Journal of Official Statistics, Vol. 24, No. 1, 2008, pp. 79–97

Probability Based Estimation Theory for Respondent Driven Sampling

Erik Volz[1] *and Douglas D. Heckathorn*[2]

Many populations of interest present special challenges for traditional survey methodology when it is difficult or impossible to obtain a traditional sampling frame. In the case of such "hidden" populations at risk of HIV/AIDS, many researchers have resorted to chain-referral sampling. Recent progress on the theory of chain-referral sampling has led to Respondent Driven Sampling (RDS), a rigorous chain-referral method which allows unbiased estimation of the target population. In this article we present new probability-theoretic methods for making estimates from RDS data. The new estimators offer improved simplicity, analytical tractability, and allow the estimation of continuous variables. An analytical variance estimator is proposed in the case of estimating categorical variables. The properties of the estimator and the associated variance estimator are explored in a simulation study, and compared to alternative RDS estimators using data from a study of New York City jazz musicians. The new estimator gives results consistent with alternative RDS estimators in the study of jazz musicians, and demonstrates greater precision than alternative estimators in the simulation study.

Key words: Respondent driven sampling; chain-referral sampling; Hansen–Hurwitz; MCMC.

1. Introduction

Chain-referral sampling has emerged as a powerful method for sampling hard-to-reach or hidden populations. Such sampling methods are favored for such populations because they do not require the specification of a sampling frame.

The lack of a sampling frame means that the survey data from a chain-referral sample is contingent on a number of factors outside the researcher's control such as the social network on which recruitment takes place. The major challenge of chain-referral sampling has been to understand an unconventional sampling process and to base estimates on the resulting data. In this article we draw on previous research on Respondent Driven Sampling (RDS) (Heckathorn 1997, 2002; Salganik and Heckathorn 2004) to show that with a few plausible assumptions about the recruitment process and the social network, it is possible to specify selection probabilities for individuals in the target population and to apply traditional probability theory to the problem of statistical inference.

[1] University of California-San Diego, Antiviral Research Center, 150 W Washington Street, San Diego, CA 92103, U.S.A. Email: erik.volz@gmail.com
[2] Cornell University, 344 Uris Hall, Ithaca, NY 14853-7601, U.S.A. Email: douglas.heckathorn@cornell.edu

Acknowledgments: This research was made possible by grants from the National Endowment for the Arts, the Centers for Disease Control and Prevention, and the National Institutes of Health/National Institute on Drug Abuse. We thank the reviewers for helpful comments and advice. Matt Salganik offered numerous suggestions which have improved the quality of the article.

The estimator presented here is similar to estimators originally proposed in the RDS literature (Heckathorn 1997, 2002; Salganik and Heckathorn 2004), although the new estimator is based on a different theoretical foundation. The classical RDS estimator is based largely on Markov chain theory and social network theory. Our new estimator relies on Markov chain sampling theory (Hastings 1970; Metropolis et al. 1953) and the theory of sampling with unequal probabilities (Hansen and Hurwitz 1943; Cochran 1977).

This article should be viewed as part of an established and growing literature on network sampling (Sirken 1998; Sudman and Kalton 1986). Birnbaum and Sirken (1965) were the first to consider sampling in affiliation networks, such as the networks of patients and health-care providers. Felix-Medina and Thompson (2004), Spreen and Zwaagstra (1994), and Rothenberg et al. (1995) have considered network sampling for hidden populations using link-tracing[3] or snowball[4] designs. Frank (1978) has considered the problem of estimating topological features of social networks given a standard random sample from a network. Work by Frank and Snijders (1994) and Thompson (1998) has focused on deriving unbiased estimates from snowball-type and link-tracing samples, and in this sense is most similar to this work.

In Section 2 we review the basics of RDS methodology. In Section 3 we introduce a new RDS estimator which offers several advantages over the traditional methods of RDS estimation. Section 4 contains an analytical comparison of the new estimator to classical RDS methodology. Section 5 contains a prospective variance estimator for estimating proportions of categorical variables, and finally Section 6 presents the results of a simulation study to compare the new and old estimators.

2. Respondent Driven Sampling

Respondent Driven Sampling (RDS) is a rigorous system of chain-referral sampling which allows statistical inference of the target population by controlling for the sources of bias usually associated with chain-referral sampling.

RDS is now being implemented in the U.S. and around the world to study hard-to-reach or "hidden" populations. The Centers for Disease Control and Prevention has announced that it will use RDS to track HIV-risk behavior among injectors in 25 cities in the U.S., and Family Health International, the largest nonprofit organization in global public health, is using it in more than a dozen countries (Lang 2004; Heckathorn et al. 2002). The main advantage of RDS is that it does not require an ordinary sampling frame. Thus it is effective for stigmatized, hidden, or hard-to-reach populations, for which the researcher lacks organizational or institutional access.

Chain-referral sampling data differs from ordinary samples in that the respondents are linked together by a chain of recruitments. In general, each respondent will have attributed to them a coupon with a unique serial number which was given to them by another

[3] Link-tracing designs are generally not respondent driven, that is, recruitment may be directed by the survey administrator, and as such, makes no assumption about random recruitment by participants. Such designs occasionally combine traditional cluster sampling or standard random sampling with chain-referral methods.

[4] Snowball usually refers to chain-referral designs which exhaustively map out social networks by allowing unlimited recruitments from each participant. This should be contrasted with the random-walk design considered in the present article.

respondent. They will also have a a limited number of coupons which they may give to other respondents. Thus it is possible to keep track of who recruited whom. Figure 1 shows an actual recruitment chain drawn from an RDS study of New York City jazz musicians (Heckathorn and Jeffri 2003).

RDS begins with the selection of an initial respondent, or "seed." Selection of the seed is typically nonrandom, such as via public venues or health centers. The seed is given a number of coupons to distribute to friends and acquaintances which can be redeemed by being interviewed. When interviewed, the new respondent is in turn given coupons to distribute, thereby perpetuating the sample chain.

Additionally, RDS requires that we keep track of the degree of each respondent. The degree of a node in a network is the number of connections to that node, i.e., the number of neighbors of that node. In the context of chain-referral sampling, the degree of an individual will be defined as the number of people that that person *could*, in principle, recruit. We consider undirected networks only, such that recruitment can take place in both directions across a social network connection.

We will assume that our chain-referral samples are with-replacement, that is, any individual may be recruited into the sample more than once. Note that with-replacement sampling in chain-referral samples bears a subtle difference to with-replacement sampling in design-based sampling methods. The selection of each sample unit is under the control of the respondents themselves, thus we are not free to resample a current respondent at any time.

In practice, the condition of with-replacement sampling is rarely met. It is possible that participation in the study may alter the acceptance rate of individuals to participate in the study again. This could be a strong confounding factor if sampling with-replacement was allowed. But if the sampling fraction is very small, we can safely use results based on sampling with-replacement to use in the case of sampling without-replacement.

In the following treatment, we assume that each respondent recruits only one neighbor, although methods have been devised to compensate for the case where respondents may

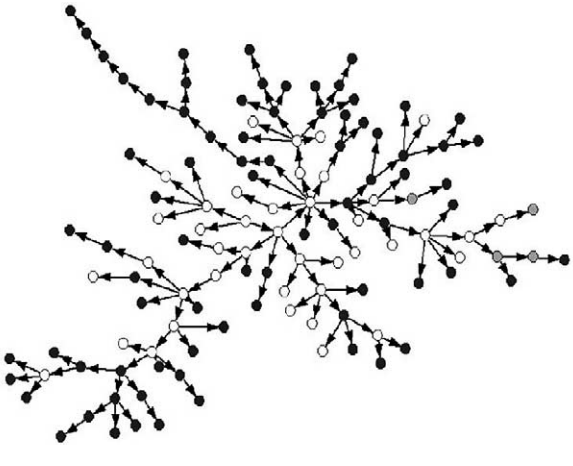

Fig. 1. Example of a recruitment chain. This recruitment chain comes from an RDS study of jazz musicians in New York City (Heckathorn and Jeffri 2003). Arrows indicate the direction of recruitment. The colors indicate the gender of each respondent: Black = Male, White = Female, Grey = Missing Data

recruit more than one neighbor. Details on this method, called *demographic adjustment*, can be found in Section 4.

Further, we assume that the sampling fraction is small, such that we can apply solutions for the sampling-with-replacement case. Refer to Table 1 for the notation used throughout this article.

In developing our theory, we will rely on the following assumptions in addition to those mentioned above:

1. *Degree*. Respondents accurately report their degree in the network.
2. *Recruitment is random*. When recruiting others, respondents select uniformly at random from their personal network.
3. *Reciprocity*. Network connections are reciprocal. Respondents recruit those with whom they have a preexisting relationship, such as acquaintances, friends, and those closer than friends. Such connections are reciprocal, e.g., my friends and acquaintances consider me to be a friend or acquaintance. Consequently, in network-theoretic terms, the potential recruitment network is undirected, so if respondent *a* can recruit *b*, then *b* can also recruit *a*. This is required by the *reciprocity model* (Heckathorn 2002; Salganik and Heckathorn 2004) upon which the original RDS estimator is based. This is formally known as the *reciprocity hypothesis*.
4. *Convergence*. Recruitment is modeled as a Markov process (MP), where the state of the MP is the last individual recruited. Transition probabilities are described in Section 3. We assume that the MP is irreducible and that each state has a finite return time. Therefore, a unique equilibrium to the MP exists and recruitment rapidly converges to this equilibrium. The implication is that after a modest number of steps, the sample composition becomes independent of the initial respondents ("seeds") who initiated the chain-referral process.

The irreducibility condition is equivalent to the condition that the social network is well-connected; that is to say, every node can be reached by a finite path from any other node. Furthermore, our social networks are assumed to be finite (though very large), so the expected return time must be finite as well.

Table 1. Notation used throughout this article

U is the set of all individuals in the population
S is the set of all individuals in the sample
A, B,. . .are disjoint sets of individuals
N_X is the number of elements in a Set X
$N_{\mathcal{G}}$ is the number of Subsets A, B,. . .
n_X is the number of sample units from Set X
P_A, P_B,. . .are the population proportions of each type, A, B, etc.
$\vec{P}$ is the vector with Elements P_A, P_B, etc.
R_{AB} is the number of recruitments from Group A to Group B
R_A is the total number of times people of Type A are recruited
$\bar{R}_A$ is the total number of recruitments from people of Type A
σ_{AB} is the estimated probability of someone from Set A selecting someone from Set B
δ_i is the degree of Individual i
δ_X is the average degree of individuals from Set X

On the surface, the irreducibility assumption may seem unrealistic, especially for large populations, where it is most likely that some units will be isolated from the network as a whole. This, however, is usually not a cause for concern. It is known from random network theory that most networks possess a so-called *giant component*, a subset of nodes such that a network path exists between any two and which occupies a nonvanishing fraction of the network as the population size goes to infinity. The giant component usually encompasses the vast majority of the population, so long as some basic conditions are met. For instance, in pure random graphs, the giant component will consist of 99% of the population if nodes have just 5 links on average. RDS studies have typically exceeded this margin comfortably. In a study of NYC jazz musicians, respondents were found to have an average degree of 109 (Heckathorn and Jeffri 2003), while in a study of gay Latinos, respondents in San Francisco had an average degree of 8 and in Chicago had an average degree of 13 (Ramirez-Valles et al. 2005). With that said, field RDS studies should come with the caveat that statistical inference is limited to the giant component, rather than the total population. But provided the giant component is very large, this is usually a minor distinction.

Furthermore, research on the *small-world* problem (Watts 1999) has led to the observation that almost all social networks have very short mean path length. Consequently, there are relatively few intermediaries between any two randomly selected individuals in most social networks. In pure random networks (Erdős and Renyi 1959; Newman et al. 2001), path length grows logarithmically with population size. It is therefore likely that the selection probability for any individual in the network will stabilize after just a few recruitments, as almost anyone in the population can be reached in a small number of steps.

Another assumption commonly called into question is that respondents recruit uniformly at random from their network neighbors. Indeed, it is difficult or impossible to enforce random recruitment among respondents, and in many cases respondents may have special reasons for selecting a particular recruit. However, nonrandom recruitment, if it occurs, will not necessarily bias our estimator. As long as recruitment is not correlated with any variable important for estimation (e.g., the study-variable or degree), the aggregate effect is for recruitment to appear uniform-random.

Nonrandom recruitment would most obviously be evidenced by skewed and nonsymmetric recruitment matrices. If, for instance, respondents preferred recruiting someone of Type A, we would expect recruitment matrices with much more weight on Elements R_{XA} than on Elements R_{AX}. In fact, this is rarely observed. By now, strong empirical evidence (Heckathorn et al. 2002) has been built up which indicates that random recruitment holds in most cases. It is nevertheless a potential source of bias that practitioners should watch out for.

3. New Estimators for Respondent Driven Sampling

It is often the case that it is more convenient to sample from a distribution other than the one we wish to use for estimation. In this case, the theory of Markov chain sampling has been developed in order to sample from arbitrary distributions. The premise is to devise a Markov process (MP) such that the equilibrium distribution of the MP is identical to the distribution one wishes to sample from. It has been shown that estimators based on Markov chain samples are asymptotically unbiased[5] (Hastings 1970).

In contrast to traditional Markov chain sampling, we are not at liberty to devise the transition probabilities between our sampling units due to the lack of a standard sampling frame. Rather the transition probabilities are imposed on us by the nature of the chain referral sample and the properties of the social network. Nevertheless, the chain-referral sample will constitute a Markov chain which fits the criteria governing the application of our theory.

In mathematical terms, a chain-referral sample is analogous to a random walk on a network. It has been shown (Salganik and Heckathorn 2004) that a random walk on a network is an MP, which in equilibrium occupies a node with probability proportional to degree. We can then infer that a chain-referral sample will select individuals in the population with probability proportional to degree.

Let $\mathscr{E}$ be the incidence matrix of the network. $\mathscr{E}$ will have elements e_{ij} where $e_{ij} = 1$ if nodes i and j are connected, and will equal zero otherwise. Note that the degree of node i, δ_i, is the i'th row sum of $\mathscr{E}$, $\sum_j e_{ij}$. If the random walk is at node i at step t, the probability of node i choosing node j is $1/\delta_i = 1/\sum_j e_{ij}$. Denote this transition probability $\sigma_{ij}^{\mathscr{E}}$, and let the matrix with these transition probabilities be called $\sigma^{\mathscr{E}}$. The random walk on the network can therefore be considered an MP with transition probabilities $\sigma^{\mathscr{E}}$.

The random walks we consider are irreducible and finite, so there must be a unique equilibrium to this MP. Furthermore the MP will converge to this equilibrium. Consider the state vector x^* with elements

$$x_i^* = \frac{\delta_i}{\sum\limits_j \delta_j} \tag{1}$$

It may be verified that x^* is an equilibrium to the MP given by $\sigma^{\mathscr{E}}$, and by our hypotheses, must also be a unique attracting equilibrium. Now that we have established that a chain-referral sample of the RDS type is a Markov chain sample, we may proceed to develop estimators for our target population. Using only the fact that RDS samples individuals with probability proportional to degree, we can develop a Hansen–Hurwitz (HH) type estimator (Hansen and Hurwitz 1943; Cochran 1977) for $\vec{P}$, the proportion of the population in each disjoint set $A, B, \ldots$. The derivation presented here uses a similar argument to that presented in (Salganik and Heckathorn 2004) to estimate the average degree in a social network from chain-referral data.

HH estimators require knowledge of the selection probabilities, p_i, the probability that individual i will be selected at any stage of the chain-referral sample. HH estimators also assume each sample element is chosen independently of the rest of the sample – an assumption which is, of course, violated by the MP model of recruitment. Thus the likeness to HH estimators is only partial. Sample units will be correlated, a fact that will

[5] By *asymptotically unbiased* we mean that any bias will be of the order $1/n$. Therefore, for meaningful sample sizes, any bias will be negligible.

not bias the estimator but which will be taken into account in Section 5 where an estimate of variance is derived.

Using the equilibrium Condition (1), the selection probabilities will be

$$p_i = \frac{\delta_i}{N\delta_U} \tag{2}$$

which we can estimate as

$$\hat{p}_i = \frac{\delta_i}{N\hat{\delta}_U} \tag{3}$$

where $\hat{\delta}_U$ is the estimate of the average degree of the total population.

The $\hat{\delta}_X$ are easy to estimate. As in (Salganik and Heckathorn 2004), we note that the average degree can be estimated as a ratio estimator of HH estimators

$$\hat{\delta}_U = \frac{\sum\limits_S \delta_i/np_i}{\sum\limits_S 1/np_i} = \frac{n}{\sum\limits_S \delta_i^{-1}} \tag{4}$$

and for just one group, e.g., the Subset A within the population

$$\hat{\delta}_A = \frac{n_A}{\sum\limits_{A\cap S} \delta_i^{-1}} \tag{5}$$

This is the well-known formula for the harmonic mean, the mean of a quantity which is being sampled with probability proportional to its size.

Now let the variable y_i be some real-valued variable of interest. Let T_y represent the total value of y in the population, $\sum_U y_i$ where y_i may represent continuous variables such as age or income, or dichotomous variables such as HIV status.

The HH estimator of the total y in the population, $\hat{T}_y$ is

$$\hat{T}_y = \frac{1}{n}\sum_S \frac{y_i}{\hat{p}_i} = \frac{1}{n}\sum_{i\in S} \frac{\hat{\delta}_U N_{yi}}{\delta_i} = \frac{\hat{\delta}_U N}{n}\sum_S \delta_i^{-1} y_i$$

If N is unknown, as is generally the case, we can still estimate the mean value of y as

$$\langle \hat{y} \rangle = \frac{\hat{\delta}_U}{n}\sum_S \delta_i^{-1} y_i \tag{6}$$

Substituting the definition of $\hat{\delta}_U$ (Equation 4), we arrive at the simple equation

$$\langle \hat{y} \rangle = \frac{\sum\limits_{i\in S} \delta_i^{-1} y_i}{\sum\limits_{i\in S} \delta_i^{-1}} \tag{7}$$

We will refer to this estimator as *RDS II* to distinguish it from the RDS estimator presented in Section 4. Essentially Equation 7 weights each case by the reciprocal of the corresponding degree value.

Suppose we are interested in estimating P_A, the proportion of the population of Type A. Let y_i be the indicator function $I_A(i)$, which takes the value 1 if $i \in A$ and 0 otherwise. Using Equation 7 we have

$$\hat{P}_A = \frac{\sum\limits_{i \in A \cap S} \delta_i^{-1}}{\sum\limits_{i \in S} \delta_i^{-1}} \tag{8}$$

There is an alternative form of Equation 8 worth mentioning, as it gives some intuition of how our estimator works. With a little manipulation we get

$$\hat{P}_A = \left(\frac{n_A}{n}\right)\left(\frac{\hat{\delta}_U}{\hat{\delta}_A}\right) \tag{9}$$

The first part of Equation 9, (n_A/n), is the proportion of the sample of Type A. If our sample were a standard random sample this would be our estimate for P_A. The second part, (δ_U/δ_A), expresses the correction due to network effects. For example, if $\delta_U > \delta_A$ we are under-sampling individuals of Type A, and consequently we inflate our estimate.

Note that the initial recruits in a chain-referral sample (i.e., the "seeds") will generally be chosen nonrandomly. It is usually prudent to exclude them from the estimator (Equation 8), as well as the estimation of average degree (Equation 5), though the estimator will be asymptotically unbiased even if they are included. The rationale for eliminating seeds is the same as that for using a "burn-in" period during an MCMC sample. Any potential bias accruing from the initial seed selection will be lessened. The recruitments made by seeds are usually included, however, in the recruitment matrix (Section 4 below). Experimental evidence for how long a burn-in period is best is currently lacking.

4. The Classical RDS Estimation Procedure and Its Relation to RDS II

In Heckathorn 1997, 2002 and Salganik and Heckathorn 2004, it was shown how to convert a chain-referral sample into a probability sample of individuals in the population and to produce unbiased estimates from chain-referral sample data. The original RDS estimator accounted for all of the sources of bias usually associated with chain-referral samples, such as oversampling well-connected individuals and nonrandom mixing in the population. Here we present a brief review of this methodology with the objective of elucidating the relationship between the new estimator (8) and the original estimator proposed in Heckathorn 1997, 2002 and Salganik and Heckathorn 2004.

The classical RDS estimator (henceforth referred to as *RDS I*) relies on the theory of network balance between subgroups in the population. The mass of network connections to and from every group can be estimated up to a constant factor. This gives us a system of balance equations for every pair of groups, which in turn can be used to solve for the relative size of each group.

Specifically, it was observed that $R_{AB}/\bar{R}_A = \sigma_{AB}$ is an unbiased estimate of the probability of someone of Type A recruiting someone of Type B. Furthermore, the connections from Group A to Group B must be equal to those from B to A by the reciprocity hypothesis. The number of connections from Group A to Group B will be proportional to $\sigma_{AB}P_A\delta_A$. Given $N_{\mathscr{G}}$ groups, this then leads to a system of $\begin{pmatrix} N_{\mathscr{G}} \\ 2 \end{pmatrix}$ balance equations:

$$\begin{aligned}
&\sigma_{12}P_1\delta_1 = \sigma_{21}P_2\delta_2 \\
&\sigma_{13}P_1\delta_1 = \sigma_{31}P_3\delta_3 \\
&\vdots \\
&\sigma_{23}P_2\delta_2 = \sigma_{32}P_3\delta_3 \\
&\vdots \\
&\sigma_{(N_{\mathscr{G}}-1)N_{\mathscr{G}}}P_{N_{\mathscr{G}}-1}\delta_{N_{\mathscr{G}}-1} = \sigma_{N_{\mathscr{G}}(N_{\mathscr{G}}-1)}P_{N_{\mathscr{G}}}\delta_{N_{\mathscr{G}}}
\end{aligned} \tag{10}$$

This system of equations can be used to solve for $\hat{\vec{P}}$, our estimate for the population proportion of each group. Of course, we must also normalize our solution by using $\sum_x P_x = 1$. This system of equations is over-determined for systems with more than two groups, such that least squares regression may be used to solve for $\hat{\vec{P}}$.

Two enhancements to *RDS I* were proposed in Heckathorn (2002), which dramatically improve the precision of the estimator. The first considered adjustments for sample data in which respondents could recruit more than one network-neighbor. In this case, it is possible that some groups in the population may systematically recruit more than other groups in the population, a phenomenon called *differential recruitment*, which can dramatically alter the composition of the sample. However, under the assumption that such data still provides us with an unbiased estimate of the group-to-group transition probabilities, σ_{AB}, we can deduce what the sample composition would be in the absence of differential recruitment.

Given the matrix of transition probabilities σ with elements σ_{AB}, the theoretical equilibrium sample composition is the vector x^* which satisfies

$$x^* = x^*\sigma \tag{11}$$

With a theoretical equilibrium sample distribution x^* and unbiased estimates of σ, we can postulate what form the recruitment matrix would take in the absence of differential recruitment. We call this matrix $\tilde{R}$, which will have elements proportional to the theoretical equilibrium composition of the recruiter-type (elements x^*_X from the vector in Equation 11) times the unbiased transition probabilities. The sum of elements of $\tilde{R}$ is equal to the sample size, so that $\tilde{R}_{AB}$ is the theoretical number of recruitments from Group A to Group B in the absence of differential recruitment.

$$\tilde{R}_{AB} = (nx_A^*)\sigma_{AB} \tag{12}$$

At this point, we may find that for some pairs of groups, $\tilde{R}_{AB} \neq \tilde{R}_{BA}$. Because RDS randomly samples connections in the network, and the number of connections between any two groups must be identical, $\tilde{R}_{AB}$ and $\tilde{R}_{BA}$ will be two point estimates for the same quantity. Therefore, a more accurate estimate can be gained by averaging the values. Averaging over all pairs of groups gives us a symmetric recruitment matrix, which we call the *data-smoothed* recruitment matrix, $\tilde{R}_{DS}$.

Transition probabilities can be recomputed from the data-smoothed recruitment matrix. Furthermore, if we assume that differential recruitment does not bias the estimated average group degrees, δ_X, then the RDS estimator can be recalculated. We will refer to this estimator as *RDS I/DS*. Simulation studies have revealed that this estimator has markedly different properties from RDS I, most notably similar accuracy and increased precision. To base estimates on the data-smoothed recruitment matrix, we assume that neither σ nor the estimated δ_X are biased by differential recruitment. This may not always be the case, but in practice has proven a reliable assumption. RDS I/DS estimates are also much closer, in general, to RDS II estimates.

This review of traditional RDS theory is pertinent, as RDS II is closely related to the classical RDS estimator, RDS I. Note that these similarities only exist when considering categorical variables, as RDS I is not adaptable to the estimation of continuous quantities.

Whenever the recruitment matrix is symmetric (that is, whenever $R_{AB} = R_{BA} \ \ \forall A, B$) the RDS I and RDS II estimators will coincide. Consequently, basing RDS estimates on the demographically adjusted and data-smoothed recruitment matrix will equalize these estimators.

To put this on firmer ground, let us collect all terms in P_A in the RDS I system of Equation 10. For any group X,

$$P_X = \frac{P_A \delta_A \bar{R}_X R_{AX}}{\delta_X \bar{R}_A R_{XA}}$$

from which it follows that

$$\sum_X P_X = 1 = P_A \frac{\delta_A}{\bar{R}_A} \sum_X \frac{R_{AX} \bar{R}_X}{R_{XA} \delta_X}$$

Neglecting initial respondents (seeds), $\bar{R}_X = n_X$. That is to say, the number of individuals of Type X recruited into the study is the same as the number of individuals in the study of Type X. Then solving for P_A we have

$$P_A = \frac{n_A}{\delta_A} \left(\sum_X \frac{R_{AX} n_X}{R_{XA} \delta_X} \right)^{-1} \tag{13}$$

Note that if $R_{AX} = R_{XA}$, then their ratio falls out of the equation. This is exactly what happens with the demographically adjusted and data-smoothed recruitment matrix. Now observe that

$$\sum_X \frac{n_X}{\delta_X} = \sum_X \sum_{i \in S \cap X} \delta_i^{-1} = \frac{n}{\hat{\delta}_U}$$

Substituting this into Equation 13 yields Equation 9, our estimator for RDS II. Thus, provided that $R_{YX} = R_{XY}$ for all groups X and Y, these two estimators will coincide.

In passing, note that a parsimonious way of expressing the demographically adjusted RDS II estimator is the following

$$\hat{P}_A = x_A^* \left(\frac{\hat{\delta}_U}{\hat{\delta}_A} \right) \tag{14}$$

Referring back to Equation 9, it is clear that the only part of the RDS II estimator that will be biased is the sample proportion of Type A, n_A/n. To correct for the bias, simply substitute the equilibrium composition x_A^* for the sample proportion of Type A.

5. Variance Estimation

The complicated design of RDS creates numerous challenges for variance estimation. It is tempting to apply the well-known variance estimator for HH estimators (Hansen and Hurwitz 1943)

$$\hat{V}_{HH}(\hat{T}_y) = \frac{1}{N^2 n(n-1)} \sum_S \left(\frac{y_i}{p_i} - \hat{T}_y \right)^2 \tag{15}$$

which when estimating the average value of y, becomes

$$\hat{V}_{HH}(\langle \hat{y} \rangle) = \frac{1}{n(n-1)} \sum_S \left(\frac{y_i \hat{\delta}_U}{\delta_i} - \langle \hat{y} \rangle \right)^2 \tag{16}$$

But as noted, the RDS II estimator is only analogous to HH estimators which assume that each sample unit is drawn independently, so that outside of a few special cases this variance estimator performs quite poorly. Recall that sample units in RDS are correlated because they are linked by a recruitment chain, something which is not considered in standard HH estimators. Thus there are multiple sources of variance in Estimator (8). Sampling with nonidentical selection probabilities is considered in the above variance estimator. But additionally, an RDS sample constitutes an MCMC sample of the social network, with transition probabilities $\sigma^{\mathscr{E}}$. In general, it is necessary to take this correlation into account when estimating variance. The sample units are connected in a recruitment chain with indices $1, \ldots, n$; considering two sample units i and j, the distance from one to the other within the recruitment chain will be $|i - j|$. The correlation between sample units will decrease with distance within the recruitment chain. For categorical variables, we can say by approximation that the probability of unit $i + 1$ being Type Y given that Unit i is

Type X is σ_{XY}, where σ is the matrix of estimated transition probabilities of the simplified Markov process as in Section 4, e.g., $\Pr[\text{Unit}\, i+1 \in Y | \text{Unit}\ i \in X]$ is estimated as $\sigma_{XY} = R_{XY}/\bar{R}_X$. Given two Units i and j, with i of Type X, the probability that Unit j is of Type Y is estimated by the matrix product of the transition probabilities: $\Pr(j \in Y | i \in X)$ is estimated by $(\sigma^{|i-j|})_{XY}$. The distance between sample units i and j, $|i-j|$, is the exponent to which the matrix σ is raised, which yields a square matrix with the same dimensions as σ, $N_{\mathscr{G}} \times N_{\mathscr{G}}$. In this context the subscript XY refers to indices corresponding to the sets X and Y, and indicates the XYth element of matrix $\sigma^{|i-j|}$. Note that using the matrix $\sigma^{|i-j|}$ is a simplification; it is not always the case that recruitment can be modelled as a first order Markov process with transition probabilities σ, and the node-specific transition probabilities $\sigma^{\mathscr{E}}$ are almost always unknown.

Below, we derive a variance estimator which accounts for nonuniform selection probabilities and the MCMC structure of the sample. These equations specifically treat the estimation of categorical variables. We conclude with the estimation of variance for $\hat{P}_A$ expressed in Equations 17, 18, and 19, as follows:

$$\hat{V}_{P_A} = \hat{V}_1 + \frac{\hat{P}_A^2}{n}\left((1-n) + \frac{2}{n_A}\sum_{i=2}^{n}\sum_{j=1}^{i-1}(\sigma^{i-j})_{AA}\right) \tag{17}$$

where

$$\hat{V}_1 = \frac{\hat{V}(Z_i)}{n} = \frac{1}{n(n-1)}\sum_S (Z_i - P_A)^2 \tag{18}$$

and

$$Z_i = \hat{\delta}_U \delta_i^{-1} I_A(i) \tag{19}$$

where $I_A(i)$ is the indicator function as in Section 3.

The derivation of Equations 17, 18, and 19 is as follows. Note that this is based on the estimation of P_A, the proportion of the population of Type A. Furthermore, this derivation is based on the approximation that $\hat{\delta}_U$ is constant.

A little rearrangement shows the form of Estimator (8) can be expressed as

$$\langle \hat{y} \rangle = \hat{P}_A = \frac{1}{n}\sum_S Z_i \tag{20}$$

We wish to find the variance

$$\begin{aligned}
&V(Z_1 + Z_2 + \cdots + Z_n)\\
&= V(Z_1 + \cdots + Z_{n-1}) + 2\operatorname{cov}(Z_1 + \cdots + Z_{n-1}, Z_n) + V(Z_n)\\
&= V(Z_1 + \cdots + Z_{n-2}) + 2\operatorname{cov}(Z_1 + \cdots + Z_{n-2}, Z_{n-1}) +\\
&2\operatorname{cov}(Z_1 + \cdots + Z_{n-1}, Z_n) + V(Z_{n-1}) + V(Z_n)\\
&\vdots\\
&= \sum_S V(Z_i) + 2\operatorname{cov}\left(\sum_{j<i} Z_j, Z_i\right)
\end{aligned}$$

Using $E(\cdot)$ to denote the expectation operator for an r.v., we have

$$\begin{aligned}\operatorname{cov}(Z_1 + \cdots + Z_{m-1}, Z_m) &= E(Z_1 + \cdots + Z_{m-1} - (m-1)E[Z])(Z_m - E[Z]) \\ &= -(m-1)E[Z]^2 + \sum_{i=1}^{m-1} E\ Z_i Z_m\end{aligned}$$

In the above equation, the expected value of the product $Z_i Z_m$ must be computed. Using the definition of Z_i (Equation 19), we have

$$E\ Z_i Z_m = \Pr[i \in A] \times E(Z_i | i \in A) \times \Pr[m \in A | i \in A] \times E(Z_m | m \in A) \tag{21}$$

The probability that Unit i is in A, $\Pr[i \in A\,]$, will be estimated as n_A/n. The expected value of Z_i given that Unit i is in A, $E(Z_i|i \in A)$ will be estimated as $\sum_{S\cap A} Z_k/n_A$. Given that Unit i is of Type A (such that $y_i = 1$), the probability of Unit m also being of Type A, $\Pr[m \in A|i \in A\,]$, will be estimated as $(\sigma^{m-i})_{AA}$.

Then EZ_iZ_m is estimated as

$$\frac{n_A}{n}\frac{\sum\limits_{S\cap A} Z_k}{n_A}(\sigma^{m-i})_{AA}\frac{\sum\limits_{S\cap A} Z_k}{n_A} = \hat{P}_A(\sigma^{m-i})_{AA}\frac{1}{n_A}\sum_{S\cap A} Z_k = \hat{P}_A(\sigma^{m-i})_{AA}\frac{\hat{\delta}_U}{\hat{\delta}_A} = \frac{n}{n_A}\hat{P}_A^2(\sigma^{m-i})_{AA}$$

where we have used $\hat{\delta}_U/\hat{\delta}_A = \hat{P}_A n/n_A$ to simplify the equation.

Continuing in this way by reducing the variance of the sum of Z_i to the sum of the variances and covariances of Z_i we find that

$$\begin{aligned}\hat{V}(Z_1 + \cdots + Z_n) &= \hat{V}(n\hat{P}_A) = n^2\hat{V}(\hat{P}_A) \\ &= n\hat{V}(Z) - \hat{P}_A^2 n(n-1) + \frac{2n\hat{P}_A^2}{n_A}\sum_{i=2}^{n}\sum_{j=1}^{i-1}(\sigma^{i-j})_{AA}\end{aligned} \tag{22}$$

Solving for $\hat{V}(\hat{P}_A)$ gives Equation 17.

Unfortunately, the variance estimator 17 is not unbiased for a couple of reasons. Firstly, $\hat{\delta}_U$ is included in each Z_i term, and will therefore affect the covariance between Z_i. This would be difficult to account for and is not included in Equation 17. But for sufficient sample size, the variance of $\hat{\delta}_U$ is generally very small, as the selection probabilities for this quantity are proportional to its size. Secondly, the transition probabilities σ are not in general known, and usually must be estimated. Although the estimation of variance is not unbiased, under most conditions it will perform quite well. Its performance is explored by simulation in the next section.

Other strategies could be pursued to derive a variance estimator, for instance by considering a linearization approximation. The RDS II estimator can be expressed as a ratio $\hat{P}_A = \hat{Y}/\hat{X}$ where $\hat{Y} = \sum_S y_i/\delta_i$ and $\hat{X} = \sum_S 1/\delta_i$, where $E\hat{X} = n/\delta_U$. Then,

$$\hat{P}_A - P_A \approx \frac{1}{E\hat{X}}\left[\hat{Y} - \frac{E\hat{Y}}{E\hat{X}}\hat{X}\right] = \frac{1}{n}\sum_S \frac{y_i - P_A}{\delta_i/\delta_U} \tag{23}$$

Taking $Z_i = (y_i - P_A)/(\delta_i/\delta_U)$ one could proceed as above.

6. Simulation Study of RDS I and RDS II

So far we have presented three estimators for RDS data, RDS I, RDS II, and RDS II/DS. In addition we have an estimate of variance for RDS II, given by Equation 17. To gain insight into the properties of these estimators we have performed computer simulations of RDS samples on random networks with known properties.

There are several technicalities in the simulation of RDS due to the complicated sample design. The population under study is represented by a random network, which can have a wide range of properties and can be generated by any of many algorithms that have been developed for the task. In addition, each node must have a value assigned to it for the study-variable, y_i. Secondly, random walks of specified length are executed on the random network by choosing a node uniformly at random from the network's giant component, and then randomly selecting a neighbor of the last node at each step of the random walk. The random walks are interpreted as RDS samples by keeping track of the degree of each node and the node's y_i value. In these simulations, we consider the estimation of P_A, such that the y_i is the indicator function for membership in Group A.

Several pieces of information are required to construct a random network:

- A specification of group sizes – that is, the size of each group A, B,. . .
- A list of degree distributions for all groups in the network
- A mixing matrix $\mathscr{A}$, where the element $[\mathscr{A}]_{XY}$ specifies the fraction of all connections in the network that exist between groups X and Y.

We proceed by randomly assigning each node a degree drawn from the corresponding degree distribution. Then we randomly match connections in the network while simultaneously satisfying the constraint specified by $\mathscr{A}$.[6]

The parameter space specified by the group sizes, degree distribution, and mixing matrix is vast. In these simulations we have kept the population size fixed at $N = 10{,}000$. The networks are divided into four groups. The variable we wish to estimate is $P_A = 0.1$, so that $N_A = 1{,}000$. The remaining three groups are equal in size: $N_X = 3{,}000$. In addition, each group has its own degree distribution. In all cases the degree distribution is Poisson, but with different parameters controlling the average degree of the group. These variables are summarized in Table 2.

The effects of both sample size and assortative mixing have been determined by experiment. Figure 2 shows the the effects of sample size on the variance of the three estimators. Five random networks were generated with mild assortative mixing ($\sigma_{AA} = .15$), and 10,000 random walks were executed on each network. The estimators were applied to each random walk, and the empirical variance of each estimator computed. The average estimate of variance (Equation 17) from these simulations is also shown in Figure 2.

[6] A more detailed description of methods for generating random networks which exhibit assortative mixing can be found in Newman (2002).

Table 2. Random networks were generated with four disjoint groups, each having the size N_X and Poisson degree distribution with average degree z

Group	N_X	z
A	1,000	32
B	3,000	40
C	3,000	48
D	3,000	56

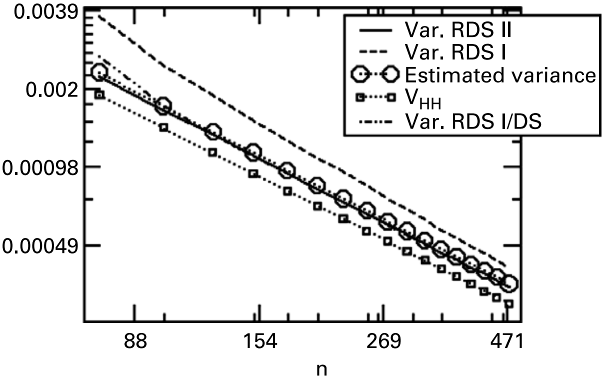

Fig. 2. Variance of three RDS estimators and mean estimated variance, based on 50,000 simulations as described in the text. Sample size is varied from 75 to 500. The data are plotted with log–log axes

As demonstrated in Section 4, RDS II and RDS I/DS coincide very closely. Both are consistently more precise than RDS I, although RDS II has greater performance than RDS I/DS for small sample size.[7]

For small sample sizes, RDS II is slightly more accurate (less biased) than RDS I/DS, as the latter methodology relies on accurate estimation of transition probabilities to perform reliably (results not shown). Both RDS II and RDS I/DS are consistently more accurate than RDS I, although because all the estimators are asymptotically unbiased, the difference disappears as sample size is increased. Salganik and Heckathorn (2004) have previously given a detailed analysis of bias for RDS I.

In this set of simulations, the variance estimator (Equation 17) shows slight but consistent bias in over-estimating the actual variance. The average coverage probability of the variance estimator for 90% confidence intervals is 91.03%. The naive variance estimator, $\hat{V}_{HH}$ (Equation 16), has an average coverage probability of 85%.

A different scenario is presented in Figure 3. Recall that σ_{AA} is the probability that someone of Type A will recruit again someone of Type A. σ_{AA} actually represents an aspect of network topology: it is the proportion of connections from nodes of Type A that

[7] Subsequent work has shown that these results are very specific to simple unbranched recruitment chains. The presence of branching (sample unit recruits more than one neighbor) can increase the performance of reciprocity-based estimators such as RDS I/DS or RDS II/DS relative to RDS II.

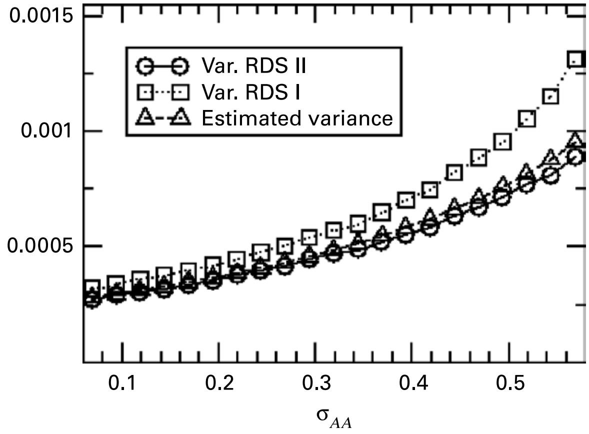

Fig. 3. Variance of RDS II and RDS I, alongside the estimated variance for RDS II based on 50,000 simulations with sample size 500 as described in the text. The mixing parameter σ_{AA} is varied from 0.069 to 0.57

go to other nodes of Type A. In these simulations we have varied this parameter from $\sigma_{AA} = 0.069$, which corresponds to essentially no assortative mixing, to $\sigma_{AA} = 0.57$ which represents a very strong preference for nodes of Type A to connect to one another at the expense of connections to nodes of other types. In all simulations, the sample size was $n = 500$.

The effect of increasing assortative mixing is the exponential increase in the variance of the estimator. In terms of MCMC sampling, this corresponds to increased sample size required for the sample to reach equilibrium.

The average coverage probability for this set of simulations is 89.997% for 90% confidence intervals. The estimated variance correctly tracks the exponential trend. The naive estimate of variance (not shown) which does not account for assortative mixing (Equation 15) grossly underestimates the actual variance and has a coverage probability which decreases from 88% to 62% as σ_{AA} is increased.

Finally, it provides a useful perspective to compare the estimators with real data. Table 3 shows RDS I, RDS I/DS, and RDS II, as applied to data from a study of 264 New York City jazz musicians (Heckathorn and Jeffri 2003). Various categorical and continuous variables are estimated. Note that for dichotomous variables such as *Gender* and *Union membership*, RDS I and RDS I/DS give identical estimates.

Table 3. RDS I, RDS I/DS, and RDS II are compared for a real data set. The data come from a survey of 264 New York City jazz musicians (Heckathorn and Jeffri 2003)

Estimator	Gender (male)	Race (white)	Gender (black)	Union membership	Age (mean)
RDS I	76.2%	53.8%	35.0%	25.1%	–
RDS I/DS	76.2%	53.2%	35.9%	25.1%	–
RDS II	72.0%	55.7%	32.8%	23.8%	42.97
Sample (Naive estimate)	73.7%	54.8%	32.8%	39.9%	45.46

For most variables, the estimators give results in line with the sample proportions. However for the *Union Membership* variable, all estimators correctly adjust for the over-counting of union members in the sample data. Although the simulation study demonstrated that the variance of RDS II and that of RDS I/DS are generally very close, individual estimates can diverge appreciably when one is not demographically adjusting the RDS II estimator, such as for the estimation of gender in Table 3.

7. Discussion

This article has further developed RDS estimation theory. A new estimator for RDS data has been presented (RDS II) which offers superior precision to prior methodology (RDS I), with the advantage of increased simplicity, analytical tractability, and analytical variance estimation. The new estimator also allows the estimation of continuous as opposed to categorical variables. The classical RDS estimator requires quite a bit of custom code in order to derive the recruitment matrix and solve the system of linear equations (see Section 4), and is restricted to the estimation of categorical variables.[8]

There are multiple issues that still need to be addressed. The theory developed here relied on the sampling-with-replacement assumption. Biases which may be introduced due to sampling without-replacement are poorly understood. When individuals are eliminated from the pool of potential recruits, not only can they not be reselected into the sample, but all avenues for recruitment that pass through them are also eliminated. If the average degree and population size are small, this can have unpredictable effects on the selection probabilities for everyone in the population.

In RDS samples, it is usually the case that respondents are allowed to recruit more than one person into the study. It is possible for this to introduce biases into the sample, for example if the number of recruits is correlated with the study variable or degree. However, these biases remain poorly understood.

The variance estimator presented here uses the known mixing properties of the population. In general, the mixing matrix will not be known, and will have to be estimated. Furthermore, it is possible for there to exist higher-order correlations between sample elements than are presented in the mixing matrix, so that the estimated covariance between sample units may actually be biased. Such problems are inevitable whenever sampling from a network with unknown composition and structure. Certainly more could be done in improving this estimate of variance, though our simulations indicate that for most applications it should perform well.

An important problem not confronted here is how to fit models such as linear and logistic regressions to RDS data. Model-fitting should incorporate sample weights as well as information about correlation between sample units.

The refinements to RDS theory outlined in this article should prove useful in exploring these problems, and as RDS is applied with greater frequency around the world, should find wide application.

[8] See http://www.respondentdrivensampling.org for downloads of RDS software for computing the classical RDS estimator.

8. References

Birnbaum, Z.W. and Sirken, M.G. (1965). Design of Sample Surveys to Estimate the Prevalence of Rare Diseases: Three Unbiased Estimates. Vital and Health Statistics, Series. 2, No. 11.

Cochran, W.G. (1977). Sampling Techniques. New York: Wiley.

Erdős, P. and Renyi, A. (1959). On Random Graphs. Publicationes Mathematicae, 6, 290–297.

Felix-Medina, M. and Thompson, S. (2004). Combining Link-Tracing Sampling and Cluster Sampling to Estimate the Size of Hidden Populations. Journal of Official Statistics, 20, 19–38.

Frank, O. (1978). Sampling and Estimation in Large Social Networks. Social Networks, 1, 91–101.

Frank, O. and Snijders, T. (1994). Estimating the Size of Hidden Populations Using Snowball Sampling. Journal of Official Statistics, 10, 53–67.

Hansen, M.H. and Hurwitz, W.N. (1943). On the Theory of Sampling from Finite Populations. The Annals of Mathematical Statistics, 14(4), 333–362.

Hastings, W.K. (1970). Monte Carlo Sampling Methods Using Markov Chains and Their Applications. Biometrika, 57, 97–109.

Heckathorn, D. (1997). Respondent Driven Sampling: A New Approach to the Study of Hidden Populations. Social Problems, 44, 174–199.

Heckathorn, D. (2002). Respondent Driven Sampling II: Deriving Valid Population Estimates from Chain-Referral Samples of Hidden Populations. Social Problems, 49, 11–34.

Heckathorn, D., Semaan, S., Broadhead, R.S., and Hughes, J.J. (2002). Extensions of Respondent-Driven Sampling: A New Approach to the Study of Injection Drug Users Aged 18–25. AIDS and Behavior, 6(1), 55–67.

Heckathorn, D. and 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, National Endowment for the Arts Research Division Report 43, Washington DC, 2003, 48–61.

Lang, S.S. (2004). New Sampling Method to Track HIV-Risk Behavior. Medical News Today, November 28, Hastings, East Sussex, U.K.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., and Teller, E. (1953). The Monte Carlo Method. Journal of Chemical Physics, 21, 1087.

Newman, M.E.J. (2002). Assortative Mixing in Networks. Physical Review Letters, 89, 208701.

Newman, M.E.J., Strogatz, S.H., and Watts, D.J. (2001). Random Graphs with Arbitrary Degree Distributions and Their Applications. Physical Review E, 64, 026118.

Ramirez-Valles, J., Heckathorn, D.D., Vázquez, R., Diaz, R.M., and Campbell, R.T. (2005). From Networks to Populations: The Development and Applications of Respondent-Driven Sampling Among IDUs and Latino Gay Men. AIDS and Behavior, 9(4), 1–16.

Rothenberg, R.B., Woodhouse, D.E., Potterat, J.J., Muth, S.Q., Darrow, W.W., and Klovdahl, A.S. (1995). Social Networks in Disease Transmission: The Colorado

Springs Study. In Social Networks, Drug Abuse, and HIV Transmission, R.H. Needle, S.G. Genser, and R.T. Trotter, II, (eds). Rockville, MD: National Institute of Drug Abuse, NIDA Research Monograph 151.

Salganik, M.J. and Heckathorn, D.D. (2004). Sampling and Estimation in Hidden Populations Using Respondent-Driven Sampling. Sociological Methodology, 34, 193–239.

Sirken, M.G. (1998). A Short History of Network Sampling. Proceedings of the American Statistical Association, Survey Research Methods Section.

Spreen, M. and Zwaagstra, R. (1994). Personal Network Sampling, Outdegree Analysis, and Multi-Level Analysis: Introducing the Network Concept in Studies of Hidden Populations. International Sociology, 9, 475–491.

Sudman, S. and Kalton, G. (1986). New Developments in the Sampling of Special Populations. Annual Review of Sociology, 12, 401–429.

Thompson, S.K. (1998). Adaptive Sampling in Graphs. Proceedings of the American Statistical Association, Survey Research Methods Section.

Watts, D.J. (1999). Small Worlds: The Dynamics of Networks Between Order and Randomness. Princeton, NJ: Princeton University Press.

Received December 2004
Revised October 2007