Forthcoming, *Sociological Methodology*

AN EMPIRICAL TEST OF RESPONDENT-DRIVEN SAMPLING: POINT ESTIMATES, VARIANCE, DEGREE MEASURES, AND OUT-OF-EQUILIBRIUM DATA*

Cyprian Wejnert

Cornell University

Word Count: ~12,600

* This research was made possible by grants from the National Science Foundation, the National Institutes of Health, and the Centers for Disease Control. I thank Douglas Heckathorn, Lisa Johnston, Matthew Salganik, Michael Spiller III, Erik Volz and two anonymous reviewers for helpful comments and advice. Please address correspondence to Cyprian Wejnert, 343 Uris Hall, Cornell University, Ithaca, NY 14853; e-mail: cwejnert@gmail.com.

Abstract:

This paper, which is the first large scale application of Respondent-Driven Sampling (RDS) to non-hidden populations, tests three factors related to RDS estimation against institutional data using two WebRDS samples of university undergraduates. First, two methods of calculating RDS point estimates are compared. RDS estimates calculated using both methods coincide closely, but variance estimation, especially for small groups, is problematic for both methods. In one method, the bootstrap algorithm used to generate confidence intervals is found to underestimate variance. In the other method, where analytical variance estimation is possible, confidence intervals tend to overestimate variance. Second, RDS estimates are found to be robust against varying measures of individual degree. Results suggest the standard degree measure currently employed in most RDS studies is among the best performing degree measures. Finally, RDS is found to be robust against the inclusion of out-of-equilibrium data. The results show that valid point estimates can be generated with RDS analysis using real data, however further research is needed to improve variance estimation techniques.

Introduction

Traditionally, sampling hidden populations - populations for which constructing a sampling frame is infeasible - has proven challenging to researchers interested in collecting probability samples. Respondent-Driven Sampling (RDS), a new network-based (i.e. snowball-type) sampling method, has been proposed as a way to sample and analyze hidden populations (Heckathorn 1997). RDS is now used to study a wide range of hidden populations in over 30 countries (Malekinejad et al. 2008). Network-based designs, which were originally introduced for the study of social networks by Coleman (1958), start with a modest number of initial respondents, or *seeds,* who provide researchers with information on their network connections; these connections then form the pool from which the second wave of respondents is drawn and so on. In RDS, however, respondents are asked to recruit peers directly, allowing referral chains to efficiently and safely penetrate social regions only accessible to insiders. Traditionally, the non-randomness of social network connections has led such samples to be viewed as convenience samples from which unbiased estimation is not possible (Berg 1988). RDS challenges this view by using data gathered during the sampling process to account for non-random social network structure and calculate unbiased population estimates (Salganik and Heckathorn 2004; Volz and Heckathorn 2008).

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 paper 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 test several methods of analyzing RDS data.

Respondent-Driven Sampling

RDS uses data on who recruited whom and the extensiveness of network connections as the basis for calculating population estimates. The theory is based on two observations of sampling connected populations (Heckathorn 2002).

First, if referral chains are sufficiently long, an equilibrium is reached where the sample composition stabilizes and becomes independent of seeds (Heckathorn 2002). Referral chain length is measured in waves, where each wave represents one recruitment or step along the chain. Because seeds do not have recruiters, they are counted as wave zero. Respondents recruited directly by seeds make up wave one. Those recruited by respondents in wave one make up wave two and so on. The number of waves required to reach equilibrium is simulated for each variable using a markov chain model in which the observed sample transition probabilities are used to calculate the sample composition of each wave based on that of the wave before. Equilibrium is reached when the sample composition from one wave to the next differs by less than 2%. To make the estimate conservative, the simulation is initialized with 100% of the sample as a single type. For example, in a sample in which males and females recruit same-sex respondents 60% of the time, the sample composition of each wave is estimated as follows: At wave zero 100% of the sample is male. At wave one 60% of the sample is male and 40% is female. At wave two 52% of the sample is male (36% recruited by males, 16% recruited by

females) and 48% is female (males and females each recruiting 24%). At wave three 50.4% of the sample is male (31.2% recruited by males, 19.2% recruited by females) and 49.6% is female (20.8% recruited by males, 28.8% recruited by females). At wave four the sample is 50.08% male and 49.92% female. Consequently, equilibrium is said to have been reached at wave four, i.e. the number of waves required for equilibrium is four. This method of calculating equilibrium is employed by RDSAT 6.0.1 function "Estimate Number of Waves Required" (Volz et. al. 2007) and used throughout this paper. In this paper "out-of-equilibrium data" refers to data collected in waves before equilibrium is reached while "in-equilibrium data" refers to data collected in waves after equilibrium is reached.

The second observation upon which RDS is based is that a sampling frame can be calculated based on two pieces of information gathered during the sampling process (Heckathorn 2002). First, each recruitment is documented. This provides the basis for controlling for bias introduced by the tendency of individuals to form social ties in a non-random way. Information regarding who recruited whom is used to quantify and account for sample bias due to non-random network structure. Second, respondents are asked how many other members of the target population they know. In a network-based sample the inclusion probability of an individual is proportional to the number of people in the target population he or she is connected to, termed his or her *degree* (Volz and Heckathorn 2008). Salganik and Heckathorn (2004) show that once a sample reaches equilibrium all ties within the target population have equal probability of being used for recruitment.

Consequently, information regarding individual degree is used to account for bias favoring high degree respondents in the sample.

RDS Estimators:

The original RDS estimator, RDS I, introduced by Heckathorn (1997) uses a two stage estimation process where data are used to make inferences about network structure and then these inferences are used to make inferences about the population. Specifically it was shown that under certain assumptions (described below) transition probabilities across groups, estimated by the sample transition probabilities, can be used along with average group degree to calculate unbiased population proportion estimates from network-based data (Salganik and Heckathorn 2004).

Under the reciprocity assumption (discussed below), 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 is averaged such that the group level matrix of who recruited whom, termed the *recruitment matrix,* is symmetric. Transition probabilities based on the data-smoothed recruitment matrix are then combined with the degree estimate (described below) to calculate an estimate for proportional group size, $\widehat{P_X^{RDS\,I}}$ (see also: Salganik and Heckathorn 2004):

$$\widehat{P_X^{RDS\,I}} = \frac{\widehat{S_{YX}}\,\widehat{D_Y}}{\widehat{S_{YX}}\,\widehat{D_Y} + \widehat{S_{XY}}\,\widehat{D_X}}, \qquad (1)$$

where $\widehat{S_{XY}}$ is the data-smoothed proportion of recruitments from group X to group Y and $\widehat{D_X}$ is the estimated average degree of group X.

Using a probability based estimation approach, Volz and Heckathorn (2008) infer that a network-based sample will select individuals in the population with probability proportional to degree and derive a new RDS estimator, $\widehat{P_X^{RDS\,II}}$:

$$\widehat{P_X^{RDSII}} = \left(\frac{n_X}{n.}\right)\left(\frac{\widehat{D.}}{\widehat{D_X}}\right) \tag{2}$$

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, $\frac{n_X}{n.}$, weighted by a correction for network effects, $\frac{\widehat{D.}}{\widehat{D_X}}$. One advantage of the RDS II estimator is that it is calculated directly from the data, removing the middle step of making inference about network structure necessary in RDS I. RDS II also allows for analysis of continuous variables, while RDS I does not.

Volz and Heckathorn (2008) show that RDS I and RDS II estimates converge when the recruitment matrix is symmetric. Thus, when data-smoothing is used, a procedure recommended for all RDS analyses and the RDSAT default, RDS I and RDS II produce equivalent estimates. A major difference is that the mathematical approach used to calculate RDS II estimates allows for analytical variance calculation, while the RDS I approach does not.

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 A recruitments are other As and the current observation is of type A, the algorithm will generate an A 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 resamples has been reached. Confidence interval tails are then taken from the distribution of these bootstrapped estimates. That is, the upper bound of a 95% confidence interval is defined as the point above which 2.5% of the bootstrapped estimate distribution falls. Consequently, the bootstrap algorithm allows for non-symmetric confidence intervals and does not provide a direct estimate of variance. All RDS I estimates and confidence intervals presented here 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.

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

$$Var\left(\widehat{P_X^{RDSII}}\right) = \widehat{V}_1 + \frac{\widehat{P_X^{RDSII}}^2}{n.}\left(\left(1-n.\right) + \frac{2}{n_X}\sum_{i=2}^{n}\sum_{j=1}^{i-1}\left(\widehat{S}_{\cdot}^{\,i-j}\right)_{XX}\right) \tag{3}$$

where

$$\widehat{V}_1 = \frac{\widehat{Var(Z_i)}}{n.} = \frac{1}{n.(n.-1)}\sum_{i=1}^{n}\left(Z_i - \widehat{P_X^{RDSII}}\right)^2 \tag{4}$$

and

$$Z_i = d_i^{-1} \widehat{D}.I_X(i) \quad (5)$$

where d_i is the degree of respondent i, $\widehat{S}.$ 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. All RDS II estimates and intervals[1] presented here are calculated using custom software corresponding to Volz and Heckathorn (2008).

In summary, RDS I and RDS II employ drastically different methods of estimating variance of convergent point estimates. This paper presents the first direct comparison of RDS I and RDS II variance estimation with real data.

Assumptions:

The original proof that the RDS estimator is asymptotically unbiased depends on a set of six assumptions (Salganik and Heckathorn 2004). This number is reduced to five assumptions in a subsequent proof by Heckathorn (2007).

1) Respondents maintain reciprocal relationships with individuals who they know to be members of the target population.
2) Each respondent can be reached by any other respondent through a series of network ties, i.e. the network forms a single component.
3) Sampling is with replacement.
4) Respondents can accurately report their personal network size or equivalently, their degree.

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

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

The first three assumptions specify the conditions necessary for RDS to be an appropriate sampling method for a population. First, in order for recruitment to occur, respondents must have access to other members of the population and be able to identify which of their peers qualify for recruitment. In addition, RDS estimates are based on a network structure in which ties are reciprocal (Heckathorn 2002). Formally, if A recruits B, then there must be a non-zero probability that B could have recruited A. Consequently, the RDS research design includes means for encouraging subjects to recruit their acquaintances or friends rather than strangers by rewarding successful recruiters and making recruitment rights scarce (Heckathorn 1997). That is, when respondents are limited in the number of recruitments they can make and given incentives for successful recruitment they are hesitant to waste valuable recruitments on strangers. Second, the population is assumed to form a single component (Salganik and Heckathorn 2004). In other words, all of the target population must be reachable from any single respondent by following a finite set of network ties. In a random network, a single component forms when individual degrees are large compared to the natural log of the population size (Bollabás 1985, see also Watts and Strogatz 1998). When respondents are allowed to recruit not merely those with whom they have a special relationship, but also any friends and acquaintances they know as members of the target population, then individual degrees are larger than that generally required for a network to form a single large component (Heckathorn 2007). Additionally, since actual social networks are never wholly random, a minimal requirement is that no social or structural barrier that

completely segregates one subgroup of the population from the rest may exist. For example, RDS can not be used to sample across castes in a culture where cross-caste interaction is forbidden. Third, the statistical theory for RDS estimation is based on a sampling-with-replacement scheme (Salganik and Heckathorn 2004). Consequently, the sampling fraction should remain small enough for such a sampling-with-replacement model to be appropriate (Heckathorn 2007).

The final two RDS assumptions are potentially the most problematic. Assumption four requires respondents to accurately provide information on their personal network size, a task that is difficult even for social network experts. Assumption five states that recruitment patterns reflect personal network composition within the target population. That is, RDS assumes respondents recruit as though they were selecting randomly from their personal networks (Heckathorn 2002), however random selection is difficult in many settings (hence the need for complex sampling and analysis techniques). For example, studies of memory suggest recency of contact may influence the accessibility of the name of a peer and therefore the likelihood of an attempted recruitment. Alternately, respondents may recruit the first eligible peer they interact with. While beyond the scope of this paper, it is possible that such non-random selection methods vary across respondents to such an extent that a random model for overall recruitment is appropriate.

More importantly, such questions about assumptions four and five (both discussed in detail below) highlight the major advantage of empirical validation over computational or analytical methods. Specifically, computational and analytical validation pose the question “can valid estimates be made given certain conditions and assumptions?” and

leave open to debate whether these conditions and assumptions are application to real data. Empirical validation avoids this additional step by directly asking "can valid estimates be made using data collected in this way?" If a method is empirically validated, it suggests either that assumptions are being met or that practical limitations to meeting assumptions do not have catastrophic affects on the analysis.

In summary, RDS estimation is based on the insight that members of many populations are known to each other so that the social connections of a small group of members can be followed to produce a population sample. Such samples, commonly known as "snowball" samples, are known to yield unrepresentative data because the social ties used to recruit new participants are not made randomly within the population. However, information on respondents and their connections can be gathered during the sampling process and used to account for any biases resulting from non-random network structure. In this way, RDS researchers are able to calculate unbiased population estimates and make inferences about the population (Heckathorn 2002).

Challenges for RDS

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[2].

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

[2] 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.

group degree estimator that is the ratio of two Hansen-Hurwitz estimators, which are known to be unbiased (Brewer and Hanif 1983). The ratio of two unbiased estimators is asymptotically unbiased with bias on the order of n^{-1}, where n is the sample size (Cochran 1977; Salganik and Heckathorn 2004). This estimator is used to correct for degree bias in RDS estimation of categorical variables.

$$\widehat{D_X} = \frac{n_X}{\sum_{i=1}^{n_X} \frac{1}{d_i}}, \tag{6}$$

where $\widehat{D_X}$ is the average degree of group X and d_i is the self reported personal degree of respondent i (Salganik and Heckathorn 2004).

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. 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 and 2) 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 paper 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

Data:

This paper compares institutional parameters to RDS estimates derived from two WebRDS samples of undergraduates at the same residential university collected in 2008 (369 recruitments, nine seeds, n = 378) and 2004 (150 recruitments, nine seeds, n = 159). WebRDS is an online variant of RDS in which respondents complete an internet survey and recruitment occurs via email. A detailed discussion of the 2004 sample and WebRDS sampling procedure is presented by Wejnert and Heckathorn (2008). Unless analytical procedures differ, this paper focuses on 2008 data and provides only a summary of Wejnert and Heckathorn's (2008) findings for the 2004 sample, referring the reader to Wejnert and Heckathorn (2008) for detailed analysis. Unlike the 2004 sample, in which one recruitment chain makes up over 70% of the data, the 2008 sample includes two large chains which make up 48.1% and 46.3% of the data respectively. Figures 1 and 2 show RDS recruitment chains for the 2008 and 2004 samples respectively.

[Figure 1: 2008 Recruitment Chains]

[Figure 2: 2004 Recruitment Chains]

Though the samples were collected using similar methods from the same university and respondents were limited to three recruitments in each study, significant

differences in compensation likely resulted in large differences in sampling speed. As described by Wejnert and Heckathorn (2008), the 2004 sample offered up to $55 for participation and completed sampling within 72 hours of the start times (see Wejnert and Heckathorn (2008) for discussion of possible biases resulting from fast recruitment). As a result, a more economic, lottery-based compensation scheme was initially used in the 2008 sample. The lottery scheme proved ineffective and was replaced with a traditional scheme were respondents could earn up to $25 for participation. Unfortunately, by the time these changes were implemented, the semester schedule had reached spring break and consequently, the majority of sampling occurred during the "crunch-time" period between spring break and finals. As a result, the 2008 sample collected 55 respondents in the first month and the remaining 323 respondents in the second month of sampling (213 respondents where collected in the final week). A comparison of respondents collected during month one to those collected in month two showed no substantive differences. Due to problems described by Wejnert and Heckathorn (2008) with fast recruitment, the 2008 sample is likely less problematic than the 2004 sample. In both studies, valid university student ID was required to pick up compensation.

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.

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:

1. Buddylist Degree: the number of students the respondent has saved on his or her instant messenger program buddylist.
2. Recruit Degree: the number of students the respondent believes she could potentially recruit for the study.

3. Email Degree: the number of students the respondent has contacted through email in the past 30 days.
4. Standard Degree: the number of students the respondent knows and has personally interacted with in the past 30 days.
5. Weighted Degree: weighted sum of the respondent's number of close friends (0.47), friends (0.50), and acquaintances (0.03), explained in detail below.

2008 Sample Degree Measures:

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

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 paper 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 1 shows population parameters and raw sample proportions for all created samples used in equilibrium analyses. 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 1: Sample Proportions and Waves Required for Equilibrium]

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 paper, 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:

$$\frac{|0.3 - 0.25|}{0.051} = 0.98. \tag{7}$$

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.

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:

$$DE_{P_X} = \frac{Var(P_X)_{RDS}}{Var(P_X)_{SRS}}, \tag{8}$$

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 to provide overall measures.

Results

Before comparing various methods of calculating estimates, I first test RDS assumptions. I then turn to comparing RDS I and RDS II and discuss the implications of various degree measures and out-of-equilibrium data.

Test of Assumptions:

The methods of testing assumptions presented here parallel those described in detail by Wejnert and Heckathorn (2008). Wejnert and Heckathorn (2008) present evidence suggesting assumptions one through three were met in 2004. Similar analyses suggest these assumptions are also met in 2008 (not shown). The difficulties of accurately measuring personal degree (assumption four) are discussed above and results comparing multiple degree measures for both samples are presented below.

The final RDS assumption is that respondents recruit as though they were selecting randomly from their personal networks (Heckathorn 2002). One method used to assess this condition is to compare recruitment patterns to self reported network

compositions for visible attributes, such as gender and race (Heckathorn et al. 2002; Wang et al. 2005). Using a χ^2 goodness of fit test, Wejnert and Heckathorn (2008) find evidence suggesting this assumption was not met for either gender or race in the 2004 sample. Following the same procedure, I find recruitment does not reflect self-reported personal network composition of gender $\left(\chi_1^2 = 42.5, p < 0.001\right)$ or race $\left(\chi_9^2 = 249.5, p < 0.001\right)$ in 2008 data either. These results contrast with previous studies (Heckathorn et al. 2002, Wang et al. 2005), which find strong association between recruitment patterns and self-reported network composition. It is impossible to know if this result is due to a failure of assumption five or inaccuracy in self-report network compositions. In either case, it shows that the data used in this paper suffer from similar or worse problems as other real RDS data.

Comparison of RDS I and RDS II:

As noted above, 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 3: RDS I and RDS II Standardized Estimates and Confidence Intervals]

Figure 3 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 solid lines, while 95% confidence intervals for RDS II estimates are represented as dashed 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 4: 2008 RDS I and RDS II Interval Inaccuracy]

[Figure 5: 2004 RDS I and RDS II Interval Inaccuracy]

Figures 4 and 5 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 4, 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 5), 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 6: RDS I and RDS II Design Effects]

Figure 6 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. Corresponding results using RDS I estimation support similar conclusions and are available from the author on request.

Comparison of Degree Measures:

Descriptive statistics and correlations for all degree measures are presented in table 2. Consistent with other on social networks, reported degree distributions are highly skewed with small numbers of respondents reporting very high degree 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.

[Table 2: Descriptive Statistics and Correlations for Degree Measures]

Interval and estimate inaccuracy scores (equation 7) for race, gender, and college based on different measures of degree are shown in figures 7 and 8 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[3] (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.

[Figure 7: 2008 Sample Degree Measures and Inaccuracy]

[Figure 8: 2004 Sample Degree Measures and Inaccuracy]

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

[3] 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.

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).

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[4]. In many cases, these same conditions are 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

[4] 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.

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[5]. 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 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

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

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 9: 2008 Inaccuracy for Waves Zero through Nine]

Figure 9 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.

On the other side of the equilibrium debate lies the question of whether early, out-of-equilibrium waves should be removed from analysis. 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 10: 2008 Inaccuracy by Earliest Wave Included]

[Figure 11: 2004 Inaccuracy by Earliest Wave Included]

Figures 10 and 11 show 2008 and 2004 RDS estimates based on the standard degree measure calculated for data starting at wave 0, 4, 7, and 10. 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 1).

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 paper 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, 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.

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

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

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

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

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.

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: 5/5/2004.

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: 3/15/2008.

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

Goel, Sharad and Matthew 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 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., Salaam Semaan, Robert S. Broadhead, and James J. Hughes. 2002. "Extensions of Respondent-Driven Sampling: A New Approach to the Study of Injection Drug Users Aged 18-25." *AIDS and Behavior* 6: 55–67.

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

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

Salganik, Matthew 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 Degani, and Douglas D. Heckathorn. 2007. Respondent-Driven Sampling Analysis Tool (RDSAT) Version 6.0.1. Ithaca, NY.

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.

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.

Cyprian Wejnert is a PhD candidate in the Department of Sociology at Cornell University. His research interests include social networks, social norms, sampling methodology, and HIV prevention in hidden populations. He has recently published articles in *Sociological Methods and Research* and *Marriage and Family Review*.

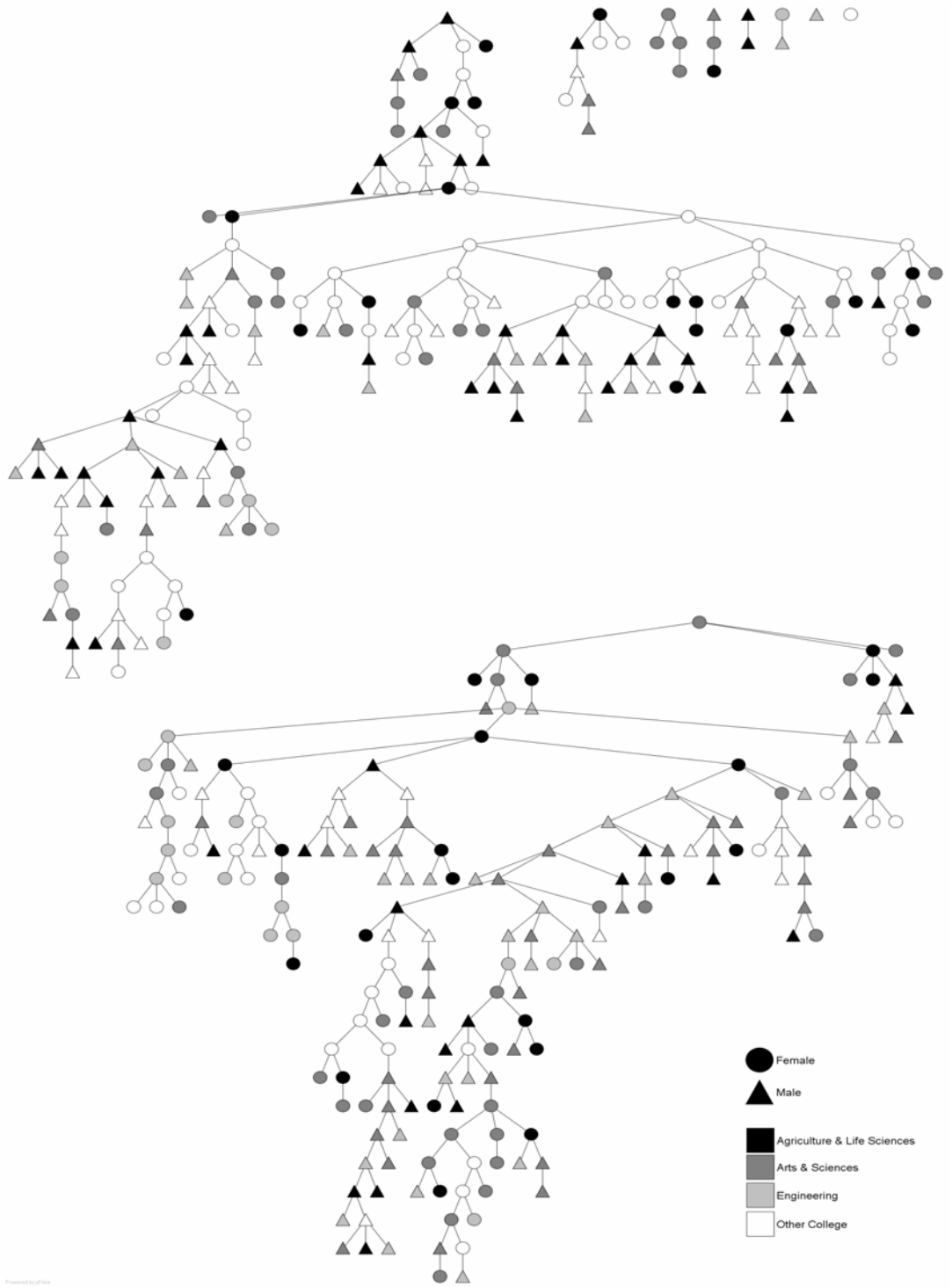

Figure 1: 2008 sample recruitment chains. Shading indicates college within the university, shapes indicate gender.

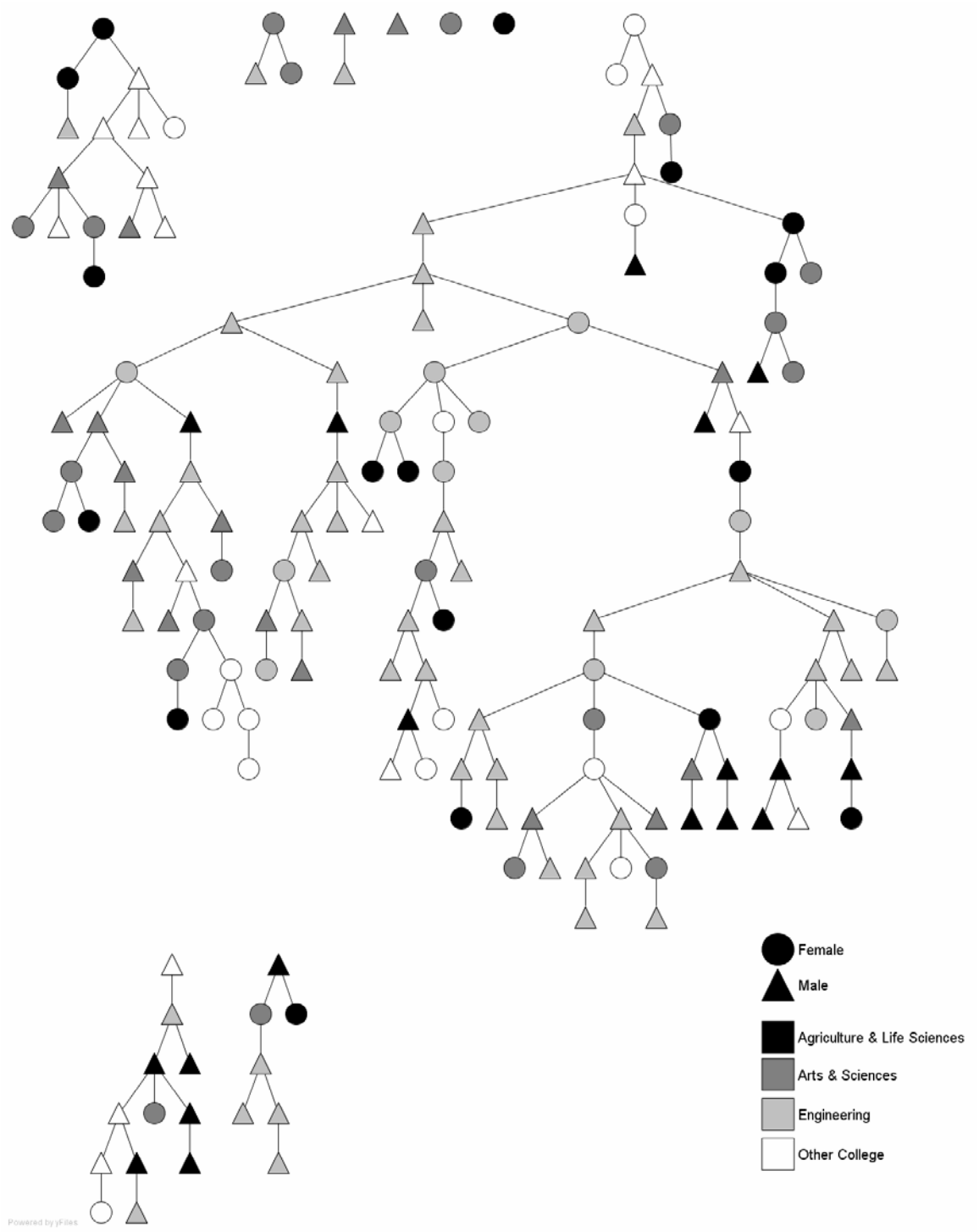

Figure 2: 2004 sample recruitment chains. Shading indicates college within the university, shapes indicate gender.

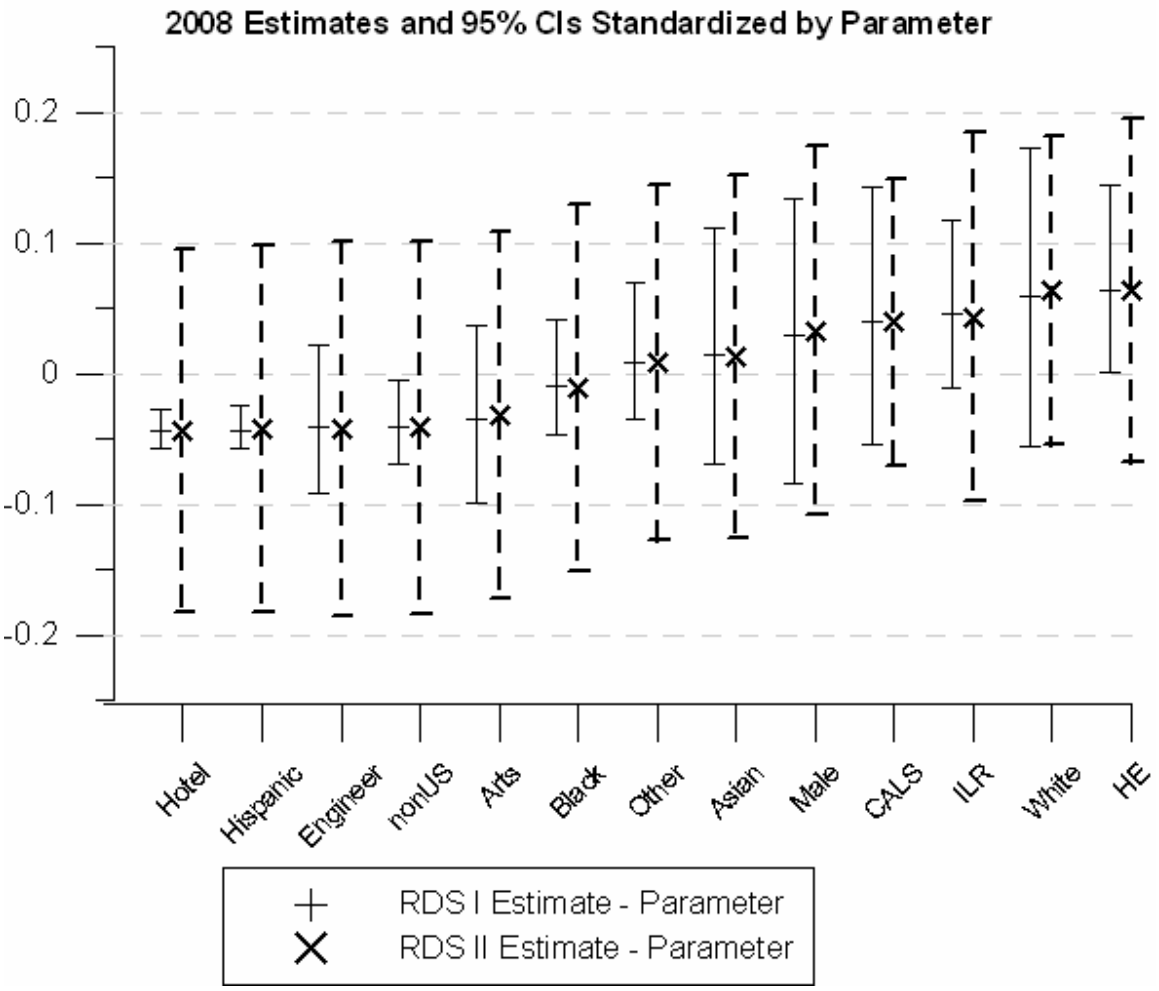

Figure 3: RDS I (+) and RDS II (x) estimates and 95% confidence intervals for 13 dichotomous variables corresponding to race, gender, and college within the university calculated using the standard degree measure from the 2008 sample. The estimates are adjusted such that the line y = 0 represents the 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 solid lines while 95% confidence intervals for RDS II estimates are represented as dashed lines.

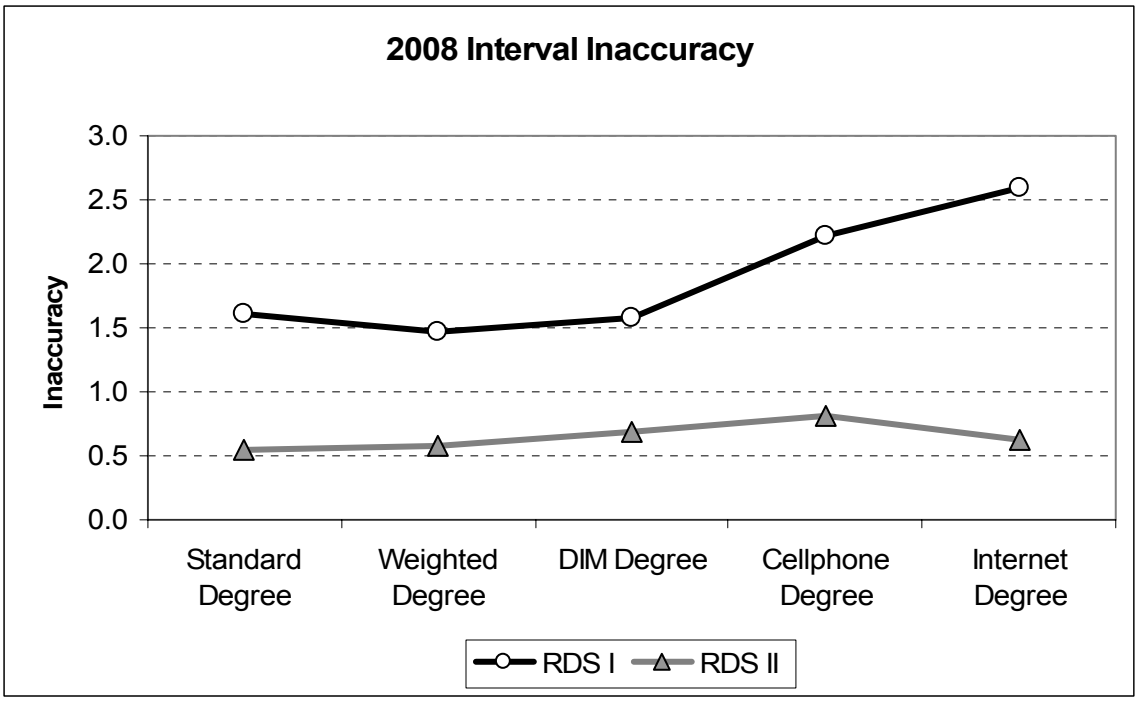

Figure 4: Interval inaccuracy for RDS I (hollow) and RDS II (solid) estimates averaged across race, gender, and college for five measures of degree in the 2008 sample. The line y = 1.96 represents 95% confidence interval bounds. Any interval with inaccuracy greater than 1.96 fails to capture the population 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.

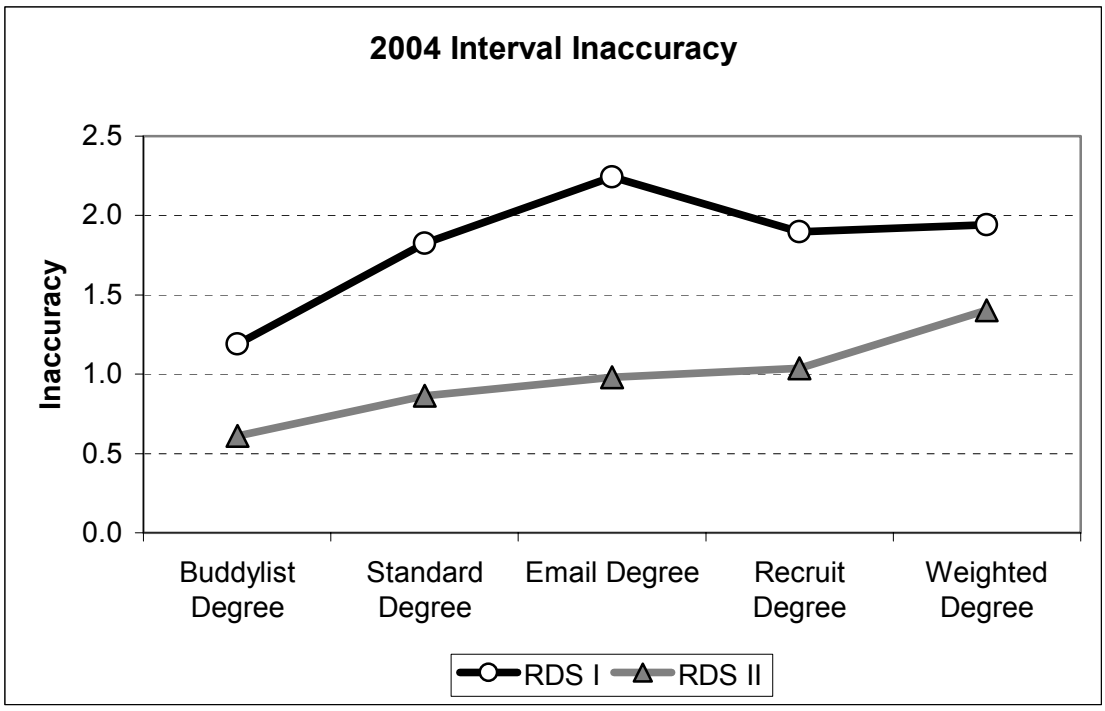

Figure 5: Interval inaccuracy for RDS I (hollow) and RDS II (solid) estimates averaged across race, gender, and college for five measures of degree in the 2004 sample. The line y = 1.96 represents 95% confidence interval bounds. Any interval with inaccuracy greater than 1.96 fails to capture the population 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.

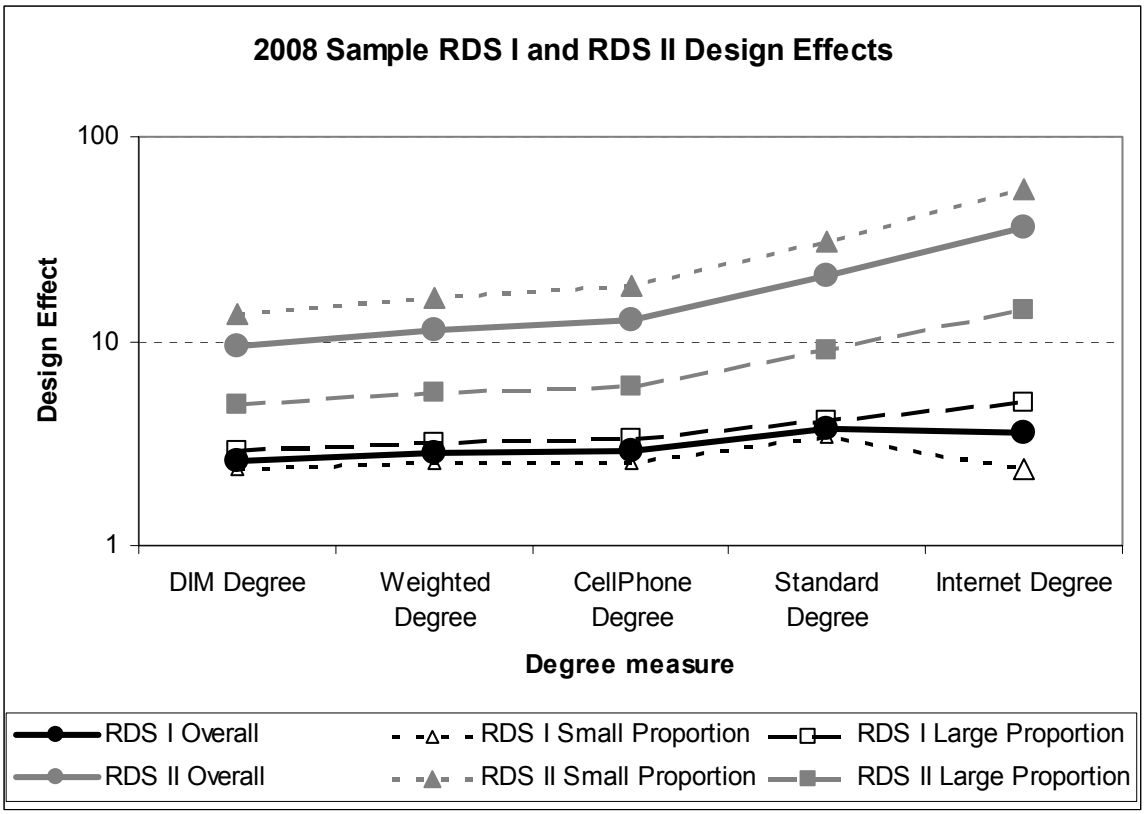

Figure 6: RDS I (black) and RDS II (gray) design effects for the overall sample, small proportion, and large proportion variables. Small proportion variables are dichotomous variables for which the population parameter is less than 0.1. Large proportion variables are those for which the population parameter is greater than 0.1.

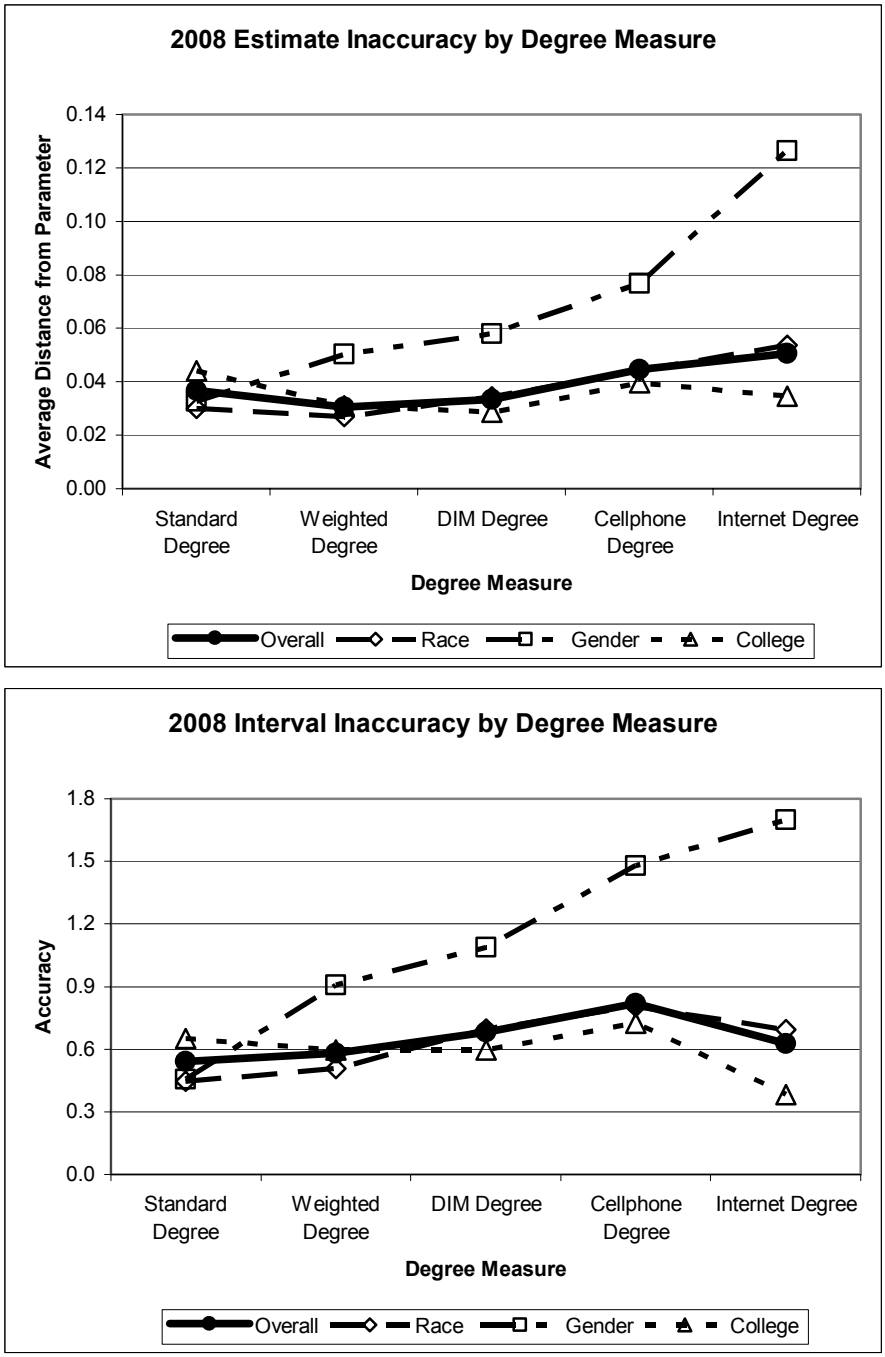

Figure 7: 2008 sample estimate and interval inaccuracy for overall, race, gender, and college based on five measures of degree. Inaccuracy scores for gender are based on a single dichotomous variable and therefore display greater variability than race, college, or

overall scores, which represent the average of scores from multiple dichotomous variables.

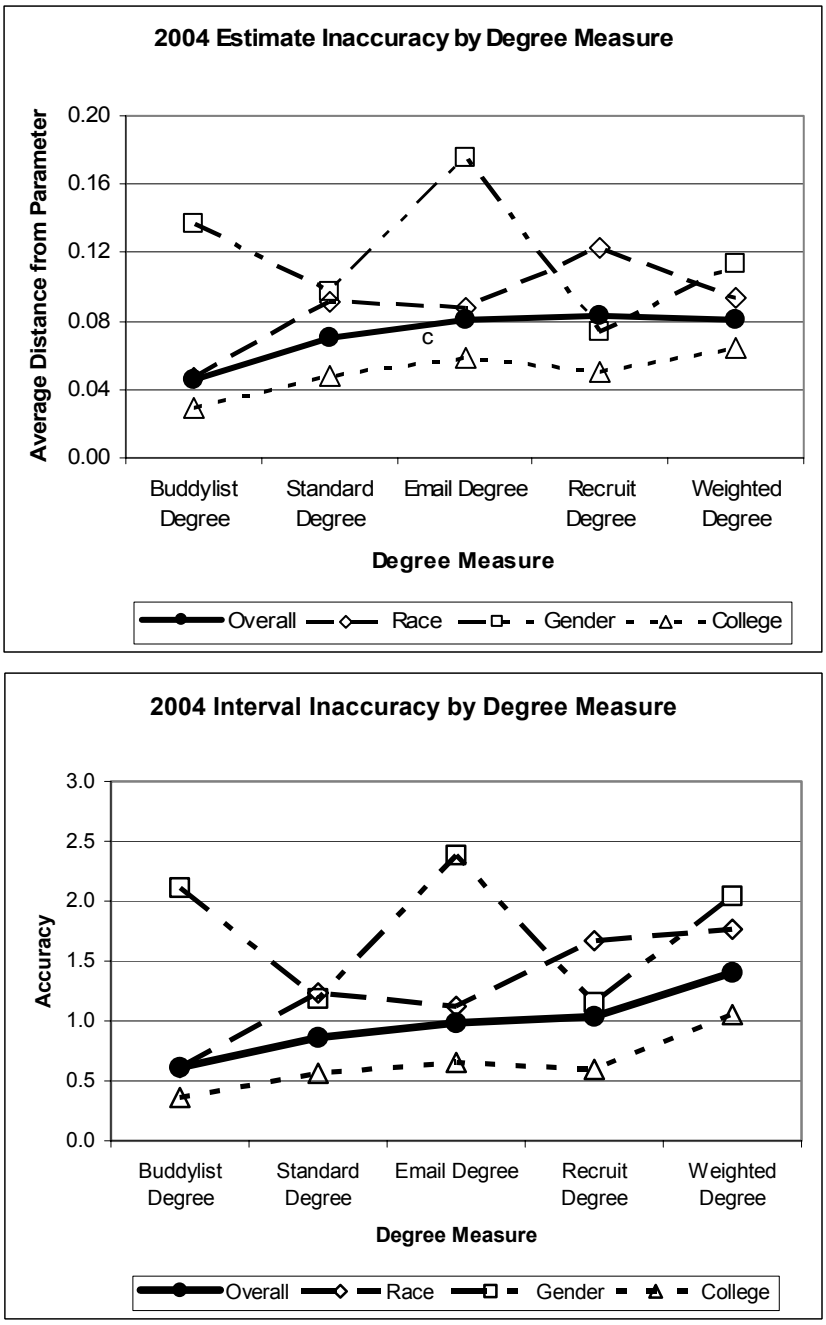

Figure 8: 2004 sample estimate and interval inaccuracy for overall, race, gender, and college based on five measures of degree. Inaccuracy scores for gender are based on a single dichotomous variable and therefore display greater variability than race, college, or

overall scores, which represent the average of scores from multiple dichotomous variables.

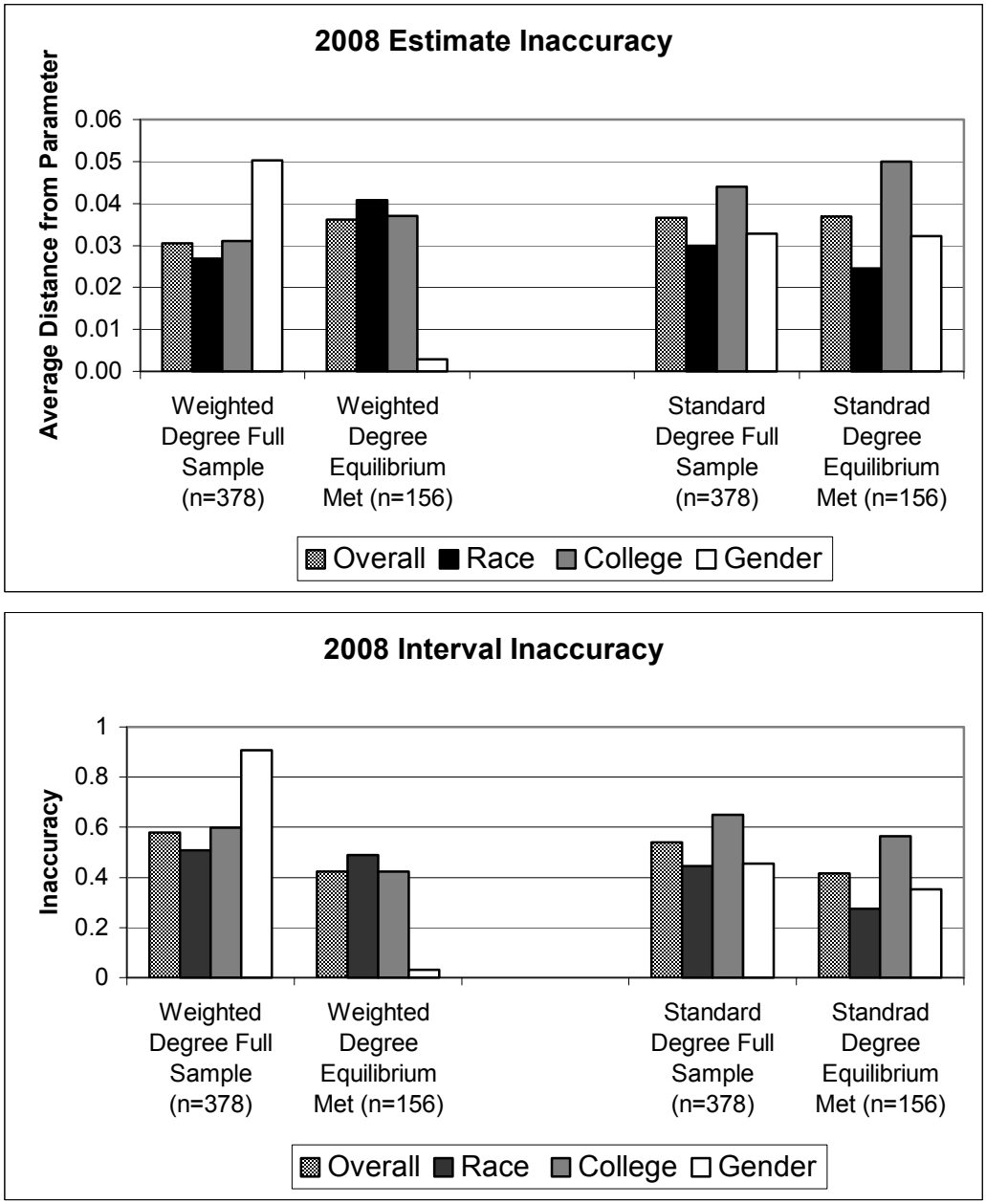

Figure 9: 2008 sample comparison of estimate and interval inaccuracy using only data collected in waves zero through nine to estimate and interval inaccuracy based on the full sample.

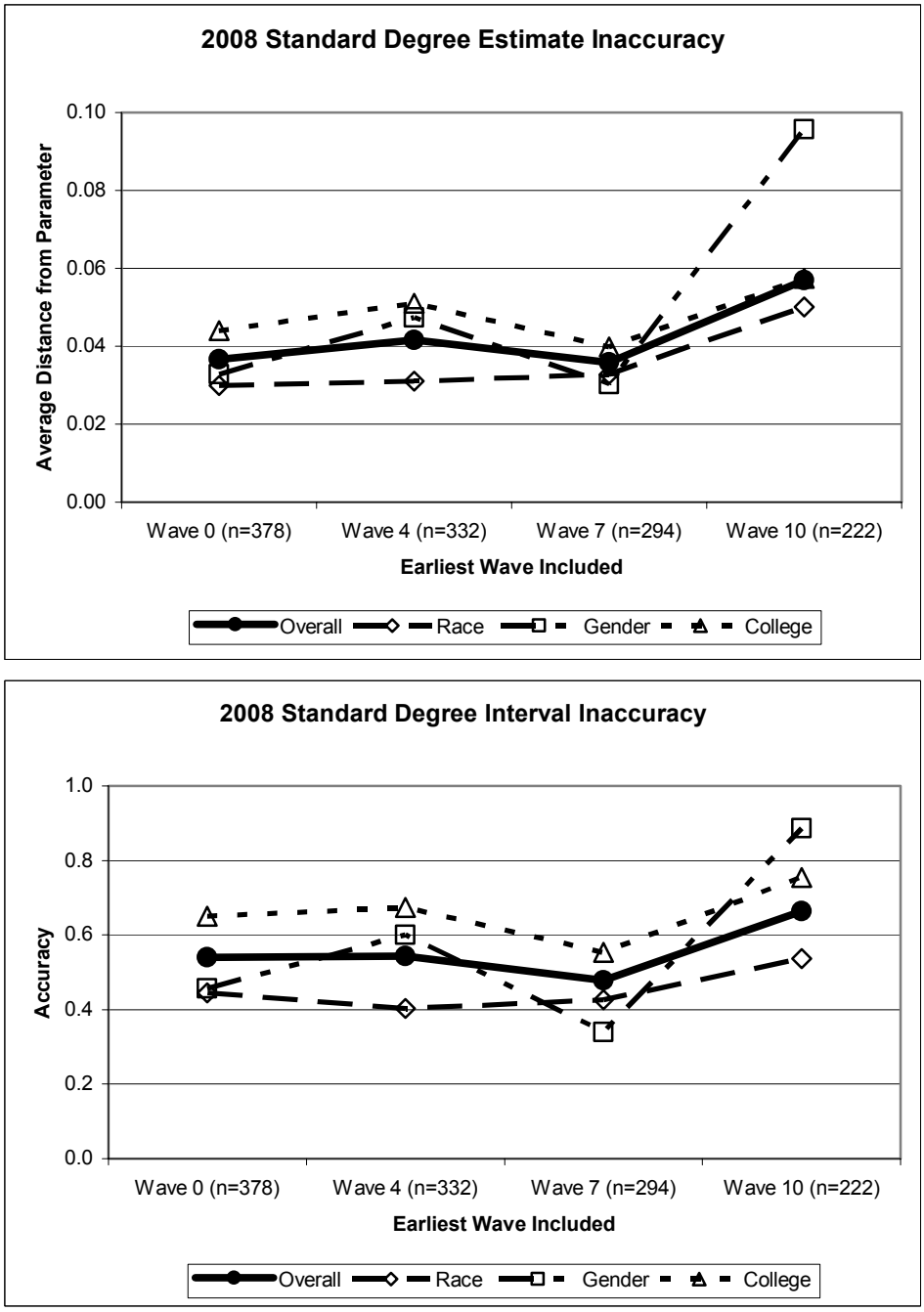

Figure 10: 2008 sample comparison of estimate and interval inaccuracy for analysis based on data starting at different waves of recruitment using the standard degree measure. Sample size used in analysis shown on x-axis.

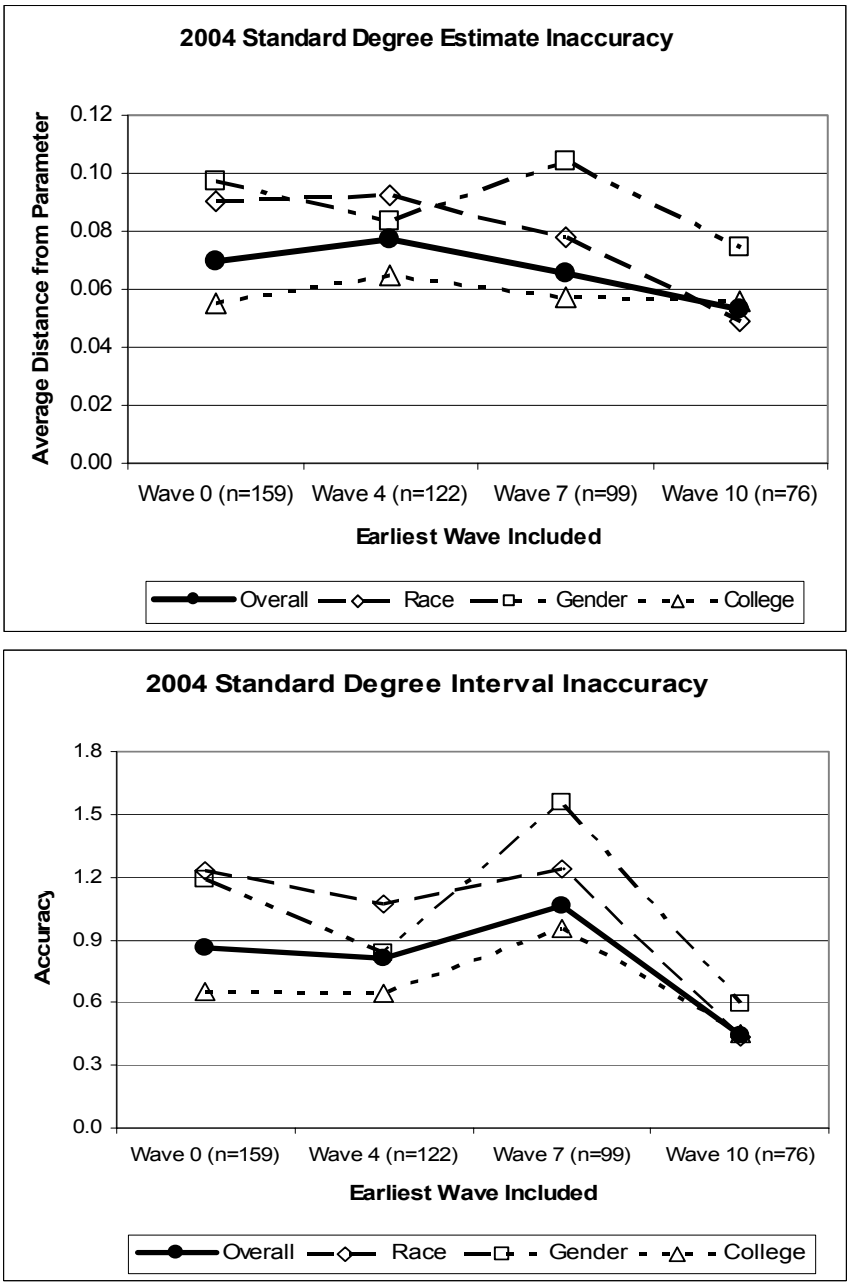

Figure 11: 2004 sample comparison of estimate and interval inaccuracy for analysis based on data starting at different waves of recruitment using the standard degree measure. Sample size used in analysis shown on x-axis.

Table 1: Sample Proportions and Waves Required for Equilibrium for All Data Sets Used in Analysis*

	2008 Sample							2004 Sample						
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)	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	9	0.182	0.154	0.147	0.157	0.154	0.158	6	0.189	0.365	0.422	0.367	0.354	0.303
Black	7	0.059	0.037	0.038	0.036	0.031	0.036	7	0.054	(0.013)	(0.012)	(0.006)	(0.01)	(0.013)
Hispanic	7	0.062	0.021	0.032	0.024	0.014	0.014	7	0.060	0.044	0.06	0.021	0.040	(0.026)
Other	8	0.051	0.035	0.032	0.033	0.031	0.036	6	0.013	0.069	0.072	0.061	0.061	0.066
White	7	0.557	0.715	0.699	0.713	0.736	0.724	6	0.685	0.509	0.434	0.537	0.535	0.592
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	3	0.505	0.597	0.578	0.590	0.596	0.618
College														
CALS	4	0.237	0.247	0.25	0.232	0.222	0.243	6	0.323	0.233	0.241	0.230	0.232	0.224
Arts	4	0.305	0.292	0.308	0.283	0.280	0.279	6	0.223	0.220	0.265	0.213	0.202	0.171
Engineer	5	0.203	0.175	0.167	0.181	0.181	0.180	7	0.197	0.352	0.277	0.393	0.414	0.434
HE	4	0.091	0.164	0.154	0.178	0.195	0.171	8	0.096	0.075	0.06	0.082	0.091	0.092
Hotel	5	0.065	0.032	0.026	0.033	0.038	0.036	9	0.058	0.069	0.12	(0.033)	(0.01)	(0.013)
ILR	6	0.062	0.085	0.083	0.090	0.082	0.086	6	0.060	0.044	0.036	0.041	0.040	0.053

() Indicates sample size is too small for RDS estimates to be calculated

*"Earliest Waves Included" defines the cut point for inclusion, data sampled in waves before the cut point are excluded from analysis.

Table 2: Descriptive Statistics and Correlations for Degree Measures from 2004 and 2008

2004 Degree Measures	N	Minimum	Maximum	Mean	Std. Dev.	Skewness Statistic	Skewness Std. Error
Standard Degree	159	0	450	74.70	68.35	2.56	0.192
Recruit Degree	159	0	10000	80.70	791.94	12.59	0.192
Email Degree	159	0	1000	19.97	80.08	11.77	0.192
Buddylist Degree	159	0	200	66.30	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 Phone	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)