Forthcoming – *Sociological Methods and Research*

WEB-BASED NETWORK SAMPLING: EFFICIENCY AND EFFICACY OF RESPONDENT-DRIVEN SAMPLING FOR ONLINE RESEARCH

Cyprian Wejnert

and

Douglas D. Heckathorn

Department of Sociology

Cornell University

May 18, 2007

WEB-BASED NETWORK SAMPLING: EFFICIENCY AND EFFICACY OF RESPONDENT-DRIVEN SAMPLING FOR ONLINE RESEARCH

Abstract

This study tests the feasibility, effectiveness, and efficiency of respondent-driven sampling (RDS) as a web-based sampling method. First, RDS referral chains were found to progress very fast. Studies with large samples can be expected to proceed up to 20 times faster than with traditional methods; in fact, to prevent bias from temporal filtering, studies with small target samples should limit hourly response rates. Second, because of the relative ease of the web-based RDS process for each recruit, compensation should remain low. High rewards appear too good to be true and may scare off potential recruits. The results also demonstrate that web-based RDS can be used with minimal staffing. Our findings, based on theoretical and empirical evidence, indicate that web-based RDS can be much faster, easier, and cheaper than both regular RDS and standard sampling methods. Additionally, RDS estimates are compared to institutional data and potential sources of bias associated with the random recruitment assumption and sample size are discussed

Introduction

Traditionally, sampling hidden populations - populations for which there exists no sampling frame, or for which constructing a sampling frame would be infeasible because of the population's small size relative to the general population and the presence of either stigma or networks that are hard for outsiders to penetrate, such as injection drug users or jazz musicians (Heckathorn 1997) - has proven challenging to researchers interested in collecting probability samples.

A traditional approach to sampling these populations involves constructing a partial sampling frame by identifying non-hidden venues or institutions through which members of the target population can be found (Semaan et al. 2001). One such method, time-space sampling, employs a two stage sampling method. First, a probability sample of venues and times frequented by the target population is selected. Next researchers travel to selected venues at the appropriate times and interview a representative sample of those in attendance. Focusing on young, Latino, men who have sex with men, Stueve et al. (2001) find time-space sampling is effective at reaching hidden populations. Because of its easily reproducible sampling frame, the method is especially well suited for tracking changes within a population across time, provided the venues with which the population is associated remain relatively stable. However, time-space sampling is biased in favor of frequent venue attendees because members of the target population who do not attend public venues are excluded from the sampling frame and venue regulars have many more chances of being interviewed than occasional patrons. Furthermore, sampling that includes small venues can prove costly. Stueve et al. (2001, p.925) write, "Our experience also indicates that surveying in small nongay settings is

very costly… however, excluding such venues also omits a potentially important subgroup that may not be accessed elsewhere." Consequently, this method's coverage is limited to those who attend public venues.

Respondent-Driven Sampling (RDS), a new network-based (i.e. snowball-type) sampling method, has been proposed as a way to sample hidden populations that overcomes the venue bias associated with time-space sampling (Heckathorn 1997). Network-based designs, which were originally introduced for the study of social networks by Coleman (1958), start with a modest number of initial respondents, or *seeds,* who provide researchers with information on their network connections; these connections then form the pool from which the second wave of respondents is drawn and so on. In RDS, however, respondents are asked to recruit peers directly, allowing referral chains to efficiently and safely penetrate social regions only accessible to insiders. Traditionally, the non-randomness of social network connections has led such samples to be viewed as convenience samples from which unbiased estimation is not possible (Berg 1988). RDS challenges this view by using data gathered during the sampling process to make inferences about social network structure, which is then used to calculate unbiased population estimates (Salganik and Heckathorn 2004).

This paper extends RDS to online research and discusses results from a web-based RDS (WebRDS) feasibility study of university students. By moving the interview location and recruitment coupon to the internet, WebRDS has the potential to sample large electronically connected populations quickly with minimal resources and staff. We first summarize the basic theory behind RDS and describe its implementation as a method of web-based sampling. Next we present our results, including a detailed

discussion of RDS assumptions and possible sources of bias related to them, comparisons to institutional data, and sampling speed. We conclude with a discussion of potential limitations and advantages of WebRDS.

Respondent-Driven Sampling

RDS consists of an enhancement of network or "snowball" sampling, in which data on who recruited whom and the extensiveness of network connections provide the basis for calculation of relative inclusion probabilities, population indicators of minimal bias, and the variability of these indicators (Heckathorn 1997, 2002, forthcoming, Salganik and Heckathorn 2004, Volz and Heckathorn forthcoming). This method is now widely employed in public health to sample drug users (Heckathorn 1997), men who have sex with men (Ramirez-Valles et al. 2005a), and sex workers (Johnston et al. 2006); it has also been used in studies of arts and culture (Heckathorn and Jeffri 2001).

RDS theory is based on two observations (Heckathorn 2002). First, if referral chains are sufficiently long—that is, if the chain-referral process consists of enough cycles of recruitment, or *waves,*—the composition of the final sample with respect to critical characteristics and behaviors will become independent of the seeds from which it began. After a certain number of waves, the sample composition becomes stable and all members of the target population have a non-zero probably of selection that is independent of seed composition (Heckathorn 2002, Salganik and Heckathorn 2004). Therefore, an important design element in RDS involves methods of increasing the length of referral chains (Heckathorn 1997).

[Figure 1 here]

While the final RDS sample composition is independent of seed composition, choice of seeds can affect the rate at which equilibrium is reached and the speed with which sampling will occur (Heckathorn 2002). Figure 1 illustrates recruitment chains by all productive seeds from the sample. It includes a single long chain that makes up over 70% of the data and five smaller chains that make up the remaining 30%. Recruitment in RDS is often dominated by large recruitment chains initiated by respondents who can be termed "super seeds" (e.g. Heckathorn 1997, Heckathorn et al. 1999, Heckathorn et al. 2002, Ramirez-Valles et al. 2005a). This need not reflect special characteristics of the individual, because it results from a positive feedback process in which, the larger a recruitment chain grows, the greater is the number of respondents working to make it grow even larger, so a "rich-get-richer" dynamic is produced in which the larger chains grow ever more quickly than the smaller chains. Hence, any reasonably productive seed stands a good chance of becoming a "super seed."

The rapidity with which sampling goals can be attained depends on the ability to begin with seeds who will be productive because only a seed who recruits can initiate a recruitment chain, and the more recruits a seed produces, the more quickly the chain can grow. According to Heckathorn and Magnani (forthcoming), seeds should be well-motivated and enthusiastic, and hence willing to try to recruit their peers; and they should also be sociometric stars, individuals whose high regard among their peers enables them to recruit their peers, while also instilling in them motivation to continue the peer recruitment process. In summary, a major advantage of RDS is that it provides population estimates starting from a small convenience sample of seeds. However, seeds with certain characteristics have been found more suitable than others by facilitating

faster production of diverse recruitment chains. The following are recommendations regarding selection of seeds: First, seeds should be diverse with respect to the factors that most strongly determine the formation of social ties within the population. Typically these are basic demographic features such as race, ethnicity, religion, caste, social status, and age. Second, because many social ties are formed based on propinquity such as living on the same street or working in the same firm, seeds should be drawn a variety of geographic areas occupied by the target population. Finally, seeds should be high-energy sociometric stars that are committed to the goals of the study. These are individuals who maintain many ties and are highly regarded within the target population. Such individuals can more easily promote participation and accelerate recruitment. Paradoxically, careful seed selection speeds the growth of recruitment chains, and thereby accelerates the point at which seed selection becomes irrelevant and thus helps reduce bias (Ramirez-Valles 2005a).

The second observation upon which RDS is based is that information gathered during the sampling process can provide the means for calculating relative inclusion probabilities. These in turn provide the means for both calculating unbiased population estimates specifying the variability of those indicators.

In traditional sampling methods, such as simple random sampling or stratified sampling, the sampling frame is constructed before the first respondent is selected. In a simple random sample, selection probabilities are equal; in a stratified sample, subgroups of special interest are oversampled and thus selection probabilities are unequal. In either case, the sample is prestratified because selection probabilities are determined before the first respondent is selected. The effects of stratification are then taken into account when

data are analyzed by using sampling weights that are equal for simple random samples and unequal for stratified samples (Ramirez-Valles et al. 2005a).

In contrast, in RDS, the sampling frame is created after sampling is complete, based on two pieces of information gathered during the sampling process (Heckathorn 1997, 2002, Salganik and Heckathorn 2004, Heckathorn forthcoming). First, each recruiter-recruit dyad is documented. This provides the basis for controlling for bias introduced by the tendency of individuals to form social ties in a non-random way. Information regarding who recruited whom is used to quantify and account for sample bias due to non-random network structure, providing the basis for the "recruitment component" of the RDS estimation weight introduced in a recent article by Heckathorn (forthcoming). Second, respondents are asked how many other members of the target population they know. In a network-based sample the inclusion probability of an individual is proportional to the number of people in the target population he or she is connected to, termed his or her *degree*. Salganik and Heckathorn (2004) show that once a sample reaches equilibrium all ties within the target population have equal probability of being used for recruitment. Consequently, an individual with 50 ties is 5 times more likely to be recruited than an individual with only 10 ties. Therefore the RDS sample may be biased toward high degree individuals. Salganik and Heckathorn (2004 p. 214-218) derive an average group degree estimator that is the ratio of two Hansen-Hurwitz estimators, which are known to be unbiased (Brewer and Hanif 1983). 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, Salganik and Heckathorn 2004). This estimator is then used to correct for degree bias in RDS estimation of categorical variables. Heckathorn

(forthcoming) removes the limitation to only categorical variables by providing a degree estimation procedure for continuous variables.

The degree estimator is combined with recruitment data to calculate an RDS estimator for proportional group size, $\widehat{P_X}$ (Salganik and Heckathorn 2004 p.218, see also Heckathorn 2002):

$$\widehat{P_X} = \frac{\widehat{S_{YX}}\widehat{D_Y}}{\widehat{S_{YX}}\widehat{D_Y} + \widehat{S_{XY}}\widehat{D_X}},$$

where S_{XY} is the proportion of recruitments from Group X to Group Y and D_X is the estimated average degree of Group X. Relative inclusion probabilities, in the form of sampling weights, are then calculated as a function of the RDS estimator and sample composition (Heckathorn forthcoming) as follows:

$$W_X = \frac{\widehat{P_X}}{C_X},$$

where W_X is the population weight for Group X, C_X is the sample proportion of Group X, and $\widehat{P_X}$ is as defined above.

The original proof that the RDS estimator is asymptotically unbiased depends on a set of six assumptions (Salganik and Heckathorn 2004). This number is reduced to five assumptions in a subsequent proof by Heckathorn (forthcoming).

1) Respondents maintain reciprocal relationships with individuals who they know to be members of the target population.
2) Respondents are all linked into a single component in the network.
3) Sampling is with replacement.

4) Respondents can accurately report their personal network size or equivalently, their degree.

5) Peer recruitment is a random selection of the recruiter's peers.

The first three assumptions specify the conditions that must be met for RDS to be an appropriate sampling method for a population. First, in order for recruitment to occur, respondents must have access to other members of the population and be able to identify which of their peers qualify for recruitment. Consequently, RDS would not be effective at sampling tax evaders or men who secretly cheat on their wives (Heckathorn forthcoming). RDS is best suited for populations structured around social interaction. In such populations, interaction with co-members provides the means for maintaining membership. For example, the relationships developed among jazz musicians when they play with different groups provide means for finding gigs and creating original music (Heckathorn and Jeffri 2003). Similarly, injection drug users rely heavily on information from other users that helps them acquire drugs and protects them from police (Abdul-Quader et al. 2006). In addition, RDS estimates are based on a network structure in which ties are reciprocal (Heckathorn 2002 p.22, Heckathorn and Salganik 2004 p.202). Formally, if Respondent A recruits Respondent B, then it is assumed that there must be a non-zero probability that Respondent B could have recruited Respondent A. Consequently, the RDS research design includes means for encouraging subjects to recruit their acquaintances or friends rather than strangers by rewarding successful recruiters and making recruitment rights scarce through quotas (Heckathorn 1997). Under these conditions, respondents have been hesitant to waste valuable recruitment rights on strangers (Heckathorn 1997, 2002). Second, the population is assumed to form a single

component (Salganik and Heckathorn 2004 p.210, see also Heckathorn forthcoming). In other words, all of the target population must be reachable from any single respondent. In a random network, a single component forms when individual degrees are large compared to the natural log of the population size (Bollobias 1985, see also Watts and Strogatz 1998). When respondents are allowed to recruit not merely those with whom they have a special relationship (e.g., musicians who perform in the same ensemble or drug users who share drugs), but also any friends and acquaintances they know as members of the target population; individual degrees are larger than that generally required for a network to form a single large component (Heckathorn forthcoming). Additionally, since actual social networks are never wholly random, there must not exist any social or structural barrier that completely segregates one subgroup of the population from the rest. For example, RDS can not be used to sample across castes in a culture where cross-caste interaction is forbidden. Similarly, a single seed can not be used to sample drug users in Atlanta and Seattle. Third, sampling is assumed to occur with replacement so that recruitments do not deplete the set of available recruits (Salganik and Heckathorn 2004 p. 209). Consequently, the sampling fraction should remain small enough for such a model to be appropriate (Heckathorn forthcoming).

The fourth assumption of RDS is that respondents can provide accurate information on their personal network size. However, since the RDS estimator is based on network size relative to other network sizes it is not affected by uniform bias in responses (Salganik and Heckathorn 2004, Heckathorn forthcoming). Additionally, studies of network indicator reliability suggest RDS style indicators are among the more reliable (Marsden 1990).

The fifth assumption for RDS analysis is that recruitment patterns reflect personal network composition within the target population. In other words, RDS assumes respondents recruit as though they were selecting randomly from their personal networks (Heckathorn 2002). Heckathorn (forthcoming) argues that the plausibility of the random recruitment assumption can be enhanced with careful research design. For example, interview locations should be in areas considered neutral to the population, transportation should be provided for distant respondents, and incentives should appeal to the entire population. Additionally, the network size question should be tailored to fit the target population. A detailed discussion of this assumption is presented below.

One important advantage of using social networks to recruit respondents is that just as rumors and other forms of information can expand at an exponential rate through the population, so too, at least potentially, can recruitment. Of course, in order for participation to occur a potential respondent must choose to participate and then actually complete the study. RDS harnesses peer pressure to increase participation by offering respondents rewards for recruiting their peers. To earn maximum compensation, a participant needs to make sure that her recruits do in fact participate. As the sample expands through the target population's peer networks, respondents who initially decline participation may eventually agree to participate as additional peers endorse the project through their recruitment efforts. In this way, what may be termed a "norm of participation" arises and recruitment is enhanced.

Despite the potential for exponential growth, RDS sampling is limited by staff size relative to the length of the interview. The original RDS studies (Heckathorn 1997, Heckathorn et al. 1999) used face-to-face interviews administered in a storefront

accessible to the target population. However, participation required making an appointment for the two-and-one-half-hour interview, there were only two interviewers and storefront hours were limited to 20 per week. Consequently, sampling proceeded rather slowly, producing 100 respondents in one year. Subsequent RDS studies with briefer interviews and more interviewers produced samples of this size in a matter of weeks. In these studies, the sample growth rate is consistently bounded by the number of respondents a site can process in one day—that is, the number of research staff and the length of the interview, and not by the number of subjects interested in participation. A recent study of drug users in New York City compiled a sample of 618 respondents for 30-minute interviews in just 13 weeks using standard RDS methods (Abdul-Quader et al. 2006).

Table 1 shows a comparison of six standard RDS sampling rates. Note that interview length, which is directly correlated with the number of interviews a site can process in one day, is a strong predictor of weekly recruitment rate. In standard RDS, sampling speed is limited by the number of interviewers available at each interview site and amount of time each interviewer spends with a respondent, i.e. the length of interview. WebRDS removes these constraints by allowing all respondents to take a survey immediately, regardless of how many respondents are simultaneously being processed. Consequently, it provides the means by which the potential for exponential growth of the sample can be realized.

[Table 1 here]

Web-Based RDS Implementation

RDS procedures begin with the selection of seeds. The present study used nine seeds; four were selected as part of a demographically diverse group, taking into account gender, college within the university, and fraternity or sorority membership. The remaining five seeds were selected from a trial sample, which was lost because of a software error (see below). These five seeds contacted the administrator when attempts to complete the trial survey failed and thus were assumed to be motivated recruiters. At the time of recruitment, no information, besides e-mail address and desire to participate, was known about these seeds. Thus, approximately half the seeds were selected for diversity across gender and college within the university and half were selected because they were thought to be highly motivated participants. Seventy-four percent of the data originates from a single seed. This "super seed" was a white female Hotel School student, and had not only the largest network of any other seed, but her degree was one and a half times larger (150) than the next highest degree seed (100). This pattern is consistent with other RDS studies, in which the productivity of a seeds is positively related to its degree.

Once the seeds were informed about the project and had agreed to participate, they received a recruitment e-mail officially informing them of the project's purpose, compensation, and recruitment process. The recruitment e-mail (Appendix 1) contained an overview of the project, the serial number used to track recruitment, a consent form, and a link to the online survey. The serial number from the seed's recruitment e-mail was automatically entered into the survey, in the manner of an auto login to a secure website. After confirming student status, the seed was administered a questionnaire.

Questionnaire completion resulted in three automated actions by the software.

First, the data were downloaded into both a master data set and multiple backup forms. Second, both the serial number and the respondent's network ID were blocked from being used on the survey again, either in combination or separately, to prevent repeat survey respondents. Finally, three new recruitment e-mails, each with a unique serial number, were sent to the seed. The seed was then asked to forward each of these e-mails to one potential recruit who met the inclusion criteria. Because these e-mails contained serial numbers, only one respondent could be recruited by each e-mail. The recruits followed the same process as the seeds. The only difference between seeds and recruits was that seeds were recruited by the administrator (and therefore had no recruiter) but the recruits—many of whom became recruiters themselves—were recruited by other respondents (see Appendix 2 for algorithm).

The number of recruitment e-mails given to each participant was determined following RDS principles and previous RDS studies, which have found that limiting the number of coupons facilitates the lengthening of recruitment chains. The number of recruitments is the target sample size minus the number of seeds (who are recruited by researchers rather than by peers). For this study, the target sample was 150. If, for example, eight coupons had been given to each respondent, the target would have been attained only after a few waves. If only one coupon had been given to each respondent, the recruitment chains might have died out because not all coupons are used. Typically, three coupons are sufficient to support the development of robust chains while permitting the growth of long chains (Heckathorn 1997).

The serial numbers sent to each respondent were recorded in a coupon manager program. Because the sampling procedure was entirely automated, respondents could not

be screened using normal face-to-face methods. Instead, a series of internal checks was embedded in the instrument to prevent self-selection and selection from outside the population.

Additional checks included in the survey were intended to improve the quality of data. Quantitative questions were required to have numerical answers. Qualitative questions required nonnumerical answers. All questions had to be answered. Respondents who did not comply were informed of the nature of the problem and given the opportunity to correct the entry before moving to the next question.

Results

The target sample was 150 peer recruitments. With the nine seeds, the sample size was 159. All the respondents except one were undergraduates, the exception being a first-year graduate student. Each answered two test questions ("what year is it?" and "who is the President?") correctly. Because the survey did not allow for it, there were no missing data.

Demographic information for the final sample is presented in Table 2.

[Table 2 here]

Feedback from participants suggests that WebRDS was adopted with relative ease. Many participants reported contacting recruits both before and after recruitment to ensure that their recruits completed the survey. Consistent with findings that RDS promotes a norm of participation, some respondents reported that the survey resulted in a brief "fad" among their peers. For the entire duration of the study, the administrator received no questions or comments from participants. A trial run of the survey failed

because of a software error; in the few hours the survey was down, the administrator received 10 messages reporting the problem, suggesting that students would have contacted the administrator had there been a problem in the final run.

Overall, 55.3% (n=88) of the 159 respondents recruited peers for the study. This is consistent with other RDS studies and the geometry of RDS recruitment networks. That is, if each recruiter has three recruits, then on average, only one-third of the respondents will have recruited. If each recruiter averages two recruits, then only half of the respondents will have recruited (Heckathorn 2002). Six of the 88 respondents who did recruit were seeds.

Testing Assumptions

As with all methodological analyses, we must consider the assumptions imposed by the method and evaluate the extent to which these assumptions have been met in the data. In this section we consider the assumptions required by RDS theory and test them using three variables: gender, race, and college within the university. Recall that when the network of the target population consists of a single large component, and equilibrium is attained, then links within the network are sampled randomly and all members of the population have no-zero probability of selection. Thus, before considering assumptions we must show that the sample reached equilibrium. This is done by simulating the number of waves required to reach equilibrium for each variable and then comparing it with the actual number of waves reached in the sample (See Heckathorn et al. 2002 Appendix for equilibrium and waves required calculations). Of our three test variables, "college within the university" has the highest simulated number of necessary waves,

nine. The longest chain in the sample has more than 18 waves, satisfying the equilibrium requirement. The standard RDS interpretation is that if equilibrium is reached within a single chain, then equilibrium is reached for the entire sample because all individuals have a non-zero probability of selection (Heckathorn 2002). Heimer (2005) questions whether equilibrium can be viewed as a sample level characteristic if multiple seeds are used, arguing that chains not in equilibrium are sampling with bias. Commenting on a two sample RDS study of Latino MSMs with similar "super seed" recruitment structures he writes, "it would be useful to compare each sample as a whole to a sub-sample that included only the participants recruited in the one long chain" (Heimer 2005 p. 405). Alternately, one could remove all respondents sampled before equilibrium is reached in their recruitment chain and perform analyses on only the subset sampled in equilibrium. While theoretically elegant, the variable specific nature of equilibrium makes this approach impractical. Because the number of waves required for equilibrium varies, estimates of each variable could be based on different sample sizes, making multivariate analysis extremely difficult. Similarly, basing estimates on only recruitments occurring after all variables have stabilized wastes much valuable data. For example, if only recruitments occurring after wave nine were included for estimation in this study, the sample size would drop from 159 to 69, resulting in estimates that do not reflect the full potential of the data. In this paper we compare estimates based on the large chain subset and full sample and discuss implications and limitations of each, but first we turn to the RDS assumptions.

The first RDS assumption is that recruiters have a preexisting relationship with their recruits, so this relationship is reciprocal (Salganik and Heckathorn 2004).

Specifically, if A is a member of B's pool of potential recruits, then B must also be a member of A's pool of potential recruits. While the instrument does not provide a direct test of this assumption, results suggest that in all cases recruits had a pre-existing social relationship to recruiters. Recruits in this study tended to be recruited by "friends" (50%, n = 75) and "close friends" (46.7%, n = 70). Only 3.3% (n = 5) of the sample was recruited by an "acquaintance," and no one reported being recruited by a "stranger". Furthermore, 56% (n = 84) reported interacting with their recruiter on a daily basis. Only 9.3% (n = 14) reported interacting with their recruiter less than once a week, and all respondents reported interacting with their recruiter at least once a month. All five respondents recruited by acquaintances reported interacting with their recruiter at least once a week. These findings suggest that in all cases recruiters and recruits had the reciprocal relationship required by RDS. Should a significant number of respondents (greater than 3% of the sample), however, report that their recruiter was a stranger, these recruitments should be removed from the data set before estimates are calculated. Additionally, recruiters must know their recruits as members of the target population. In this case, the target population was a non-hidden student body which forms a residential community largely closed off from the greater population. Therefore, identifying eligible recruits was not problematic.

The second assumption for RDS is that respondents are all linked by a single component. The mean sample degree was 66 known students; when the RDS degree estimator is used, the mean estimated degree is 40 known students per individual. In a population of approximately 13,000 students, this mean degree is sufficient to suggest most students can be reached through the network from any other student (Bollobias

1985, see also Watts and Strogatz 1998). Third, the sampling fraction, 159/13,000, is sufficiently small for a sampling with replacement approximation to be used. Fourth, respondents must be able to accurately report their network sizes. Research comparing self-report network size indicators has had limited success (Bell et al. 2007, McCarty et al. 2001). To avoid problems with self report network sizes, a *buddy list degree* measure was used. Buddy list degree is defined as the number of students the respondent has on her instant messenger buddy list (similar to an email address book). This number is stored and displayed by the software for each user and is therefore not subject to recall or self-report bias. At the time of sampling, instant messenger programs were the primary means of communication among students on campus. Many respondents reported contacting potential recruits with instant messenger to confirm interest in participation before recruiting them, an action that helps ensure reciprocity (Heckathorn 2002). Based on this measure, the mean degree was 66.29 buddies (S.E. = 43.05), with a minimum of 6 and maximum of 200 (at time of sampling instant messenger buddy lists were limited to 200 buddies).

Lastly, the fifth RDS assumption states that recruitment patterns reflect personal network composition, such that respondents recruit as though they were selecting randomly from their personal networks (Heckathorn 2002). One method that can be used to assess the extent to which unbiased recruitment occurs, is to ask respondents about the composition of their personal networks with respect to visible attributes, such as gender and race and compare these self-reports to the actual recruitment patterns (Heckathorn et al. 2002, Wang et al. 2005).

[Tables 3A & 3B here]

We asked respondents how many males, females, Asians, Whites, and "Others" they knew and, using the self-report degree as expected values, tested the likelihood that recruitment was not random across gender and race using a χ^2 goodness of fit test. Table 3A shows results for gender. On average, male students reported knowing approximately 41 males and 38 females. Women reported knowing approximately 31 males and 41 females. These averages are converted into probabilities and used to calculate expected recruitments such that the number of expected recruitments for males and females equals the number of recruitments actually made by male (n=96) and female (n=54) respondents respectively. The analysis of the race variable is done in a similar way. In both cases, the results suggest non-random recruitment. For gender, recruitment by males appears to have been heavily favored toward other males ($\chi_1^2 = 5.449, p = 0.020$). When race is examined, Asian students tended to recruit non-randomly, favoring Others over Whites ($\chi_4^2 = 9.462, p = 0.051$). Consequently, the random recruitment assumption does not appear to have been satisfied in this sample. This finding contrasts with previous studies (Heckathorn et al. 2002, Wang et al. 2005), in which a strong association was found between recruitment patterns and self-reported network composition. We will return to this issue below. It is important to note that these methods test the likelihood that recruitment was not random. The p-value, therefore, can not be interpreted as the probability that recruitment was random.

The self-report to recruitment χ^2 comparison is useful for visible traits, but is not well suited for the nonvisible or hidden characteristics that make up the majority of variables in most studies. One method of testing random recruitment could be to compare the number of cross group recruitments from group A to B to the recruitments from B to

A (Ramirez-Valles et al. 2005). Under the reciprocity model, these should be equal if recruitment is random and all groups recruit equally effectively, a condition that is not satisfied in most RDS studies (Heckathorn forthcoming). However, while comparing cross-recruitment does not require additional self report data, it does not provide enough information to fully test the random recruitment assumption. Specifically, differential ingroup recruitment, such as over recruiting of males by males, can not be explored through comparison of cross-recruitment counts. We leave development of a test of random recruitment that is both powerful and widely applicable open for future research.

Comparing RDS Estimates to Institutional Data

Table 4 shows the RDS estimates and sample proportions for the full sample (n=159) and the large component subset (n=117), as well as the true population parameters (Cornell University 2004) for race, gender, and college within the university. First, note that in all but three cases,[1] the full sample RDS estimate is a better approximation of the true proportion than the sample proportion. In one case, the Other category for race, the RDS estimate underestimates the true proportion. It is unclear why this is the case. One possibility is that, since both the institutional estimate and the RDS estimate are catch all categories, they measure different segments of the population. For example, the institutional data on race/ethnicity includes a "foreign national" category. This category, which makes up 7.2% of the student body, is included in Table 4 as "Other"; however, of these foreign nationals, many are likely to be Asian or White.

[Table 4 here]

In a second case, the sample proportion for students in the College of Agriculture

and Life Sciences, falls within 1% of the true population proportion, and the RDS full sample estimate falls within 4% of the true parameter. However, the 95% confidence interval for this category captures the true proportion easily, suggesting that the lack of improvement is due to an especially accurate sample proportion. Lastly, the gender variable, which is discussed in detail in the following sections, failed the random recruitment assumption (see above). Thus it is not surprising that the RDS estimate for gender is not accurate.

Estimates based on only the 19 wave large component display a similar pattern of results when compared to their sample proportions with two exceptions. Estimates for Hotel and Industrial Labor Relations students performed worse than the sample proportions. However, the differences are small and are likely due to reduced sample size in the subset data.

For each estimate, a 95% confidence interval, shown in Table 4, can be used as a test of validity. In eight of ten categories for both the full and large component samples, the 95% confidence interval captures the true population proportion. Furthermore, in five of these cases, the true population proportion falls within one standard deviation of the full sample RDS estimate. In the large component sample, three true proportions lie within one standard deviation of the estimate. Estimates for the two colleges discussed above are less precise but still captured the true proportions within a 95% confidence interval. Additionally, estimates based on the large component appear better for gender and worse for Asians while the corresponding full sample estimates are visa verse. These differences are likely the result of random variation within the sample as the true proportions fall near a confidence interval bound in both variables and remain so for the

full and the large component sample. While it is possible that analysis based solely on respondents in long recruitment chains could lead to different estimates, our data suggest that estimates are consistent across both methods. As a final note, it is interesting that while the large component's seed is a white, female, Hotel School student, the large component *under* sampled all three of these groups.

Web-Based RDS Sampling Speed

The target sample size was reached within 72 hours, with the final 50 respondents surveyed in four hours, much faster than with standard RDS sampling techniques. Several factors may account for this acceleration. First, since the entire process of being recruited, being interviewed, and then recruiting others can be conducted at the respondent's computer, the total turnaround time from being recruited to recruiting can be brief, in this study as low as 25 minutes per recruit. Standard RDS takes more time because each recruit must find time to come in for an interview and personally recruit new respondents, who then need to come in for interviews. Second, the university setting is ideal for online information transfer. 130 respondents contacted after the study reported checking their e-mail on average nine times per day (s.d. = 10.6), and 16 (12.3%) reportedly checked their e-mail continuously or more than 20 times per day. Finally, because the survey is entirely automated using a server with high bandwidth, there is no practical limit on how many surveys can be processed at once. The average time between recruitment and recruit survey submission was approximately four hours.

[Figure 2 here]

To quantify the sampling speed, we consider the number of recruits by both time

and wave of recruitment. Because each new recruit adds three recruitment e-mails to the system, we expect recruitment to grow exponentially as sampling progresses. If every member of a population were to behave identically with respect to recruitment and participation, we would expect this growth to be a function of both time and recruitment wave. However, because each respondent behaves differently, some recruitments take longer, and others may not occur at all. Figure 2 shows recruitments by wave for the WebRDS study. The best-fit curve, a linear function of wave, explains just over 5% of the variation ($R^2 = 0.0542$). The slope is pulled negative by the last two waves, which were cut short when our target sample size was reached. Removal of these waves produces a linear model with positive slope but no increase in explained variance ($R^2 = 0.048$). Two aspects of the sampling structure contribute to this pattern. First, exponential growth is expected by wave for each recruitment chain. Inherent in this assumption is that minor differences in recruitment rate at early stages of recruitment are magnified at later stages. Therefore, unless respondent behavior is uniform, any chains containing more productive respondents in early waves will expand much more rapidly, smothering the other chains as the target sample size is reached. Chains with less productive early respondents will be able to reach only a modest number of waves before the large chain exhausts the target sample size—that is, when growth of the recruitment chains is terminated. In early stages, multiple chains are recruiting, and that activity contributes to the overall number of recruits. In our sample, only one chain had more than six waves, corresponding to the dip observed in Figure 2 at wave 6, at which point all other chains have died out and only the recruitments from the large chain remain. Figure 3 shows recruitments by wave for only the large chain. Here an exponential model fits the data

well ($R^2 = .8617$).[2]

[Figure 3 here]

Although the recruitment rate by wave is sensitive to variation in respondent behavior, recruitment rate by time is simply a function of the number of active coupons circulating in the population, which is less sensitive to recruitment and behavioral variance. After the initial 24 hours, WebRDS processed one 20-minute survey every 13 minutes. The final 50 surveys were completed at three-minute intervals, suggesting that a much larger sample could easily be collected in one week. Figure 4 shows recruitment by time of day by day and total recruitment by day. Not surprisingly, recruitment varies by time of day: few recruitments occur between midnight and noon on each day, reflecting the schedule of university undergraduates. Additionally, the number of recruits more than doubles each successive day with 15 on Friday, 43 on Saturday, and 101 on Sunday. Consequently, a predicted sample size of more than 4,000 respondents could be recruited in one week. However, as the sample progresses, it includes an ever-increasing portion of the population, making nonsampled recruits more scarce and in turn slowing the recruitment process as it approaches saturation. For example, a sample size of 4,000 students would include more than 30% of the university's student enrollment (13,000 students). We conservatively estimate that 1,000 respondents from our target population could be sampled in one week. A larger target population, however, would not suffer from this slowing-down effect until much larger samples were reached.

[Figure 4 here]

Efficiency and Ease of Use

In addition to being a fast sampling method, WebRDS is also highly efficient. Respondents were compensated $10 for completing the study and $15 for each of their recruits who completed the survey, for a total possible compensation of $55. However, because each respondent completes one survey and has one recruiter, the actual cost to the researcher is $25 per respondent. We found this level of compensation to be too large and often interpreted as a potential scam, likely because it was confused with SPAM e-mails offering implausible sums of money for minimal effort.[3] We speculate that $5 for completing the survey and $10 for recruitment would have been just as effective, or perhaps more so, in soliciting respondents. Furthermore, approximately 20% of respondents did not bother to collect their compensation, suggesting that at least some participated for other reasons, such as being involved in the latest fad.

More importantly, whereas other sampling methods require at least one proctor to administer the survey and often one or more interviewers, the cost of WebRDS in person-hours is minimal. After the online survey has been set up and tested, the researcher need only identify and contact a modest number of seeds to begin sampling. Once sampling has started, no effort is needed on the part of the researcher except to download the completed data set. In traditional sampling methods, the researcher must identify every member of her sample using a predefined sampling frame and persuade each one to participate. RDS requires the identification and recruitment of only five to ten seeds; the identification and recruitment of the remaining sample is then done entirely by the respondents. Since the final RDS sample is independent of seeds once it reaches equilibrium, these five to ten seeds can be selected based on convenience instead of a

probabilistic sampling frame.

Random Recruitment and the Gender Bias

Two variables in the data seem problematic for RDS estimation. Both gender and race potentially fail the random recruitment assumption and both produce borderline estimates and confidence intervals. However, an inconsistency in racial classification between institutional and survey data suggests that the institutional estimates for Whites and Asians actually represent the estimate's lower bound, while the institutional estimates for Others represents an upper bound (see above). Furthermore, of the two, the gender variable is more consistently problematic than race estimation. Therefore, we focus on the gender variable as an illustration of limitations of WebRDS.

Above, the random recruitment assumption is tested by comparing recruitment to self report network composition. The plausibility of this hypothesis can be assessed by comparing the RDS population estimates with an alternative procedure in which self-reported network composition substitutes for transition probabilities (as in the χ^2 comparison above), and degree for each respondent is estimated as number of males plus the number of females known to them. With these adjustments, the estimated proportion of females increases from 37% (the original RDS estimate) to 47% (the self report estimate); a figure that closely approximates the institutional data showing 50% females (Cornell University 2004). This result suggests that the self reports represent the true network composition and that WebRDS recruitment was biased in favor of men. Specifically, males disproportionately recruited other males, whereas females recruited in a manner consistent with self-reported network composition. Current data is not sufficient

to provide an explanation for male recruitment behavior, but two theoretically distinct possibilities can be differentiated. One possibility is that males selectively recruited other males. A second possibility is that there was a gender difference in non-response bias such that male efforts to recruit females were successful less frequently than male efforts to recruit other males. In many RDS studies, respondents are asked about their unsuccessful recruitment efforts in a post-interview questionnaire. However, these questions were not included in the current study because its principal aim was to assess the feasibility of a web-based RDS study.

Several distinct mechanisms could have produced each of these types of bias from either selective recruitment or differential non-response by gender. First, selective recruitment of males by males could have reflected differences in e-mail usage. If usage among males is greater, males would have a distinct participation advantage over females. However, among a convenience subsample of the data[4], males reported checking their email on average 7.69 times per day (n=78) while women reported checking email 11.12 times per day (n=52), suggesting that if there was an email usage gender bias, it favored women. If differential email usage was problematic and well documented, unbiased estimates could be calculated by incorporating daily email usage into the RDS degree estimate.

Second, high incentives, especially for recruitment, may have caused recruitment to occur along a social *business* network, where recruits are not selected from the set of all peers, but from the set of all peers with which the recruiter would engage in small business transactions such as borrowing/ lending money or co-purchasing an expensive item. While a complete discussion of gender bias in business exchange is beyond the

scope of this paper, recruitment along business lines could bias males toward recruiting other males. For example, Van Emmerik (2006 p.25) finds men are more effective at creating *hard social capital,* or the accumulation of task-oriented resources that can be used to achieve valued outcomes, than are women (see also Ibarra 1993). Future studies can avoid this problem in one of two ways. Incentives can be held low so that respondents don't view their participation as a business transaction. Alternately, the degree question can be altered in such a way to solicit the number of peers with which the respondent would engage in small business transactions.

Third, the above-noted discrepancies between institutional data and the RDS estimates may result from a "temporal filtering effect." This is a source of bias that has long been recognized in time-space sampling, as in Sudman's (1980) classic study of patrons of shopping centers. In such sampling, the period of sampling must be extensive enough to include respondents who become accessible at different times. For example, a sample drawn on a weekend may differ from that drawn during work hours (see also Sudman and Kalton 1986; Sudman et al. 1988). Consequently, there may have been a gender difference in non-response bias.

The speed with which a WebRDS study can be conducted makes this issue relevant. Our feasibility study began on a Friday. Because of the accelerating nature of the sampling process, fully 90.6% (n=144) of the surveys were completed on the weekend (Saturday and Sunday). Consequently, respondents who were inaccessible that weekend—whether because they failed to check their e-mail or because they were off-campus—dropped out of the sampling frame. While it is unlikely that such a bias would favor one gender over another, differential access to email may explain the oversampling

of Engineering students (see Table 4). Engineering students, who make up 19.6% of the student population, represent 35.2% of our WebRDS sample. On average, sampled Engineers checked their email 11.15 times per day (n = 47) while non-engineers reported checking email 2.54 times fewer (8.61, n = 83). Furthermore, Engineers reported spending significantly more time in class and on school work than non-Engineers. On average, engineers spent 57.1 (n=56) hours in class or studying per week, while non-engineers devoted 46.8 hours per week (n=103). These findings suggest that Engineers selected into the WebRDS study were more likely to be checking their email and/or doing school work, which often requires email, on the weekend than non-Engineers.

Thus, a plausible source of Engineer oversampling bias is a temporal filtering effect due to email usage differentials during weekends. If that is the case, one solution is to ensure that the sampling process continues for at least one week, or ideally for at least two weeks, so that respondents can be reached irrespective of the times they are online. This could be done by introducing a delay between completion of the survey and e-mailing the recruitment coupons. There are, however, types of studies in which a different trade-off between speed and avoiding bias would be appropriate. Researchers tracking an emerging epidemic, for example, might consider a temporal filtering bias an acceptable price to pay for speeding up the sampling process.

Conclusion

WebRDS has several limitations. First and foremost, WebRDS requires individuals in the target population to have e-mail: individuals who are not electronically connected cannot be recruited or recruit others in their networks. Furthermore, the speed

of recruitment can be problematic for small samples. In populations where e-mail usage is highly variable, the sampling period must remain open long enough for light e-mail users to check their in-boxes and reply. Moreover, RDS estimation assumes that each respondent is a unique individual, and therefore steps must be taken to avoid duplicate participation in the sample. In standard RDS, where respondents are interviewed face to face, finding unique identifying features is straightforward; however, in online environments, one individual can have multiple e-mail addresses and a savvy user can easily disguise himself. Although further research is needed to develop better methods of preventing duplication, one possible measure is to keep compensation incentives low and the apparent probability of being caught high (note the "you will not be paid" warnings in Appendix 1). This way, the effort required to self-recruit will appear to be larger than the return from self-recruitment.

Second, successive observations in RDS are not independent because characteristics of recruiters and recruits tend to be correlated (Heckathorn 2002). Therefore, our comparisons with institutional data may be reasonable, but not precise because of a more than unitary design effect. After comparing the design effects of several standard RDS studies, Salganik (2006) recommends RDS samples be at least double that which would be required for a comparable simple random sample (SRS) design. Table 5 shows trait specific design effects for college, gender, and race. Consistent with Salganik's (2006) recommendation, the mean design effect is near two for each variable. Therefore, while our sample size is 159 students, the RDS estimates are comparable to SRS estimates based on a sample of about 80 students. Consequently, differences between estimates and institutional data may have resulted simply from either

small sample size or the interaction of small sample size with other factors.

[Table 5 here]

Despite these limitations, WebRDS provides a useful means of reaching hidden and non-hidden electronically connected populations quickly. While WebRDS is well-suited for any study interested in drawing samples quickly, two specific applications seem particularly appropriate. First, web-based RDS can be used for short-term serial cross sectional studies. Many serial cross-sectional studies have been multiyear endeavors, with waves at annual or longer intervals. However, many changes, both natural and induced, take place over the course of several weeks or months. Because of its sampling speed, WebRDS could be used at monthly or bimonthly intervals to monitor, for example, the effects of policy implementation on a community. Second, WebRDS is suitable for case-control studies conducted during infectious disease outbreaks to compare infected individuals with a representative sample of noninfected controls. Identifying a set of suitable controls is frequently a time-consuming process that limits the speed with which patterns of infection can be identified. When outbreaks occur in universities or other institutions that employ a proprietary e-mail system, web-based RDS could be used to draw the control sample quickly and thus accelerate treatment and containment measures.

Appendix 1: Recruitment E-mail

Sleep, School, or Social Life: A Study of Social Networks

Dear Student,

Congratulations! You have been invited to participate in a 20 minute online survey in exchange for up to $55. Your Password is: _____

IF YOU HAVE ALREADY COMPLETED THE SURVEY: Thank you! In order to reap the maximum compensation please forward each of these invitations to one other Cornell student to be recruited for participation in the study (for a total of 3 recruits). Remember, you will be compensated $15 for each recruit *that successfully completes the survey,* so it is in your best interest not to recruit strangers. DO NOT forward any one invitation to more than one person as multiple uses of the same password will invalidate the data and YOU WILL NOT BE PAID.

Please **identify yourself to your recruits** as they will need to know who recruited them to complete the survey.

If you do not know your recruit's e-mail click here: http://directory.cornell.edu/dsgw/bin/lang?context=cupb

IF YOU HAVE NOT YET COMPLETED THE SURVEY: Please read the following information carefully. Your complete participation in this study should not exceed 25 minutes.

You are invited to participate in a research study to empirically validate Respondent Driven Sampling (RDS) as an analytical tool for the study of social structure. You have been selected as a possible participant because you are a member of the Cornell community. We ask that you read the following and ask any questions you have before agreeing to be in the study.

Background information: This study uses a new sampling procedure called Respondent Driven Sampling (RDS) to analyze affiliation patterns in a population. RDS has been a seminal contribution to the study of hidden populations and network structure. Most notably RDS is paving the way for new lines of research in AIDS prevention by providing a way of reaching at risk groups that is both more effective and more efficient than previous methods. Our study is aimed at extending the ground breaking role of RDS to the study of non-hidden populations by providing a real-world validation of RDS methods and the first RDS analysis of a non-hidden population.

Procedures: If you agree to participate in the study, we will ask you to do the following 2 steps:

Step 1: Follow the link below to our survey website. After entering your unique password provided above (we recommend you cut and paste directly from this e-mail) fill out the online survey. The survey will only take 10-15 minutes to complete.

Step 2: After successful completion of the survey you will be sent three invitations to participate identical to this one, with the exception that each invite will contain a different password. We ask that you forward each invite to 1 member of the Cornell student body. For our purposes, it is crucial that no password is used more than once and that no individual person participates in the study more than once, any attempted duplication of subjects or passwords will invalidate the data and YOU WILL NOT BE PAID. You will be compensated for each person you recruit (forward an invite to) who successfully completes the survey (see Compensation below) therefore it is in your best interest to recruit subjects who are not strangers to you.

Once you have forwarded your invites, your participation is complete.

Online Survey: {link goes here}

Risks and Benefits of being in the study: We do no anticipate any direct risks or benefits for you participating in this study beyond those encountered in your daily life.

Compensation: You will receive $10 for successful completion of the online survey (Step 1). You will

receive an additional $15 for each subject you recruit (step 2) *who successfully completes the online survey*, for a possible total of $55. **NOTE**: Any attempted duplication or falsification of passwords, recruits (including self recruitment), or subjects will invalidate the data and YOU WILL NOT BE PAID.

Your decision whether or not to participate is completely voluntary and will not effect your current or future relations with any institution. If you decide to participate, you are free to withdraw at any point. Failure to complete the online survey, however, will result in forfeit of any compensation.

You must be at least 18 years of age to participate in this study.

Privacy: The records of this study will be kept private. In any sort of report we might publish, we will not include any information that will make it possible to identify you. Research records will be kept in a locked file; only the researchers will have access to the records, HOWEVER, in order for us to gain a deep understanding of Cornell social networks it is necessary to reveal your name (and only your name) to future participants in the study. As this is an online survey, we cannot guarantee 100% protection from third parties, however all possible precautions will be taken to protect your privacy.

Contacts and Questions: The researcher conducting this study is XXX who is working under the supervision of XXX. Please direct any questions you may have to:

XXX

E-mail: XXX

XXX

E-mail: XXX

If you have any questions or concerns regarding your rights as a subject in this study, you may contact the University Committee on Human Subjects (UCHS) at 607-255-5138, or access their website at

http://osp.cornell.edu/Compliance/UCHS/homepageUCHS.htm.

Appendix 2: Algorithm for Online RDS

The algorithm starts when a seed submits a survey. Upon survey submission:

→ Make validity checks (valid password, unused network ID, correct format for answers, etc.).

If checks fail, inform user and return him to survey.

If validity checks pass,

→ Save data in database.

→ Record and block respondent's network ID.

→ Record and block respondent's password.

→ Generate 3 new recruitment e-mails with new passwords.

→ Record these passwords as valid for recruitment.

→ Send new recruitment e-mails to respondent.

Table 1: Comparison of Six Standard RDS Sampling Rates

Study Site	Middletown, CT	Meriden, CT	New London, CT	Chicago	San Francisco	NYC
Study Population	IDU	IDU	IDU	Latino MSM	Latino MSM	DU
Compensation (Interview / 1 recruitment)	$4 / $10	$4 / $10	$4 / $10	$50 / $20	$50 / $20	$20 / $10
Total Sample	100	100	100	100	100	618
Total Sampling Time (Weeks)	52	52	52	13	7	13
Weekly Sample Size	1.9	1.9	1.9	7.7	14.3	47.5
Interview Length	2 hr	2 hr	2 hr	2 hr	2 hr	.5 hr

IDU - Injection drug users
DU - Drug Users
MSM - Men who have sex with men

Table 2: Demographic Characteristics of the Final Sample

		Seeds n = 9	Sample n = 159	Percentage
Age				
	18	0	10	6.3
	19	2	37	23.3
	20	3	41	25.8
	21	2	50	31.4
	22	2	18	11.3
	>22	0	3	1.8
Gender				
	Male	5	95	59.7
	Female	4	64	40.3
Race				
	White	6	81	50.9
	Asian	2	58	36.5
	Other	1	20	12.6
Year				
	Freshman	0	13	8.2
	Sophomore	4	51	32.1
	Junior	2	40	25.2
	Senior	3	51	32.1
	5+	0	4	2.5
College				
	Agriculture & Life Sciences	3	35	22
	Arts & Sciences	4	37	23.3
	Engineering	0	56	35.2
	Architecture	0	1	0.6
	Hotel	2	11	6.9
	Industrial & Labor Relations	0	7	4.4
	Human Ecology	0	12	7.5
Housing				
	Dormitory	4	67	40.9
	Other University Housing	0	2	1.2

	Fraternity	1	7	4.4
	Sorority	0	5	3.1
	Shared Apartment w/ Friends	3	63	39.6
	Room/Apartment w/o Friends	1	10	6.3
	Other	0	5	3.1
Fraternity / Sorority Membership				
	None	7	131	82.4
	Fraternity	1	15	9.4
	Sorority	1	13	8.2
High School				
	Public	4	124	78
	Private	5	35	22

Table 3A: Random Recruitment Test by Gender

RDS Recruitment Matrix

	Male	Female	Total
Male	66	30	96
Female	25	29	54
			150

Self-Report Mean Degree

	Male	Female	Total
Male	40.99	38.24	79.23
Female	30.87	41.1	71.97
			151.2

Expected Recruitment Matrix

	Male	Female	Total
Male	54.760	41.240	96
Female	26.030	27.970	54
			150

Chi-Square Test, df=1

	Male	Female	Total
Male	2.307	3.063	5.371
Female	0.041	0.038	0.079
		Statistic	5.449
		p-value:	0.02

Table 3B: Random Recruitment Test by Race

RDS Recruitment Matrix

	White	Asian	Other	Total
White	55	13	7	75
Asian	14	39	9	62
Other	6	4	3	13
				150

Self-Report Mean Degree

	White	Asian	Other	Total
White	54.64	10.25	6.84	71.73
Asian	25.38	35.74	5.43	66.55
Other	63.2	23.05	27.1	113.4
				251.6

Expected Recruitment Matrix

	White	Asian	Other	Total
White	57.131	10.717	7.152	75
Asian	23.645	33.296	5.059	62
Other	7.248	2.644	3.108	13
				150

Chi-Square Test, df=4

	White	Asian	Other	Total
White	0.079	0.486	0.003	0.569
Asian	3.934	0.977	3.071	7.982
Other	0.215	0.696	0.004	0.915
			Statistic	9.465
			p-value:	0.051

Table 4: RDS Population Estimates Based on Full Sample (n=159) and Large Component (n=117)

		True Proportion	Full Sample Estimate, P n=159 (95% CI)	Large Component Estimate, P n=117 (95% CI)	Full Sample Proportion n=159	Large Component Sample Proportion n=117
College	Agriculture	0.222	**0.264**** (.16, .38)	**0.254**** (.16, .38)	0.22	0.205
	Arts & Sciences	0.322	**0.257*** (.16, .35)	**0.267*** (.14, .35)	0.233	0.222
	Engineering	0.196	**0.266*** (.18, .36)	**0.263*** (.17, .38)	0.352	0.402
	Hotel	0.058	**0.063**** (.02, .15)	**0.014*** (.00, .059)	0.069	0.034
	Industrial & Labor Relations	0.06	**0.067**** (.03, .14)	**.109*** (.05, .21)	0.044	0.043
	Human Ecology	0.096	**0.08**** (.04, .12)	**0.09**** (.03, .14)	0.075	0.085
Gender	Male	0.503	**0.634** (.52, .73)	**0.605*** (.49, .72)	0.597	0.573
	Female	0.497	**0.366** (.27, .48)	**0.395*** (.28, .51)	0.403	0.427
Race	White[+]	0.594	**0.645**** (.52, .76)	**0.603**** (.44, .73)	0.509	0.487
	Asian†	0.164	**0.258*** (.16, .38)	**0.314** (.19, .475)	0.365	0.393
	Other‡	0.242	**0.097** (.05, .14)	**0.083** (.04, .13)	0.126	0.12

* True value captured by RDS estimate CI

** True value captured by RDS estimate ± 1 s.d.

[+] True institutional proportion of White students may be as high as 0.665.

† True institutional proportion of Asian students may be as high as 0.236.

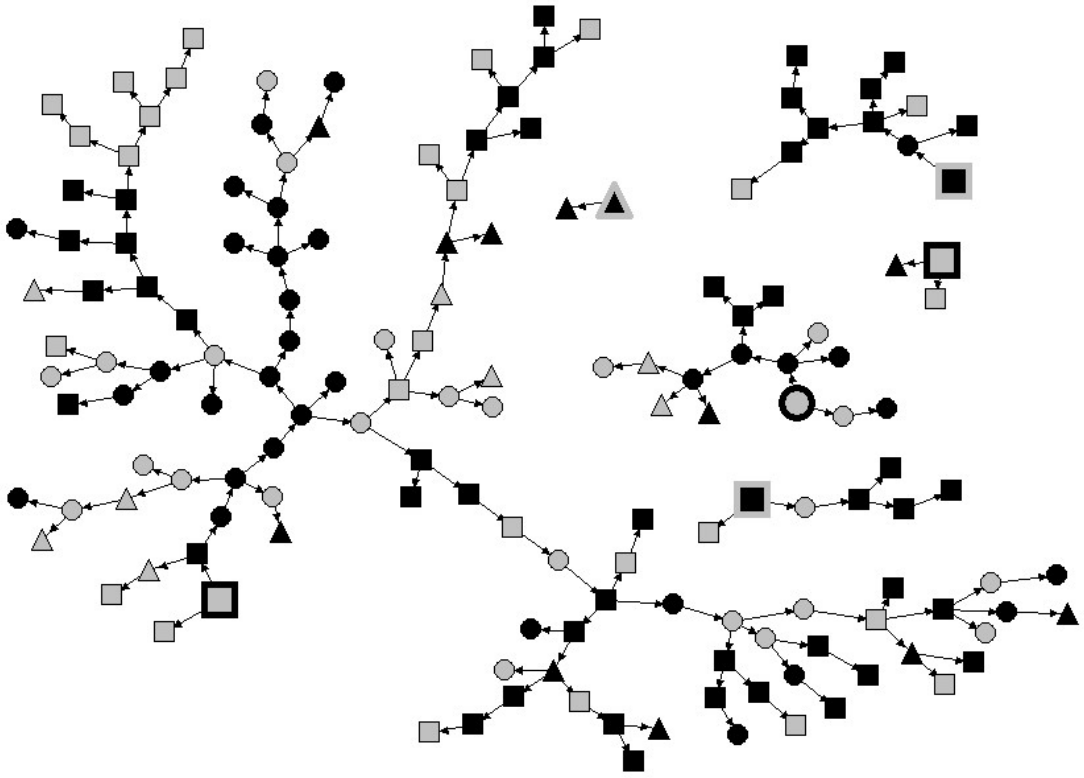
‡ True institutional proportion of Other students may be as low as 0.169

Estimates calculated using RDSAT 5.6

Table 5: WebRDS Design Effects

		Design Effect	n
College	Agriculture	2.2	35
	Arts & Sciences	1.7	37
	Engineering	1.6	56
	Hotel	2.5	11
	Industrial & Labor Relations	1.8	7
	Human Ecology	0.9	12
Gender	Male	1.8	95
	Female	1.8	64
Race	White	2.5	81
	Asian	2.3	58
	Other	0.9	20

Figure 1: RDS recruitment chains of all 6 productive seeds (enlarged and highlighted with opposite color borders). The nodes are color-coded for gender (male = black, female = gray) and shape-coded for race/ethnicity (White = square, Asian = circle, Other = triangle).



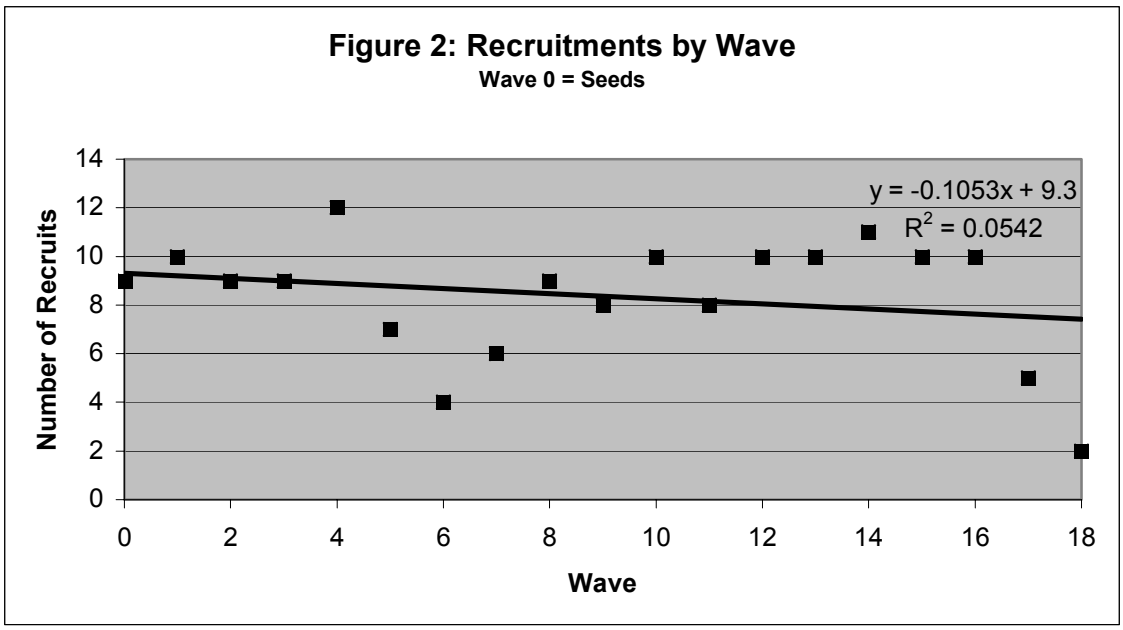


Figure 2: Recruitments by Wave
Wave 0 = Seeds
y = -0.1053x + 9.3
R2 = 0.0542
Number of Recruits
Wave
0
2
4
6
8
10
12
14
16
18

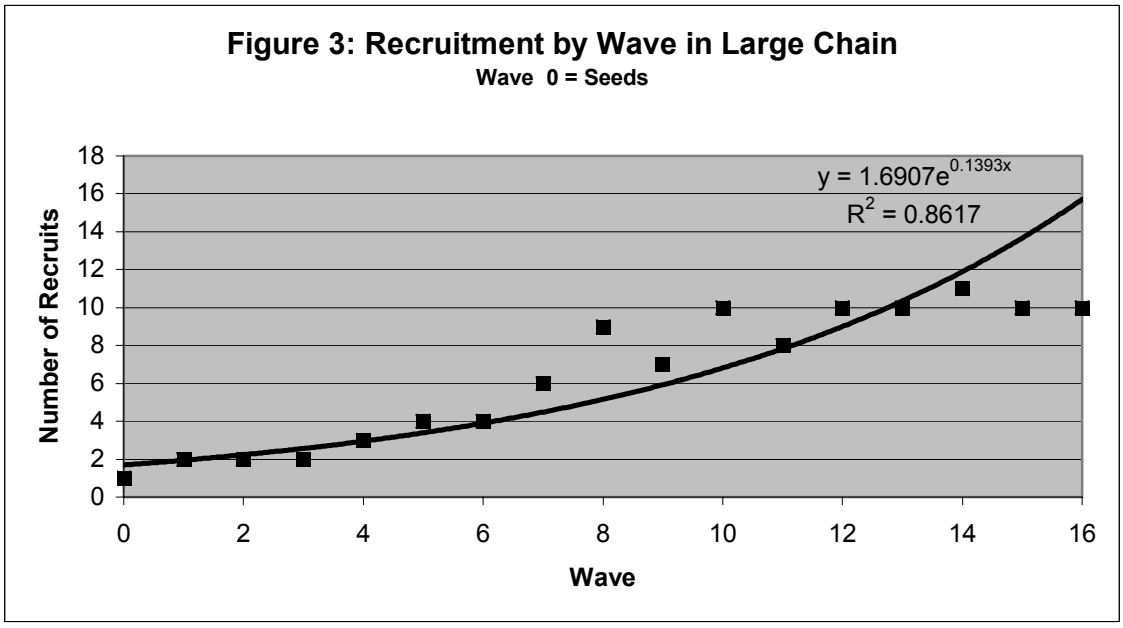


Figure 3: Recruitment by Wave in Large Chain
Wave 0 = Seeds
y = 1.6907e0.1393x
R2 = 0.8617
Number of Recruits
Wave
0
2
4
6
8
10
12
14
16
18

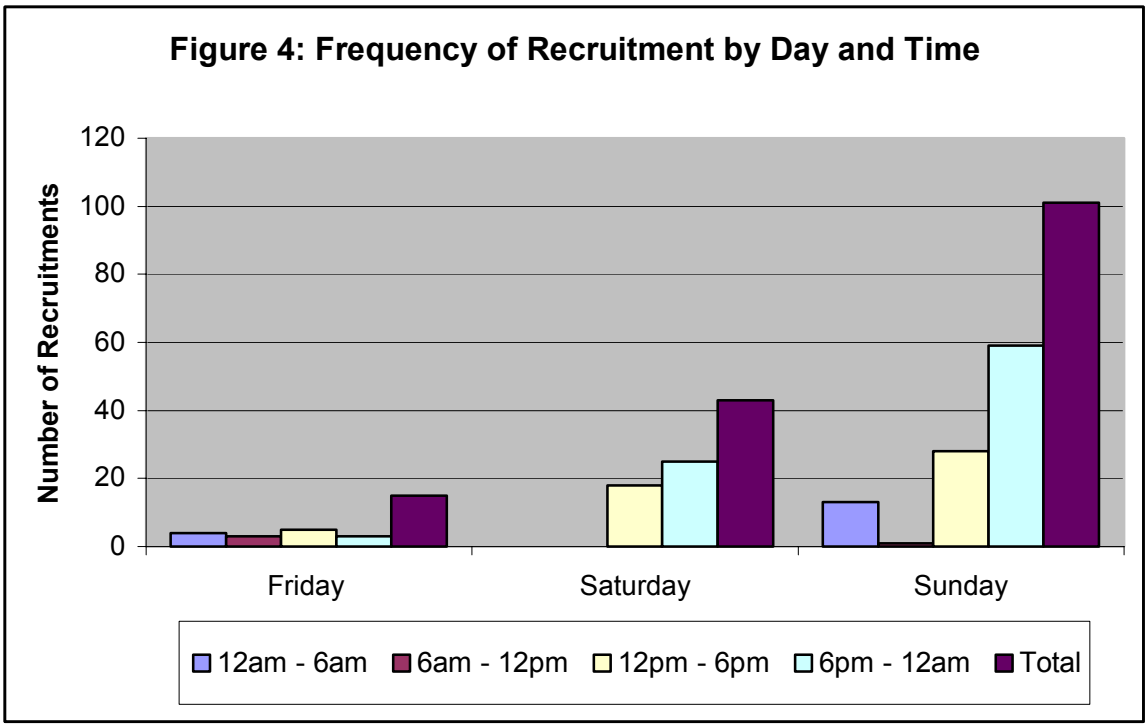


Figure 4: Frequency of Recruitment by Day and Time
Number of Recruitments
120
100
80
60
40
20
0
Friday
Saturday
Sunday
12am - 6am
6am - 12pm
12pm - 6pm
6pm - 12am
Total

References

Abdul-Quader, Abu S., Douglas D. Heckathorn, Courtney McKnight, Heidi Bramson, Chris Nemeth, Keith Sabin, Kathleen Gallagher, and Don C. Des Jarlais. 2006. "Effectiveness of Respondent Driven Sampling for Recruiting Drug Users in New York City: Findings from a Pilot Study." *AIDS and Behavior* 9: 403–408.

Bell, David C., Benedetta Belli-McQueen, Ali Haider. 2007. "Partner Naming and Forgetting: Recall of Network Members." *Social Networks* 29: 279-299.

Berg, S. 1988. "Snowball Sampling." In *Encyclopedia of Statistical Sciences* 8: 528-532. S. Kotz and N. I. Johnson eds. New York: Wiley.

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

Bollabás, Béla. 1985. *Random Graphs*. Cambridge: Cambridge University Press.

Cochran, William G. 1977. *Sampling Techniques.* 3d ed. New York: Wiey.

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

Cornell University. 2004. "Enrollment at a Glance." Ithaca, NY: Cornell University, Division of Planning and Budget. Available at http://dpb.cornell.edu/F_Undergraduate_Enrollment.htm, accessed May 5, 2004.

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

------. 2002. "Respondent-Driven Sampling II: Deriving Valid Population Estimates From Chain Referral Samples of Hidden Populations." *Social Forces* 49: 11–34.

------. Forthcoming. "Extensions of Respondent-Driven Sampling: Analyzing Continuous

Variables and Controlling for Differential Degree" *Sociological Methodology*.

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

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

Heckathorn, Douglas D. and Robert Magnani. Forthcoming. "Snowball and Respondent-Driven Sampling" in *Behavioral Surveillance Surveys: Guidelines for Repeated Behavioral Surveys in Populations at Risk for HIV*.

Heckathorn, Douglas D., Robert S. Broadhead, Denise L. Anthony, and David L. Weakliem. 1999. "AIDS and Social Networks: Prevention through Network Mobilization." *Sociological Focus* 32: 159–79.

Heckathorn, Douglas D., Salaam Semaan, Robert S. Broadhead, and James 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: 55–67.

Heimer, Robert. 2005. "Critical Issues and Further Questions About Respondent-Driven Sampling: Comment on Ramirez-Valles et al. (2005)" *AIDS and Behavior* 9: 403-408.

Ibarra, Herminia. 1993, "Personal Networks of Women and Minorities in Management: a Conceptual Framework. *Academy of Management Review* 18: 56-87.

Johnston, Lisa, Keith Sabin, Mai Thu Hien, and Pham Thi Huong. 2006. "Effectiveness of Respondent-Driven Sampling to Recruit Female Sex Workers in Two Cities in

Vietnam." *Journal of Urban Health* 83: i16-i28..

Marsden, Peter V. 1990. "Network Data and Measurement." *Annual Review of Sociology* 16: 435-463.

McCarty, Christopher, Peter D. Killworth, H. Russell Bernard, Eugene C. Johnsen, and Gene A. Shelley. 2001. "Comparing Two Methods for Estimating Network Size." *Human Organization* 60: 28-39.

McPherson, Miller and Lynn Smith-Lovin. 1987. "Homophily in Voluntary Organizations: Status Distance and the Composition of Face-to-Face Groups." *American Sociological Review* 52: 370-379.

Ramirez-Valles, Jesus, Douglas D. Heckathorn, Raquel Vázquez, Rafael M. Diaz, and Richard T. Campbell. 2005a. "From Networks to Populations: The Development and Application of Respondent-Driven Sampling among IDUs and Latino Gay Men." *AIDS and Behavior* 9: 387–402.

------. 2005b. "Evaluating Respondent-Driven Sampling: Response to Heimer." *AIDS and Behavior* 9: 403–408.

Salganik, Mathew J. 2006. "Variance Estimation, Design Effects, and Sample Size Calculations for Respondent-Driven Sampling." *Journal of Urban Health* 83: i98-i112.

Salganik, Mathew J., and Douglas D. Heckathorn. 2004. "Sampling and Estimation in Hidden Populations Using Respondent Driven Sampling." *Sociological Methodology* 34: 193–239.

Semaan, Salaam, Jennifer Lauby, Jon Liebman. 2002. "Street and Network Sampling in Evaluation Studies of HIV Risk-Reduction Interventions." *AIDS Review* 4: 213-

223.

Stueve, Ann, Lydia N. O'Donnell, Richard Duran, Alexi San Doval, Julianna Blorne. 2001. "Time-Space Sampling in Minority Communities: Results with Young Latino Men who have Sex with Men." *American Journal of Public Health* 91: 922-926.

Sudman, Seymour. 1980. "Improving the Quality of Shopping Center Sampling." *Journal of Marketing Research* 17: 423–31.

Sudman, Seymour, and Graham Kalton. 1986. "New Developments in the Sampling of Special Populations." *Annual Review of Sociology* 12: 401–29.

Sudman, Seymour, Monroe G. Sirken, and Charles D. Cowan. 1988. "Sampling Rare and Elusive Populations." *Science* 240: 991–96.

Van Emmerick, I. J. Hetty. 2006. "Gender Differences in the Creation of Different Types of Social Capital: A Multilevel Study." *Social Networks* 28:24-37.

Volz, Erik, and Douglas D. Heckathorn. Forthcoming. "Probability-Based Estimation Theory for Respondent-Driven Sampling." *Journal of Official Statistics.*

Wang, Jichuan, Robert G. Carlson, Russel S. Falck, Harvey A. Siegal, Ahmmed Rahman, and Linna Li. 2005. "Respondent Driven Sampling to Recruit MDMA Users: A Methodological Assessment." *Drug and Alcohol Dependence* 78: 147–57.

Watts, Duncan J. and Steven H. Strogatz. 1998. "Collective Dynamics of 'Small-World' Networks." *Nature* 393: 440-442.

Notes

[1] Since the proportion of each gender can be calculated directly from the other, gender is considered one estimate.

[2] Waves 17 and 18 are excluded because they were cut short by the end of the study and thus greatly underestimate the number of recruitments that would have occurred had the target sample of 150 not been reached. It is likely that wave 15 and 16 recruitments are also reduced by this constrain, potentially causing the leveling off of recruitment visible in later waves.

[3] Although still problematic, web-based RDS recruitment e-mails are less likely to look like SPAM than population-based e-mail surveys because the message comes from a known peer. Moreover, in this study, most respondents contacted their potential recruits by phone, text messaging, or in person to tell them to expect the survey.

[4] Respondents were asked about email usage when they arrived to pick up incentives. Approximately 20% of the sample did not pick up incentives personally and email usage information could not be gathered. Additionally, the question did not differentiate between weekday and weekend email usage, however we speculate that large differences in email usage are possible between weekdays and weekends.