understand how \( G_n(\lambda_2) \) depends on the sample size \( n \). If \( G(\lambda_2) = O(n^{-1}) \) as \( n \to \infty \), then the design effect for \( \hat{\mu} \) is bounded. Otherwise, the design effect grows with \( n \).

Subsections 4.1 and 4.2 give lower and upper bounds to \( G_n(z) \), respectively. These subsections illustrate the upper and lower bounds with \( m \)-trees. These are trees in which every participant refers exactly \( m \) future participants, until the final round, in which there are zero referrals. Section 4.3 applies the bounds to random trees in which each participant refers an iid number of participants (i.e. Galton-Watson trees).

4.1 Lower bounds for \( G(\lambda_2) \)

**Fact 4.1.** Select two nodes \( I, J \) from \( T \) uniformly at random. Define the random variable \( D = d(I, J) \) to be the graph distance in \( T \) between \( I \) and \( J \). Define \( \| J \| = d(0, J) \) to be the distance from \( J \) to the root. Then
\[
G(z) \geq z^{\mathbb{E}D} \geq \max(z^{d(T)}, z^{2\mathbb{E}\|J\|}) \geq \min(z^{d(T)}, z^{2\mathbb{E}\|J\|}) \geq z^{2h(T)},
\]
where \( \mathbb{E}\|J\| \) is the average distance from the seed node, \( d(T) \) is the diameter of the \( T \), and \( h(T) \) is the height of the tree.

Define \( \beta = 1/\lambda_2^2 \). If \( T \) is an \( m \)-tree, with \( m > \beta \), then
\[
G(\lambda_2) \geq \lambda_2^{2h(T)} \geq \lambda_2^{2\log_m n} = n^{-\log_m \lambda_2^2},
\]
where the second inequality follows from the fact that when \( T \) is an \( m \)-tree, \( h(T) \leq \log_m n \). This implies that the design effect grows when \( m > \beta \). However, for \( m < \beta \), this lower bound is not tight.

**Fact 4.2.** When \( z > 0 \), \( G(z) \geq n^{-1} \).

This bound matches the bound in inequality (12) when \( m = \beta \). Together Facts 4.1 and 4.2 show that when \( m < \beta \), the design effect does not converge to zero, and when \( m > \beta \) the design effect grows with \( n \). Said another way, when \( m < \beta \), the estimators do not converge faster than the standard \( \sqrt{n} \) rate, and when \( m > \beta \), the estimators converge slower than the standard \( \sqrt{n} \) rate. These are lower bounds. The next section gives a matching upper bound under an additional “balanced” condition on \( T \).
4.2 Upper bound for $G(\lambda_2)$

Upper bounding $G$ requires a more global assumption about the “balance” of $T$. Note that $G(\lambda_2)$ is small when $d(I, J)$ is likely to be big (i.e. $\hat{\mu}$ has a smaller variance when most distances are large). Define $\|I\| = d(0, I)$ to be the distance from the root to node $I$. For $\tau, \sigma \in T$, define $\tau \wedge \sigma \in T$ to be the most recent common ancestor of $\sigma$ and $\tau$. The formula

$$d(I, J) = \|I\| + \|J\| - 2\|I \wedge J\|$$

shows that most pairwise distances $d(I, J)$ are large when $\|I\|$ is large for most nodes and when $\|I \wedge J\|$ is small for most pairs. In essence, the balanced condition is a way of ensuring that $\|I \wedge J\|$ is small.

For $\sigma \in T$, define $\mathcal{A}(\sigma)$ as the set of ancestors of $\sigma$, that is the nodes in $T$ that fall along the shortest path between $\sigma$ and the root (for convenience, include $\sigma \in \mathcal{A}(\sigma)$). Define the descendants of $\tau \in T$ in the $n$th generation as

$$D_n(\tau) = \{\sigma : d(0, \sigma) = n \text{ and } \tau \in \mathcal{A}(\sigma)\}.$$ 

Because $0$ is the root node, $D_n(0)$ contains all nodes in generation $n$ and $|D_n(0)|$ is the number of nodes in generation $n$. A tree $T$ grows at rate $m$ if there exist positive constants $c$ and $\bar{c}$ such that for all $n$,

$$c \leq \frac{|D_n(0)|}{m^n} \leq \bar{c}.$$ 

Notice that this implies $T$ is an infinite tree. The results below study $T_h$, the induced subgraph of $T$ that is formed by all nodes $\tau$ with $\|\tau\| \leq h$.

Suppose that $T$ grows at rate $m$. For $\tau \in T$ with $\|\tau\| = k$, define

$$c_{\tau} = \sup_n \frac{|D_n(\tau)|}{m^{n-k}}.$$ 

Because $|D_n(\tau)| \leq |D_n(0)|$ and the tree is assumed to grow at rate $m$, these constants are finite; $c_{\tau} \leq \bar{c} m^k < \infty$. However, under a sequence of $\tau_n$, $c_{\tau_n}$ could be unbounded. A tree satisfies the balanced assumption if there exists a constant $c$ such that for all $n$,

$$|D_n(0)|^{-1} \sum_{\|\tau\|=n} c_{\tau}^2 \leq c < \infty.$$ 

That is, the second moment of the $c_{\tau}$’s is uniformly bounded across all generations. For example, $m$-trees grow at rate $m$ and satisfy the balanced assumption because $c_{\tau} = 1$ for all $\tau$.

**Theorem 4.1.** Let $T$ be an infinite tree that grows at rate $m$. Define $T_h$ as the node induced subgraph of $T$ that contains all nodes $\tau \in T$ satisfying $\|\tau\| \leq h$. Define $G_h$ as in Definition 2 with tree $T_h$. If $T$ satisfies the balanced assumption, then

$$G_h(\lambda_2) \leq \begin{cases} 
  c (\log n) n^{-1} & \text{if } m < \beta \\
  c (\log n)^2 n^{-1} & \text{if } m = \beta \\
  c (\log n) n^{-\alpha} & \text{if } m > \beta,
\end{cases}$$

where $\beta = \lambda_2^{-2}$, $\alpha = \log_m \lambda_2^{-2}$, and $c$ is a constant that could depend on $m$ and $\lambda_2$, but is independent of $n$. 

13
The growth rate assumption implies that $T_h$ has $n = O(m^h)$ nodes. So, Fact 4.1 yields matching lower bounds; the $\beta$ threshold is identical and the rates differ only by $\log n$ terms.

The upper bound in Theorem 4.1 comes from upper bounding $\mathbb{P}(d(I,J) = k)$, where $I$ and $J$ are nodes selected uniformly at random from $T$. First, condition on $\|I\|$ and $\|J\|$. Then, $d(I,J) = \|I\| + \|J\| - 2\|I \cap J\|$ is determined by $\|I \cap J\|$. In order to use the fact that $I$ and $J$ are independent,

$$\mathbb{P}(\|I \cap J\| = \ell \mid \|I\| = a, \|J\| = b) = \sum_{\tau: |\tau| = \ell} \mathbb{P}(\tau = I \cap J \mid \|I\| = a, \|J\| = b) \leq \sum_{\tau: |\tau| = \ell} \mathbb{P}(\tau \in \mathcal{A}(I) \mid \|I\| = a) \mathbb{P}(\tau \in \mathcal{A}(J) \mid \|J\| = b).$$

These terms are related to $c_2^2$. So, the balance condition provides a bound. Finally, the growth rate assumption provides bounds for $\mathbb{P}(\|I\| = a)$. Appendix D contains the full proof for Theorem 4.1.

To see the necessity of the balanced condition, suppose that a tree grows at rate $m > 1$ and every generation $t - 1$ contains a single node that produces all nodes in generation $t$. Because $m > 1$ there is a non-vanishing probability that $I$ and $J$ come from the final generation $h$. On this event, $I$ and $J$ have the same parent and $D = d(I,J) = 2$. As $h$ grows, $G$ will not decay. So, to have an upper bound that decays as the tree grows must prevent this unbalanced tree. An assumption similar to the balanced condition has appeared previously; Proposition 3.3 in Lyons [1990] implies that a balanced tree is “quasi-spherical.”

### 4.3 Galton-Watson $T$

A Galton-Watson (GW) tree is a random tree parameterized by its offspring distribution. Starting at the root node, and iterating through all future generations, each node takes an iid draw from the offspring distribution and produces this many offspring. This process is highly studied with several classical results (e.g. Athreya and Ney [1972]).

Let $\xi$ be a generic draw from the offspring distribution and denote $\mathbb{E}\xi = m$. The results from the last section identified the critical threshold $m \geq \lambda_2^{-2}$ using $m$-trees (i.e. $\text{Var}(\xi) = 0$). This section shows that this threshold also applies to GW trees with $\text{Var}(\xi) > 0$, under the assumption that $\mathbb{E}\xi^4 < \infty$.

To have a positive probability that the tree generates an infinite number of nodes, the results below require that $m > 1$. Denote $T_h$ as the sub-tree of $T$ that includes all nodes within distance $h$ of the root.

**Theorem 4.2.** [Kesten and Stigum, 1966] Suppose $T$ is a random Galton-Watson tree. Let $\xi$ be a single draw from the offspring distribution; presume that $m = \mathbb{E}(\xi) > 1$ and $\mathbb{E}(\xi \log \xi) < \infty$. Conditioned on the survival of the GW process, $T$ grows at rate $m$, a.s..

See Lyons et al. [1995] for a conceptual proof of the Kesten-Stigum Theorem.

Under the conditions of Theorem 4.2, $T_h$ has $n = O(m^h)$ nodes. As such, the function $G_h$ built from $T_h$ will have the same lower bound as $m$-trees (see the discussion after Fact 4.1). A matching upper bound on $G_h$ requires a fourth moment assumption on $\xi$.

**Lemma 4.1.** Suppose $T$ is a random Galton-Watson tree. Let $\xi$ be a single draw from the offspring distribution; presume that $m = \mathbb{E}(\xi) > 1$ and $\mathbb{E}(\xi^4) < \infty$. Conditioned on the survival of the GW process, $T$ satisfies the conditions of Theorem 4.1 (i.e. it grows at rate $m$ and it is balanced, a.s.).
In the context of RDS, the offspring distribution is typically bounded by three. As such, the assumption that $E(\xi^4) < \infty$ is certainly satisfied. The proof relies on the fact that $c_\tau$ is the supremum of the standard Galton-Watson martingale. Then, it relies upon the $L^p$ maximal inequality for martingales. The full proof is contained in the appendix, Section E. The following theorem combines the previous results, identifying a critical threshold for the design effect, $DE(\hat{\mu})$.

**Theorem 4.3.** Suppose $T$ is a random Galton-Watson tree. Let $\xi$ be a single draw for the offspring distribution; presume that $m = E(\xi) > 1$ and $E(\xi^4) < \infty$. Conditioned on the survival of the GW process. Define $T_h$ as the node induced subgraph of $T$ that contains all nodes $\tau \in T$ satisfying $\|\tau\| \leq h$. Let $P$ be a Markov transition matrix on $G$ that is reversible with respect to its stationary distribution $\pi$. Under the $(T_h, P)$-walk on $G$, let $\hat{\mu}_h$ be constructed with the samples from $T_h$. If $\rho_2^2(y, f_2) > 0$ and $\lambda_2 > 0$, then

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

(14)

where $DE$ is defined in Equation (5), $\asymp$ is equality up to $(\log n)^2$ terms, $\beta = \lambda_2^2$, and $\alpha = \log_m \lambda_2^2$.

The proof combines the previous results. It is contained in Section E of the appendix. Note that after making the appropriate substitutions, Theorem 4.3 holds under the infinite population model from Example 2.

While it would be exceedingly interesting to allow the number of referrals of node $\tau \in T$ to depend on $X_\tau$ or $Y_\tau$, potentially confounding the results, our results do not allow for this. For Theorem 2.1 to hold, $T$ must be independent of the Markov process.

## 5 The gap between sampling with and without replacement

The $(T, P)$-walk on $G$ is a Markov model for the network driven sampling mechanism. This model allows for nodes to be sampled multiple times. However, several applications of network driven sampling do not allow “with replacement sampling.” In classical sampling, when the sample size $n$ is small with respect to the population size $N$, “with replacement sampling” can approximate “without replacement sampling”. This section extends this notion to network driven sampling. This section shows that under certain conditions, the rate of resampling for the $(T, P)$-walk on $G$ is not affected by the critical threshold $m > \lambda_2^2$.

Define the number of repeated pairs as

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

The argument below studies the asymptotic behavior of $E(R_n)$ as $n$ and $N$ grow in tandem. In short, $n = o(\sqrt{N})$ (with some additional assumptions) implies that the rate of $E(R_n)$ does not depend on $\lambda_2$. This argument shows that, in expectation, the critical threshold does not create additional repeated pairs.

**Proposition 5.1.** Under the $(T, P)$-walk on $G$, suppose that $G$ is undirected and $P$ is a simple random walk. If node degrees in $G$ are less than $D$, then

$$E(R_n) \geq n/D.$$
The proof is based on the fact that if $X_\tau = i$, 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 F of the appendix.

As Proposition 5.1 shows, the $(T, P)$-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, $m > \lambda^2$. Because $R_n$ counts pairs of repeats, $\mathbb{E}(R_n)$ could grow at rate $n^2$. The next result shows that under certain conditions, this is not the case; when ignoring log terms, Theorem 5.1 provides a matching upper bound on $\mathbb{E}(R_n)$.

**Theorem 5.1.** Consider a sequence of samples $\{X_\tau : \tau \in T_n\}$ that are sampled from a $(T_n, P_N)$-walk on $G_N$, where $n$ and $N$ are both growing. Suppose that the sequence $T_n$ satisfies the conditions of Theorem 4.1; that is, there is a balanced infinite tree $T$ that grows at rate $m$ and $T_n$ 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 $\lambda_\ell$ that exceed the critical threshold $1/\sqrt{m}$ is bounded by $k$ for all $N$; and (3) $n = o(\sqrt{N})$, then

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

Notice that condition (1) is implied by the sparse graph assumption in Proposition 5.1. Importantly, this upper bound does not depend on $\lambda_2$. So, under the conditions of these results, $\lambda_2$ and the critical threshold do not effect the rate of $\mathbb{E}(R_n)$.

The key to proving Theorem 5.1 is the relationship between the trace of a matrix and its eigenvalues. First, notice that

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

Then, this probability relates to the trace of $P$.

$$\mathbb{P}(X_\sigma = X_\tau) = \sum_{i \in G} \pi_i \mathbb{P}(X_\tau = i | X_\sigma = i) = \sum_{i \in G} \pi_i P_{di}^{d(\sigma, \tau)} \leq cN^{-1} \text{tr}(P^{d(\sigma, \tau)}) = cN^{-1} \sum_\ell \lambda_\ell^{d(\sigma, \tau)}$$

Sum over $\sigma \neq \tau$ (from Equation (15)), exchange summations, and express the terms as $G$ functions. Theorem 4.1 bounds these terms. The full proof is contained in Section F of the appendix.

### 5.1 A simulation of without replacement sampling and the $(T, P)$-walk on $G$

This simulation compares the $(T, P)$-walk on $G$ to without replacement network sampling. The simulations are performed on networks simulated from the Stochastic Blockmodel. The ten panels in Figure 2 correspond to ten different networks that are simulated from ten different Stochastic Blockmodels. Each of the ten networks has $N = 10k$ nodes, equally balanced between group zero and group one. 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 $p + r$. Figure 2 controls the parameters $p$ and $r$ with the expected degree $rN + pN/2$ and the second eigenvalue of $\mathcal{P} = \mathbb{E}(D)^{-1}\mathbb{E}(A)$,

$$\lambda_2(\mathcal{P}) = \frac{1}{2(r/p) + 1}$$

where expectations are under the Stochastic Blockmodel (cf example on page 1893 of Rohe et al. [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 $\lambda_2(\mathcal{P})$.

The trees are simulated to have $m = 2$. So, Theorem 4.3 suggests that design effect should grow when $\lambda_2$ exceeds $1/\sqrt{2} \approx .7$. 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 $(T,P)$-walk on $G$ has a critical threshold somewhere between .6 and .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. This confirms that without replacement network sampling has a similar critical threshold.

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 $\lambda_2 = .6$ and increasing when $\lambda_2 = .8$). The specific simulation settings are described in Section F.1.

![With and without replacement sampling display a similar critical threshold](image)

Figure 2: In the left most plots, both lines are roughly flat. In the right most plots, both lines are quickly increasing. This shows that the $(T,P)$-walk on $G$ and without replacement network driven sampling have a critical threshold somewhere between $\lambda_2(\mathcal{P}) = .6$ and $\lambda_2(\mathcal{P}) = .8$.

## 6 Empirical illustration

Theorem 2.1 expresses the variance of the RDS derived estimator as a function of the eigenstructure of the referral process $P$ and the function $G$, which is computed from the sampling tree $T$. To illustrate the various ways that $P$ and $T$ can influence the variance, the next two subsections investigate the properties of $P$ and $G$ for previously recorded data. Subsection 6.1 examines the political blog network from Adamic and Glance 2005. Subsection 6.2 examines the function $G$ computed from 18 different referral trees that have been previously published in the RDS literature.
6.1 Eigenstructure of the political blog network

Adamic and Glance [2005] recorded the hyperlink network among popular political blogs after the 2004 US presidential election between Bush and Kerry. Numerous publications on the Stochastic Blockmodel have studied this network (e.g. Karrer and Newman [2011] and Chen et al. [2012]). The nodes in this data have a feature that indicates whether the blog supported Kerry or Bush. A connection between the blogs corresponds to a hyperlink. All edges in the networks have been made symmetric and the following analysis uses the largest connected component in the 2-core of the networks. Taking this network as $G$, this experiment samples the network with a $(T, P)$-walk on $G$. The node feature $y(i)$ is a binary variable indicating whether blog $i$ supports Kerry. The population average $\mu_{\text{true}} \approx 0.5$ is the proportion of blogs which support Kerry.

Theorem 2.1 allows for the computation of the exact design effect, conditional on the sampling tree $T$. Figure 3 provides the results for three trees sampled from the Galton-Watson distribution (see Section 4.3 for a definition). In this network, the critical threshold is $\lambda_2^{-2} \approx 1.2$. Each of the three trees has a different average number of referrals that each participant provides 1, 1.2, and 3.

![Design effects for the blog network](image)

Figure 3: When the referral rate is less than the critical threshold ($m = 1$, solid line), then the design effect does not grow. When the referral rate exceeds the critical threshold ($m = 2$, short dashed line), the design effect grows with the sample size. In the middle, the referral rate matches the critical threshold.

Figure 3 shows very large design effects. The reason for this is explored in Figure 4 which displays (1) the network of political blogs, with nodes colored by the politician that they support, and (2) the leading eigenvectors $f_2, \ldots, f_6$ of the simple random walk Markov transition matrix $P$. To emphasize the structure of the eigenvectors, the figures dichotomize the individual elements into +/- values, represented by black and white. The title of each plot gives the corresponding eigenvalue. Because the Kerry-Bush divide creates a strong bottleneck (only about 10% of edges cross the political partition), the bottleneck is represented in the leading eigenspace of the $P$. As such, the Volz-Heckathorn estimator for “the proportion of blogs which are Kerry blogs” is likely to have a large standard error.
The leading eigenvectors of $P$ on the political blog network

Figure 4: The left panel displays the political blog network. The nodes on the left are Kerry blogs (colored blue). The nodes on the right are Bush blogs (colored red). The remaining three panels display the second, fourth, and fifth eigenvectors of the Markov transition matrix corresponding to the simple random walk. For easier visualization, each vector is dichotomized into +/- parts represented as black and white on the nodes. After this dichotomization, the third eigenvector appears similar to the second, but is driven by a few outliers. For this reason, it is not plotted. The second eigenvalue is .89 and its eigenvector has correlation .82 with $y$, the indicator of Kerry blogs. This suggests that RDS derived estimators for the proportion of Kerry blogs will have a large variance. Three outlying nodes have been removed for easier visualization.

6.2 The design effect with empirical referral trees

In practice, the function $G$ is exactly computable because $T$ is observed. Using 18 trees that were published in previous RDS papers, Figure 5 presents the function $G$ as a function of $\lambda_2 \in [0, 1]$. To ensure a fair comparison across different sample sizes, the vertical axis represents the design effect under the assumption that $y = f_2$, so that the design effect is exactly $nG(\lambda_2)$ (cf Example 1). In practice, the design effect could be estimated by creating an estimate for $\lambda_2$ (cf Section 3).

The legend gives the number of samples in each tree. 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 et al., 2015]. The remaining ten trees come from a study of 25 villages in rural Uganda [McCreesh et al., 2012].

As expected, all of the lines are upward sloping, indicating that stronger bottlenecks produce larger design effects. Moreover, the larger trees are more sensitive to the bottleneck strength.

The right panel of Figure 5 further illustrates this result. It presents the function $nG(\lambda_2)$ for a set of 2-trees, where every node has exactly two offspring. Here again, the larger trees are more sensitive to stronger bottlenecks; the design effect explodes when $\lambda_2 > 1/\sqrt{2}$. However, so long as $\lambda_2 < 1/\sqrt{2} \approx .7$, the design effect is largely insensitive to the bottleneck.

7 Discussion

This paper studies network driven sampling using the $(T, P)$-walk on $G$. Under this approach, only the referrals are random. The network and node features are fixed. Theorem 2.1 gives a closed form expression for the variance of $\hat{\mu}$ when it is constructed from a network driven sample. This
Figure 5: The left panel presents the design effect $nG(\lambda_2)$ for 18 empirical referral trees. The right panel presents the same results for several 2-trees. Each line corresponds to a tree. On the horizontal axis, $\lambda_2$ represents the strength of a potential bottleneck. The legend gives the number of nodes in each tree. All lines are upward sloping. Moreover, larger trees are more sensitive to larger values of $\lambda_2$. In practice, one must estimate $\lambda_2$. The right panel illustrates how 2-trees have a critical threshold at $\lambda_2 = 1/\sqrt{2} \approx .7$. 
expression combines the following three elements: (1) the eigenfunctions and eigenvalues $f_\ell, \lambda_\ell$ of the Markov transition matrix $P$, (2) the covariance between the node feature and the eigenfunctions $\langle y, f_\ell \rangle_\pi$, and (3) the bushiness of the referral tree as measured by the function $G$. Corollary 2.1 shows that if $\rho_\pi^2(y, f_2) > 0$, then the design effect is asymptotically proportional to $nG_n(\lambda_2)$. Using two example models, Section 3 proposes two paths towards estimating the variance from observed data.

The formula for $\text{Var}_{RDS}(\hat{\mu})$ in Theorem 2.1 is conditional on the sampling tree $T$. To study the asymptotic behavior of the design effect requires that the tree grows. The results in Section 4 identify the fundamental features of the tree that characterize this asymptotic behavior. Under the assumption that the tree is balanced, the growth rate $m$ determines the asymptotic rates for the design effect. The specifics of the tree only appear in the constants. When the tree is balanced and grows at rate $m$, the upper and lower bounds identify the same critical threshold. If $m < \beta = \lambda_2^2$, then ignoring log terms, the design effect is $O(1)$. However, if $m > \beta$, then ignoring log terms, the design effect is $O(n^{1-\alpha})$, where $\alpha = \log_m \beta \in (0, 1)$. In order for the upper and lower bound to match, the balanced assumption is necessary. To understand the stringency of the balanced condition, Section 4.3 shows that if $T$ is a random Galton-Watson tree with $E\xi^4 < \infty$, where $\xi$ is a draw from the offspring distribution, then the tree is balanced a.s.. The lower bound does not require the balanced assumption; the classical Kesten-Stigum theorem shows that the lower bound only requires $E\xi \log \xi < \infty$.

Section 5 examines how well the “with replacement” $(T, P)$-walk on $G$ approximates a “without replacement” network driven sampling mechanism. Proposition 5.1 and Theorem 5.1 give matching lower and upper bounds on $E(R_n)$, the expected number of repeated pairs in a $(T, P)$-walk on $G$. So long as $n = o(\sqrt{N})$, and some further technical conditions, These bounds show that $\lambda_2$ and the critical threshold do not affect the rate of $E(R_n)$. As such, the critical threshold does not create additional repeated pairs. Subsection 5.1 presents a simulation comparing the $(T, P)$-walk on $G$ to a without replacement network driven sample. Under the simulation settings, both the with replacement and without replacement samples displayed a similar critical threshold.

To illustrate the analytic results for the $(T, P)$-walk on $G$, Section 6 studies simulated and empirical data. Subsection 6.1 uses $(T, P)$-walk on $G$ to sample the political blog network. Because the node covariate is strongly related to the leading eigenspace, network driven estimators have an excessively large design effect. Figure 5 in Subsection 6.2 gives the design effect for synthetic 2-trees and empirically observed referral trees from a wide range of RDS studies. In both, the larger trees display a greater sensitivity to the bottlenecks in the social network.

Taken together, this research presents a disheartening bias-variance tradeoff that arises in the process of data collection. Applications of RDS initialize with a convenience sample (i.e. not the stationary distribution). So, estimators are potentially biased. However, because the process converges to the stationary distribution exponentially fast in the number of steps, the VH estimator is asymptotically unbiased. For this reason, practitioners strive for longer chains. To obtain long chains, while ensuring that the process does not die, many studies obtain one very large recruitment tree; often there are several other trees which are much smaller. In this respect, the critical threshold and Figure 5 present a disheartening conflict; the larger trees are more sensitive to the strength of the bottleneck $\lambda_2$ and they have larger design effects. While longer chains are preferred for an unbiased sample, if these long chains are achieved by a bushy structure in $T$, then the variance can be excessive relative to the number of sample obtained. As such, there is a fundamental conflict between obtaining an unbiased sample and small standard errors. Goel and Salganik [2009] realized
this conflict under a 2-tree and a simplified network model. Future research should study various
data collection procedures to manage this conflict. For example, both the bias and variance would
decrease if (1) implementation prevented T from growing geometrically, while (2) ensuring that the
referral process does not go extinct.
## 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 \( \{X(i) : i \in 0, 1, \ldots \} \) be a Markov chain with the same transition matrix \( P \) that is initialized from \( \pi \). Define \( d(\sigma, \tau) \) as the graph distance between nodes \( \sigma \) and \( \tau \) in \( \mathbb{T} \).

**Lemma A.1.** If the transition matrix \( P \) is reversible, then for any two nodes \( \sigma \) and \( \tau \) in the referral tree,

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

**Proof.** Let \( p = \sigma \land \tau \) be the most recent common ancestor of \( \sigma \) and \( \tau \). By the reversibility of the process,

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

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

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

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

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

The following is a proof of Theorem 2.1.

**Proof.**

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

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

For ease of notation, let \( t = d(\sigma, \tau) \). From Lemma A.1 (and suppressing the \( \text{RDS} \) 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 1.1), with the fact that \( f_1 \) is a constant vector and \( \lambda_1 = 1 \) [Levin et al., 2009],

#### 23
\[
E(y(X(0))y(X(t))) = \sum_{u,v \in G} y(u)y(v)P(X(0) = u, X(t) = v) \\
= \sum_{u,v \in G} y(u)y(v)\pi_u\pi_v \sum_{\ell=1}^N \lambda^j_{\ell}f_\ell(u)f_\ell(v) \\
= \sum_{u,v \in G} y(u)y(v)\pi_u\pi_v \left(1 + \sum_{\ell=2}^N \lambda^j_{\ell}f_\ell(u)f_\ell(v)\right) \\
= \left(\sum_{u \in G} y(u)\pi_u\right)^2 + \sum_{\ell=2}^N \lambda^j_{\ell} \left(\sum_{u \in G} y(u)\pi_u f_\ell(u)\right)^2 \\
= (Eg(X(0)))^2 + \sum_{\ell=2}^N \lambda^j_{\ell} \langle y, f_\ell \rangle^2 \pi.
\]

Terms cancel. So,

\[
Cov(y(X_\sigma), y(X_\tau)) = \sum_{\ell=2}^N \lambda^d(\sigma, \tau) \langle y, f_\ell \rangle^2 \pi.
\tag{16}
\]

Then,

\[
Var_{RDS}(\hat{\mu}) = n^{-2} \sum_{\sigma, \tau \in T} Cov(y(X_\sigma), y(X_\tau)) \\
= n^{-2} \sum_{\sigma, \tau \in T} \sum_{\ell=2}^N \lambda^d(\sigma, \tau) \langle y, f_\ell \rangle^2 \pi \\
= n^{-2} \sum_{\ell=2}^N \langle y, f_\ell \rangle^2 \pi \sum_{\sigma, \tau \in T} \lambda^d(\sigma, \tau) \\
= \sum_{\ell=2}^N \langle y, f_\ell \rangle^2 \pi \mathcal{G}(\lambda_\ell).
\]

\[\square\]

B Proof of Corollary 2.1

The proof of Corollary 2.1 uses two lemmas.

Lemma B.1. For \(\sigma^2 = Var_{RDS}(Y_0)\),

\[
\sigma^2 = \sum_{j=2}^N \langle y, f_j \rangle^2 \pi.
\]
This proof is given on page 342 of Levin et al. [2009] and is repeated here for completeness.

Proof.
\[
\sum_{j=2}^{N} \langle y, f_j \rangle_{\pi}^2 = \sum_{j=1}^{N} \langle y, f_j \rangle_{\pi}^2 - (\mathbb{E}_{\pi} Y_0)^2 = \mathbb{E}_{\pi}(Y_0^2) - (\mathbb{E}_{\pi} Y_0)^2 = \text{Var}_{\pi}(Y_0) = \text{Var}_{RDS}(Y_0).
\]

Lemma B.2. For any \((T, P)\)-walk on \(G\) that satisfies the conditions of Theorem 2.1, \(G(\lambda_k) \geq 0\) for all \(k = 1, \ldots, N\).

Proof. Define \(y = f_k\), then \(0 \leq \text{Var}_{RDS}(\hat{\mu}) = G(\lambda_k)\). For \(k = 1\), \(G(1) = 1\).

Now, a proof of Corollary 2.1

Proof. By the definition of the ordering, \(|\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_N|\) and the assumption \(\lambda_2 > 0\), it follows that \(\lambda_2 \geq |\lambda_\ell|\) for \(\ell > 2\). This implies \(\lambda_2^d > \lambda_\ell^d\) for any \(d\). It then follows that \(G(\lambda_2) \geq G(\lambda_\ell)\). So,

\[
\text{Var}_{RDS}(\hat{\mu}) = \sum_{\ell=2}^{N} \langle y, f_\ell \rangle_{\pi}^2 G(\lambda_\ell) \leq G(\lambda_2) \sum_{\ell=2}^{N} \langle y, f_\ell \rangle_{\pi}^2 = G(\lambda_2) \sigma^2.
\]

Because \(G(\lambda_\ell) \geq 0\) for all \(\ell\),

\[
\text{Var}_{RDS}(\hat{\mu}) = \sum_{\ell=2}^{N} \langle y, f_\ell \rangle_{\pi}^2 G(\lambda_\ell) \geq \langle y, f_2 \rangle_{\pi}^2 G(\lambda_2).
\]

To convert to DE, divide by \(\text{Var}_{\pi}(\hat{\mu}) = \text{Var}_{RDS}(Y_0)/n = \sigma^2/n\).

C Proof of lower bounds

The following is a proof of Fact 4.1

Proof. The first inequality follows directly from Jensen’s inequality. The next inequalities use

\[
\mathbb{E}D \leq \mathbb{E}(\|I\| + \|J\|) = 2\mathbb{E}\|J\| \leq 2h(T).
\]

Also, notice that \(\mathbb{E}D \leq d(T) \leq 2h(T)\). The result follows from the restriction that \(|z| < 1\).

The following is a proof of Fact 4.2

Proof. As before, denote \(D = d(I, J)\).

\[
G(z) = \sum_{k=0}^{d} z^k \mathbb{P}(D = k) \geq z^0 \mathbb{P}(D = 0) = n^{-1}.
\]
D Proof of upper bounds

The proof will use the following fact about a finite geometric series:

\[
\sum_{k=0}^{2h} x^k = \frac{1 - x^{2h+1}}{1 - x} \leq \begin{cases} 
(1 - x)^{-1} & \text{if } x < 1 \\
 x^{2h+1}(x - 1)^{-1} & \text{if } x > 1.
\end{cases}
\]  

(17)

The following is a proof of Theorem 4.1.

Proof. An upper bound on \(P(d(I, J) = k)\) provides an upper bound on \(G_h(\lambda_2)\).

\[
P(d(I, J) = k) = \sum_{j=0}^{k} P(d(I \land J, I) = k - j \cap d(I \land J, J) = j)
= \sum_{j=0}^{k} \sum_{\ell=0}^{[h-k/2]} P(d(I \land J, I) = k - j \cap d(I \land J, J) = j \mid \|I\| = k - j + \ell, \|J\| = j + \ell) \times P(\|I\| = k - j + \ell)P(\|J\| = j + \ell).
\]

First, bound the terms on \(\|I\|\) and \(\|J\|\) with the growth rate assumption,

\[
P(\|I\| = k - j + \ell)P(\|J\| = j + \ell) \leq a^2b^2m^{k+2\ell-2h}.
\]

To bound the \(I \land J\) term, define

\[
c_{\tau,n} = \frac{\mathcal{D}_{\|\tau\|+n(\tau)}}{m^n}.
\]

A key idea in what follows is that \(\tau = I \land J \implies \tau \in \mathcal{A}(I) \cap \mathcal{A}(J)\). Then, because \(I\) and \(J\) are independent, this probability breaks apart into two terms.

\[
P(d(I \land J, I) = k - j \cap d(I \land J, J) = j \mid \|I\| = k - j + \ell, \|J\| = j + \ell)
= \sum_{\tau:|\tau|=\ell} P(\tau = I \land J \mid \|I\| = k - j + \ell, \|J\| = j + \ell)
\leq \sum_{\tau:|\tau|=\ell} P(\tau \in \mathcal{A}(I) \mid \|I\| = k - j + \ell)P(\tau \in \mathcal{A}(J) \mid \|J\| = j + \ell)
= \sum_{\tau:|\tau|=\ell} \frac{\mathcal{D}_{k-j+\ell}(\tau)}{\mathcal{D}_{k-j+\ell}(0)} \frac{\mathcal{D}_{j+\ell}(\tau)}{\mathcal{D}_{j+\ell}(0)}
\leq a^2 \sum_{\tau:|\tau|=\ell} c_{\tau,k-j}c_{\tau,j}m^{-2\ell}.
\]

By the definition of \(c_{\tau}\),

\[
\sum_{j=0}^{k} c_{\tau,k-j}c_{\tau,j} \leq kc_{\tau}^2.
\]
So,
\[
P(d(I, J) = k) \leq c \sum_{j=0}^{k} \sum_{\ell=0}^{[h-k/2]} \sum_{\tau:|\tau|=\ell} c_{\tau,k-j} c_{\tau,j} m^{-2\ell} m^{k+2\ell-2h}
\]
\[
\leq cm^{k-2h} \sum_{\ell=0}^{[h-k/2]} m^\ell (D_\ell(0))^{-1} \sum_{\tau:|\tau|=\ell} c_\tau^2.
\]

By the balanced assumption, there is a constant \(c < \infty\) such that for all \(\ell\),
\[
(D_\ell(0))^{-1} \sum_{\tau:|\tau|=\ell} c_\tau^2 < c.
\]

So, use equation (17) and let the constant depend on \(m\),
\[
P(d(I, J) = k) \leq cm^{k-2h} \sum_{\ell=0}^{[h-k/2]} m^\ell
\]
\[
\leq ckm^{k-2h} m^{h-k/2+1}
\]
\[
= ckm^{k/2-h}.
\]

By the growth rate assumption, \(m^{-h} \leq cn^{-1}\) and \(h \leq c \log n\). So,
\[
G_h(z) = \sum_{k=0}^{2h} z^k P(d(I, J) = k)
\]
\[
\leq c \sum_{k=0}^{2h} z^k km^{k/2-h}
\]
\[
\leq cn^{-1} \log n \sum_{k=0}^{2h} (\sqrt{mz})^k.
\]

When \(mz^2 = 1\), the sum contributes \(2h \leq c \log n\) and the rate is \(n^{-1}(\log n)^2\). Using the fact about geometric series in equation (17), for \(mz^2 \neq 1\),
\[
G_h(z) \leq cn^{-1} \log n (\sqrt{mz})^{2h+1}.
\]

When \(mz^2 < 1\), the leading term gives the rate because the fraction converges to a constant. However, when \(mz^2 > 1\), the fraction explodes.
\[
n^{-1} (\sqrt{mz})^{2h+1} \sqrt{mz} - 1 \leq cn^{-1} (mz^2) h \leq cz^{2h} = cz^2 \log m n = cn^2 \log m z.
\]

\[\square\]
The following is a proof of Lemma 4.1.

**Proof.** Because trees that go extinct are balanced, it is not necessary to condition on survival. The proof below shows that if $T$ is generated from the GW with a finite fourth moment, then it is balanced a.s..

Each node $\tau \in T$ generates an identically distributed GW tree below it. Denote

$$Z_n^\tau = |\mathcal{D}_{\|\tau\|+n}(\tau)|, \quad W_n^\tau = \frac{Z_n^\tau}{m_n}, \quad W_n^\tau = \sup_n W_n^\tau, \text{ and } W^\tau = \lim W_n^\tau.$$  

Across all values of $\tau$, $W_n^\tau$ are identically distributed. Moreover, within a single generation of the tree (i.e. $\tau \in \mathcal{D}_h(0)$), $W_n^\tau$ are independent. The same holds for $Z_n^\tau, W_n^\tau$ and $W^\tau$. So, dropping the superscript $\tau$ will correspond to a generic iid draw from the same distribution.

The values $W_n^\tau$ correspond to the $c_\tau$’s in the balanced assumption. We wish to bound

$$|\mathcal{D}_h(0)|^{-1} \sum_{\|\tau\|=h} (W_n^\tau)^2 < C(\omega),$$

where $C$ is a random variable that does not depend on $h$.

**Lemma E.1.** Under the conditions of the theorem, $\mathbb{E}W_+^4 < \infty$.

A proof of this lemma is given following the proof of the theorem.

Using Borel-Cantelli, the argument below will show that for $\mu = \mathbb{E}W_+^2$ and $\epsilon > 0$,

$$\mathbb{P} \left( \left\{ |\mathcal{D}_h(0)|^{-1} \sum_{\|\tau\|=h} (W_n^\tau)^2 > \mu + \epsilon \right\} \text{ i.o. in } h \right) = 0. \quad (18)$$

As such, a.s. there exists a variable $C(\omega)$ that satisfies the balanced condition.

Denote $Z_k = Z_k^0$. Let $\mathcal{F}_h$ denote the filtration $\sigma(Z_0, Z_1, \ldots, Z_h)$. By Chebyshev’s inequality,

$$\sum_{h=1}^\infty \mathbb{P} \left( \sum_{\|\tau\|=h} (W_n^\tau)^2 > Z_{h-1}(\mu + \epsilon) \right)
= \sum_{h=1}^\infty \mathbb{E} \mathbb{P} \left( \sum_{\|\tau\|=h} (W_n^\tau)^2 > Z_{h-1}(\mu + \epsilon) \mid \mathcal{F}_{h-1} \right)
\leq \sum_{h=1}^\infty \mathbb{E} \left\{ 1\{Z_{h-1} \neq 0\} \sum_{\|\tau\|=h} (W_n^\tau)^2 - \mu \right\} \mathbb{E} \left( \frac{((W_n^\tau)^2 - \mu)^2}{Z_{h-1}^2 \epsilon^2} \mid \mathcal{F}_{h-1} \right)
= \sum_{h=1}^\infty \mathbb{E} \left\{ 1\{Z_{h-1} \neq 0\} \mathbb{E} \left( \frac{(W_n^2 - \mu)^2}{Z_{h-1}^{\epsilon^2}} \mid \mathcal{F}_{h-1} \right) \right\}. \quad (19)$$

Then,

$$\mathbb{E} \left( \frac{(W_n^2 - \mu)^2}{Z_{h-1}^{\epsilon^2}} \right) \leq \mathbb{E}(W_+^4) < c \epsilon^2 < \infty.$$
Using this to bound equation (19),

$$\leq c \sum_{h=1}^{\infty} \mathbb{E} \left( 1 \{ Z_{h-1} \neq 0 \} Z_{h-1}^{-1} \right).$$

By Theorem 1 in Ney and Vidyashankar [2003], there exists some constant $\rho < 1$ and some other constant $c$ such that

$$\mathbb{E} \left( 1 \{ Z_{h} \neq 0 \} Z_{h}^{-1} \right) \leq c \rho^h.$$

Because this is a summable sequence, Borel-Cantelli implies the desired result.

The following is a proof of Lemma E.1

Proof. Define

$$W_{+n} = \max_{0 \leq m \leq n} W_m.$$

By the Monotone Convergence Theorem, $\mathbb{E} W_{+n}^4 \to \mathbb{E} W_{+}^4$. So, it is enough to show that $\sup_n \mathbb{E} W_{+n}^4 < \infty$.

By the $L^p$ maximum inequality (e.g. Theorem 5.4.3 in Durrett [April 21, 2013]), $W_n = \mathbb{E}(W|F_n)$, and Jensen’s inequality,

$$\mathbb{E} W_{+n}^4 \leq c \mathbb{E} W_{+}^4 = c \mathbb{E}(\mathbb{E}(W|F_n)^4) \leq c \mathbb{E}(W^4).$$

Bingham and Doney [1974] shows that $\mathbb{E}(\xi^4) < \infty$ implies $\mathbb{E}(W^4) < \infty$, concluding the proof.

Next a proof of Theorem 4.3.

Proof. First, a proof of the upper bound. From Corollary 2.1, $DE \leq n \mathbb{G}_h(\lambda_2)$. By the Kesten-Stigum Theorem, $\mathbb{T}$ grows at rate $m$. From Lemma 4.1 $\mathbb{T}$ is balanced. So, Theorem 4.1 gives upper bounds for $\mathbb{G}_h$. Multiplying the bounds by $n$ yields the upper bound on $DE$ given in Equation (14).

For the lower bound, $DE \geq cn \mathbb{G}_h(\lambda_2)$ for a generic positive constant, $c > 0$. By Fact 4.1, $\mathbb{G}_h(\lambda_2) \geq \lambda_2^h$. By the Kesten-Stigum Theorem, $\mathbb{T}$ grows at rate $m$. So, $n \geq c \sum_k m^k \geq cm^h$. So, $h \leq \log_m n - c$. Performing the algebra analogous to Equation (12), yields $\mathbb{G}_h(\lambda_2) \geq cn^{-\log_m \lambda_2^{-2}}$. Multiplying by $n$ and combining this with Fact 4.2 yields the lower bound.

F Sampling with replacement results in Section 5

The following is a proof of Proposition 5.1.

Proof. Let $\sigma'' \in \mathbb{T}$ denote a node that is distance two away from $\sigma \in \mathbb{T}$. Let $\sigma' \in \mathbb{T}$ be the intermediate node between $\sigma$ and $\sigma''$. Because $G$ is undirected and $P$ is a simple random walk, $P$ is reversible. So, the direction of the edges between $\sigma, \sigma'$, and $\sigma''$ does not matter.

$$\mathbb{E}(R_n) = \sum_{\sigma \neq \tau} \mathbb{P}(X_\sigma = X_\tau) \geq \sum_{\sigma} \mathbb{P}(X_\sigma = X_{\sigma''}) = \sum_{\sigma} \mathbb{E} \frac{1}{\text{deg}(X_{\sigma'})} \geq \sum_{\sigma} \frac{1}{D} = \frac{n}{D}.$$
The following is a proof of Theorem 5.1

Proof. Using properties of the trace function,

\[
\mathbb{P}(X_\sigma = X_\tau) = \sum_{i \in G} \mathbb{P}(X_\sigma = i)\mathbb{P}(X_\tau = i | X_\sigma = i)
\]

(20)

\[
= \sum_{i \in G} \pi_i P^{d(\sigma, \tau)}_{ii}
\]

(21)

\[
\leq cN^{-1} \text{tr}(P^{d(\sigma, \tau)})
\]

(22)

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

(23)

Then,

\[
\mathbb{E}(R_n) = \sum_{\sigma \neq \tau} \mathbb{P}(X_\sigma = X_\tau)
\]

(24)

\[
= cN^{-1} \sum_\ell \sum_{\sigma \neq \tau} \lambda_\ell^{d(\sigma, \tau)}
\]

(25)

\[
= cN^{-1} \sum_\ell (n^2 G_\ell - n)
\]

(26)

\[
= cN^{-1} n^2 \left( G(\lambda_1) + \sum_{\ell \in \mathcal{A}} G(\lambda_\ell) + \sum_{\ell \in \mathcal{B}} G(\lambda_\ell) - cN/n \right),
\]

(27)

where

\[
\mathcal{A} = \{ \ell > 1 : |\lambda_\ell| \geq m^{-1/2} \} \text{ and } \mathcal{B} = \{ \lambda : |\lambda_\ell| < m^{-1/2} \}.
\]

From properties of the Markov transition matrix, \(\lambda_1 = 1\). So, \(G(\lambda_1) = 1\). By assumption (2), \(|\mathcal{A}| \leq k\) for some constant \(k\). By Theorem 4.1, \(\ell \in \mathcal{A}\) implies \(G(\lambda_\ell) = O((\log n)n^{-\alpha})\) for \(\alpha = \log_m \lambda_2^{-2} > 0\). Similarly, \(\ell \in \mathcal{B}\) implies \(G(\lambda_\ell) = O((\log n)n^{-1})\). Substituting these values,

\[
\mathbb{E}(R_n) \leq \frac{cn^2}{N} + k(\log n) \frac{cn^2}{N} + c(\log n)n.
\]

By assumption (3), the first two terms converge to zero, leaving the third and final term which yields the result. \(\square\)

F.1 Simulation settings for Figure 2

Figure 2 compares the \((T, P)\)-walk on \(G\) to without replacement network driven sampling. The simulation first generates \(T\) as a Galton-Watson tree with offspring distribution \(1 + \text{Binomial}(2, 1/2)\). Trees are sampled until a tree reaches 2000 nodes; while only 500 samples are kept, it will become clear why \(T\) must be initialized to be large that 500. This tree is seeded with a node selected from the stationary distribution. The simulation iterates through the tree, one node at a time in the fashion of a breadth first search, filling the future generations of \(T\) by sampling without replacement from the viable friend list; a friend is viable if it has not yet appeared in \(T\). If \(\sigma \in T\) should produce three referrals, but \(X_\sigma\) does not have that many viable friends, then the tree is pruned accordingly; all viable friends are sampled and the remaining descendants in \(T\) are removed;
this happens infrequently in the simulation. Once 500 nodes are sampled, the remaining nodes in $T$ are pruned. This pruned tree is then used to run the $(T, P)$-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, $(4n)^{-1}$.

**G Bibliography**

**References**

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

A. S. Abdul-Quader, D. D. Heckathorn, C. McKnight, H. Bramson, C. Nemeth, K. Sabin, K. Gallagher, and D. C. Des Jarlais. Effectiveness of respondent-driven sampling for recruiting drug users in new york city: findings from a pilot study. *Journal of urban health*, 83(3):459–476, 2006.

L. A. Adamic and N. Glance. The political blogosphere and the 2004 us election: divided they blog. In *Proceedings of the 3rd international workshop on Link discovery*, pages 36–43. ACM, 2005.

D. Aldous and J. Fill. Reversible markov chains and random walks on graphs. 2002. *Unfinished manuscript.*, 2002.

K. B. Athreya and P. E. Ney. *Branching processes*. Springer, 1972.

K. Avrachenkov, B. Ribeiro, and D. Towsley. Improving random walk estimation accuracy with uniform restarts. In *Algorithms and Models for the Web-Graph*, pages 98–109. Springer, 2010.

I. Benjamini and Y. Peres. Markov chains indexed by trees. *The Annals of Probability*, pages 219–243, 1994.

N. Bingham and R. Doney. Asymptotic properties of supercritical branching processes i: The galton-watson process. *Advances in Applied Probability*, pages 711–731, 1974.

A. Birnbaum. On the foundations of statistical inference. *Journal of the American Statistical Association*, 57(298):269–306, 1962.

A. Chen, A. Amini, P. Bickel, and E. Levina. Fitting community models to large sparse networks. *arXiv preprint arXiv:1207.2340*, 2012.

F. Chung. *Spectral graph theory*. Number 92. Amer Mathematical Society, 1997.

A. Dembo, P. Mörters, and S. Sheffield. Large deviations of markov chains indexed by random trees. In *Annales de l’Institut Henri Poincare (B) Probability and Statistics*, volume 41, pages 971–996. Elsevier, 2005.

R. Durrett. *Probability: theory and examples. Edition 4.1*. Cambridge university press, April 21, 2013. URL http://www.math.duke.edu/$\sim$rtd/PTE/PTE4_1.pdf
C. J. Geyer. Practical markov chain monte carlo. *Statistical Science*, pages 473–483, 1992.

K. J. Gile. Improved inference for respondent-driven sampling data with application to hiv prevalence estimation. *Journal of the American Statistical Association*, 106(493), 2011.

K. J. Gile and M. S. Handcock. Respondent-driven sampling: An assessment of current methodology. *Sociological methodology*, 40(1):285–327, 2010.

K. J. Gile and M. S. Handcock. Network model-assisted inference from respondent-driven sampling data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 2015. ISSN 1467-985X. doi: 10.1111/rssa.12091. URL [http://dx.doi.org/10.1111/rssa.12091](http://dx.doi.org/10.1111/rssa.12091).

K. J. Gile, L. G. Johnston, and M. J. Salganik. Diagnostics for respondent-driven sampling. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 178(1):241–269, 2015.

M. Gjoka, M. Kurant, C. T. Butts, and A. Markopoulou. Walking in facebook: A case study of unbiased sampling of osns. In *INFOCOM, 2010 Proceedings IEEE*, pages 1–9. IEEE, 2010.

S. Goel and M. J. Salganik. Respondent-driven sampling as markov chain monte carlo. *Statistics in medicine*, 28(17):2202–2229, 2009.

S. Goel and M. J. Salganik. Assessing respondent-driven sampling. *Proceedings of the National Academy of Sciences*, 107(15):6743–6747, 2010.

L. A. Goodman. Snowball sampling. *The annals of mathematical statistics*, pages 148–170, 1961.

M. S. Handcock and K. J. Gile. Comment: On the concept of snowball sampling. *Sociological Methodology*, 41(1):367–371, 2011.

M. S. Handcock, K. J. Gile, et al. Modeling social networks from sampled data. *The Annals of Applied Statistics*, 4(1):5–25, 2010.

D. D. Heckathorn. Respondent-driven sampling: a new approach to the study of hidden populations. *Social problems*, pages 174–199, 1997.

P. Holland, K. Laskey, and S. Leinhardt. Stochastic blockmodels: First steps. *Social Networks*, 5(2):109–137, 1983.

L. G. Johnston, Y.-H. Chen, A. Silva-Santisteban, and H. F. Raymond. An empirical examination of respondent driven sampling design effects among hiv risk groups from studies conducted around the world. *AIDS and Behavior*, 17(6):2202–2210, 2013.

G. L. Jones et al. On the markov chain central limit theorem. *Probability surveys*, 1(299-320):5–1, 2004.

B. Karrer and M. Newman. Stochastic blockmodels and community structure in networks. *Physical Review E*, 83(1):016107, 2011.

H. Kesten and B. P. Stigum. A limit theorem for multidimensional galton-watson processes. *The Annals of Mathematical Statistics*, pages 1211–1223, 1966.
M. Khabbazian, B. Hanlon, Z. Russek, and K. Rohe. Novel sampling design for respondent-driven sampling. In preparation, 2015.

A. Kohut, S. Keeter, C. Doherty, M. Dimock, and L. Christian. Assessing the representativeness of public opinion surveys. Pew Research Center, Washington, DC, 2012.

V. Krishnamurthy, M. Faloutsos, M. Chrobak, L. Lao, J.-H. Cui, and A. G. Percus. Reducing large internet topologies for faster simulations. In NETWORKING 2005. Networking Technologies, Services, and Protocols; Performance of Computer and Communication Networks; Mobile and Wireless Communications Systems, pages 328–341. Springer, 2005.

M. Kurant, A. Markopoulou, and P. Thiran. On the bias of bfs (breadth first search). In Teletraffic Congress (ITC), 2010 22nd International, pages 1–8. IEEE, 2010.

J. Leskovec and C. Faloutsos. Sampling from large graphs. In Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 631–636. ACM, 2006.

D. A. Levin, Y. Peres, and E. L. Wilmer. Markov chains and mixing times. American Mathematical Soc., 2009.

R. Lyons. Random walks and percolation on trees. The Annals of Probability, 18(3):pp. 931–958, 1990. ISSN 00911798. URL http://www.jstor.org/stable/2244410

R. Lyons and Y. Peres. Probability on trees and networks. 2005. URL http://mypage.iu.edu/~rdlyons/prbtree/prbtree.html

R. Lyons, R. Pemantle, Y. Peres, et al. Conceptual proofs of l log l criteria for mean behavior of branching processes. The Annals of Probability, 23(3):1125–1138, 1995.

N. McCreesh, S. Frost, J. Seeley, J. Katongole, M. N. Tarsh, R. Ndunguse, F. Jichi, N. L. Lunel, D. Maher, L. G. Johnston, et al. Evaluation of respondent-driven sampling. Epidemiology (Cambridge, Mass.), 23(1):138, 2012.

W. W. Neely. Statistical Theory for Respondent-Driven Sampling. PhD thesis, University of Wisconsin–Madison, 2009.

P. E. Ney and A. N. Vidyashankar. Harmonic moments and large deviation rates for supercritical branching processes. Annals of Applied Probability, pages 475–489, 2003.

Y. Peres. Probability on trees: an introductory climb. In Lectures on probability theory and statistics, pages 193–280. Springer, 1999.

B. Ribeiro and D. Towsley. Estimating and sampling graphs with multidimensional random walks. In Proceedings of the 10th ACM SIGCOMM conference on Internet measurement, pages 390–403. ACM, 2010.

K. Rohe, B. Yu, and S. Chatterjee. Spectral clustering and the high dimensional stochastic block-model. The Annals of Statistics, 39(4):1878–1915, 2011.

M. J. Salganik. Variance estimation, design effects, and sample size calculations for respondent-driven sampling. Journal of Urban Health, 83(1):98–112, 2006.
M. J. Salganik and D. D. Heckathorn. Sampling and estimation in hidden populations using respondent-driven sampling. *Sociological methodology*, 34(1):193–240, 2004.

D. Stutzbach, R. Rejaie, N. Duffield, S. Sen, and W. Willinger. On unbiased sampling for unstructured peer-to-peer networks. *IEEE/ACM Transactions on Networking (TON)*, 17(2):377–390, 2009.

C. L. Szwarcwald, P. R. B. de Souza Júnior, G. N. Damacena, A. B. Junior, and C. Kendall. 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 Journal of Acquired Immune Deficiency Syndromes*, 57:S129–S135, 2011.

A. M. Verdery, T. Mouw, S. Bauldry, and P. J. Mucha. Network structure and biased variance estimation in respondent driven sampling. *arXiv preprint arXiv:1309.5109*, 2013.

E. Volz and D. D. Heckathorn. Probability based estimation theory for respondent driven sampling. *Journal of Official Statistics*, 24(1):79, 2008.

U. Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007.

W. Yang. Some limit properties for markov chains indexed by a homogeneous tree. *Statistics & Probability Letters*, 65(3):241–250, 2003.