REGRESSION MODELING OF DATA COLLECTED USING RESPONDENT-DRIVEN SAMPLING

A Thesis

Presented to the Faculty of the Graduate School

of Cornell University

In Partial Fulfillment of the Requirements for the Degree of

Master of Arts

by

Michael W. Spiller

May 2009

ABSTRACT

Respondent-Driven Sampling (RDS) is a snowball-type sampling method used to survey hidden populations. To date, analyses of RDS data have primarily consisted of estimating population proportions and their variance because of the special complexities RDS data pose for regression analysis. This paper discusses those complications, focusing on the role of homophily (differential affiliation) in the recruitment process and respondent clustering at multiple potential levels of aggregation. It proposes two techniques for confronting these problems: entering recruiter characteristics directly into recruit-level regression models and estimating fixed- or random-effects models at the levels where significant clustering is observed. An empirical example demonstrates the modeling process, and a six-step procedure for regression modeling of RDS data is presented.

BIOGRAPHICAL SKETCH

Michael Spiller was born in Wichita Falls, Texas. After graduating from Midwestern State University in 2004, he married his best friend Sally. He has spent his life looking for answers; if you have any, please let him know.

This, my first real piece of scholarly work, is dedicated to the women who have always shown faith in me: my grandmothers Dorothy and Marjorie, my mother Diane, and my wife Sally.

ACKNOWLEDGMENTS

This work would not have been possible without the excellent guidance and wisdom of Douglas Heckathorn, Steve Morgan, and Kim Weeden. No less responsible are those who always had my back: my amazing family, Brian, Cyprian, Jason, and Sally above all.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

REGRESSION MODELING OF DATA COLLECTED USING RESPONDENT-DRIVEN SAMPLING

1. Introduction

Respondent-Driven Sampling (RDS) is a snowball-type sampling method designed to study hidden populations - those populations for which a sampling frame cannot be constructed. It belongs to a class of sampling methods called "adaptive link-tracing" designs that significantly improve on traditional snowball sampling by recording the links (or recruitments) among respondents and estimating population parameters using a mathematical model of the sampling process and the social network underlying it (Thompson and Frank 2000). RDS was developed by Douglas Heckathorn in the late 1990s, and has since undergone significant theoretical advances in parameter and variance estimation (Salganik and Heckathorn 2004; Heckathorn 2007; Volz and Heckathorn 2008; Wejnert and Heckthorn 2008). Additionally, RDS has been used extensively by the U.S. Center for Disease Control and Prevention (CDC) to study populations at risk for HIV, AIDS, and other sexually transmitted infections.

To date, most analyses of RDS data have consisted of estimating population proportions and their variance because of the special complexities RDS data pose for regression analysis. This paper discusses those complications, focusing on the role of homophily (differential affiliation) in the recruitment process and respondent clustering at multiple potential levels of aggregation. It proposes two techniques for confronting these problems: entering recruiter characteristics directly into recruit-level regression models and estimating fixed- or random-effects models at the levels where significant clustering is observed. An empirical example demonstrates the modeling process, and a six-step procedure for regression modeling of RDS data is presented.

This thesis assumes that readers are familiar with sampling theory, have a basic understanding of Respondent-Driven Sampling, and have experience building and diagnosing regression models.[1]

2. Respondent-Driven Sampling Data

Before delving into the details of modeling RDS data, a brief description of RDS sampling in practice and RDS' basic assumptions is in order.

Prior to sampling, a researcher must erect the physical infrastructure through which sampling will take place. This infrastructure consists of the sites where interviews will take place, which are typically rented storefronts or space in community institutions such as churches, community colleges, or senior centers.[2] A researcher must first decide which geographic areas will be included in sampling, ensuring that interview sites are located such that all respondents will have reasonable access to a site (which can require care when a city is highly geographically segregated, such as New York City's five boroughs). Since one of the central RDS assumptions is that recruits are selected at random from his/her recruiter's pool of potential recruits (Salganik and Heckathorn 2004; Heckathorn 2007), it is imperative that sites be in neutral locations (to prevent, for example, racially homogenous neighborhoods from deterring the subset of respondents who would feel uncomfortable).

Once the sampling infrastructure is in place, the researcher locates population members to serve as "seeds" for the recruitment trees. Typically, the researcher will choose multiple heterogeneous seeds for each site, interview them, remunerate them for their time, and give them a fixed number of uniquely numbered dollar-bill sized

[1] A detailed description of RDS population estimation may be found in the Appendix.
[2] However, some sampling designs have interviewers that travel to respondents, such as Jeffri and Heckathorn's study of New York City aging artists (Spiller, Heckathorn, and Jeffri 2007).

coupons with which to recruit population members they know into the sample.[3] When subsequent recruits come to the site, they are interviewed, issued coupons, remunerated, and the process is repeated. Any respondent who has recruited successfully may come to a site and receive modest remuneration for each recruitment he/she made; this "dual-incentive" structure motivates respondents to recruit and helps to ensure that respondents only recruit population members known to them (a recruiter is not paid if their recruit does not qualify for the study or does not know their recruiter). In general, the researcher has either a target sample size or a fixed pool of resources with which to pay respondents, so the approximate final sample size is known. As the sample reaches the target size, respondents are issued fewer and, eventually, no coupons with which to recruit. When sampling is complete, the researcher has data on each site's geographical location, which site the respondent was interviewed at, who recruited the respondent, who the respondent recruited, and all other information collected in the survey. Figure 1 displays the sample recruitment trees (respondents who share a seed are members of the same recruitment tree) from Jesus Ramirez-Valles' Chicago study of Latino men who have sex with men.

[3] Some studies use a more complex approach where sample members with certain characteristics are given more (or fewer) recruitment coupons.

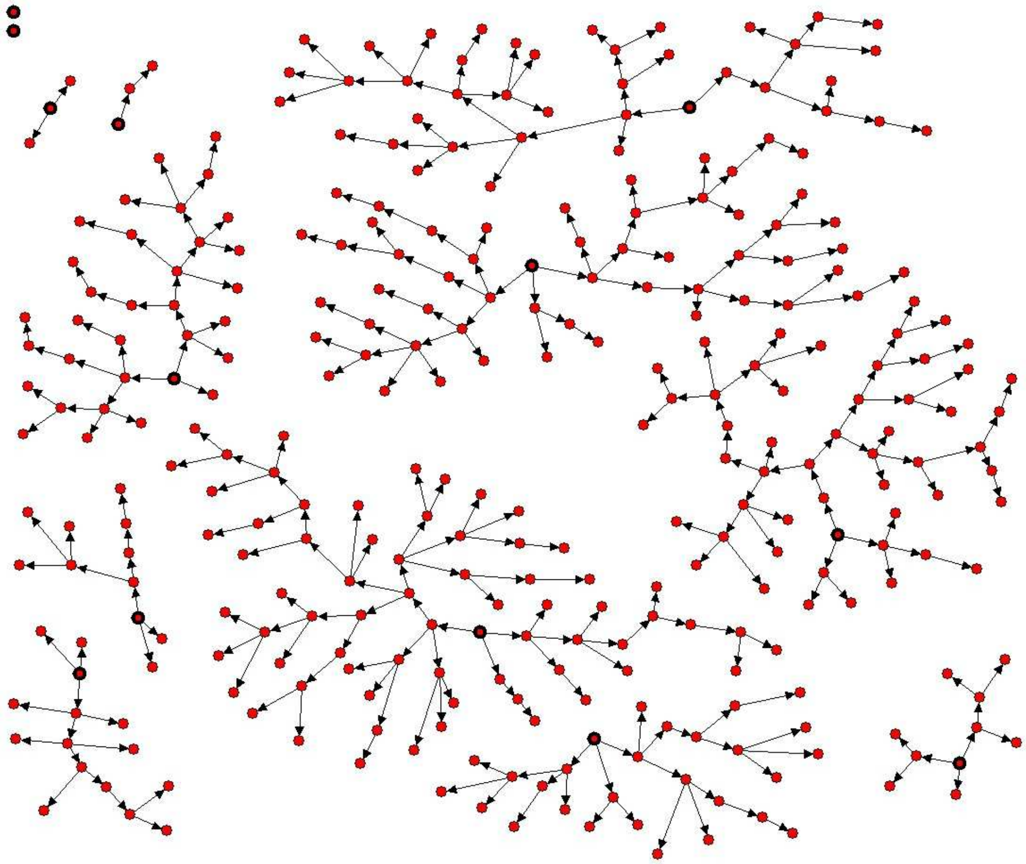

Figure 1 RDS recruitment trees from a 320-person Chicago study of Latino men who have sex with men (Ramirez-Valles et al. 2005). Seeds are denoted by large-rimmed nodes.

A key assumption required for RDS is that the social network being sampled forms one giant component (Salganik and Heckathorn 2004; Heckathorn 2007). If the network is comprised of many small, disconnected clusters, it is likely that some clusters would not be linked to the overall network at all (and would therefore be highly unlikely to become members of an RDS sample). Fortunately, work in network graph theory indicates that most nodes are members of one large component even in relatively sparse graphs (Newman 2003). Additionally, work on the "small world problem" shows that in most real-world social networks any two nodes are linked through a relatively small number of steps (Watts and Strogatz 1998; Watts 1999; Dodds et al., 2003). A central task for modelers of RDS data is assessing whether this assumption was met during the sampling process. If it was, there should be little or no affiliation (homophily) based on geographic location or interview site because the social network reaches across geographic boundaries. Additionally, there should be little or no relationship between a respondent's characteristics and which recruitment tree a respondent was in because all of the recruitment trees were sampling from the same well-mixed network. If significant differences are observed between respondents in different geographic areas, interview sites, or recruitment trees, the researcher should consider carefully whether he/she was actually sampling multiple unconnected (or barely connected) social networks.[4]

3. Theoretical Concerns

The primary concern for modelers of RDS data is adjusting for the lack of independence among respondents. Standard regression models assume that individual-level errors are not correlated with the independent variables in the model (implying that observations are independently sampled from the population). Because some respondents in the RDS sample recruit more than one other respondent, this assumption does not hold for RDS data. In this case, we are treating respondents who share a recruiter as being a cluster. Dependence results in higher-than-expected between-cluster variance (also known as "over-dispersion") and lower-than-expected

[4] If this is the case, the researcher may obtain population inferences by employing RDS estimation on each smaller network and using weights to recombine the estimates (although population estimates will then rely significantly on the chosen weights).

within-cluster variance, which virtually guarantees that standard errors (and therefore significance tests) will be incorrect. Almost all of the time, they will be too small because they will overestimate the amount of information within clusters relative to true random sampling. Note that marginal coefficient estimates will still be consistent under the assumption that the unobserved between-cluster heterogeneity is uncorrelated with the independent variables in the model (Greene 2003).[5] If this assumption does not hold, marginal coefficients will be inconsistent and fixed-effects estimators are required to obtain consistent estimates.

The recruitment process suggests that the dependence among observations is strongest at the recruiter-recruit dyad level, but most respondents are both recruits and recruiters so dyads do not form mutually exclusive groups (which is required for most regression adjustment strategies). In most RDS studies respondents can make more than one recruitment, so we should expect to observe strong clustering by shared recruiter if recruiters' social networks are more homogeneous than the entire social network being sampled. A central concern of the adjustment strategy outlined below is addressing clustering by shared recruiter.

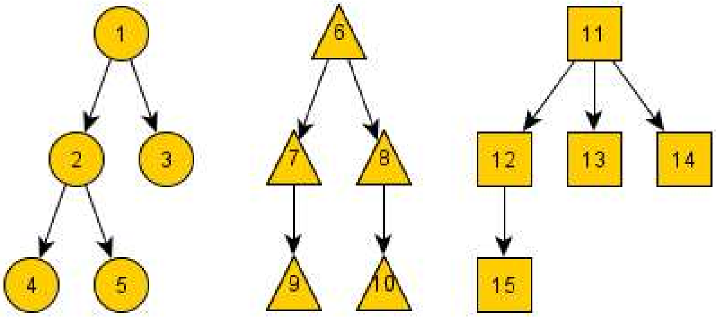

Figure 2 Recruitment trees in a hypothetical RDS sample.

[5] RDS populations estimates are biased on the order of [1/sample size] (Salganik and Heckathorn 2004).

Figure 2 displays three recruitment trees from a hypothetical RDS sample. Nodes {1}, {6}, and {11} are the seeds selected by the researcher. As noted above, we would expect the strongest dependence among respondents at the recruitment dyad level, but they do not form mutually exclusive groups (and therefore are not amenable to adjustment through regression). Some of the recruitment dyads in Figure 2 are nodes {1, 2}, {1, 3}, {7, 9}, {11, 12}, and {11, 13}. The clustering adjustment strategy proposed by this thesis focuses on adjusting at the shared recruiter level; examples of nodes in Figure 2 that share a recruiter are {2, 3}, {4, 5}, and {12, 13, 14}. Note that the number of members in a shared recruiter "group" will always be between one and the number of coupons respondents are given to recruit with.

Significantly complicating the modeling of dependence among respondents is the unobservable grouping in the social network from which RDS is sampling. Since the population network consists of real social groups (i.e., churches, neighborhoods, clubs, etc.), sampling would be most efficient and transparent if it could sample these groups and then respondents within them (similar to sampling U.S. census tracts then respondents within them, although social groups are not mutually exclusive). This sampling approach is impossible because a sampling frame cannot be constructed, but the modeler would ideally adjust for clustering at this level because it is likely that respondents would be more similar within these groups than even a recruiter-recruit dyad across groups.[6] Unfortunately, there is no theoretical reason to expect information collected in an RDS sample to consistently map onto the grouping structure of the population social network. Note that this unobservable grouping is what we are assuming is uncorrelated with the model independent variables, as mentioned above.

[6] However, see Handcock, Raftery, and Tantrum 2007 and Newman 2006 for work on modeling the latent clusters (or community structure) underlying an observed social network.

The second major concern for modelers of RDS data is that the model is consonant with the RDS population estimation approach (i.e., addresses its assumptions and faithfully represents its statistical power). Because the modeler is primarily concerned with the relationships among respondents' characteristics, it is not necessary for the model to accurately reproduce the population estimates, but it is desirable for the model to take into account the information used for population estimation. The information necessary to replicate population estimates from sample data (but not to replicate variance estimates) is fully captured by respondents' probability of inclusion in the sample. In RDS, the inclusion probability has two components: the number of potential recruit(er)s a respondent knows in the target population and the recruitment characteristics of a respondent's variable-specific group (i.e., the characteristics of African-Americans in the sample for the variable race). The first component is stable across variables, but the second is not. Ideally, we could estimate the second component for all variables simultaneously, but the matrix would be too sparse for reliable estimation without severe loss of information from collapsing variables.

Fortunately, there is a deep literature from which to draw on modeling clustered (or correlated) data and on survey estimation that adjusts for non-unitary inclusion probabilities. Unfortunately, this literature is almost universally oriented toward classical multi-stage survey sampling (and non-response, post-stratification raking, and the myriad factors survey analysts must adjust for), so drawing appropriate lessons for modeling RDS data is more complex than one would hope.

4. Preliminary Steps: Examining the Data

The modeler's first step in identifying underlying network structure is to assess whether or not the sample has mixed across geographic area and interview sites. If the underlying network is geographically integrated, we would expect to see geographic area (and site of interview) randomly distributed within and across recruitment trees. If the underlying network is completely segregated, we would expect no mixing across geographic areas such that all members of a recruitment tree would be interviewed in the same area (or at the same site).[7] There are a few approaches available to assess geographic and site mixing; the most preferable is examination of homophily using the standard RDS estimation approach (homophily is the tendency to associate with those similar to oneself).[8] If geographic area homophily is high, the modeler will need to consider including geography as a factor in his/her regression model. If there is more than one site in any geographic area, the modeler should examine homophily by site to make sure that the sample is not segregated by site within area (i.e., to make sure the network is truly structured by geographic area and not at some finer level). If there is not mixing across sites, the sample should be divided into multiple samples for population estimation (as in Heckathorn 1997), which will avoid the problems posed by the giant-component assumption. Additionally, the modeler should consider estimating a fixed-effects model on geographic area, interview site, or recruitment tree as a regression strategy if he/she believes that between-cluster variation at any of these levels is correlated with a model's independent variables.

[7] If there is exactly one site per geographic area, geographic area and site are equivalent. Note that homophily cannot be calculated for recruitment trees because, by design, there can be no cross-tree recruitment (so clustering at the recruitment tree level should be evaluated using alternative strategies).

[8] The RDS Analysis Tool is the easiest way to estimate population parameters and homophily; it is available for download at www.respondentdrivensampling.org. A useful alternative to examining homophily is one-way ANOVA; however, there is usually significant imbalance in number of respondents per site, so with small sample sizes we may not have enough information to conclude (or reject) that site-level distributions were drawn from the same parent distribution.

Before the modeler begins building the model itself, it is important to examine the dimensions along which respondents sort themselves (as indicated by homophily). In well-mixed and homogeneous populations, it is possible to observe very little or no homophily, making the modeler's job simpler. On the other hand, most populations are structured by age, income, and/or race (along with other population-specific characteristics), so modelers should anticipate adjusting for at least some observable population sorting. To determine the extent of sorting, the modeler should analyze homophily for the characteristics mentioned above, population-specific characteristics known to the modeler, and any other variables which might be included in the regression model "crossed" with the outcome variable. Because homophily will only be relevant to consistent coefficient estimation in the model if it is related to the outcome variable, overall homophily for each variable is less informative than homophily by the outcome variable. For example, if the model outcome were whether or not a respondent was HIV positive, the modeler would examine race homophily by HIV status rather than overall race homophily. A complementary strategy would be to examine whether the sample is "self-weighting," which occurs when there is not significant homophily, differences in group-level degree distributions, or differential recruitment. If the sample is self-weighting, RDS weights will equal one.

Finally, the modeler will need to assess the degree to which the sample reveals the social groups that form the network RDS is sampling. As discussed above, it would be virtually impossible to directly observe membership in these groups because a researcher could not construct a list of groups a priori. Because of the stochastic nature of the sampling process (primarily due to the random recruitment assumption), the extent to which the sample maps onto the underlying social groups is also a stochastic process (i.e., if one sampled the same population repeatedly, the degree to which the sample reveals social groups would be a stochastic outcome based on an unknown population sampling distribution). In light of this, the modeler should not attempt to detect the actual social groups underlying the sample, but should instead examine the aggregate degree of respondent similarity (intra-class correlation) at different levels of observable grouping. As noted above, the lower level of grouping is among recruits who share a recruiter; the higher level of grouping is by recruitment tree. The modeler should examine clustering at both of these levels and anticipate adjusting the regression model if significant clustering is detected.[9] It should be emphasized that for consistent coefficient estimation, we are only concerned with clustering that is related to the model outcome variable; if the clustering is independent of the outcome variables, adjustments will not impact the coefficient estimates and are therefore unnecessary (although they will affect variance estimation).

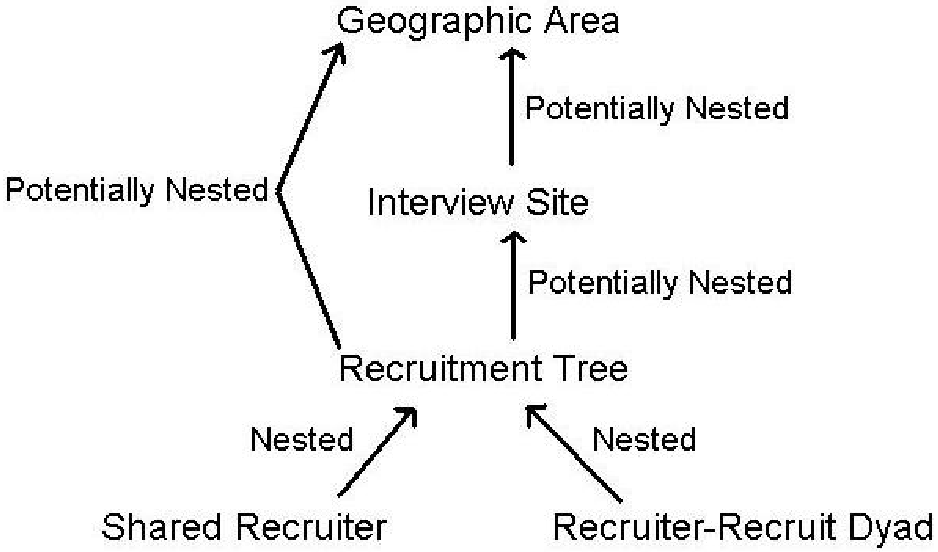

Figure 3 Potential grouping/clustering levels in RDS and their nesting structure.

[9] Because each recruiter is also a recruit, recruitment dyads do not form mutually exclusive groups and are therefore inappropriate for cluster adjustment strategies. Recruiter impact on recruit outcome variables will be addressed below.

Figure 3 displays the levels at which the modeler could observe grouping in RDS data and their nesting structure. As discussed above, the potential nesting of recruitment trees, interview sites, and geographic area are determined by both the sampling infrastructure and the actual sampling process (e.g., a researcher might anticipate recruitment across interview sites but not observe it in the final sample). The modeler must examine grouping at all these levels to determine the appropriate adjustments to the regression model. Note that the "potential nesting" links in Figure 3 will only obtain if the sample did not mix well (i.e., usually they will not be nested); if this is the case, the modeler no longer has the option of examining homophily by area/site because there was no cross-recruitment.

5. Regression Modeling of RDS Data

Once the modeler has identified the variables along which respondents sort themselves and the levels at which there is significant clustering, he/she may start building the model itself. This process is akin to a balancing act, as either significant clustering or homophily may lose its strength when conditioned on the recruit-level covariates in the model. Because clustering will largely determine the variance estimation approach, it is possible that the modeler will proceed relatively far into the modeling process before uncovering information that suggests another estimation approach would be better-suited. As has been oft noted, the process of modeling stands on the line between statistics and art.

The modeler should first examine the association between his/her set of potential predictors and the outcome variable, retaining those that exhibit relatively significant association.[10] After running the basic model, the modeler should employ

[10] This paper is concerned with building a best-fitting descriptive model; if one were particularly interested in the effect of certain variables they should be included even if displaying insignificance at

the diagnostic tools appropriate for the model type (e.g., linear, binary, count, etc.), exploring transformations of the predictor variables and interactions between variables as necessary. It is at this point that sorting variables should be taken into account. If there are no variables related to the dependent variable exhibiting high homophily, the modeler has strong evidence that the population from which the sample was drawn did not strongly sort itself and he/she can be confident that sorting along independent variables will not generate inconsistent coefficient estimates (although he/she should still examine the effect of including recruiter's value of the dependent variable; its effect is unpredictable from homophily analysis).

Because homophily operates most transparently at the recruiter-recruit level, I propose to adjust for it by entering a respondent's recruiter's values for the homophilous variables as predictors in the regression model.[11] It is reasonable to expect respondent-level characteristics to exhibit stronger effects on the outcome variable than recruiter-level characteristics, so the modeler should decide whether to include recruiter-level predictors by evaluating their significance conditional on the respondent-level predictors (by adding them into the base model instead of examining univariate associations with the outcome variable). Note that one may include respondent and recruiter values for the same variable. If the modeler has transformed the recruit-level value, he/she should strongly consider transforming the recruiter-level value as well. Additionally, if both are significant the modeler should consider adding an interaction between them. For example, if both a respondent's race and recruiter's race significantly predict the outcome variable, it is possible that the unique

this stage. If one is interested in testing causal claims, there is much more work to be done (see Morgan and Winship 2007 for an enlightening discussion).

[11] McPherson and Smith-Lovin 1987 describes two types of homophily: induced, which is a function of the people available to a respondent to associate with; and choice, which is a function of whom a respondent chooses to associate with given the potential candidates. Since induced homophily cannot be observed by our sample, the only option is to adjust for structuring variables as if they are completely driven by choice (i.e., at the recruiter-recruit level).

combinations of recruiter-respondent race could also play an important role and should therefore be investigated through an interaction term. Also note that seeds do not have recruiters, so they will be excluded from the analysis. Because recruitments of seeds are also excluded from RDS population estimation, excluding them from the model will more accurately reflect the general RDS estimation approach (seeds are also excluded from degree estimation because they were not recruited into the sample by another sample member).

As the modeler begins examining recruit- and recruiter-level variables and interactions, it is possible for the number of predictors to rapidly increase. Just as two (or more) highly correlated respondent-level variables can interfere with accurate modeling, so can highly correlated recruiter-level and/or highly correlated inter-level variables. For example, if there were 100% racial homophily (such that every recruiter only recruited someone of the same race) recruiter race and respondent race would be perfectly collinear. Most RDS samples exhibit significant homophily along a relatively small number of variables, so a modeler should be suspicious of collinearity if there are many significant recruiter-level/inter-level predictor variables or other unstable significance patterns. Note that if homophily is too severe, the sample must be split as described above.

Once the modeler has built the respondent-level base model and added appropriate recruiter-level and inter-level variables, he/she must decide how to adjust for clustering at the shared recruiter and recruitment tree levels described in Figure 3. If there is not significant clustering at these levels, the modeler can be confident that there is not unobserved heterogeneity that render coefficient estimates inconsistent. However, if the modeler observes significant clustering at one or both levels, he/she will need to select an appropriate adjustment strategy. If the sample did not mix geographically such that interview site or geographic nesting is observed, adjustments for area and site would switch from the homophily-based adjustments concerned with coefficient consistency described above to the clustering-based adjustments for shared recruiter and recruitment tree used to inflate the standard errors.

Generally, there are three parametric options for modeling clustered data (with mutually exclusive clusters): “within-group” estimators (or fixed-effect), “between-group” estimators (or population-averaged), and “random effects” estimators (see Wooldridge 2002 or other graduate-level econometrics textbooks for a detailed description of these estimation approaches).[12] Within-group estimators are equivalent to estimating a standard regression model for each of the clusters independently and computing a weighted average of the coefficients (note that variables that do not vary within clusters cannot be evaluated). Since we will be adjusting for shared recruiter and many recruiters only make a single recruit, the within-group estimator would lose extreme amounts of statistical power because there can be no within-group variance in a one-respondent group.[13] Therefore, if the modeler believes that there is unobserved heterogeneity between shared-recruiter clusters that is correlated with the model independent variables, fixed-effect estimators at this level will lose so much information as to be useless and consistent coefficient estimation must rely on non-parametric approaches.

Between-group estimators are equivalent to collapsing each cluster’s information into means and variances and estimating a standard regression model on the cluster means. They would be appropriate if unobserved within-cluster

[12] Potentially useful non-parametric estimators include Generalized Estimating Equations (also known as Generalized Method of Moments) and Feasible Generalized Least Squares (which is an alternative estimation strategy to maximum likelihood for random effects models). However, these approaches require correctly specifying the error variance-covariance matrix, which is beyond the scope of this paper and a topic of future research.

[13] One need not be concerned with random-effects models on small clusters; see Gelman and Hill 2007:275-276 for discussion.

heterogeneity was believed to be correlated with model independent variables, but the random-recruitment assumption asserts that this will not be true.

Random-effects estimators are a weighted combination of within- and between-group estimators; they assume that neither between- nor within-cluster variance is correlated with the model independent variables. It models the cluster-level over-dispersion by assuming that cluster-specific characteristics are drawn from a parametric distribution, usually the Gaussian/normal distribution (Rabe-Hesketh and Skrondal 2005; Gelman and Hill 2007). Generally, estimates of regression coefficients will not be sensitive to misspecification of the mixing distribution (Neuhaus et al. 1994, cited in Pendergast et al. 1991). There is not a closed-form solution for maximizing the likelihood so numerical approximation algorithms such as Newton-Raphson and Gauss-Hermite adaptive quadrature are used (Rabe-Hesketh and Skrondal 2005). Since both within- and between-group estimators entail loss of information, random effects estimators are generally preferred if the assumption of non-correlated heterogeneity obtains. However, models with many random effects can become sensitive to the numerical approximation algorithm and take significant amounts of computing time, so care is needed.

The computational complexity of a random-effects adjustment strategy will depend primarily on whether the levels of significant clustering are nested. If recruitment trees are nested within interview site (i.e., all members of each recruitment tree interviewed at the same site), the modeler will only need to include one adjustment each for interview site and recruitment tree. However, if some respondents were interviewed at different sites than their recruiter (i.e., site and tree are “crossed”), the modeler will need to include an adjustment for each site plus one for recruitment tree (Rabe-Hesketh and Skrondal 2005). The nesting issue is similar at the interview site within geographic area level: if all respondents within a geographic area were interviewed at the same site, the modeler will have significantly fewer adjustments to make than if some respondents within an area interviewed at a different site than their recruiter. As the number of such "crossed" adjustments can increase rapidly, the modeler must be careful not to push the computational burden to the point of estimate instability.[14]

Although the modeler will want to adjust for clustering at the shared recruiter level in order to model the most variance, he/she should first consider whether a fixed-effects approach at a higher clustering level is more appropriate. If the modeler observed that there was not mixing across interview sites or geographic areas, he/she should consider fixed-effect approaches because there is almost certainly unobserved heterogeneity that is correlated with model independent variables.

If clustering is observed at the shared recruiter or recruitment tree levels (and is not believed to be correlated with model independent variables), the modeler should begin by adding a random intercept to the model for the lowest level of significant clustering. If the variance of the random intercept is significant and the model fit is improved, the modeler should keep it in the model (however, there is still debate on the value of random-effect significance tests; see Pinheiro and Bates 2000 for discussion). If the modeler observes significant clustering at the lowest level and the lowest level is nested in the higher level, he/she may proceed by adding a random intercept at the higher level, examining the significance of the random effect's variance and evaluating the change in model fit using log-likelihood test. If the higher random effect is significant and fit improved, the modeler should keep it in the model. As long as the lower clusters are nested within higher clusters, the modeler may proceed by adding higher random effects and evaluating the model as above. He/she

[14] If the estimates are unstable, they will change as the number of integration/quadrature points is altered. Basically, the integration and/or likelihood spaces have become so complex that we cannot be sure the numerical approximation algorithms are solving them correctly.

should stop when additional random effects do not increase model fit or the estimator becomes unstable due to computational burden.

If, however, the modeler observes clustering at multiple levels and lower levels are not nested in the higher ones, he/she will face some difficult decisions. Because the lowest level of observed clustering will capture the most variance, it should be retained in the model. However, choosing whether or not to include higher clustering levels is complicated by the computational instability that results from models including too many random effects. After adding the lowest-level random effect, the modeler should estimate the model with the lowest level random effect "crossed" with the next clustering level up. If there are too many random effects for stable estimation, the modeler should switch to estimating the lowest level random effect crossed with the second significant clustering level above it. If this is also unstable, he/she should continue upward in levels to cross with the lowest level. Once a stable model is found, the modeler should evaluate the change in model fit from the lowest level only random effects model to the crossed random effects model, retaining the crossed random effects if there is significant improvement in fit. Note that one must choose a covariance pattern among random effects at different levels; theory may guide a modeler to expect a specific covariance pattern or he/she may evaluate different covariance patterns and select the best-fitting.

Once the modeler has selected and evaluated respondent-level covariates, recruiter-level covariates, and the appropriate clustering adjustment strategy, he/she has one final concern: how to (or whether to) weight the data/model. As noted above, the RDS weights represent a respondent's probability of inclusion into the sample, so they are appropriate for Horvitz-Thompson inverse probability weighting (Salganik and Heckathorn 2004; see Thompson 2002 for a discussion of Horvitz-Thompson weighting in general). Because these weights do not include adjustments for non-response or post-stratification, they are not subject to many of the difficulties involved in complex survey weighting. However, as in complex survey weighting, the weights alone will not accurately reproduce the variance in responses due to the complicated sampling design (so simply estimating a weighted OLS regression would not generate accurate standard errors). Error heteroskedasticity can be addressed with use of the Hubert-White “sandwich” estimator (Greene 2003). There is ongoing debate about regression and survey weights in the statistical community (for a recent review, see Gelman 2007 and responses), so deciding whether and how to weight is not a simple issue. In their regression model of RDS data, Ramirez-Valles et al. (2008) cite Winship and Radbill’s 1994 article that espouses estimating the model with and without weights and using weights only if there are significant differences in the results. The justification for this is that the variance of the weights unnecessarily decreases the precision of the regression estimates if the weights are solely a function of the independent variables. However, the complicated nature of RDS design suggests to this author that the weights will rarely be exclusively based on the independent variables in the model, and we know the weights will still underestimate standard errors if there is significant clustering. If between-cluster over-dispersion is not observed, weights should be used with standard maximum likelihood equations and the sandwich standard error estimator.

Additionally, there is significant debate about how to weight models with random effects (and “multi-level” models, which are another name for models with nested random effects), so it is not clear how to incorporate the weights if one adjusts for clustering using random-effects estimators (see Pfefferman et al. 1998 and Rabe-Hesketh and Skrondal 2006 for discussion). There seems to be agreement that one must weight at each random effect level, so a modeler could use the estimated mean degree for each cluster as the basis for cluster-level weights (in addition to including the respondent-level weights).[15] However, it is not clear that the cluster-specific mean degree is the appropriate metric for defining each cluster's probability of inclusion: if one created the cluster-level weights solely as a function of mean degree, cluster size would not impact the weights at all. Given that one recruit in a cluster was recruited, however, the probability of inclusion for any other potential members of that cluster is then exclusively a function of the recruiter's degree (i.e., if the recruiter is selecting randomly from his pool of recruits, each member of the pool has the same probability of inclusion). Additionally, if a recruiter has a small degree, the assumption of sampling with replacement becomes considerably less defendable. This suggests that the weights should also reflect the number of recruits in a cluster. A reasonable middle path would be constructing the weights as a function of the cluster's mean degree then adjusting them for the number of recruits in the cluster.

The model-building procedure described above has primarily been concerned with estimating unbiased and consistent parameter estimates for the regression model. Variance estimation for these parameters will be quite complex; the current RDS population estimator relies on a custom bootstrapping algorithm because the variance cannot be derived analytically (although see Volz and Heckathorn 2008 for recent advances). Adjusting for clustering will inflate standard errors relative to those assuming simple random sampling (SRS) because it will address the redundancy of information within cluster (i.e., there will be less variance among three respondents who share a recruiter than among three randomly chosen respondents, effectively decreasing the sample size). However, this inflation will not reflect the entire sampling design.

[15] In Stata 10.1, the built-in random effects estimators only allow one set of cluster-level weights (StataCorp 2007). However, Rabe-Hesketh, Skrondal, and Pickles' GLLAMM package allows weights at each level of the model (Rabe-Hesketh, Skrondal, and Pickles 2004).

One potential approach to estimating standard errors would be to multiply the SRS standard errors by the estimated design effect (see Salganik 2006), but this method is a rather crude omnibus approach to estimating coefficient standard errors. If one inflates the standard errors by adjusting for clustering using random-effects, multiplying by the design effect would in essentially be "doubling down" on the inflation (and would therefore overestimate the standard errors). Another approach would be to calculate replicate weights to faithfully recreate the design-based variance (as is common for complex surveys), but this option is beyond the scope of this thesis and a topic for future research. For now, we can safely conclude that the lower bound of the true standard errors will be those estimated by a weighted model using the sandwich estimator. In random-effects models, the lower bound of true standard errors will be those estimated by the weighted model.

This section of the thesis has examined regression modeling of data collected using RDS (or any adaptive link-tracing sampling design), paying close attention to adjusting the regression model for observable sample network characteristics. The considerations discussed above imply a six-step approach for regression modeling of RDS data:

1. Evaluate respondent clustering at the geographic area, interview site, recruitment chain, and common-recruiter cluster levels using one-way ANOVA and/or "empty" random-effects models.

2. Determine which variables contribute to sample structuring (i.e., have high homophily) and are related to the model's dependent variable using the RDS Analysis Tool.

3. Build and diagnose a base regression model using respondent predictors and appropriate interactions.

4. Guided by the variables that from Step 2, add recruiter-level predictors and evaluate interactions between recruiter and respondent variables.

5. Guided by the clustering uncovered in Step 1, decide whether fixed-effects at the geographic area, interview site, or recruitment tree level is appropriate. If it is not, add random effects, starting at the lowest clustering level and proceeding upward (keeping in mind whether the random effects are nested or crossed). Ensure that the estimates are stable by varying the number of random-effects integration/quadrature points.

6. If a normal model is employed, use a weighted model with sandwich estimator standard errors. If random-effects models are used, explore the effects of weighting the regression model, keeping in mind that weights must be applied at each cluster level for random-effects models (if possible). If significant differences are observed between weighted and unweighted results, report both in your work.

As noted above, the modeling process may entail looping back at one or more steps, so this should serve as an outline of the tasks that need to be addressed (rather than a comprehensive, ordered list/algorithm).

6. An Empirical Example

An example often conveys more information than any amount of exposition, so this section of the thesis will build two models using a two-city sample of 643 Latino men who have sex with men collected by Jesus Ramirez-Valles and associates (see Ramirez-Valles et al. 2005 for a complete description; see Figure 1 above for a network diagram of the Chicago sample). Because social and public health researchers are often concerned with modeling binary outcome data[16], the models will examine two such variables: whether a respondent has had unprotected anal intercourse in the past 12 months and whether a respondent is HIV positive.

Table 1 displays the sample descriptive statistics. The sample was almost evenly split between Chicago and San Francisco. About 21% of respondents report having unprotected anal sex in the past twelve months, and 26% of respondents report being HIV positive. 11% of the sample has never been tested for HIV, and around half of respondents report having a sexually transmitted infection other than HIV. Approximately half of respondents were in a relationship at the time of the survey, and 69.5% identify as "gay, homosexual, or queer." Almost a quarter of respondents were born in the United States, and about half of the sample had a high school diploma or less and the other half had been to either technical school or college. 65% of respondents make less than $20,000 annually, and the sample's mean age is 35 years.

[16] The simple descriptions of within-/fixed-, between-, and random-effects models above generally refer to models with continuous outcomes. For binary and other nonlinear outcomes, the within- and between-group estimators take a different form. For example, in Stata 10.1 the "xtlogit" fixed effects command calls the clustered logit estimator and the "xtlogit" population averaged command calls the Generalized Estimating Equations estimator (which, as noted above, requires correctly specifying the error variance-covariance matrix).

Table 1 Descriptive Statistics of Ramirez-Valles' sample of 643 Latino men who have sex with men.

R City of residence	Freq.	Percent
San Francisco	323	50.23
Chicago	320	49.77
Total	643	100

R had unprotected anal intercourse in past 12 months	Freq.	Percent
No	397	78.61
Yes	108	21.39
Total	505	100

R HIV Positive	Freq.	Percent
No	473	73.56
Yes	170	26.44
Total	643	100

R HIV Unknown	Freq.	Percent
No	571	88.8
Yes	72	11.2
Total	643	100

R have Sexually Transmitted Infection (other than HIV)	Freq.	Percent
No	307	50.74
Yes	298	49.26
Total	605	100

R in relationship	Freq.	Percent
No	324	50.39
Yes	319	49.61
Total	643	100

R identifies as "gay, homosexual, or queer"	Freq.	Percent
No	196	30.48
Yes	447	69.52
Total	643	100

R born in U.S.	Freq.	Percent
No	492	77.24
Yes	145	22.76
Total	637	100

R Education	Freq.	Percent
Less than high school	172	26.75
High Scool/GED	149	23.17
Technical or vocational school	59	9.18
Some college	158	24.57
College degree	86	13.37
Graduate degree	19	2.95
Total	643	100

R Annual Income category	Freq.	Percent
Less than $10,000	260	40.44
$10,000 to $14,999	101	15.71
$15,000 to $19,999	71	11.04
$20,000 to $24,999	63	9.8
$25,000 to $29,999	57	8.86
$30,000 to $34,999	35	5.44
$35,000 to $39,999	27	4.2
More than $40,000	29	4.51
Total	643	100

R Age in years	Mean	Std. Dev.
	35.14	9.74

The first step in building the regression models is determining the degree to which the sample exhibits clustering. Table 2 displays the results of one-way

ANOVAs for three potential clustering levels: shared recruiter, recruitment tree, and 3-digit zip code.

Table 2 Results of one-way ANOVA tests for clustering.

One-Way ANOVA Tests						
Unprotected sex		Within SS	Between SS	Total SS	F-value	p-value
	Shared Recruiter	26.25	53.45	79.70	1.43	0.00
	Recruitment Tree	78.71	6.19	84.90	1.29	0.15
	3-Digit Zip Code	81.40	2.08	83.48	0.77	0.72
HIV Positive		Within SS	Between SS	Total SS	F-value	p-value
	Shared Recruiter	47.50	71.55	119.05	1.45	0.00
	Recruitment Tree	106.37	18.69	125.05	3.46	0.00
	3-Digit Zip Code	115.00	8.95	123.95	2.82	0.00

Table 2 reveals very different clustering patterns for the two outcome variables: unprotected sex is strongly clustered at the shared recruiter level but displays no clustering at the other three levels. In contrast, HIV positive displays significant clustering at all three levels. Based on these results, we can be confident that adjusting at the shared recruiter level will take care of unprotected sex; however, we will need to investigate all three clustering levels for HIV positive (remember that the recruitment tree and zip code levels are not conditional on adjusting for shared recruiter in Table 2; clustering at these levels may wash away after we adjust for shared recruiter). Table 3 displays another approach to measuring clustering: running “empty” random effects models on the outcome variables. The outcome measure, Rho, is the “intra-class correlation” (or the proportion of the total variance that is between clusters).

Table 3 Results of Empty Random Effects Model tests for clustering.

Empty Random Effects Models				
Unprotected Sex		Recruiter	Tree	3-Digit Zip
	Rho	0.275	0.024	0.000
	Rho SE	0.127	0.030	0.000
HIV Positive		Recruiter	Tree	3-Digit Zip
	Rho	0.332	0.289	0.070
	Rho SE	0.082	0.099	0.062

Table 3 confirms the results in Table 2; clustering is only observed for the shared recruiter level for unprotected sex. HIV positive clustering is observed at all three levels, but it is strongest at the shared recruiter level.

The next step in the modeling process is to examine the variables along which sample members sorted themselves (i.e., which exhibit high homophily). As noted above, it is only necessary to adjust for homophilous variables related to the model's outcome variable. Table 4 displays the homophily for the descriptive variables above crossed with the outcome variables by city (since there is no cross-city recruitment, homophily analyses must be run separately).

Table 4 Homophily of descriptive variables by outcome variables by city. (Homophily values over .15 are asterisked.)

Unprotected Sex Homophily (crossed with predictors)								
	Chicago				**San Francisco**			
	unprotected=0		**unprotected=1**		**unprotected=0**		**unprotected=1**	
Variable	**0**	**1**	**0**	**1**	**0**	**1**	**0**	**1**
Gay ID	0.237*	0.165*	-1*	0.056	-0.076	0.099	0.059	0.125
HIV Positive	-0.005	0.191*	0.01	0.327*	0.015	0.233*	-0.303*	-1*
HIV Unknown	0.06	-0.147	0.045	-1*	0.074	-1*	0.058	-1*
Club Drug Use	0.027	0.076	-1*	0.147	0.121	0.044	0.045	0.039
Hard Drug Use	0.13	0.228*	0.119	-1*	0.029	0.114	0.077	-1*
Relationship	0.042	0.004	-1*	0.134	0.039	0.004	0.073	-0.313*
Born in U.S.	0.163*	0.268*	0.093	0.135	0.235*	0.026	0.146	0.023
STD	-0.024	0.055	-0.481*	0.081	-0.026	-0.098	0.028	0.011
Unemployed	0.077	0.019	0.085	-1*	0.031	-0.371*	0.049	-1*

Table 4 (Continued)

HIV Positive Homophily (crossed with predictors)								
	Chicago				San Francisco			
	hivpos=0		hivpos=1		hivpos=0		hivpos=1	
Variable	**0**	**1**	**0**	**1**	**0**	**1**	**0**	**1**
Gay ID	0.069	0.193*	0.063	0.262*	0.119	0.2*	-0.534*	0.242*
Unprotected	-0.005	0.01	0.191*	0.327*	0.015	-0.303*	0.233*	-1*
Club Drug Use	0.037	0.001	0.25*	0.134	0.132	0.037	0.138	0.185*
Hard Drug Use	0.253*	0.22*	0.273*	0.035	0.125	0.155*	0.3*	0.008
Relationship	-0.015	-0.025	0.209*	0.194*	-0.047	0.132	0.175*	0.046
Born in U.S.	0.293*	0.402*	0.236*	0.147	0.227*	0.112	0.308*	-1*
STD	0.17*	0.008	0.043	0.242*	0.032	0.042	0.144	0.042
Unemployed	0.281*	-0.137	0.285*	0.174*	0.041	-0.386*	0.167*	0.058

Table 4 reveals the complex sorting patterns in each city. Because homophily estimates become unstable as cell size decreases (every homophily entry of -1 indicates this instability), these estimates should serve only as a guide. For the unprotected sex outcome, the only variable showing consistently non-zero homophily is HIV positive. For the HIV positive outcome, hard drug use and being in a relationship exhibit consistent homophily. Note that the homophily for the outcome variables themselves is not displayed; the estimates will not convey much information about sorting conditional on the model covariates, so one should test them in the models themselves.

After examining the sample data, clustering, and homophily patterns, we are ready to begin building the models themselves. First, we will build and diagnose models with respondent-level predictors only, including interaction terms and transformations of independent variables as necessary. Table 5 displays the respondent-level models for each of our two outcome variables.

Table 5 Respondent-level logistic regression models for Unprotected sex and HIV Positive outcomes.

Unprotected Sex Respondent-level Model

Variable		Model 1
Gay ID	Odds Ratio	**2.44**
	Standard Error	0.75
HIV Positive		**0.78**
		0.23
HIV Unknown		**1.88**
		0.70
STI		**1.57**
		0.39
Chicago		**1.07**
		0.26
Relationship		**1.67**
		0.40
Education		**1.20**
		0.09
Born in U.S.		**2.38**
		0.64
Constant		**0.04**
		0.02
	Log Likelihood	-222.41
	N	472

HIV Positive Respondent-Level Model

Variable		Model 1
Gay ID	Odds Ratio	**3.027081**
	Standard Error	0.774674
STD		**2.35173**
		0.51776
Chicago		**0.511007**
		0.112668
Relationship		**0.627237**
		0.135585
Income		**0.742501**
		0.041092
Age		**1.711121**
		0.161169
Age squared		**0.994023**
		0.00113

Table 5 (Continued)

Constant		**5.25E-06**
		1.01E-05
	Log Likelihood	-272.356
	N	605

In the Unprotected Sex model, we observe that identifying oneself as “gay, homosexual, or queer” is strongly related to having had unprotected anal sex in the past 12 months (an increase in odds by a factor of 2.44). The model also indicates that being HIV positive is associated with a decrease by a factor of .78 in the odds of having had unprotected anal sex. Not knowing one’s HIV status and having an STI are both positively related to having had unprotected anal sex, but the associations are not significant at the traditional levels (using sandwich estimator standard errors). Chicago residents exhibit slightly higher rates of unprotected anal sex than do San Francisco residents. Note that the HIV positive, HIV unknown, STI, and Chicago variables are retained even though they are non-significant; I am interested in these predictors regardless of their significance. Being in a relationship is associated with an increase in the odds of having unprotected anal sex by a factor of 1.67; if we assume that these relationships are stable, this result is consonant with the idea that respondents are less concerned about the consequences of unprotected sex with their long-term partners than with others. Higher levels of education are moderately associated with higher rates of unprotected sex, and being born in the United States is the second strongest predictor of unprotected sex after identifying as gay (with an increase in the odds by a factor of 2.38).

The HIV positive model indicates that identifying as gay is strongly associated with being HIV positive (an increase in odds by a factor of 3.03). Additionally, having a sexually transmitted infection other than HIV is positively related to having HIV. Both living in Chicago, being in a relationship, and higher income are negatively associated with being HIV positive, although it is unclear what might be driving these associations. Finally, increased age is positively associated with having HIV, but the association decreases over time (as indicated by the negative coefficient on the age-squared transformed variable).

After building and diagnosing the respondent-level models, we are ready to consider adding recruiter-level predictors. Table 6 displays the results from the unprotected sex models testing different recruiter-level predictors.

Table 6 Logistic Models including respondent- and recruiter-level predictor variables for Unprotected sex outcome.

Unprotected Sex Models Adding in Recruiter-level Predictors

Variable		Model 1	Model 2	Model 3
Gay ID	Odds Ratio	**2.44**	**2.21**	**2.36**
	Standard Error	0.75	0.69	0.76
HIV Positive		**0.78**	**0.89**	**0.91**
		0.23	0.26	0.27
HIV Unknown		**1.88**	**2.05**	**2.02**
		0.70	0.78	0.76
STD		**1.57**	**1.44**	**1.45**
		0.39	0.37	0.37
Chicago		**1.07**	**1.08**	**1.06**
		0.26	0.28	0.27
Relationship		**1.67**	**1.60**	**1.59**
		0.40	0.41	0.40
Education		**1.20**	**1.16**	**1.18**
		0.09	0.09	0.09
Born in U.S.		**2.38**	**2.61**	**2.42**
		0.64	0.74	0.67
Recruiter hard drug			**0.45**	
			0.17	
Recruiter unprotected				**0.94**
				0.20

Table 6 (Continued)

Constant		**0.04**	**0.05**	**0.04**
		0.02	0.02	0.02
	Log Likelihood	-222.41	-207.11	-209.76
	N	472	447	447

As noted above, the only variable consistently exhibiting homophily crossed with the outcome was hard drug usage. Model 2 reveals that the recruiter-level hard drug variable is associated with a significant decrease in the probability of having had unprotected sex in the past twelve months. We will retain the recruiter hard drug predictor in our model. On the other hand, recruiter-level unprotected sex in the past 12 months displays a weak relationship with unprotected sex, so we will discard it. Note that the sample size drops when recruiter-level predictors are added because seeds have no recruiter values (unless one wants to consider the researcher as the seeds' recruiter, which would require assuming that the relationships between researcher and seeds are as meaningful as the relationships between non-seed recruiters and their recruits).

Table 7 displays the results for the HIV positive models with recruiter-level predictors. As noted above, relationship and hard drug usage crossed with HIV Positive displayed high homophily, so the effects of these variables and the recruiter-level outcome variable are tested.

Table 7 Logistic Models including respondent- and recruiter-level predictor variables for HIV Positive sex outcome.

HIV Positive Models Adding in Recruiter-level Predictors					
Variable		**Model 1**	**Model 2**	**Model 3**	**Model 4**
Gay ID	Odds Ratio	**3.03**	**2.46**	**2.80**	**2.71**
	Standard Error	0.77	0.63	0.73	0.71

Table 7 (Continued)

STD		**2.35**	**2.22**	**2.27**	**2.34**
		0.52	0.51	0.51	0.54
Chicago		**0.51**	**0.55**	**0.51**	**0.52**
		0.11	0.13	0.12	0.12
Relationship		**0.63**	**0.57**	**0.59**	**0.58**
		0.14	0.13	0.13	0.13
Income		**0.74**	**0.77**	**0.76**	**0.76**
		0.04	0.05	0.04	0.04
Age		**1.71**	**1.64**	**1.68**	**1.68**
		0.16	0.15	0.16	0.16
Age squared		**0.99**	**0.99**	**0.99**	**0.99**
		0.00	0.00	0.00	0.00
Recruiter HIV positive			**2.59**		
			0.59		
Recruiter relationship				**0.66**	
				0.14	
Recruiter hard drug					**0.51**
					0.17
Constant		**0.00**	**0.00**	**0.00**	**0.00**
		0.00	0.00	0.00	0.00
	Log Likelihood	-272.36	-252.02	-259.11	-258.32
	N	605	574	574	574

Table 7 informs us that having an HIV Positive recruiter is strongly associated with being HIV Positive (an increase in odds of a factor of 2.59). While having a recruiter in a relationship or having a recruiter who uses hard drugs are both negatively associated with the odds of being HIV Positive, they do not exhibit significance and the log likelihood is significantly higher for the model including recruiter HIV Positive. Recruiter HIV Positive will be retained in the model.

Now that we have examined and selected from the potential recruiter-level predictors, we will address clustering in the two models. Table 8 exhibits the results for adding a shared recruiter random intercept to the earlier unprotected sex model.

Table 8 Logistic Models including respondent- and recruiter-level predictor variables and a shared recruiter random intercept for Unprotected sex outcome.

Unprotected Sex Models Testing Random Effects

Variable		Model 2	Model 3 (Recruiter)
Gay ID	Odds Ratio	**2.21**	**2.81**
	Standard Error	0.69	1.19
HIV Positive		**0.89**	**0.90**
		0.26	0.35
HIV Unknown		**2.05**	**2.49**
		0.78	1.21
STD		**1.44**	**1.53**
		0.37	0.50
Chicago		**1.08**	**1.15**
		0.28	0.41
Relationship		**1.60**	**1.78**
		0.41	0.56
Education		**1.16**	**1.18**
		0.09	0.13
Born in U.S.		**2.61**	**3.18**
		0.74	1.22
Recruiter hard drug		**0.45**	**0.38**
		0.17	0.18
Constant		**0.05**	**0.02**
		0.02	0.02
	Log Likelihood	-207.11	-204.63
	N	447	447
	Rho	n/a	0.35
	Rho SE	n/a	0.15

Table 8 informs us that the shared recruiter random intercept gives us a significant boost in log-likelihood, reveals that 35% of the total variance is between shared-recruiter cluster, and demonstrates that adjusting for clustering can cause shifts in the coefficient values (the maximum likelihood coefficient estimates take a different form under random-effects models, leading to unpredictable shifts in the estimates; both estimates are consistent). For example, the odds ratios on the gay identification, HIV unknown, and born in U.S. variables all increased sizably.

Table 9 displays the results from the HIV Positive outcome random effects models; recall that we exhibited significant clustering at the shared recruiter, recruitment tree, and zip-code levels in our test above.

Table 9 Logistic Models including respondent- and recruiter-level predictor variables and random intercepts for HIV Positive outcome.

HIV Positive ModelsTesting Random Effects

Variable		Model 2	Model 3	Model 4	Model 5	Model 6
-			**Recruiter**	**Tree**	**Recruiter & Tree**	**3-Digit Zip**
Gay ID	Odds Ratio	**2.46**	**2.61**	**2.46**	**2.61**	**2.46**
	Standard Error	0.63	0.75	0.64	0.75	0.65
STD		**2.22**	**2.30**	**2.22**	**2.30**	**2.19**
		0.51	0.56	0.51	0.56	0.50
Chicago		**0.55**	**0.54**	**0.55**	**0.54**	**0.54**
		0.13	0.14	0.13	0.14	0.12
Relationship		**0.57**	**0.57**	**0.57**	**0.57**	**0.56**
		0.13	0.14	0.13	0.14	0.13
Income		**0.77**	**0.76**	**0.77**	**0.76**	**0.76**
		0.05	0.05	0.05	0.05	0.05
Age		**1.64**	**1.68**	**1.64**	**1.68**	**1.63**
		0.15	0.18	0.16	0.18	0.16
Age squared		**0.99**	**0.99**	**0.99**	**0.99**	**0.99**
		0.00	0.00	0.00	0.00	0.00
rec_hivpos		**2.59**	**2.73**	**2.59**	**2.73**	**2.59**
		0.59	0.69	0.58	0.69	0.59
Constant		**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
		0.00	0.00	0.00	0.00	0.00
	Log Likelihood	-252.02	-251.59	-252.02	-251.59	-248.71
	N	574	574	574	574	566
	Rho		0.091	0.000	n/a	0.0000
	Rho SE		0.102	0.000	n/a	0.0000

Table 9 indicates that the neither the random intercepts for shared recruiter, recruitment tree, or recruitment tree and shared recruiter offer any improvement in model fit. The random intercept on 3-digit zip code does offer a slight improvement in log likelihood, but the coefficient estimates do not shift from the normal model and the rho value is virtually zero. Based on the lack of parameter changes and the dictum that simpler is almost always better, I will choose to reject all the models containing random intercepts. These results make clear the tentative status of the clustering analysis performed in the first step.

Our final step in building the models is examining the effects that weighting the data has on our results. As noted above, weights should be used when standard estimators are used. There is significant debate on how to incorporate weights into random-effects models, so the decision of whether to use weights is more tentative in that scenario. Table 10 displays the results of applying weights to our random-intercept model of Unprotected Sex.[17]

Table 10 Random-effects Logistic Models including respondent- and recruiter-level predictor variables and testing weights for Unprotected Sex outcome.

Unprotected Sex Models Comparing Cluster-level Random Effects weights

		Model 3	**Model 7**	**Model 8**
Variable		Unweighted	Mean network size weights	Number of recruitments weights
Gay ID	Odds Ratio	**2.81**	**1.75**	**1.62**
	Standard Error	1.19	0.69	0.65
HIV Positive		**0.90**	**1.52**	**1.63**
		0.35	0.59	0.64
HIV Unknown		**2.49**	**1.93**	**2.12**
		1.21	0.94	1.07
STD		**1.53**	**1.53**	**1.40**
		0.50	0.49	0.46
Chicago		**1.15**	**1.14**	**1.08**
		0.41	0.40	0.38
Relationship		**1.78**	**1.16**	**0.75**
		0.56	0.36	0.24
Education		**1.18**	**1.23**	**1.14**
		0.13	0.13	0.13

[17] These models were estimated using Stata's "xtlogit" random-effects command. As noted above, this command only allows weights to be applied at the cluster level.

Table 10 (Continued)

Born in U.S.		**3.18**	**3.88**	**3.87**
		1.22	1.48	1.58
Recruiter hard drug		**0.38**	**0.23**	**0.17**
		0.18	0.12	0.10
Constant		**0.02**	**0.03**	**0.06**
		0.02	0.02	0.04
	Log Likelihood	-204.63	-201.54	-212.54
	N	447	447	447
	Rho	0.35	0.32	0.37
	Rho SE	0.15	0.18	0.20

Table 10 indicates that there are significant effects to applying cluster-level weights to our random-effects model. The weights in Model 7 are composed exclusively of the cluster-specific mean degree estimates; the weights in Model 8 alter those in Model 7 to reflect the lower probability of observing multi-member clusters (i.e., the probability of observing one member of a cluster is 1/[mean_degree], whereas the probability of observing all three members of a three-member cluster is (1/[mean_degree])^3)). The change in log likelihood and coefficient estimates does not exhibit a detectable pattern across clusters. Additionally, the direction of the HIV Positive coefficient changes, which makes me suspicious of these results. Knowing the state of the literature and observing the unpredictable changes across models, I will retain the unweighted model as my final model (noting in my results that cluster-level weights significantly altered the coefficient values).

Table 11 displays the results of applying weights to our logistic models of HIV Positive. It indicates that there is a significant shift in log likelihood and in model coefficients when we apply the respondent-level weights. Some variables' coefficients are amplified, and others' coefficients are moderated. Additionally, the effect of our recruiter-level predictor is attenuated significantly. Because the coefficient patterns are consistent across models (i.e., are all in the same direction and similar magnitude) and the literature is clearer about weighting in individual level models, I will choose to retain the weighted model as my final model.

Table 11 Logistic Models including respondent- and recruiter-level predictor variables and testing weights for HIV Positive outcome.

HIV Positive Models Comparing respondent-level weights

Variable		Model 2	Model 7
Gay ID	Odds Ratio	**2.46**	**3.31**
	Standard Error	0.63	1.16
STD		**2.22**	**2.85**
		0.51	0.92
Chicago		**0.55**	**0.32**
		0.13	0.11
Relationship		**0.57**	**0.52**
		0.13	0.16
Income		**0.77**	**0.73**
		0.05	0.07
Age		1.64	1.56
		0.15	0.21
Age squared		0.99	1.00
		0.00	0.00
rec_hivpos		2.59	1.84
		0.59	0.58
Constant		0.00	0.00
		0.00	0.00
	Log Likelihood	-252.02	-223.49
	N	574	574

As the above makes clear, there are not hard-and-fast rules for choosing among models in the later stages of the modeling process. If modelers report conflicting results and honestly represent their confidence in the models' conclusions, they have done due diligence and can be confident in the validity of their results.

This section has demonstrated the six steps of the RDS modeling process. As the examples have made clear, there are many contingencies to deal with as one moves through the steps, so more care is needed than if one were modeling data collected using a simple random sample. Additionally, the conclusions one would draw based on the above models are sensitive to the different assumptions and adjustments, so simply modeling RDS data as if were collected as a random sample is not an option.

7. Summary and Conclusion

In conclusion, researchers must be particularly careful when modeling data collected with a complex sampling design. Respondent-Driven Sampling is one such design that is further complicated by the network-based properties of the method. In order to estimate consistent parameter estimates, the modeler must take into account the homophily inherent in the recruiting process and the network clustering underlying the sample. This thesis has described how modelers can glean information about recruitment homophily and clustering from sample data, and it suggests adjusting for the former by entering recruiter values directly as predictors and the latter by employing fixed- or random-effects estimators as appropriate. The thesis continues by noting the extreme complexity and ongoing debate about RDS variance estimation and how this applies to regression models of RDS data, proposing a simple estimate of the standard error lower bounds. Additionally, it briefly remarks on the ongoing debate about survey weighting of random-effects and multi-level models, making some simple suggestions for choosing whether or not and how to weight. The thesis concludes by modeling two different outcome variables from the same sample, resulting in two very different adjustment strategies.

This thesis represents an initial approach to the regression modeling of data collected using RDS (or other adaptive link-tracing sampling designs). It examines the theoretical considerations modelers must take into account and suggests a six-step guide for modeling. Clearly this is only the first step in modeling RDS data, and much more research is needed (particularly on variance estimation and regression weighting); however, the procedure outlined in the thesis should serve as a reasonable guide for modelers until such research is available.

APPENDIX

Respondent-Driven Sampling: Basic Concepts

Respondent-Driven Sampling (RDS) is a “chain-referral” or “link-tracing” method for sampling “hidden” populations (those populations for which no sampling frame can be constructed). Similar to snowball sampling, RDS begins with researchers non-randomly selecting initial participants or “seeds” for the study. After the seeds are interviewed, they are given multiple (usually three or four) numbered dollar bill-sized coupons with which they can recruit friends and acquaintances who are also population members into the study. Respondents are remunerated both for taking the survey and for each successful recruitment they make.

The first insight that advances RDS beyond snowball sampling is that, after a number of recruitment cycles (or “waves”), the sample composition will become independent of the initial seeds. This point is termed “equilibrium,” and it is usually reached within five or fewer waves. If the sample has attained equilibrium, dyadic ties (recruitments) will be randomly sampled from the network with equal probability, and any population member’s probability of inclusion is proportional to her number of ties within the population (Heckathorn 1997; Heckathorn 2002; Salganik and Heckathorn 2004).

RDS estimation relies fundamentally on information about the network structure of the population, which is gathered by recording the characteristics of both recruiter and recruit for every recruitment in the sample. Thus, the unit of analysis for the RDS estimator is the recruitment dyad, defined by the recruiter and recruit values for the variable of interest. For example, a recruitment of a female by a male could be termed an “MF” tie, and a recruitment of a male by a female could be termed an “FM” tie. From sample recruitment data, a “recruitment matrix” is created that displays the number of ties to and from each group (or value on the variable of interest). Unbiased transition probabilities, or the probability that a given recruiter will recruit someone of each different group, can be estimated from the recruitment matrix and used for RDS population proportion estimation (see Heckathorn 2007 for proof).

The link between the information collected in the sample and RDS estimation theory is the “Reciprocity Model” (Heckathorn 2002). Under this model, it is assumed that every relationship a recruiter has that could lead to recruitment is symmetric (i.e., the relationship is strong enough that each member would recognize the other as a potential recruit). This assumption has been examined by asking sample respondents what type of relationship they have with their recruiter, and more than 98% of relationships are friends or acquaintances in most RDS studies (indicating that this assumption holds).

In this formulation, the number of ties from group X to group Y (Txy) in a two-group system can be expressed as the product of four parameters:

$$Txy = NPxDxSxy$$

where N is the population size, P_x is the proportional size of group X, D_x is the average network size of group X, and S_{xy} is the proportion of X-originating ties that are to group Y members (see Heckathorn 2007 for a detailed discussion). When the reciprocity assumption is met, the number of ties from group X to group Y must equal the number from group Y to group X. Because the two groups’ proportional sizes sum to one, the following two equations represent the system:

$$1 = Px + Py$$

$$NPxDxSxy = NPyDySyx$$

When solved for group X's proportional size, P_x, it can be seen that the RDS population estimator depends exclusively on four pieces of information, the proportion of cross-group ties going each direction and each group's mean degree:

$$\hat{P}x = \frac{SyxDy}{SyxDy + SxyDx}$$

Note that the total population size, N, drops out of the solution; RDS can only estimate population proportions, not population sizes (however, population size estimation using RDS and the "capture-recapture" method is described in Heckathorn and Jeffri 2003).

Because RDS estimation requires discrete groups, analyzing continuous variables entails partitioning the variable into groups (and therefore requires some loss of information). Heckathorn (2007) provides a means for determining how much a continuous variable must be aggregated based on the desired average number of recruitments in the cells of the recruitment matrix:

$$AL = \sqrt{\frac{n}{n_c}}$$

where AL is the number of groups the variable is partitioned into, n is the sample size, and n_c is the mean number of cases per cell. Analysis suggests that there is a "zone of convergence" for the estimates of average cell sizes 12 ± 4 (Heckathorn 2007). Fortunately, the "dual component" weighting approach requires no loss of information for mean estimation of continuous variables.

Average Group Network Size Estimation

Each individual's network size (or "degree") is estimated by asking respondents how many members of the target population they know (e.g., "How many intravenous drug users between the ages of 18 and 65 do you know, who also know you, and who you have seen in the last month?"). Work by Marsden (1990) indicates that degree estimators of this type are fairly robust, and the RDS population estimator is only responsive to *relative* degree (not absolute degree), so uniform biases in self-reported network size do not alter RDS population estimates. However, further work developing and testing network size questions is certainly worth pursuing.

Because respondents with large network sizes are oversampled (they have more potential recruitment paths leading to them), a standard arithmetic mean would overestimate group X's true mean network size. To correct for this, the ratio of two Hansen-Hurwitz estimators (a type of harmonic mean) is employed:

$$\widehat{D_x} = \frac{n_x}{\sum_{i=1}^{n_x} \frac{1}{D_i}}$$

where $\widehat{D_x}$ is group X's estimated mean degree, n_x is the number of group X respondents in the sample, and D_i is respondent i's self-reported degree. Work by Salganik and Heckathorn (2004) shows that this estimator is asymptotically unbiased, which means bias is on the order of 1/[sample size] (Cochran 1977). The RDS estimator includes degree estimates in both the numerator and denominator. Because the ratio of asymptotically unbiased estimators is asymptotically unbiased, the RDS estimator is also asymptotically unbiased.

Note that this degree estimator does not control for differential recruitment by degree. That is, if respondents of high degree also recruit more effectively, the estimated mean degree will still be biased upward. Differential recruitment by degree is usually very mild in RDS studies due to the strict quota on the number of recruitments each recruiter can make, but certain sampling scenarios make it desirable to relax this restriction. Fortunately, applying the individualized weighting strategy described below to the degree variable and weighting the RDS estimator eliminates this source of bias (see Heckathorn 2007 for details).

Continuous Variable Mean Estimation

The RDS sampling weight for any group X is the ratio between the RDS population proportion estimate and the sample proportion,

$$W_x = \frac{\hat{P}_x}{C_x}$$

The first step toward "individualizing" the sampling weight is to separate it into two elements, one quantifying the role of differential group recruitment and the other the role of differences in group average network size (Heckathorn 2007). This can be achieved by estimating what the sample proportions would have been without systematic differences in recruitment (homophily). Heckathorn (2002) presents an appropriate approach, which involves modeling the recruitment process as a "first-order Markov process." Each group corresponds to a state in the fixed state space, and the recruitment proportions (from the recruitment matrix) represent the transition probabilities. As noted above, when the assumptions are met the sample reaches an "equilibrium" composition that is independent of the state with which recruitment began (the seed's group). The equilibrium proportions for a two-group system can be calculated as

$$\hat{E}x = \frac{Syx}{Syx + Sxy}$$

This equation is equivalent to the RDS population estimator without an average degree component. If the two groups in the system have equal average degrees, the RDS population estimator reduces to the above equilibrium estimator. Whether or not there are differences in degree, this estimator successfully projects what the sample composition would have been without differential recruitment and will allow us to disaggregate the RDS sampling weight.

The RDS sampling weight can be separated into the "degree component," DC_x, and the "recruitment component," RC_x. DC_x is calculated as

$$DC_x = \frac{\hat{P}_x}{\hat{E}_x}$$

As noted above, when degrees are equal $\hat{P}_x = \hat{E}_x$ and, therefore, $DC_x = 1$. If Group X has a greater mean degree than Group Y, it is oversampled, so $\hat{P}_x < \hat{E}_x$ and $DC_x < 1$. Conversely, if Group X had a smaller mean degree than Group Y, $DC_x > 1$. The recruitment component RC_x is calculated as

$$RC_x = \frac{\hat{E}_x}{C_x}$$

If recruitment effectiveness (homophily) is equal for Groups X and Y, $RC_x = 1$. If Group X recruited more effectively than Group Y, $\hat{E}_x < C_x$ and $RC_x < 1$.

After partitioning (aggregating) and analyzing a continuous variable, the degree and recruitment weight components are calculated. They will be identical for all members of a group. To adjust the weights for continuous variable mean estimation, the group-level degree component is replaced with an individual-level degree measure (termed a “multiplicity adjustment”) as follows:

$$DC^i = K\frac{1}{D^i}$$

where DC^i is individual *i*’s degree component, K is a positive constant, and D^i is individual *i*’s degree. If we impose the constraint that the individualized weights sum to *n*, the degree component for individual *i* in group X is

$$DC^i = \frac{n}{\sum_{j=1}^{n} \frac{1}{D^j} RC^j} \cdot \frac{1}{D^i}$$

where the summation is over all individuals in the sample. This approach gives us the tools to eliminate bias due to differential recruitment by degree and to estimate the population mean of a continuous variable.

RDS Assumptions

Overall, determinants of RDS validity can be distilled to six core assumptions about the sampling process (Heckathorn 2007):

1. Respondents know each other as members of the target population, so recruitment ties are reciprocal.
2. Sampling occurs with replacement.
3. The population of interest is linked by a single-component network.
4. Recruitment chains extend deep enough into the network to overcome initial seed selection bias (i.e., obtain equilibrium).
5. Respondents can accurately report their network size.

6. Recruitment is a random selection from eligible members of the recruiter's network.

Assumption one corresponds to the basic assumption of the Reciprocity Model that every recruitment relationship is symmetrical and all sample participants are actually population members (see above). Assumption two is required by the mathematics underlying RDS estimation; in practice, this assumption holds as long as the sampling fraction is not greater than approximately 10% (Cochran 1977).

The third assumption is necessary because the proof of random tie sampling requires every member of the population to have a non-zero inclusion probability (Salganik and Heckathorn 2004). If the network is comprised of many small, disconnected clusters, it is likely that some clusters would not be linked to the overall network at all. Fortunately, work in network graph theory indicates that most nodes are members of one large component even in relatively sparse graphs (Newman 2003). Additionally, work on the "small world problem" shows that in most real-world social networks any two nodes are linked through a relatively small number of steps (Watts and Strogatz 1998; Watts 1999; Dodds et al., 2003).

The fourth assumption specifies the equilibrium requirements for overcoming initial seed-selection bias, but it also serves to bolster assumption three. Specifically, there is some degree of clustering in all real-world networks; if all seeds for a study were members of non-central clusters, multiple waves would be required to assure that the primary component is tapped and all members of the population have non-zero inclusion probabilities.

The fifth assumption is required because average group degree estimates are a central component of the RDS estimator. As noted above, RDS is only sensitive to relative degree, so uniform inflation or deflation of degree is non-problematic. Additional work in this area would be helpful.

The sixth assumption asserts that respondents recruit as though they are recruiting randomly from their pool of potential recruits. If taken literally on the individual level, this assumption seems unlikely to hold (i.e., it seems highly implausible that anybody would list their contacts and randomly select whom they would recruit). However, there is little reason to suspect that respondents could or would collude to selectively recruit in the same way. RDS studies have compared self-reported network composition with recruitment behavior and found a strong relationship (Heckathorn et al. 2002, Wang et al. 2005). Additionally, one would expect inter-group recruitment patterns to be asymmetric if this assumption was not met (i.e., the number of male to female recruitments would not be approximately equal to the number of female to male recruitments in a sample if there were recruitment bias on gender). Ramirez-Valles et al. (2005b) found that differences in recruitment patterns were not significantly asymmetric across groups.

REFERENCES

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

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

Gelman, Andrew. 2007. “Struggles with Survey Weighting and Regression Modeling.” *Statistical Science*, 22(2):153-164.

Gelman, Andrew and Jennifer Hill. 2007. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. New York, NY: Cambridge University Press.

Greene, William H. 2003. *Econometric Analysis, 5th Ed*. Upper Saddle River, NJ: Prentice Hall.

Handcock, Mark S., A.E. Raftery, and J.M. Tantrum. 2007. “Model-based Clustering for Social Networks." *Journal of the Royal Statistical Society A*, 170(2): 301-354.

Hausman, J.A. 1978. “Specification tests in econometrics.” *Econometrica* 46(6):1251-1271.

Heckathorn, Douglas 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(1):11–34.

--------------. 2007. “Extensions of Respondent-Driven Sampling: Analyzing Continuous Variables and Controlling for Differential Recruitment.” *Sociological Methodology*, 37(1):151-207.

Heckathorn, Douglas D. and Joan Jeffri. 2003. “Social Networks of Jazz Musicians.” Pp. 48-61 in *Changing the Beat: A Study of the Worklife of Jazz Musicians.* Vol. 3. *Respondent-Driven Sampling: Survey Results by the Research Center for Arts and Culture*, National Endowment for the Arts Research Division Report #43. Washington, D.C.

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

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

McPherson, J.M. and L. Smith-Lovin. 1987. “Homophily in voluntary organizations: status distance and the composition of face-to-face groups.” *American Sociological Review* 52:370–79.

Morgan, Stephen L. and Christopher Winship. 2007. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. New York, NY: Cambridge University Press.

Newman, M.E.J. 2003. ‘‘The Structure and Function of Complex Networks.’’ *SIAM Review* 45:167–256.

-----------. 2006. “Finding community structure in networks using the eigenvectors of matricies.” *Physical Review E*, 74:036104.

Neuhaus, J.M., J.D. Kalbfleisch, and W.W. Hauck. 1994. “Estimation in mixed-effects models.” *Canadian Journal of Statistics*, 22(1):139-148.

Pendergast, Jane, Stephen J. Gange, Michael A. Newton, Mary J. Lindstrom, Mari Palta, and Marian R. Fisher. 1991. “A Survey of Methods for Analyzing Clustered Binary Response Data.” *International Statistical Review*, 64(1):89-118.

Pfeffermann, D, C.J. Skinner, D.J. Holmes, H. Goldstein, and J. Rasbash. 1998. “Weighting for unequal selection probabilities in multilevel models.” *Journal of the Royal Statistical Society B*, 60(1):23-40.

Pinheiro, J., and D. Bates. 2000. *Mixed Effects Models in S and S-Plus*. New York: Springer.

Rabe-Hesketh, Sophia and Anders Skrondal. 2005. *Multilevel and Longitudinal Modeling Using Stata*. College Station, TX: Stata Press.

Rabe-Hesketh, Sophia and Anders Skrondal. 2006. “Multilevel modeling of complex survey data.” *Journal of the Royal Statistical Society A*, 169(4):805-827.

Rabe-Hesketh, Sophia, A. Skrondal, and A. Pickles. 2004. *GLLAMM Manual*. U.C. Berkeley Division of Biostatistics Working Paper Series. Working Paper 160.

Ramirez-Valles, Jesus, Douglas D. Heckathorn, Raquel Vazquez, 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(4): 387-402.

----------. 2005b. “The Fit between Theory and Data in Respondent-Driven Sampling: Response to Heimer.” *Aids and Behavior*, 9(4):409-414.

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(6):1036-1042.

Salganik, Matthew J. 2006. "Variance estimation, design effects, and sample size calculations for respondent-driven sampling." *Journal of Urban Health*, 83:98-111.

Salganik, Matthew 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. 2007. “Social Networks of Aging Artists.” In *Above Ground: Information of Artists III: Special Focus New York City Aging Artists*, Ed. Joan Jeffri.

StataCorp. 2007. *Stata Statistical Software: Release 10*. College Station, TX: StataCorp LP.

Thompson, S.K. 2002. *Sampling*, Second Edition. New York: Wiley.

Thompson, Steven K. and Ove Frank. 2000. “Model-based estimation with link-tracing sampling designs.” *Survey Methodology*, 26(1):87–98.

Volz, Erik and Douglas D. Heckathorn. 2008. “Probability based estimation theory for Respondent Driven Sampling.” *Journal of Official Statistics,* 24(1):79-97.

Wang, Jichuan, R.G. Carlson, R.S. Falck, H.A. Siegal, A. Rahman, and L. Li. 2005. “Respondent-Driven Sampling to Recruit MDMA Users: A Methodological Assessment.” *Drug and Alcohol Dependence* 78:147-157.

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

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

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

Winship, Chris and L Radbill. 1994. “Sampling Weights and Regression Analysis.” *Sociological Methods and Research*, 23(2):230-257.

Wooldridge, J.M. 2002. *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: MIT Press.