RESPONDENT-DRIVEN SAMPLING AS MARKOV CHAIN MONTE CARLO

SHARAD GOEL AND MATTHEW J. SALGANIK

Abstract. Respondent-driven sampling (RDS) is a recently introduced, and now widely used, technique for estimating disease prevalence in hidden populations. The sample is collected through a form of snowball sampling where current sample members recruit future sample members. In this paper we observe that respondent-driven sampling can be viewed as Markov chain Monte Carlo (MCMC) importance sampling. By establishing this connection, we are able to draw on the MCMC literature to address key RDS implementation and analysis issues. It was believed that the social network structure of the hidden population affected both the bias and variance of RDS estimates, but the precise nature of the relationship was unknown. Here we clarify that intuition by relating both to the second largest eigenvalue of the network transition matrix. In particular, we show segregation within the population effectively reduces sample size. We also show that sample size is effectively reduced by a sample design feature which allows sample members to recruit multiple future sample members. The paper concludes with suggestions for implementing and evaluating results from respondent-driven sampling studies.

1. Introduction

Researchers in the social and biological sciences often wish to estimate the prevalence of some behavior or characteristics in a particular social group. When a sampling frame for that community exists, well-established methods can be used. For many populations of interest, however, no such sampling frame exists. Examples of these so-called "hidden" or hard-to-reach populations include injection drug users, men who have sex with men, and commercial sex workers; three groups that are of particular interest to public health researchers working to understand and control the spread of HIV (Magnani et al., 2005).

In studying hidden populations, researchers have been forced to apply problematic sampling methods that in many situations lead to estimates of unknown bias and variance (for a review see Semaan et al. (2002) or Magnani et al. (2005)). Respondent-driven sampling (RDS) (Heckathorn, 1997, 2002; Salganik & Heckathorn, 2004) is a new technique that addresses many of the shortcomings of previous methodologies. Despite being introduced quite recently, respondent-driven sampling is attracting a great deal of interest in the public health community, mostly in the context of HIV research. Already scholars have field-tested respondent-driven sampling to assess its feasibility in a range of settings including injection drug users in North America (Abdul-Quader et al., 2006; McKnight et al., 2006; Frost et al., 2006; Robinson et al., 2006), Asia (Yeka et al., 2006), Eastern Europe (Simic et al.,

Date: August 22, 2007.

Key words and phrases. homophily, importance sampling, Markov chain Monte Carlo, respondent-driven sampling, spectral gap, variance estimation.

2006), and the former Soviet Union (Platt et al., 2006); sex workers in Asia (Johnston et al., 2006; Yeka et al., 2006) and Eastern Europe (Simic et al., 2006); men who have sex with men in North America (Ramirez-Valles et al., 2005); as well as several other populations (Heckathorn & Jeffri, 2003; Wang et al., 2005, 2006). In addition, other studies have compared the results of respondent-driven sampling to other more established sampling methods (Ramirez-Valles et al., 2005; Robinson et al., 2006; Platt et al., 2006). The results of these studies have generally—but not always—been positive, leading the Center for Disease Control and Prevention (CDC) to use respondent-driven sampling for the National HIV Behavioral Surveillance (NHBS) studies of injection drug users (Abdul-Quader et al., 2006; Lansky et al., 2007). This multi-site, multi-year program will generate additional RDS data, and will likely further stimulate the use of RDS among academics and governmental organizations.

A respondent-driven sample is selected using a chain-referral method in which current sample members recruit future sample members. This type of sampling is also called snowball, random-walk, or link-tracing (Coleman, 1958; Goodman, 1961; Erickson, 1979; Klovdahl, 1989; Spreen, 1992; Snijders, 1992; Frank & Snijders, 1994; Thompson & Frank, 2000), and can be considered a form of adaptive sampling (Thompson & Seber, 1996; Thompson & Collins, 2002). To collect a respondent-driven sample, one begins by selecting a set of seeds in the target population. After participating, the seeds are asked—and 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. A more detailed description of the sampling method, and its relation to other methods is available elsewhere (Heckathorn, 1997; Salganik & Heckathorn, 2004; Magnani et al., 2005; McKnight et al., 2006). The sample process leads to recruitment networks like the one in Figure 1 that resulted from a study of drug users in New York City (Abdul-Quader et al., 2006). The sample began with eight seeds and grew to include 618 drug users within 13 weeks.

Early RDS estimation was based on an indirect procedure that used the sample to make inferences about the target population's social network (Heckathorn, 2002; Salganik & Heckathorn, 2004), which in turn was used to make prevalence estimates. A new estimator proposed by Volz & Heckathorn (2007) allows for direct estimation from sample to population without the intermediate network step. Their estimator allows for estimating both prevalence (What proportion of drug injectors in Atlanta have HIV?) as well as means of real valued functions (What is the average injection frequency of drug injectors in Atlanta?).

This paper connects respondent-driven sampling to Markov chain Monte Carlo, and analyzes how RDS estimates are affected by both the community structure of a hidden population and the recruitment procedure. We find that sample size is effectively reduced by: 1) social network segregation within the hidden population, and 2) a study design in which participants recruit multiple individuals. In Section 2 we show that RDS sampling and estimation can be viewed as Markov chain Monte Carlo (MCMC) importance sampling. While MCMC algorithms are typically computer-driven, a novel feature of RDS is that state transitions consist of individuals physically recruiting others in the hidden popultion. In Section 3 we

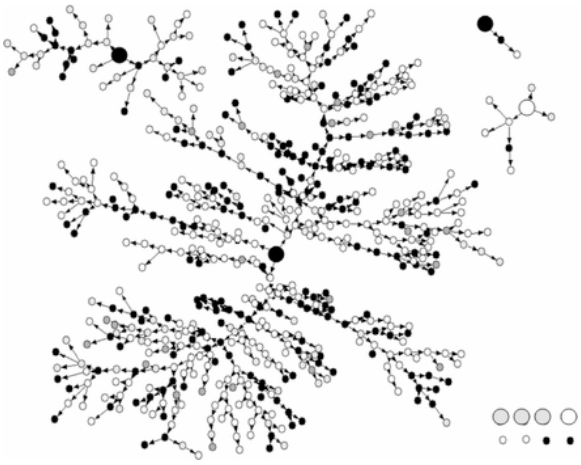

FIGURE 1. Recruitment networks from a study of drug users in New York City. The eight seeds are larger than the others nodes and all nodes are color coded by race as follows: Non-Hispanic black = white, Non-Hispanic White = light gray, Hispanic = black, Other = Dark gray. This figure was originally published in Abdul-Quader et al. (2006).

draw on the established body of MCMC research to analyze a particular, illustrative example in detail. We show the hidden population's social network structure affects both the bias and variance of RDS estimates. Importantly, as that example shows, segregation in any part of the network may affect RDS estimates of quantities that are not directly related to the source of segregation. For example, racial segregation may degrade RDS estimates of gender composition because segregation anywhere increases the correlation between RDS samples, effectively decreasing sample size—an observation that was previously overlooked in the RDS literature. We furthermore explore the effect of multiple recruitments on bias and variance, an issue that is important for RDS, but has not been considered previously and does not typically arise in traditional MCMC applications. We show that "thick," as opposed to "thin," recruitment chains increase the statistical dependence between samples, and consequently worsen RDS estimates. Section 4 shows that results derived for the particular example of Section 3 are qualitatively quite general. Section 5 concludes with recommendations for users of respondent-driven sampling and suggestions for future research. We have relegated most proofs and technical details to appendices.

2. Respondent-Driven Sampling as MCMC

Respondent-driven sampling (Heckathorn, 1997, 2002; Salganik & Heckathorn, 2004) is a form of snowball sampling 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 Markov chain Monte Carlo importance sampling, and make the connection to RDS precise.

Markov Chain Monte Carlo. Markov chain Monte Carlo was popularized by the introduction of the Metropolis algorithm (Metropolis et al., 1953), and has been extensively applied in a variety of fields, including physics, chemistry, biology and statistics. MCMC has also been the subject of several book-length treatments, e.g. Gilks et al. (1996); Kendall et al. (2005); Liu (2001).

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 NYC). 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:

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

In terms of RDS, $K(v_i, v_j)$ is the probability that any individual v_i recruits another 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 at v_i. Under this assumption, there is a unique distribution $\pi : V \to \mathbb{R}$—called the *stationary distribution*—satisfying

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

That is, if $X_0, X_1, X_2, \ldots$ 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.

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 $f : V \to \mathbb{R}$, the sample mean

$$\frac{1}{n} \sum_{i=0}^{n-1} f(X_i) \tag{2.1}$$

gives an unbiased estimate not of the uniform mean, but of $\mathbb{E}_\pi f = \sum_{i=1}^{n} f(v_j)\pi(v_j)$. The insight behind importance sampling Marshall (1956) is that the weighted sample mean

$$\frac{1}{n} \sum_{i=0}^{n-1} \frac{f(X_i)}{N \cdot \pi(X_i)} \tag{2.2}$$

produces an unbiased estimate of the uniform 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 importance sampling can be used to estimate the infected population proportion $p = |D|/N = \mu_f$ by setting $f(v_i) = 1$ if $v_i \in D$ and $f(v_i) = 0$ otherwise.

It is often convenient to replace (2.2) by the asymptotically unbiased importance sampling estimator

$$\hat{\mu} = \frac{1}{\sum_{i=0}^{n-1} 1/\pi(X_i)} \sum_{i=0}^{n-1} \frac{f(X_i)}{\pi(X_i)}. \tag{2.3}$$

The considerable advantage of (2.3) over (2.2) is that the importance weights $1/\pi(X_i)$ only need to be evaluated up to a multiplicative constant. In many applications, including RDS, this simplification is essential.

Respondent-Driven Sampling. Importance sampling allows estimation of p given samples $X_0, X_1, \ldots$ from any fixed distribution π. RDS generates such samples via Markov chain Monte Carlo. The link between respondent-driven sampling and Markov chain Monte Carlo has been hinted at in the literature (Salganik & Heckathorn, 2004; Thompson, 2006; Volz & Heckathorn, 2007). Here we make that connection explicit.

Consider a social network $G = (V, E)$ where nodes $x \in V$ represent individuals in the population, and are either infected or healthy. We assume symmetric weighted edges $e \in E$ in the graph connect friends, and we write $W(x, y) = W(y, x)$ for the weight of the edge between nodes x and y. It is imperative that the network be connected: There must exist a path of mutual friends connecting every two individuals in the population. For some populations (e.g. drug users) this is reasonable, for others (e.g. people who cheat on their taxes) this may not hold.

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 $W_{\{x\}}$.

Random walk on the weighted graph G is defined by the kernel $K(x, y) = W(x, y)/W_x$. In the setting of RDS, $K(x, y)$ is the probability that individual x recruits individual y. Assuming the network is connected, the walk has a unique stationary distribution

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

Consequently, for $X_0, X_1, X_2, \ldots$ a realization of the chain with $X_0 \sim \pi$, and $f : V \to \mathbb{R}$ any function, the importance sampling estimator (2.3) of the uniform mean μ_f reduces to

$$\hat{\mu}_f = \frac{1}{\sum_{i=0}^{n-1} 1/W_{X_i}} \sum_{i=0}^{n-1} \frac{f(X_i)}{W_{X_i}} \tag{2.4}$$

(W_V in the numerator and denominator cancel). The RDS estimator (2.4) was recently introduced in Volz & Heckathorn (2007), and will likely supplant the original RDS estimator introduced in Salganik & Heckathorn (2004). In the particular case of estimating disease prevlance, by setting $f(v_i) = 1$ if v_i is infected and $f(v_i) = 0$

otherwise, (2.4) simplifies to

$$\hat{p} = \frac{1}{\sum_{i=0}^{n-1} 1/W_{X_i}} \sum_{X_i \text{ infected}} \frac{1}{W_{X_i}}. \tag{2.5}$$

To evaluate the RDS estimators (2.4) and (2.5) 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 friends uniformly at random. In this case, W_x equals the degree, i.e. number of friends, of node x. Ascertaining the size of an individual's social network is itself a challenging problem (see e.g. Marsden (1990); McCarty et al. (2001); Zheng et al. (2006)), and is a potential source of non-sampling error.

In contrast to estimates from classical snowball sampling (see e.g. Erickson (1979); Spreen (1992)), which are based on the sample mean (2.1), RDS estimates weight samples proportional to their assumed probability of selection. In the case where all nodes have the same degree, this snowball estimate is equivalent to the RDS estimate (2.4) together with the uniform recruitment assumption.

In the above, we start the walk in stationarity: $X_0 \sim \pi$. Assuming the network has at least one triangle, the RDS estimator $\hat{\mu}$ is asymptotically unbiased regardless of the starting distribution. Moreover, there is a central limit theorem for $\hat{\mu}$:

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

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 (see e.g. Liu (2001)).

2.1. **Estimating Distributions.** Respondent-driven sampling is generally applied to study incidence rates of specific diseases or behavior. The sampling process, however, also gives information about the distributions of individual characteristics, e.g. the distribution of age in the hidden population. Suppose each node is classified into exactly one of k categories $\{1, 2 \ldots, k\}$, and denote the category of node v by $C(v)$. Define $f_i : V \to \mathbb{R}$

$$f_i(v) = \begin{cases} 1 & C(v) = i \\ 0 & \text{otherwise} \end{cases}.$$

Then the uniform mean of f_i satisfies $\mu_{f_i} = p_i$ where p_i is the proportion of nodes in category i. By (2.4),

$$\hat{p}_i = \frac{1}{\sum_{j=0}^{n-1} 1/W_{X_j}} \sum_{C(X_j)=i} \frac{1}{W_{X_j}} = \frac{1}{Z} \sum_{C(X_j)=i} \frac{1}{W_{X_j}}$$

where

$$Z = \sum_{c=1}^{k} \sum_{C(X_j)=c} \frac{1}{W_{X_j}}$$

is such that $\hat{p}_1 + \cdots + \hat{p}_k = 1$.

In the case of estimating degree distribution, i.e. $C(v) = \deg(v)$, under the uniform recruitment assumption $W_{X_j} = \deg(X_j)$, we have the simplified expression

$$\hat{p}_i = \frac{1}{Z} \frac{\#\{X_j : \deg(X_j) = i\}}{i}.$$

3. Effects of Segregation on RDS estimates: An Example

Social network ties tend to form between similar people, a tendency that sociologists call homophily (McPherson et al., 2001). One consequence of homophily is that social networks generally have "community structure," that is, they often can be partitioned into relatively homogenous groups where there are frequent ties within groups, but few ties between groups (Newman, 2006). We show this network structure can increase the bias and variance of RDS estimates by effectively reducing the sample size. Before establishing general results, we develop results for a particularly simple network that allows us to isolate this network effect clearly. Our example, while motivated by the qualitative features of real social networks, is not intended to be an "accurate" model of any particular network. Rather, it is intended to provide insight by allowing for exact and interpretable results. In Section 4 we generalize these results beyond this specific example.

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 $0 < c < 1/2$. That is, within-group associations are stronger than between-group associations. We parameterize homophily, by c—as c decreases the tendency for within-groups ties becomes stronger. Note that this definition of homophily is not entirely consistent with other definitions (for example, Abdul-Quader et al. (2006); Heckathorn (2002)), but it links naturally to the general notion of *conductance* which is more fully described in Section 4.

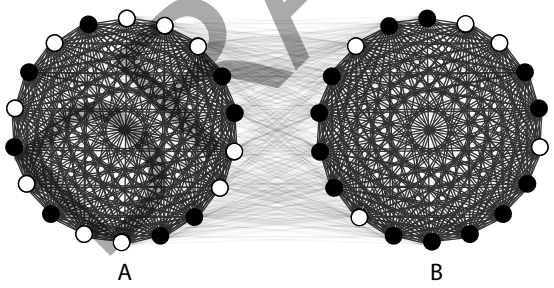

Figure 2. The population is divided into two equally sized groups, A and B. There is an edge between every pair of nodes in the network, but within-group edges have higher weight than between-group edges. White indicates infected nodes.

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 the network segregation affects the RDS estimate $\hat{p}$ of the infected population proportion p. We can imagine this by considering the case of street-based and agency-based sex workers in Belgrade, two groups that have been found to have little contact (Simic et al., 2006). If these two groups had different rates of HIV infection, then the weak connections between the groups could lead to both bias and high variance for the RDS estimated HIV prevalence for sex-workers as a whole.

In our example situation, RDS is based on the following Markov chain:

$$K(x,y) = \begin{cases} 2(1-c)/N & x,y \in A \text{ or } x,y \in B \\ 2c/N & x \in A, y \in B \text{ or } x \in B, y \in A \end{cases} . \tag{3.1}$$

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

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

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 (2.5) 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, 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 chain after l steps.

Lemma 3.1. *For $0 < c < 1/2$, the l-step distribution of the walk defined at* (3.1) *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 in-group recruitment. For example, if the initial seed is chosen from A, then for $c = .1$, after 5 steps the chain is still about twice as likely to be in A than in B.

Here, β_1, the second largest eigenvalue of the transition matrix K, is seen to control the rate of convergence of the chain to its equilibrium distribution. This phenomenon is true for general chains (see Section 4), and as we show below, β_1 also affects both the bias and variance of the population proportion estimate.

3.1. **Bias and Variance.** 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 degree of network segregation that is induced by the homophily parameter c and the length of the recruitment chains. This illustrates the fact that the initial seed becomes important in segregated populations.

Lemma 3.2. *Consider the walk defined at* (3.1)*. 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 segregation.

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 $c \to 0$, i.e. as the two populations become completed segregated, $\text{bias}(\hat{p}) \to (p_A - p_B)/2$. This makes sense because 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.

A key point is that unlike simple random samples, the samples in an RDS study are *dependent*. In a segregated population, it is more likely that individuals refer people who are in their same social subgroup. Intuitively, in this situation, we gain less information than if that recruit were chosen from the entire population. The result of this dependence is an effective reduction in the sample size which increases the variance of the estimates no matter how the initial seeds are chosen. The dependence between samples is quantified by their covariance.

Lemma 3.3. *Consider the walk defined at* (3.1)*. Suppose* $X_0^1, X_1^1, \ldots$ *and* $X_0^2, X_1^2, \ldots$ *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* $i, j \geq 0$

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

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* $X_0, X_1, \ldots$ *defined at* (3.1)*. If* $X_0 \sim \pi$*, then the variance of* $\hat{p}$ *satisfies*

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

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

Again we see the spectral gap $1 - \beta_1$ affects RDS estimates. A naive estimate of the variance assumes samples are uncorrelated, yielding only the first term $(p - p^2)/n$. In particular, it does not take into account possible segregation in the hidden

population. The simple variance estimate $V(\hat{p}) = (p - p^2)/n$ is related to the true variance by the design effect (Lohr, 1999)

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

For example, for $c = .1$, $p_A = .3$ and $p_B = .1$, $\text{Var}(\hat{p}) \approx 1.5 \times V(\hat{p})$. Accordingly, confidence intervals determined by the true variance are $\sqrt{1.5} \approx 1.2$ times wider. Put another way, segregation in this example effectively reduces sample size by a third: A sample size of 500 collected via RDS corresponds to a sample size of 335 collected via simple random sampling.

3.2. **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 allowed to recruits multiple individuals, as seen in recruitment chains from the New York City study of drug injectors (Figure 1). With multiple recruitment, chain lengths are shortened, even if total sample size is large. Consequently, there is significant dependence between all samples, increasing the variance of RDS estimates—a finding that was previously overlooked in the RDS literature.

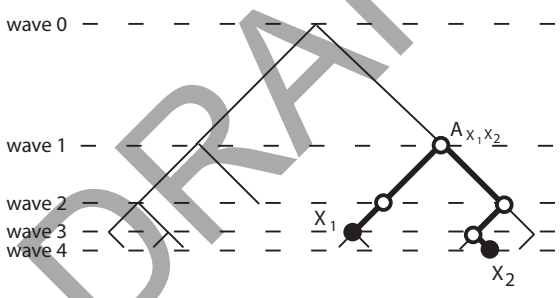

FIGURE 3. $A_{x_1x_2}$ is the most recent common ancestor of X_1 and X_2, and consequently $\text{Cor}(f_D(X_1), f_D(X_2)) = \left(\frac{p_A - p_B}{2}\right)^2 \beta_1^{2+3}$.

We assume, again, 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 3, observe that $A_{x_1x_2}$ 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 $A_{x_1x_2}$, 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.

	Number of recruits			
	0	1	2	3
Probability	1/3	1/6	1/6	1/3

TABLE 1. Multiple recruitment offspring distribution based on a study of injection drug users in Tijuana and Ciudad Juarez (Frost et al., 2006).

Lemma 3.4. *Consider the walk defined in* (3.1)*. Suppose a recruitment tree is chosen according to a probability distribution ν on the set of n-node trees, and RDS samples are correspondingly collected. Then the variance of $\hat{p}$ satisfies*

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

where L_k is the set of pairs of samples distance k apart.

Lemma 3.4 shows that, for a given social network structure, the further apart the samples the lower the variance. That is, "thin," as opposed to "thick," recruitment chains lead to improved estimates. For example, consider a recruitment graph that is generated as a branching process with offspring distribution based on data from the Frost et al. (2006) study of injection drug users in Tijuana and Ciudad Juarez. 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 sixth recruited two others, and the remaining third recruited three other participants (table 1). In this case, it seems difficult to find an analytic expression for $\mathbb{E}|L_k|$, but simulation reveals that for our common parameter values ($c = 0.1$, $p_A = 0.1$, and $p_B = 0.3$), and a sample size of $n = 500$, we have $\mathrm{Var}(\hat{p}) \approx 3.7 \times V(\hat{p})$. In other words, segregation 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.

3.3. **Summary.** The examples discussed in this section shows the effect of both the social network and the recruitment network on RDS estimates in a simplified setting which reasonably maps onto many settings in which RDS may be used. To summarize our findings, for this hypothetical population we compare three sampling situations: simple random sampling, RDS with single recruitment, and RDS with multiple recruitment. We use parameters $p_A = 0.1$, $p_B = 0.3$, $c = 0.1$ and $n = 500$.

Simple random sampling. Since $p = 0.2$, the variance of $\hat{p}$ is $(p - p^2)/500 = .00032$ and its standard deviation is approximately .0179. Consequently, the 95% confidence interval for the estimate is approximately $\hat{p} \pm 3.5\%$. The variance, in this case, was independent of the network structure.

RDS – Single Recruitment. For $c = 0.1$ the second largest eigenvalue satisfies $\beta_1 = 1 - 2c = .8$. If the samples are the result of a single, long chain (without multiple recruitment) in which $X_0 \sim \pi$, then $\mathrm{Var}(\hat{p})$ is given by Corollary 3.1, which yields a standard deviation of approximately .0219. The 95% confidence

	95% CI	Effective Sample Size
Simple random sampling	$\hat{p} \pm 3.5\%$	500
RDS – Single Recruitment	$\hat{p} \pm 4.3\%$	335
RDS – Multiple Recruitment	$\hat{p} \pm 6.7\%$	136

TABLE 2. Results for $p_A = .1$, $p_B = .3$, $c = 0.1$ and $n = 500$.

interval is then $\hat{p} \pm 4.3\%$, the same interval one would get from drawing 335 simple random samples.

RDS – Multiple Recruitment. Assume multiple recruitment that follows a branching process with offspring distribution based on the recruitment data from the Frost et al. (2006) study of injection drug users in two Mexican cities (see table 1). Simulation shows that $(1/n) \sum_{k=1}^{n-1} \beta_1^k \mathbb{E}|L_k| \approx 21.3$. Lemma 3.4 then shows the standard deviation of the estimate is approximately .0343, yielding a confidence interval $\hat{p} \pm 6.7\%$. This level of confidence corresponds to effectively 136 simple random samples, or 204 samples collected via RDS with single recruitment.

4. General Bounds on Variance

Now we move beyond the specific example of Section 3, and consider the general case. For computer-based applications, there are several methods for estimating the variance of Markov chain Monte Carlo estimators (see e.g. Geyer (1992); Kendall et al. (2005)), which are in turn used to construct confidence intervals. For example, window estimators of the lagged autocovariance, and the method of batch means. It is, however, unclear to what extent those estimators can be adapted to RDS, where sample sizes are small (≈ 500) and chain lengths short (≈ 10). A method of inference that relies on simulating multiple MCMC sequences is developed in Gelman & Rubin (1992); Brooks & Gelman (1998), and seems to be a particularly fruitful direction to explore given the multiple seed design aspect of RDS. Also, a bootstrapping technique specifically designed for RDS is discussed in Salganik (2006). In this section we take another approach: We apply results that relate variance to the geometry of the underlying social network. In general, this tack facilitates only crude bounds, but offers insight into what features of the social network structure affect RDS estimates.

In Section 4.1, we review the relationship between variance and the spectral gap $1 - \beta_1$. In Section 4.2, we connect the variance of RDS estimates to conductance, a measure of network segregation that is estimable from RDS data. See the beginning of Appendix A for definitions of the terms used in this section.

4.1. Eigenvalue Bounds.

Here we apply an analytic result of Diaconis & Basseti (2007) to express the variance of RDS estimators for general walks on networks with symmetric edge weights in terms of the eigenvalues and eigenfunctions of the transition matrix.

Theorem 4.1 (Diaconis & Basseti (2007))**.** *Consider a reversible Markov chain* (K, π) *with orthonormal basis of eigenfunctions* $\psi_0, \psi_1, \ldots, \psi_{N-1}$ *and associated eigenvalues*

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

Let $X_0^1, X_1^2, \ldots$ and $X_0^2, X_1^2, \ldots$ be independent realizations of the K chain with $X_0^1 = X_0^2 \sim \pi$. Then for any function g

$$Cov(g(X_i^1), g(X_j^2)) = \sum_{k=1}^{N-1} a_k^2 \beta_k^{i+j} \tag{4.1}$$

where $a_k = \langle g, \psi_k \rangle_\pi$. In particular,

$$Cov(g(X_i^1), g(X_j^2)) \leq \beta_1^{i+j} \, Var_\pi g.$$

For our recurring example defined at (3.1), there are only two non-zero eigenvalues, i.e. $\beta_k = 0$ for $k \geq 2$. Since $\psi_1 = 1_A - 1_B$ (see the proof of Lemma 3.1), for f_D indicating infected individuals we have $a_1 = (p_A - p_B)/2$. Consequently,

$$\mathrm{Cov}(f_D(X_i^1), f_D(X_j^2)) = \left(\frac{p_A - p_B}{2} \right)^2 \beta_1^{i+j} \tag{4.2}$$

as was also shown by the probabilistic argument of Lemma 3.3. The covariance bound of Theorem 4.1 and the covariance of our example network given in (4.2) are quite similar in that the second largest eigenvalue is the dominant term.

In general, as the spectral gap decreases, i.e. as β_1 approaches 1, the covariance between samples increases. In this sense, the network structure as characterized by the spectral gap is crucial for determining the quality of RDS estimates. Furthermore, when recruitment chains are thick, as opposed to thin, samples are typically connected via short paths in the recruitment tree, resulting in a smaller exponent $i + j$ in (4.1), and accordingly larger covariance between samples.

4.2. **Bounding Eigenvalues by Conductance.** The bounds of Theorem 4.1 were in terms of the eigenvalues and eigenfunctions of the Markov chain derived from the underlying social network. These bounds can be hard to interpret, and require detailed knowledge of the network which is usually not available in applications. Here we recall the Cheeger inequality (Cheeger, 1970), 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*. One of the early applications of this technique to finite Markov chains was in the analysis of an MCMC algorithm for approximating the permanent of a 0-1 matrix (Jerrum & Sinclair, 1989).

As shown in Simic et al. (2006), in Belgrade there is segregation between street-based and agency-based sex workers. Roughly, we quantify segregation between these groups as the probability of between-group recruitment—high probability corresponds to low segregation. Although an exact answer depends on knowledge of the entire social network, in Section 4.3 we discuss an estimation procedure based on data readily available in RDS studies. The following definitions make precise this notion of segregation; uniform recruitment corresponds to $W(x, y) = 1$.

Definition 4.1. Given a weighted graph with nodes V and edge weights W, the transition probability $P_{S \to S^c}$ from a subset $S \subset V$ to $S^c = V \setminus S$ is

$$P_{S \to S^c} = \sum_{x \in S, y \notin S} W(x, y) \Big/ \sum_{x \in S, y \in V} W(x, y).$$

The definition of $P_{S \to S^c}$ 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

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

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 4.2. The conductance of a weighted graph is

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

Returning to the example of Section 3, consider the network defined at (3.1) with within-group edge weights $1 - c$ and between-group weights c. For any subset S the denominator in the definition of $P_{S\to S^c}$ 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,

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

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.

Notably, assuming 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 $p_A = .3$, $p_B = .1$ and $c = .1$, $I(D, D^c) = .76$. In particular, the conductance of the diseased set is large regardless of the extent of segregation between the two groups A and B. Although there is no direct segregation between infected and healthy individuals, a bottleneck elsewhere in the population (namely, between A and B) significantly impacts RDS estimates of the disease prevalence. Previous work on RDS has overlooked this observation.

4.3. **Estimating Conductance.** Although calculating conductance exactly requires detailed knowledge of the social network structure, it can nonetheless be estimated from data readily available in RDS studies. Instead of sampling individuals, one can view the RDS procedure as sampling network ties, i.e. edges. Consider the *random directed edges* $E_i = (X_i, X_{i+1})$, $0 \leq i \leq n - 2$ where $X_0, \ldots, X_{n-1}$ are samples derived from the Markov chain started in equilibrium. For any given edge $e = (v_a, v_b)$

$$\begin{aligned}\mathbb{P}(E_i = e) &= \pi(v_a)K(v_a, v_b) \\ &= \frac{W_{v_a}}{W}\frac{W(v_a, v_b)}{W_{v_a}} \\ &= \frac{W(v_a, v_b)}{W}.\end{aligned}$$

That is, each directed edge is chosen with probability proportional to its weight. Under the uniform recruitment assumption, each edge is equally likely to be chosen.

This process of selecting edges is itself a Markov chain on the set of edges, with stationary distribution proportional to edge weight.

To estimate conductance, first consider the sample statistics $e_{S\to S^c} = \#\{E_i : X_i \in S, X_{i+1} \notin S\}$, $e_{S^c\to S} = \#\{E_i : X_i \notin S, X_{i+1} \in S\}$, and $m = \#\{X_i \in S\}$. Now, $(e_{S\to S^c})/(n-1)$ estimates the probability of sampling a node in S who then recruits a node in S^c. That is,

$$\mathbb{E}\left[\frac{e_{S\to S^c}}{(n-1)}\right] = \sum_{x\in S, y\notin S} \frac{W_x}{W}\cdot\frac{W(x,y)}{W_x} = \sum_{x\in S, y\notin S}\frac{W(x,y)}{W}$$

By symmetry of edge weights, $e_{S^c\to S}/(n-1)$ estimates that same probability, i.e.

$$\mathbb{E}\left[\frac{e_{S^c\to S}}{(n-1)}\right] = \sum_{x\notin S, y\in S}\frac{W(x,y)}{W} = \sum_{x\in S, y\notin S}\frac{W(x,y)}{W}.$$

Averaging these two estimates for efficiency, we have

$$\mathbb{E}\left[\frac{e_{S\to S^c}+e_{S^c\to S}}{2(n-1)}\right] = \sum_{x\in S, y\notin S}\frac{W(x,y)}{W}. \tag{4.3}$$

Furthermore, m/n estimates the probability of sampling a node in S, i.e.

$$\mathbb{E}\left[\frac{m}{n}\right] = \sum_{x\in S, y\in V}\frac{W_x}{W}\cdot\frac{W(x,y)}{W_x} = \sum_{x\in S, y\in V}\frac{W(x,y)}{W}. \tag{4.4}$$

Since $P_{S\to S^c}$ is the conditional probability of recruiting a node in S^c starting from a node in S, we take the ratio of (4.3) to (4.4), defining the estimator

$$\hat{P}_{S\to S^c} = \frac{e_{S\to S^c}+e_{S^c\to S}}{2m}.$$

Note that $\hat{P}_{S\to S^c}$ is a consistent estimator by the law of large numbers for Markov chains (see e.g. Liu (2001)). Finally, since $I(S,S^c) = \max\{P_{S\to S^c}, P_{S^c\to S}\}$, we have the estimator

$$\hat{I}(S,S^c) = \frac{e_{S\to S^c}+e_{S^c\to S}}{2\min\{m, n-m\}}. \tag{4.5}$$

While (4.5) only estimates the conductance for a given partition (S,S^c), the complete network conductance is the minimum of $I(S,S^c)$ taken over all partitions. Although it is difficult to directly approximate this minimum, it seems reasonable to focus on partitions tied to demographic characteristics known to create bottlenecks, e.g. race, gender, religion, class, occupation, neighborhood, etc.

Intuitively, it is clear that bottlenecks increases the variance of RDS estimates. The connection to eigenvalues is given by Cheeger's inequality.

Theorem 4.2 (Cheeger's inequality)**.** *The second largest eigenvalue β_1 and the conductance I are related by*

$$1-2I \le \beta_1 \le 1-\frac{I^2}{2}.$$

(For a proof, see e.g. Saloff-Coste (1996).) Hence, the spectral gap $1-\beta_1$ satisfies

$$\frac{I^2}{2} \le 1-\beta_1 \le 2I.$$

Conductance is related to segregation, and hence reasonable that it would affect RDS estimates. Cheeger's inequality makes precise this relationship by linking

conductance to the spectral gap. In RDS studies, it would be valuable to include estimates of conductance as a means of quantifying social network segregation.

5. Conclusion

In this paper we connect RDS to MCMC, and show that the geometry of both the social network and the recruitment network affects RDS estimates. We conclude by describing some of the implications of these findings for the practice of RDS.

Segregation. When undertaking a study using RDS, it is necessary to consider whether the procedure allows for sufficiently accurate estimates given the network structure of the population and the proposed study budget. In particular, past work focussed on segregation between infected and healthy individuals. Segregation anywhere in the network, however, impacts the quality of RDS estimates. Thus, conductance, as discussed in Section 4.2, is a useful measure for quantifying the extent of segregation in a hidden population. We hope future theoretical and empirical work continues to explore the design effect of RDS as a function of network structure, particularly taking care to develop procedures which require only limited information about the network.

Multiple recruitment. The multiple recruitment feature of RDS was developed to help ensure that sampling chains not die out even when some subjects did not recruit. However, this design feature may greatly diminish the accuracy of RDS estimates by increasing the correlation between sample units, an observation previously overlooked in the literature. Since the specific structure of recruitment chains impacts RDS estimates, by themselves larger sample sizes are not always more accurate than smaller samples sizes, contrary to intuition from simple random sampling. While it is currently common practice to provide subjects with 3 recruitment coupons each, RDS would benefit from techniques that make it practical to reduce that number. It would be useful to continue investigating the trade-off between allocating resources to produce thin (as opposed to thick) chains, and larger samples.

Monitoring convergence. As with computer-based MCMC, it would be useful to monitor the convergence of RDS estimates during the data collection phase, something that is not currently done. The use of multiple seeds, a common design feature, creates parallel chains that should make it possible to monitor convergence by adapting existing MCMC procedures such as Gelman & Rubin (1992); Brooks & Gelman (1998). If estimates do not seem to be converging, researchers could attempt to collect a larger sample than originally planned.

Burn in. During the data analysis phase, researchers should consider discarding data from early sample waves, just as researchers using computer-driven MCMC often discard the first half of their draws during the so called "burn-in" phase (Gelman et al., 2004). Given the expense of collecting RDS data this may be an extreme suggestion, but, at a minimum, researchers could check the stability of the estimates with varying degrees of discarded data.

Uniform recruitment. Past RDS work has assumed that participants recruit randomly from their friends, even though in practice this is probably not the case. Our results, however, are entirely in terms of edges weights, as opposed to degree, and so lend themselves to relaxing the uniform recruitment assumption. With a better understand of the nature of non-random recruitment, non-uniform edge weights could yield improved RDS estimates.

Acknowledgments

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

Appendix A. Technical Details and Proofs for Section 3

We think of (K, π) as an operator on $L^2(\pi)$ – the space of functions $f : V \to \mathbb{R}$ 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 $\psi \in L^2(\pi)$ an eigenfunction for K with eigenvalue λ if $K\psi = \lambda\psi$.

Random walks on weighted graphs are *reversible*, i.e. they satisfy the detailed balance equation

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

Reversibility is equivalent to $K : L^2(\pi) \to L^2(\pi)$ being self-adjoint. Consequently, reversible walks are diagonalizable in an orthonormal basis of real eigenfuctions. That is, there exist eigenfunctions $\psi_0, \psi_1, \ldots, \psi_{N-1}$ with corresponding real eigenvalues

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

such that $\langle\psi_i, \psi_j\rangle = \delta_{ij}$. For details of the above functional analytic view, see e.g. Saloff-Coste (1996).

Lemma 3.1. For $0 < c < 1/2$, the l-step distribution of the walk defined at (3.1) 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. $\psi_1(x)$ is 1 on A and -1 on B), $\beta_1 = 1 - 2c$
(3) The $N - 2$ dimensional subspace of functions $\psi : V \to \mathbb{R}$ 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 Saloff-Coste (1996) shows that for reversible walks

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

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 2 non-zero eigenvalues (and their corresponding eigenfunctions as we have written them down are orthonormal), we have

$$\begin{aligned}\frac{K_n(x,y)}{\pi(y)} &= \beta_0^n \psi_0(x)\psi_0(y) + \beta_1^n \psi_1(x)\psi_1(y) \\ &= 1 + \beta_1^n \psi_1(x)\psi_1(y).\end{aligned}$$

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

Lemma 3.2. Consider the walk defined at (3.1). 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. Let f_D indicate infection, i.e. $f(v_i) = 1$ if v_i is infected, and $f(v_i) = 0$ otherwise. First observe that for $X_0 \in A$

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

Then,

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

□

Lemma 3.3. Consider the walk defined at (3.1). Suppose $X_0^1, X_1^1, \ldots$ and $X_0^2, X_1^2, \ldots$ 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 $i, j \geq 0$

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

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,

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

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 $x_0 \in A$

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

By symmetry, for $x_0 \in B$

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

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

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

The result now follows because,

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

□

Corollary 3.1. Consider the walk $X_0, X_1, \ldots$ defined at (3.1). If $X_0 \sim \pi$, then the variance of $\hat{p}$ satisfies

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

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

Proof. First note that

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

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

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

where we use the fact that

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

□

Lemma 3.4. Consider the walk defined in (3.1). Suppose a recruitment tree is chosen according to a probability distribution ν on the set of n-node trees, and RDS samples are correspondingly collected. Then the variance of $\hat{p}$ satisfies

$$\mathrm{Var}(\hat{p}) = \frac{p - p^2}{n} + \frac{(p_A - p_B)^2}{2n^2} \sum_{k=1}^{n-1} \beta_1^k \mathbb{E}|L_k|$$

where L_k is the set of pairs of samples distance k apart.

Proof. First observe that (A.1) 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

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

Now summing the covariance terms, we have

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

from which the result follows. □

Appendix B. Technical Details and Proofs for Section 4

Consider a reversible Markov chain (K,π) on V. Then there is an orthonormal basis of eigenfunctions $\psi_0,\psi_1,\ldots,\psi_{N-1}$ with eigenvalues $1=\beta_0\geq\cdots\geq\beta_{N-1}$. Let $f:V\to\mathbb{R}$ be any function, and express f in the basis $\{\psi_k\}$:

$$f=\sum_{k=0}^{N-1}a_k\psi_k \qquad a_k=\langle f,\psi_k\rangle_\pi.$$

Suppose $X_0^1,X_1^1,X_2^1,\ldots$ and $X_0^2,X_1^2,X_2^2,\ldots$ 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 $i,j\geq 0$

$$\begin{aligned}\mathbb{E}f(X_i^1)f(X_j^2) &= \mathbb{E}\left(\mathbb{E}[f(X_i^1)f(X_j^2)|X_0]\right)\\ &= \mathbb{E}\left(\mathbb{E}[f(X_i^1)|X_0]\cdot\mathbb{E}[f(X_j^2)|X_0]\right)\\ &= \mathbb{E}\left(\sum_{k=0}^{N-1}a_k\mathbb{E}[\psi_k(X_i^1)|X_0]\cdot\sum_{k=0}^{N-1}a_k\mathbb{E}[\psi_k(X_j^2)|X_0]\right).\end{aligned}$$

Now,

$$\begin{aligned}\mathbb{E}[\psi_k(X_i^1)|X_0] &= \sum_{y\in V}K_i(X_0,y)\psi_k(y)\\ &= K^i\psi_k(X_0)\\ &= \beta_k^i\psi_k(X_0).\end{aligned}$$

Analogously, $\mathbb{E}[\psi_k(X_j^2)|X_0]=\beta_k^j\psi_k(X_0)$, so

$$\begin{aligned}\mathbb{E}f(X_i^1)f(X_j^2) &= \mathbb{E}\left(\sum_{k=0}^{N-1}a_k\beta_k^i\psi_k(X_0)\cdot\sum_{k=0}^{N-1}a_k\beta_k^j\psi_k(X_0)\right)\\ &= \sum_{k=0}^{N-1}a_k^2\beta_k^{i+j}\end{aligned}$$

where the last equality follows from orthonormality since $X_0\sim\pi$. Furthermore, since $\psi_0\equiv 1$, $a_0=\mathbb{E}f=\mathbb{E}f(X_i^1)=\mathbb{E}f(X_j^2)$. Consequently,

$$\mathrm{Cov}(f(X_i^1),f(X_j^2))=\sum_{k=1}^{N-1}a_k^2\beta_k^{i+j}.$$

Theorem 4.1 (Diaconis & Basseti (2007))**.** Consider a reversible Markov chain (K, π) with orthonormal basis of eigenfunctions $\psi_0, \psi_1, \ldots, \psi_{N-1}$ and associated eigenvalues

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

Let $X_0^1, X_1^2, \ldots$ and $X_1^2, X_2^2, \ldots$ be independent realizations of the K chain with $X_0^1 = X_0^2 \sim \pi$. Then for any function g

$$\text{Cov}(g(X_i^1), g(X_j^2)) = \sum_{k=1}^{N-1} a_k^2 \beta_k^{i+j}$$

where $a_k = \langle g, \psi_i \rangle_\pi$. In particular,

$$\text{Cov}(g(X_i^1), g(X_j^2)) \leq \beta_1^{i+j} \text{Var}_\pi g.$$

Proof. The inequality follows from the fact that the $L^2(\pi)$ norm of g satisfies

$$\|g\|_2^2 = \sum_{k=0}^{N-1} a_k^2.$$

Since $\psi_0 \equiv 1$, $a_0 = \mathbb{E}_\pi g$. Consequently,

$$\sum_{k=1}^{N-1} a_k^2 = \|g\|_2^2 - \mathbb{E}_\pi g = \text{Var}_\pi g.$$

□

Recall the importance sampling RDS estimator

$$\frac{1}{n} \sum_{i=0}^{n-1} f(X_i) \frac{W_V}{N W_{X_i}} = \frac{1}{n} \sum_{i=0}^{n-1} g(X_i) \qquad g(x) = f(x) \frac{\overline{W}}{W_x}$$

where we assume $X_0, X_1, \ldots$ starts in stationarity. Furthermore, for

$$J = \frac{1}{N} \sum_{i=1}^{N} f(v_i)$$

the mean of f

$$\begin{aligned} \text{Var}_\pi(g) &= \mathbb{E}_\pi g^2 - (\mathbb{E}_\pi g)^2 \\ &= \sum_{x \in V} f^2(x) \frac{\overline{W}_V^2}{W_x^2} \cdot \frac{W_x}{W_V} - J^2 \\ &= \frac{1}{N} \sum_{x \in V} f^2(x) \frac{\overline{W}_V}{W_x} - J^2. \end{aligned}$$

As a heuristic, if $W_x \geq \overline{W}_V$ for x such that $f(x)$ is relatively large, then $\text{Var}_\pi(g) \leq \text{Var}_u(f)$ where u is the uniform distribution. For example, in the case of estimating disease prevalence (i.e. $f_D(v) = 1$ if v is infected and $f_D(v) = 0$ otherwise), the variance of the estimate is affected by the typical degree of infected individuals. In fact, if infected people tend to have higher than average degree, independent sampling from the stationary distribution π leads to lower variance than sampling from the uniform distribution u, i.e. simple random sampling.

References

Abdul-Quader, A. S., Heckathorn, D. D., McKnight, C., Bramson, H., Nemeth, C., Sabin, K., Gallagher, K. & DesJarles, D. C. (2006). Effectiveness of respondent-driven sampling for recruiting drug users in New York City: Findings from a pilot study. *Journal of Urban Health* **83**, 459–476.

Brooks, S. P. & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Geographical Statistics* **7**, 434 – 455.

Cheeger, J. (1970). A lower bound for the smallest eigenvalue of the Laplacian. In *Symposium in Honor of S. Bochner*. Princeton University Press.

Coleman, J. S. (1958). Relational analysis: The study of social organization with survey methods. *Human Organization* **17**, 28–36.

Diaconis, P. & Basseti, F. (2007). Examples comparing importance sampling and the metropolis algorithm. Preprint.

Erickson, B. H. (1979). Some problems of inference from chain data. *Sociological Methodology* **10**, 276–302.

Frank, O. & Snijders, T. A. B. (1994). Estimating the size of hidden populations using snowball sampling. *Journal of Official Statistics* **10**, 53–67.

Frost, S. D. W., Brouwer, K. C., Cruz, M. A. F., Ramos, R., Ramos, M. E., Lozada, R. M. & Strathdee, C. M.-R. S. A. (2006). 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* **83**, 83–97.

Gelman, A., Carlin, J. B., Stern, H. S. & Rubin, D. B. (2004). *Bayesian Data Analysis*. Boca Raton: Chapman & Hall, 2nd ed.

Gelman, A. & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science* **7**, 457–471.

Geyer, C. J. (1992). Practical markov chain monte carlo. *Statistical Science* **7**, 473–511.

Gilks, W., Richardson, S. & Spiegelhalter, D. (1996). *Markov Chain Monte Carlo in Practice*. Chapman & Hall.

Goodman, L. (1961). Snowball sampling. *Annals of Mathematical Statistics* **32**, 148–170.

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

Heckathorn, D. D. (2002). Respondent-driven sampling II: Deriving valid population estimates from chain-referral samples of hidden populations. *Social Problems* **49**, 11–34.

Heckathorn, D. D. & Jeffri, J. (2003). Social networks of jazz musicians. In *Changing the Beat: A Study of the Worklife of Jazz Musicians, Volume III: Respondent-driven sampling: Survey Results by the Research Center for Arts and Culture*, Research Division Report 43. Washington, DC: National Endowment for the Arts, pp. 48–61.

Jerrum, M. & Sinclair, A. (1989). Approximating the permanent. *SIAM Journal on Computing* **18**, 1149–1178.

Johnston, L. G., Sabin, K., Hien, M. T. & Huong, P. T. (2006). Assessmnt of respondent driven sampling for recruiting female sex workers in two Vietnamese cities: Reaching the unseen sex worker. *Journal of Urban Health* **83**, 16–28.

Kendall, W. S., Liang, F. & Wang, J.-S., eds. (2005). *Markov chain Monte Carlo: Innovations and Applications*. World Scientific Publishing Co. Pte. Ltd.

Klovdahl, A. (1989). Urban social networks: Some methodological problems and possibilities. In *The Small World*, M. Kochen, ed. Norword, NJ: Ablex Publishing, pp. 176–210.

Lansky, A., Abdul-Quader, A. S., Cribbin, M., Hall, T., Finlayson, T. J., Garffin, R. S., Lin, L. S. & Sullivan, P. S. (2007). Developing an HIV behavioral survelillance system for injecting durg users: The national HIV behavioral surveillance system. *Public Health Reports* **122**, 48–55.

Liu, J. (2001). *Monte Carlo Strategies in Scientific Computing*. Springer Series in Statistics.

Lohr, S. L. (1999). *Sampling: Design and Analysis*. Pacific Grove: Duxbury Press.

Magnani, R., Sabin, K., Saidel, T. & Heckathorn, D. (2005). Review of sampling hard-to-reach and hidden populations for HIV surveillance. *AIDS* **19**, S67–S72.

Marsden, P. V. (1990). Network data and measurement. *Annual Review of Sociology* **16**, 435–463.

Marshall, A. (1956). The use of multi-stage sampling schemes in monte carlo computations. In *Symposium on Monte Carlo Methods*, M. Meyer, ed. Wiley, New York.

McCarty, C., Killworth, P. D., Bernard, H. R., Johnsen, E. C. & Shelley, G. A. (2001). Comparing two methods for estimating network size. *Human Organization* **60**, 28–39.

McKnight, C., Jarles, D. D., Bramson, H., Tower, L., Abdul-Quader, A. S., Nemeth, C. & Heckathorn, D. (2006). Respondent-driven sampling in a study of drug users in New York City: Notes from the field. *Journal of Urban Health* **83**, 54–59.

McPherson, M., Smith-Lovin, L. & Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annu. Rev. Sociol.* **27**, 415–44.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M., Teller, A. H. & Teller, E. (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics* **21**, 1087–1091.

Newman, M. E. J. (2006). Finding community structure in networks using the eigenvectors of matricies. *Physical Review E* **74**, 036104.

Platt, L., Wall, M., Rhodes, T., Judd, A., Hickman, M., Johnston, L. G., Renton, A., Bobrova, N. & Sarang, A. (2006). Methods to recruit hard-to-reach groups: Comparing two chain referral sampling methods of recruiting injection drug users acrross nine studies in Russia and Estonia. *Journal of Urban Health* **83**, 39–53.

Ramirez-Valles, J., Heckathorn, D. D., Vázquez, R., Diaz, R. M. & Campbell, R. T. (2005). From networks to populations: The development and application of respondent-driven sampling among IDUs and Latino gay men. *AIDS and Behavior* **9**, 387–402.

Robinson, W. T., Jan M. H. Risser, S. M., Becker, A. B., Rehman, H., Jefferson, M., Griffin, V., Wolverton, M. & Tortu, S. (2006). Recruiting injection drug users: A three-site comparison of results and experiences with respondent-driven and targeted sampling procedures. *Journal of Urban Health*

83, 29–38.

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

Salganik, M. J. & Heckathorn, D. D. (2004). Sampling and estimation in hidden populations using respondent-driven sampling. *Sociological Methodology* **34**, 193–239.

Saloff-Coste, L. (1996). Lectures on finite markov chains. In *Lectures on Probability Theory and Statistics*, P. Bernard, ed., vol. 1665 of *Lecture Notes in Mathematics*. Ecole d'Eté de Probabiltés de Saint-Flour XXVI, Springer.

Semaan, S., Lauby, J. & Liebman, J. (2002). Street and network sampling in evaluation studies of HIV risk-reduction interventions. *AIDS Reviews 2002* **4**, 213–223.

Simic, M., Grazina, L., Platt, L., Baros, S., Andjelkovic, V., Novotny, T. & Rhodes, T. (2006). Exploring barriers to 'respondent-driven sampling' in sex workers and drug-injecting sex workers in Eastern Europe. *Journal of Urban Health* **83**, 6–15.

Snijders, T. A. B. (1992). Estimation on the basis of snowball samples: How to weight? *Bulletin de Méthodologie Sociologique* **36**, 59–70.

Spreen, M. (1992). Rare populations, hidden populations, and link-tracing designs: What and why? *Bulletin de Méthodologie Sociologique* **36**, 34–58.

Thompson, S. K. (2006). Targeted random walk designs. *Survey Methodology* **32**, 11–24.

Thompson, S. K. & Collins, L. M. (2002). Adaptive sampling in research on risk-related behaviors. *Drug and Alcohol Dependence* **68**, S57–S67.

Thompson, S. K. & Frank, O. (2000). Model-based estimation with link-tracing sampling designs. *Survey Methodology* **26**, 87–98.

Thompson, S. K. & Seber, G. A. F. (1996). *Adaptive Sampling*. New York: John Wiley & Sons.

Volz, E. & Heckathorn, D. D. (2007). Probability-based estimation theory for respondent-driven sampling. *Journal of Official Statistics* In press.

Wang, J., Carlson, R. G., Falck, R. S., Siegal, H. A., Rahman, A. & Li, L. (2005). Respondent-driven sampling to recruit MDMA users: A methodological assessment. *Drug and Alcohol Dependence* **78**, 147–157.

Wang, J., Falck, R. S., Li, L., Rahman, A. & Carlson, R. G. (2006). Respondent-driven sampling in the recruitment of illicit stimulant drug users in a rural setting: Findings and technical issues. *Addictive Behaviors* **32**, 924–937.

Yeka, W., Maibani-Michie, G., Prybylski, D. & Colby, D. (2006). Application of respondent driven sampling to collect baseline date on FSWs and MSM for HIV risk reduction interventions in two urban centers in Papua New Guinea. *Journal of Urban Health* **83**, 60–72.

Zheng, T., Salganik, M. J. & Gelman, A. (2006). How many people do you know in prison?: Using overdispersion in count data to estimate social structure in networks. *JASA* **101**, 409 – 423.

(Sharad Goel) Yahoo! Research, 111 W. 40th Street, 17th Floor, New York, NY 10018
E-mail address: goel@yahoo-inc.com

(Matthew J. Salganik) Department of Sociology, Princeton University, Princeton, NJ 08544
E-mail address: mjs3@princeton.edu