5

SAMPLING AND ESTIMATION IN HIDDEN POPULATIONS USING RESPONDENT-DRIVEN SAMPLING

*Matthew J. Salganik**
Douglas D. Heckathorn†

Standard statistical methods often provide no way to make accurate estimates about the characteristics of hidden populations such as injection drug users, the homeless, and artists. In this paper, we further develop a sampling and estimation technique called respondent-driven sampling*, which allows researchers to make asymptotically unbiased estimates about these hidden populations. The sample is selected with a snowball-type*

This research was made possible by grants from the Centers for Disease Control and Prevention (U62/CCU114816-01), the National Institute on Drug Abuse (RO1 DA08014), and the National Endowment for the Arts. For MJS this material is based on work supported under a National Science Foundation Graduate Research Fellowship, Cornell's IGERT program in nonlinear systems funded by NSF Grant DGE-9870681, and a Fulbright Fellowship with support from the Netherland-America Foundation, which allowed him to spend the year at the ICS/Sociology department at the University of Groningen, where most of this paper was written. We would like to thank Marijtje van Duijn, Henk Flap, Andrew Gelman, Steve Morgan, Michael Schweinberger, Tom Snijders, and Frans Stokman for valuable comments on earlier versions of this paper. Peter Dodds and Steve Thompson provided helpful conversation. Some parts of this paper were presented at the 2003 Sunbelt International Social Network Conference, 2003 American Sociological Association Annual Meeting, and the 2003 Joint Statistical Meetings.

Direct correspondence to Matthew J. Salganik at <mjs2105@columbia.edu> and Douglas D. Heckathorn at <douglas.heckathorn@cornell.edu>.

*Columbia University

†Cornell University

design that can be done more cheaply, quickly, and easily than other methods currently in use. Further, we can show that under certain specified (and quite general) conditions, our estimates for the percentage of the population with a specific trait are asymptotically unbiased. We further show that these estimates are asymptotically unbiased no matter how the seeds are selected. We conclude with a comparison of respondent-driven samples of jazz musicians in New York and San Francisco, with corresponding institutional samples of jazz musicians from these cities. The results show that some standard methods for studying hidden populations can produce misleading results.

1. INTRODUCTION

The problem of collecting accurate information about the behavior and composition of social groups arises in many areas of research. In most cases, standard sampling and estimation techniques, developed over the past 70 years, provide a means for collecting such information. However, there are a number of important groups for which these techniques are not applicable.

For example, injection drug users and men who have sex with men are populations of great interest to researchers because the behavior of these groups affects the spread of HIV/AIDS and other diseases. Unfortunately, standard sampling and estimation techniques cannot be used on these populations. A recent report by the World Health Organization cited inability to monitor the behavior and HIV seroprevalence of at-risk subpopulations, like injection drug users and men who have sex with men, as a major weakness in current HIV prevention efforts (World Health Organization, 2000). This limitation of current statistical methodology has received attention not just from public health researchers, but also from statisticians and sociologists.

Standard sampling and estimation techniques require the researcher to select sample members with a known probability of selection. In most cases this requirement means that researchers must have a sampling frame, a list of all members in the population. However, for many populations of interest such a list does not exist.

A researcher wishing to study a population without a sampling frame could attempt to construct such a frame. However, for a number of populations this frame construction is made impractical or impossible by, first, the small size of the target population and,

second, the difficulty of locating members of the target population. This difficulty could be caused by the sensitive nature of the behaviors in the population (for example, injection drug users) or simply because members of the target population are difficult to distinguish from members of the general population (for example, jazz musicians). These special populations that cannot be studied using standard sampling and estimation techniques are called *hidden populations*. A short list of examples includes injection drug users, men who have sex with men, artists, commercial sex workers, illegal immigrants, participants in some social movements, draft resisters, and the homeless.

Inability to collect information about hidden populations has complicated existing studies and forced researchers to focus on other problems. Imagine a researcher who wishes to begin a study of, for example, injection drug users. This researcher is immediately faced with a huge problem—how to get a sample from the hidden population from which one can generalize.

A first thought might be to attempt to construct a sampling frame and then, once the frame is complete, select people with known probability of selection from the frame. This approach brings the problem back under the purview of standard sampling and estimation. However, for reasons discussed earlier, frame construction is extremely expensive and probably impossible—imagine trying to make a list of all the injection drug users in a large city.

Another approach would be to reach a large number of people via random digit dialing and then screen them for membership in the hidden population. Again, this approach is extremely costly and potentially very inaccurate. For example, if 1 percent of a city is a member of the hidden population it would take approximately 50,000 screening interviews to yield a sample of size 500. Further, it is unlikely that the resulting sample would be a simple random sample from the hidden population because many members are not reachable by phone, or would not reveal their behavior to an unknown interviewer over the phone. The nature and the magnitude of these biases would be unknown, and thus there would be no way to generalize from this type of sample to the entire hidden population.

A more efficient way to collect information about the behavior and composition of hidden populations is to take a sample of target population members in an institutional setting—for example, injection drug users in a drug rehabilitation program. This can provide researchers

with valuable information, but since the members of the population who enter institutional settings are a nonrandom sample from the hidden population, it is impossible to use samples from institutional settings to make accurate estimates about the entire hidden population. For example, Watters and Cheng (1987) found that in San Francisco, injection drug users outside of drug treatment programs were twice as likely to be infected with HIV as injectors in treatment programs.

Because of the problems with these three approaches, they are rarely used. The two most common approaches, targeted sampling and time-space sampling, will be discussed more thoroughly in Section 2. These two methods, as well as those previously discussed, often fail to provide researchers with a way of making accurate estimates about a hidden population based on a sample. These methods also have another common feature: they implicitly treat members of the population as discrete, atomized units. By doing so, these methods fail to make use of an important feature of many hidden populations—they are made of real people connected in a network of relationships.

Switching to this network perspective gives us a fresh and novel approach to the study of hidden populations. Using the extra information that is available in the social network allows one to design a sampling and estimation scheme that, in many cases, is both cheaper and more accurate than existing methods commonly in use.

This new sampling and estimation method, called *respondent-driven sampling*, is a variation of chain-referral sampling methods that were first introduced by Coleman (1958) under the name *snowball sampling*.[1] The basic idea behind these methods is that respondents are selected not from a sampling frame but from the friendship network of existing members of the sample. The sampling process begins when the researchers select a small number of seeds who are the first people to participate in the study. These seeds then recruit others to participate in the study. This process of existing sample members recruiting future sample members continues until the desired sample size is reached.

Experience with chain-referral methods has shown them to be effective at penetrating hidden populations. However, researchers also correctly realized that the promise of chain-referral methods was tempered by the difficulty of making statistical inferences from this type of sample.

[1]Other variations of snowball sampling have also appeared under the names *link-tracing sampling* and *random-walk sampling*.

This difficulty occurs because chain-referral methods produce samples that are not even close to simple random samples. For example, in a simple random sample all people have the same probability of selection. However, in chain-referral samples this is far from the case. Since people recruit their friends, those with many friends are more likely to be included in the sample than social isolates.

Another main concern among researchers centers around the choice of seeds (the first people to be included in the sample). Since all people in the sample are indirectly recruited by the seeds, researchers believed that any small bias in selecting the seeds would be compounded in unknown ways as the sampling process continued.

Both of these specific problems occur because chain-referral designs fall outside of the realm of traditional probability sampling, where units are selected from a sampling frame with known probability of selection. Instead, because of a lack of sampling frame and unknown probability of selection, chain-referral samples have been considered to be nonprobability or convenience samples "which can only be assessed by subjective evaluation" (Kalton 1983).

Because of the nature of the chain-referral samples, numerous researchers have questioned the estimates that can be drawn from them (Welch 1975; Erickson 1979; Kalton 1983; Kalton and Anderson 1986; Berg 1988; Spreen 1992; Friedman 1995; Eland-Goosensen et al., 1997). This research is fairly well summarized by Berg (1988) when he writes, "as a rule, a snowball sample will be strongly biased toward inclusion of those who have many interrelationships with or are coupled to, a large number of individuals. In the absence of knowledge of individual inclusion probabilities in different waves of the snowball sample, unbiased estimation is not possible." It would be fair to say that the conventional wisdom among sociologists, public health researchers, and statisticians is that chain-referral sampling holds great promise for a number of problems, especially the study of hidden populations, but that it is so hopelessly biased that it cannot be used to make reliable estimates.

We believe that previous researchers have been overly pessimistic about chain-referral samples. In this paper we will show that it is possible to make unbiased estimates about hidden populations from these types of samples. Further, we will show that these estimates are asymptotically unbiased no matter how the seeds are selected.

2. EXISTING METHODS FOR STUDYING HIDDEN POPULATIONS

Most studies about the characteristics of hidden populations use either targeted sampling or time-space sampling.[2] These approaches differ in their strengths and weaknesses, but one problem that they both share is that, for many hidden populations, they allow no systematic or principled way to use the information collected to make inferences about the population from which they are drawn.

In targeted sampling (Watters and Biernacki 1989) researchers use a number of different outreach techniques to attract a sample of people in the hidden population. This technique is also sometimes called "street outreach" because it generally involves sending field-workers into to the streets to find, and recruit, members of the hidden population. Some of these studies also make use of some type of chain-referral technique to recruit additional people into the sample.

Targeted sampling does succeed in giving researchers access to a large sample of noninstitutional members of the population. However, the targeted sample is clearly not a random sample where all people have the same probability of selection. For example, in studies of injection drug users, safety concerns often require that recruitment occur only during the day when most drug scenes are less active. Also, injection drug users who do not congregate in public are almost certain to be missed in the sampling process. There is no way to know the magnitude of these selection biases, so it is not possible to generalize from the sample to the target population.

A more refined alternative to targeted sampling is time-space sampling (Muhir et al. 2001). Under this procedure, ethnographic fieldwork is done to construct a sampling frame identifying times when members of the target population gather at specific locations—for example, Tuesday afternoon from 2 PM

[2]There have also been a number of approaches to estimating the size of hidden populations. This research has shown that estimating the size of a population can be different from estimating its behavior and characteristics. Some examples are based on snowball sampling (Frank 1979; Frank and Snijders 1994; Dávid and Snijders 2002), network scale-up designs (Killworth et al. 1998a, b), multiplicity sampling (Sirken 1970), and capture-recapture (Sudman, Sirken, and Cowan 1988; Hogan 1993; Heckathorn et al. 2002; Heckathorn and Jeffri 2003).

to 6 PM at a specific park. These specific venue-day-time segments are the primary sampling units. These units are randomly selected, in some cases with probabilities based on the expected sample yield at the location, and members of the target population entering the venue are intercepted and interviewed.

Because the venue-day-time segments are sampled with a known probability, it is possible to use the sample to make statistical inferences about the population that attends the identified venues. Unfortunately, all venues are not accessible. For example, some studies do not sample from venues with low expected sample yields because it is prohibitively costly (Stueve et al. 2001). Additionally, private venues are generally not accessible to the researchers and in some studies safety concerns further limit the choice of venues (see Kanouse et al. [1999] for an example).

Time-space sampling produces probability samples of the population that attend venue-day-time segments, which are accessible to researchers. However, in some situations this venue-attending population differs in unknown ways from the true population of interest. For example, drug injectors who appear in public places accessible to researchers probably are not representative of all drug injectors.[3] These coverage problems introduce unknown bias into the estimates, a bias that could be substantial in some situations and minimal in others depending on the population under study.

In summary, most current studies of hidden populations require researchers to expend a lot of effort to collect a sample from which they can not generalize to the population of interest. Instead, they present summary statistics about the sample and then leave interpretation to the reader. As we will see in Section 9, this approach can lead people to make extremely misleading conclusions. Clearly, we would prefer a way to collect a sample from which it is possible to accurately generalize to the population.

[3]One potential solution to this coverage problem is to change the definition of the population of interest. For example, we could study brothel-based sex workers if all brothels are accessible for researchers. However, if there are differences between the brothel-based sex workers and nonbrothel-based sex workers, it must be made clear that the estimates refer only to brothel-based sex workers and not sex workers in general.

3. A NEW APPROACH TO THE STUDY OF HIDDEN POPULATIONS

In this paper we develop a new approach to studying hidden populations that is based on a technique called *respondent-driven sampling* (Heckathorn 1997, 2002). In respondent-driven sampling, a sample is collected using a chain-referral procedure. That is, respondents are selected not from a sampling frame but from the social network of existing members of the sample.

Unlike a conventional probability sampling design, where each unit has a known and constant probability of selection, respondent-driven sampling is based on an adaptive sampling design where the selection procedure is affected by the realized network in the population (Thompson and Seber 1996; Thompson and Frank 2000). Therefore, special estimation procedures must be used.

Rather than attempting to directly estimate from the sample to the population, as in traditional sampling and estimation, as shown in Figure 1(a), respondent-driven sampling uses an indirect method. First, the sample is used to make estimates about the social network connecting the population. Then, this information about the social network is

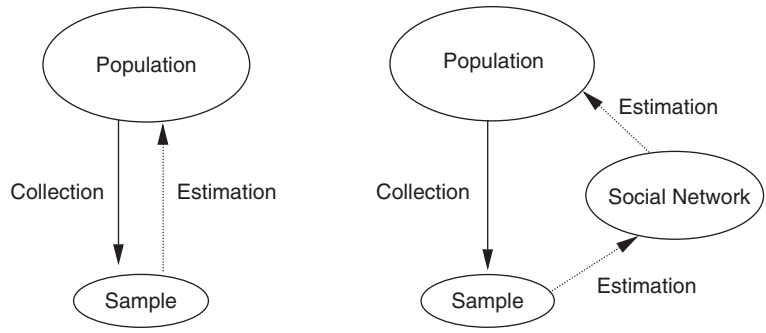


FIGURE 1. Schematic representations showing the differences between traditional sampling and estimation and respondent-driven sampling. By not attempting to estimate directly to from the sample to the population, respondent-driven sampling avoids many of the well-known problems with estimation from a chain-referral sample.

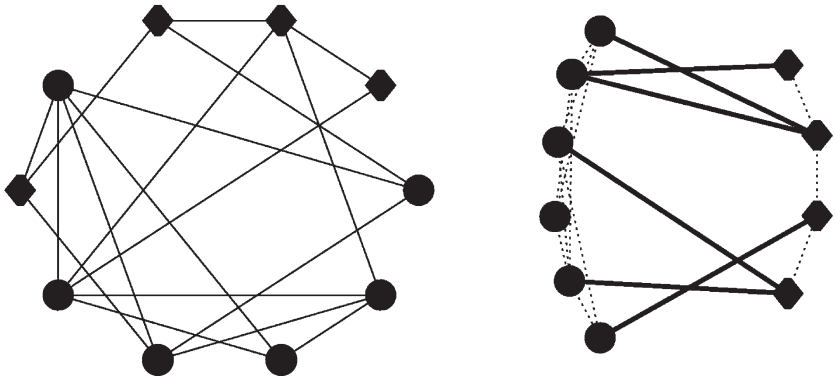
used to derive the proportion of the population in different groups (for example, HIV+ or HIV−). This process is illustrated in Figure 1(b).

To explain how respondent-driven sampling works, in Section 4 we present the method for using information about the social network to make estimates about the characteristics of the population. In Section 5, we then describe how to collect a sample that can be used to estimate the necessary information about the social network. Section 6 presents the estimation techniques and conditions under which they are unbiased. These analytic results are supported by simulation results presented in Section 7. In Section 8 we present results showing that the estimates are asymptotically unbiased no matter how the seeds are selected. We conclude in Section 9 with some real data from a study of jazz musicians that shows that institutional samples and naive snowball samples can produce misleading results.

4. USING THE SOCIAL NETWORK TO MAKE ESTIMATES ABOUT POPULATION PROPORTIONS

The key idea behind the estimation procedure is that the estimates do not come from the sample proportions. Rather, the sample is used to make estimates about the network connecting the population. Then, using information about this network, we can derive the population proportion in different groups. By not attempting to estimate directly from the sample to the population, we avoid many of the well-known problems with chain-referral samples. To estimate from the social network to the population, we will use the reciprocity model that was introduced in Heckathorn (2002).

Consider a hypothetical population such as that shown in Figure 2(a). The population is made up of two groups of people (for example, HIV+ and HIV−), and we would like to estimate the proportion of the population in each of these groups. By redrawing the same population in a different way, as shown in Figure 2(b), we are able to notice that the number of ties from someone in group A to someone in group B, which in this example is 6, is the same as the number of ties from someone in B to someone in group A. This statement may seem quite trivial, but it turns out to be very useful



(a) A visual representation of a hypothetical population

(b) An alternative representation of the same population

FIGURE 2. Two different representations of the same hypothetical population made up of two types of people—circles (●) and diamonds (◆). In (b) the population is drawn to emphasize the difference between within-group friendships (represented by dotted lines) and between-group friendships (represented by solid lines). Using just information about the network structure and equations (8) and (9), we are able to correctly estimate that 60 percent of the population are circles and 40 percent are diamonds.

because there are two different ways to calculate this number of cross-group ties. First, we must begin with some notation.

We can store all the information about the social network in an adjacency matrix **X**. That is, $x_{ij} = 1$ if there is a directed edge between person i and j, otherwise $x_{ij} = 0$. In this paper we will consider only reciprocal relations, so if it is the case that $x_{ij} = 1$, then it is also the case that $x_{ji} = 1$. For convenience, we will call these relationships friendships, although other relationships are possible.

An important property of a person is the number of friends that he or she has. We define the degree of a person i, d_i, to be the number of friendships that involve person i, $d_i = \Sigma_j \, x_{ij}$. The total number of friendships radiating from people in group A, R_A, is the sum of the degree of all people in group A and is defined as

$$R_A = \sum_{i \in A} d_i = N_A \cdot D_A, \tag{1}$$

where N_A is the number of people in group A and D_A is the average degree of people in group A.

Now consider, for a given network $\mathbf{X}$, the probability that if we follow a randomly chosen friendship beginning with a person in group A that we cross groups and end up at someone in group B. We can define this probability, $C_{A,B}$, as

$$C_{A,B} = \frac{T_{AB}}{R_A}, \tag{2}$$

where T_{AB} is the number of ties that contain a person in group A and a person in group B.

Since we are considering only reciprocal ties, we know that the number of relationships from group A to group B, in this example 6, is the same as the number of relationships from group B to group A. We can calculate that number two different ways: (1) the number of friendships radiating from group A, R_A, times the probability that one of those relationships will lead to someone in group B, $C_{A,B}$ or (2) the number of friendships radiating from group B, R_B, times the probability that one of those relationships will involve someone from group A, $C_{B,A}$. That is,

$$R_A \cdot C_{A,B} = T_{AB} \tag{3}$$

$$R_B \cdot C_{B,A} = T_{AB}. \tag{4}$$

Setting equations (3) and (4) equal to each other and using the definition of R_A and R_B, we can write

$$N_A \cdot D_A \cdot C_{A,B} = N_B \cdot D_B \cdot C_{B,A}. \tag{5}$$

Note that equation (5) brings together both information about the characteristics of the nodes and characteristics of the network. Thus, we can begin to see how we can infer properties of the nodes from information about the network.

However, even if we had complete information about the social network—that is, if we knew D_A, D_B, $C_{A,B}$, and $C_{B,A}$—we still have one equation with two unknowns—N_A and N_B, the sizes of the population in group A and group B. If we divide both sides of equation (5) by N, the total size of the population, then we can rewrite

equation (5) in terms of population proportions, PP_A and PP_B. This allows us to add a second constraint—that the sum of the population proportions must be 1. So now we have

$$PP_A \cdot D_A \cdot C_{A,B} = PP_B \cdot D_B \cdot C_{B,A} \tag{6}$$

$$PP_A + PP_B = 1, \tag{7}$$

where PP_A is the *proportion* of the population in group A and PP_B is the *proportion* of the population in group B.

Now we have a system with two equations and two unknowns. Using ordinary algebra, we can derive that

$$PP_A = \frac{D_B \cdot C_{B,A}}{D_A \cdot C_{A,B} + D_B \cdot C_{B,A}} \tag{8}$$

$$PP_B = \frac{D_A \cdot C_{A,B}}{D_A \cdot C_{A,B} + D_B \cdot C_{B,A}}. \tag{9}$$

Examination of equations (8) and (9) reveals that we can recover the population proportions, PP_A and PP_B, with knowledge only of the network structure connecting the population.

Again we can return to the hypothetical population in Figure 2. Using the values $D_{\bigcirc} = 4$, $D_{\diamond} = 3$, $C_{\bigcirc,\diamond} = 0.25$, and $C_{\diamond,\bigcirc} = 0.5$ as well as equations (8) and (9), we can correctly estimate that 60 percent of the population are circles and 40 percent are diamonds. It is important to note that these equations are true for any network structure that contains only reciprocal relations.

We can see now that it is possible to estimate the proportion of the population in group A and group B, but only if we know some other information about the network connecting people in these groups. Next we will discuss methods for collecting a sample that can be used to estimate this network information.

5. COLLECTING THE SAMPLE

A respondent-driven sample is collected via a chain-referral design. These designs were originally introduced by Coleman (1958), and

later elaborated by Goodman (1961) in order to study characteristics of social networks. Work by Snijders (1992) has further reinforced the point that chain-referral designs are more appropriate for inference about the structure of the network than the characteristics of the people in the network. However, much confusion has arisen because people have ignored this advice and attempted to use chain-referral samples to make estimates directly about the population, and not the network connecting the populations.

To conduct a respondent-driven sample, one begins by selecting a set of s initial seeds that are chosen based on preexisting contact with the study population. These seeds are paid to be interviewed and form wave 0 of the sample.

Interviews for these studies can be organized in a number of ways depending on the target population. A study of injection drug users conducted interviews in a storefront office in a location accessible to the target population (Heckathorn 2002). However, in a study of jazz musicians, interviewers traveled to locations convenient to the respondents (Heckathorn and Jeffri 2001). In some cases telephone interviews may be preferred, as in a study of Vietnam War draft resisters (Hagan 2001).

No matter how the interview is conducted, each seed in wave 0 is supplied with c unique recruitment coupons similar to the one in Figure 3. Subjects are told to give these coupons to other people they know in the target population. Because each coupon is unique, it can be used to trace the recruitment patterns in the population. When a new member of the target population participates in the study, the recruiter of that person is paid an additional bonus. Thus, subjects are paid to participate and to recruit others.

FIGURE 3. An example of a recruitment coupon given to subjects.

The recruits of people in wave 0 form wave 1. Upon completion of the screening, a person from wave 1 is provided with c recruitment coupons and the process continues. The sampling continues in this way until the desired sample size is reached.

The choice of the number of recruitment coupons given to each subject, c, depends on the expected recruitment behavior of the people in the sample. For the sampling process to continue, each participant must recruit on average at least one new participant. So, the choice of c should be large enough so that the recruitment continues even if some subjects choose not to recruit. However, it is also desirable to have a small c because it means that the sample can go through many waves before the desired sample size is reached.

If the seeds are not drawn as desired (as explained in Section 6.1), many sampling waves are preferable because they allow the sampling process to explore parts of the network that may have had a zero probability of being included as a seed. These long chains are important because all members of the hidden population must have a nonzero chance of being included in the estimation procedure. Work on the "small-world" problem (Watts and Strogatz 1998; Watts 1999; Dodds et al., 2003) seems to indicate that in most networks the average path length between any two people is often quite short. Thus, we can be fairly confident that with reasonably long recruitment chains all members of the population have a nonzero probability of being reached even if the seeds occupy an unusual place in the network structure. Also, in the case where the seeds are not selected as desired, many recruitment waves are preferable because they allow the sampling process to converge to an equilibrium (as we will see in Section 8).

In other chain-referral type sampling techniques—for example, the random-walk design by Klovdahl (1989)—respondents are asked to list the people they know in the population under study, and then the researcher uses a randomization device to select the people to be included in the next wave of the sample. While appropriate for Klovdahl's study of the social network of a large city, there are problems with this approach to sampling in hidden populations.

First, since the populations under study are often subject to social stigma and/or criminal prosecution, respondents are often hesitant to give researchers information about their colleagues in the population. This bias, sometimes called masking (Erickson 1979),

can cause respondents to provide inaccurate information to researchers leading to unknown biases in the sample selection process. Additionally, once the next member of the sample is selected via the randomization device, it is often difficult to locate the person so that they can be interviewed (McGrady et al. 1995; Eland-Goosensen et al., 1997). Finally, this procedure is sometimes forbidden in the United States because, by requiring respondents to divulge sensitive information about their peers, it violates federal guidelines for the protection of human subjects (Heckathorn et al., 2002).

Respondent-driven sampling solves these problems by allowing the subjects to do the recruitment. Thus, participants are not required to divulge any sensitive information to the researcher, and the researcher does not have to spend time looking for the named recruitee. One concern with having the subjects do the recruitment is that it could introduce unknown bias into the sampling process if the recruitments are not random. However, there is some evidence that is consistent with the idea that respondents recruit randomly from their friends (Heckathorn et al. 2002).

In addition to the substantive information a researcher desires to collect, two additional pieces of information must be collected from each participant. These additional pieces of information are needed to implement the estimation procedure.

First, the researcher must collect the degree of each person in the sample.[4] Second, the researcher must use the coupons to record the recruitment patterns so that each sample member can be linked to the person who recruited them. These two pieces of information, the self-reported degree and the observed recruitment patterns, will be critical for the estimation procedures that will be described more fully in Section 6.

There are two additional important steps in the sample selection process: nonduplication and population membership verification. It is a common problem in studies of drug injectors that some respondents attempt to participate in the study multiple times in order to earn additional money (Biernacki and Waldorf 1981). This duplication can affect the quality of the data and thus accuracy of the estimates.

[4]It is known that there are a number of difficulties in collecting the degree of a person accurately (McCarty et al. 2001). This source of nonsampling error is not considered in this paper.

It is an important, but difficult, problem to ensure that each person can participate only once. Because the respondents in many hidden populations may be unwilling or unable to provide photographic identification, some other methods must be used to prevent duplication. Previous researchers have had success using biometric measurements that are not threatening to the respondents (Heckathorn, Broadhead, and Sergeyev 2001).

Another important step in ensuring the quality of the sample data is to verify that sample members are indeed members of the target population. Again, there is a possibility that people who are not in the population under study may attempt to participate in the study for financial reasons. Methods of population verification must be tailored to the specific population under study. For example, urinalysis can be used to verify recent use of certain drugs (Kral et al. 1998).

This sampling methodology was designed to be implementable in the real world and does not require an unrealistic amount of funding, time, or cooperation by subjects. Because the design is relatively simple and robust, it has already been used successfully in a number of studies. Also, a recent review article by Semaan, Lauby, and Liebman (2002) found that respondent-driven sampling is cheaper, quicker, and easier to implement than other methods that have been used for sampling hidden populations in evaluations of HIV risk-reduction interventions.

At the end of the sampling procedure, the researcher has a sample that consists of people who are seeds, people who are not seeds, and a set of recruitments. In the next section, we will discuss how to use the information in the sample to make estimates about the social network. This information about the social network will be used to make estimates about the population.

6. USING THE SAMPLE TO MAKE ESTIMATES ABOUT THE SOCIAL NETWORK

Once we have selected a sample, we must then have a procedure to use information from that sample to make estimates about the social network from which it was drawn. Previous work on estimating proprieties of networks have often required a random sample of nodes (Frank 1981). However, it is not possible to collect such a sample from a

hidden population. This impossibility forces us to make estimates from a chain-referral sample, which is notoriously difficult.

6.1. *Assumptions*

In order to make estimates from a respondent-driven sample, we must first assume some things about the population under study and the way that the recruitment occurs. By making these assumptions explicit, we allow for them to be tested and for research to be done about the nature of the bias introduced by their violation.

It is helpful to think of the sample selection as a process that alternates between the selection of nodes and the selection of edges. That is, nodes are first drawn to form wave 0 of the sample. Then these nodes choose edges that define recruitment period 1. The edges in recruitment period 1 determine the nodes drawn in wave 1. The process continues in this way with nodes selecting edges that in turn select nodes until the desired sample size is reached.

To make our assumptions more clear, we must first develop some notation. We begin by setting up two indicator functions: one for the nodes, $NI(j)_{w=x}$, which tells us whether a given node, j, is selected in wave x and one for the directed edges, $EI(e_{j\to k})_{r=x}$, which tells us whether a given directed edge $e_{j\to k}$ is selected in recruitment period x. These indicator functions are defined as follows:

$$NI(j)_{w=x} = \begin{cases} 1 & \text{if node } j \text{ is selected in wave x} \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

$$EI(e_{j\to k})_{r=x} = \begin{cases} 1 & \text{if edge } e_{j\to k} \text{ is selected in recruitment x} \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

First, throughout this paper we will consider only the case of sampling with replacement, even though, in practice, the actual sampling is sometimes without replacement. This assumption simplifies the calculations because it eliminates changes in the population as the sampling progresses.[5]

[5]Preliminary simulations for the case of sampling without replacement show that it has a very small effect on the estimates when the sample size is small relative to the population. However, this question deserves further research.

Next, we will assume that the network of the hidden population forms one connected component. That is, we assume that there is a path between every person and every other person. This may seem like a restrictive assumption, but a number of results from random graph theory predict that even for very sparse graphs, almost all nodes belong to one giant component (Newman, 2003b). For many hidden populations, like jazz musicians or injection drug users, the friendship network used in the recruitment is dense enough that this assumption is reasonable.[6] It is important to note that the assumption that the network forms only one component is the only assumption that we need to make about the structure of the network. This lack of potentially untestable network assumptions makes respondent-driven sampling applicable to many different types of hidden populations.

Additionally, we assume that all respondents receive and use one coupon,[7] and that when respondents recruit others, they recruit randomly from all edges that involve them. Some readers may doubt this random recruitment assumption, but there is some empirical evidence that is consistent with this assumption (Heckathorn et al. 2002). To specify this recruitment assumption more clearly, we can write

$$\Pr[EI(e_{j\to k})_{r=x+1} = 1 | NI(j)_{w=x} = 1] = \frac{1}{d_j}. \tag{12}$$

Finally, we will assume, initially, that the seeds are drawn with probability proportional to their degree. That is, a person with 10

[6]However, given this assumption, it may not be wise to do the recruitment over the sexual or drug-sharing networks, which are less dense and potentially consist of multiple components. Also, this network assumption means that respondent-driven sampling is not appropriate for groups like tax evaders, who do not have frequent contact with each other.

[7]These results easily extend to the case of multiple coupons as long as all coupons are used. However, this extension requires messier algebra, and so for ease of presentation we presented the one coupon case. However, once multiple coupons are allowed, new possibilities emerge for heterogeneous recruitment behavior within a group and heterogeneous recruitment behavior between groups. These situations clearly deserve further research.

friends is twice as likely to be a seed as a person with 5 friends. This assumption can be expressed mathematically as[8]

$$\Pr[NI(j)_{w=0} = 1] = \frac{d_j}{\sum_{i \in N} d_i}. \tag{13}$$

We chose to make this assumption because in studies of hidden populations, the people drawn as seeds are often those who are known to the researchers. These more well-known people tend to have more friends than average. Thus, it seems reasonable to assume that a person's chance of being drawn as a seed increases with his or her degree. In order to increase the generality of respondent-driven sampling, this assumption about the seeds being drawn with probability proportional to degree will be relaxed in Section 8.

6.2. *Consequences of the Assumptions*

Since we have assumed that the seeds are drawn with probability proportional to degree, we can make some conclusions about the the probability that certain edges will be drawn in recruitment period 1 and then the probability that certain nodes will be drawn in wave 1.

A relationship $e_{j \to k}$ can be selected in recruitment period 1 only if the node from which it points, node j, is selected in wave 0. Using the indicator function notation and the rules of conditional probability, we can calculate the probability that a given relationship, $e_{j \to k}$, will be drawn in recruitment period 1 as follows:

$$\begin{aligned} &\Pr[EI(e_{j \to k})_{r=1} = 1] = \\ &\quad \Pr[NI(j)_{w=0} = 1] \cdot \Pr[EI(e_{j \to k})_{r=1} = 1 | NI(j)_{w=0} = 1]. \end{aligned} \tag{14}$$

[8]For equation (13) the probabilities should not be considered the overall probability of being selected as a seed but the draw probability in each one of the s draws that select the seeds. To see this further, compare equation (13) with the overall probability that node j will be selected as a seed, which is $1 - \left(\frac{(\sum_{i \in N} d_i) - d_j}{\sum_{i \in N} d_i}\right)^s$.

Since we have assumed that the seeds were drawn with probability proportional to their degree and that people choose randomly from their relationships when making a recruitment, we can rewrite equation (14) as

$$\Pr[EI(e_{j\to k})_{r=1} = 1] = \frac{d_j}{\sum_{i\in N} d_i} \cdot \frac{1}{d_j}, \tag{15}$$

which can be simplified to,

$$\Pr[EI(e_{j\to k})_{r=1} = 1] = \frac{1}{\sum_{i\in N} d_i}. \tag{16}$$

Equation (16) tells us that if the nodes in wave 0 are drawn with probability proportional to degree, then each relationship has the same probability of being drawn in recruitment period 1.

This fact also gives us some information about the draw probabilities of the nodes in wave 1. The probability that a node j will be drawn in wave 1 is equal to the sum of the probabilities that the d_j relationships leading to it will be drawn in recruitment period 1. That is,

$$\Pr[(NI(j)_{w=1} = 1] = \sum_{d_j} \frac{1}{\sum_{i\in N} d_i} = \frac{d_j}{\sum_{i\in N} d_i}. \tag{17}$$

Equation (17) shows us that if the seeds are drawn with probability proportional to degree, then the nodes in wave 1 will also be drawn with probability proportional to degree. Repeating this argument iteratively shows that, if the nodes in wave 0 are drawn with probability proportional to degree, then the nodes in all successive waves also will be drawn with probability proportional to degree.

This argument can also be repeated iteratively to show that, if the seeds are drawn with probability proportional to degree, then the

probability that a specific edge, $e_{j\to k}$, will be drawn in recruitment period x is constant and equal for all edges:[9]

$$\Pr[EI(i)_{r=x} = 1] = \frac{1}{\sum_{i\in N} d_i}. \tag{18}$$

It is important to note that these arguments apply for any network structure with reciprocal ties and thus are quite general. The consequences of the assumptions will prove useful when considering our estimators.

6.3. *Making the Estimates*

Using the results derived in the previous section, we can now derive estimators for specific network properties and show that these estimates are asymptotically unbiased. First, we show how to use the observed recruitment behavior to estimate the probability of cross-group connections, $C_{A,B}$ and $C_{B,A}$. Second, we will estimate the average degree for people in the groups, D_A and D_B, using the self-reported network-size information. Third, we will combine these estimates to estimate the proportion of the population falling into one of two distinct groups, PP_A and PP_B.

6.3.1. *Estimating* $C_{A,B}$ and $C_{B,A}$

Recall that we wanted to use the information from the sample to estimate some properties of the network connecting the population. The first thing that we would like to estimate is the probability that if we follow a random friendship beginning in group A that we cross groups and end up in group B. One way to estimate this probability, $C_{A,B}$, would be to ask respondents what percentage of their friends fall into certain groups. However, this is not possible because for

[9]As was pointed out by a reviewer, equal draw probabilities for all edges does not mean that these probabilities are independent. The edge draw probabilities are correlated. That is, $Pr[EI(e_{i\to j})_{r=t} = 1, EI(e_{j\to k})_{r=t+1} = 1] \neq Pr[EI(e_{i\to j})_{r=t} = 1] \cdot Pr[EI(e_{j\to k})_{r=t+1} = 1]$. This dependence in the edge selection process could introduce difficulties into the estimation of certain network properties. However, for the properties of interest in this paper it does not present a problem.

many items of interest—for example HIV status—respondents may not have enough information about their peers to make accurate assessments. The accuracy of this self-reported data is also doubtful.

Instead of basing the cross-group friendship probabilities on self-reported data, we base them on actual behavior. When one respondent recruits another, this behavioral link represents a network link that can be verified by asking the recruitee to characterize the relationship to the recruiter. Verification requires that the recruitee identify the recruiter as an acquaintance, friend, or closer than friend, rather than as a stranger. Only verified links should be used for estimation.

To be able to construct an estimate for $C_{A,B}$ and $C_{B,A}$, we must know something about how the recruitments we observed are selected from the set of possible recruitments. In Section 6.2, we showed that each edge, $e_{j \to k}$ is equally likely to be selected in each recruitment period. That is, we showed that the recruitments we observe are a random sample of all possible recruitments.

Recall that for each sample we observe a set of recruitments. These recruitments can be divided into four groups—recruitments from a person in group A to another person in group A, r_{AA}, recruitments from a person in group A to a person in group B, r_{AB}, etc.

Since the observed recruitments are a random sample from all edges, unbiased estimates for $C_{A,B}$ and $C_{B,A}$ are

$$\widehat{C_{A,B}} = \frac{r_{AB}}{r_{AA} + r_{AB}} \tag{19}$$

$$\widehat{C_{B,A}} = \frac{r_{BA}}{r_{BB} + r_{BA}}. \tag{20}$$

We now turn our attention to deriving estimators for D_A and D_B, the average degree of nodes in group A and group B.

6.3.2. *Estimating* D_A and D_B

Recall that during the sampling process we collected the degree of each member of our sample. If we tried to estimate the mean degree of people in group A by taking the mean degree of the people in the sample, our estimate would be too high, in some cases much too high because chain-referral methods overrepresent people with high degree

(Erickson 1979; Kalton and Anderson 1986; Eland-Goosensen et al. 1997).

Because the simple mean is not a good estimator, a different estimation procedure is required. That is, we must take the sample data and somehow adjust it so that it yields accurate information about the population.[10] Two distinct approaches can be used to motivate the exact form of this adjustment, and, as we will see, both of these approaches lead us to the same estimator. We can then see that this estimator is asymptotically unbiased.

One approach for constructing an estimator for the average degree, which we call the degree distribution approach, is motivated by the degree distributions of the sample and the population. We begin by recalling that in Section 6.2 we showed that if our assumptions are met, then the nodes are drawn with probability proportional to their degree in all waves. We can use this fact, together with the observed sample degree distribution, $q_A(d)$, to estimate population degree distribution $p_A(d)$. This population degree distribution can be used to estimate the average degree of people in group A, D_A.

Previous work on unrelated problems (Feld 1991; Newman, Strogatz, and Watts, 2001; Newman 2003a) has found that if nodes are drawn with probability proportional to their degree, the sample degree distribution, $q_A(d)$, is given by

$$q_A(d) = \frac{d \cdot p_A(d)}{\sum_{d=1}^{max(d)} d \cdot p_A(d)}, \tag{21}$$

[10]When attempting to estimate the average degree of people in group A, D_A, we recommend that researchers do not include the degree of the seeds. The seeds are selected with a different and potentially unknown mechanism, and they should really be treated differently from the data that are collected via the chain referrals. In practice, since the number of seeds is small (about 5) relative to the sample size (about 500), only a small amount of data are not used. For the simulations presented in this paper, all estimates of average degree exclude the degrees of the seeds. It is somewhat unusual not to use all of the information collected when making the estimates from a sample, but this is not without precedent when estimating after adaptive sampling procedures (for example, see Thompson [1990]).

where $p_A(d)$ is the population degree distribution and $\sum_{d=1}^{max(d)} d \cdot p_A(d)$ is a normalizing constant to ensure that $q_A(d)$ sums to 1.

Equation (21) allows us to predict the sample degree distribution given the population degree distribution. However, we have the opposite problem. That is, we know the sample degree distribution and want to predict the population degree distribution. Since, as described in equation (21), $d \cdot p_A(d)$ is proportional to $q_A(d)$, it is also the case that $p_A(d)$ is proportional to $\frac{1}{d} \cdot q_A(d)$. So, if a sample has a degree distribution, $q_A(d)$, then the population degree distribution, $p_A(d)$, can be estimated as

$$\widehat{p_A(d)} = \frac{\frac{1}{d} \cdot q_A(d)}{\sum_{d=1}^{max(d)} \frac{1}{d} \cdot q_A(d)}, \tag{22}$$

where $\sum_{d=1}^{max(d)} \frac{1}{d} \cdot q_A(d)$ is a normalizing constant.

From this predicted population distribution, $\widehat{p_A(d)}$, we can estimate the average degree in the population, D_A, by recalling that the average of a discrete probability density function $f(x)$ is $\sum_{x=0}^{\infty} x \cdot f(x)$. Since we are using a population distribution approach, we will annotate our estimator, $\widehat{D_A}$, with the superscript *dist*:

$$\widehat{D_A^{dist}} = \sum_{d=1}^{max(d)} d \cdot \widehat{p_A(d)}. \tag{23}$$

In equation (23) the summation in the estimator is indexed by degree and not by sample element. If we rewrite equation (23) indexed by sample element (see Appendix B for the derivation), we get

$$\widehat{D_A^{dist}} = \frac{n_A}{\sum_{i=1}^{n_A} \frac{1}{d_i}}. \tag{24}$$

The estimator presented in equation (24) at first appears unlike an estimator for a population mean because the sample size, n_A, is in the numerator and not the denominator.

However, as we will show, this estimator is equivalent to something that is more familiar—the ratio of two Hanson-Hurwitz estimators,[11] one that estimates R_A, the total number of stubs radiating from people in group A, and one that estimates N_A, the number of people in group A.

We write this Hansen-Hurwitz based estimator for $\widehat{D_A}$ with the superscript hh,

$$\widehat{D_A^{hh}} = \frac{\widehat{R_A}}{\widehat{N_A}} = \frac{\frac{1}{n_A}\sum_{i=1}^{n_A}\frac{1}{p_i}\cdot d_i}{\frac{1}{n_A}\sum_{i=1}^{n_A}\frac{1}{p_i}}, \tag{25}$$

where p_i is the probability that person i will be selected on a specific draw.

At first it seems we have no hope for working with equation (25) because the draw probabilities, p_i, are unknown. However, because people are drawn with probability proportional to degree, we do know that the relative draw probabilities for two nodes, i and k, will be

$$\frac{p_k}{p_i} = \frac{d_k}{d_i} \quad \forall \quad i,k. \tag{26}$$

So for each person we can rewrite the draw probability in terms of a chosen reference person k. Thus, using equation (26) to rewrite equation (25), we get

$$\widehat{D_A^{hh}} = \frac{\frac{1}{n_A}\sum_{i=1}^{n_A}\frac{d_k}{d_i\cdot p_k}\cdot d_i}{\frac{1}{n_A}\sum_{i=1}^{n_A}\frac{d_k}{d_i\cdot p_k}}. \tag{27}$$

[11]Hansen-Hurwitz estimation is a standard procedure when estimating from data where the sampling is with replacement and the units have unequal draw probabilities (Hansen and Hurwitz 1943; Cochran 1977; Brewer and Hanif, 1983). The basic idea, similar to the Horvitz-Thompson estimator, is that the procedure weights each sample element by the inverse of its draw probability. That is, units with a small chance of being selected are counted more.

Since $\frac{d_k}{p_k}$ is constant, we can remove it from the numerator and denominator. Also, canceling the $\frac{1}{n_A}$ terms, we are left with

$$\widehat{D_A^{hh}} = \frac{n_A}{\sum_{i=1}^{n_A} \frac{1}{d_i}}. \tag{28}$$

Note that equation (28) does not require the unknown draw probabilities, p_i, for each node. It requires only information that is collected in the sample.

Also, surprisingly, equations (24) and (28) are identical. That is, we can see that these two very different approaches—one based on degree distributions and one based on Hansen-Hurwitz estimators—have led us to the same estimator. Since these estimators are equal, we will drop the superscripts *hh* and *dist* and refer to the estimator as simply $\widehat{D_A}$.

The numerator and denominator of $\widehat{D_A}$ are both Hansen-Hurwitz estimators that are known to be unbiased (Brewer and Hanif 1983). It is also the case that the ratio of two unbiased estimators is asymptotically unbiased with bias on the order of n^{-1}, where n is the sample size (Cochran 1977). So as n_A, the sample size in group A, gets large, $E[\widehat{D_A}] \to D_A$. Generally, this bias is considered negligible in samples of moderate size (Cochran 1977).

We have now shown that two distinct approaches to estimate the average degree for group A are equivalent and that this estimator, $\widehat{D_A}$, is asymptotically unbiased.

6.3.3. *Putting It All Together to Estimate* PP_A and PP_B

Now that we have estimated certain information about the social network that connects the population, we can use that information to estimate PP_A and PP_B. Plugging equations (28) and (19) into the equations that express the population proportions in terms of network information (equations [8] and [9]), we can now estimate the population proportion as

$$\widehat{PP_A} = \frac{\widehat{D_B} \cdot \widehat{C_{B,A}}}{\widehat{D_A} \cdot \widehat{C_{A,B}} + \widehat{D_B} \cdot \widehat{C_{B,A}}}. \tag{29}$$

Again we have a ratio of asymptotically unbiased estimates that is also asymptotically unbiased. Equation (29) is the central result of this paper. With it, we are claiming that we can make an asymptotically unbiased estimate of the proportion of the population with a specific trait based only on data collected during respondent-driven sampling.

7. COMPUTER SIMULATIONS

In the previous section, we have presented a number of analytic results arguing that our proposed population proportion estimator is asymptotically unbiased. These analytic arguments can be further supported with numerical simulation. Also, simulation will allow us to consider questions that are not analytically tractable, like those considered in Section 8.

In these simulations, we generate a population and then repeatedly take samples (with replacement) from that population. The properties of our proposed estimators can be evaluated by averaging the estimates that come from these many samples. For example, if the average of these estimates matches the true population value, then we have further support for our argument that the estimation procedures we have developed are unbiased. The algorithms used to perform these simulations are described in Appendix A.

In Figure 4 we can see the results of these numerical simulations for a wide range of population proportion values. The average value of the estimates, represented by the circles, is the same as the true value, represented by the dotted line. It seems that averaging these estimates over many repetitions does in fact yield the true population proportion in group A, thus providing strong support for our argument that these estimates are asymptotically unbiased.

Because the estimate is only asymptotically unbiased, and not strictly unbiased, we might be interested in knowing the rate of convergence and the magnitude of the resulting bias. Since our estimate for population proportions is a ratio of two asymptotically unbiased estimates, we know that the bias is on the order of n^{-1}. Thus, as the sample size gets big, the bias gets small very quickly.

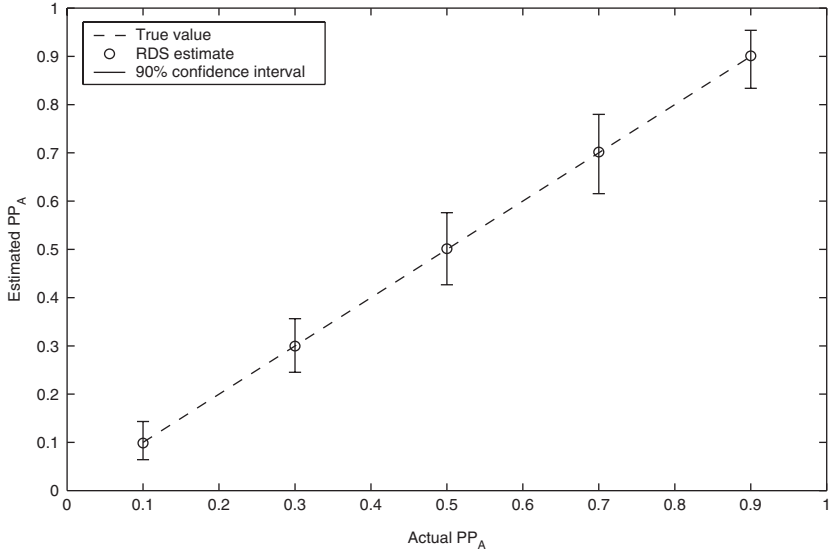


FIGURE 4. Numerical simulations of equation (29) supporting our analytic argument that the respondent-driven sampling population proportion estimates are asymptotically unbiased. With a sample size of only 500, the bias is negligible. The vertical lines show 90 percent confidence intervals around the estimate.

Numerical simulations from a population with parameters described in Appendix A are presented in Figure 5. We see that, in this case, the rate of convergence is rapid and the magnitude of the bias is small. With a sample size of 500, the bias is less than one-tenth of 1 percent (0.001) and therefore not a serious cause for concern. However, more general results about the convergence properties, and the resulting bias, would be desirable.

8. A NOTE ON THE CHOICE OF SEEDS

So far in our derivations we have assumed that the seeds were drawn with probability proportional to degree. This assumption deserves further consideration. Previously, numerous researchers have doubted estimates from chain-referral samples because it is not possible to select the seeds randomly from the target population (Erickson 1979; Spreen 1992; Snijders 1992; Friedman 1995). In a review article

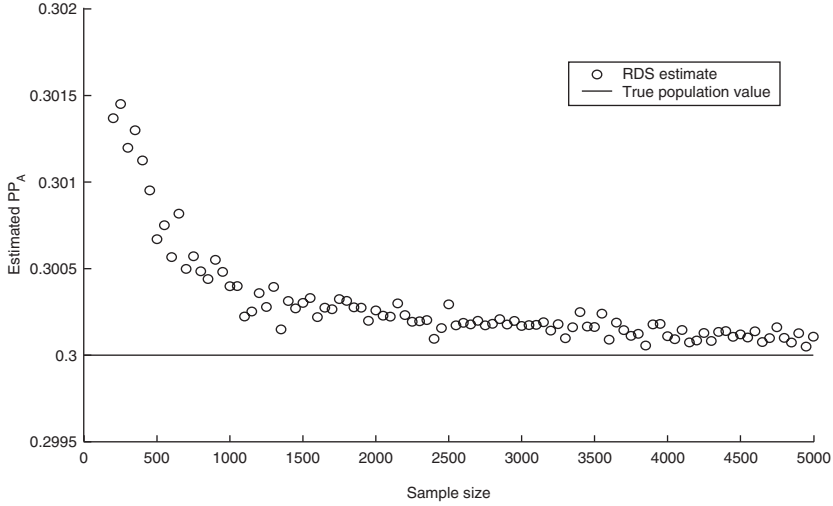


FIGURE 5. Numerical simulations supporting our argument that the rate of convergence to the true estimate is quite rapid. The circles represent the average estimate from 100,000 samples. In this case, the magnitude of the bias is also small. For a sample size of 500 the respondent-driven sampling estimate is within one-tenth of 1 percent (0.001) of the true population value.

on chain-referral methods, Spreen (1992) summarizes this and other literature: "[T]he central question in the methodological discussion about sampling and analyzing hidden populations is basically: How to draw a random (initial) sample?" Given that the choice of seeds has long been considered a critical weakness of chain-referral methods, it is natural to devote further attention to this topic.

Most previous chain-referral methods have assumed that the seeds are a simple random sample from the population. However, when actually implementing chain-referral sampling, researchers often select people known to them to serve as seeds. These known individuals are far from representative, and usually have much higher degrees than other members of the hidden population. Unlike other chain-referral methods, in respondent-driven sampling the estimates are unbiased if the seeds are drawn with probability proportional to degree—much closer to what actually happens in the field.

However, even though our assumption is closer to what actually happens, it is still the case that it is unlikely that the seeds will be drawn exactly with probability proportional to degree. Fortunately, it turns out that using a Markov chain argument we can show that *our estimates are asymptotically unbiased no matter how the seeds are chosen*.

8.1. *Convergence of Estimates for Arbitrary Choice of Seeds: Analytic Results*

Previously in this paper, we have shown that our estimates are asymptotically unbiased if the seeds are drawn with probability proportional to degree. This assumption about the seeds selection mechanism was important because it allowed us to derive two important conclusions: (1) that nodes are drawn with probability proportional to their degree in all waves and (2) that edges are drawn with uniform probability. These conclusions were critical in showing that our estimates were asymptotically unbiased.

However, it turns out that even if the seeds are not drawn with probability proportional to degree, the sampling process converges to one in which people are drawn with probability proportional to degree. So, after a number of waves the draw probability for the nodes will converge to the situation that we needed to make our estimates unbiased.

We can see this by setting up a Markov chain on the nodes[12] as in Thompson (2003). A Markov chain is a natural way of modeling this problem because the draw probabilities for nodes in wave w are only dependent on the draw probabilities of nodes in wave $w-1$.

To begin, let's construct a $1 \times N$ vector, $\vec{\pi}^{(t)}$, which stores the draw probability for each person j in wave t. That is,

$$\vec{\pi}_j^{(t)} = \Pr[NI(j)_{w=t} = 1]. \tag{30}$$

[12]In a previous work on respondent-driven sampling, Markov chains have been used on the population groups (Heckathorn 1997, 2002). These previous Markov chain arguments showed that the sample proportions converged independent of the seeds. This is different than the claim here, which is that the respondent-driven sampling estimates converge to the true population value.

Now let's construct an $N \times N$ transition probability matrix, $\mathbf{P}$, which stores the probability that if we are at node i at time w then we will be at node j at time $w+1$. That is,

$$p_{ij} = \begin{cases} 1/d_i & \text{if } x_{ij} = 1 \\ 0 & \text{otherwise.} \end{cases} \tag{31}$$

Since we assumed that the network is connected and that every person is reachable from every other person, then this transition probability matrix $\mathbf{P}$ describes a regular Markov chain.[13] This type of Markov chain has a number of nice properties. In this case, the most useful is that for all possible initial probability vectors (for any way of selecting the seeds), the draw probability in later waves converges to a unique probability vector, $\vec{\pi}^{(*)}$, which satisfies the following self-consistency equation (Häggström 2002):

$$\vec{\pi}^{(*)}\mathbf{P} = \vec{\pi}^{(*)}. \tag{32}$$

The only probability vector $\vec{\pi}^{(*)}$ where equation (32) holds (see Appendix C for proof) is

$$\vec{\pi}_j^{(*)} = \frac{d_j}{\sum_{i \in N} d_i} \tag{33}$$

So we can now see that even if the seeds are not drawn with probability proportional to degree, the selection of the nodes in the later waves more and more closely approximates probability proportional to degree. As $\vec{\pi}^{(t)} \to \vec{\pi}^{(*)}$, the selection process for the sample becomes closer and closer to desired. As the sample size gets larger, the

[13]In order for the Markov chain to be regular, it must be irreducible and aperiodic. Because the network consists of one connect component, we know that the chain is irreducible. Further, if the network contains one triangle, then there are paths of both odd and even length from each state back to itself. This condition ensures aperiodicity. Since the networks under study are large, and since social networks have a strong tendency toward transitivity, the existence of at least one triangle has probability approaching 1. So, since the chain is irreducible and aperiodic, it is regular (Häggström 2002).

people selected as desired make up a larger and larger part of the sample and so our estimate converges to the true population value.

8.2. *Convergence of Estimates for Arbitrary Choice of Seeds: Simulation Results*

Asymptotic convergence results are reassuring, but we will never be able to select an infinitely large sample of people. Therefore, it is important to consider the rate of convergence. Also, it is important to note that this convergence, due to the seeds not being drawn as desired, is different, and potentially more serious than, the previous problems of convergence that were caused by the use of ratio estimation.

Previously, some work has been done on the rate of convergence to $\vec{\pi}^{(*)}$ in the case of random walks on networks (Lovász 1993) and in Markov chain Monte Carlo methods[14] (Cowles and Carlin 1996). However, we are interested in the rates of convergence of our estimates, not the vector $\vec{\pi}^{(*)}$. Simulation using the algorithm in Appendix A will be used to explore rates of convergence.

To begin to explore convergence, we can consider a family of different ways to draw the seeds:

$$\Pr[NI(j)_{w=0} = 1] = \frac{(d_j)^{\alpha}}{\sum_{i \in N} (d_i)^{\alpha}}. \tag{34}$$

When $\alpha = 1$, the seeds are drawn with probability proportional to degree. However, it may be the case that the degree of the nodes has a smaller or larger effect on the chance that a node is drawn as a seed. For example, the nodes could be drawn with probability proportional to degree squared ($\alpha = 2$) or the nodes could be drawn independent of degree—that is, via simple random sampling ($\alpha = 0$).

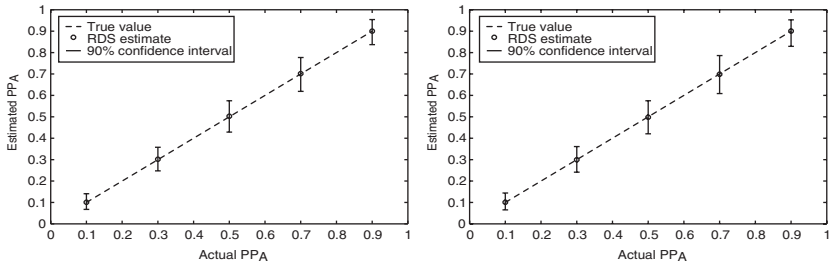
A natural first question to ask is, for a typical network structure such as the one described in Appendix A, and a reasonable sample size (say, 500 people): How close will our estimate be to the

[14]Readers familiar with Markov chain Monte Carlo (MCMC) methods will notice a similarity between our convergence argument and the method of selecting random draws from a difficult distribution. However, unlike in MCMC, we do not discard the data collected during the convergence process.

true population value even if the seeds are not drawn as assumed? For a wide range of population proportion values, we have presented simulation results when $\alpha = 0$ and $\alpha = 2$ in Figure 6. We see that for a sample size of only 500, the estimates are all within one-half of one percentage point (0.005) of the true population values. That is, for this network structure, our estimates are essentially unbiased when we selected a reasonable sample size, even though in both cases we violate the assumption that the seeds are drawn with probability proportional to degree.

Another natural question to ask is, how fast does our estimate converge? For this question we can offer only preliminary results. In Figure 7 we present simulation results for a population with 30 percent of its members in group A, and where the seeds are drawn with probability proportional to degree squared, $\alpha = 2$ (the other population parameters appear in Appendix A). Note that even for small sample sizes of fewer than 200 people, the average estimate has a small bias—a little more than two-tenths of one percentage point (0.002). As the sample size gets larger, the bias gets smaller. In this case, for a sample size of 500 the bias is about one-tenth of one percentage point (0.001), or well within rounding error and not a serious cause for concern.

The figures are provided as illustrations—the actual rates of convergence will depend on the characteristics of the population and



(a) Seeds selected with probability proportional to degree squared

(b) Seeds selected randomly from the entire population

FIGURE 6. Simulation results for a wide range of population proportion values. In (a) the seeds are drawn with probability proportional to degree squared ($\alpha = 2$). In (b) the seeds are drawn independent of degree ($\alpha = 0$). In both cases, the selection of the seeds violates our assumptions but has almost no effect on the estimates. The vertical lines show 90 percent confidence intervals.

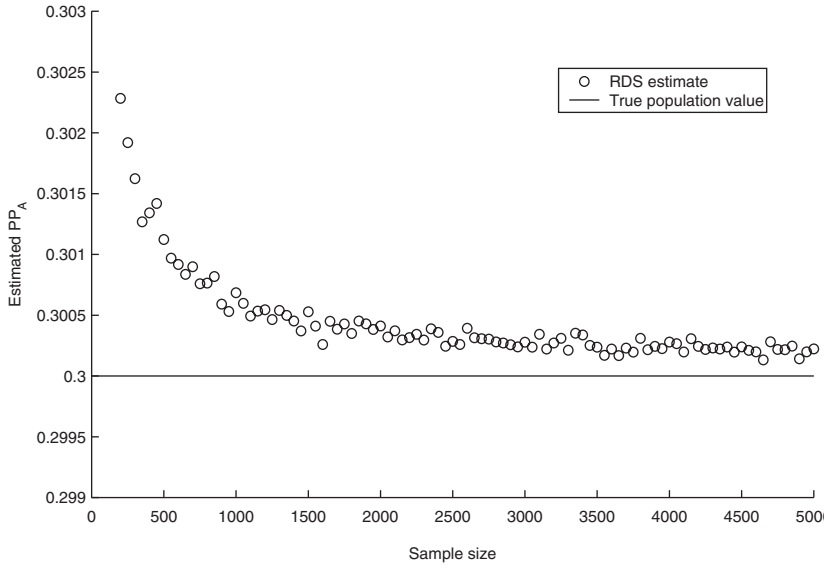


FIGURE 7. Numerical simulations showing that the respondent-driven sampling estimate converges to the true value even if the seeds are not drawn as desired. (In this case the seeds are drawn with probability proportional to degree squared.) The circles represent the average estimate from 100,000 samples. Note that, in this case, for all sample sizes greater than 200, the bias is less than three-tenths of 1 percent (0.003) and not a serious cause for concern.

the method of selecting the seeds. Further research in this area is called for.

9. A WARNING ABOUT IGNORING SAMPLE DESIGN

Up until now, in almost all research on hidden populations, researchers have collected a sample with some nonrandom sample design and then presented summary statistics about the sample they collected—for example, the proportion of the sample that is HIV positive. There is never an explicit claim that this sample can be generalized to the population, but that claim is often implicit. This procedure can lead researchers to draw very misleading conclusions (Thompson and Frank 2000; Thompson and Collins 2002).

We can see this problem more clearly if we look at some real data collected from jazz musicians in San Francisco and New York. A major problem for arts organizations is that it is often difficult to get accurate information about the communities they seek to represent and support (Throsby and Mills 1989). In the case of jazz musicians, there is no sampling frame from which one can draw a sample.

To resolve this situation, researchers took a sample of jazz musicians via respondent-driven sampling in several American cities (Heckathorn and Jeffri 2001). A random sample from the American Federation of Musicians was also taken and screened for jazz musicians. With this data it is possible to compare the estimates from two methods currently in use—institutional sampling and chain-referral sampling ignoring the sample design—with the estimates from respondent-driven sampling.

Some of the data from this study are presented in Table 1. It is important to note the differences between the estimates depending on the data source and the method of estimation. These differences are not merely of academic concern. Results from studies of hidden populations are often used to set public policy. If these policies are set based on incorrect information, they could waste valuable resources or even have effects counter to the goal of the policy.

In San Francisco, using the union sample we would conclude that 81.6 percent of the city's jazz musicians have had their work played on the radio. However, the respondent-driven sampling estimate, which, based on the arguments presented in this paper we believe to be accurate, is that only 31.4 percent have had their work played on the radio. This same pattern, overrepresentation of successful musicians in the union sample, is also observed in New York, but to a smaller extent. This overrepresentation of successful musicians in the union could have been predicted ahead of time, and in some sense serves as a reality check for the respondent-driven sampling procedure. However, even though the direction of the bias could have been predicted ahead of time, it would have been hard to guess its magnitude.

In other situations, the relationship between the estimate from the union sample and respondent-driven sampling estimate is quite difficult to guess ahead of time. For example, in San Francisco the union sample overestimates the percentage of females by about 10 percentage points. But in New York we observe the opposite; females

TABLE 1
Three Different Estimates of the Characteristics of Jazz Musician Populations

Jazz musicians in New York			
Characteristic	Union Sample ($n = 415$)	Chain-Referral Sample ($n = 251$)	RDS Estimate ($n = 251$)
Union member	100	39.6	25.4
Female	16.1	26.8	23.7
Solo only	19.9	8.8	11.6
Received airplay	79.6	81.6	74.9
Jazz musicians in San Francisco			
Characteristic	Union Sample ($n = 237$)	Chain-Referral Sample ($n = 221$)	RDS Estimate ($n = 221$)
Union member	100	11.2	4.0
Female	21.9	14.5	11.1
Solo only	15.0	9.2	21.0
Received airplay	81.6	48.7	31.4

Note: The data reflect different conclusions researchers would make depending on the source of data used. The estimates from the union sample and chain-referral sample are sample means. The respondent-driven sampling (RDS) estimates use the data from the chain-referral sample and the estimation techniques presented in this paper.

are underestimated in the union sample by almost 7 percentage points. The magnitude—or even direction—of this bias could not have been guessed at ahead of time. Because of these types of problems, the sample proportion from institutional samples should be viewed as a bad estimate of the true population proportion.

Another technique currently used by researchers of hidden populations is to select a sample via some nonrandom sample design and then to ignore the sample design and treat the sample as if it were a simple random sample. In these cases, the proportion of the sample with a given characteristic (for example, HIV positive) is used as an estimate of the proportion of the population with that characteristic. By comparing this naive estimate, which ignores the sample design, with the respondent-driven sampling estimate, which correctly accounts for the sample design, we can see that ignoring the sample design can produce misleading estimates.

As one might expect, the sample proportion from a chain-referral sample overestimates those with many contacts in the population—for example, union members. In New York, the naive estimate from the chain-referral sample estimates that 39.6 percent of jazz musicians are in the union. However, when we properly account for the method used to collect the sample, we estimate that only 25.4 percent are union members. We observe a similar pattern, where the naive chain-referral estimate (11.2 percent) is larger than the respondent-driven sampling estimate (4.0 percent), in San Francisco.

The sample proportion from the chain-referral sample also underestimates those with fewer relationships in the population—for example, musicians who perform solo. In San Francisco, the chain-referral estimate is that only 9.2 percent of jazz musicians perform solo. The respondent-driven sampling estimate indicates that this was an underestimate, and estimates the actual value to be 21.0 percent. A similar bias, but of different magnitude, is observed in the New York data.

We can draw a number of conclusions from this jazz musician data. Most importantly, failure to successfully consider the sampling design can lead researchers to make misleading conclusions. The magnitudes, and sometimes even the directions, of these biases are often difficult to predict ahead of time.

10. CONCLUSIONS

For many years researchers have known that chain-referral samples are an extremely useful way to collect samples from hidden populations. However, they have also considered chain-referral samples mere convenience samples, so hopelessly full of bias as to be useful only for exploratory purposes. However, we have shown here that when handled properly, chain-referral samples can produce estimates that are asymptotically unbiased.

We have also shown that our estimation method is robust to the actual selection mechanism for the seeds, even though this is traditionally considered to be one of the largest problems with chain-referral methods.

There are several other nice properties of respondent-driven sampling. First, the sample gives us information about not just the

people in the population but also the network connecting them. Currently, researchers are beginning to explore ways of extracting useful social network information from the sample (Heckathorn and Jeffri 2001). Given the role that networks play in the transmission of disease, this information could prove extremely useful in public health studies.

Another desirable property of respondent-driven sampling is that sample data can be combined with institutional data to estimate the size of a hidden population (Heckathorn and Jeffri 2001). Previous methods for estimating the sizes of hidden populations did not also allow for unbiased estimates of population composition.

Respondent-driven sampling is also cheaper, quicker, and easier to implement than other methods commonly used to study hidden populations (Semaan, Lauby, and Liebman 2002). This is a significant advantage because it means that for a given amount of resources, respondent-driven sampling allows researchers to have more study sites or larger sample sizes than other methods.

There are also still many open questions. In order to produce analytic results about the properties of the respondent-driven sampling estimators we had to make assumptions that in some cases may be violated. For example, nonrandom recruitment of friends could influence the estimates in unknown ways. Additionally, differential recruitment success by different types of people could bias the sample of edges that we observe. Nonsampling error related to estimating subject degree could also introduce bias into the estimates. Empirically checking the reasonableness of the assumptions and further research related to the robustness of the estimation procedure are both problems worthy of further study.

We hope this work opens up new possibilities for researchers interested in carrying out substantive research on hidden populations. We also hope that this will spur further work not just about respondent-driven sampling but chain-referral methods in general. These methods have great promise for answering a number of important questions.

APPENDIX A: ALGORITHMS USED FROM COMPUTER SIMULATION

It is possible to simulate the respondent-driven sampling procedure in order to provide support for the arguments offered in this paper. The simulation model can also be used to explore questions that are not analytically tractable.

There are five different population parameters that we vary in the simulations. We can vary the number of people in the population, N, and the proportion of the population in group A, PP_A. We can also vary the super-population degree distribution, $\Gamma(d)$, from which the node degrees are drawn. That is, for a given degree d, $\Gamma_A(d)$ is the probability that a node in group A will be assigned degree d. When the number of nodes is large, the difference between the super-population distribution, $\Gamma(d)$, and the population degree distribution, $p(d)$, becomes small. There is some evidence that degree distributions of friendship networks are right skewed (Killworth et al. 1998a,b; McCarty et al. 2001), so unless otherwise stated we will draw from an exponentially distributed super-population. However, the mean for group A and group B can be different.

Finally, we can vary the interconnectedness of the two groups. In some situations the friendship structure may be highly homophilous (McPherson, Smith-Lovin, and Cook 2001)—that is, people in group A are more likely than chance to be connected to people in group A. We formalize this with a parameter I, which represents the interconnectedness of the two groups. I is defined as the ratio of the actual number of cross-group ties to the possible number of cross-group ties, where the number of possible cross-group ties is limited by the $min(R_A,R_B)$. That is,

$$I = \frac{T_{AB}}{min(R_A, R_B)}. \tag{A.1}$$

To actually construct the network, we modify the algorithm presented in Malloy and Reed (1995) and Newman, Strogatz, and Watts (2001). First, the nodes are created and assigned to each group. Then for each node i, we draw a random number d_i from $\Gamma_A(d)$ if $i \in A$ or from

$\Gamma_B(d)$ if $i \in B$. We then assign d_i "stubs" to node i.[15] The stubs represent the friendships to be connected with other nodes. Then each d_i is increased by 1 to avoid any members of the population having degree 0.

Once the stubs are set, the required number of cross-group ties is calculated, $T_{A,B} = I \cdot min(R_A, R_B)$. Then, a randomly chosen stub from group A is paired with a randomly chosen stub from group B until the desired number of cross-group edges has been added. The remaining stubs are meant for within-group edges, so they are randomly connected to other stubs in their group. There is a check before each edge is added that prevents self-loops and multiple edges between the same two nodes.

Occasionally, the graph generation procedure can get "stuck". That is, there are stubs remaining but no way to connect them without adding self-loops or multiple edges between the same nodes. In that case, the algorithm removes all the edges and begins adding them again from the beginning. The algorithm described here ensures that we are randomly sampling from the ensemble of all graphs with the specified parameters.

Once the network is complete, we can simulate the sampling process. First, a set of s seeds is chosen to make up wave 0. Then we randomly sample from the $s \cdot c$ coupons that are eligible to recruit people for wave 1.

Once a coupon is selected, we choose a friend of the person who holds that coupon. This new recruitee is now a member of wave 1 of the sample. We continue choosing randomly from the set of coupons of people in wave 0 until all the people from wave 0 have used all their c coupons. At this point there are $s \cdot c \cdot c$ coupons to be used to recruit wave 2. We then begin randomly sampling from the set of coupons to be used to recruit wave 2. This process continues until the desired sample size is reached.

In Section 6.1, we assumed that people were selected with replacement. So in the simulation, each time a person is selected they are considered a new member of the sample. In this way the composition of the population does not change as the sampling procedure progresses.

Unless otherwise shown, all simulations are conducted with the following population and sampling parameters:

[15]All random numbers were generated using routines from *Numerical Recipes* (Press et al. 1992).

- Population size (N): 10000
- Proportion of the population in group A (PP_A): 0.3
- Super-population degree distribution for group A ($\Gamma_A(d)$): exponential with mean 20
- Super-population degree distribution for group B ($\Gamma_B(d)$): exponential with mean 10
- Interconnectedness (I): 0.6
- Sample size (n): 500
- Replicate samples generated for each population (r): 1000
- Seeds (s): 5
- Coupons (c): 2

APPENDIX B: DERIVATION OF $\widehat{D_A^{dist}}$

In the text we made the claim that the estimator $\widehat{D_A^{dist}}$ takes on a more convenient form when it is rewritten indexed by sample element. We will show that derivation here.

First, we begin with the equation

$$\widehat{D_A^{dist}} = \sum_{d=1}^{max(d)} d \cdot \widehat{p_A(d)}. \tag{B.1}$$

By plugging in the definition of $\widehat{p_A(d)}$ we get,

$$\widehat{D_A^{dist}} = \sum_{d=1}^{max(d)} \frac{q_A(d)}{\sum_{d=1}^{max(d)} \frac{1}{d} \cdot q_A(d)} \tag{B.2}$$

Since the denominator is constant, we can pull it outside of the summation and write

$$\widehat{D_A^{dist}} = \frac{1}{\sum_{d=1}^{max(d)} \frac{1}{d} \cdot q_A(d)} \cdot \sum_{d=1}^{max(d)} q_A(d). \tag{B.3}$$

Because $q_A(d)$ is a probability distribution, we know that $\sum_{d=1}^{max(d)} q_A(d) = 1$. Therefore we can rewrite

$$\widehat{D_A^{dist}} = \frac{1}{\sum_{d=1}^{max(d)} \frac{1}{d} \cdot q_A(d)}. \tag{B.4}$$

Now let's consider the frequency distribution, $f_A(d)$, which records the number of people in group A of degree d observed in the sample. It is the case that $\frac{f_A(d)}{n_A} = q_A(d)$. We can now rewrite in terms of $f_A(d)$:

$$\widehat{D_A^{dist}} = \frac{1}{\sum_{d=1}^{max(d)} \frac{1}{d} \cdot \frac{f_A(d)}{n_A}}. \tag{B.5}$$

Since it is the case that $\frac{1}{d} \cdot f_A(d) = \sum_{i \in M} \frac{1}{d}$ where $M = \{i | d_i = d\}$, we can substitute and rewrite the sum indexed by sample element i. Moving n_A to the numerator we get

$$\widehat{D_A^{dist}} = \frac{n_A}{\sum_{i=1}^{n_A} \frac{1}{d}}. \tag{B.6}$$

This completes the derivation.

APPENDIX C: PROOF OF CONVERGENCE OF MARKOV CHAIN

In the text we argued that the vector $\vec{\pi}^{(t)}$ converges to a unique vector $\vec{\pi}^{(*)}$ such that,

$$\vec{\pi}^{(*)}\mathbf{P} = \vec{\pi}^{(*)} \tag{C.1}$$

We then argued that the only such vector, $\vec{\pi}^{(*)}$ that satisfies equation (C.1) is

$$\vec{\pi}_j^{(*)} = \frac{d_j}{\sum_{i \in N} d_i}. \tag{C.2}$$

We now rewrite equation (C.1) in expanded form

$$\begin{bmatrix} \pi_1^{(*)} & \pi_2^{(*)} & \cdots & \pi_N^{(*)} \end{bmatrix} \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1N} \\ p_{21} & p_{22} & \cdots & p_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ p_{N1} & p_{N2} & \cdots & p_{NN} \end{bmatrix} = \begin{bmatrix} \pi_1^{(*)} & \pi_2^{(*)} & \cdots & \pi_N^{(*)} \end{bmatrix}. \tag{C.3}$$

We can now check to see if it is the case that

$$\sum_{i=1}^{N} \vec{\pi}_i^{(*)} \cdot p_{ij} = \vec{\pi}_j^{(*)}. \tag{C.4}$$

To verify equation (C.4), we can begin by ignoring all terms in the sum where $p_{ij} = 0$. These are cases where nodes i and j are not friends, so there is no possibility of person i directly recruiting person j. We can then rewrite both sides of equation (C.4) using the definitions of the transition probability matrix $\mathbf{P}$ and the probability vector $\vec{\pi}_j^{(*)}$,

$$\sum_{\{i|x_{ij}=1\}} \frac{d_i}{\sum_{i \in N} d_i} \cdot \frac{1}{d_i} = \frac{d_j}{\sum_{i \in N} d_i}, \tag{C.5}$$

which simplifies to

$$\sum_{\{i|x_{ij}=1\}} \frac{1}{\sum_{i \in N} d_i} = \frac{d_j}{\sum_{i \in N} d_i}. \tag{C.6}$$

By recalling that the sum on the left side of equation (C.6) is over d_j terms, we can see that the equation does indeed hold.

So now we have shown that for any selection of the seeds, $\vec{\pi}^{(0)}$, the probability of a node being selected converges to a value which is proportional to its degree.

REFERENCES

Berg, S. 1988. "Snowball Sampling." Pages 528–532 in *Encyclopedia of Statistical Sciences*, vol. 8, edited by S. Kotz and N. L. Johnson. New York: Wiley.

Biernacki, P., and D. Waldorf, 1981. "Snowball Sampling: Problems and Techniques of Chain Referral Sampling." *Sociological Methods and Research* 10(2):141–63.

Brewer, K. R. W., and M. Hanif. 1983. *Sampling with Unequal Probability*. New York: Springer-Verlag.

Cochran, W. G. 1977. *Sampling Techniques*. 3d ed. New York: Wiley.

Coleman, J. S. 1958. "Relational Analysis: The Study of Social Organization with Survey Methods." *Human Organization* 17:28–36.

Cowles, M. K., and B. P. Carlin, 1996. "Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review." *Journal of the American Statistical Association* 91:883–904.

Dávid, B., and T. A. B. Snijders, 2002. "Estimating the Size of the Homeless Population in Budapest, Hungary." *Quality and Quantity* 36:291–303.

Dodds, P. S., R. Muhamad, and D. J. Watts. 2003. "An Experimental Study of Search in Global Social Networks." *Science* 301:827–29.

Eland-Goosensen, M., L. Van De Goor, E. Vollemans, V. Hendriks, and H. Garretsen. 1997. "Snowball Sampling Applied to Opiate Addicts Outside the Treatment System." *Addiction Research* 5(4):317–30.

Erickson, B. H. 1979. "Some Problems of Inference from Chain Data." Pp. 276–302 in *Sociologal Methodology*, Vol. 10, edited by Karl F. Schuessler. San Francisco, CA: Jossey-Bass Publishers.

Feld, S. L. 1991. "Why Your Friends Have More Friends Than You Do." *American Journal of Sociology* 96(6):1464–77.

Frank, O. 1979. "Estimation of Population Totals by Use of Snowball Samples." Pp. 319–47 in *Perspectives on Social Network Research*. edited by P. Holland and S. Leinhardt. New York: Academic Press.

———. 1981. "A Survey Graph Analysis." Pp. 110–155 in *Sociological Methodology*, Vol. 12, edited by Samuel Leinhardt. San Francisco, CA: Jossey-Bass Publishers.

Frank, O., and T. A. B. Snijders. 1994. "Estimating the Size of Hidden Populations Using Snowball Sampling." *Journal of Official Statistics*, 10(1):53–67.

Friedman, S. R. 1995. "Promising Social Network Research Results and Suggestions for a Research Agenda." Pp. 196–215 in *Social Networks, Drug Abuse, and HIV Transmission, NIDA Research Monograph Number 151*. Washington, DC: National Institute on Drug Abuse.

Goodman, L. 1961. "Snowball Sampling." *Annals of Mathematical Statistics* 32(1):148–70.

Hagan, J. 2001. *Northern Passage: American Vietnam War Resisters in Canada*. Cambridge, MA: Harvard University Press.

Häggström, O. 2002. *Finite Markov Chains and Algorithmic Applications*. Cambridge, England: Cambridge University Press.

Hansen, M. H., and W. N. Hurwitz. 1943. "On the Theory of Sampling from Finite Populations." *Annals of Mathematical Statistics* 14(4):333–62.

Heckathorn, D. D. 1997. "Respondent-Driven Sampling: A New Approach to the Study of Hidden Populations." *Social Problems* 44(2):174–99.

———. 2002. "Respondent-Driven Sampling II: Deriving Valid Population Estimates from Chain-Referral Samples of Hidden Populations." *Social Problems* 49(1):11–34.

Heckathorn, D. D., R. S. Broadhead, and B. Sergeyev. 2001. "A Methodology for Reducing Respondent Duplication and Impersonation in Samples of Hidden Populations." *Journal of Drug Issues* 31(2):543–64.

Heckathorn, D. D., and J. Jeffri. 2001. "Finding the Beat: Using Respondent-Driven Sampling to Study Jazz Musicians." *Poetics* 28:307–29.

———. 2003. "Social Networks of Jazz Musicians." Pp. 48–61 in *Changing the Beat: A Study of the Worklife of Jazz Musicians, Volume 3. Respondent-Driven Sampling: Survey Results by the Research Center for Arts and Culture*, Research Division Report 43. Washington, DC: National Endowment for the Arts.

Heckathorn, D. D., S. Semaan, R. S. Broadhead, and J. J. Hughes. 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.

Hogan, H. 1993. "The 1990 Post-enumeration Survey: Operations and Results." *Journal of the American Statistical Association* 88(423):1047–60.

Kalton, G. 1983. *Introduction to Survey Sampling*. Beverly Hills, CA: Sage.

Kalton, G., and D. W. Anderson. 1986. "Sampling Rare Populations." *Journal of the Royal Statistical Society,* Series A, 149:65–82.

Kanouse, D. E., S. H. Berry, N. Duan, J. Lever, S. Carson, J. F. Perlamn, and B. Levitan. 1999. "Drawing a Probability Sample of Female Street Prostitutes in Los Angeles County." *Journal of Sex Research* 36(1):45–51.

Killworth, P. D., E. C. Johnson, C. McCarty, G. A. Shelley, and H. R. Bernard. 1998a. "A Social Network Approach to Estimating Seroprevalence in the United States." *Social Networks* 20:23–50.

Killworth, P. D., C. McCarty, H. R. Bernard, G. A. Shelley, and E. C. Johnson. 1998b. "Estimation of Seroprevalence, Rape, and Homelessness in the United States Using a Social Network Approach." *Evaluation Review* 22(2):289–308.

Klovdahl, A. 1989. "Urban Social Networks: Some Methodological Problems and Possibilities." Pp. 176–210 in *The Small World*, edited by M. Kochen. Norword, NJ: Ablex Publishing.

Kral, A. H., R. N. Bluthenthal, R. E. Booth, and J. K. Watters. 1998. "HIV Seroprevalence Among Street-Recruited Injection Drug and Crack Cocaine Users in the 16 US Municipalities." *American Journal of Public Health* 88(1):108–13.

Lovász, L. 1993. "Random Walks on Graphs: A Survey." Pp. 1–46 in *Combinatorics, Paul Erdös Is Eighty*, Vol. 2, edited by D. Miklós, D. Sós, and T. Szöni. Budapest, Hungary: Uános Bolyai Mathematical Society.

Malloy, M., and B. Reed. 1995. "A Critical Point for Random Graphs with a Given Degree Sequence." *Random Structures and Algorithms* 6:161–79.

McCarty, C., P. D. Killworth, H. R. Bernard, E. C. Johnsen, and G. A. Shelely. 2001. "Comparing Two Methods for Estimating Network Size." *Human Organization* 60(1):28–39.

McGrady, G. A., C. Marrow, G. Myers, M. Daniels, M. Vera, C. Mueller, E. Liebow, A. Klovdahl, and R. Lovely. 1995. "A Note on Implementation of Random-Walk Design to Study Adolescent Social Networks." *Social Networks* 17:251–55.

McPherson, M., L. Smith-Lovin, and J. M. Cook. 2001. "Birds of a Feather: Homophily in Social Networks." *Annual Review of Sociology* 27:415–44.

Muhir, F. B., L. S. Lin, A. Stueve, R. L. Miller, W. L. Ford, W. D. Johnson, and P. J. Smith. 2001. "A Venue-Based Method for Sampling Hard-to-Reach Populations." *Public Health Reports* 116(suppl. 1):216–22.

Newman, M. E. J. 2003a. "Ego-Centered Networks and the Ripple Effect." *Social Networks*, 25:83–95.

———. 2003b. "The Structure and Function of Complex Networks." *SIAM Review* 45:167–256.

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

Press, W. H., S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. 1992. *Numerical Recipes in C: The Art of Scientific Computing*. 2d ed. Cambridge, England: Cambridge University Press.

Semaan, S., J. Lauby, and J. Liebman. 2002. "Street and Network Sampling in Evaluation Studies of HIV Risk-Reduction Interventions." *AIDS Reviews* 2002 4:213–23.

Sirken, M. G. 1970. "Household Surveys with Multiplicity." *Journal of the American Statistical Association* 65(329):257–66.

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.

Stueve, A., L. N. O'Donnell, R. Duran, A. S. Doval, and J. Blome. 2001. "Time-Space Sampling in Minority Communities: Results with Young Latino Men Who Have Sex with Men." *American Journal of Public Health*, 91(6):922–26.

Sudman, S., Sirken, M. G., and Cowan, C. D. 1988. "Sampling Rare and Elusive Populations." *Science* 240(4855):991–96.

Thompson, S. K. 1990. "Adaptive Cluster Sampling." *Journal of the American Statistical Association* 85(412):1050–59.

———. 2003. "Markov Chain Sampling Designs in Graphs." Technical Report 03-01, Pennsylvania State University, Deparment of Statistics, University Park, PA 16802.

Thompson, S. K., and L. M. Collins. 2002. "Adaptive Sampling in Research on Risk-Related Behaviors." *Drug and Alcohol Dependence* 68:S57–S67.

Thompson, S. K., and O. Frank. 2000. "Model-Based Estimation with Link-Tracing Sampling Designs." *Survey Methodology* 26(1):87–98.

Thompson, S. K., and G. A. F. Seber. 1996. *Adaptive Sampling*. New York: Wiley.

Throsby, D., and Mills, D. 1989. *When Are You Going to Get a Real Job? An Economic Study of Australian Artists*. Sydney, Australia: Australia Council.

Watters, J. K., and P. Biernacki. 1989. "Targeted Sampling: Options for the Study of Hidden Populations." *Social Problems* 36(4):416–30.

Watters, J. K., and Y.-T. Cheng. 1987. "HIV-1 Infection and Risk Factors Among Intravenous Drug Users in San Francisco: Preliminary Findings." *Contemporary Drug Problems* 14:397–410.

Watts, D. J. 1999. "Networks, Dynamics, and the Small World Phenomenon." *American Journal of Sociology* 105(2):493–527.

Watts, D. J., and S. H. Strogatz. 1998. "Collective Dynamics of 'small-world' networks." *Nature* 393:440–42.

Welch, S. 1975. "Sampling by Referral in a Dispersed Population." *Public Opinion Quarterly* 39(2):237–45.

World Health Organization. 2000. *Second-Generation Surveillance for HIV: The Next Decade*. WHO/CDS/CSR/EDC/2000.5. Geneva: World Health Organization and UNAIDS Working Group on Global HIV/AIDS and STI Surveillance.