# EXTENSIONS OF RESPONDENT-DRIVEN SAMPLING: WEB-BASED RDS, EMPIRICAL VALIDATION, AND THE DUAL HOMOPHILY MODEL

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

In Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Cyprian Wejnert

May 2009

# EXTENSIONS OF RESPONDENT-DRIVEN SAMPLING: WEB-BASED RDS, EMPIRICAL VALIDATION, AND THE DUAL HOMOPHILY MODEL

Cyprian Wejnert, Ph. D.

Cornell University 2009

This dissertation makes contributions to Respondent-Driven Sampling (RDS) and the study of social networks. RDS is a new network-based method of collecting and analyzing data from hidden populations in a statistically viable way. The first chapter provides an introduction to RDS procedures and estimation. After describing the operating procedures, the chapter introduces the statistical theory behind RDS, including the models assumptions and how it accounts for sources of bias commonly associated with network samples. It then compares two distinct families of RDS estimator, RDS I and RDS II, by describing the evolution of all seven RDS estimators. Chapter Two introduces WebRDS, an online version of RDS that has been shown to produce samples in record speeds, and describes the two WebRDS samples on which the remaining analyses are based. Chapter Three provides an in depth empirical test of RDS estimators and confidence intervals. While RDS estimation has been validated analytically and computationally, it has not been empirically tested on a population with known parameters. Chapter Three utilizes RDS data on university undergraduates to compare the accuracy of RDS point and variance estimates across two estimation techniques (RDS I and RDS II), self-report measures of degree, and multiple cut-points for excluding early wave data. The chapter RDS I and RDS II estimates to be accurate and convergent, but estimates of variance to be problematic in opposite ways.

The RDS I bootstrap method tends to under estimate variance, while RDS II analytical variance estimation provides an over estimate. For both methods, the problem is exacerbated in small groups. Differences in degree measure and cutting early wave data resulted in only minor differences in the estimation. Chapter Four presents the Dual Homophily Model, which breaks a common measure of homophily into two components, one due to relational preferences and one due to differential degree. Applications of the model, including examples where standard homophily measures miss important differences between groups, are discussed.

## BIOGRAPHICAL SKETCH

Cyprian Wejnert was born in Poznan, Poland on March 27, 1981 and immigrated to the United States at the age of five. He attended elementary school in Statesboro, GA and high school in Ithaca, NY. As the son of a Sociologist, he grew up with a strong interest in and awareness of social issues. He earned his BA in Biological Statistics from Cornell University in 2003. In 2006 he earned his MA in Sociology at Cornell University. In 2008 he married Katherine E. Newman, a paleoceanographer. His work has been published in *Sociological Methodology*, *Sociological Methods and Research*, and *Marriage and Family Review*.

for Kate, my sails

# ACKNOWLEDGMENTS

I am deeply grateful to the members of my special committee, Dr. Douglas Heckathorn, Dr. Michael Macy, Dr. Stephen Morgan, and Dr. Richard Swedberg, for their advice, excitement, motivation, and willingness to listen to my ideas. I would especially like to thank my special committee chair, Dr. Heckathorn. His generosity and help made this project possible. His work and passion for research lead me to the social sciences.

I am grateful to the Sociology Department staff, Sharon Sandlin, Sue Meyer, and Alice Murdock, for providing the critical bridge between academia and the real world with a friendly smile.

I am thankful to my wife, Katherine Wejnert, and family, Barbara Wejnert, Richard Depue, and Camille Wejnert-Depue, for their continuing support and undying love. Without them, there would be no point.

Lastly, I would like to thank my fellow Sociology graduate students. They have served as reviewers, critics, colleagues, and most of all friends. I am especially grateful to my close friends Jason Perelshteyn and Michael “Trey” Spiller III, who kept me sane these past years.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

# INTRODUCTION: RESPONDENT-DRIVEN SAMPLING OPERATIONAL PROCEDURES, EVOLUTION OF ESTIMATORS, AND TOPICS FOR FUTURE RESEARCH

## *Introduction*

Respondent-Driven Sampling (RDS) is a method for drawing and analyzing probability samples of hidden, or "hard-to-reach," populations. Populations such as these can be difficult to sample using standard survey research methods for two reasons: First, they lack a sampling frame, that is, an exhaustive list of population members from which the sample can be drawn. Second, constructing a sampling frame is not feasible because one or more of the following are true: (a) the population is such a small part of the general population, that locating them through a general population survey would be prohibitively costly; (b) because the population has social networks that are difficult for outsiders to penetrate, access to the population requires personal contacts; and/or (c) membership in the population is stigmatized, so gaining access requires establishing trust. Populations with these characteristics are important to many research areas including public health studies of HIV and other infectious disease, sociological studies of the welfare of marginalized or low income groups, and network studies of large populations.

RDS is now widely used to study a wide range of hidden populations in the U.S including jazz musicians (Heckathorn and Jeffri 2001), aging artists (Spiller et al. 2008), drug users (Abdul-Quader et al. 2006), men who have sex with men (Ramirez-Valles et al. 2005), and Latino migrant workers (Kissinger et al. 2008). Internationally,

over 120 studies in 30 countries have used RDS to study HIV/AIDS and other sexually transmitted infections (Malekinejad et al. 2008).

RDS accesses members of hidden populations through their social networks, employing a variant of snowball or "chain-referral" sampling. As in all such samples, the study begins with a set of initial respondents who serve as *seeds*. These seeds then recruit their acquaintances, friends, or relatives who qualify for inclusion in the study to form the first *wave*. The first wave respondents then recruit the second wave, who in turn recruit the third wave, and so forth. By allowing respondents to recruit new participants directly, RDS removes the need for researchers to locate population members, penetrate social networks through personal contacts, or establish trust within stigmatized populations.

While snowball sampling has been used for decades (Coleman 1958), the resultant data have generally been viewed as convenience samples because respondents are not sampled in a random way. RDS challenges this view by applying a mathematical model that weights the sample to compensate for the fact that it was not obtained in a simple random way (Salganik and Heckathorn 2004). Consequently, RDS provides researchers a method of harnessing the advantages of snowball sampling without sacrificing the ability to make unbiased population estimates.

In this chapter I first present operational procedures used in collecting RDS data. I then outline the progression of two families of the RDS estimator and discuss RDS network analysis techniques. The chapter concludes with a discussion of limitations, ongoing projects, and directions for future RDS development.

## *Operational Procedures of RDS*

RDS operational procedures primarily consist of recruiting seeds, setting incentives, and collecting data necessary for RDS analysis. Additionally, it is important for operating procedures to promote long recruitment chains and minimize recruitment by strangers.

### Seed Selection

As in all chain-referral samples, the sampling process in RDS begins with the selection of an initial set of respondent group members or seeds (Heckathorn 1997). The seeds complete the survey interview and are then asked to recruit a specified number of additional respondents to be interviewed, who in turn recruit a subsequent wave of respondent group members, and so on until a target sample size has been reached. Because the ultimate sample composition under RDS does not depend upon the characteristics of the seeds chosen, it is not necessary that the seeds be randomly chosen. However, because the rate at which sample composition becomes independent of seeds is increased if the seeds chosen are diverse with regard to key characteristics, choosing a diverse set of seeds increases the efficiency of the sampling operation.

Given that recruitment chains grow only if seeds actually recruit, it is also important that seeds be well motivated. Ideally, seeds should be sociometric stars that are committed to the goals of the study. These characteristics fit the "volunteers" with which many snowball samples begin, and have traditionally been seen as a source of bias in these samples. In contrast, in RDS, given that seed selection becomes irrelevant only if seeds succeed in spawning expansive recruitment chains, starting with high energy seeds does not add to bias but instead reduces it by speeding the recruitment process.

**Incentives**

RDS relies on dual incentives to encourage participation and achieve sufficiently long referral chains (Heckathorn 1997). First, respondents are rewarded for participating in an interview. Second, respondents are given a modest reward for each peer they recruit into the study. For example, in a recent U.S. study of drug users conducted by De Jarlais et al. (2007), respondents were paid $20 for participating in a survey interview and an additional $10 for each drug user they successfully recruited (i.e., whose recruits subsequently appear to be interviewed and fulfill the study criteria for inclusion). These recruits were in turn paid to be interviewed and for each successful recruit.

The size of incentive is determined on a setting-by-setting basis, but in general should be of sufficient in size to encourage participation by respondent group members, but not so large as to encourage participation by imposters. Excessive rewards could also encourage coercive recruitment.

Idealistic motives for recruiting peers are also emphasized. In this way, respondents are provided not only with a means to earn respondent fees, but also a means to help peers by giving them the opportunity to benefit from participation in the study. It is emphasized to subjects that by recruiting peers, they are undertaking a task that in most other studies is carried out by public health professionals; and the rewards they receive are recognition for their having succeeded at this important task. Rewards for recruiters are a useful means for promoting peer recruitment and thereby producing the large recruitment chains upon which the RDS method depends.

**Recruitment Quotas and the Referral Process**

As discussed below, sampling bias is minimized in RDS by having long referral chains. In order to encourage longer referral chains and promote greater socio-metric depth, recruitment quotas are used in order to limit the ability of

population members with large personal networks to dominate a given sample (Heckathorn 1997). Consider, for example, what would happen without quotas. If one respondent recruited 10 peers, who each recruited 10 peers, the sample size would quickly grow huge, e.g., starting from a single seed (wave 0) to 10, then 100, then 1,000, and 10,000 by wave four. In contrast, if each recruited only two peers, the growth would be much slower, e.g., from the single seed, to two, then four, then eight, and 16 by wave four. Thus, for any given sample size, restrictive recruitment quotas produce recruitment chains with more waves. Quotas are also useful because they make recruitment rights scarce and hence too valuable to waste on strangers.

Choosing the proper recruitment quota involves a tradeoff. If the quota is too small, recruitment may die out because some subjects fail to recruit and others do not fulfill their quotas. Furthermore, restrictive recruitment quotas slow the recruitment process, because they prevent energetic recruiters from contributing as much as they could. Therefore, quotas should be small, but not oppressively so. In most RDS applications in the U.S. to date, the quota has been set at three or four recruits per recruiter, and only about two-thirds of recruitment rights have usually been exercised, so when the quota is three, the average number of recruits per subject is two. In general, an initial quota of three recruits per respondent group member is recommended.

In many applications of RDS, recruitment quotas have been implemented by providing subjects with paper money-sized recruitment coupons (see Figure 1). The coupon includes information on how to contact the project and a map to the interview site. Each coupon also includes a unique serial number. This is useful for determining how much each subject should be paid for recruitment. More importantly, it is also useful for documenting who recruited whom, a piece of information that is crucial for calculating RDS population estimates. The serial number also ensures that only the

subject to whom it was given can be rewarded for the recruitment, so the recruitment coupons cannot circulate as though they were an alternative form of money.


Figure 1: Example of a recruitment coupon employed in an RDS study of Connecticut injection drug users (IDUs). Note that the front includes a serial number, and the back includes a map to the interview site.

**Data Requirements**

RDS analysis has special data requirements because each analysis requires not only information on the focal variable, but also two additional items of information that function to provide the sampling frame from which post-stratification weights are calculated. These are:

- Cross-group recruitment (e.g., proportional recruitment of HIV positives by HIV negatives, and recruitment of HIV negatives by HIV positives)
- Estimated mean network sizes (e.g., the estimated mean network sizes of HIV positives and negatives).

The reason why every RDS study must keep track of who recruited whom is so these cross-group recruitment proportions can be calculated, and the reason why each respondent must be asked about their personal network size, or *degree,* is so estimated mean degree by group can be calculated. These are then used to calculate unbiased population estimates.

A typical question for measuring personal degree in a study of injection drug users (IDUs) is: "How many people do you know personally who inject, that is people you know, who also know you, and that you have seen at least once in the last 6 months?" Note that it is essential that this question be framed so subjects are asked about the number of peers they know who fit the screening criteria for the study, because the aim of this question is to find out how many potentially recruitable persons the respondent knows. It is also important that the question make clear that these are not people the respondent has heard about, but persons with whom the respondent has a personal relationship. Finally, an interval for most recent contact should be specified to exclude persons known only in the distant past.

In order to provide a test of two key assumptions for analysis (discussed below) two addition items of information are collected:

- Relationship of recruit to recruiter: This can be assessed using the following question asked of each recruit: How can your relationship to your recruiter be best described: As closer than a friend; As a friend; As an acquaintance; As a stranger; etc.?
- Proportional distribution of networks: This is generally added as a follow up to the degree question by asking: How many of these [answer to network question] people are white? Black? Hispanic? Male? Female? Etc.

These last items allow the researcher to test two key assumptions of RDS: reciprocity (that each respondent knows his or her recruiter) and random recruitment

from among one's peers (that the composition of recruitment is representative of the composition of personal networks).

**WebRDS**

Wejnert and Heckathorn (2008) introduce an online version of RDS, termed *WebRDS*. WebRDS studies follow similar operating procedures as regular RDS, except the interview is replaced by a web-based survey and recruitment occurs through an electronic medium such as email. Among populations that are well connected electronically, WebRDS provides several advantages over regular RDS. First, because there is no need for an interview location or staff, the operating cost in terms of manpower and capital is minimal. Once seeds have been contacted and the survey has been set up, the researcher need only distribute incentives and download the data. Second, because respondents can be recruited, complete the survey, and recruit peers from their personal computer, the sampling speed can be especially fast. In their WebRDS study of university undergraduates, Wejnert and Heckathorn (2008) were able to collect a sample of 159 surveys in 72 hours. While not yet tested, WebRDS also has the potential to sample online communities without geographical limitation.

WebRDS has several limitations. First, the anonymity of the internet makes uniquely identifying respondents and therefore preventing study exploitation through repeat participation difficult. Similarly, limiting false positives, that is respondents who are not members of the target population, presents a challenge. Finally, because respondents are never physically in contact with the researcher, distribution of incentives can be problematic. While further research is needed to fully remove these limitations, incentives and their distribution can be designed to provide some safeguards. For example, in a WebRDS study of university undergraduates, Wejnert and Heckathorn (2008) required respondents to pick up incentives in person and

present a valid university student ID. Alternately, online studies could provide gift cards to sellers who only sell products of interest to the target population. WebRDS is presented in detail in Chapter Two.

**RDS in Small World Networks**

Much recent work on social networks has focused on networks with small world properties. Small world networks are characterized by two conditions; First, at the local level, a connection neighborhood is preserved such that individuals close to each other are more likely to form ties. Second, at the global level, a significant number of far reaching or random ties exists such that any node can be reached from any other in a small number steps along the network (Watts 1999). Consequently, small world networks include high levels of local clustering and global connectivity.

Because these networks are widely regarded as representative of "real-world" networks (Amaral et al. 2002) it is important to consider whether such network structure is beneficial to the application of RDS. Below I show that the level of homophily, a measure of clustering, is related to the rate at which RDS data become independent of the seeds from which sampling began. Thus, at first glance it seems small world social networks may be problematic for RDS. However, homophily is only problematic for RDS at levels so high as to preclude connectivity across groups, which does not occur in small world networks.

In most cases, connectivity observed in small world networks guarantees that all individuals in the network have non-zero probability of selection (Newman et al. 2002), a requirement of RDS analysis. Furthermore, local clustering characteristic of small world networks aids the use of social incentives for recruitment and promotes the emergence of a norm of participation among respondents. Consequently, global connectivity and local clustering present in small world networks are both beneficial

in promoting efficient RDS sampling and conducting unbiased estimation based on RDS data.

As a final note, it is important to point out that, while beneficial, small world networks are not the ideal structure for RDS applications. The ideal structure for RDS is the random network, because RDS applied to a random network provides a random sample of the population. However, as is widely noted, random networks are rare in real life.

### *Analysis and Estimation*

RDS is based on a mathematical model of the recruitment process which functions somewhat like a corrective lens, controlling for the distorting effects of network structure on the sampling process to produce an unbiased estimate of population characteristics. This procedure includes controls for four biases that are inherent in any snowball sample:

- The seeds cannot be recruited randomly, because if that were possible, the population would not qualify as "hidden" in the first place. Generally, the seeds are respondents to whom researchers have easy access, a group that may not be representative of the full target population. Consequently, the seeds introduce an initial bias.
- Respondents recruit their acquaintances, friends, and family members, whom they tend to resemble in income, education, race/ethnicity, religion, and other factors. The implication of this "homophily" principle is that by recruiting those whom they know, respondents do not recruit randomly. Instead recruitments are shaped by the social network connecting the target population.

- Respondents who are well-connected tend to be over-sampled, because more recruitment paths lead to them. Therefore, respondents who have larger social networks are over-sampled.
- Population subgroups vary in how effectively they can recruit, so the sample reflects disproportionately the recruitment patterns of the most effective recruiters. For example, in AIDS prevention research, HIV positives generally recruit more effectively, and also tend to recruit other positives, so positives tend to be over-sampled.

RDS employs a Markov chain model to approximate the recruitment process. This model is based on two observations (Heckathorn 2002): (1) if recruitment chains are sufficiently long, an equilibrium is reached in which the sample composition is independent of the initial seeds; (2) information gathered during the sampling process can be used to account for sampling bias.

**Equilibrium**

The first observation is recognizing that if referral chains are sufficiently long; that is, if the chain-referral process consists of enough waves or cycles of recruitment, the composition of the final sample with respect to key characteristics and behaviors will become independent of the seeds from which it began. In other words, after a certain number of waves, the sample compositions stabilize, remaining unchanged during further waves, and this sample composition is independent of the seeds from which sampling began. This point at which the sample composition becomes stable is termed the *equilibrium*.

Figure 2 illustrates this process. Figure 2 (top) uses university data on peer recruitment by gender and fraternity/sorority membership to project what the sample composition would have been had sampling begun with only non-fraternity males. The seeds (wave 0) would have all been non-fraternity males, but their percentage

declines to 51% in wave 1, 30 in wave 3, and stabilizes at 26%. Figure 2 (bottom) projects what would have happened had all the seeds been from female sorority members. The percent of non-fraternity males among the seeds (wave 0) would be 0%, but this would increase to 9% in wave 1, 19% in wave 3, 24% in wave 5, and stabilize at 26%. Note that after the first several waves, the sample composition is the same whether the seeds were all non-fraternity males or all female sorority members. The same would be true had the all seeds been drawn from other groups, or any combination of groups. The implication is that if recruitment chains are sufficiently long, the selection of seeds becomes irrelevant, so lengthening recruitment chains provides the means for overcoming bias from the choice of seeds (Ramirez-Valles et al. 2005).

The number of waves required to reach equilibrium varies based on the level of segmentation, or *homophily* (discussed in detail below), present in the population. Figure 3 shows the relationship between homophily and the number of waves required for equilibrium to be attained when all the seeds are drawn from the same group. The curve is accelerating (that is, as one moves to the right it becomes more steep). When homophily is zero, equilibrium is attained in only a single wave, because irrespective of group membership, each subject recruits randomly from the target population. As homophily grows, so too does the number of waves required for equilibrium to be attained, because it takes an increasing number of waves to break out of the initial group. In the extreme case of 100% homophily, recruitment chains could never break out of that group, so equilibrium would never be attained (such a case is a violation of the second RDS assumption, discussed below). In Figure 2, homophily varies from 0.233 among non-fraternity males to 0.628 among fraternity males and equilibrium is reached in five waves.

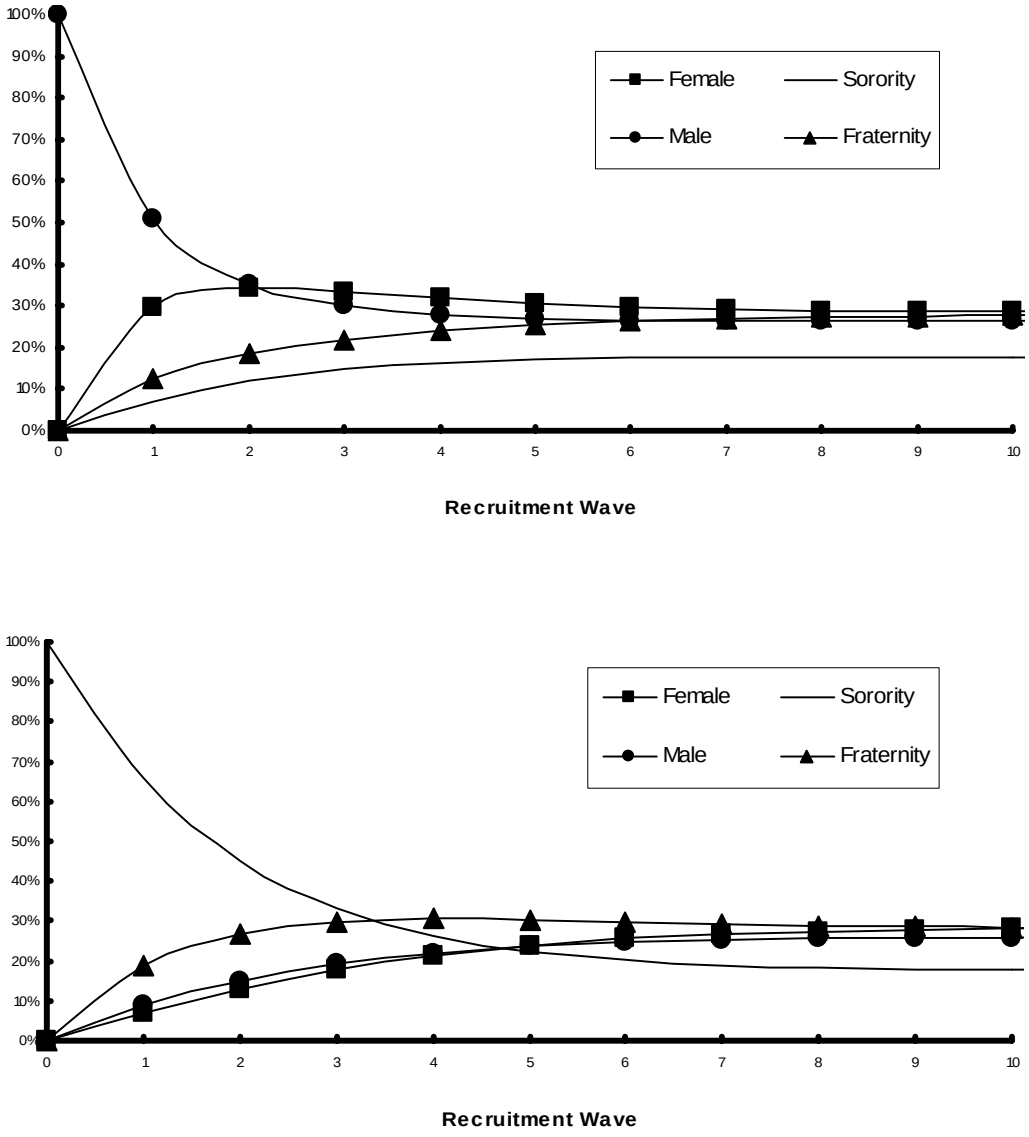



Figure 2: Simulated equilibrium of RDS sample composition. In the top graph the sample begins with 100% non-fraternity males. In the bottom graph, sampling begins with 100% female sorority members. However, the sample reaches equilibrium after only a few waves in both graphs and the subsequent sample composition is the same, regardless of starting point.

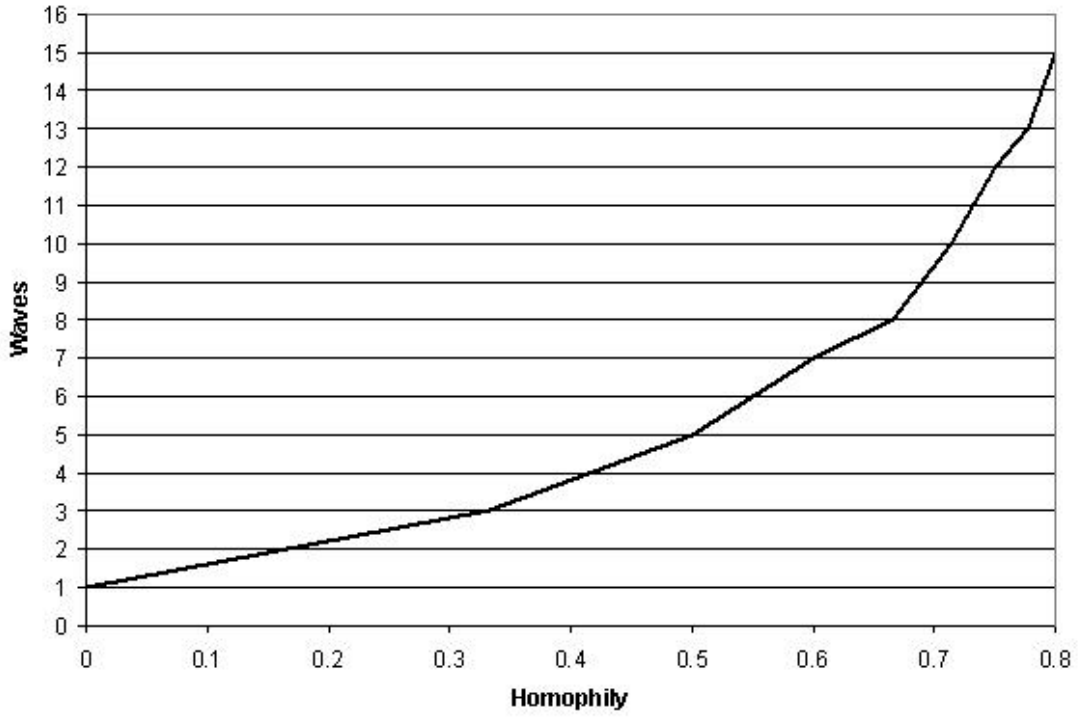



Figure 3: Change in the number of waves required for sample composition to reach equilibrium as homophily increases: A worst case projection based on drawing all seeds from a single group

Fortunately for RDS analysis, homophily levels tend not to be extreme. For example, based on the limited number of currently available studies of US injection drug users (IDUs), homophily tends to be greatest by race and ethnicity, in the 0.3 to 0.55 range. Homophily among US IDUs tends to be lower for HIV risk behavior, such as syringe sharing and condom use, where homophily is in the 0.1 to 0.3, and homophily by HIV status is generally less than 0.1. Therefore, only a modest number of waves would be required for equilibrium to be attained, even if one were to adopt the worst possible strategy for selecting seeds, taking them all from the most insular (i.e., most homophilous) group. Of course, fewer waves are required if the seeds are diverse.

**Recruitment Information**

The second observation upon which RDS is based is that gathering information during the sampling process can provide the means for constructing a sampling frame from which inclusion probabilities can be calculated. This in turn provides the means to verify that population estimates are unbiased and to determine the variability of these indicators.

Recall that in traditional sampling methods such as cluster sampling, construction of the sampling frame comes before the first respondent is selected. In a simple random sample, selection probabilities are equal, and in a stratified sample subgroups of special interest are over-sampled so selection probabilities are unequal. In either case, the sample is pre-stratified because selection probabilities are determined before the first respondent is selected. The effects of stratification are then taken into account when data are analyzed using sampling weights that are equal for a simple random sample and unequal for stratified samples (Ramirez-Valles et al. 2005)

In contrast, in RDS, the sampling frame is created after sampling based on special information gathered during the sampling process. This special information involves three elements:

- Who recruited whom? This provides the basis for controlling for bias introduced by the tendency of subjects to recruit those like themselves. Therefore, an important element in the RDS research design is documenting recruiter/recruit relationships.
- How well connected is each respondent within the target population, that is, what is the subject's personal degree? Information on how many persons each subject knows who fit the eligibility criteria for the study provides the means for controlling for bias toward over-sampling those with larger personal networks.

- Do the recruiter and recruit know one another, or are they strangers? The analytics upon which RDS population estimates are based depends on the recruiter and recruit knowing one another, so the RDS research design includes means for encouraging subjects to recruit those they already know. This includes rewards for recruiters and making recruitment rights scarce through quotas, so valuable recruitment rights will not be wasted on strangers. Asking recruits about their relationship to their recruiter is also useful so recruitments by strangers can be flagged for possible elimination from the data set.

Based on this information, relative inclusion probabilities in the form of sampling weights are calculated using the statistical theory upon which RDS is based. This occurs only after sampling has been completed, a process known as post-stratification. Because RDS does not require a sampling frame before sampling can begin, it can be implemented quickly. Because a sampling frame is available when RDS data are analyzed, the RDS method provides the benefits of other probability sampling methods.

**Assumptions**

It has been shown that if the assumptions upon which RDS is based are satisfied, RDS estimates are asymptotically unbiased (Salganik and Heckathorn 2004). The model is based on five assumptions. The first three specify the conditions under which RDS is an appropriate sampling method:

- Respondents must know one another as members of the target population. Peer recruitment is a feasible sampling strategy only if this condition is satisfied. Consequently, RDS would not be suitable for sampling tax cheats, who can be friends and not know they share membership in that hidden population. On the other hand, it is suitable for sampling

populations linked by a “contact pattern,” such as musicians who perform together or drug users who purchase drugs together.

- Ties must be reciprocal and dense enough to sustain the chain-referral process. For populations linked by a contact pattern or those that form a single community, this is rarely problematic.
- Sampling is assumed to occur with replacement, so recruitments do not deplete the set of respondents available for future recruitment. Consequently, the sampling fraction should be small enough for a sampling-with-replacement model to be appropriate.

The final two assumptions are required by the statistical model on which estimation is based:

- Respondents can accurately report the number of peers they could potentially recruit for the study. Studies of the reliability of network indicators suggest that the RDS network question is one of the more reliable indicators (Marsden 1990); furthermore, the RDS population estimator depends not on absolute but on relative degree, so variations that inflate or deflate the reports in a linear manner have no effect on the estimates. However, violation of this assumption is a source of potential bias (Wejnert and Heckathorn 2008).
- Respondents recruit as though they are choosing randomly from their networks. That is the composition of recruitment is representative of the composition of personal networks. This is based on the expectation that respondents would lack an incentive or ability to coordinate to selectively recruit any particular group. The plausibility of this assumption is enhanced, in part, by appropriate research design. For example, if a research site were located in a high-crime neighborhood, recruiting

> residents of the neighborhood might be easy, but recruiting peers from more comfortable neighborhoods might prove difficult, so sampling would be non-random because it excluded the latter group. However, if research identifies neutral turf in which all potential respondents feel safe, the random recruitment assumption is made more plausible. Similarly, if incentives are offered that are salient to respondents from all income groups (e.g., a choice between receiving a monetary reward and making a contribution to a charity of the respondent's choice), the random recruitment assumption is made more plausible (Ramirez-Valles et al. 2005).

There is no direct way to test if respondents accurately report their number of peers or if they then recruit randomly from that pool. However, it is possible to test if respondents' recruitment patterns accurately reflect their self-report network composition by gathering additional data on personal network composition based on easily identifiable traits, such as gender or race. While, some studies have found strong association between recruitment patterns and self-reported network composition (Heckathorn et al. 2002; Wang et al. 2005), others have found significant differences between self-report network composition and recruitment patterns (Wejnert and Heckathorn 2008; Wejnert in press). Whether these differences are due to a failure of random recruitment or a failure of accurate reporting of network composition requires further research. Methods of testing this and other RDS assumptions are presented in detail in Chapter Two.

### *RDS Estimators*

Over a decade of research has gone into refining and enhancing RDS population proportion point estimators. Table 1 compares seven published RDS estimators and their contributions to RDS theory.

The first RDS estimator, described by Heckathorn (1997), is limited to nominal variables and uses the Markov chain equilibrium proportion, $\widehat{E_X}$ , as the estimator.

$$\widehat{E_X} = \frac{\widehat{S_{YX}}}{\widehat{S_{YX}} + \widehat{S_{XY}}} \tag{1.1}$$

where, $\widehat{S_{XY}}$ is the transition probability (discussed below) from group X to group Y. Heckathorn (1997) also showed that an RDS sample is self-weighting if homophily is uniform across groups. While this estimator paved the way for future RDS estimators it does *not* account for all major sources of sampling bias and is no longer used as a method of population estimation with RDS data.

**RDS I Estimators**

In 2002, Heckathorn introduced the reciprocity model, which assumes a reciprocal relationship between recruiter and recruit. That is, if X recruited Y there is a non-zero probability that Y could have recruited X. Using this model, Heckathorn (2002) presents an improved estimator that controls for differences in homophily and average degree across groups. In the first version of this estimator, 2002A, linear least squares are used to solve a system of over determined equations to calculate estimates for variables of more than two categories. In a second estimator, 2002B, the reciprocity model provides means for calculating multi-category estimates. Under reciprocity, the number of ties or recruitments from group X to group Y equals the number of ties or recruitments from group Y to group X. However, in a finite sample,

this is not always the case. Thus, Heckathorn (2002) improves the estimate of cross-group ties through a process known as *data-smoothing*, in which the number of cross-group recruitments from X to Y and Y to X are averaged such that the recruitment matrix is symmetric. The data-smoothed recruitment matrix is then used to calculate transition probabilities, $\widehat{S_{XY}}$ . The data-smoothing method is recommended over linear least squares because it produces narrower confidence intervals around RDS estimates.

These reciprocity-based estimators provide the foundation for a family of RDS estimators, termed *RDS I* estimators. RDS I estimators employ a two stage estimation process where the data are first used to make inferences about network structure in the form of transition probabilities (based on the recruitment matrix) and estimates of average group degree (based on self-reported degrees). These inferences are then used to calculate a population proportion estimate for each group, $\widehat{P_X^{RDS\,I}}$ ,

$$\widehat{P_X^{RDS\,I}} = \frac{\widehat{S_{YX}}\,\widehat{D_Y}}{\widehat{S_{YX}}\,\widehat{D_Y} + \widehat{S_{XY}}\,\widehat{D_X}} \tag{1.2}$$

where $\widehat{D_X}$ is the estimated average degree of group X. These estimators all control for differences in average degree and homophily across groups and thus differ substantially from the estimator developed by Heckathorn (1997). Unfortunately, the two stage estimation process complicates variance calculations. To date, RDS I estimates rely on a bootstrap algorithm to estimate confidence intervals around the estimate (Heckathorn 2002; Salganik 2006).

Table 1: Comparison of RDS estimators. LLS = Linear Least Squares

| RDS Estimator | Information Employed | Theoretical Foundation | Limitations | Variance Estimation | Distinctive Contribution |
|---|---|---|---|---|---|
| Heckathorn 1997 | Recruitment Matrix | Markov equilibrium | Limited to nominal variables; Degree not accounted for | None | Sample is self-weighting when homophily is uniform |
| Heckathorn 2002A | Recruitment Matrix; Self-Reported Degrees | Reciprocity model with LLS (RDS I) | Limited to nominal variables; Restricted by assumptions | Bootstrap | Controls for differences in degree and homophily across groups |
| Heckathorn 2002B | Recruitment Matrix; Self-reported degrees | Reciprocity model with data-smoothing (RDS I) | Limited to nominal variables; Restricted by assumptions | Bootstrap | Data-smoothing yields narrower confidence intervals than LLS |
| Salganik and Heckathorn 2004 | Recruitment Matrix; Self-reported degrees | Reciprocity model based estimator (RDS I) | Limited to nominal variables; Restricted by assumptions | Bootstrap | Proof estimate asymptotically unbiased; Average group degree estimate |
| Heckathorn 2007 | Recruitment Matrix; Self-reported degrees | Dual-Component estimator (RDS I) | Restricted by assumptions | Bootstrap | Analysis of continuous variables; Controls for differential recruitment |
| Volz and Heckathorn 2008A | Recruitment Matrix; Self-reported degrees | Probability based estimator (RDS II) | Does not control for differential recruitment; Restricted by assumptions | Analytic | Analytically tractable estimator; Analysis of continuous variables |
| Volz and Heckathorn 2008B | Recruitment Matrix; Self-reported degrees | Probability based estimator with data-smoothing (RDS II) | Limited to nominal variables; Restricted by assumptions | Analytic | Convergence between RDS I and RDS II; Controls for differential recruitment |

While Heckathorn (2002) details much of the underlying theory and estimation procedures for RDS I estimation, several improvements have been made. Salganik and Heckathorn (2004) derive an unbiased estimate of average group degree and prove that the RDS I estimator is asymptotically unbiased, which means that bias is on the order of 1/[sample size], so bias is trivial in samples of meaningful size (Cochran 1977). In 2007, Heckathorn developed a dual-component version of the RDS I estimator which calculates a sampling weight based on equation (1.2) that can be applied to the sample proportion to estimate population proportion. This estimator not only controls for differences in degree and homophily across groups, but also separates their effects on the sampling weight into recruitment (homophily) and degree components. The dual component estimator allows for analysis of continuous variables and controls for differential recruitment that occurs if some groups recruit more effectively than others (Heckathorn 2007).

In a study of IDUs, Frost et al. (2006) compare estimates generated using data-smoothing (Heckathorn 2002) and degree adjustment (Salganik and Heckathorn 2004) to those based on unadjusted data and find RDS I estimates are sensitive to differences in the estimation model applied. Such sensitivity to adjustments is not unexpected because both adjustments represent a theoretical improvement in estimation for which there would be no need without sensitivity to them. Consequently, estimates based on Frost et al.'s (2006) "smoothed-adjusted" model, which corresponds to Heckathorn's (2007) dual component estimator are likely the most reliable.

**RDS II Estimators**

Using a probably-based estimation approach, Volz and Heckathorn (2008) introduce a second family of RDS estimators, $\widehat{P_X^{RDS\,II}}$ , termed *RDS II* estimators.

$$\widehat{P_X^{RDS\,II}} = \left(\frac{n_X}{n}\right)\left(\frac{\widehat{D}}{\widehat{D_X}}\right) \tag{1.3}$$

where $n_X$ is the number if respondents in group X, $n$ is the total number of respondents, and $\widehat{D}$ is the overall average degree. Essentially, the estimate is the sample proportion weighted by a correction for network effects. RDS II estimators are calculated directly from the data, removing the middle step of making inference about network structure necessary in RDS I. More importantly, the mathematical approach used to calculate RDS II estimates allows for analytical variance calculation. Currently, there two are versions of the RDS II estimator. The first, 2008A, allows analysis of continuous variables, but does not adjust for differential recruitment. The second, 2008B, uses data-smoothing to adjust for differential recruitment, but can not be used to analyze continuous data (data-smoothing is only applicable to nominal data).

**RDS I vs. RDS II Estimation**

As expected of two unbiased estimators of the same parameter, Volz and Heckathorn (2008) show that when data-smoothing is used, RDS I and RDS II estimators are convergent. Consequently, beyond the mathematical approach used in their calculation, the primary difference between RDS I and RDS II estimation is the method in which estimate variances and confidence intervals are calculated.

Confidence intervals for RDS I are estimated using a specialized bootstrap algorithm (Heckathorn 2002; see also Salganik 2006). The algorithm generates a resample of dependent observations based on the sample transition matrix. That is, if 70% of type X recruitments are other Xs and the current observation is of type X, the algorithm will generate an X as the next observation in the resample with probability 0.7. This process continues until the resample reaches the original sample size. RDS I

estimates are then calculated and the process is repeated until the specified number of re-samples has been reached. Confidence interval tails are then taken from the distribution of these bootstrapped estimates. For example, the upper bound of a 95% confidence interval is defined as the value above which 2.5% of the bootstrapped estimates fall. Consequently, the bootstrap algorithm allows for non-symmetric confidence intervals and does not provide a direct estimate of variance.

Confidence interval bounds for RDS II estimates are based on the RDS II variance estimator (Volz and Heckathorn 2008):

$$Var\left(\widehat{P_X^{RDS\,II}}\right) = \widehat{V}_1 + \frac{\widehat{P_X^{RDS\,II}}^2}{n}\left(\left(1-n\right) + \frac{2}{n_X}\sum_{i=2}^{n}\sum_{j=1}^{i-1}\left(\widehat{S}_{\cdot}^{\,i-j}\right)_{XX}\right) \tag{1.4}$$

where

$$\widehat{V}_1 = \frac{\widehat{Var\left(Z_i\right)}}{n} = \frac{1}{n\left(n-1\right)}\sum_{i-1}^{n}\left(Z_i - \widehat{P_X^{RDS\,II}}\right)^2 \tag{1.5}$$

and

$$Z_i = d_i^{-1}\widehat{D}\,I_X\left(i\right) \tag{1.6}$$

where $d_i$ is the degree of respondent i, $\widehat{S}_{\cdot}$ is the matrix of transition probabilities, and $I_X(i)$ is an indicator function which takes the value 1 if $i \in X$ and 0 otherwise. While the estimate is not unbiased, Volz and Heckathorn (2008) find it closely approximates unbiased estimates of variance in their simulations.

To date, few studies have directly compared the two methods. However, a study by Wejnert (in press), which compares 95% confidence intervals generated by RDS I and RDS II for real data with known parameters, finds both variance estimation

methods lacking, albeit in different ways. That is, confidence intervals based on RDS II are generally wider, more consistent across variables, and more likely to capture population parameters than their RDS I counterparts, however, analysis of design effects suggests RDS II overestimates variance, in some cases by a large amount. Furthermore, the RDS I bootstrap procedure used to estimate confidence intervals was found to underestimate variance, especially for small groups. In Wejnert's (in press) analysis, 95% confidence intervals calculated based on bootstrapped variance fail to capture the parameter more often than the 5% suggested by the interval, while those calculated using RDS II display a capture rate that resembles what would be expected from an ideal variance estimate (Wejnert in press). Chapter Three is based on Wejnert's (in press) analysis.

More generally, computational work testing both RDS I (Salganik and Heckathorn 2004) and RDS II (Volz and Heckathorn 2008) estimators suggests they perform well. Using real data on men who have sex with men, Kendall et al. (2008) find RDS produced a sample with wider inclusion of relevant demographic groups than time-location sampling or other snowball methods. When comparing RDS I estimates to known population parameters, Wejnert and Heckathorn (2008) conclude RDS estimation is reasonable, but not precise. Using two data sets, including that used by Wejnert and Heckathorn (2008), Wejnert (in press) tests both RDS I and RDS II estimates and find both to be reasonably accurate and that problems with confidence intervals described by Wejnert and Heckathorn (2008) are likely due to variance estimation procedures and not point estimation.

### *Social Network Analysis with RDS*

A currently underused feature of RDS data is the presence of network information ideal for analyses of social network structure. RDS has two advantages

that make it especially efficient for social network analysis. First, estimates of homophily and average degree allow inferences on large networks using survey data. Studying large networks with current techniques is problematic; ego-centric samples are unlikely to include connected respondents; database records, such as email networks, often lack important demographic variables; and saturated data are simply impractical for large networks. The second advantage is that every respondent has at least one documented behavioral tie (recruitment) to another respondent in the data. Including respondents' alters in the data allows for analysis of network structure based on private characteristics unknown to a respondent's immediate ties, avoiding what Erickson (1979) calls *masking*, where respondents project their own views onto their friends in self-report studies. RDS also provides greater range of analysis because tie and node characteristics can be collected independently and combined during analysis. Finally, because respondents are only asked information about themselves or their recruiter, who has already provided informed consent through his or her own participation, many ethical human subjects concerns often associated with network analysis are avoided (Kadushin 2005; Klovdahl 2005).

**Average Group Degree**

Salganik and Heckathorn (2004) 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). In addition to providing a correction for degree bias in RDS estimation of categorical variables, the estimator provides a measure of group centrality.

$$\widehat{D_X} = \frac{n_X}{\sum_{i=1}^{n_X} \frac{1}{d_i}} \tag{1.7}$$

where $\widehat{D_X}$ is the average degree of group X, $n_X$ is the sample size of nodes in group X, and $d_i$ is the self reported personal degree of individual *i* (Salganik and Heckathorn 2004).

This estimator can be used to study important network characteristics, such as connectedness and centrality. For example, in a study of New York City aging artists, Spiller et al. (2008) find that artists tend to lose connections to the art community as they age. However, a small proportion of aging artists remain involved in the community and maintain far reaching contact networks, such that even those maintaining less than five network ties likely associated with someone who is very well connected.

**Homophily and Affiliation**

As stated above, network-based samples, like RDS, are biased by the non-random nature of social network ties used to make recruitments. RDS network analysis makes use of this bias to measure a common friendship tendency constraining social network structure: the tendency for individuals to associate with specific alters based on the characteristics of those alters. A special form of this tendency, termed homophily, concerns "the principle that contact between similar people occurs at a higher rate than among dissimilar people" and has been shown to be a powerful mechanism by which affiliations deviate from random mixing (McPherson et al. 2001, p.416). Evidence for the homophily effect is extensive across a wide range of variables. Strong instances of homophily have been found according to race and ethnicity, age, gender, educational aspiration, drug use, musical tastes, political

identification, religion, and behavior (see McPherson et al. 2001 for an extensive review).

RDS homophily can be calculated for any variable in the data set by comparing a standardized measure of the difference between affiliation patterns observed among respondents and the affiliation patterns that would result from random mixing (Heckathorn 2002). Specifically, homophily is calculated from the estimated proportion of in-group ties and that which would be expected from random mixing, in which in-group ties would merely reflect the group's proportional size (Heckathorn 2002).

$$\widehat{H_X} = \frac{\widehat{S_{XX}} - \widehat{P_X}}{1 - \widehat{P_X}} \quad if \quad \widehat{S_{XX}} \geq \widehat{P_X}$$
$$\widehat{H_X} = \frac{\widehat{S_{XX}} - \widehat{P_X}}{\widehat{P_X}} \quad if \quad \widehat{S_{XX}} < \widehat{P_X} \tag{1.8}$$

where $\widehat{S_{XX}}$ is the transition probability of in-group recruitments made by group X, $\widehat{P_X}$ is the estimated proportion of the population contained in group X, and $\widehat{H_X}$ is the homophily of group X. The measure was first introduced by Coleman (1958) as what he termed an index of "inbreeding bias" and later independently derived by Fararo & Sunshine (1964) as part of their work on biased net theory. RDS homophily can be calculated for any partition of categorical variables and ranges from negative one to positive one. Positive homophily indicates a group with disproportionate in-group ties, suggestive of preference. Homophily near zero indicates a non-group, i.e. the variable in question is not of social importance to the network structure. Negative homophily, or *heterophily*, indicates disproportionately few in-group ties, suggestive of avoidance (Heckathorn 2002).

Intermediate levels of homophily are defined in a parallel manner. For example, a homophily of 0.12 means that the respondents form their networks as though 12% of the time they form a tie to another person like themselves, and the rest of the time they form ties through random mixing, that is, forming ties in proportion to population composition. Negative homophilies are defined similarly. For example, a homophily of -0.16 means that the respondents form their networks as though 16% of the time they form a tie to someone unlike themselves, and the rest of the time form network connections in proportion to population composition.

The RDS homophily measure depends on the population proportion of each group, providing a better measure of departure from random mixing than earlier methods, such as Krackhardt and Stern's (1988) E-I index, which depend on the proportion of in-group ties compared to that of out-group ties. In studies where groups represent equal portions of the population these methods are not problematic; however, in populations where group sizes differ, random mixing will generate more ties to individuals in larger groups than smaller groups.

In RDS theory, the homophily estimator reflects the strength of association to one's own group beyond random mixing. A generalization, termed *affiliation*, expresses the strength of association between differing groups, where a positive value for two groups indicates a greater proportion of cross-linking ties than random mixing would produce and a negative value indicates fewer cross-linking ties (Heckathorn 2002). Hence, the affiliation index provides a measure of preference or avoidance for any cell in the matrix. It can measure, for example, not only whether Whites prefer or avoid other Whites (homophily or heterophily), but whether and to what extent they interact with Blacks, Asians, or Hispanics. RDS network measures differ from other indices, which identify groups by structural measures, such as density and transitivity

(Wasserman and Faust 1994), by focusing on actor characteristics and identifying which characteristics significantly influence the network.

$$\widehat{A_X} = \frac{\widehat{S_{XY}} - \widehat{P_X}}{1 - \widehat{P_X}} \quad if \quad \widehat{S_{XY}} \geq \widehat{P_X}$$
$$\widehat{A_X} = \frac{\widehat{S_{XY}} - \widehat{P_X}}{\widehat{P_X}} \quad if \quad \widehat{S_{XY}} < \widehat{P_X} \tag{1.9}$$

where $\widehat{A_{XY}}$ is the affiliation preference of group X for group Y. In calculating homophily, the $\widehat{S_{XX}}$ term is simply the transition probably from group X to itself observed in the data. In calculating the affiliation, RDS' assumption of reciprocity between recruiter and recruit becomes significant for the $\widehat{S_{XY}}$ term. Consequently, data-smoothing is used in calculation of the $\widehat{S_{XY}}$ term in equation (1.9). Note that because data-smoothing does not alter the diagonal entries of the transition matrix, it does not alter calculation of homophily.

These measures can be used to analyze macro level social network structures. For example, in the aging artist study, Spiller et al. (2008) find distinctly different structures between professional and non-professional artists. Affiliation of professional artists is centered on participation in the artistic community whereas affiliation patterns of non-professional artists resemble those of the general population.

In summary, RDS provides a random sample of ties based on behavioral network data, i.e. recruitments (Salganik & Heckathorn 2004), that can be used to make social network inferences at the micro level by comparing characteristics of certain types of ties with others, at the group level through estimates of average group degree, and at the macro level through homophily and affiliation analysis.

## *The Future of RDS*

In little over a decade its effectiveness and ease of use has made RDS the emerging *de facto* method for sampling hard-to-reach populations world wide. RDS data has been collected in hundreds of studies in over 30 nations on six continents. However, while new data sets and sampling lessons continue to emerge, the development and enhancement of methods to statistically analyze such data is limited to a small handful of researchers and RDS specific analytical techniques remain largely underdeveloped. While the first decade of research has been dedicated to optimizing RDS sampling procedures, current research is focused on expanding RDS statistical analysis in three directions: variance estimation, multivariate analysis, and network analysis. Additionally, the uses and applications of WebRDS, especially its potential for very fast sampling and researching online communities, need to be further tested.

### Variance Estimation

Developing an improved estimate of variance is the primary motivation behind the RDS II family of estimators. By recalculating the estimate using a probability based approach, RDS II opened the door for analytical calculation of variance (Volz and Heckathorn 2008). While the variance estimator presented by Volz and Heckathorn (2008) is not without problems (Wejnert in press), it represents the crucial first step toward an analytical variance estimate.

One problem affecting both RDS I and RDS II variance estimation methods is multiple recruitment. The RDS II analytical variance formula assumes a maximum of one recruitment per respondent. Similarly, the RDS I bootstrap method, simulates samples following a chain in which each respondent makes one recruitment. Projects are currently under way to improve both variance measures by removing the single recruitment assumption.

**Multivariate Analysis**

A major limitation of RDS is the lack of RDS specific methods of multivariate analysis. Heckathorn's (2007) dual-component estimator makes an important contribution to future multivariate analysis techniques by splitting the sampling weight and deriving an individual level degree component. However, because the recruitment component is based on group level calculations, the method falls one step short of the holy grail of multivariate RDS analysis: an individual level sampling weight applicable across all variables.

Multivariate analysis with RDS data is currently the most widely anticipated and researched next step for RDS research. Several methods are under development (e.g. Plat et al. 2006; Philbin et al. 2008), however the current recommendation is to apply sampling weights based on the dependent variable as an overall sampling weight. In addition, work by Winship and Radbill (1994) finds that under certain conditions, regression analysis based on unweighted data provides greater precision than analysis using weighted data. Such an approach is employed by Ramirez-Valles et al. (2008) in an RDS study of Latino men who have sex with men. More research is needed to validate these techniques and further develop new multivariate analysis techniques for RDS data.

**Social Network Analysis**

Currently, RDS researchers can easily make inferences regarding group-level network structure and centrality using the RDS degree and homophily/affiliation measures based only the information required for normal RDS analysis (Heckathorn 2002). Unfortunately, this information is greatly underused in many RDS studies, including research in which networks play a vital role in the research topic, such as studies of HIV transmission.

While multiple RDS network inferences exist, the full potential of RDS as a method of network analysis has yet to be developed. Research is currently being conducted to improve the network analysis capacity of RDS. In one project, Heckathorn, Frost, and others are building a simulation environment to understand what information about network structure can be ascertained from the RDS sampling process. By using observed network information to construct a family of model networks with consistent structural features, researchers hope to provide new information about network attributes that can be incorporated into RDS estimates, providing improved variance estimates.

**WebRDS**

The potential of WebRDS has yet to be fully explored. Projects are currently underway using WebRDS to study both online and electronically connected real-world communities. The data and lessons learned from these projects will provide information on the method's ability to sample various populations, the speed with which samples can be collected, and the factors influencing efficiency and efficacy of the method.

**Validation of Existing Techniques**

Finally, empirical and computational testing and validation of analytical techniques are being conducted. For example, researchers are using simulation experiments to explore the performance of the RDS sampling process and RDS estimators on empirical and simulated networks when assumptions about network structure and recruitment behavior are systematically relaxed. Simulating RDS with parameters drawn from real RDS data sets will be used to further refine estimates and guidelines about when RDS can be successfully applied. Additionally, more empirical work in which RDS is applied to known populations and estimates are compared to

true population parameters is needed to confirm that RDS estimation provides valid estimates in practice as well as in theory.

### *Conclusion*

RDS combines an efficient chain-referral sampling method with a statistical method of analysis that corrects for the fact data are collected in a non-random way to provide unbiased population estimates. It has been widely used in the fields of public health and sociology to study hidden populations such as those at risk for HIV, artistic communities, and impoverished groups. Additionally, RDS has been shown to be an effective method of analyzing social network structure and has been successfully implemented as an online sampling method. Further information, along with specialized software for conducting RDS analysis, is available, free of charge from the RDS website: RespondentDrivenSampling.org.

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.

Amaral, L.A.N, A. Scala, and H.E. Stanley. 2000. "Classes of Small World Networks." *Proceedings of the National Academy of Sciences* 97: 11149-11152.

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

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

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

Des Jarlais, Don C., Kamyar Aresteh, Theresa Perlis, Holly Hagan, Abu Adbul-Quader, Douglas D. Heckathorn, Courtney McKnight, Heidi Bramson, Chris Nemeth, Lucia V. Torian, and Samuel R. Friedman. 2007. "Convergence of HIV seroprevalence among injecting and non-injecting drug users in New York City." *AIDS* 21: 231-235.

Erickson, B.H., 1979. "Some Problems of Inference from Chain Data." *Sociological Methodology* 10: 276-302.

Fararo, T. J., Sunshine, M.H., 1964. A Study of a Biased Friendship Net. Syracuse University Youth Development Center. Syracuse, NY.

Frost, Simon D. W., Kimberly C. Brouwer, Michelle A Firestone Cruz, Rebeca Ramos, Maria Elena Ramos, Remedios M. Lozada, Carlos Magis-Rodriquez, and Steffanie A Strathdee. 2006. "Respondent-Driven Sampling of Injection Drug Users in Two U.S.-Mexico Border Cities: Recruitment Dynamics and

Impact on Estimates of HIV and Syphilis." *Journal of Urban Health* 83: i83-i97.

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

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

------. 2007. "Extensions of Respondent-Driven Sampling: Analyzing Continuous Variables and Controlling for Differential Degree" *Sociological Methodology* 37: 151-207.

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

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.

Kadushin, C., 2005. "Who benefits from network analysis: ethics of social network research." *Social Networks* 27: 139-153.

Kendall, Carl, Ligia R. F. S. Kerr, Rogerio C. Gondim, Guilherme L. Werneck, Raimunda Hermelinda Maia Macena, Marta Kerr Pontes, Lisa G. Johnston, Keith Sabin, and Willi Farland. 2008. "An Empirical Comparison of Respondent-Driven Sampling, Time Location Sampling, and Snowball Sampling for Behavioral Surveillance in Men Who Have Sex with Men, Fortaleza, Brazil." *AIDS and Behavior* 12: s97-s104.

Kissinger, Patricia, Nicole Liddon, Lisa Longfellow, Erin Curtin, Norine Schimdt, Oscar Salinas, Jaun Cleto, and Douglas Heckathorn. 2008. "HIV/STI Risk

Among Latino Migrant Workers in New Orleans Post-Hurricane Katrina." Presented at the Annual CDC STD Prevention Conference, Chicago, IL.

Klovdahl, A.S., 2005. "Social network research and human subjects protection: Towards more effect infectious disease control." *Social Networks* 27: 119-137.

Krackhardt, D. and R. Stern. 1988. "Informal Networks and Organizational Crises: An Experimental Simulation." *Social Psychology Quarterly* 51: 123-140.

Malekinejad, Mohsen, Lisa G. Johnston, Carl Kendall, Ligia R. F. S. Kerr, Marina R. Rifkin, and George W. Rutherford. 2008. "Using Respondent-Driven Sampling Methodology for HIV Biological and Behavioral Surveillance in International Settings: A Systematic Review." *AIDS and Behavior* 12: 105-130.

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

McPherson, M., Smith-Lovin, L., Cook, J.M., 2001. "Birds of a feather: Homophily in Social networks." *Annual Review of Sociology* 27: 415-444.

Newman, Mark, Duncan Watts, and Stephen Strogatz. 2002. "Random Graph Models of Social Networks." *Proceedings of the Academy of Sciences* 99: 2566-2572.

Philbin, Morgan, Robin A. Pollini, Rebecca Ramos, Remedios Lozada, Kimberly C. Brouwer, Maria Elena Ramos, Michelle Firestone-Cruz, Patricia Case, Steffanie A. Strathdee. 2008. "Shooting Gallery Attendance Among IDUS in Tijuana and Ciudad Juarez, Mexico: Correlates, Prevention Opportunities, and the Role of the Envirnoment." *AIDS and Behavior* 12: 552-560.

Platt, Lucy, Natalia Bobrova, Tim Rhodes, Anneli Uuskula, John V. Parry, Kristi Ruutel, Ave Talu, Katri Abel, Kristina Rajaleid, and Ali Judd. 2006. "High HIV Prevalence Among Injecting Drug Users in Estonia: Implications for Understanding the Risk Environment." *AIDS* 20: 2120-2123.

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

Ramirez-Valles, Jesus, Dalia Garcia, Richard T. Campbell, Rafael M. Diaz, and Douglas D. Heckathorn. 2008. “HIV Infection, Sexual Risk Behavior, and Substance Use Among Latino Gay and Bisexual Men and Transgender Persons.” *American Journal of Public Health* 98: 1036-1042.

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.

Spiller, Michael W., Douglas D. Heckathorn, and Joan Jeffri. 2008. “The Social Networks of Aging Visual Artists.” In *Above Ground: Information on Artists III: Special Focus on New York City Aging Artists.* Research Center for Arts and Culture, p. 29-69, New York.

Volz, Erik, and Douglas D. Heckathorn. 2008. “Probability-Based Estimation Theory for Respondent-Driven Sampling.” *Journal of Official Statistics* 24: 79-97.

Wang, Jichuan, Robert G. Carlson, Russell 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.

Wasserman, Stanley and Katherine Faust. 1994. *Social Network Analysis*. Cambridge University Press, Cambridge, MA.

Watts, Duncan. 1991. *Small Worlds: The Dynamics of Networks between Order and Randomness*. Princeton University Press, Princeton, NY.

Wejnert, Cyprian. In press. "An Empirical Test of Respondent-Driven Sampling: Point Estimates, Variance, Measures of Degree, and Out-of-Equilibrium Data." *Sociological Methodology*.

Wejnert, Cyprian and Douglas D. Heckathorn. 2008. "Web-Based Networks Sampling: Efficiency and Efficacy of Respondent-Driven Sampling for Online Research." *Sociological Methods and Research* 37: 105-134.

Winship, Christopher and Larry Radbill. 1994. "Sampling Weights and Regression Analysis." *Sociological Methods and Research* 23: 230-257.

# CHAPTER 2

# WEB-BASED RESPONDENT-DRIVEN SAMPLING

## *Introduction to Data Used Throughout This Document*

Empirical analyses conducted in the remaining chapters will be based on two samples of Cornell University undergraduates collected using Web-Based Respondent-Driven Sampling (WebRDS) in 2004 and 2008. WebRDS is an online variant of RDS in which respondents complete an internet survey and recruitment occurs via email. Figure 4 and Figure 5 show WebRDS recruitment chains for the 2008 and 2004 samples, respectively, shape coded for gender and color coded for college within the university. The 2008 sample is made up of 369 recruitments and nine seeds for a total sample size of 378 respondents. The 2004 sample includes 150 recruitments and nine seeds for a total sample size of 159 respondents. The university is a highly selective, research school of over 13,000 undergraduate students located in an isolated rural town in central New York State. At the undergraduate level, the university is divided into seven colleges, three colleges are considered private and four are technically part of the New York State University (SUNY) system. While some minor differences are observed in the proportion of students choosing not to disclose their race to the university between 2004 and 2008, there is no reason to believe the population of students in 2004 differed significantly from that of 2008. The four year interval between surveys does, however, ensure a near complete turnover of students between 2004 and 2008. Both studies where conducted in the spring semester.

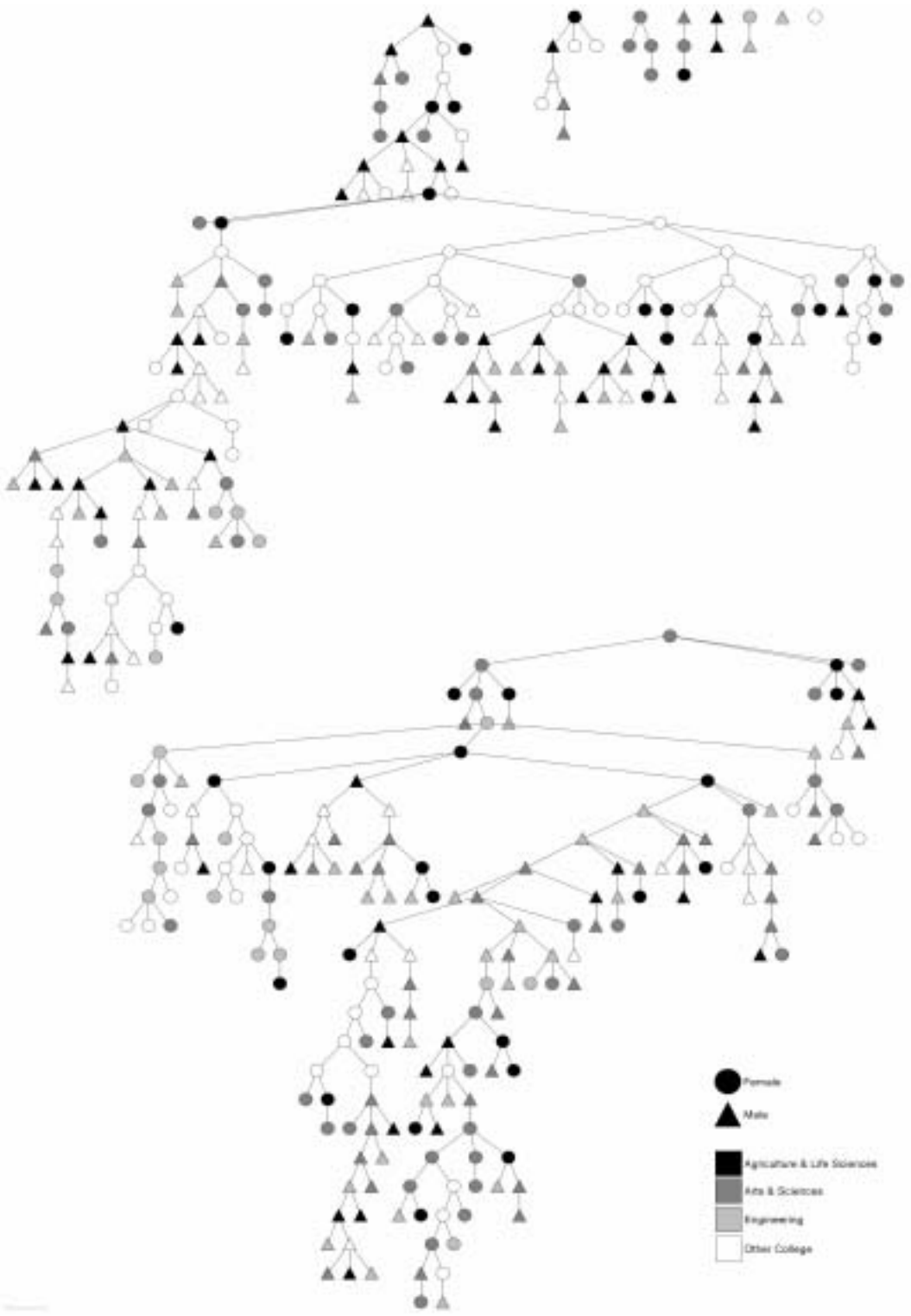



Figure 4: 2008 sample recruitment chains. Colors indicate college within the university, shapes indicate gender.

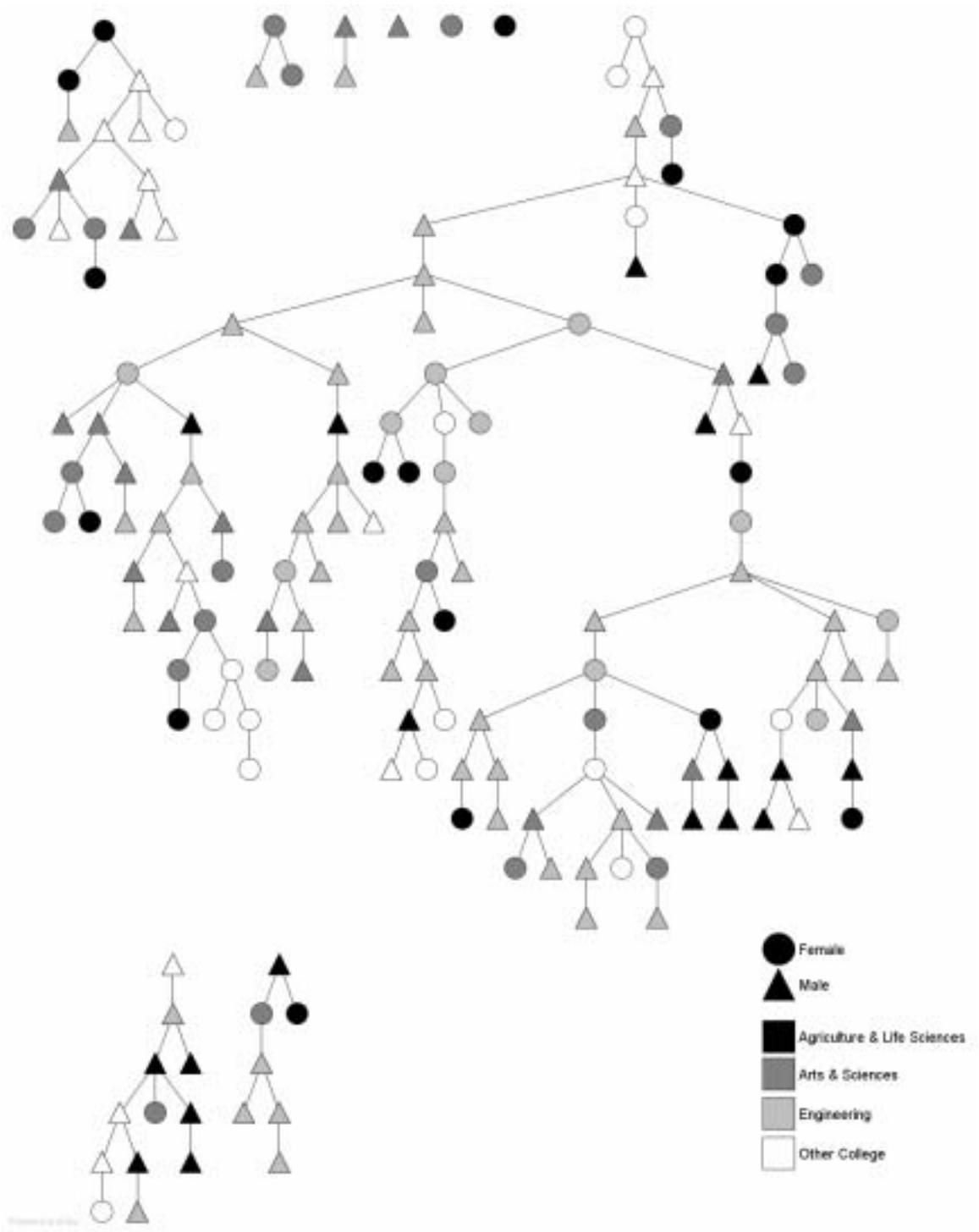



Figure 5: 2004 sample recruitment chains. Colors indicate college within the university, shapes indicate gender.

### *Web-Based RDS Implementation*

WebRDS implementation consists of two parts: setup and implementation. First, the study needs to be set up using specialized WebRDS software. This includes building an online questionnaire, writing a recruitment email, and setting WebRDS specific options such as the number of coupons given to each respondent for recruitment and setting the maximum sample size. Finally, as with all RDS studies, a small set of seeds need to be selected for participation. Seed selection for WebRDS studies does not differ from that of other RDS studies; seeds should be comprised of a relatively small group of highly motivated, well connected or respected members of the target population who are representative of the diversity within the target population. Once an appropriate pool of seeds has been informed about the project and agreed to participate, the WebRDS study can begin.

Beginning with the seeds, each respondent receives a recruitment e-mail officially informing them of the project's purpose, compensation, and recruitment process. The recruitment e-mail contains 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 each recruitment e-mail is automatically entered into the survey, in the manner of an automatic login to a secure website. After following a link to the survey, each respondent logs in and is administered a questionnaire.

Questionnaire completion results in three automated actions by the software. First, the respondent's data is downloaded into both a master data set and multiple backup forms. Second, both the email serial number and the respondent's login ID are 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, are sent to the respondent. The respondent is then asked to forward each of these e-mails to one potential recruit who meets the inclusion criteria.

Because these e-mails contain serial numbers, only one respondent can be recruited by each e-mail. This procedure is followed for other seeds and recruits. The only difference between seeds and recruits is that seeds are recruited by the administrator (and therefore had no recruiter) while recruits—many of whom became recruiters themselves—are recruited by other respondents

Finally, once participation is complete, compensation is distributed to respondents based on the number of recruitments each respondent makes. This can be done either in person by setting up a distribution site or through electronic means such as Paypal or online gift certificates.

### *Web-Based RDS Implementation at Cornell University*

#### Seed Selection

Each of the two Cornell University samples used nine seeds. In the 2004 sample four seeds 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. 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 2004 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. Seeds for the 2008 sample were selected as part of a demographically diverse group, taking into account race, gender, year in college, and college within the university from three sections of an undergraduate sociology class. Seeds were identified as follows: the researcher provided a brief description of the project during class and

passed around a sign up sheet where any interested students provided, along with an email address, information about their race, gender, year, and college. A diverse subset of this list was then contacted, through email, and asked if they were still interested in the study. Any students replying to this email were included as seeds. Students not replying were not included in the study nor were they contacted again. In this way, 2008 seeds were considered both diverse and motivated study participants.

Seventy-four percent of the 2004 data originate from a single seed. This "super seed" was a white, female, Hotel 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 seed is positively related to its degree. The 2008 sample has two such "super seeds", one producing forty-nine percent of the sample and the producing forty-seven percent. However, these seeds differ from the norm because they have approximately median degree by multiple measures compared to other seeds. Only fourteen percent of the 2008 sample originates from the remaining seven seeds.

**WebRDS Software and Recruitment Quotas:**

Specialized pilot software was written specifically for the 2004 survey. As part of the 2008 project, specialized WebRDS software was written and combined with Web Survey Toolbox[1], a freely available, open-source online survey builder, to provide customizable WebRDS software.

[1] Available for download at: http://sourceforge.net/projects/jspsurveylib/

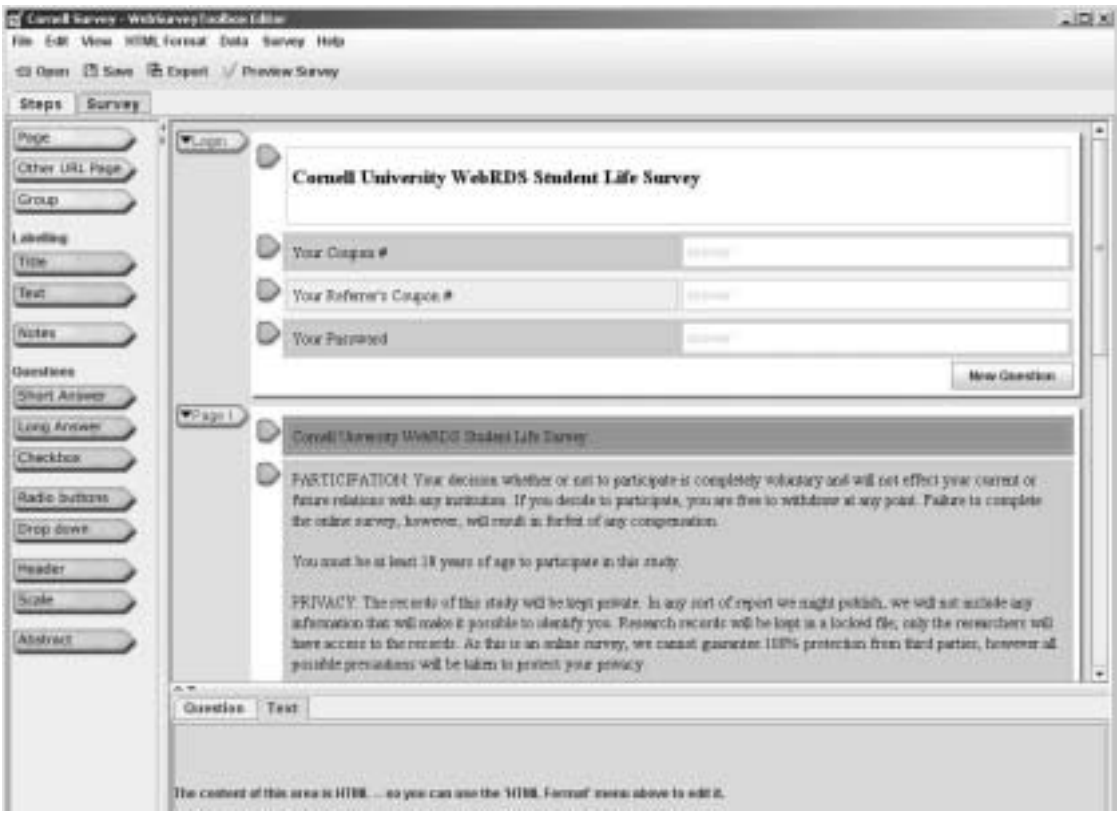



Figure 6: Building a WebRDS survey questionnaire with WebSurvey Toolbox.

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 these studies the target sample was 150 in 2004 and 370 in 2008. In both studies, respondents were given three recruitment coupons. Typically, this is sufficient to support the development of robust chains while permitting the growth of long chains (Heckathorn 1997).

Overall, 55.3% (n=88) of the 2004 respondents recruited peers for the study. Similarly 51.1% (n=193) of the 2008 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).

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 non-numerical 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.

**Efficiency and Ease of Use:**

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 2004 respondents reported that the survey resulted in a brief “fad” among their peers. For the entire duration of both studies, the administrator received no questions or comments from participants. A trial run of the 2004 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 actual surveys.

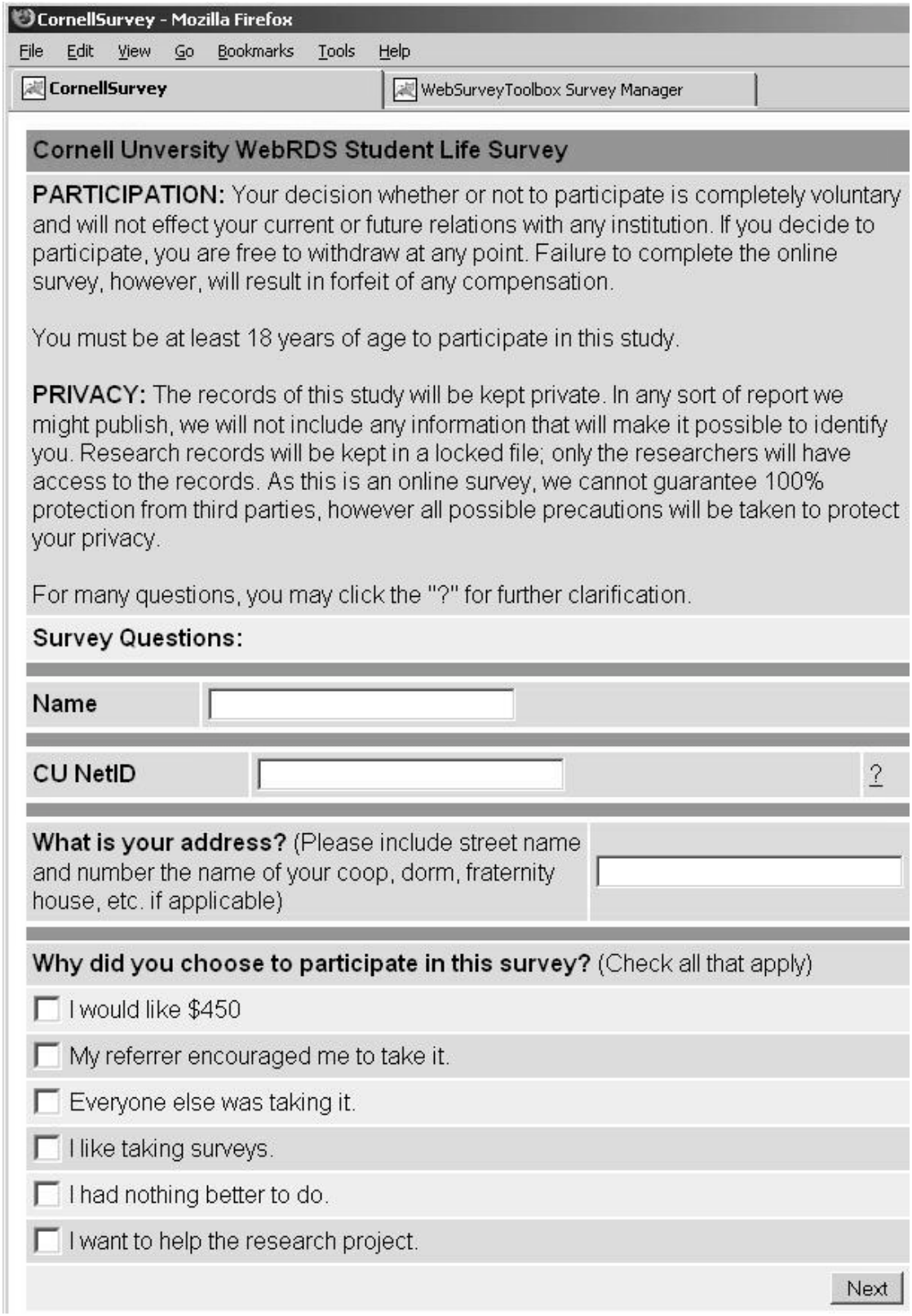



Figure 7: Screenshot of page 1 of the 2008 Cornell survey

For the researcher, the cost of conducting a WebRDS study is relatively minimal and consists primarily in compensating respondents for participation and recruitment. In 2004, 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 was $25 per respondent. This level of compensation was found to be too large and often interpreted as a potential scam, likely because it was confused with SPAM[2] e-mails offering implausible sums of money for minimal effort. Wejnert and Heckathorn (2008) speculate that $5 for completing the survey and $10 for recruitment would have been just as effective, or perhaps more so, in soliciting respondents. Following this recommendation, a more economic, lottery based compensation scheme was initially used in the 2008 sample. Under the lottery, respondents were awarded one lottery entry for participation and an additional entry for each successful recruitment. The lottery consisted of 12 $450 awards (one award for every 84 entries). Consequently, a respondent who only competes the survey would receive a 1 in 84 chance of winning while a respondent who participates and makes the maximum three successful recruitments would have a 1 in 21 chance of winning $450. The average per respondent cost of this scheme is about $11. Unfortunately, the lottery scheme proved an ineffective incentive and was replaced with a traditional scheme were respondents could earn $10 for participation and $5 for each recruitment with a maximum possible compensation of $25 and an

[2] Although still problematic, Web 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.

average cost of $15 per respondent to the researcher. The effects of a lower incentive in 2008 as well as the failed lottery method on sampling time are discussed below.

As a final note on respondent compensation, approximately 20% of 2004 respondents and 25% of 2008 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, and further reducing WebRDS costs. For example, in 2008 the actual compensation cost was approximately $13 per respondent, including a single $450 payment made to one of seven initial respondents choosing to remain in the lottery. Excluding this failed lottery tax, the 2008 compensation cost would have been approximately $8 per respondent.

Beyond the cost of respondent compensation, WebRDS requires minimal additional resources. 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 monitor recruitment and 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.

### *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 I consider the assumptions required by RDS theory (discussed in Chapter One) and test them using three variables: gender, race, and college within the university. Because it has the smaller sample size, I focus first on the 2004 sample.

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 X is a member of Y's pool of potential recruits, then Y must also be a member of X's pool of potential recruits. While the survey 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 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 in 2004 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 (Bollobas 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 degree. A discussion of the difficulties of accurately measuring personal degree (assumption four) is presented in Chapter Three as are results comparing multiple degree measures for both samples. 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).

Table 2: Chi-squared test for random recruitment by gender in 2004 data.

**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 | 2.307115 | 3.063472 | 5.371 |
| Female | 0.040757 | 0.03793 | 0.079 |
| | | Statistic | 5.449 |
| | | p-value: | 0.02 |

I 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 $\chi^2$ goodness of fit test. Table 2 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 (Table 3), 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 the 2004 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. It is important to note that these methods are used to test whether the random recruitment assumption was

not met. The p-value, therefore, can not be interpreted as the probability that recruitment was random.

Table 3: Chi-squared test for random recruitment by race in 2004 data.

**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 | 71.7 |
| Asian | 25.38 | 35.74 | 5.43 | 66.6 |
| Other | 63.2 | 23.05 | 27.1 | 113 |
| | | | | 252 |

**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 |

The self-report to recruitment $\chi^2$ comparison is useful for visible traits, but is not well suited for the non-visible 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 X to Y to the recruitments from Y to X (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 2007). 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 in-group recruitment, such as over recruiting of males by males, can not be explored through comparison of cross-recruitment counts.

Similar analyses suggest assumptions one through three are similarly met in 2008. A discussion of the difficulties of accurately measuring personal degree (assumption four) is presented in Chapter Three as are results comparing multiple

degree measures for both samples. The final RDS assumption is that recruitment patterns reflect personal network composition, such that respondents recruit as though they were selecting randomly from their personal networks (Heckathorn 2002). Following the same procedure outlined above, I find recruitment does not reflect self-reported personal network composition of gender $\left(\chi_1^2 = 42.5, p < 0.000\right)$ or race $\left(\chi_9^2 = 249.5, p < 0.000\right)$ in the 2008 data either. It is impossible to know if this result is due to a failure of assumption five or inaccuracy in self-report network compositions. I leave development of a test of random recruitment that is both powerful and widely applicable open for future research.

### *Analysis of Sampling Speed*

**2004 Sample**

The 2004 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.

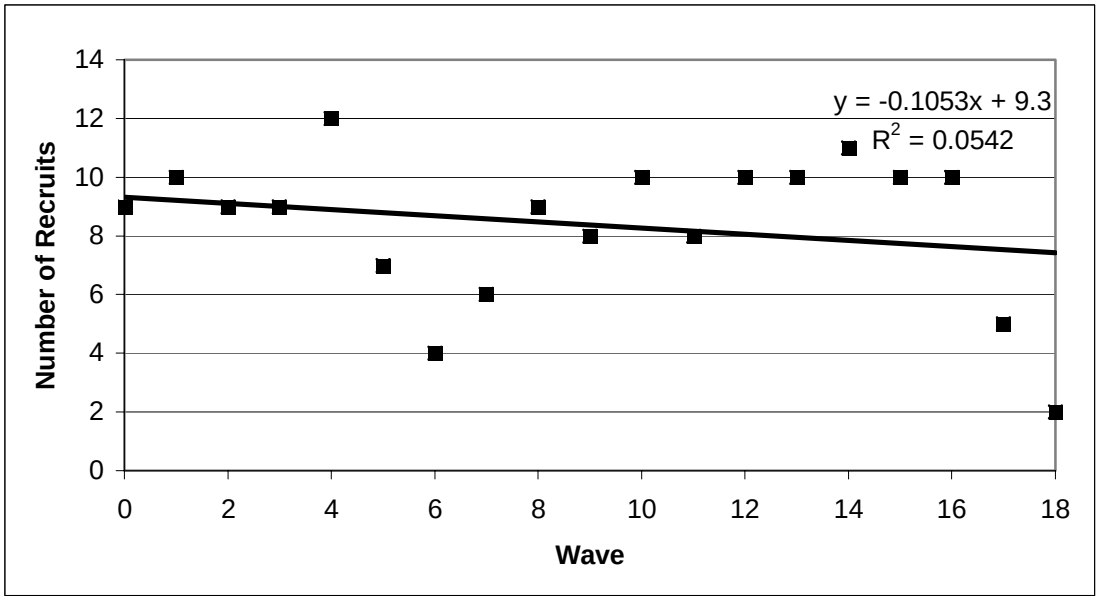



Figure 8: Number of overall recruitments during each wave. Seeds are counted as waves zero.

To quantify the sampling speed, I 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, I 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 8 shows recruitments by wave for the 2004 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 the 2004 sample, only one chain had more than six waves, corresponding to the dip observed in Figure 8 at wave six, at which point all other chains have died out and only the recruitments from the large chain remain. Figure 9 shows recruitments by wave for only the large chain. Here an exponential model fits the data well ($R^2 = .8617$).[3]

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 10 shows recruitment by time of day by day and total recruitment by day. Not

[3] Waves 17 and 18 are excluded because they were cut short by the end of the study and thus 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.

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 non-sampled 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.

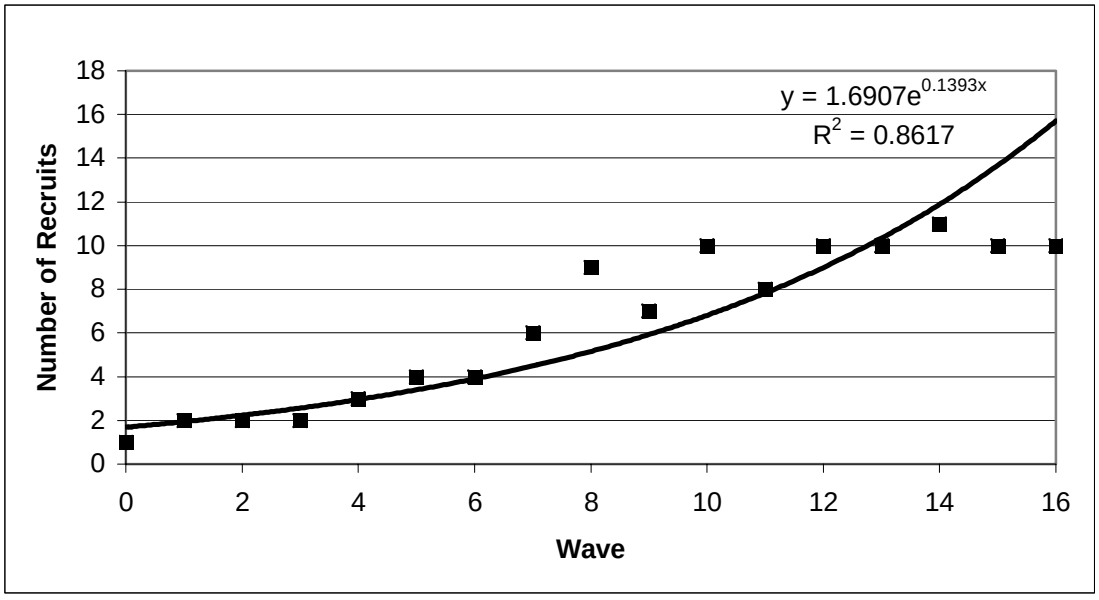



Figure 9: Number of recruitments made in each wave of the large recruitment chain. Seeds counted as wave zero.

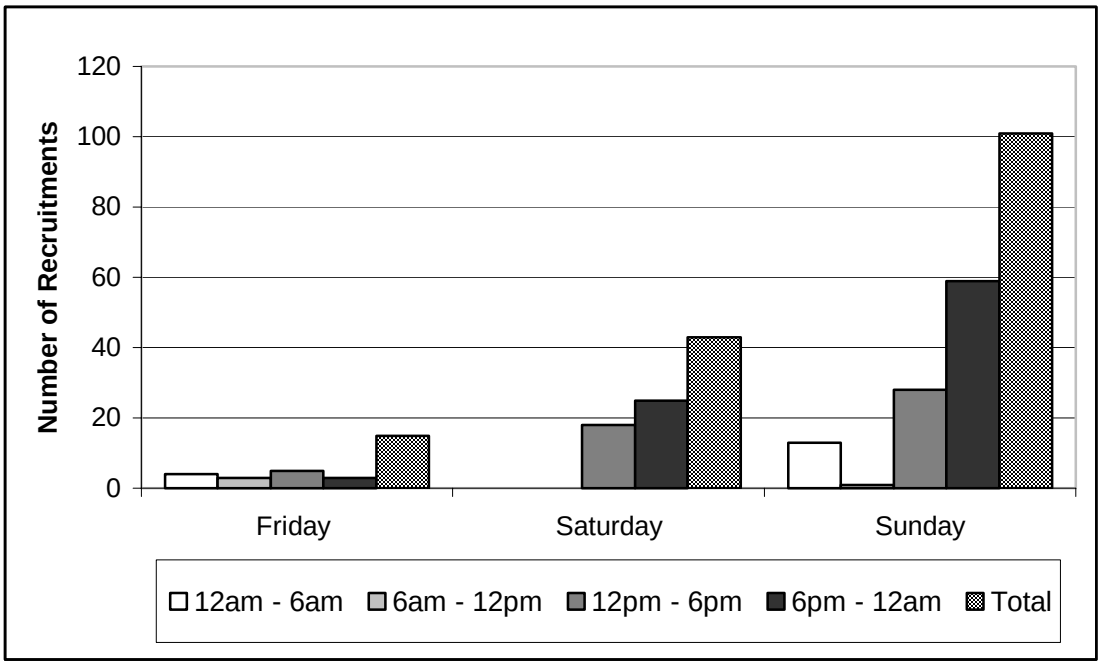



Figure 10: Frequency of recruitment by day and time.

**2008 Sample**

Understanding the sampling speed of the 2008 study is more difficult. While the 2004 study was able to sample 159 students in 72 hours, the 2008 study remained open for nearly ten weeks before the desired sample size of 378 respondents was met. However, consistent with results from 2004, over fifty percent (n=201) of 2008 respondents were surveyed in the final 96 hours of the study. Figure 11 shows a weekly summary of events and number of surveys completed.

Weekly Timeline of 2008 Survey Recruitment and Procedures

| | |
|---|---|
| Week 1: | Sampling begins using lottery incentive system<br>Recruitment coupons valid for 1 week<br>10 seed emails sent out<br>3 of 10 seeds complete survey<br>15 surveys completed (including 3 seeds) |
| Week 2: | 6 additional seeds contacted (incentive preference noted)<br>Incentive changed from lottery to traditional scheme<br>Office hours for incentive pickup begin<br>6 of 6 additional seeds complete survey<br>8 surveys completed (including 6 seeds) |
| Week 3: | All respondents informed of new incentive scheme and procedures<br>22 surveys completed |
| Week 4: | Coupon expiration removed<br>All unused coupons unexpired<br>1 survey completed |
| Week 5: | 8 surveys completed |
| Week 6: | Spring Break, campus closed<br>2 surveys completed |
| Week 7: | All respondents contacted regarding unexpired/unused coupons and pending payments<br>55 surveys completed |
| Week 8: | 26 surveys completed |
| Week 9: | Upperclassmen offered extra incentive ($10) to recruit within 48 hours<br>3 upperclassmen recruit and earn extra incentive<br>69 survey completed |
| Week 10: | Final 77 hours of study<br>172 surveys completed |

Figure 11: Weekly timeline of 2008 survey recruitment, procedural changes, and academic events.

As discussed above, compensation for the 2008 study began as a lottery. However this incentive was abandoned after one week of sampling produced 15 completed surveys and participation by only three of 10 initial seeds. At the start of week two, six additional seeds were contacted, the majority of whom reported preference for a more traditional incentive, and the incentive was changed to the pay-for-participation scheme discussed above. At the start of week three, office hours were scheduled for incentive distribution and all previous respondents were informed of the

incentive change by email. Incentives were distributed several times a week during the duration of the survey to provide legitimacy for the project. I.e. a student is considered more likely to participate if his recruiter has cash in hand to prove the study is legitimate. The change appeared promising and led to 22 completed surveys in week three. However participation declined to a trickle in weeks four and five, which led up to spring break in week six.

Spring break, a nine day period during which the university shuts down and nearly all undergraduates leave, posed a strong possibility of stopping the survey completely. This was especially likely given that recruitment coupons had been set to expire one week after being sent to a respondent. To combat this problem, the expiration time was removed from all outstanding and future coupons and all respondents with outstanding coupons were contacted via email on the first day after break (week seven) reminding them about the study. These measures resulted in 81 completed surveys during weeks seven and eight.

During week eight preliminary analysis of data revealed an over sample of first year students (54%). In week nine all upperclassmen with outstanding coupons were offered an extra $10 incentive to recruit within 48 hours. While only three respondents qualified for the incentive, this was enough to reduce to proportion of freshmen recruited after this incentive offer to 34 percent of new respondents.

The observed sampling speed in 2008 differs sharply from that observed in 2004 and requires further discussion. Specifically, what lead to such stark differences between the sample and what factors can be used by researchers to control the rate at which WebRDS sampling occurs.

While the studies occurred four years apart, they were conducted at the same time of year and on the same population, thus it is unlikely the 2008 sample population was significantly different from that in 2004 in a manner that would affect

sampling speed. The largest contributing factor is likely the difference in incentive. First the lottery system was not effective at generating recruitment. Second, the maximum incentive a respondent could earn in 2004 ($55) was more than double that one could earn in 2008 ($25). Another potentially significant difference was the incentive structure. In 2004 a higher incentive was offered for recruitment ($15) than for completing the survey ($10) while in 2008 respondents earned more for completing the survey ($10) than for recruitment ($5). Offering a high incentive for recruitment is beneficial in several ways. First, it shifts the focus of incentive minded participants onto persuading their recruits to participate and more successfully promotes the norm of participation discussed Wejnert and Heckathorn (2008). Furthermore, because each respondent completes one survey and has one recruiter, the study can advertise a higher potential incentive for the same cost. For example, if the 2008 study had offered $5 for survey completion and $10 for recruitment, the maximum incentive a respondent could earn would be $35 (compared to $25 under the current scheme) while keeping the average cost per respondent the same ($15). Offering such an incentive was considered, but ruled out for fears students would be hesitant to complete a survey for $5. Thus, while offering a higher incentive is likely the best way to increase sampling speed, offering a higher incentive for recruitment at the expense of the participation incentive may provide a cost efficient method of accomplishing a similar goal. More research is needed to confirm or disconfirm these hypotheses.

### *Conclusion*

WebRDS is primarily limited by the requirement that individuals in the target population have frequent access to 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 him or herself. 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. This way, the effort required to self-recruit will appear to be larger than the return from self-recruitment.

Despite these limitations, WebRDS potentially 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, WebRDS 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 non-infected 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, WebRDS could be used to draw the control sample quickly and thus accelerate treatment and containment measures.

Third, WebRDS is currently being developed as a way to study online communities. As use of the internet expands, virtual communities of individuals known to each other only as online screen names are becoming ever more relevant to a wide range of Sociological and health outcomes. For example, many high school aged bi-/gay curious males are turning to internet chat rooms and other anonymous forms of virtual interaction to gain information bi-/gay lifestyles and relationships. WebRDS provides a way to access and study such communities that are not accessible, even to members, offline.

Finally, WebRDS can be used in conjunction with online networking sites, such as Friendster or Facebook, to study social networks and the propagation of information through them. While the existence of such sites and their direct data on social connections has proven invaluable to network researchers, the data suffer from a high degree of false positives. That is, many individuals are listed as “friends” of others who they may not know currently, directly, or at all. By applying WebRDS to such networks, researchers guarantee that only active ties are being sampled. RDS adjustments can then be applied to the data and comparisons to the complete network data can be made to gain a richer understanding of the network structure and how information, in the form of recruitment, propagates through it.

REFERENCES

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.

------. 2007. “Extensions of Respondent-Driven Sampling: Analyzing Continuous Variables and Controlling for Differential Degree” *Sociological Methodology* 37: 151-208.

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.

Ramirez-Valles, Jesus, Douglas D. Heckathorn, Raquel Vázquez, Rafael M. Diaz, and Richard T. Campbell. 2005. “Evaluating Respondent-Driven Sampling: Response to Heimer.” *AIDS and Behavior* 9: 403–408.

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.

Wejnert, Cyprian and Douglas D. Heckathorn. 2008. “Web-Based Networks Sampling: Efficiency and Efficacy of Respondent-Driven Sampling for Online Research.” *Sociological Methods and Research* 37: 105-134.

# CHAPTER 3

# EMPIRICAL TEST OF RESPONDENT-DRIVEN SAMPLING ESTIMATION

## *Introduction*

While RDS estimators have been shown to be asymptotically unbiased computationally and analytically, critics have questioned the plausibility of meeting RDS assumptions with real data (Heimer 2005) and suggested that design effects of RDS studies maybe impractically high (Goel and Salganik 2008). This chapter analyzes RDS estimates calculated for a known population. By focusing on a known population, it is possible to compare RDS estimates to true institutional parameters and compare several suggested methods of analyzing RDS data.

## *Challenges and Concerns for Respondent-Driven Sampling*

While RDS has been used successfully to study a wide range of hidden populations and estimates have been shown to be unbiased analytically and computationally, questions remain as to whether RDS theory and assumptions can be realistically applied to real data. Such questions include: Is variance estimation accurate? Can assumptions about random recruitment and accurate degree reporting be met? What should be done with out-of-equilibrium data?

### Design Effects and Variance Estimation

Variance estimation for RDS estimates remains largely underdeveloped and has been described by some as “the new frontier” for RDS researchers. Unfortunately, because successive observations in RDS are not independent (Heckathorn 1997), RDS variance is difficult to estimate. To date, few studies of RDS design effects, which measure increase in variance due to sampling method, have been conducted. After comparing RDS confidence interval widths based on the RDS I bootstrapping

technique to expected interval widths under a simple random sample design (SRS) with the same proportions, Salganik (2006) recommends RDS samples be at least double that which would be required for a comparable SRS design, consistent with design effects greater than two. Using the same method, Wejnert and Heckathorn (2008) report an average estimated design effect of 3.14 in their study of university students. However, using simulated data and the RDS II estimator, Goel and Salganik (2008) find RDS design effects may reach above 20, an outcome that suggests RDS analysis may produce essentially random estimates[4].

**Degree Estimation**

Measuring degree for RDS analysis presents three challenges.

First, according to Salganik and Heckathorn (2004) RDS respondents are chosen with probability proportional to degree, inflating the sample arithmetic mean degree above the population mean degree. Salganik and Heckathorn (2004) 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, shown in equation (1.7), is used to correct for degree bias in RDS estimation of categorical variables.

Second, RDS theory assumes that respondents can accurately report their degree. While studies of degree indicator reliability suggest RDS style indicators are among the more reliable (Marsden 1990), this assumption is not without controversy.

[4] Goel and Salganik (2008) do not use the RDS II variance estimator. Instead they calculate the RDS II estimate for simulated RDS data and observe the estimate's variability over repeated trials, thus their results apply to the point estimate and not the variance estimate.

Self-report data on individual degree is often limited by poor respondent recall and research comparing self-report degree indicators has had limited success (McCarty et al. 2001; Bell et al. 2007). Additionally, ambiguous terms increase individual level variation in responses. For example, self-reported data on friendship closeness are problematic because the distinction between "friend" and "close friend" may vary across individuals and groups (Fischer 1982). To reduce self-report error, RDS degree questions define interpersonal associations behaviorally within a temporal frame by asking the number of individuals who meet a specified standard with whom the respondent has engaged in a specified behavior over a short period of time. For example, "How many university undergraduates do you know personally (i.e., you know their name and they know yours, and you have interacted with them in some way in the last 14 days)?" While careful question wording likely reduces self-report error in degree estimation, it is unclear how large this reduction is. Fortunately, because both RDS estimator equations (equations (1.2) and (1.3)) include measures of degree in the numerator and denominator, they rely on relative, not absolute, degree reports. Thus, if respondents uniformly inflate or deflate degree, the estimator is unaffected.

Another method for reducing respondent-recall error could be to solicit information for which the respondent does not need to rely on memory alone and use this information as a proxy for his or her degree. Many electronic means of communication, such as cell phones, store information on users' contacts. In these cases, the user can simply look up the number of his or her contacts, without relying on memory. Of course, such methods are not without drawbacks. First, respondents are not likely to use any one method of electronic contact equally, allowing for underestimation of degree for individuals who do not use the method regularly or at all. Second, contacts within such lists are rarely categorized, so respondents who refer

to them likely provide information on their entire list of contacts, not the preferred subset of potential recruits. Finally, the presence of a contact on such a list does not necessarily mean a relationship between individuals exists. Old friends with whom the respondent no longer has contact or those who were only contacted once may remain on such lists indefinitely. However, these limitations may be a small price to pay if they provide usable information that is more effective than self-reports.

The third challenge for RDS degree estimation is the random recruitment assumption. While this criterion is often viewed as an unrealistic assumption about individual behavior, the assumption can be rephrased as an assumption that recruitment occurs randomly from those individuals who comprise the recruiter's degree. Thus, if the recruitment process is adequately understood and the degree question is specified accordingly, the random recruitment assumption is more likely to be met. For example, respondents may only recruit close ties that they trust; those with whom they discuss important matters; those they know will participate in the study; or simply the first person they see. In each of these cases more direct degree questions (i.e. "how many undergraduates do you discuss import matters with?") would solicit more appropriate subsets of potential recruits. If respondents rely on more than one method of recruitment, e.g. some recruit those with whom they discuss important matters and others recruit randomly, the researcher can ask multiple degree questions and which method is used for each recruitment and then weight the degree data accordingly.

These challenges to estimating RDS degree present an empirical question: Does choice of degree question affect RDS estimates and, if so, how can one identify which questions are most appropriate?

**Out-of-Equilibrium Data**

Equilibrium in RDS studies has been a hot topic for RDS theorists and field research alike since the original RDS publication in 1997 (Heckathorn). The Markov chain model on which RDS is based argues that after a, usually modest, number of waves, sample composition stabilizes and becomes independent of the initial seeds from which the sample was taken. Often this is interpreted as meaning that once a sample has gone through enough waves to reach equilibrium, it has stabilized and analysis can be performed. A stricter interpretation is that data collected before reaching equilibrium are biased by seeds and therefore the sampling truly starts only after equilibrium is reached. The question then is what to do with data collected before equilibrium is reached. Some theorists have suggested excluding early, pre-equilibrium waves from analysis altogether (Salganik 2006). Others point to practical limitations of such an approach, citing that the rate at which at which equilibrium is attained is variable specific and excluding out-of-equilibrium data would produce univariate population estimates based on one sample with varying sample sizes (Wejnert and Heckathorn 2008). As a simplification, one could exclude all data sampled before a certain cutoff. Such a method would essentially expand Volz and Heckathorn's (2008) recommendation that seeds be excluded from analysis to exclude early waves as well. Alternately, for studies where a majority of the sample originates from one seed, Heimer (2005) suggests calculating estimates based only on data gathered in the longest chain. Wejnert and Heckathorn (2008) adapt this approach in their analysis of the 2004 data analyzed in this chapter and find only stochastic differences between estimates based on full data vs. long chain data.

While the debate over out-of-equilibrium data largely stems from differences between methodological theory, where analysis is governed by very specific rules and assumptions about the data, and methodological practice, where all data are valuable

and no data are perfect, several empirical questions can help elucidate the debate. First, are there substantial differences between estimates calculated using data that have just reached equilibrium and data that have been primarily sampled after reaching equilibrium? Second, what effect does excluding out-of-equilibrium or early wave data from analysis have on the estimates and/or confidence intervals? Finally, is there an optimum cutoff for including or excluding data gathered in early waves?

## *Methods*

### RDS Analysis

Analysis is carried out on three categorical variables included in both 2004 and 2008 samples: race (White, Black, Hispanic, Other, and Non-U.S. Citizen [2008 sample only]), gender (male, female), and college within which each student is enrolled (Agricultures and Life Sciences [CALS], Arts and Sciences [Arts], College of Engineering [Engineer], Human Ecology [HE], Hotel Administration [Hotel], and Industrial Labor Relations [ILR]). All variables are dichotomized and analyzed independently. In cases where the number of respondents in a category, such as Hispanic students, becomes too small to estimate, analysis of all categories in that variable can fail if they are analyzed as a single, multi-category variable. Dichotomization of all categories reduces estimation failure to only the affected category. Differences between estimates based on dichotomized categories and those based on the complete variable are minor and non-systematic. In the dichotomization, all non-group respondents, including those labeled as “missing” are coded as part of the non-group. Including missing values as members of the non-group increases the number of recruitments in the 2008 sample by six for race and one for college. There are no missing data in the 2004 sample.

Unless stated otherwise, all RDS I estimates and confidence intervals are calculated using RDSAT 6.0.1 (Volz et al. 2007) with alpha level 0.025 (consistent with a 95% confidence interval), 10,000 re-samples for bootstrap, and default settings for all other options. All RDS II estimates and intervals[5] are calculated using custom software corresponding to Volz and Heckathorn (2008).

**Degree Measures**

For each sample, estimates are calculated based on five different measures of degree. In all cases, respondents were asked to provide the number of undergraduates enrolled at the university who meet the stated criteria however, due to changes in technology and lessons learned from the 2004 sample, the degree questions asked in 2008 differ from those asked in 2004. The following degree measures are used in the comparisons:

2004 Sample Degree Measures:

- Buddylist Degree: the number of students the respondent has saved on his or her instant messenger program buddylist.
- Recruit Degree: the number of students the respondent believes she could potentially recruit for the study.
- Email Degree: the number of students the respondent has contacted through email in the past 30 days.
- Standard Degree: the number of students the respondent knows and has personally interacted with in the past 30 days.
- Weighted Degree: weighted sum of the respondent's number of close friends (0.47), friends (0.50), and acquaintances (0.03).

[5] I am grateful to Erik Volz for his help calculating RDS II estimates and variance. Any errors are my own.

2008 Sample Degree Measures:

- Internet Degree: the number of different students the respondent has saved on any internet networking software, such as MySpace, FaceBook, Instant Messenger, etc.
- Discuss Important Matters (DIM) Degree: the number of students the respondent discusses important matters with.
- Cell Phone Degree: The number of students the respondent has stored in his cell phone contact list.
- Standard Degree: the number of students the respondent knows and has personally interacted with in the past 14 days.
- Weighted Degree: weighted sum of standard degree (0.19) and discuss important matters degree (0.81).

Degree measures used in the 2004 sample represent a diverse range of possible networks used for recruitment. First, the number of buddies a student has saved on his or her instant messenger program represents the primary means of online communication available to students. At the time of sampling, high speed internet was available to all students in every building on campus and nearly every student's home, but the wide range of networking software and sites, such as MySpace, FaceBook, and gmail chat, had not yet become popular and students primarily used AOL Instant Messenger for online communication and texting. The number of buddies is clearly displayed by the software for each user. Many respondents reported contacting potential recruits through instant messenger to confirm interest in participation before forwarding a recruitment email (Wejnert and Heckathorn 2008). Second, respondents were asked to report the number of students they could potentially recruit for the study. This question is intended as the most direct measure of degree according to

RDS theory and assumptions described above. Third, because recruitment occurred via email, respondents were asked the number of students with whom they had communicated through email in the past 30 days. Fourth, respondents were asked the number of students they knew personally with whom they had interacted in the past 30 days. This format, where a tie is behaviorally defined within a specified time frame, is referred to as the “standard" measure because it follows the behaviorally and temporally defined individual degree question format used in nearly all RDS studies. Finally, respondents reported the number of “close friends”, “friends”, and “acquaintances” they have at the university. Additionally, each respondent was asked to categorize her recruiter as a “close friend”, “friend”, “acquaintance”, or “stranger”. Excluding the seeds, who have no recruiter, approximately 47% reported being recruited by a “close friend”, 50% by a “friend”, and 3% by an “acquaintance”. Each respondent’s reported number of close friends, friends, and acquaintances is weighted by these percentages and summed to provide a weighted measure of individual degree.

Degree measures used in analysis of 2008 data are similar, but differ in several ways. First, respondents reported the number of different students they have saved on any online communication software. However, by this time, many options existed for online networking and respondents may not have been able to look up their degree as easily as in 2004. Next, the number of potential recruits question was replaced with a report of the number of students with whom the respondent discusses important matters. The “discuss important matters” question (here after referred to as “DIM degree”) has been used extensively in social network studies and found effective at capturing close ties (Burt 1985; Marsden 1987; McPherson et al. 2006). Third, respondents were asked the number of students saved in their cell phone address book. At the time of sampling, cell phones had become the primary method of communication among students. Fourth, the temporal constraint used in the standard

degree question was reduced from 30 days to 14 days due to the potential speed with which recruitment can occur on campus (Wejnert and Heckathorn 2008). Finally, weighted degree is calculated based on the proportion of students reporting being recruited by someone with whom they discuss important matters to provide a more objective classification than the friendship categories used in 2004. Nearly 81% of respondents reported being recruited by someone with whom they discuss important matters. Consequently, the weighted degree measure is the weighted sum of DIM and standard degree measures.

Degree measurement in both studies is designed to maintain a realistic scenario applicable to many RDS studies. All measures rely on self-reports and are susceptible to any problems associated with such measures. For measures where respondents could look up their degree, such as the buddylist measure, there is no guarantee that respondents did not answer from memory nor is it guaranteed that all students used such methods of communication equally or at all. Additionally, while respondents were asked to limit their answers to students at the university in all measures, no checks were imposed nor were the answers vetted in any way to conform to this requirement. Consequently, there is no reason to suspect the degree measures employed in this chapter are unlike those that could be used in other RDS studies.

**Analysis of Equilibrium**

To answer questions related to equilibrium, multiple datasets were created to exclude respondents surveyed before or after specific waves of interest. Table 4 and Table 5 show population parameters and raw sample proportions for all created samples used in equilibrium analyses for 2008 and 2004 data, respectively. The datasets were created using waves as cut points, for example, column five (earliest waves included = 4) refers to a data set in which all respondents sampled before wave four are excluded from analysis. The table also shows the estimated number of waves

required to reach equilibrium for each variable. Between three and nine waves were required for equilibrium, with an average of 6.4 waves, for variables analyzed in the 2004 sample. Variables analyzed in the 2008 sample required four to nine waves, with an average of 6.2 waves, to reach equilibrium. Thus, equilibrium is said to be reached for all analyzed variables by wave nine of sampling in each sample. When seeds are counted as wave zero, there are 18 waves of recruitment in the 2004 sample and 23 waves of recruitment in 2008.

Table 4: 2008 sample proportions and waves required to reach equilibrium for all data sets used in analysis.

| Variable | Waves Required for Equilibrium | Population Parameter | Full Sample (n = 378) | Equilibrium Met Sample (n = 156) | Earliest Waves Included = 4 (n = 332) | Earliest Waves Included = 7 (n = 294) | Earliest Waves Included = 10 (n = 222) |
|---|---|---|---|---|---|---|---|
| **Race** | | | | | | | |
| Asian | 9 | 0.182 | 0.154 | 0.147 | 0.157 | 0.154 | 0.158 |
| Black | 7 | 0.059 | 0.037 | 0.038 | 0.036 | 0.031 | 0.036 |
| Hispanic | 7 | 0.062 | 0.021 | 0.032 | 0.024 | 0.014 | 0.014 |
| Other | 8 | 0.051 | 0.035 | 0.032 | 0.033 | 0.031 | 0.036 |
| White | 7 | 0.557 | 0.715 | 0.699 | 0.713 | 0.736 | 0.724 |
| nonUS | 9 | 0.089 | 0.037 | 0.045 | 0.036 | 0.034 | 0.032 |
| **Gender** | | | | | | | |
| Male | 6 | 0.511 | 0.516 | 0.539 | 0.532 | 0.531 | 0.545 |
| **College** | | | | | | | |
| CALS | 4 | 0.237 | 0.247 | 0.25 | 0.232 | 0.222 | 0.243 |
| Arts | 4 | 0.305 | 0.292 | 0.308 | 0.283 | 0.280 | 0.279 |
| Engineer | 5 | 0.203 | 0.175 | 0.167 | 0.181 | 0.181 | 0.180 |
| HE | 4 | 0.091 | 0.164 | 0.154 | 0.178 | 0.195 | 0.171 |
| Hotel | 5 | 0.065 | 0.032 | 0.026 | 0.033 | 0.038 | 0.036 |
| ILR | 6 | 0.062 | 0.085 | 0.083 | 0.090 | 0.082 | 0.086 |

Table 5: 2004 sample proportions and waves required to reach equilibrium for all data sets used in analysis. ( ) indicates sample size is too small for estimate calculation.

| Variable | Waves Required for Equilibrium | Population Parameter | Full Sample (n = 159) | Equilibrium Met Sample (n = 83) | Earliest Waves Included = 4 (n = 122) | Earliest Waves Included = 7 (n = 99) | Earliest Waves Included = 10 (n = 76) |
|---|---|---|---|---|---|---|---|
| **Race** | | | | | | | |
| Asian | 6 | 0.189 | 0.365 | 0.422 | 0.367 | 0.354 | 0.303 |
| Black | 7 | 0.054 | (0.013) | (0.012) | (0.006) | (0.01) | (0.013) |
| Hispanic | 7 | 0.060 | 0.044 | 0.06 | 0.021 | 0.040 | (0.026) |
| Other | 6 | 0.013 | 0.069 | 0.072 | 0.061 | 0.061 | 0.066 |
| White | 6 | 0.685 | 0.509 | 0.434 | 0.537 | 0.535 | 0.592 |
| **Gender** | | | | | | | |
| Male | 3 | 0.505 | 0.597 | 0.578 | 0.590 | 0.596 | 0.618 |
| **College** | | | | | | | |
| CALS | 6 | 0.323 | 0.233 | 0.241 | 0.230 | 0.232 | 0.224 |
| Arts | 6 | 0.223 | 0.220 | 0.265 | 0.213 | 0.202 | 0.171 |
| Engineer | 7 | 0.197 | 0.352 | 0.277 | 0.393 | 0.414 | 0.434 |
| HE | 8 | 0.096 | 0.075 | 0.06 | 0.082 | 0.091 | 0.092 |
| Hotel | 9 | 0.058 | 0.069 | 0.12 | (0.033) | (0.01) | (0.013) |
| ILR | 6 | 0.060 | 0.044 | 0.036 | 0.041 | 0.040 | 0.053 |

## Population Parameters

Population parameters are calculated using published frequency data of university enrollment for fall semester of the academic year in which the sample was taken (Cornell 2004; 2008). While both RDS samples were collected in the spring, it is unlikely that university spring enrollment differs from that of the fall in any significant, systematic way. Population parameters are calculated for gender, college within the university, and race as follows. Gender proportions are calculated as the number of males or females enrolled divided by the total number of students enrolled. Similarly, college proportions are calculated as the number of students enrolled in each college divided by the sum of students enrolled in each college excluding the approximately 40 students (less than 0.3% of all students) enrolled as “internal transfer division”. Students enrolled in the College of Art, Architecture, and Planning, which make up approximately 4% of the student population and are excluded from analysis due to low prevalence in the samples, are included in the divisor for other college

parameters. Consequently, population parameters for the six colleges reported do not sum to 100%. However, this does not present a problem for estimation comparison because each college is analyzed as an independent dichotomous variable and therefore the estimated proportions need not sum to 100%.

Finally, calculation of population parameters for race is more complex because of two key differences between the institutional and survey categorizations of race. First, the institutional data treat "Foreign Nationals" as a separate catch all category and present racial categories for US nationals only. Thus, there could be a significant number of respondents who self-identify as one race on the survey but are counted as "Foreign Nationals" in the institutional data. Second, the institutional data include a "US citizen, race unreported" category which becomes problematic if some races are more likely to withhold their racial status from the university than others. While no further information is available, it is unlikely racial information is withheld randomly.

These additional categories in the institutional data are especially problematic for analysis of 2004 data, which do not include either category. In this chapter, individuals in the "US citizen, unreported" and "Foreign Nationals" institutional data categories are not counted as part of the student body in 2004 and excluded from parameter calculation. For example, the population parameter for blacks is the proportion of black students out of all students who are US nationals and reported their race to the university. While excluding approximately 13% of the student body, this method is arguably better than Wejnert and Heckathorn's (2008) method, which includes all non-whites or non-Asians under a single "under-represented minority (URM)" category and implicitly assumes that all foreign nationals and all US nationals who do not report their race are neither white nor Asian.

To avoid this discrepancy between survey and institutional data, two additions were made in 2008. First, a "prefer not to answer" option was included in the race

question (neither survey allowed unanswered questions). Second, in a separate question, respondents were asked if they are U.S. citizens/permanent residents. All respondents reporting they are not U.S. citizens/permanent residents (n = 14) make up 3.7% of the survey data and are coded as a separate "nonUS" racial category that corresponds to the "Foreign Nationals" category in the institutional data, which make up 7.9% of the student body. Eleven of these 14 respondents racially identified themselves as "Asian". Only two of 378 respondents chose the "prefer not to answer" racial option, suggesting that students are more willing to provide racial information to a survey than to university officials and removing the ability to include a "US citizen, unreported" racial category in the 2008 analysis. Consequently, individuals in the "US citizen, unreported" institutional data category, which make up 11% of students, are not counted as part of the denominator and are excluded from parameter calculation in both 2004 and 2008.

**Measuring Estimate Accuracy**

In their institutional comparisons, Wejnert and Heckathorn (2008) report whether or not population parameters are captured by the 95% confidence interval, a method that combines the accuracy of RDS estimates and confidence intervals into a single measure. In order to test RDS estimates and confidence intervals separately, I use two continuous measures based on the absolute difference between the estimate and the parameter. These measures are termed estimate and interval *in*accuracy because lower values correspond to better estimates. Estimate inaccuracy is defined as the absolute difference between parameter and estimate. While not standardized, estimate inaccuracy removes any possible confounding effects of RDS variance estimation, which may be flawed, and provides a measure of inaccuracy dependent only on the estimate. An estimate is considered good if it has estimate inaccuracy less than 0.05 and acceptable if estimate inaccuracy is less than 0.1.

Interval inaccuracy is intended to measure the inaccuracy of the confidence interval around RDS estimates and is defined as the estimate inaccuracy standardized by the standard error of the estimate. For RDS II estimates, this is straightforward; however, for RDS I, potentially non-symmetric confidence intervals are taken directly from the bootstrapped distribution without first estimating variance (Salganik 2006). Thus, for RDS I estimates, interval inaccuracy is defined as the estimate inaccuracy standardized by the distance from the estimate to the 95% confidence interval tail closest to the parameter divided by 1.96, which serves as an approximation of the bootstrapped standard error. Thus, if the estimate underestimates the parameter, standardization is based on the upper bound, if the parameter is overestimated, the lower bound is used. The standardization is an estimate of the single tail standard error and ensures that all confidence intervals that fail to capture the institutional parameter will have interval inaccuracy greater than 1.96. For example, if a hypothetical group makes up 25% of the population and the RDS estimate is 30%, with 95% CI (20, 50), then the estimate inaccuracy is $|0.3-.25|=0.05$, the standardizing value is $(0.3-0.2)/1.96=0.051$, and the interval inaccuracy is:

$$\frac{|0.3-.025|}{0.051}=0.98 \qquad (3.1)$$

Inaccuracy scores are calculated for each dichotomous category and then averaged with other categories of that variable to provide averaged inaccuracy scores for race, gender, and college. Because gender is already dichotomous, inaccuracy scores for males are reported (in all cases inaccuracy scores for males and females are equivalent).

The proportional nature of this analysis ensures that all differences between observed and expected measures are less than one, thus, they are not squared in order to avoid artificially deflating these differences.

**Estimating Design Effect**

Estimated design effects, $DE_{P_X}$ , for both RDS I and RDS II are defined in a manner consistent with Salganik (2006) as follows:

$$DE_{P_X} = \frac{Var(P_X)_{RDS}}{Var(P_X)_{SRS}} \tag{3.2}$$

where $Var(P_X)_{RDS}$ is the RDS estimated variance and $Var(P_X)_{SRS}$ is the expected variance for a simple random sample of equal size. Design effects are calculated for each dichotomous variable and then averaged according to partitions defined in the text.

## *Results*

In addition to the results presented here, tests of assumptions for both samples are presented in Chapter Two.

**Comparing RDS I and RDS II**

As noted in Chapter One, there are two forms of the RDS estimator, RDS I and RDS II, each employing a different approach to variance estimation. While simulations conducted by Volz and Heckathorn (2008) suggest RDS II may provide better estimates than RDS I, to date no empirical comparison on a known population has been done.

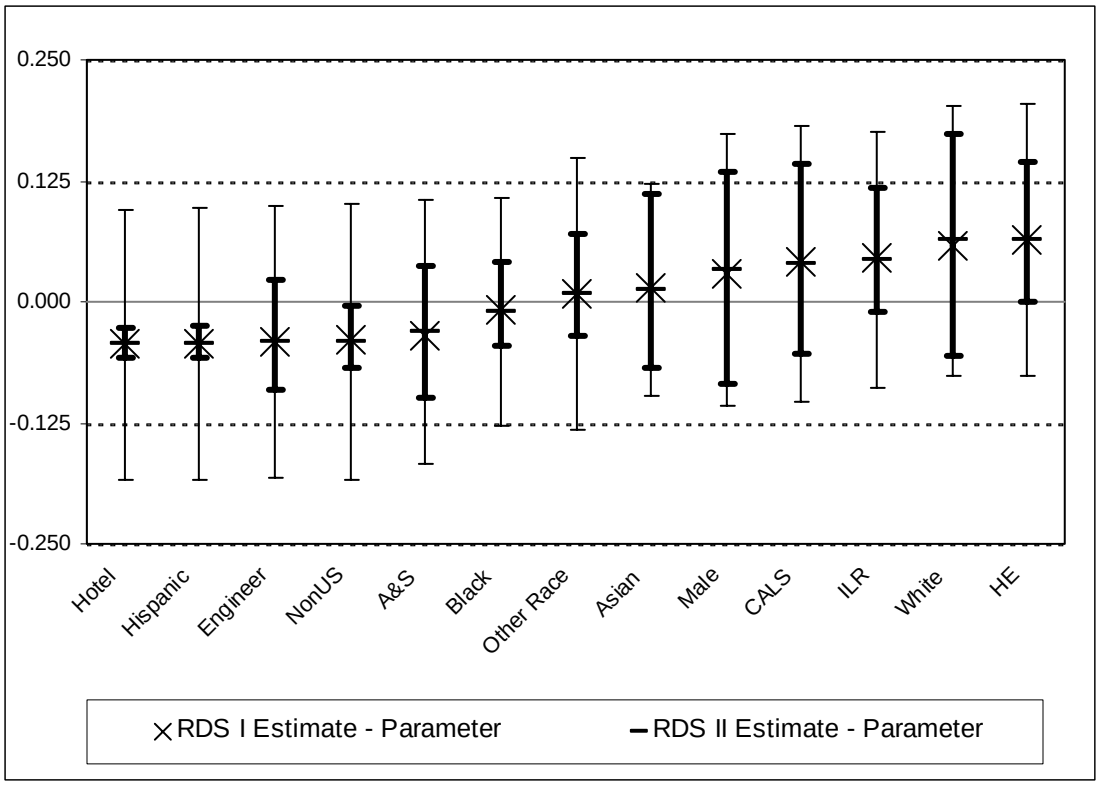



Figure 12: RDS I (”x”s) and RDS II (dashes) estimates and 95% confidence intervals for 13 dichotomous variables corresponding to race, gender, and college within the university calculated using the standard degree measure from the 2008 sample. The estimates are adjusted such that the line y = 0 represents the population parameter for each variable. Each estimate’s distance from the axis represents its distance from the true parameter. 95% confidence intervals for RDS I estimates are represented as heavy solid lines while 95% confidence intervals for RDS II estimates are represented as thin solid lines.

Figure 12 shows RDS I and II estimates for 13 dichotomous variables corresponding to race, gender, and college calculated using 2008 standard degree measure data. The estimates are adjusted such that the line y = 0 represents the population parameter for each variable. Consequently, an estimate’s distance from the axis represents its distance from the true parameter. 95% confidence intervals for RDS I estimates are represented as heavy solid lines, while 95% confidence intervals for RDS II estimates are represented as thin solid lines. First, note that the markers for RDS I and RDS II estimates coincide closely, mathematically:

$\sum_{1}^{13}\left|\widehat{P_i^{RDS\,I}} - \widehat{P_i^{RDS\,II}}\right| = 0.0215$. Second, the estimates themselves provide reasonable approximations, generally falling within ±0.05 of the population parameter.

However, RDS I and RDS II confidence intervals differ substantially. RDS II intervals are wider, in some cases much wider, and more consistent across variables than their RDS I counterparts. Furthermore, while RDS I intervals fail to capture the true parameter in four of the 13 variables (Hotel, Hispanic, nonUS, and HE), RDS II intervals capture all 13 parameters easily.

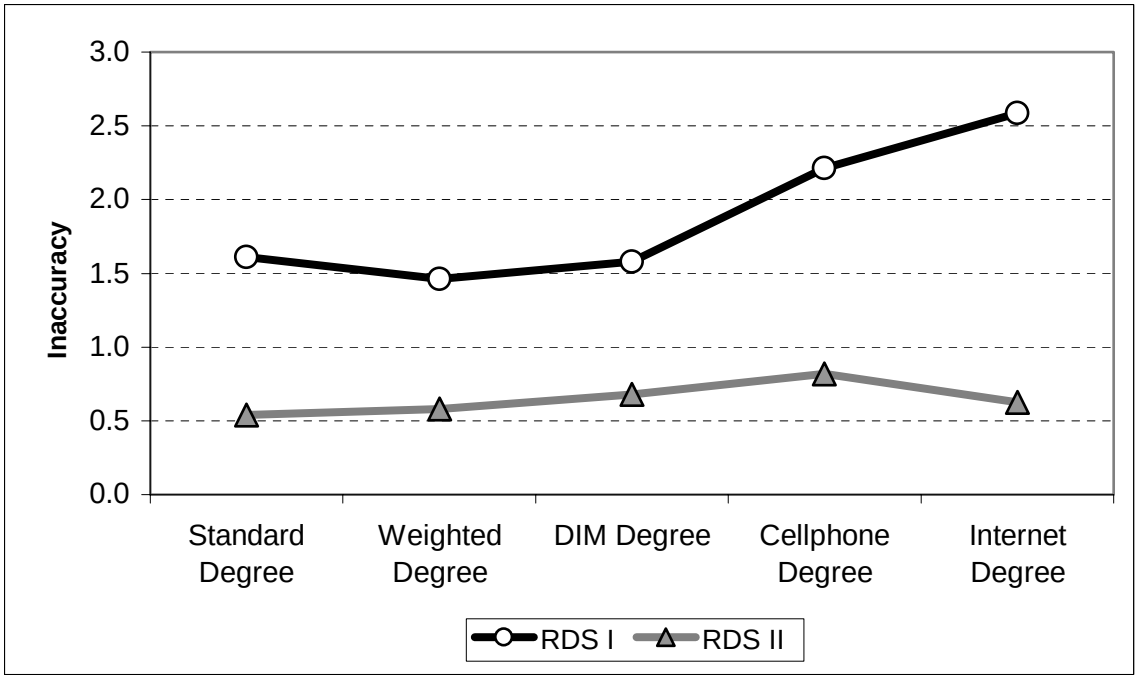



Figure 13: 2008 sample interval inaccuracy for RDS I (black) and RDS II (gray) estimates averaged across race, gender, and college for five measures of degree. The line y = 1.96 represents 95% confidence interval bounds. Any interval with inaccuracy greater than 1.96 fails to capture the population parameter. Confidence intervals based on RDS II variance estimation are wider and more likely to capture population parameters on average than intervals based on RDS I variance estimation.

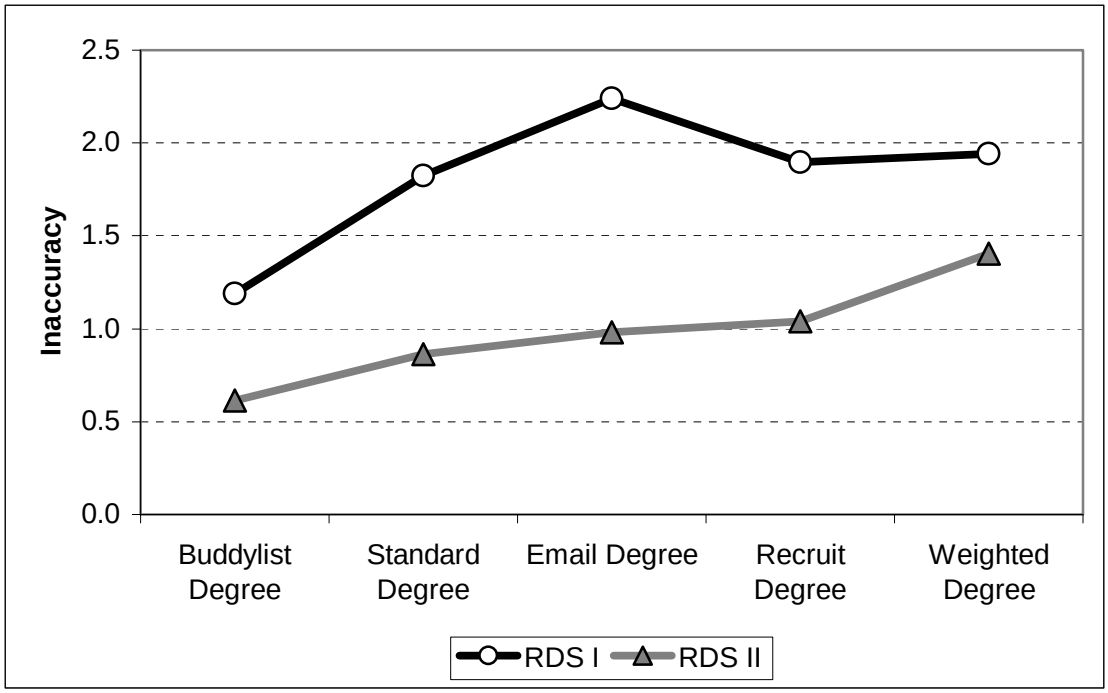



Figure 14: 2004 sample interval inaccuracy for RDS I (black) and RDS II (gray) estimates averaged across race, gender, and college for five measures of degree. The line y = 1.96 represents 95% confidence interval bounds. Any interval with inaccuracy greater than 1.96 fails to capture the population parameter. Confidence intervals based on RDS II variance estimation are wider and more likely to capture population parameters on average than intervals based on RDS I variance estimation.

Figure 13 and Figure 14 show interval inaccuracy averaged across all variables for 2008 and 2004 estimates calculated using five measures of degree. An interval inaccuracy score less than 1.96 signals that, on average, 95% confidence interval bounds include the population parameter. The close correspondence between RDS I and RDS II estimates ($r = 0.9970$ for 2004 data and $r = 0.9998$ for 2008 data) suggests that differences in interval inaccuracy between RDS I and RDS II are largely due to differences in variance estimation. The graphs show that in all cases the overall interval inaccuracy of RDS II is less than RDS I and therefore less susceptible to type I error. The RDS II overall interval inaccuracy generally falls well within the 1.96 cutoff for parameter inclusion while RDS I interval bounds tend to hover dangerously

close to the parameter. Of the 65 dichotomous estimates used to generate Figure 13, which are by no means independent, 63 (96.9%) RDS II 95% confidence intervals capture the parameter, while only 42 (64.6%) parameters are captured by RDS I confidence intervals. In the 2004 sample (Figure 14), RDS II intervals capture 45 of 55 (81.8%) parameters, while RDS I intervals only capture 36 (64.6%) parameters.

**RDS I and RDS II Design Effects:**

While wider confidence intervals decrease the probability that parameters are not captured by confidence intervals (type I error), excessively wide intervals reduce the precision with which inferences regarding the population can be made (type II error). Using the design effect terminology, RDS I and RDS II variance estimation is compared to variation expected from simple random samples of similar size. Results presented above, which find 95% confidence intervals succeed in capturing population parameters in fewer than 95% of cases, suggest the bootstrap variance estimation procedure used in RDS I underestimates variance. Furthermore, the problem appears to arise predominantly in cases where the estimated variable represents a small portion of the population. While RDS II variance estimation does not appear to suffer from underestimation, it is possible that variance is over estimated using the RDS II variance estimator.

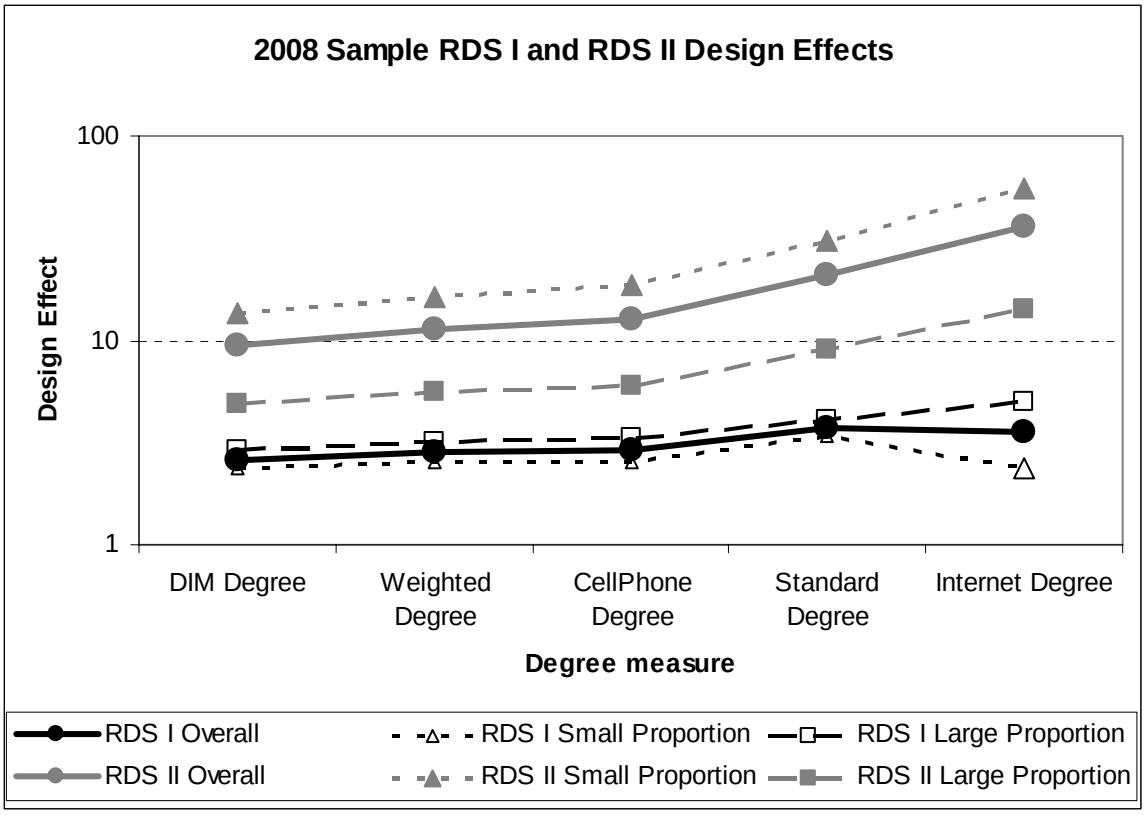



Figure 15: RDS I (black) and RDS II (gray) design effects for the overall sample, small proportion, and large proportion variables. Small proportion variables are dichotomous variables for which the population parameter is less than 0.1. Large proportion variables are those for which the population parameter is greater than 0.1.

Figure 15 plots RDS I and RDS II design effects calculated for 2008 data using five degree measures. Results for the 2004 sample (not shown) are similar. Overall lines show the design effect averaged across all 13 variables for each degree measure used, "small proportion" is the average design effect of seven dichotomous variables that compose less than 10% of the population, and "large proportion" is the average design effect of the six remaining variables which make up over 10% of the population. As expected, overall design effect of RDS I is smaller than that for RDS II. Averaged across all variables and degree measures, the design effect is 3.1 for RDS I and 18.1 for RDS II. A design effect of 3.1 means an RDS estimate has variance three times as large as that of a simple random sample. In other words, an RDS sample

would require sample size three times larger than a simple random sample to achieve the same statistical power. The results suggest groups that make up a small proportion of the population may be the primary culprits. In RDS I calculation, groups making up less than 10% of the population tend to have lower design effects (average DE = 2.6) than groups making up more than 10% of the population (average DE = 3.7). However, in RDS II the opposite is true. Small groups tend to have very large design effects (average DE = 26.8) while larger groups have smaller design effects (average DE = 7.9).

It is important to note that these design effects are calculated using the same data, for estimates that are highly convergent, and therefore neither overall design effect should be attributed to RDS in general or this sample in particular as a way to calculate power or sample size. These results merely show that neither RDS I nor RDS II variance estimation is error free and suggest the largest difficulties arise with small groups. In small groups, RDS I confidence intervals sometimes fail to capture the true parameter while RDS II intervals tend to have large design effects.

In summary, RDS I and RDS II point estimates are found to coincide closely with each other, a result that is consistent with Volz and Heckathorn's (2008) work. However, confidence intervals based on RDS II are generally wider, more consistent across variables, and more likely to capture population parameters than their RDS I counterparts. Furthermore, the RDS I bootstrap procedure used to estimate confidence intervals was found to underestimate variance, especially for small groups. In this analysis, 95% confidence intervals calculated based on bootstrapped variance fail to capture the parameter more often than the 5% suggested by the interval, while those calculated using RDS II display a capture rate that resembles what would be expected from an ideal variance estimate. On the other hand, analysis of design effects suggests RDS II overestimates variance, in some cases by a large amount. Both estimation

procedures seem to have significant problems with variance estimation of small groups, such as those making up less than 10% of the population, albeit in opposite ways.

Discussion of various degree measures and their effect on estimation is presented below. Because it is generally better to overestimate variance rather than underestimate it, results presented in the remainder of this paper are calculated using RDS II estimation. Analyses based on RDS I estimation support similar conclusions.

**Comparison of Degree Measures:**

Descriptive statistics and correlations for all degree measures are presented in Table 6. Consistent with other work on social networks, reported degree distributions are highly skewed with small numbers of respondents reporting very high degrees for all measures. Respondents surveyed in 2008 tended to report higher degrees than those in 2004, however the small sample size and the unique sampling method make statistical comparison difficult. In some cases, self-report degree measures include unreasonably high outliers. For example, in 2004 one respondent reported 10,000 potential recruits, 10,000 friends, and 100,000 acquaintances at the university which has less than 14,000 students. In such cases, it is common to truncate the degree distribution by pulling in a small percentage of the outlying degrees when calculating RDS I estimates. Estimates were calculated for non-truncated, 1%, 5%, and 10% truncation for buddylist and standard degree in 2004 and standard and weighted degree in 2008. Pulling in degree outliers had no effect, positive or negative, on estimates in 2004 or 2008 (not shown). Finally, all 2008 degree measures are significantly correlated with each other ($p < 0.01$). Reported degrees in 2004 display less positive correlation; however, when the one extreme outlier described above is removed, all 2004 degree measures are significantly correlated with each other ($p < 0.01$, not shown). While all degree measures are positively correlated, the correlations

are not large enough to make choice of degree measure trivial. Consequently, it is important to know how estimates based on various degree measures compare.

Interval and estimate inaccuracy scores for race, gender, and college based on different measures of degree are shown in Figure 16 and Figure 17 for 2008 and 2004 samples respectively. In the 2008 sample, the best estimates are those produced by standard degree, weighted degree, and DIM degree measures, which all capture the true parameter and are within 0.1 of the parameter for all dichotomous variables. Of the three, the weighted degree measure is the best by a small margin because, on average, it produces estimates that are closer to the parameters[6] (estimate inaccuracy = 0.031) than both DIM degree (estimate inaccuracy = 0.034) and standard degree (estimate inaccuracy = 0.037). Furthermore, its interval inaccuracy (0.580) is slightly higher than that of standard degree (0.541) suggesting that confidence intervals based on weighted degree are narrower than those based on standard degree. The difference, however, is expectedly small given that weighted degree is calculated from DIM and standard degree measures proportionally weighted by the number of respondents reporting discussing important matters with their recruiter.

[6] Recall that inaccuracy scores for gender are based on a single dichotomous variable and therefore display greater variability than race, college, or overall scores which represent the average of scores from multiple dichotomous variables.

Table 6: Descriptive statistics and correlations for degree measures in 2004 and 2008.

| 2004 Degree Measures | N | Minimum | Maximum | Mean | Std. Dev. | Skewness Statistic | Skewness Std. Error |
|---|---|---|---|---|---|---|---|
| Standard Degree | 159 | 0 | 450 | 74.7 | 68.35 | 2.56 | 0.192 |
| Recruit Degree | 159 | 0 | 10000 | 80.7 | 791.94 | 12.59 | 0.192 |
| Email Degree | 159 | 0 | 1000 | 19.97 | 80.08 | 11.77 | 0.192 |
| Buddylist Degree | 159 | 0 | 200 | 66.3 | 44.58 | 1.13 | 0.192 |
| Weighted Degree | 159 | 2.59 | 3547 | 49.13 | 281.34 | 12.33 | 0.192 |

| 2008 Degree Measures | N | Minimum | Maximum | Mean | Std. Dev. | Skewness Statistic | Skewness Std. Error |
|---|---|---|---|---|---|---|---|
| Standard Degree | 378 | 3 | 1000 | 103.91 | 104.99 | 3.68 | 0.125 |
| Cell Phone Degree | 377 | 0 | 300 | 58.66 | 47.82 | 2.16 | 0.126 |
| Internet Degree | 377 | 0 | 900 | 128.67 | 141.61 | 1.75 | 0.126 |
| DIM Degree | 377 | 0 | 150 | 11.99 | 14.69 | 4.14 | 0.126 |
| Weighted Degree | 377 | 1.33 | 198.1 | 29.47 | 25.78 | 2.59 | 0.126 |

2004 Degree Correlations

| | Recruit | Email | Buddylist | Weighted |
|---|---|---|---|---|
| Standard | -0.01 | 0.06 | 0.50** | 0.03 |
| Recruit | | 0.98** | 0.25** | 0.99** |
| Email | | | 0.29** | 0.98** |
| Buddylist | | | | 0.28 |

2008 Degree Correlations

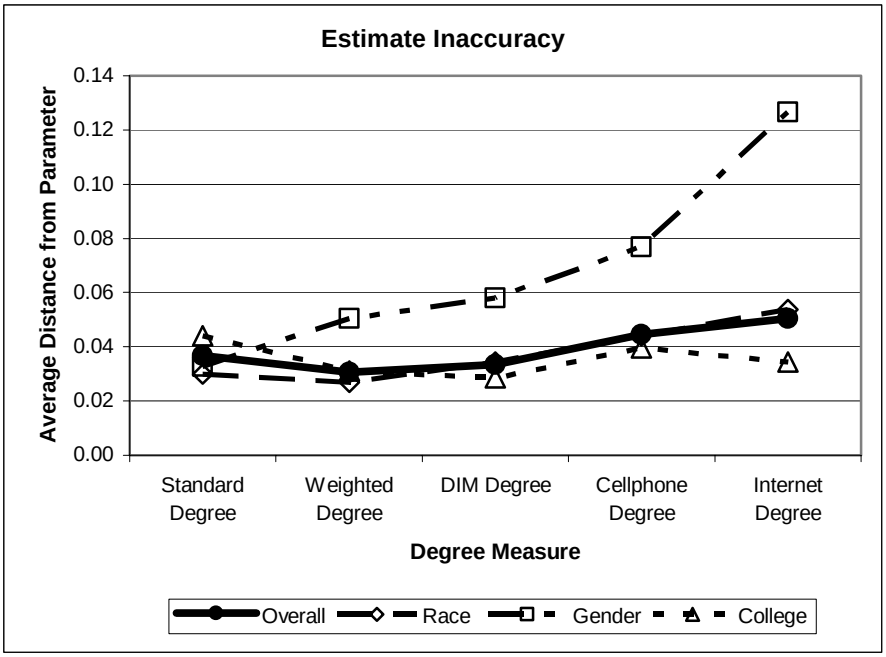
| | Cell | Internet | DIM | Weighted |
|---|---|---|---|---|
| Standard | 0.42** | 0.32** | 0.26** | 0.90** |
| Cell Phone | | 0.46** | 0.33** | 0.48** |
| Internet | | | 0.16** | 0.32** |
| DIM | | | | 0.66** |

** Correlation is significant at the 0.01 level (2-tailed)

In the 2004 sample, the buddylist degree measure provides the best overall estimates. In all but one case (gender), the estimate is within 0.1 of the true parameter. As described by Wejnert and Heckathorn (2008), the 2004 sample largely over-sampled Asian students, biasing racial estimates. The buddylist degree measure is able to compensate for this bias because of its direct connection to the method of

recruitment. However, two degree measures intended to be directly associated with the recruitment process performed poorly. Recruit degree, which solicits the number of students a respondent might possibly recruit likely proved confusing to answer accurately for respondents who had not yet attempted to make recruitments. Estimates based on email degree performed even worse, possibly because email is rarely used for communication among undergraduates and likely had little to do with who respondents chose to recruit.

It is important to note that estimates calculated based on weighted degree in 2004 perform worse than those in 2008 because the two measures are inherently different. 2004 weighted degree is a function of the number of "close friends", "friends", and "acquaintances" respondents reported proportionally weighted by the number of recruitments made by "close friends", "friends" and "acquaintances". Consequently, 2004 weighted degree is based on terms that are largely subjective and likely interpreted differently from one respondent to another while the 2008 weighted degrees are based on DIM questions, which have been found to be interpreted consistently across respondents (Burt 1985).



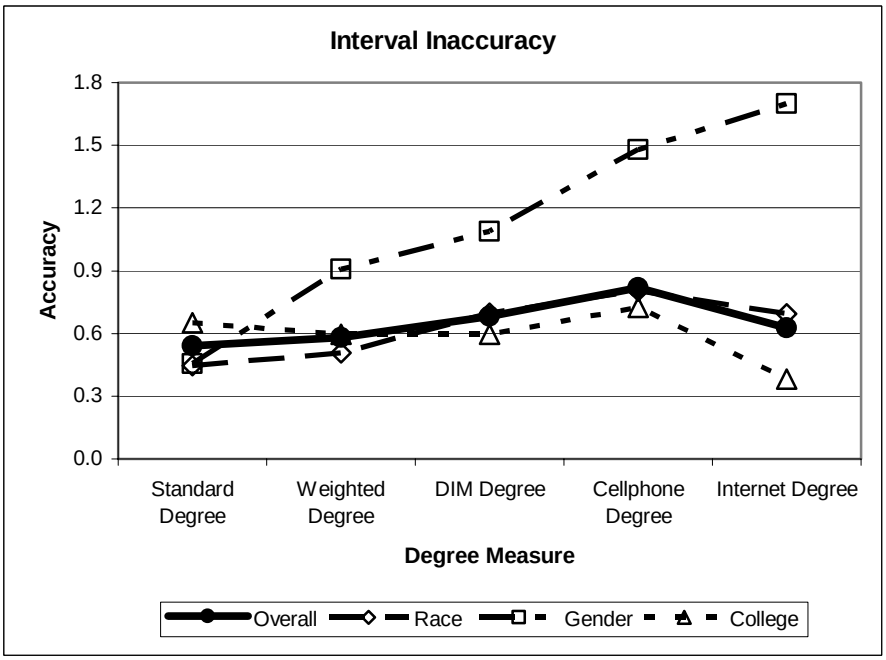
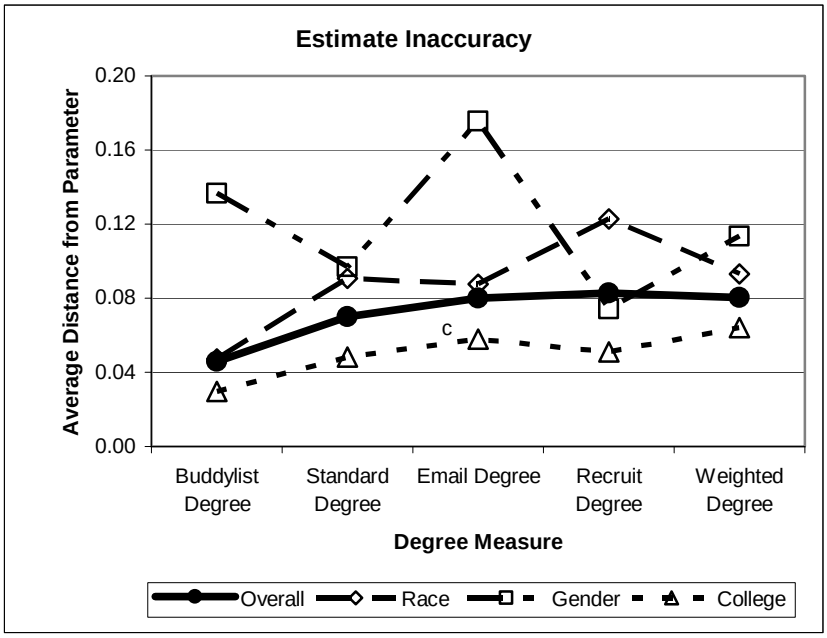




Figure 16: 2008 sample estimate and interval inaccuracy for overall, race, gender, and college based on five measures of degree. Inaccuracy scores for gender are based on a single dichotomous variable and therefore display greater variability than race, college, or overall scores, which represent the average of scores from multiple dichotomous variables.



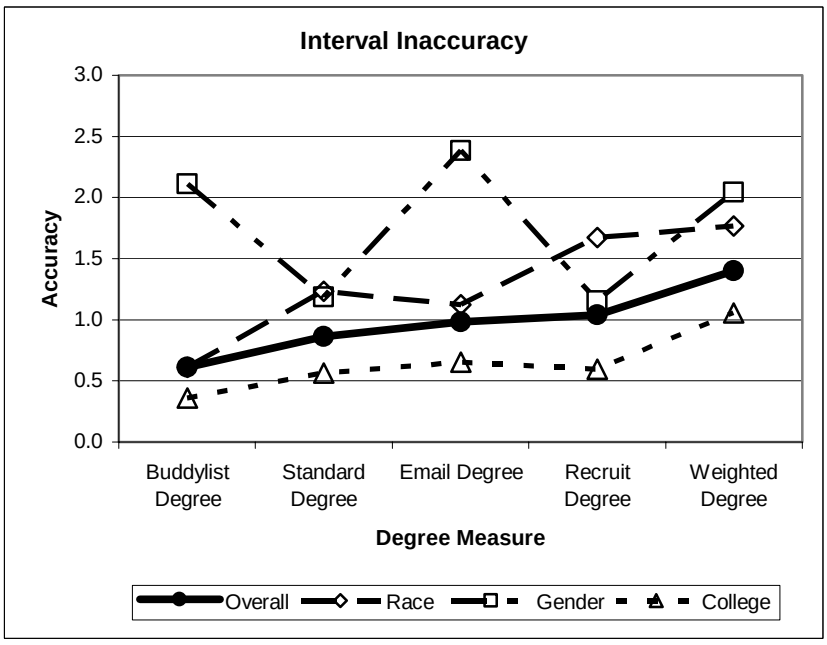




Figure 17: 2004 sample estimate and interval inaccuracy for overall, race, gender, and college based on five measures of degree. Inaccuracy scores for gender are based on a single dichotomous variable and therefore display greater variability than race, college, or overall scores, which represent the average of scores from multiple dichotomous variables.

While results show that estimates calculated using different measures of degree may differ substantially, the question remains how can one identify which measures will provide the best estimates? This analysis suggests that the best degree measures are those directly tied to recruitment choice and ability. In the 2004 sample, the buddylist measure provides precisely this. However, without access to population parameters, estimates based on the buddylist degree look more anomalous than promising and would likely have been discounted as inaccurate. The measure was included based on extensive prior knowledge of communication methods among students attending the university at the time of sampling, which is not often available to RDS researchers. Furthermore, the 2004 sample represents a unique case in which instant messenger programs constrained recruitment by displaying a pool of immediately available students from which to recruit and the speed of recruitment nearly guaranteed the recruits of anyone who waited more than a few hours to recruit would not get to participate. Consequently, while measures such as buddylist degree are difficult to identify, instances such as this are rare and highly unlikely to occur in any community in which members can interact through multiple media.

The weighed degree measure used in 2008, on the other hand, is both easy to identify and measure. As discussed above, RDS procedures reduce recruitment of strangers by making recruitment both valuable and scarce. This effect likely extends beyond strangers and encourages respondents to recruit individuals with whom they will have repeated interaction and trust to participate after accepting a coupon, i.e. those with whom they are closely tied[7]. In many cases, these same conditions are

[7] While not relevant to WebRDS studies and beyond the scope of this paper, the desire to recruit those who are likely to participate also favors recruitment of strangers waiting outside interview locations to solicit coupons from participants. Researchers should take any steps possible to reduce such recruitment.

necessary, if not sufficient, for the discussion of important matters. 2008 respondents reported a mean of approximately12 and maximum of 150 students with whom they discuss important matters, approximately $1/10^{th}$ the mean (104) and maximum (1000) number of students they reported knowing and interacting with in the past 14 days. However, over 81% reported being recruited by someone with whom they discuss important matters, suggesting that respondents may be recruiting from smaller, tighter circles than just those individuals they know. Fortunately, the questions necessary for calculating this degree measure, "how many Xs do you discuss important matters with?" and "do you discuss important matters with your recruiter?", are easily included on any questionnaire and applicable to any population in any setting[8]. Weighted degree is then easily calculated based on the proportion of respondents reporting being recruited by someone with whom they discuss important matters.

Finally, it is important to note that while the standard degree measure, which is commonly used in RDS studies, does not produce the best estimates in either sample, it does quite well. In 2004 it is second only to buddylist degree and in 2008 its estimates are statistically equivalent to both weighted degree and DIM degree. Therefore, studies in which only the standard degree measure is used are likely to produce equally valid estimates.

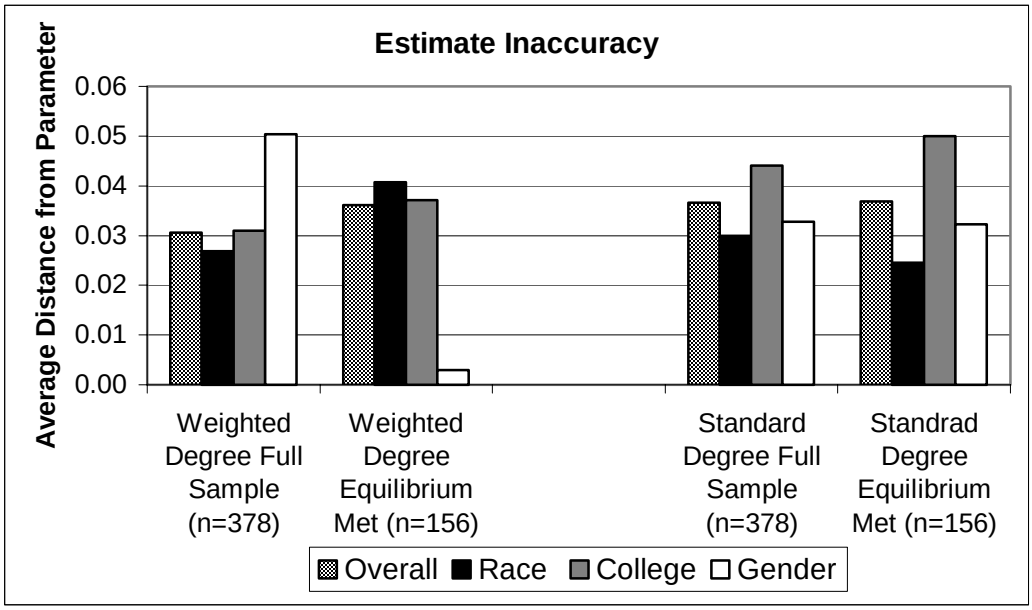
**Effects of Out-of-Equilibrium Data**

The standard RDS interpretation is that if equilibrium is reached within a single recruitment chain, then equilibrium is reached for the entire sample because all individuals have a nonzero probability of selection. A corollary of this interpretation is that once enough waves have been gathered to reach equilibrium, sampling can stop

[8] The definition of an "important matter" may vary across populations, but the trust necessary to discuss it remains relatively constant.

and analysis can begin. In most RDS studies, sampling is terminated based not on the number of waves reached, but on the overall sample size. However, if the required number of waves is not reached within the target sample size, it is recommended that sampling continue until such time. In such cases, it is important to know whether estimates derived from a sample that includes just enough waves to reach equilibrium provide adequate results. Above, the most waves required to reach equilibrium in both samples is found to be nine waves. Consequently, following the stop-when-equilibrium-is-reached approach, sampling would have stopped after wave nine.

Figure 18 compares estimate and interval inaccuracy using only data collected in waves zero through nine for the 2008 sample to inaccuracy based on the full sample. Unfortunately, the results are confounded by a reduction in sample size from 378 to 156 when only early wave data (Equilibrium Met) are used. As a result, early wave point estimates are more variable and confidences intervals are wider than those based on the full data. Consequently, wider confidence intervals are reflected as improved interval inaccuracy in early wave data compared to the full sample, while more variable estimates lead to inconsistent differences in estimate inaccuracy. Thus, while there is no evidence here to suggest a sample that has just reached equilibrium would produce worse estimates than a sample of equal size collected primarily after reaching equilibrium, further research is needed to disentangle the effects of out-of-equilibrium data versus reduction of sample size. Results from 2004 data reflect a similar pattern and are available from the author on request.



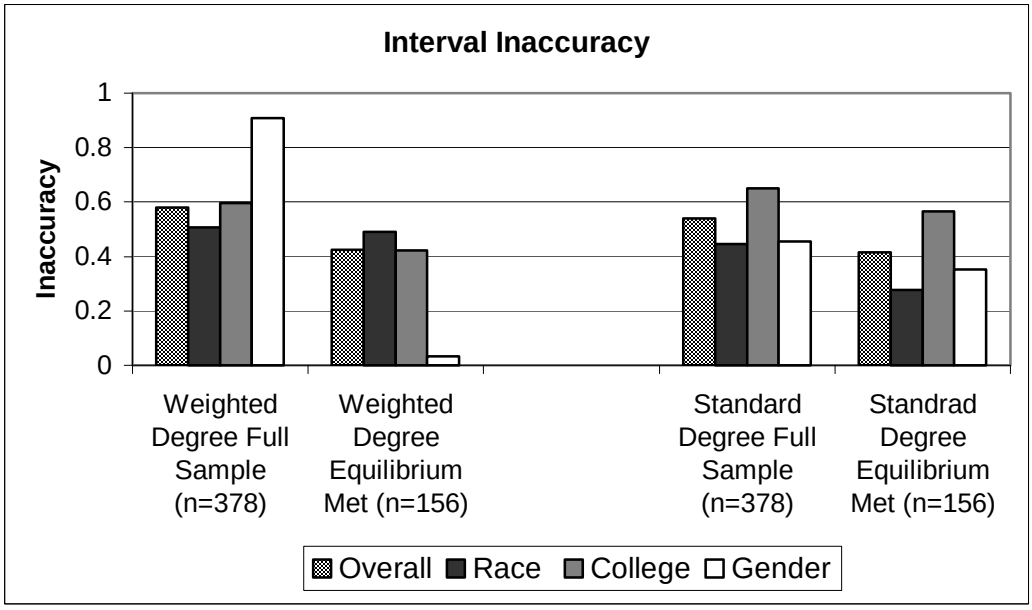




Figure 18: 2008 sample comparison of estimate and interval inaccuracy using only data collected in waves zero through nine to estimate and interval inaccuracy based on the full sample.

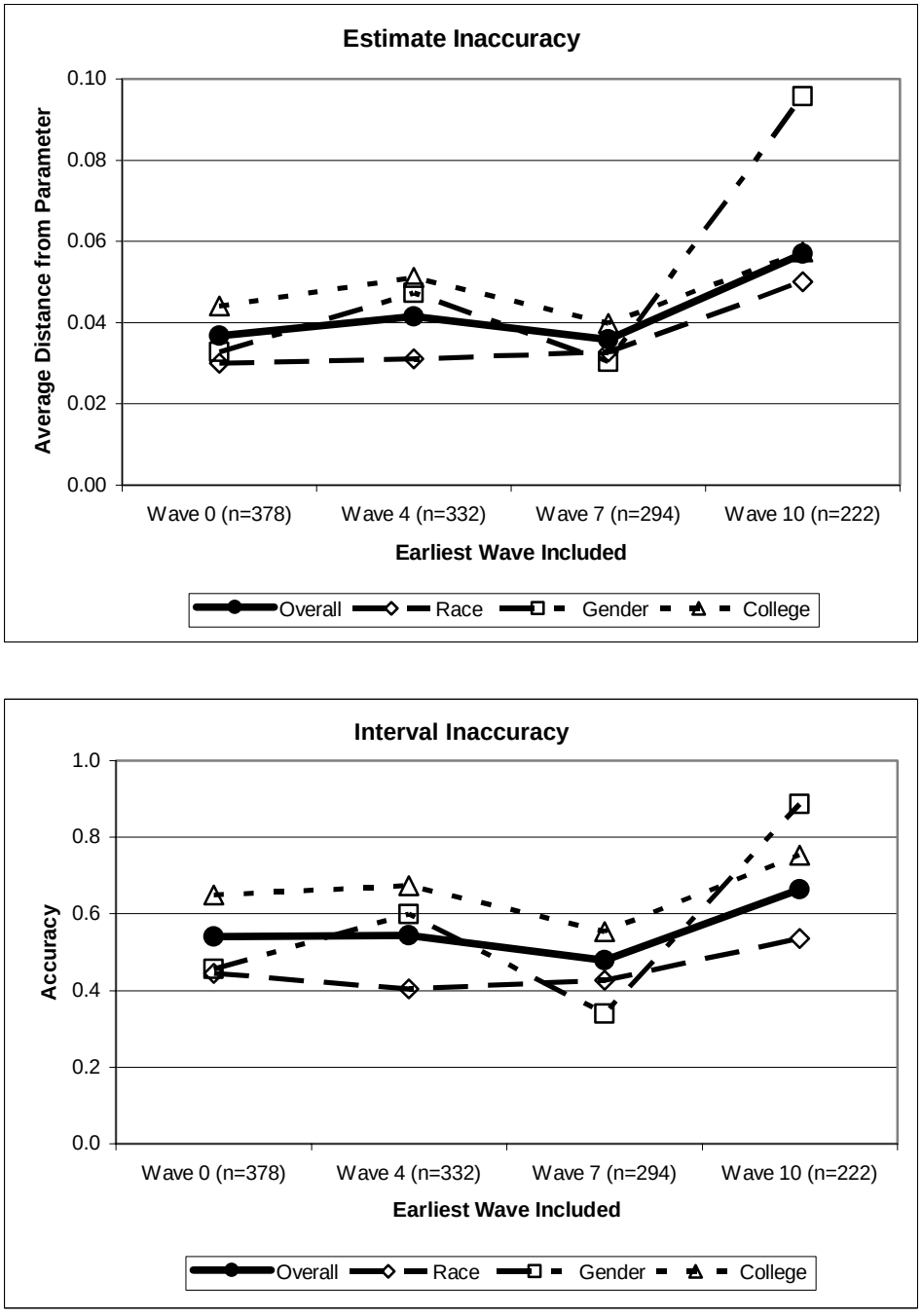
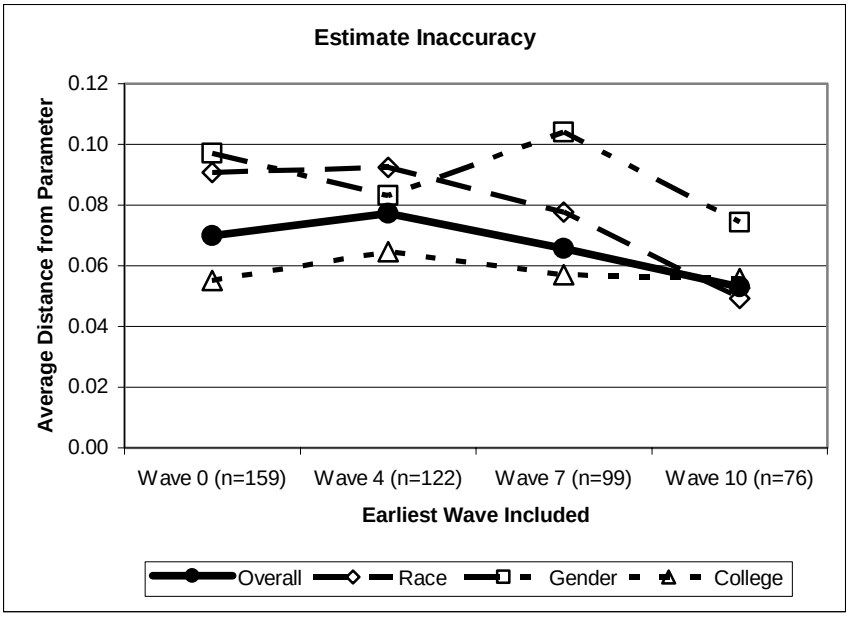



Figure 19: 2008 sample comparison of estimate and interval inaccuracy for analysis based on data starting at different waves of recruitment using the standard degree measure. Sample size used in analysis shown on x-axis.



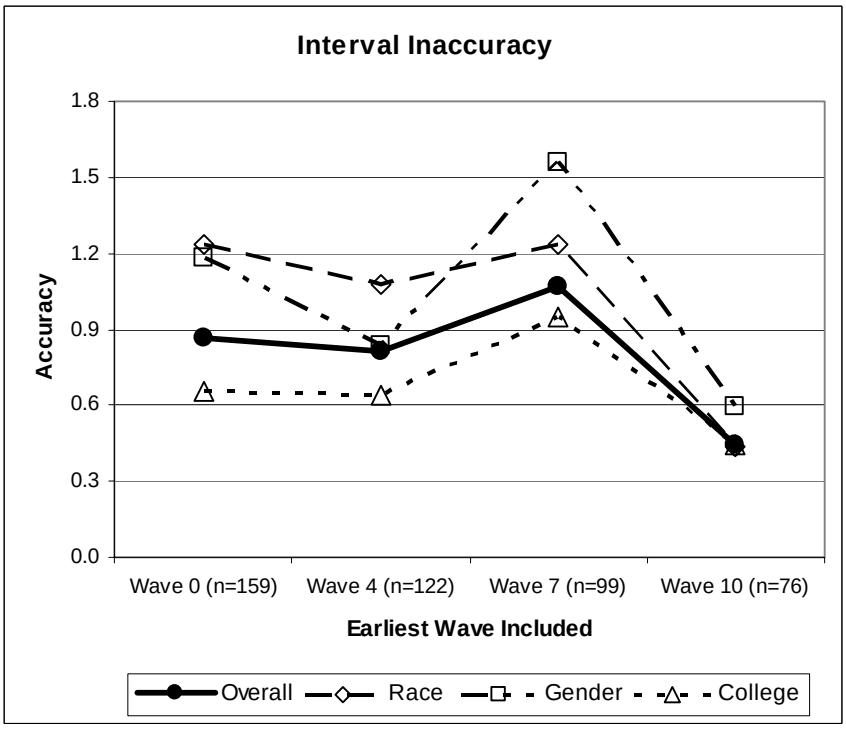




Figure 20: 2004 sample comparison of estimate and interval inaccuracy for analysis based on data starting at different waves of recruitment using the standard degree measure. Sample size used in analysis shown on x-axis.

On the other side of the equilibrium debate lies the question of whether early, out-of-equilibrium waves should be removed from analysis. The above results suggest that early waves do not impose a strong bias on estimates, however, theoretically, recruitments made after equilibrium is reached represent a random sample of network ties, while those made before equilibrium may be biased by seed selection. The relevant question, therefore, is whether the gain from analyzing only in-equilibrium data is greater than the loss inflicted by reduction in sample size when early waves are thrown out?

Figure 19 and Figure 20 show 2008 and 2004 RDS estimates based on the standard degree measure calculated for data starting at wave zero, four, seven, and 10, respectively. Results based on other measures of degree suggest similar conclusions. The results for 2008 and 2004 differ considerably. In 2008, estimate and interval inaccuracy remain relatively stable until only wave 10 and higher data are included, at which point both estimates and intervals become less accurate. Based on the 2008 data alone, the trade-off between sample size and equilibrium appears to favor keeping all data to maximize sample size.

The effect of excluding early-wave data is more complex in the 2004 sample. Overall estimate inaccuracy based on 2004 standard degree suggests RDS estimation may improve as early waves are excluded from analysis. However this conclusion is not supported by interval inaccuracy measures, which exhibit no consistent trend. At least two possible factors help explain such erratic results. First, as early waves are excluded the already modest sample size is reduced by at least 20 respondents at each step. When only data collected in wave seven or higher are used, the analysis is based on only 99 respondents, at wave ten the sample size is 76. Second, as more data are removed from analysis, estimates of variables with small proportions, which tend to be most problematic, can not be calculated and do not influence inaccuracy (see Table 5).

In summary, the results do not provide sufficient evidence to suggest that including out-of-equilibrium data in RDS analysis has a significant negative effect on RDS estimation. However, it is important to note that these results are intended as a practical guide to researchers seeking to get the most out of their data and not as a theoretical conclusion on the importance of equilibrium. Therefore, theoretical or computation work observing the effect of equilibrium in a vacuum that finds improved estimates when early waves are excluded is not necessarily flawed.

### *Discussion*

Overall, results from this study suggest that RDS estimates are reasonable, but better methods of estimating variance of the estimates are needed. The study has several limitations.

First, the study population, which was chosen because population parameters are easily available, is not representative of populations commonly studied using RDS, which are often stigmatized, hard-to-reach, and at risk for HIV/AIDS. Furthermore, while most RDS studies use recruitment coupons and include face-to-face or computer aided interviews conducted at a location operated by researchers, this study used WebRDS where participation and recruitment can occur from a personal computer. Thus the study lacks some difficulties common to other RDS studies such as risk to the respondent of being identified as a stigmatized population member or transportation to and from a survey site. However, because the study and analysis presented does not use any methods or information beyond that normally available to RDS researchers during data collection or analysis, it suffers from many of the same problems found in other studies and the findings presented here regarding variance, degree measures, and out-of-equilibrium data are likely applicable to a wide range of real world RDS applications.

Second, reliable institutional data exist for gender and college within the university, but institutional data for the race variable did not match up perfectly with study categories in 2004. In the 2008 sample, categories for foreign national and non-response were added to the survey, however a greater proportion of students are apparently willing to provide information regarding race on a survey than on official university documents. In addition, the institutional category "foreign national" may represent a broader subset of students than that used on the survey (non-U.S. citizen or permanent resident).

A general limitation of RDS is its youth as a sampling and analysis method. For example, while considerable work on point estimates has been done, other parameters of interest to researchers, such as correlation coefficients or regression coefficients remain underdeveloped. This chapter addresses the one parameter that is well understood, however more research is needed to further develop other RDS specific parameter estimation.

Finally, while beyond the scope of the paper, the third assumption of RDS, that sampling is with replacement, deserves further investigation. Most RDS studies, including this one (see Chapter Two), argue that because the sampling fraction is small relative to the study population, a sampling with replacement approach is appropriate. However, each observation is taken from a pool of respondents known to the recruit, not the entire population. Under the reciprocity assumption, a respondent's recruiter is in that pool, thus if the pool is small, the removal of one's recruiter may be significant for analysis. Further research is needed to confirm or disconfirm this hypothesis.

### *Conclusion*

This paper makes three contributions to empirical RDS analysis. First, estimates and variance calculated using RDS I and RDS II methods are compared. RDS estimates calculated using RDS I and RDS II coincide closely, but variance estimation, especially for small groups, is problematic in opposite directions. The bootstrap algorithm used to generate RDS I confidence intervals is found to underestimate variance of groups making up less than 10% of the population to such an extent that confidence intervals often fail to capture population parameters. Conversely, intervals calculated using RDS II's analytical variance estimate easily capture population parameters, but tend to overestimate variance of small groups to such an extent that design effects above 20 can be observed.

Second, RDS estimates are found to be relatively robust against varying measures of individual degree. The standard degree measure currently included in most RDS studies is found to be among the better, but not best, performing degree measures. The study finds respondents disproportionately recruit close tie individuals, such as those with whom they discuss important matters.

Finally, the results do not provide sufficient evidence to suggest that including out-of-equilibrium data in RDS analysis has a negative effect on RDS estimation. There was not sufficient evidence to show estimates generated using predominantly out-of-equilibrium data are problematic. Furthermore, excluding early waves of recruitment did not improve estimates, suggesting that the reduction in sample size involved in excluding early waves is not worth the potential benefit to estimates.

# REFERENCES


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

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

Burt, Ronald S. 1985. “General Social Survey Network Items.” *Connections* 8: 119-123.

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

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.

Cornell University. 2008. “Enrollment at a Glance.” Ithaca, NY: Cornell University, Division of Planning and Budget. Available at http://dpb.cornell.edu/F_Undergraduate_Enrollment.htm, accessed March 15, 2008.

Fischer, C. S.1982. *To Dwell Among Others.* Chicago: University of Chicago Press.

Goel, Sharad and Mathew J. Salganik. 2008. “Simulation Studies of Respondent-Driven Sampling Under (Reasonably) Realistic Conditions.” *unpublished manuscript*.

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

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.

Marsden, Peter V. 1987. “Core Discussion Networks of Americans.” *American Sociological Review* 52: 122-131.

------. 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, Lynn Smith-Lovin, and Matthew E. Brashears. 2006. “Social Isolation in America: Changes in Core Discussion Networks over Two Decades.” *American Sociological Review* 71: 353-375.

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.

Volz, Erik, Cyprian Wejnert, Ismail Deganii, and Douglas D. Heckathorn. 2007. Respondent-Driven Sampling Analysis Tool (RDSAT) Version 6.0.1. Ithaca, NY: Cornell University.

Volz, Erik, and Douglas D. Heckathorn. 2008. “Probability-Based Estimation Theory for Respondent-Driven Sampling.” *Journal of Official Statistics* 24: 79-97.

Wejnert, Cyprian and Douglas D. Heckathorn. 2008. “Web-Based Networks Sampling: Efficiency and Efficacy of Respondent-Driven Sampling for Online Research.” *Sociological Methods and Research* 37: 105-134.

# CHAPTER 4

# THE DUAL HOMOPHILY MODEL

## *Introduction*

The notion that "birds of a feather" flock together has been a well known element of popular culture and social science for decades. Strong instances of homophily have been found according to race and ethnicity (e.g. Mollica et al. 2003; Moody 2001; Marsden 1987), age (e.g. Feld 1982; Fischer 1982), gender (e.g. Marsden 1987; Tuma and Hallinan 1979), educational aspiration (e.g. Tuma and Hallinan 1979; Kandel 1978), drug use (e.g. Heckathorn and Rosenstein 2002; Kandel 1978), political identification, religion, behavior, attitudes, and behavior (McPherson et al. 2001). Homophily has also been found to exist where it is unknown to the actors themselves. For example, in a study of jazz musicians, where information about personal income is a tightly held secret, Heckathorn and Jeffri (2003) found that actors are homophilous on income, despite having no information on the income of others in the network. However, while empirical studies of homophily are abundant, to my knowledge, theoretical discussion of its measurement and makeup is limited.

In this chapter, I build on the work of Fararo and Sunshine (1964) and present a dual homophily model that divides homophily into two components, one due to relational preference and one due to differential degree. I present formulas for these components, show how homophily can be expressed as a function of them, and provide an empirical example of how they can be used to enrich research on social networks using an RDS sample of university undergraduates.

### *Measures of Affiliation*

Studies of affiliation patterns have focused on either of two measures of affiliation (Blau 1977; McPherson et Al. 2001). One, that we will term *Blau Homophily,* focuses only on crosscutting ties— that is, the number of social network linkages to out-group members relative to the total number of linkages to in-group and out-group members. This is the approach embraced by Blau (1977, 1994) and others operating in the macrostructural tradition. This is also the focus of Putnam's (2000) studies of social capital, in which within-group ties are viewed as the source of social cohesion (i.e., *bonding social capital*), and cross-group ties promote cross-group integration (i.e., *bridging social capital*).

A problem faced by this approach is that cross-group ties reflect not merely social cohesion or solidarity but also relative group size. Consider, for example, a term devoid of social meaning: whether one was born in an odd- or an even-numbered month. Given that about the same number of people is born each month, and that birth month is socially irrelevant, one can expect that the odd and even groups will have about 50% in-group ties and 50% out-group ties. In Putnam's terms, bridging and bonding social capital are equal. If one then considers only those born in December, the results are different, with 1/12th in-group ties and 11/12ths out-group ties. Here, bridging social capital is greater than bonding social capital, so the December group appears atomized. In contrast, the non-December group has mostly in-group ties, so it could appear xenophobic. Obviously, such conclusions are nonsense because the characteristic is devoid of social meaning, and the number of in-group versus out-group ties depends merely on relative group size. In recognition of this issue, Blau (1994) and his associates have properly focused much of their attention on the association between group size and crosscutting ties. For example, the principal

hypothesis of Blau's macrostructural analysis is that as group size decreases, the proportion of crosscutting ties will increase.

Formal network analysts have adopted a more informationally demanding approach to analyzing affiliation patterns (Coleman 1958; Fararo and Sunshine 1964; Rapoport 1979; Fararo and Skvoretz 1984) that is based not only on measures of cross-cutting ties, but also relative group sizes. This more complex approach to analyzing affiliation patterns may be termed *Network Homophily*. It uses as a baseline the notion of a randomly connected network, in which structure is defined as existing within a network only when affiliation patterns depart from those that would be produced by random mixing. The result is a means for measuring the strength of in-group affiliation that compensates for the effects of group size. That is, random mixing would create a non-structured network in which the proportion of in-group ties for each group equals that group's proportional size. Consequently, each group's proportion of cross-cutting reflects the proportional size of other groups relative to the referent group.

Coleman (1958) and Fararo and Sunshine (1964) independently defined equivalent indices of network homophily. The index value is one if all ties are to the in-group and zero if ties are formed by random mixing; intermediate values have a straightforward interpretation. For example, a value of 0.55 indicates that the actor forms ties as though 55% of the time a tie is formed to the in-group, and 45% of the time a tie is formed irrespective of group membership. As thus defined, the index provides a measure of the strength of in-group affiliation bias that has the benefit of compensating for the effects of group size. Subsequently, this index was generalized to accommodate affiliation based on complementarity, in which ties are preferentially formed not to the in-group but to the out-group (Heckathorn 2002). The range of index values was thereby extended to negative values, where an index value of –1

indicates that all ties are formed to the out-group, and intermediate values have an interpretation consistent with that for positive values. For example, a value of -0.33 indicates that networks are structured as though 33% of the time a link is formed to out-group members, and the other 67% of the time a link is formed independent of group identity[9].

**Network Homophily**

In this context, network homophily is equivalent to what is referred to as "homophily" in Chapter One. Consequently, the equations for computing network homophily (Heckathorn 2002) are presented in equation (1.8). Homophily is positive if $S_{XX} > P_X$ and negative if $S_{XX} < P_X$.

**Relational Homophily**

By using proportional group size to estimate availability, the network homophily measure represents an improvement to the Blau homophily measure that was based only on cross-cutting ties. This enhancement, however, assumes that there is no difference in average degree across groups (Fararo and Sunshine 1964, p.63), as is the case for terms such as marriage in a monogamous system, were the network size for a married individual is one. For example, if group X and group Y each make up 50% of a population, it is assumed that they are equally available for tie formation. However, if members of group X have four times as many network ties as members of group Y, then group X could be considered more available because 80% of ties within the network are connected to members of group X.

[9] An alternate metric, reviewed in detail by Gower and Legendre (1986) and used by Krackhardt (1990) and Ibarra (1992), measures homophily by calculating a correlation coefficient based on counts of existing and non-existing in-group and cross-group ties. This method, however, is limited because it requires complete network data and is restricted to dichotomous variables.

To control for such bias in differential degree, the proportional number of ties associated with a group, $S_X$, can be substituted for the proportional size of the group equation (1.8).

$$S_X = \frac{T_X}{T} \tag{4.1}$$

where $T$ is the total number of ties in the network, $T_X$, is the number of ties connected to members of group X and $S_X$ is as defined above. Because homophily calculated using this baseline is a direct measure of affiliation preferences of the group it is termed the *relational homophily*, $RH_X$, of the group.

$$\begin{aligned} RH_X &= \frac{S_X - S_{XX}}{S_X - 1} \;\; if \;\; S_X \leq S_{XX} \\ RH_X &= \frac{S_{XX}}{S_X} - 1 \;\; if \;\; S_X > S_{XX} \end{aligned} \tag{4.2}$$

This measure, which is also bonded between -1 and 1, represents the homophily of a group after controlling for both population proportion and differential degree. If all groups have equal average degree then: $H_X = RH_X$.

**Degree Homophily**

While useful in some settings, the treatment of differential degree as an availability bias to be accounted for when measuring homophily may not be appropriate or desirable in other cases. A common misconception among homophily research is that a group that exhibits high homophily does so at the expense of forming ties with out-group members and that homophily therefore produces a social barrier to cross-group interaction and integration. While a social barrier to cross-group interaction does exist when homophily approaches unity, cross-group interaction is not

blocked in most cases of homophily. In order for moderate homophily to produce a barrier the formation of a tie to one individual must produce an opportunity cost equal to one tie formed to someone else. In other words, when a relationship is made it uses up one of a fixed number of ties those individuals will maintain. In some instances, such as monogamous marriage or teammates on a fixed roster, this is certainly the case; however, in many social networks individuals are not strictly limited in the number of ties they can make.

This is not to say that the formation of a tie does not incur an opportunity cost; on the contrary, tie formation incurs an opportunity cost that is a highly variable function of the social environment and the individuals involved. Often, the ease with which one tie can be maintained, and therefore the total number of ties that can exist at one time is directly related to the presence or absence of other ties in the network (Simmel 1964; Wasserman and Faust 1994). Therefore, relational preference is not only constrained by relative group degree, but also directly influences it. Consequently, a complete quantification of homophily requires not only a measure accounting for differential degree, but also a direct measure of homophily accounted for by differential degree. The latter is termed *degree homophily*.

**Social Capital**

Drawing on Putnam's (2000) terminology, relational homophily serves as an indicator of the type of social capital present within the group. That is, positive relational homophily represents a high degree of social cohesion or bonding social capital, while negative relational homophily suggests a high level of out-group preference, or bridging social capital. The benefit of such a contextualization of social capital is that it represents a sliding scale on which neither extreme is beneficial; what Woolcock and Narayan (2000) would call the trade off between playing offense (bridging social capital) and defense (bonding social capital). At the positive extreme,

group members have a high degree of social cohesion, but are virtually isolated from outside influences, interaction, and resources. At the negative extreme, members avoid each other, maximizing access to outside resources at the total expense of having a social safety net associated with social cohesion (Pescosolido and Georianna 1989).

While relational homophily provides a measure of the type of social capital present in a group, degree homophily measures the amount of social capital available to the group relative to other groups. This two-dimensional representation of social capital based on the dual homophily model allows for not only an understanding how much social capital is present within a group, but also it what ways is it harmful or beneficial. Employing a sports analogy, degree homophily quantifies the level of resources or depth a team has relative to its competition, while relational homophily measures to what extent those resources are concentrated in offense or defense.

### *Measuring Homophily with RDS*

In order to quantify degree homophily I turn to recent theoretical work on RDS disentangling the effects of affiliation and degree on the RDS population estimator (Heckathorn 2007). The advantage of using RDS theory is that a considerable body of work regarding the effect of affiliation and degree on population estimates can be drawn on and applied to homophily estimation. For example, relational homophily is defined above as departure from random mixing where random mixing is defined as the proportional number of ties associated with a group or, equivalently, as the homophily if all groups had equal average degree. In RDS theory, the equilibrium proportion, $E_X$ , of a group serves is an estimate of population proportion under the assumption that all groups have equal average degree (Heckathorn 2002). Consequently, a new formula for relational homophily, based on RDS theory is defined as follows:

$$RH_X^{RDS} = \frac{E_X - S_{XX}}{E_X - 1} \quad if \quad E_X \leq S_{XX}$$
$$RH_X^{RDS} = \frac{S_{XX}}{E_X} - 1 \quad if \quad E_X > S_{XX} \tag{4.3}$$

In order for equation (4.3) to hold, $E_X$ must be shown to equal $S_X$ as defined in equation (4.2) above.

From equation (4.1):

$$S_X = \frac{T_X}{T} \tag{4.4}$$

Expanding $T_X$ and $T$:

$$S_X = \frac{P_X D_X}{\sum_i (P_i D_i)} \tag{4.5}$$

where $P_X$ is the population proportion of X and $D_X$ is the average degree of members of X.

Equation (4.5) provides a measure of the proportional number of ties held by a group that can be calculated using RDS measures and used as a baseline for measuring relational homophily. In theory, this should produce the same results as the RDS relational homophily measure, suggesting that:

$$S_X = E_X \tag{4.6}$$

where $E_X$ is the equilibrium proportion of group X from RDS. Heckathorn (1997) shows that in a two category system:

$$E_X = \frac{S_{YX}}{S_{YX} + S_{XY}} \tag{4.7}$$

Equation (4.5) can be expanded based on previous RDS work (Heckathorn 2002, Salganik and Heckathorn 2004):

$$S_X = \frac{D_X \bullet \dfrac{S_{XY} D_Y}{S_{XY} D_Y + S_{YX} D_X}}{D_X \bullet \dfrac{S_{XY} D_Y}{S_{XY} D_Y + S_{YX} D_X} + D_Y \bullet \dfrac{S_{YX} D_X}{S_{YX} D_X + S_{XY} D_Y}} \tag{4.8}$$

which reduces to:

$$S_X = \frac{S_{YX} D_Y D_X}{\left(S_{XY} + S_{YX}\right) D_Y D_X} = \frac{S_{YX}}{S_{XY} + S_{YX}} = E_X \tag{4.9}$$

Thus $S_X = E_X$ and equations (4.3) and (4.2) are equivalent for a two category system. To maintain consistency with RDS terminology, equation (4.3) will be used from here out.

Following Heckathorn's (2007) dual weights estimation model for RDS data, degree homophily, $DH_X$, of a group can be calculated as follows:

$$\begin{aligned} DH_X &= \frac{P_X - E_X}{P_X - 1} \;\; if \;\; P_X \le E_X \\ DH_X &= \frac{E_X}{P_X} - 1 \;\; if \;\; P_X > E_X \end{aligned} \tag{4.10}$$

As with network homophily and relational homophily, degree homophily is bounded between -1 and 1 and represents the extent to which affiliation departs from random mixing due to differences in average degree across groups.

### *Calculating Homophily from Its Components*

In order to formalize the relationship between network homophily and relational and degree homophilies a formula is needed for network homophily as a function of its components. Following the terminology of Fararo and Sunshine (1964) we can treat each homophily as the probability that an event, defined as a tie being made to the in-group with probability one, occurs. Conversely, the probability that the event does not occur, i.e. the tie is made randomly, is defined by subtracting the event from unity. Thus we can calculate homophily as the difference from unity of the probability that neither degree nor affiliation bias events occur, where the probability that neither event occurs is the product of the probabilities of each event not occurring. When both homophily components are of similar sign, that is when $H_X$, $RH_X$, and $DH_X$ are all positive or all negative, this relationship is straight forward as shown in equation (4.11). All homophily events are positive when $S_{XX} > E_X > P_X$, and negative when $P_X > S_{XX} > E_X$.

$$\begin{aligned} H_X &= 1-\left(\left(1-\left|RH_X\right|\right)\bullet\left(1-\left|DH_X\right|\right)\right) \;\; if \;\; S_{XX} > E_X > P_X \\ H_X &= -1+\left(\left(1-\left|RH_X\right|\right)\bullet\left(1-\left|DH_X\right|\right)\right) \;\; if \;\; P_X > S_{XX} > E_X \end{aligned} \tag{4.11}$$

In this equation the absolute value of relational and degree homophily is used in order to preserve the properties of a probability. This formula can be simplified to show that homophily is the sum of relational and degree homophilies reduced by the interaction of effects.

$$H_X = RH_X + DH_X - \left(\left|RH_X\right| \bullet \left|DH_X\right|\right) \quad if\ S_{XX} > E_X > P_X$$
$$H_X = RH_X + DH_X + \left(\left|RH_X\right| \bullet \left|DH_X\right|\right) \quad if\ P_X > S_{XX} > E_X \tag{4.12}$$

The relationship between homophily and its components becomes complicated when relational and degree homophilies have opposite sign, an event that occurs frequently. For example, in a two category system, both groups will always have relational homophilies of similar sign and degree homophilies of opposite sign. The complication occurs due to differences in the criteria deciding the equation used to calculate each type of homophily. The RH equation is determined by the relationship between $E_X$ and $S_{XX}$, DH is determined by $P_X$ and $E_X$, and $H_X$ is determined by $P_x$ and $S_{XX}$, thus there are six possible combinations and six different expressions for the relationship between homophily and its components[10].

Intuitively, altering relative average degree is more complicated than altering affiliation preference. For example, when a group adds an ingroup tie, DH and AH are both increased some constant amount. However, when a group adds an outgroup tie, AH is increase a constant amount, but DH is increased a variant amount depending on the number of other groups there are in the system. In a two category there is no change in DH because the added tie increases the degree of both groups in the system by an equal amount. In an N-category system, adding an outgroup tie increases the overall average degree of the out-group by a factor of $1/(n-1)$ as much as it increases average degree for the in-group.

[10] An alternate method of calculating RH and DH is to determine the proper equation using the criteria of homophily, i.e. Px relative to Sxx. Under this method, equation (4.12) holds in all cases. However, RH and DH are not bounded between -1 and 1.

The four remaining equations for expressing homophily as a function of its components are:

$$H_X = RH_X - \frac{P_X \bullet DH_X - P_X \bullet RH_X \bullet DH_X}{P_X - 1} \qquad if\ S_{XX} > P_X > E_X$$

$$H_X = DH_X - RH_X - RH_X \bullet DH_X + \frac{RH_X}{P_X} \qquad if\ P_X > S_{XX} > E_X \tag{4.13}$$

$$H_X = DH_X + RH_X \bullet DH_X - \frac{P_X \bullet RH_X}{P_X - 1} \qquad if\ E_X > S_{XX} > P_X$$

$$H_X = RH_X - DH_X - RH_X \bullet DH_X + \frac{DH_X + RH_X \bullet DH_X}{P_X} \quad if\ E_X > P_X > S_{XX}$$

Unfortunately, these relationships do not simplify or present an intuitive mathematical pattern. Further research is needed to better understand these relationships.

### *Practical Applications and Contributions of the Dual Homophily Model*

#### Finer Homophily Measure

The dual homophily model's primary contribution is a finer measure of homophily and social capital. By separating out the effects of affiliation preference and degree on network homophily it is possible to tell a more complete story about the structure of social interaction. Figure 21 shows three analyses, based on the 2008 data, which exemplify the benefits of the dual homophily model.

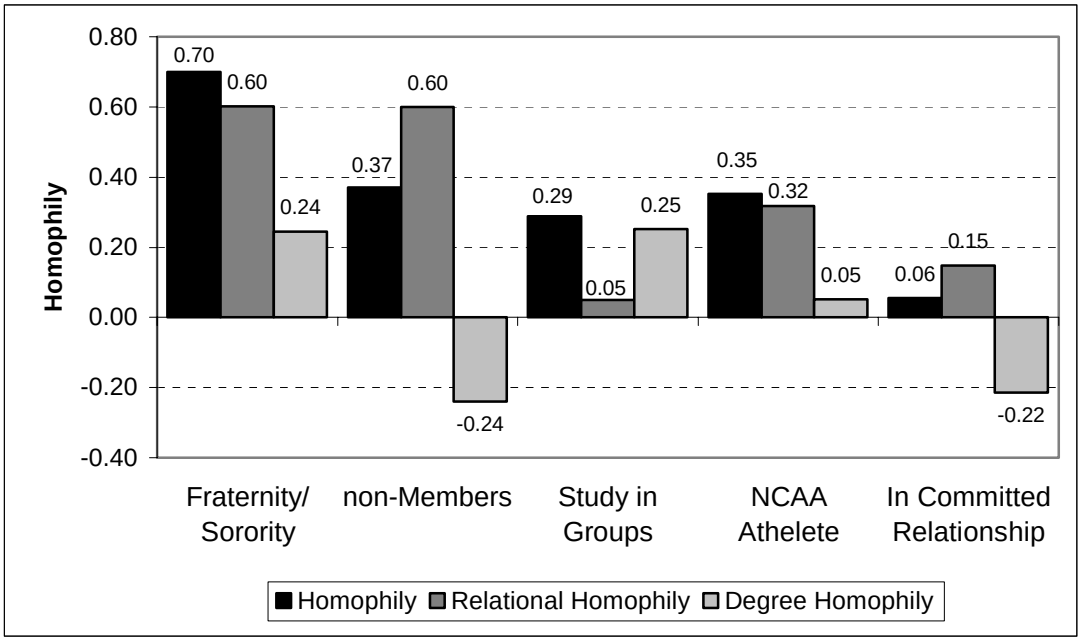



Figure 21: Four analyses of homophily that exemplify advantages of the dual homophily model. First, low relative degree of non-fraternity or sorority members masks high in-group preference. Second, while students who study in groups and NCAA athletes display similar homophily, the underlying association trends differ greatly. Finally, relational and degree forces acting in opposite directions combine to give the appearance that being in a committed relationship does not affect social associations.

First, a comparison of fraternity/sorority members to non-members reveals that each group has a high level of relational homophily (0.6) that blocks interaction and resource sharing between the groups. However, fraternity/sorority members, by maintaining a higher level of degree homophily (0.24) enjoy more social capital and status than non-members.

Second, a comparison of network homophily of students who regularly study in groups and those who are NCAA athletes suggests each group has similar network structure, with observed network homophily of approximately 0.3. However, the source of this homophily is different in each group. Students who regularly study in groups have trivia relational homophily (RH = 0.05) and substantial degree homophily

(DH = 0.25), suggesting that while their connections have an in-group tendency, it is the result of increased degree and not affiliation preference. On the other hand, NCAA athletes display the opposite trend. They have trivial degree homophily (DH = 0.05), and strong relational homophily (RH = 0.32). Consequently, there is evidence to suggest that students who study in groups maintain contacts in addition to their normal associations, presumably those with whom they study, and enjoy greater access to social resources because of these contacts than their counterparts; however their associations remain balanced so that neither form of social capital dominates. NCAA athletes, on the other hand, associate with each other at the expense of forming ties with non-athletes, providing them with a high level of bonding capital at the expense of access to outside resources. Such distinctions can not be made using the coarser network homophily measure.

In the final example, students in committed relationships have trivial homophily (0.06), suggesting that being in a committed relationship does not influence association patterns. However, the dual homophily model reveals a deeper story in which students in committed relationships maintain substantially smaller networks (DH = -0.22) than other students and have a slight preference toward in-group connections (RH = 0.15). Thus, while these students would otherwise appear integrated into the greater population, the dual homophily model reveals a pattern of in-group bonding and low degree consistent with an isolated or marginalized group.

**Ease of Calculation**

While the dual homophily model as a whole provides an improved understanding and quantification of social capital and network segmentation. A major advantage of relational homophily, as presented here, is that it can be calculated using only the transition probabilities across groups, greatly reducing data constraints.

For example, Laumann and Youm (1999) use a complex analysis to provide a social networks explanation of the disparity between STD prevalence among whites and African Americans. They hypothesize that the disparity may be due to differences in sexual contact between individuals who are "peripheral" (those who have had one partner in the past 12 months), "adjacent" (those who have had two or three partners in the past 12 months), and "core" (those who have had four or more partners in the past 12 months) to the sexual network. Specifically, white peripheral members have much lower rates of sexual contact with more promiscuous adjacent and core members than their African American counter parts. While their analysis goes beyond simply documenting the network structure, applying *relational affiliation*, $RA_{XY}$ , a generalization of relational homophily that calculates relational preference of all groups for all groups, to their data provides a clear simple picture of their finding. The formula for relational affiliation parallels that of affiliation preference presented in equation (1.9):

$$\begin{aligned} RA_{XY} &= \frac{E_X - S_{XY}}{S_{XY} - 1} \quad if \quad E_X \le S_{XY} \\ RA_{XY} &= \frac{S_{XY}}{E_X} - 1 \quad if \quad E_X > S_{XY} \end{aligned} \tag{4.14}$$

Table 7 shows relational affiliation calculated using only the contact matrix provided in Table 2 of Laumann and Youm (1999). The network patterns are striking and consistent with Laumann and Youm's (1999) analysis. First, white and African American sexual networks do not intersect, as evidenced by strong negative relational affiliation across race (non-shaded area), a network condition required to maintain racial prevalence disparity. Second, Laumann and Youm's (1999) key finding regarding differences in white and black peripheral members' sexual contacts is

clearly visible in Table 7. White peripherals display highly negative relational affiliation with whites who have had more than one partner in the past 12 months (top-left, bold, shaded area). This network barrier protects them from risk associated with the behaviors of more promiscuous whites. African American peripherals, on the other hand, have near zero relational affiliation scores with other African Americans, a pattern consistent with random mixing (lower-right, bold, shaded area). Thus, while both white and black peripherals prefer in group sexual contact, white sexual contact patterns form a barrier to disease spread from higher risk individuals, while black sexual contact patterns do not.

Table 7: Relational affiliation patterns for sexual contact by race and location in network based on Laumann and Youm (1999) data. WP = White periphery; WA = White adjacent; WC = White core; AP = African American periphery; AA = African American adjacent; AC = African American core.

| | WP | WA | WC | AP | AA | AC |
|---|---|---|---|---|---|---|
| WP | 0.79 | **-0.66** | **-0.81** | -0.89 | -0.99 | -0.98 |
| WA | **-0.66** | 0.34 | 0.26 | -0.95 | -0.86 | -0.72 |
| WC | **-0.81** | 0.30 | 0.36 | -0.97 | -0.89 | -0.77 |
| AP | -0.89 | -0.95 | -0.97 | 0.75 | **0.04** | **0.03** |
| AA | -0.99 | -0.86 | -0.89 | **0.06** | 0.40 | 0.36 |
| AC | -0.98 | -0.72 | -0.77 | **0.05** | 0.40 | 0.29 |

The above reanalysis of Laumann and Youm's (1999) data replicates their findings in a way that is clearer, simpler, and requires less data than the analysis they present.

### *Conclusion*

This chapter extends theoretical work on homophily through the dual homophily model. By quantitatively breaking network homophily into its components I am able to address more detailed questions regarding the interaction of social network structure and group level social capital. Specifically, relational homophily is

a term that captures the extent to which affiliation is based on like associating with like after controlling for differential group size and degree; consequently it measures the relative levels of two opposing categories of social capital in the group. If relational homophily is near 1, the group has high bonding social capital, but low bridging social capital. If it is near -1, the group has high bridging, but low bonding social capital. A score near 0 suggests the group maintains a good balance of each type of capital. In contrast, degree homophily is a term that captures the extent to which differences in affiliation patterns reflect differential status, as reflected in differences in network sizes; consequently it measures a group's total level of social capital relative to other groups in the population, a form of social inequality. Thus the dual homophily model serves both as an improved methodological tool for the study of network structure and a two level model for thinking about the magnitude and direction of groups' social capital.

# REFERENCES


Blau, Peter M. 1977. *Inequality and Heterogeneity.* New York: Free Press.

------. 1994. *Structural Context of Opportunities.* Chicago, IL: University of Chicago Press.

Coleman, J. S. 1958. "Relational analysis: The study of social organizations with survey methods." *Human Organizations* 17:28-36.

Fararo, T. J. and M. H. Sunshine. 1964. *A Study of a Biased Friendship Net.* New York: Syracuse University Press.

Fararo, Thomas J., and John Skvoretz. 1984. “Biased Networks and Social Structure Theorems: Part II.” *Social Networks* 6:223–58.

Feld, Scott L. 1982. “Social structural determinants of similarity among associates” *American Sociological Review* 47:797-801

Fischer, C. S. 1982. *To Dwell Among Others.* Chicago: University of Chicago Press.

Gower, J. C. and P. Legendre. 1986. “Metric and Euclidean Properties of Dissimilarity Coefficients.” *Journal of Classification* 3: 5-48.

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

------. 2002. "Respondent-driven sampling II: Deriving valid population estimates from chain-referral samples of hidden populations." *Social Problems* 49:11-34.

------. 2007. “Extensions of respondent-driven sampling: Analyzing continuous variables and controlling for differential recruitment.” *Sociological Methodology* 37:151-208.

Heckathorn, D. D. and J. Jeffri. 2003. "Jazz networks: Using respondent-driven sampling to study stratification in two jazz communities." Presented at the

*Annual Meeting of the American Sociological Association.* Atlanta, GA. August 2003.

Heckathorn, D. D. and J. E. Rosenstein. 2002. Group Solidarity as the Product of Collective Action: Creation of Solidarity in a Population of Injection Drug Users. *Advances in Group Processes,* Vol. 19.

Ibarra, Herminia. 1992. ”Homophily and Differential Returns: sex Differences in Network Structure and Access in an Advertising Firm.” *Administrative Science Quarterly* 37: 422-447.

Kandel, D. B. 1978 “homophily, selection, and socialization in adolescent friendships” *American Journal of Sociology* 84: 427-436.

Krackhardt, David. 1990. “Assessing the Political Landscape: Structure, Cognition, and Power in Organizations.” *Administrative Science Quarterly* 35: 342-369.

Laumann, Edward and Yoosik M.A. Youm. 1999. “Race/Ethnic Group Differences in the Prevalence of Sexually Transmitted Diseases in the United States: A Network Explanation.” *Sexually Transmitted Diseases* 26: 250-261.

Marsden, P. V. 1987 Core Discussion Networks of Americans. *American Sociological Review* 52: 122-131.

McPherson, M., L. Smith-Lovin, and J. M. Cook. 2001. "Birds of a feather: Homophily in social networks." *Annual Review of Sociology* 27:415-444.

Mollica, Kelly, Gray, Barbara, and Trevino, Linda. (2003) “Racial Homophily and Its Persistence in Newcomers’ Social Networks” *Organization Science.* 14: 123-136.

Moody, James. 2001. “Race, School Integration, and Friendship Segregation in America” *American Journal of Sociology* 107: 679-716.

Pescosolido, Bernice A. and Sharon Georgianna. 1989. “Durkheim, Suicide, and Religion: Toward a Network Theory of Suicide.” *American Sociological Review* 54: 33-48.

Putnam, Robert D. 2000. Bowling Alone : The Collapse and Revival of American Community. New York, NY : Simon & Schuster.

Rapoport, Anatol. 1979. “A Probabilistic Approach to Networks.” *Social Networks* 2:1–18.

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

Simmel, Georg. 1964. *The Sociology of George Simmel.* London: Free Press of Glencoe. Edited and translated by Wolff, K. H.

Tuma N. B. and N. Z. Hallinan. 1979. “The effect of sex, race, and achievement on schoolchildren’s friendships.” *Social Forces* 57: 1265-1285.

Wasserman, S. and K. Faust. 1994. *Social Network Analysis*. Cambridge, MA: Cambridge University Press.

Woolcock, Michael and Deepa Narayan. 2000. “Social Capital: Implications for Development Theory, Research, and Policy.” *The World Bank Research Observer* 15: 225-249.