

Respondent-driven sampling as Markov chain Monte Carlo

Sharad Goel^{1, †} and Matthew J. Salganik^{2, *, †}

¹*Yahoo! Research, 111 W. 40th Street, New York, NY 10018, U.S.A.*

²*Department of Sociology and Office of Population Research, Princeton University, Princeton, NJ 08544, U.S.A.*

SUMMARY

Respondent-driven sampling (RDS) is a recently introduced, and now widely used, technique for estimating disease prevalence in hidden populations. RDS data are collected through a snowball mechanism, in which current sample members recruit future sample members. In this paper we present RDS as Markov chain Monte Carlo importance sampling, and we examine the effects of community structure and the recruitment procedure on the variance of RDS estimates. Past work has assumed that the variance of RDS estimates is primarily affected by segregation of healthy and infected individuals. We examine an illustrative model to show that this is not necessarily the case, and that bottlenecks anywhere in the networks can substantially affect estimates. We also show that variance is inflated by a common design feature in which the sample members are encouraged to recruit multiple future sample members. The paper concludes with suggestions for implementing and evaluating RDS studies. Copyright © 2009 John Wiley & Sons, Ltd.

KEY WORDS: hard-to-reach populations; hidden populations; HIV surveillance; importance sampling; Markov chain Monte Carlo; respondent-driven sampling; social networks; spectral gap

1. INTRODUCTION

The Joint United Nations Program on HIV/AIDS (UNAIDS) estimates that there are between 30 and 35 million people living with HIV/AIDS worldwide, and that between 2 and 4 million people were newly infected in 2007. In most countries outside of sub-Saharan Africa, these infections are concentrated in three subpopulations: men who have sex with men, injection drug users, and sex workers and their sexual partners [1]. Consequently, there is general consensus among epidemiologists that better data about disease prevalence and risk behaviors within these key subpopulations are critical for understanding and controlling the spread of the disease [2–5].

Unfortunately, because these subpopulations lack appropriate sampling frames, are relatively small, and their members often desire to remain anonymous, they are difficult to study with standard

*Correspondence to: Matthew J. Salganik, Department of Sociology and Office of Population Research, Princeton University, Wallace Hall, Princeton, NJ 08544, U.S.A.

†E-mail: mjs3@princeton.edu

‡E-mail: goel@yahoo.inc.com

sampling methods. For this reason they are often called ‘hidden’ or ‘hard-to-reach.’ A variety of sampling approaches have been tried to study these hidden populations, but in many cases they produce estimates of unknown bias and variance [5, 6]. The resulting uncertainty about key subpopulations has complicated public health efforts to evaluate prevention programs and allocate resources effectively.

Respondent-driven sampling (RDS) is a new approach for sampling from hidden populations that is rapidly gaining in popularity: A recent review identified more than 120 RDS studies worldwide [7], including populations as diverse as men who have sex with men in Uganda [8], sex workers in Vietnam [9], and injection drug users in the former Soviet Union [10]. Furthermore, the U.S. Centers for Disease Control and Prevention (CDC) recently selected RDS for a 25-city study of injection drug users that is part of the National HIV Behavioral Surveillance System [11]. Because CDC decisions often influence global public health standards, RDS is likely to become increasingly common in the study of hidden populations.

RDS data are collected through a snowball mechanism, in which current sample members recruit future sample members.[§] An RDS study begins by recruiting a small number of people in the target population to serve as seeds. After participating, the seeds are asked—and often provided financial incentive—to recruit other people that they know in the target population. The sampling continues in this way with current sample members recruiting the next wave of sample members until the desired sample size is reached.[¶] The process results in recruitment networks like the one in Figure 1 from a study of drug users in New York City; the sample began with eight seeds and grew to include 618 people in 13 weeks [27]. Under certain strong assumptions described later in the paper, these RDS data can then be used to produce asymptotically unbiased estimates about the hidden population (e.g. estimates of the proportion of drug users in New York who have HIV).

Despite the widespread use of RDS and its potential to address important public health questions, the statistical foundations of RDS remain poorly understood. This paper presents RDS as Markov chain Monte Carlo (MCMC) importance sampling, and analyzes how RDS estimates are affected by both the community structure of the hidden population and the recruitment procedure. We show that the variance of the RDS estimator is increased by: (1) ‘bottlenecks’ between different groups in the hidden population and (2) a study design in which participants recruit multiple individuals.

Our paper is organized as follows. In Section 2 we show that RDS sampling and estimation can be viewed as MCMC importance sampling. While MCMC algorithms are typically computer-driven, a novel feature of RDS is that the state transitions consist of individuals physically recruiting others in the hidden population. In Section 3 we analyze a particular, illustrative network model in detail. This example shows that the structure of the hidden population’s social network can significantly impact both the bias and variance of RDS estimates, a phenomenon that is well understood in the statistical MCMC community but that has been overlooked in the RDS literature. Importantly, bottlenecks in any part of the network may affect RDS estimates of quantities that are not directly related to the source of the bottleneck. For example, a bottleneck between racial groups may degrade RDS estimates of gender composition. This suggests that the variance of RDS estimates is likely larger than previously believed. In Section 4 we explore the effect of multiple

[§]This type of sampling is also sometimes called chain-referral, random-walk, or link-tracing [12–19], and can be considered as a form of adaptive sampling [20, 21].

[¶]Although it is beyond the scope of this paper, it is critical to note that there are serious logistical and ethical complications involved in running an RDS study [22–26]. To give just one example, in studies of injection drug users, it is common for non-drug users to attempt to participate in the study in order to earn money [22, 26].

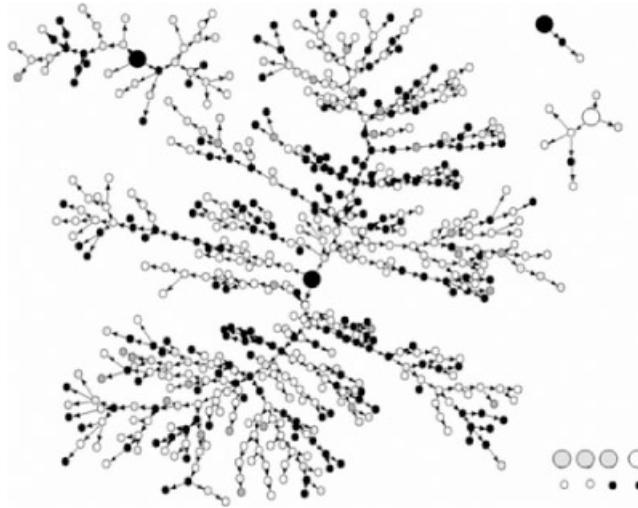


Figure 1. Recruitment networks from a study of drug users in New York City. The eight seeds are larger than the others nodes and all nodes are color coded by race/ethnicity. This figure was originally published in [27].

recruitment on variance, an issue that is important for RDS, but has not been considered previously and typically does not arise in MCMC applications. We show that ‘thick’ as opposed to ‘thin’ recruitment chains increase the statistical dependence between samples, and consequently worsen RDS estimates. Section 5 summarizes the results and concludes with recommendations for users of RDS. We have relegated most proofs and technical details to Appendices A and B. Appendix C reviews conductance, a formal measure of bottlenecks in networks.

2. RDS AS MCMC

RDS [28–30] is a form of snowball sampling often used to estimate the proportion of a population with a specific characteristic. Although in this paper we talk about estimating the proportion p of infected individuals, we could more generally be estimating the occurrence of any characteristic or behavior. Here we review MCMC importance sampling and make the connection to RDS precise.

2.1. Markov chain Monte Carlo

MCMC was popularized by the introduction of the Metropolis algorithm [31] and has been applied extensively in a variety of fields, including physics, chemistry, biology, and statistics. MCMC has also been the subject of several book-length treatments [32–35].

Behind all MCMC methods is a Markov chain on a state space V . In the context of RDS, V is the population from which we sample (e.g. drug injectors in New York City). We confine ourselves to the case where V is a finite population of size N , and so identify the chain with a kernel $K(v_i, v_j)$ that gives the probability of transition from state v_i to state v_j :

$$K(v_i, v_j) \geq 0, \quad \sum_{v_j \in V} K(v_i, v_j) = 1$$

In terms of RDS, $K(v_i, v_j)$ is the probability that any individual v_i recruits an individual v_j . The chain is *irreducible* if for every pair of points v_i, v_j , there is positive probability of eventually reaching v_j starting from v_i . Under this assumption, there is a unique distribution $\pi: V \rightarrow \mathbb{R}$ —called the *stationary distribution*—satisfying

$$\sum_{v_i \in V} \pi(v_i) K(v_i, v_j) = \pi(v_j)$$

That is, if X_0, X_1, X_2, \dots is a realization of the chain with $X_0 \sim \pi$, then $X_i \sim \pi$ for $i \geq 0$. Consequently, by starting the chain in equilibrium, the walk can be used to generate dependent samples from the distribution π .

2.2. Importance sampling

As shown above, a chain-referral sampling method can be used to draw dependent samples from the population V with distribution π :

$$\mathbb{P}(X_i = v_j) = \pi(v_j)$$

That is, on each draw individual v_j has probability $\pi(v_j)$ of being chosen. Then for any function $f: V \rightarrow \mathbb{R}$, the sample mean

$$\frac{1}{n} \sum_{i=0}^{n-1} f(X_i) \quad (1)$$

gives an unbiased estimate not of the population mean, but of $\mathbb{E}_\pi f = \sum_{i=1}^N f(v_i) \pi(v_i)$. That is, because units are selected with unequal probability, the sample mean is not a consistent estimator of the population mean. As is common in the survey sampling literature [36], the idea behind importance sampling [37] is that the weighted sample mean

$$\frac{1}{n} \sum_{i=0}^{n-1} \frac{f(X_i)}{N \cdot \pi(X_i)} \quad (2)$$

produces an unbiased estimate of the population mean μ_f of f since

$$\begin{aligned} \mathbb{E}_\pi \left(\frac{f(X_i)}{N \cdot \pi(X_i)} \right) &= \sum_{i=1}^N \frac{f(v_i)}{N \cdot \pi(v_i)} \pi(v_i) \\ &= \frac{1}{N} \sum_{i=1}^N f(v_i) \end{aligned}$$

In particular, if $D \subseteq V$ is the subset of infected individuals, then (2) can be used to estimate the disease prevalence $p = |D|/N = \mu_f$ by setting $f(v_i) = 1$ if $v_i \in D$ and $f(v_i) = 0$ otherwise.

It is often necessary to replace (2) by the asymptotically unbiased importance sampling estimator

$$\hat{\mu}_{\text{IS}} = \frac{1}{\sum_{i=0}^{n-1} 1/\pi(X_i)} \sum_{i=0}^{n-1} \frac{f(X_i)}{\pi(X_i)} \quad (3)$$

The considerable advantage of (3) over (2) is that the importance weights $1/\pi(X_i)$ only need to be evaluated up to a multiplicative constant (e.g. one does not need to know N). In many applications, including RDS, this simplification is essential.

2.3. Respondent-driven sampling

Importance sampling allows estimation of p given samples X_0, X_1, \dots from any fixed distribution π . RDS generates such samples via a recruitment process akin to MCMC. The link between RDS and MCMC has been noted previously [29, 38, 39]; here we make that connection explicit.

Consider a social network $G = (V, E)$ where nodes $x \in V$ represent individuals in the population that are either infected or healthy, and $e \in E$ represent edges in the network. We assume symmetric weighted edges (i.e. symmetric relationships) and we write $W(x, y) = W(y, x)$ for the weight of the edge between nodes x and y .^{||} Further, we assume that the network is connected (i.e. that there exists a path between every pair of individuals in the population).

For a subset of individuals $A \subseteq V$ we use the notation

$$W_A = \sum_{x \in A} \sum_{y \in V} W(x, y)$$

to denote the weight of A . For singleton sets, we write W_x instead of $W_{\{x\}}$.

We model the RDS sampling procedure as a random walk on the weighted graph G defined by the kernel $K(x, y) = W(x, y)/W_x$, where $K(x, y)$ is the probability that individual x recruits individual y .^{**} Assuming that the network is connected (i.e. the chain is irreducible), the walk has a unique stationary distribution

$$\pi(x) = \frac{W_x}{W_V}$$

Consequently, for X_0, X_1, X_2, \dots a realization of the chain with $X_0 \sim \pi^{\dagger\dagger}$ and $f: V \rightarrow \mathbb{R}$ any function, the importance sampling estimator (3) of the population mean μ_f reduces to

$$\hat{\mu}_f = \frac{1}{\sum_{i=0}^{n-1} 1/W_{X_i}} \sum_{i=0}^{n-1} \frac{f(X_i)}{W_{X_i}} \quad (4)$$

The RDS estimator (4) was recently introduced in [38], and will likely supplant the RDS estimator introduced in [29]. In the case of estimating disease prevalence, by setting $f(v_i) = 1$ if v_i is infected and $f(v_i) = 0$ otherwise, (4) simplifies to

$$\hat{p} = \frac{1}{\sum_{i=0}^{n-1} 1/W_{X_i}} \sum_{X_i \text{ infected}} \frac{1}{W_{X_i}} \quad (5)$$

To evaluate the RDS estimators (4) and (5) one still needs to know the weights W_{X_i} . Typically, researchers set uniform edge weights, $W(x, y) = 1$, corresponding to the assumption that participants recruit their contacts uniformly at random and that all contacts approached agree to

^{||}This assumption of symmetric relationship may not hold in some situations, for example caste societies.

^{**}Note that unlike our model, in practice RDS is conducted as sampling without replacement (i.e. those who participate cannot participate again).

^{††}Although real RDS studies often use multiple seeds, for simplicity we assume that only one seed is selected. However, most of our results can be generalized to handle multiple seeds without significant complication and remain qualitatively unchanged.

participate. Throughout the paper we refer to this as the *uniform recruitment assumption*.^{‡‡} In this case, W_x equals the degree of node x (i.e. her number of contacts).^{§§}

In contrast to estimates from snowball sampling based on the sample mean (1),^{¶¶} RDS estimates weight samples proportional to their assumed probability of selection. In the case where all nodes have the same degree, the sample mean estimate is equivalent to the RDS estimate (4) given the uniform recruitment assumption.

In the above, we start the walk in stationarity: $X_0 \sim \pi$ (i.e. the initial seed is drawn according to the stationary distribution). However, if the walk is aperiodic (i.e. if the network is not bipartite), then the RDS estimator $\hat{\mu}$ is asymptotically unbiased regardless of the starting distribution. Moreover, there is a central limit theorem for $\hat{\mu}$:

$$\sqrt{n}(\hat{\mu}_n - \mu) \rightarrow N(0, \sigma_f^2) \quad (6)$$

for any initial distribution on X_0 .^{|||} The variance σ_f^2 depends on the variance of f and the autocorrelation structure of the chain, and can be difficult to estimate in practice [34].

We hasten to point out that these results regarding the asymptotic behavior of RDS estimates hinge critically on the validity of the modeling assumptions. In particular, these results require that participants recruit a single individual^{***} chosen uniformly at random from their network of contacts, and that participants can be recruited into the sample multiple times (i.e. sampling with replacement). Furthermore, even if all of the appropriate conditions are met, the asymptotic theory says little regarding the performance of RDS in small samples ($n \approx 500$). As we show, in the case of small samples, the social network structure is of central importance.

^{‡‡}This uniform recruitment assumption is almost certainly not met in most studies because of selection bias in the decisions of recruiters on whom to enlist and of recruits on whether to participate. For example, in an RDS study of injection drug users in Chicago, Scott found that some participants were more likely to recruit their close contacts, a problem that was exacerbated by the financial incentives inherent in the RDS design [26]. Furthermore, in studies of men who have sex with men in Brazil, Ukraine, and Estonia, it was noted that some members of the target population actively avoided participating because they did not want to take the HIV test that was part of the study [24].

^{§§}Ascertaining an individual's degree is itself a challenging problem [40, 41], particularly in the context of RDS [24], and constitutes a source of non-sampling error.

^{¶¶}For more sophisticated estimators dealing with data collected via snowball sampling and more general link tracing designs, see [13, 17–19, 39, 42, 43].

^{|||}The Markov chain central limit theorem is typically applied to the unbiased estimate (2); one can, however, generalize that result to apply to (4). First observe that

$$\begin{aligned} \sqrt{n}(\hat{\mu}_n - \mu) &= \frac{\sqrt{n}}{\sum_{i=0}^{n-1} 1/W_{X_i}} \sum_{i=0}^{n-1} \frac{f(X_i) - \mu}{W_{X_i}} \\ &= \left[\frac{1}{(1/n) \sum_{i=0}^{n-1} W_V / N W_{X_i}} \right] \left[\frac{1}{\sqrt{n}} \sum_{i=0}^{n-1} \frac{f(X_i) - \mu}{N W_{X_i} / W_V} \right] \end{aligned}$$

By the Markov chain strong law of large numbers, the first term in the final expression converges to 1 almost surely. The result now follows by applying the usual Markov chain central limit theorem to the second expression with the function

$$g(X_i) = \frac{f(X_i) - \mu}{N W_{X_i} / W_V}$$

^{***}The case of multiple recruitment will be treated in Section 4 of this paper, where it is shown that (6) no longer holds in general.

3. EFFECTS OF COMMUNITY STRUCTURE

It is well understood that the bias and variance of MCMC estimates are critically affected by the structure of the network underlying the random walk. However, past work on RDS has focused on only one structural feature: bottlenecks between infected and uninfected individuals (Figure 2) [28, 29, 44].^{†††} That is, it was previously believed that as long as there were sufficient connections between infected and uninfected individuals, the RDS estimates would be reasonably precise. While this structural feature is certainly a concern, taken in isolation it underestimates the effect of network structure on the variance of RDS estimates. Even when infected and uninfected individuals are relatively well-connected, bottlenecks in other parts of the network can lead to large variance.

To illustrate this point, we analyze RDS on two network models in detail. Our examples, while motivated by the qualitative features of real social networks, are not intended to be accurate models of any specific social network. Rather, they provide insight by allowing for exact and interpretable results.

3.1. Two network models

3.1.1. A two-group model. Consider a population V consisting of two groups, A and B , of equal size $N/2$. Edges exist between every pair of individuals, however *within-group* edges have weight $1 - c$ while *between-group* edges have weight c where $0 < c < 1/2$ (see Figure 3(a)).^{†††} That is, within-group relationships are stronger than between-group relationships. In this model, c parameterizes homophily based on group membership—the well-observed social tendency for people to form ties to others who are similar [45]; as c increases, the tendency for within-group ties decreases.^{§§§}

Let p_A and p_B denote the proportion of infected individuals within the two groups, and let $D \subseteq V$ be the subset of infected people. Since we are assuming $|A| = |B|$, the proportion of infected individuals in the entire population is $p = (p_A + p_B)/2$. If the two groups have different infection rates, $p_A \neq p_B$, then, as we show, the network bottleneck between the two groups affects the RDS estimate, even though infected and uninfected individuals are well-connected.

We can imagine this more concretely by considering the case of street-based and agency-based sex workers in Belgrade, two groups that have been found to have little contact [46]. If these two groups had different HIV prevalence, then the weak connections between the groups could lead to high variance for the RDS-estimated HIV prevalence for sex workers as a whole because sometimes the sampling would get stuck in one group and sometimes it would get stuck in the other. Further, if the seeds are not selected from the stationary distribution, the bottleneck between groups can lead to biased estimates.

^{†††}We use the term ‘bottleneck’ in a strictly non-technical sense to refer to features of the network that lead to a propensity for the sampling procedure to get ‘stuck’ in certain regions. For example, we would say there is a bottleneck between groups A and B in the two-group example in Figure 3(a). A more formal notion of ‘bottlenecks’ is conductance, which is elaborated upon in Appendix C, but which is not essential to the main exposition.

^{†††}This network model allows for self-edges, which means that it allows for self-recruitment during the sampling process. This assumption departs from the actual RDS sampling process, but has minimal effect on the qualitative results.

^{§§§}This parameterization differs slightly from other parameterizations of homophily (for example, [27, 28]), but links naturally to the general notion of *conductance*, which is more fully described in Appendix C.

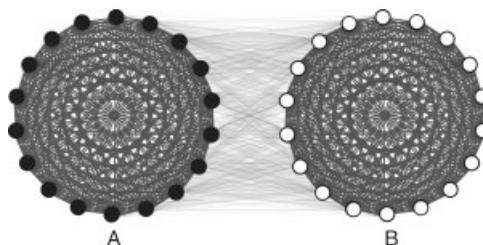


Figure 2. Hypothetical network with an edge between every pair of nodes, where within-group edges have higher weight than between-group edges. Here the two groups are defined by infection status, and a bottleneck exists between healthy and infected individuals.

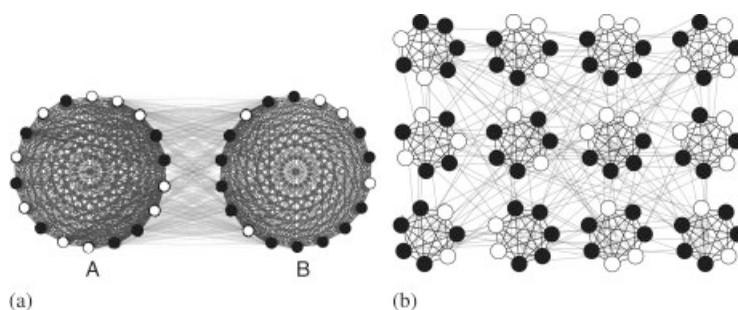


Figure 3. Hypothetical networks with an edge between every pair of nodes, where within-group edges have higher weight than between-group edges. In the two-group model the population is divided into two equally sized groups that differ in disease prevalence. In the multi-group model, the population is divided into many smaller equally sized subgroups that also differ in disease prevalence. In terms of RDS, these two models are equivalent: (a) two-group model and (b) multi-group model.

3.1.2. A multi-group extension. Although one could plausibly detect and potentially compensate for the simple network bottleneck in the two-group model, more subtle—and hence harder to diagnose—structural features can also lead to high variance of RDS estimates.

Consider the multi-group network depicted in Figure 3(b). This more general model aims to capture the fact that many real social networks partition into relatively homogenous subgroups where there are stronger ties within subgroups than between subgroups—a feature that sociologists call ‘cohesive subgroups’ [47] and physicists call ‘community structure’ [48]. In the multi-group model, N nodes are divided into subgroups of m nodes; all nodes are connected, but within-subgroup edges have weight 1, and between-subgroup edges have weight $b < 1$.^{†††} The subgroups themselves come in two varieties, A and B , with exactly half the subgroups of type A and the other half type B . Type A subgroups have a proportion p_A of their nodes infected, and type B subgroups have a proportion p_B infected. In the case of contagious diseases, such clumping of cases within subgroup is particularly likely [49].

^{†††}As with the two-group model, the multi-group model allows for self-edges, meaning that nodes are allowed to recruit themselves.

Despite their apparent differences, the multi-group model is in fact equivalent to the two-group model: For every value of $c < \frac{1}{2}$ in the two-group model, there is a corresponding value of b in the multi-group model such that the RDS estimator \hat{p} has the same distribution under both network models (details are provided in Appendix A).

3.2. Analyzing the models

Here we consider the bias and variance properties of RDS on the network models discussed above. Without loss of generality, we consider only the two-group model.

In the two-group model, RDS is based on the following Markov chain:

$$K(x, y) = \begin{cases} 2(1-c)/N, & x, y \in A \text{ or } x, y \in B \\ 2c/N, & x \in A, y \in B \text{ or } x \in B, y \in A \end{cases} \quad (7)$$

Written as a matrix,

$$K = \left[\begin{array}{c|c} 2(1-c)/N & 2c/N \\ \hline 2c/N & 2(1-c)/N \end{array} \right]$$

where the matrix is partitioned into blocks of size $N/2 \times N/2$.

K has stationary distribution $\pi(x) = W_x / W_V = 1/N$ that is uniform over V , since

$$W_x = \sum_y W(x, y) = (1-c)N/2 + cN/2 = N/2$$

is independent of x (i.e. each unit has equal probability of selection). Furthermore, since the weight of each node is the same, the RDS estimator (5) simplifies to

$$\hat{p} = \frac{1}{\sum_{i=0}^{n-1} 1/W_{X_i}} \sum_{X_i \in D} \frac{1}{W_{X_i}} = \frac{\#\{X_i \in D\}}{n}$$

which is the usual estimator for simple random samples. Unlike simple random samples, however, the samples X_i are not independent, and the social network structure of the population affects RDS estimates.

To analyze \hat{p} , we derive an explicit expression for the distribution K_l of the state of the chain after l steps.

Lemma 3.1

For $0 < c < \frac{1}{2}$, the l -step distribution of the walk defined in (7) is

$$K_l(x, y) = \begin{cases} (1 + \beta_1^l)/N, & x, y \in A \text{ or } x, y \in B \\ (1 - \beta_1^l)/N, & x \in A, y \in B \text{ or } x \in B, y \in A \end{cases}$$

where β_1 is the second largest eigenvalue of the transition matrix K , which in this case is equal to $1 - 2c$.

Although the equilibrium distribution $\pi(x) = 1/N$ is uniform over V , after any finite number of steps the chain is more likely to be in the group from which the initial sample was chosen due to

preferential within-group recruitment. For example, if the initial seed is chosen from A , then for $c=0.1$, after five steps the chain is still about twice as likely to be in A than in B .

Here the second largest eigenvalue β_1 of the transition matrix is seen to control the rate of convergence of the chain to its equilibrium distribution. This phenomenon is true for general chains [50], and as we show below, β_1 also affects both the bias and variance of the RDS estimate.

In our example, the RDS estimator \hat{p} is unbiased if the initial sample X_0 is chosen from the stationary distribution π . If instead X_0 is chosen uniformly from group A , then \hat{p} is biased (although still asymptotically unbiased), and moreover, the bias depends on the bottleneck that is induced by the homophily parameter c and the length of the recruitment chains. This illustrates the fact that the network location of the seed becomes increasingly important in populations with bottlenecks between groups.

Lemma 3.2

Consider the walk defined in (7). For an initial sample X_0 chosen uniformly from group A and a referral chain of size n

$$\mathbb{E}\hat{p} = p + (p_A - p_B) \frac{1 - \beta_1^n}{4nc}$$

where $\beta_1 = 1 - 2c$.

From Lemma 3.2 we know that the estimator \hat{p} has bias

$$\text{bias}(\hat{p}) = (p_A - p_B) \frac{1 - (1 - 2c)^n}{4nc} \approx \frac{p_A - p_B}{4nc} = \frac{p_A - p_B}{2n(1 - \beta_1)}$$

that depends on the homophily c , the length of the chain n , and the difference in infection proportions between the two groups. The *spectral gap* $1 - \beta_1 = 2c$ captures the effect of network structure. Note that this also shows that even though RDS estimates are asymptotically unbiased, as is often claimed in the literature, there can be substantial bias when the seeds are not selected from the stationary distribution and the sample size is small.

In a population with $c=0.1$, a referral chain of length 10 that has initial seed chosen uniformly from group A has bias approximately $(p_A - p_B)/5$. As $c \rightarrow 0$ (i.e. as the two populations become completely disconnected), $\text{bias}(\hat{p}) \rightarrow (p_A - p_B)/2$. In this extreme case, RDS erroneously estimates only p_A instead of $p = (p_A + p_B)/2$.

In the above we estimated the bias of \hat{p} given that the initial seed was chosen uniformly from group A . Now we assume that the seed is chosen uniformly from the entire population (so that \hat{p} is unbiased) and analyze its variance.

In populations with community structure, it is more likely that individuals refer people who are in their same social subgroup. Intuitively, in this situation we gain less information from each recruit than if that recruit was chosen randomly from the entire population. The result of this dependence is an effective reduction in sample size. That is, the variance of RDS estimates is larger than the variance of estimates based on a simple random sample of the same nominal size.

The dependence between samples is quantified by their covariance.

Lemma 3.3

Consider the walk defined in (7). Suppose X_0^1, X_1^1, \dots and X_0^2, X_1^2, \dots are two independent realizations of the walk with $X_0^1 = X_0^2 \sim \pi$. That is, both chains begin at the same vertex v , which is

drawn from the stationary distribution π . Then for $i, j \geq 0$

$$\text{Cov}(f_D(X_i^1), f_D(X_j^2)) = \left(\frac{p_A - p_B}{2} \right)^2 \beta_1^{i+j}$$

where $\beta_1 = 1 - 2c$ and

$$f_D(v_i) = \begin{cases} 1, & v_i \text{ infected} \\ 0, & \text{otherwise} \end{cases}$$

Corollary 3.1

Consider the walk X_0, X_1, \dots defined in (7). If $X_0 \sim \pi$, then the variance of \hat{p} satisfies

$$\text{Var}(\hat{p}) = \frac{p - p^2}{n} + \frac{(p_A - p_B)^2 \beta_1}{2n(1 - \beta_1)} - \frac{(p_A - p_B)^2 (\beta_1 - \beta_1^{n+1})}{2n^2(1 - \beta_1)^2}$$

where $\beta_1 = 1 - 2c$ and n is the sample size.

Again we see that the spectral gap $1 - \beta_1$ affects RDS estimates. A naive estimate of the variance (i.e. the variance under simple random sampling) assumes samples are uncorrelated, yielding only the first term $(p - p^2)/n$. In particular, it does not take into account possible community structure in the hidden population. For example, for $c = 0.1$, $p_A = 0.3$, and $p_B = 0.1$, $\text{Var}(\hat{p})$ is approximately 1.5 times the variance of estimates from a simple random sample. Accordingly, confidence intervals determined by the true variance are $\sqrt{1.5} \approx 1.2$ times wider. Put another way, community structure in this example effectively reduces sample size by a third: RDS estimates based on a sample of 500 individuals have the same variance as estimates based on a simple random sample of 335.

4. EFFECTS OF MULTIPLE RECRUITMENT

Above, we have been assuming that RDS estimates are based on a single, long run of the chain. In practice, this approach is difficult to implement since some sample members do not recruit others, causing the chains to terminate. Instead, in order to ensure that the chains continue, each respondent is encouraged to recruit multiple individuals, as seen in recruitment chains from the Abdul-Quader *et al.* study of drug users in New York City (Figure 1) [27]. In this study, as with almost all RDS studies, participants were encouraged to recruit up to three others [7]. For a given sample size, however, chain lengths are shorter under multiple recruitment than single recruitment. Consequently, multiple recruitment increases the dependence between participants, and in turn increases the variance of RDS estimates—a concern that was previously overlooked. Observe that multiple recruitment is a different source of dependence than that which directly arises from network community structure; but, as we show, the two interact: community structure amplifies the problems caused by multiple recruitment.^{||||}

^{||||}Multiple recruitment does not yet appear to have been applied in traditional, computer-driven applications of MCMC. In those contexts, there is no worry of chain termination, and so the primary motivation for multiple recruitment in RDS studies is no longer relevant. However, one could still imagine utility in branching off multiple random walkers to explore ‘good’ regions of the state space. In that case, the tradeoffs described here would potentially apply.

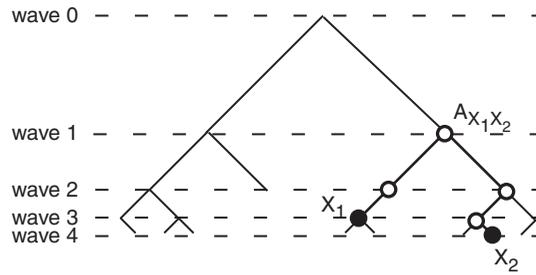


Figure 4. A_{x_1,x_2} is the most recent common ancestor of X_1 and X_2 .

We examine the effects of multiple recruitment on the two-group and multi-group network models described in Section 3. As before, we assume that the initial sample X_0 is drawn from the stationary distribution. To compute the covariance between $f_D(X_1)$ and $f_D(X_2)$ in Figure 4, observe that A_{x_1,x_2} is the most recent common ancestor of X_1 and X_2 . Consequently, X_1 and X_2 result from independent runs of the chain started at A_{x_1,x_2} , and so we are in the situation of Lemma 3.3. That is,

$$\text{Cov}(f_D(X_1), f_D(X_2)) = \left(\frac{p_A - p_B}{2}\right)^2 \beta_1^{2+3}$$

In general, for two samples X_i and X_j , this argument shows that

$$\text{Cov}(f_D(X_i), f_D(X_j)) = \left(\frac{p_A - p_B}{2}\right)^2 \beta_1^{l(i,j)}$$

where $l(i, j)$ is the length of the unique path between X_i and X_j in the recruitment tree.

Lemma 4.1

Consider the walk defined in (7). Suppose a recruitment tree is chosen according to a probability distribution λ on the set of n -node trees, and RDS samples are correspondingly collected. Then the variance of \hat{p} satisfies

$$\text{Var}(\hat{p}) = \frac{p - p^2}{n} + \frac{(p_A - p_B)^2}{2n^2} \sum_{k=1}^{n-1} \beta_1^k \mathbb{E}_\lambda L(k)$$

where $L(k)$ is the number of pairs of samples distance k apart.

Lemma 4.1 shows that, for a given social network structure, the further apart the sample units, the lower the variance. That is, ‘thin,’ as opposed to ‘thick,’ recruitment chains lead to improved estimates. Furthermore, observe the key role again played by the second largest eigenvalue β_1 . As β_1 increases (i.e. as the spectral gap $1 - \beta_1$ decreases), the variance of RDS increases. In other words, community structure amplifies the effects of multiple recruitment.

With single recruitment, the Markov chain central limit theorem (6) shows that the variance of \hat{p} decreases as $1/n$, where n is the sample size. With multiple recruitment, however, the variance may decrease as $1/n^\delta$ for $\delta < 1$. That is, multiple recruitment may lead to asymptotically slower decay of error in the RDS estimator. To see this effect, we analyze the variance of the RDS estimator

on the two-group model in the case where each sample member recruits exactly two other sample members.

In order to apply Lemma 4.1, we first estimate the path length distribution.

Theorem 4.1

Suppose T_H is a complete binary tree of height $H \geq 1$ (i.e. T_H has $2^{H+1} - 1$ nodes and each non-leaf node has exactly two children). Let $L_H(k)$ be the number of pairs of nodes distance k apart. Then for $1 \leq k \leq 2H$

$$\frac{1}{4}2^{H+k/2} \leq L_H(k) \leq 2k2^{H+k/2}$$

Since $2^H \approx n$, Theorem 4.1 shows that, ignoring log factors, $L_H(k) \approx n2^{k/2}$. Now we estimate the variance of \hat{p} .

Theorem 4.2

Consider the walk defined in (7). Suppose the recruitment tree is a complete binary tree of height $H \geq 1$, in which case $n = 2^{H+1} - 1$. If $\beta_1 > \sqrt{2}/2$, then the variance of \hat{p} satisfies

$$\begin{aligned} \frac{p-p^2}{n} + \left[\frac{1}{n^{\log_2 1/\beta_1^2}} - \frac{2}{n} \right] \left(\frac{\beta_1 \sqrt{2}(p_A - p_B)}{32(\beta_1 \sqrt{2} - 1)} \right) &\leq \text{Var}(\hat{p}) \\ &\leq \frac{p-p^2}{n} + \frac{4 \log_2 n}{n^{\log_2 1/\beta_1^2}} \left(\frac{p_A - p_B}{1 - \beta_1 \sqrt{2}} \right)^2 \end{aligned}$$

Ignoring log factors, Theorem 4.2 shows that for $\beta_1 > \sqrt{2}/2$, $\text{Var}(\hat{p}) \approx 1/n^{\log_2 1/\beta_1^2}$. In particular, for $\sqrt{2}/2 < \beta_1 < 1$, we have $\log_2 1/\beta_1^2 < 1$. Furthermore, this exponent decreases (i.e. decay is slower) as β_1 increases.

Above we analyze a deterministic recruitment tree; now we consider a more realistic stochastic recruitment procedure that is modeled as a branching process**** with offspring distribution based on data from the Frost *et al.* study of injection drug users in Tijuana and Ciudad Juarez [52]. In that study, three coupons were provided to each participant and approximately one-third of the participants recruited no other participants, one-sixth recruited one other, another one-sixth recruited two others, and the remaining one-third recruited three other participants (Table I). In this case, while it seems difficult to find an analytic expression for $\mathbb{E}L(k)$, the path length distribution can be estimated by simulation. Combining these simulation results that estimate $\mathbb{E}L(k)$ with our common network parameter values ($c=0.1$, $p_A=0.1$, and $p_B=0.3$) and a sample size of $n=500$, we have $\text{Var}(\hat{p})$ is approximately 3.7 times the variance under simple random sampling. In other words, community structure and multiple recruitment substantially reduce the effective RDS sample size. In this example, a sample size of 500 collected via RDS with multiple recruitment corresponds to a sample size of just 136 people collected via simple random sampling.

****Further complications arise when the offspring distribution is a function of the characteristics of the individuals. For example, in an RDS study of jazz musicians in New York, female participants recruited an average of 1.4 people while male participants recruited an average of 0.85 [51].

Table I. Multiple recruitment offspring distribution based on a study of injection drug users in Tijuana, Mexico and Ciudad Juarez, Mexico [52].

	Number of recruits			
	0	1	2	3
Probability	$\frac{1}{3}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{3}$

5. CONCLUSION

5.1. Summary

Our network models illustrate the effects of both the social network and multiple recruitment on RDS estimates in a stylized setting that attempts to mimic situations in which RDS may be used. To summarize our findings for the two-group model (and equivalently, the multi-group model), we compare three sampling situations: simple random sampling, RDS with single recruitment, and RDS with multiple recruitment. Figure 5 shows the distribution of \hat{p} in these three cases.

Simple random sampling. Since $p=0.2$, the variance of \hat{p} is $(p-p^2)/500=0.00032$ and its standard deviation is approximately 0.0179. Consequently, the 95 per cent confidence interval for the estimate is approximately $\hat{p}\pm 3.5$ per cent. The variance, in this case, was independent of the network structure.

RDS—single recruitment. For $c=0.1$ the second largest eigenvalue satisfies $\beta_1=1-2c=0.8$. If the samples are the result of a single, long chain (without multiple recruitment) starting at the stationary distribution, then $\text{Var}(\hat{p})$ is given by Corollary 3.1, which yields a standard deviation of approximately 0.0219. The 95 per cent confidence interval is then $\hat{p}\pm 4.3$ per cent, the same interval one would get from a simple random sample of size 335.

RDS—multiple recruitment. Assume multiple recruitment follows a branching process with offspring distribution based on the recruitment data from the Frost *et al.* study of injection drug users in two Mexican cities (see Table I) [52]. Simulation shows that

$$\frac{1}{n} \sum_{k=1}^{n-1} \beta_1^k \mathbb{E}L(k) \approx 21.3$$

where the expectation is conditional on the recruitment tree having size $n=500$ (i.e. we disregard trees that go extinct prematurely). Lemma 4.1 then shows that the standard deviation of the estimate is approximately 0.0343, yielding a confidence interval $\hat{p}\pm 6.7$ per cent. This level of variance corresponds to a simple random sample of size 136 or an RDS sample with single recruitment of size 204.

5.2. Implications and directions for further research

We conclude by describing some of the specific implications of these findings for the practice of RDS.

Community structure. Past RDS work focused on the bottleneck between infected and healthy individuals. Bottlenecks anywhere in the network, however, impact the quality of RDS estimates. While preexisting knowledge may alert researchers to some bottlenecks (e.g. those between street-based and brothel-based sex workers), we suspect that it is difficult in practice to detect and

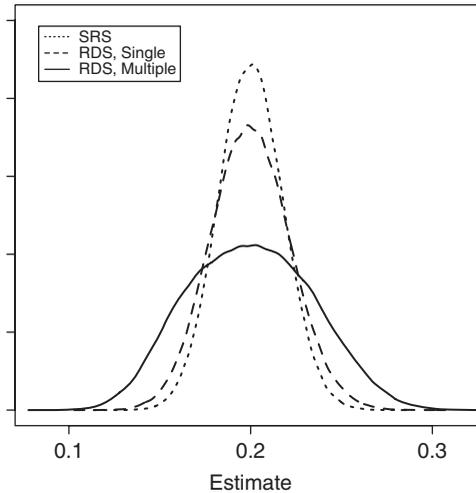


Figure 5. Comparing the results from simulated simple random sampling, RDS with single recruitment and RDS with multiple recruitment ($p_A=0.1$, $p_B=0.3$, $c=0.1$, $n=500$). In this case, all three produce unbiased estimates, but the variance of the RDS estimates is larger, much larger in the case of RDS with multiple recruitment.

to adjust for bottlenecks that exist in networks with complex community structure, such as the multi-group network in Figure 3(b). We hope future theoretical and empirical work continues to explore the bias and variance of the RDS estimators as a function of network structure, particularly taking care to develop procedures that require only limited information about the underlying social network.

Multiple recruitment. The multiple recruitment feature of RDS was developed to help ensure that sampling chains survive even when some subjects do not recruit. However, this design feature may diminish the accuracy of RDS estimates by increasing the dependence between sample units. Since the specific structure of recruitment chains impacts RDS estimates, larger sample sizes do not always produce more accurate estimates than smaller samples sizes, contrary to intuition from simple random sampling. While it is currently common practice to provide subjects with three recruitment coupons each, RDS would benefit from techniques that make it practical to reduce that number.

Assumptions. The properties of the RDS estimator rest on a number of assumptions, many of which may not be met in practice. For example, the sampling with replacement assumption is violated by design in virtually all RDS studies. Also problematic is the uniform recruitment assumption (i.e. that participants recruit their contacts uniformly at random and that all contacts approached agree to participate). For example, de Mello *et al.* found evidence of non-random recruitment in their study of men who have sex with men in Campinas, Brazil [53], and similar results have been reported elsewhere in the literature [24, 26, 54]. We hope that future work develops diagnostics to detect violations of these assumptions and explores the effects of such violations on RDS estimates.

High variance. Much of excitement around RDS in the public health community has focused on the fact that, under certain strong assumptions, the estimates are asymptotically unbiased. This paper highlights the variance of these estimates. In some cases, the variance of RDS estimates

may be so large that the estimates themselves are of little value. Prior work suggested that RDS researchers should assume a design effect of 2; that is, that RDS samples should be twice as large as would be needed under simple random sampling [44]. The results in this paper suggest that this rule of thumb should probably be revised upward. We hope that future work will provide further guidance to researchers about the sample sizes needed for their studies.

Beyond the specific results from this paper, clarifying the connection between RDS and MCMC allows future researchers to harvest ideas from the vast MCMC literature. For example, RDS researchers could consider discarding data from early sample waves, just as researchers using computer-driven MCMC often discard a portion of their draws during the so-called ‘burn-in’ phase [55]. This possibility is especially important because RDS seeds are almost certainly not chosen from the stationary distribution. Another potential avenue for future work would be to modify existing MCMC convergence diagnostics so as to monitor the convergence of RDS estimates. For example, the use of multiple seeds, a common feature of RDS studies, creates parallel chains that could lead to natural convergence measures [56–58]. One nice feature of this approach would be that researchers could run these diagnostics while the study is underway, and thereby potentially detect problems while there is still time to correct them. These suggestions represent just a few of the possible improvements to RDS, improvements that may ultimately allow researchers to better study hidden populations.

APPENDIX A: EQUIVALENCE OF THE TWO-GROUP AND MULTI-GROUP MODELS

Random walks on the two-group and multi-group network models of Section 3 are equivalent in the following sense. Let f indicate infection status: $f(v_i)=1$ if v_i is infected and $f(v_i)=0$ otherwise. Suppose N is even and divisible by m , and that the between-subgroup edge weight c in the two-group model satisfies $c < \frac{1}{2}$. Set the between-subgroup edge weight b in the multi-group model to

$$b = \frac{cm}{N(1/2 - c) + cm} \quad (\text{A1})$$

Finally, let X_0, X_1, \dots and $\tilde{X}_0, \tilde{X}_1, \dots$ denote RDS samples from the two networks, respectively, with either X_0 and \tilde{X}_0 chosen uniformly from a type- A subgroup or X_0 and \tilde{X}_0 chosen uniformly from a type- B subgroup. That is, X_0, X_1, \dots and $\tilde{X}_0, \tilde{X}_1, \dots$ are RDS samples on the two networks conditioned to start in the same subgroup type. Then, $f(X_0), f(X_1), \dots$ has the same distribution as $f(\tilde{X}_0), f(\tilde{X}_1), \dots$.

To prove this equivalence, it is sufficient to show that with b satisfying (A1), the probability of transition from any node in an A subgroup to any node in a B subgroup is the same in both models. In the two-group model, this between-group transition probability is

$$\frac{cN/2}{cN/2 + (1-c)N/2} = c$$

and in the multi-group model, the between-group transition probability is

$$\frac{bN/2}{b(N-m) + m}$$

Consequently, the two models are equivalent for b such that

$$\frac{bN/2}{b(N-m)+m} = c \quad (\text{A2})$$

Solving for b in (A2), we have

$$b = \frac{cm}{N/2 - c(N-m)} = \frac{cm}{N(1/2 - c) + cm}$$

establishing the equivalence.

To better understand the equivalence between these two models, we examine their limit behavior to build intuition; as we have shown, even the finite network models are equivalent. Observe that the probability of transitioning out of one of the small m -node subgroups is

$$\frac{b(N-m)}{b(N-m)+m} = \frac{1}{1 + \frac{m}{b(N-m)}} = \frac{1}{1 + \frac{N(1/2 - c) + cm}{c(N-m)}} \rightarrow 2c \quad \text{as } N \rightarrow \infty$$

In the limit, the two models are equivalent when transition out of a subgroup in the multi-group model occurs with probability $2c$. Since the number of subgroups $N/m \rightarrow \infty$, the probability of transitioning to a type-A subgroup, given that one has transitioned out of one's current subgroup, is $\frac{1}{2}$. Consequently, transition between *subgroup-types* (i.e. type-A or type-B subgroups) occurs with probability c .

APPENDIX B: FURTHER TECHNICAL DETAILS AND PROOFS

Suppose K is the kernel of an irreducible finite Markov chain and π its stationary distribution. We think of (K, π) as an operator on $L^2(\pi)$ —the space of functions $f: V \rightarrow \mathbb{R}$ with inner product

$$\langle f, g \rangle = \sum_{x \in V} f(x)g(x)\pi(x)$$

and corresponding norm

$$\|f\|_2^2 = \sum_{x \in V} f^2(x)\pi(x)$$

Then for $f \in L^2(\pi)$

$$Kf(x) = \sum_{y \in V} K(x, y)f(y)$$

We call $\psi \in L^2(\pi)$ as an eigenfunction for K with eigenvalue λ if $K\psi = \lambda\psi$.

Simple random walks on weighted graphs are *reversible*; that is, they satisfy the detailed balance equation

$$\pi(x)K(x, y) = \pi(y)K(y, x)$$

Reversibility is equivalent to $K: L^2(\pi) \rightarrow L^2(\pi)$ being self-adjoint. Consequently, reversible walks are diagonalizable in an orthonormal basis of real eigenfunctions. That is, there exist eigenfunctions

$\psi_0, \psi_1, \dots, \psi_{N-1}$ with corresponding real eigenvalues

$$1 = \beta_0 \geq \beta_1 \geq \dots \geq \beta_{N-1} \geq -1$$

such that $\langle \psi_i, \psi_j \rangle = \delta_{ij}$. For details of the above functional analytic view, see Saloff-Coste [50].

Lemma 3.1

For $0 < c < \frac{1}{2}$, the l -step distribution of the walk defined in (7) is

$$K_l(x, y) = \begin{cases} (1 + \beta_1^l)/N, & x, y \in A \text{ or } x, y \in B \\ (1 - \beta_1^l)/N, & x \in A, y \in B \text{ or } x \in B, y \in A \end{cases}$$

where $\beta_1 = 1 - 2c$.

Proof

The eigenfunctions and eigenvalues of K are:

- (1) $\psi_0(x) \equiv 1$, $\beta_0 = 1$.
- (2) $\psi_1(x) = 1_A(x) - 1_B(x)$ (i.e. $\psi(x)$ is 1 on A and -1 on B), $\beta_1 = 1 - 2c$.
- (3) The $N - 2$ dimensional subspace of functions $\psi: N \rightarrow \mathbb{R}$ such that

$$\sum_{x \in A} \psi(x) = \sum_{x \in B} \psi(x) = 0$$

These functions have eigenvalue $\lambda = 0$.

Lemma 1.2.9 of [50] shows that for reversible walks

$$\frac{K_l(x, y)}{\pi(y)} = \sum_i \beta_i^l \psi_i(x) \psi_i(y)$$

where $\{\psi_i\}$ is an $L^2(\pi)$ orthonormal basis of eigenfunctions for K with corresponding eigenvalues β_i .

In our case, since there are only two non-zero eigenvalues (and their corresponding eigenfunctions as we have written them down are orthonormal), we have

$$\begin{aligned} \frac{K_l(x, y)}{\pi(y)} &= \beta_0^l \psi_0(x) \psi_0(y) + \beta_1^l \psi_1(x) \psi_1(y) \\ &= 1 + \beta_1^l f_1(x) f_1(y) \end{aligned}$$

The result follows since $\psi_1(x) \psi_1(y) = 1$ if x and y are in the same group and $\psi_1(x) \psi_1(y) = -1$ if x and y are in different groups. \square

Lemma 3.2

Consider the walk defined in (7). For an initial sample X_0 chosen uniformly from group A and a referral chain of size n

$$\mathbb{E} \hat{p} = p + (p_A - p_B) \frac{1 - \beta_1^n}{4nc}$$

where $\beta_1 = 1 - 2c$.

Proof

The result follows from the distribution calculation of Lemma 3.1. First observe that for $X_0 \in A$

$$\begin{aligned} \mathbb{P}(f(X_i) = 1) &= \mathbb{P}(f(X_i) = 1, X_i \in A) + \mathbb{P}(f(X_i) = 1, X_i \in B) \\ &= \mathbb{P}(X_i \in A) \cdot \mathbb{P}(f(X_i) = 1 | X_i \in A) \\ &\quad + \mathbb{P}(X_i \in B) \cdot \mathbb{P}(f(X_i) = 1 | X_i \in B) \\ &= p_A \mathbb{P}(X_i \in A) + p_B \mathbb{P}(X_i \in B) \\ &= p_A \frac{1 + (1 - 2c)^i}{2} + p_B \frac{1 - (1 - 2c)^i}{2} \end{aligned}$$

Then,

$$\begin{aligned} \mathbb{E} \hat{p} &= \frac{1}{n} \sum_{i=0}^{n-1} \mathbb{P}(f(X_i) = 1) \\ &= \frac{1}{n} \sum_{i=0}^{n-1} p_A \frac{1 + (1 - 2c)^i}{2} + p_B \frac{1 - (1 - 2c)^i}{2} \\ &= \frac{p_A + p_B}{2} + \frac{p_A - p_B}{2n} \sum_{i=0}^{n-1} (1 - 2c)^i \\ &= p + \frac{p_A - p_B}{2n} \cdot \frac{1 - (1 - 2c)^n}{2c} \quad \square \end{aligned}$$

Lemma 3.3

Consider the walk defined in (7). Suppose X_0^1, X_1^1, \dots and X_0^2, X_1^2, \dots are two independent realizations of the walk with $X_0^1 = X_0^2 \sim \pi$. That is, both chains begin at the same vertex v , which is drawn from the stationary distribution π . Then for $i, j \geq 0$

$$\text{Cov}(f_D(X_i^1), f_D(X_j^2)) = \left(\frac{p_A - p_B}{2} \right)^2 \beta_1^{i+j}$$

where $\beta_1 = 1 - 2c$ and

$$f_D(v_i) = \begin{cases} 1, & v_i \text{ infected} \\ 0, & \text{otherwise} \end{cases}$$

Proof

Since the walks begin in stationarity, $X_i^1 \sim \pi$ and $X_i^2 \sim \pi$ for all i . Consequently, $\mathbb{E} f_D(X_i^1) \mathbb{E} f_D(X_j^2) = p^2$. To calculate $\mathbb{E} f_D(X_i^1) f_D(X_j^2)$, observe that X_i^1 and X_j^2 are conditionally independent given X_0^1 . So,

$$\mathbb{E}(f_D(X_i^1) f_D(X_j^2) | X_0^1 = x_0) = \mathbb{E}(f_D(X_i^1) | X_0^1 = x_0) \mathbb{E}(f_D(X_j^2) | X_0^2 = x_0)$$

Now

$$\begin{aligned}\mathbb{E}(f_D(X_i^1)|X_0^1=x_0) &= \mathbb{P}(X_i^1 \in I \cap A | X_0^1=x_0) + \mathbb{P}(X_i^1 \in I \cap B | X_0^1=x_0) \\ &= p_A \mathbb{P}(X_i^1 \in A | X_0^1=x_0) + p_B \mathbb{P}(X_i^1 \in B | X_0^1=x_0)\end{aligned}$$

By Lemma 3.1, for $x_0 \in A$ and $k=1, 2$

$$\mathbb{E}(f_D(X_i^k)|X_0^k=x_0) = p_A \frac{1+\beta_1^i}{2} + p_B \frac{1-\beta_1^i}{2}$$

and for $x_0 \in B$

$$\mathbb{E}(f_D(X_i^k)|X_0^k=x_0) = p_A \frac{1-\beta_1^i}{2} + p_B \frac{1+\beta_1^i}{2}$$

Consequently, for $x_0 \in A$

$$\begin{aligned}\mathbb{E}(f_D(X_i^1)f_D(X_j^2)|X_0^1=x_0) &= \left(p_A \frac{1+\beta_1^i}{2} + p_B \frac{1-\beta_1^i}{2} \right) \left(p_A \frac{1+\beta_1^j}{2} + p_B \frac{1-\beta_1^j}{2} \right) \\ &= \frac{1}{4} (p_A^2 (1+\beta_1^i)(1+\beta_1^j) + 2(1-\beta_{i+j}) p_A p_B + p_B^2 (1-\beta_1^i)(1-\beta_1^j))\end{aligned}$$

By symmetry, for $x_0 \in B$

$$\begin{aligned}\mathbb{E}(f_D(X_i^1)f_D(X_j^2)|X_0^1=x_0) &= \frac{1}{4} (p_B^2 (1+\beta_1^i)(1+\beta_1^j) \\ &\quad + 2(1-\beta_{i+j}) p_A p_B + p_A^2 (1-\beta_1^i)(1-\beta_1^j))\end{aligned}$$

Finally, since $X_0^1 \sim \pi$

$$\begin{aligned}\mathbb{E}f_D(X_i^1)f_D(X_j^2) &= \mathbb{E}(f_D(X_i^1)f_D(X_j^2)|X_0^1 \in A) + \mathbb{E}(f_D(X_i^1)f_D(X_j^2)|X_0^1 \in B) \\ &= p^2 + \left(\frac{p_A - p_B}{2} \right)^2 \beta_1^{i+j}\end{aligned}\tag{B2}$$

The result now follows because

$$\begin{aligned}\text{Cov}(f_D(X_i^1), f_D(X_j^2)) &= \mathbb{E}f_D(X_i^1)f_D(X_j^2) - \mathbb{E}f_D(X_i^1)\mathbb{E}f_D(X_j^2) \\ &= p^2 + \left(\frac{p_A - p_B}{2} \right)^2 \beta_1^{i+j} - p^2 \\ &= \left(\frac{p_A - p_B}{2} \right)^2 \beta_1^{i+j}\end{aligned}\quad \square$$

Corollary 3.1

Consider the walk X_0, X_1, \dots defined in (7). If $X_0 \sim \pi$, then the variance of \hat{p} satisfies

$$\text{Var}(\hat{p}) = \frac{p-p^2}{n} + \frac{(p_A - p_B)^2 \beta_1}{2n(1-\beta_1)} - \frac{(p_A - p_B)^2 (\beta_1 - \beta_1^{n+1})}{2n^2(1-\beta_1)^2}$$

where $\beta_1 = 1 - 2c$ and n is the sample size.

Proof

First note that

$$\begin{aligned} \text{Cov}(f_D(X_i), f_D(X_j)) &= \mathbb{E}(f_D(X_i) - p)(f_D(X_j) - p) \\ &= \mathbb{E}(\mathbb{E}[(f_D(X_i) - p)(f_D(X_j) - p) | X_i]) \\ &= \left(\frac{p_A - p_B}{2}\right)^2 \beta_1^{j-i} \end{aligned}$$

where the last equality follows from Lemma 3.3 since $X_i \sim \pi$. So,

$$\begin{aligned} \text{Var}(\hat{p}) &= \frac{1}{n^2} \sum_{i=0}^{n-1} \text{Var}(f_D(X_i)) + \frac{2}{n^2} \sum_{0 \leq i < j \leq n-1} \text{Cov}(1_I(X_i), 1_I(X_j)) \\ &= \frac{p-p^2}{n} + \frac{(p_A - p_B)^2}{2n^2} \sum_{0 \leq i < j \leq n-1} \beta_1^{j-i} \\ &= \frac{p-p^2}{n} + \frac{(p_A - p_B)^2}{2n^2} \cdot \frac{(n-1)\beta_1 - n\beta_1^2 + \beta_1^{n+1}}{(1-\beta_1)^2} \end{aligned}$$

where we use the fact that

$$\sum_{1 \leq i < j \leq M} x^{j-i} = \frac{(M-1)x - Mx^2 + x^{M+1}}{(1-x)^2} \quad \square$$

Lemma 4.1

Consider the walk defined in (7). Suppose a recruitment tree is chosen according to a probability distribution λ on the set of n -node trees, and RDS samples are correspondingly collected. Then the variance of \hat{p} satisfies

$$\text{Var}(\hat{p}) = \frac{p-p^2}{n} + \frac{(p_A - p_B)^2}{2n^2} \sum_{k=1}^{n-1} \beta_1^k \mathbb{E}_\lambda L(k)$$

where $L(k)$ is the number of pairs of samples distance k apart.

Proof

Let $l(i, j)$ denote the length of the unique path between X_i and X_j in the recruitment tree. Observe that (B1) shows that

$$\begin{aligned} \mathbb{E}f_D(X_i)f_D(X_j) &= \mathbb{E}(\mathbb{E}[f_D(X_i)f_D(X_j)|l(i, j)]) \\ &= p^2 + \left(\frac{p_A - p_B}{2}\right)^2 \sum_{k=1}^m \beta_1^k \mathbb{P}(l(i, j) = k) \end{aligned}$$

and so

$$\text{Cov}(f_D(X_i), f_D(X_j)) = \left(\frac{p_A - p_B}{2}\right)^2 \sum_{k=1}^n \beta_1^k \mathbb{P}(l(i, j) = k)$$

Now summing the covariance terms, we have

$$\begin{aligned} \sum_{i < j} \text{Cov}(f_D(X_i), f_D(X_j)) &= \left(\frac{p_A - p_B}{2}\right)^2 \sum_{k=1}^{n-1} \beta_1^k \sum_{i < j} \mathbb{P}(l(i, j) = k) \\ &= \left(\frac{p_A - p_B}{2}\right)^2 \sum_{k=1}^{n-1} \beta_1^k \mathbb{E}L(k) \end{aligned}$$

from which the result follows. □

Theorem 4.1

Suppose T_H is a complete binary tree of height $H \geq 1$ (i.e. T_H has $2^{H+1} - 1$ nodes, and each non-leaf node has exactly two children). Let $L_H(k)$ be the number of pairs of nodes distance k apart. Then for $1 \leq k \leq 2H$

$$\frac{1}{4}2^{H+k/2} \leq L_H(k) \leq 2k2^{H+k/2}$$

Proof

For any node n_i of T_H , let $D(n_i) \mapsto \{0, \dots, H\}$ be the height of n_i in the tree. That is, $D(n_i) = 0$ for leaf nodes, and $D(n_j) = D(n_i) + 1$ if n_j is the parent of n_i . In particular, the root has height H . For any node n_i and $k \geq 1$, define $L_H(k, n_i)$ to be the number of pairs of nodes (a, b) that is distance k apart, and such that: (1) the shortest path between a and b includes n_i and (2) n_i is the highest node on that shortest path.

We start by approximating $L_H(k, n_i)$. If $D(n_i) < k/2$, then any path through n_i such that n_i is the highest node on the path has length at most $2D(n_i)$. Consequently, in this case, $L_H(k, n_i) = 0$. Now suppose $k/2 \leq D(n_i)$. First note that n_i cannot be a leaf node. Suppose k is even. Then the left child of n_i has $2^{k/2-1}$ descendants that are distance $k/2$ from n_i . The same is true for the right child of n_i . Consequently, pairing the left and right descendants,

$$L_H(k, n_i) \geq 2^{k/2-1} 2^{k/2-1} = 2^{k-2}$$

If k is odd, then $(k+1)/2 \leq D(n_i)$ since $D(n_i)$ is an integer. Now, the left child of n_i has $2^{(k+1)/2-1}$ descendants that are distance $(k+1)/2$ from n_i and the right child has $2^{(k-1)/2-1}$ descendants that are distance $(k-1)/2$ from n_i . Again pairing the left and right descendants, we have $L_H(k, n_i) \geq 2^{k-2}$.

To upper bound $L_H(k, n_i)$, suppose $k_1, k_2 \geq 1$ and $k_1 + k_2 = k$. Then the left child of n_i has at most 2^{k_1-1} descendants distance k_1 from n_i and the right child of n_i has at most 2^{k_2-1} descendants distance k_2 from n_i (note: if the height of n_i is too small, then, for example, there might be no descendants distance k_1 from n_i , but the upper bounds still hold). Pairing these left and right descendants generates at most 2^{k-2} paths, and adding up all such paths with $1 \leq k_1 \leq k-1$ gives at most $(k-1)2^{k-2}$ paths. Finally, n_i has at most 2^k descendants distance k from itself. Consequently,

$$L_H(k, n_i) \leq (k-1)2^{k-2} + 2^k = (k+3)2^{k-2}$$

If k is even, then there are $2^{H-k/2+1} - 1$ nodes with $D(n_i) \geq k/2$. If k is odd, then there are $2^{H-(k+1)/2+1} - 1$ such nodes. Finally, since every path has a unique highest node,

$$L_H(k) \geq [2^{H-(k+1)/2+1} - 1]2^{k-2} \geq \frac{2^{H+k/2}}{4}$$

and the upper bound follows since

$$L_H(k) \leq [2^{H-k/2+1} - 1](k+3)2^{k-2} \leq 2k2^{H+k/2} \quad \square$$

Theorem 4.2

Consider the walk defined in (7). Suppose the recruitment tree is a complete binary tree of height $H \geq 1$, in which case $n = 2^{H+1} - 1$. If $\beta_1 > \sqrt{2}/2$, then the variance of \hat{p} satisfies

$$\begin{aligned} \frac{p-p^2}{n} + \left[\frac{1}{n^{\log_2 1/\beta_1^2}} - \frac{2}{n} \right] \left(\frac{\beta_1 \sqrt{2}(p_A - p_B)}{32(\beta_1 \sqrt{2} - 1)} \right) &\leq \text{Var}(\hat{p}) \\ &\leq \frac{p-p^2}{n} + \frac{4 \log_2 n}{n^{\log_2 1/\beta_1^2}} \left(\frac{p_A - p_B}{1 - \beta_1 \sqrt{2}} \right)^2 \end{aligned}$$

Proof

The proof relies on evaluating the variance expression given in Lemma 4.1 with the path length bounds of Theorem 4.1. First we show the upper bound by examining the sum in Lemma 4.1. Since $2^H \leq n$, we have

$$\begin{aligned} \frac{1}{2n^2} \sum_{k=1}^{n-1} \beta_1^k \mathbb{E}_\lambda L(k) &\leq \frac{2^H}{n^2} \sum_{k=1}^{2H} k \beta_1^k 2^{k/2} \\ &\leq \frac{1}{2^H} \sum_{k=1}^{2H} k (\sqrt{2} \beta_1)^k \\ &= \frac{2H(\sqrt{2} \beta_1)^{2H+2} - (2H+1)(\sqrt{2} \beta_1)^{2H+1} + 1}{2^H (1 - \sqrt{2} \beta_1)^2} \\ &\leq H \left(\frac{2\beta_1}{1 - \beta_1 \sqrt{2}} \right)^2 \beta_1^{2H} \end{aligned}$$

$$\begin{aligned} &\leq \log_2 n \left(\frac{2}{1 - \beta_1 \sqrt{2}} \right)^2 \beta_1^{2 \log_2 n} \\ &= \log_2 n \left(\frac{1}{1 - \beta_1 \sqrt{2}} \right)^2 \frac{4}{n^{\log_2 1 / \beta_1^2}} \end{aligned}$$

The upper bound now follows. The lower bound is analogous. Since $n < 2^{H+1}$, by Theorem 4.1 we have

$$\begin{aligned} \frac{1}{2n^2} \sum_{k=1}^{n-1} \beta_1^k \mathbb{E}_\lambda L(k) &\geq \frac{1}{32 \cdot 2^H} \sum_{k=1}^{2^H} (\beta_1 \sqrt{2})^k \\ &= \frac{1}{32 \cdot 2^H} \frac{(\beta_1 \sqrt{2})^{2^{H+1}} - \beta_1 \sqrt{2}}{\beta_1 \sqrt{2} - 1} \\ &= \frac{\beta_1 \sqrt{2}}{32(\beta_1 \sqrt{2} - 1)} \left[\beta^{2^H} - \frac{1}{2^H} \right] \\ &\geq \frac{\beta_1 \sqrt{2}}{32(\beta_1 \sqrt{2} - 1)} \left[\frac{1}{n^{\log_2 1 / \beta_1^2}} - \frac{2}{n} \right] \quad \square \end{aligned}$$

APPENDIX C: CONDUCTANCE

In analyzing the network models of Section 3, we see that the key quantity is the spectral gap $1 - \beta_1$. In some sense, β_1 encapsulates the degree of community structure. However, β_1 can be hard to interpret exactly, and direct approximation requires detailed knowledge of the network, which is usually not available in applications. Here we recall the Cheeger inequality [59], a classical result in differential geometry that relates bounds on the second largest eigenvalue β_1 to the geometry of the Markov chain as quantified by *conductance*. This relationship between conductance and the spectral gap is often used in the analysis of Markov chains.

As shown in Simic *et al.*, in Belgrade there is little contact between street-based and agency-based sex workers [46]. Roughly, we quantify the bottleneck between these groups as the probability of cross-group recruitment—high probability of cross-group recruitment corresponds to low segregation.

The following definitions make the notion of a network bottleneck precise. For these definitions, recall that uniform recruitment corresponds to equal edge weights $W(x, y) = 1$.

Definition C.1

Given a weighted graph with nodes V and edge weights W , the transition probability $P_{S \rightarrow S^c}$ from a subset $S \subset V$ to $S^c = V \setminus S$ is

$$P_{S \rightarrow S^c} = \frac{\sum_{x \in S, y \notin S} W(x, y)}{\sum_{x \in S, y \in V} W(x, y)}$$

The definition of $P_{S \rightarrow S^c}$ is not symmetric (e.g. the probability of transition from street-based sex workers to agency-based sex workers is not necessarily the same as that from agency-based sex workers to street-based sex workers). For any partition of the population into sets S and S^c , the conductance of that partition is defined to be

$$I(S, S^c) = \max\{P_{S \rightarrow S^c}, P_{S^c \rightarrow S}\}$$

By definition, the conductance of a partition is symmetric: $I(S, S^c) = I(S^c, S)$. Finally, the conductance I of the entire social network is the conductance of the most segregated partition. In other words, conductance is a measure of how hard it is, in the worst case, to leave a set of nodes.

Definition C.2

The conductance of a weighted graph is

$$I = \min_{S \subset V} I(S, S^c)$$

Returning to the two-group network model of Section 3, consider the network defined in (7) with within-group edge weights $1 - c$ and between-group weights c . For any subset S the denominator in the definition of $P_{S \rightarrow S^c}$ equals $|S|N/2$. Since $W(x, y) \geq c$ we have

$$\sum_{x \in S, y \notin S} W(x, y) \geq |S|(N - |S|)c$$

and consequently,

$$P_{S \rightarrow S^c} \geq \frac{c(N - |S|)}{N/2}$$

Since either $|S| \leq N/2$ or $|S^c| \leq N/2$, this shows that $I(S, S^c) \geq c$, and hence $I \geq c$. Now, for $S = A$ (or $S = B$), we see $I(S, S^c) = c$. Consequently, in this example, the conductance I equals the homophily c .

Further, assuming that no more than half the population is infected, a straightforward calculation shows that for the set of diseased individuals D

$$I(D, D^c) = 1 - \frac{p_A^2 + p_B^2 - c(p_A - p_B)^2}{p_A + p_B} \geq 1 - \frac{p_A^2 + p_B^2}{p_A + p_B}$$

For the values used in our previous examples— $p_A = 0.3$, $p_B = 0.1$, and $c = 0.1$ —it is the case that $I(D, D^c) = 0.76$. In other words, the conductance of the diseased set is large, meaning that it is relatively easy to transition between diseased and healthy individuals. But, the bottleneck elsewhere in the population, namely between groups A and B , significantly increases the variance of the RDS estimates of disease prevalence.

It is also important to note that while it may be possible based on preexisting knowledge to be aware of the bottleneck between groups A and B in the two-group model, it seems more difficult to identify the bottlenecks in the multi-group model. In other words, in many practical situations it may be hard for researchers to know of the bottlenecks.

Conductance is clearly related to the community structure of a network, and hence it is reasonable that it would affect RDS estimates. We have also seen that the variance of RDS estimates is related to the second largest eigenvalue β_1 of the underlying Markov chain. The correspondence between these geometric and algebraic points of view is made precise by Cheeger's inequality.

Theorem C.1 (Cheeger's inequality)

The second largest eigenvalue β_1 and the conductance I are related by

$$1 - 2I \leq \beta_1 \leq 1 - \frac{I^2}{2}$$

For a proof, see Saloff-Coste [50]. Hence, the spectral gap $1 - \beta_1$ satisfies

$$\frac{I^2}{2} \leq 1 - \beta_1 \leq 2I$$

Conductance, however, does not completely determine the variance of RDS estimates. For example, a given target population has a single value of conductance, but the variance of RDS estimates of HIV prevalence may be higher or lower than the variance of RDS estimates of needle sharing prevalence.

ACKNOWLEDGEMENTS

This work was partially supported by the Department of Mathematics at the University of Southern California, the Institute for Social and Economic Research and Policy (ISERP) at Columbia University, and the Applied Statistics Center at Columbia University. The authors thank Edo Airoldi, Andrew Gelman, Doug Heckathorn, Erik Volz, and the anonymous reviewers for helpful comments.

REFERENCES

1. UNAIDS/WHO. *AIDS Epidemic Update*. UNAIDS: Geneva, 2007.
2. UNAIDS/WHO. *Guidelines for Second Generation HIV Surveillance*. UNAIDS/WHO: Geneva, 2000.
3. Pisani E, Lazzari S, Walker N, Schwartlander B. HIV surveillance: a global perspective. *Journal of Acquired Immune Deficiency Syndroms* 2003; **32**(S1):3–11.
4. Rehle T, Lazzari S, Dallabetta G, Asamoah-Odei E. Second-generation HIV surveillance: better data for better decision-making. *Bulletin of the World Health Organization* 2004; **82**(2):121–127. DOI: 10.2471/BLT.00.00000.
5. Magnani R, Sabin K, Saidel T, Heckathorn D. Review of sampling hard-to-reach and hidden populations for HIV surveillance. *AIDS* 2005; **19**(S2):S67–S72. DOI: 10.1097/01.aids.0000172879.20628.e1.
6. Semaan S, Lauby J, Liebman J. Street and network sampling in evaluation studies of HIV risk-reduction interventions. *AIDS Reviews* 2002; **4**:213–223.
7. Malekinejad M, Johnston L, Kendall C, Kerr L, Rifkin M, Rutherford G. Using respondent-driven sampling methodology for HIV biological and behavioral surveillance in international settings: a systematic review. *AIDS and Behavior* 2008; **12**(S1):105–130. DOI: 10.1007/s10461-008-9421-1.
8. Kajubi P, Kanya MR, Raymond HF, Chen S, Rutherford GW, Mandel JS, McFarland W. Gay and bisexual men in Kampala, Uganda. *AIDS and Behavior* 2008; **12**(3):492–504. DOI: 10.1007/s10461-007-9323-7.
9. Johnston LG, Sabin K, Hien MT, Huong PT. Assessment of respondent driven sampling for recruiting female sex workers in two Vietnamese cities: reaching the unseen sex worker. *Journal of Urban Health* 2006; **83**(7):16–28. DOI: 10.1007/s11524-006-9099-5.
10. Platt L, Wall M, Rhodes T, Judd A, Hickman M, Johnston LG, Renton A, Bobrova N, Sarang A. Methods to recruit hard-to-reach groups: comparing two chain referral sampling methods of recruiting injection drug users across nine studies in Russia and Estonia. *Journal of Urban Health* 2006; **83**(7):39–53. DOI: 10.1007/s11524-006-9101-2.
11. Lansky A, Abdul-Quader AS, Cribbin M, Hall T, Finlayson TJ, Garffin RS, Lin LS, Sullivan PS. Developing an HIV behavioral surveillance system for injecting drug users: the National HIV Behavioral Surveillance System. *Public Health Reports* 2007; **122**(S1):48–55.
12. Coleman JS. Relational analysis: the study of social organization with survey methods. *Human Organization* 1958; **17**:28–36.
13. Goodman L. Snowball sampling. *Annals of Mathematical Statistics* 1961; **32**(1):148–170. DOI: 10.1214/aoms/1177705148.

14. Erickson BH. Some problems of inference from chain data. *Sociological Methodology* 1979; **10**:276–302. DOI: 10.2307/270774.
15. Klodahl A. Urban social networks: some methodological problems and possibilities. In *The Small World*, Kochen M. (ed.). Ablex Publishing: Norwood, NJ, 1989; 176–210.
16. Spreen M. Rare populations, hidden populations, and link-tracing designs: what and why? *Bulletin de Méthodologie Sociologique* 1992; **36**:34–58.
17. Snijders TAB. Estimation on the basis of snowball samples: how to weight? *Bulletin de Méthodologie Sociologique* 1992; **36**:59–70.
18. Frank O, Snijders TAB. Estimating the size of hidden populations using snowball sampling. *Journal of Official Statistics* 1994; **10**(1):53–67.
19. Thompson SK, Frank O. Model-based estimation with link-tracing sampling designs. *Survey Methodology* 2000; **26**(1):87–98.
20. Thompson SK, Seber GAF. *Adaptive Sampling*. Wiley: New York, 1996.
21. Thompson SK, Collins LM. Adaptive sampling in research on risk-related behaviors. *Drug and Alcohol Dependence* 2002; **68**:S57–S67. DOI: 10.1016/S0376-8716(02)00215-6.
22. Heckathorn DD, Broadhead RS, Sergeyev B. A methodology for reducing respondent duplication and impersonation in samples of hidden populations. *Journal of Drug Issues* 2001; **31**:543–564.
23. McKnight C, Jarlais DD, Bramson H, Tower L, Abdul-Quader AS, Nemeth C, Heckathorn D. Respondent-driven sampling in a study of drug users in New York City: notes from the field. *Journal of Urban Health* 2006; **83**(7):54–59. DOI: 10.1007/s11524-006-9102-1.
24. Johnston L, Malekinejad M, Kendall C, Iuppa I, Rutherford G. Implementation challenges to using respondent-driven sampling methodology for HIV biological and behavioral surveillance: field experiences in international settings. *AIDS and Behavior* 2008; **12**(S1):131–141. DOI: 10.1007/s10461-008-9413-1.
25. Semaan S, Santibanez S, Garfein R, Heckathorn D, Des Jarlais D. Ethical and regulatory considerations in HIV prevention studies employing respondent-driven sampling. *International Journal of Drug Policy* 2008; **20**(1):14–27. DOI: 10.1016/j.drugpo.2007.12.006.
26. Scott G. ‘They got their program, and I got mine’: a cautionary tale concerning the ethical implications of using respondent-driven sampling to study injection drug users. *International Journal of Drug Policy* 2008; **19**(1):42–51. DOI: 10.1016/j.drugpo.2007.11.014.
27. Abdul-Quader AS, Heckathorn DD, McKnight C, Bramson H, Nemeth C, Sabin K, Gallagher K, DesJarles DC. Effectiveness of respondent-driven sampling for recruiting drug users in New York city: findings from a pilot study. *Journal of Urban Health* 2006; **83**(3):459–476. DOI: 10.1007/s11524-006-9052-7.
28. Heckathorn DD. Respondent-driven sampling II: deriving valid population estimates from chain-referral samples of hidden populations. *Social Problems* 2002; **49**(1):11–34. DOI: 10.1525/sp.2002.49.1.11.
29. Salganik MJ, Heckathorn DD. Sampling and estimation in hidden populations using respondent-driven sampling. *Sociological Methodology* 2004; **34**:193–239. DOI: 10.1111/j.0081-1750.2004.00152.x.
30. Heckathorn DD. Respondent-driven sampling: a new approach to the study of hidden populations. *Social Problems* 1997; **44**(2):174–199. DOI: 10.1525/sp.1997.44.2.03x0221m.
31. Metropolis N, Rosenbluth AW, Rosenbluth M, Teller AH, Teller E. Equations of state calculations by fast computing machines. *Journal of Chemical Physics* 1953; **21**(6):1087–1091. DOI: 10.1063/1.1699114.
32. Gilks W, Richardson S, Spiegelhalter D. *Markov Chain Monte Carlo in Practice*. Chapman & Hall: London, 1996.
33. Kendall WS, Liang F, Wang JS (eds). *Markov Chain Monte Carlo: Innovations and Applications*. World Scientific Publishing: Singapore, 2005.
34. Liu J. *Monte Carlo Strategies in Scientific Computing*. Springer: New York, 2001.
35. Robert CP, Casella G. *Monte Carlo Statistical Methods* (2nd edn). Springer: New York, 2004.
36. Thompson SK. *Sampling* (2nd edn). Wiley: New York, 2002.
37. Marshall A. The use of multi-stage sampling schemes in Monte Carlo computations. In *Symposium on Monte Carlo Methods*, Meyer M (ed.). Wiley: New York, 1956; 123–140.
38. Volz E, Heckathorn DD. Probability-based estimation theory for respondent-driven sampling. *Journal of Official Statistics* 2008; **24**(1):79–97.
39. Thompson SK. Targeted random walk designs. *Survey Methodology* 2006; **32**(1):11–24.
40. McCarty C, Killworth PD, Bernard HR, Johnsen EC, Shelley GA. Comparing two methods for estimating network size. *Human Organization* 2001; **60**(1):28–39.
41. Zheng T, Salganik MJ, Gelman A. How many people do you know in prison?: using overdispersion in count data to estimate social structure in networks. *Journal of the American Statistical Association* 2006; **101**(474):409–423. DOI: 10.1198/016214505000001168.

42. Frank O. Network sampling and model fitting. In *Models and Methods in Social Network Analysis*, Carrington PJ, Scott J, Wasserman S (eds). Cambridge University Press: Cambridge, 2005.
43. Handcock MS, Gile KJ. Modeling networks from sampled data. *Annals of Applied Statistics* 2009; in press.
44. Salganik MJ. Variance estimation, design effects, and sample size calculations for respondent-driven sampling. *Journal of Urban Health* 2006; **83**(7):98–112. DOI: 10.1007/s11524-006-9106-x.
45. McPherson M, Smith-Lovin L, Cook JM. Birds of a feather: homophily in social networks. *Annual Review of Sociology* 2001; **27**:415–444. DOI: 10.1146/annurev.soc.27.1.415.
46. Simic M, Grazina L, Platt L, Baros S, Andjelkovic V, Novotny T, Rhodes T. Exploring barriers to ‘respondent-driven sampling’ in sex workers and drug-injecting sex workers in Eastern Europe. *Journal of Urban Health* 2006; **83**(7):6–15. DOI: 10.1007/s11524-006-9098-6.
47. Wasserman S, Faust K. *Social Network Analysis: Methods and Applications*. Cambridge University Press: Cambridge, U.K., 1994.
48. Newman MEJ. Finding community structure in networks using the eigenvectors of matrices. *Physical Review E* 2006; DOI: 10.1103/PhysRevE.74.036104.
49. Friedman SR, Kottiri BJ, Neaigus A, Curtis R, Vermund SH, Des Jarlais DC. Network-related mechanisms may help explain long-term HIV-1 seroprevalence levels that remain high but do not approach population-group saturation. *American Journal of Epidemiology* 2000; **152**(10):913–922. DOI: 10.1093/aje/152.10.913.
50. Saloff-Coste L. Lectures on finite Markov chains. In *Ecole d’Été de Probabilités de Saint-Flour XXVI*, Bernard P (ed.). Lectures on Probability Theory and Statistics, Lecture Notes in Mathematics, vol. 1665. Springer: Berlin, 1996.
51. Heckathorn DD. Extensions of respondent-driven sampling: analyzing continuous variables and controlling for differential recruitment. *Sociological Methodology* 2007; **37**:151–208. DOI: 10.1111/j.1467-9531.2007.00188.x.
52. Frost SDW, Brouwer KC, Cruz MAF, Ramos R, Ramos ME, Lozada RM, Strathdee CMRSA. Respondent-driven sampling of injection drug users in two U.S.–Mexico border cities: recruitment dynamics and impact on estimate of HIV and Syphilis prevalence. *Journal of Urban Health* 2006; **83**(7):83–97. DOI: 10.1007/s11524-006-9104-z.
53. de Mello M, de Araujo Pinho A, Chinaglia M, Tun W, Júnior AB, Ilário MCFJ, Reis P, Salles RCS, Westman S, Díaz J. Assessment of risk factors for HIV infection among men who have sex with men in the metropolitan area of Campinas City, Brazil, using respondent-driven sampling. *Technical Report*, Population Council, 2008.
54. Johnston LG, Khanam R, Reza M, Khan SI, Banu S, Alam MS, Rahman M, Azim T. The effectiveness of respondent driven sampling for recruiting males who have sex with males in Dhaka, Bangladesh. *AIDS and Behavior* 2008; **12**(2):294–304. DOI: 10.1007/s10461-007-9300-1.
55. Gelman A, Carlin JB, Stern HS, Rubin DB. *Bayesian Data Analysis* (2nd edn). Chapman & Hall: Boca Raton, 2004.
56. Gelman A, Rubin DB. Inference from iterative simulation using multiple sequences. *Statistical Science* 1992; **7**(4):457–471. DOI: 10.1214/ss/1177011136.
57. Brooks SP, Gelman A. General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics* 1998; **7**(4):434–455. DOI: 10.2307/1390675.
58. Lynch SM. *Introduction to Applied Bayesian Statistics and Estimation for Social Scientists*. Springer: New York, 2007.
59. Cheeger J. A lower bound for the smallest eigenvalue of the Laplacian. *Symposium in Honor of S. Bochner*. Princeton University Press: Princeton, NJ, 1970; 195–199.