DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 32, No. 1, 2016, pp. 29–73, http://dx.doi.org/10.1515/JOS-2016-0002

Respondent-Driven Sampling – Testing Assumptions: Sampling with Replacement

Vladimir D. Barash[1], *Christopher J. Cameron*[2], *Michael W. Spiller*[3], *and Douglas D. Heckathorn*[4]

Classical Respondent-Driven Sampling (RDS) estimators are based on a Markov Process model in which sampling occurs with replacement. Given that respondents generally cannot be interviewed more than once, this assumption is counterfactual. We join recent work by Gile and Handcock in exploring the implications of the sampling-with-replacement assumption for bias of RDS estimators. We differ from previous studies in examining a wider range of sampling fractions and in using not only simulations but also formal proofs. One key finding is that RDS estimates are surprisingly stable even in the presence of substantial sampling fractions. Our analyses show that the sampling-with-replacement assumption is a minor contributor to bias for sampling fractions under 40%, and bias is negligible for the 20% or smaller sampling fractions typical of field applications of RDS.

Key words: Respondent-driven sampling; hidden populations; sampling with replacement.

1. Introduction

Respondent-Driven Sampling (RDS) has become the method of choice for studies of hidden and hard-to-reach populations, yet important questions regarding the method remain unresolved. RDS is a form of network sampling paired with an estimation strategy, where individuals are treated as network nodes and their social relationships are treated as edges.

Drawing an RDS sample involves several steps. First, when sampling from a hidden population, one begins with a convenience sample of initial respondents who serve as "seeds". Seeds can be identified by key informants who are drawn from organizations where the target population congregates, or they may self-identify by volunteering for the study. Second, initial respondents each recruit several peers, who compose the sample's first "wave". Third, the first-wave recruits each recruit several peers, who form the sample's second wave. Thus the first-wave recruits become the recruiters of the second wave. Fourth, the sample expands in this recursive manner, wave by wave, with the prior wave's recruits becoming the recruiters of the subsequent wave, until the desired sample size has been reached.

[1] Graphika Inc., 116 West 23rd Street, 5th Floor, New York NY 10011, U.S.A. Email: vlad.barash@graphika.com
[2] Cornell University – Sociology, 344 Uris Hall Ithaca, New York 14853, U.S.A. Email: cjc73@cornell.edu
[3] Cornell University – Sociology, 344 Uris Hall, Ithaca, New York 14853, U.S.A. Email: mws24@cornell.edu
[4] (Corresponding author) Cornell University – Sociology, 344 Uris Hall Ithaca, New York 14853, U.S.A. Email: douglas.heckathorn@cornell.edu
Acknowledgments: This research was made possible by a grant from the National Institutes of Health/National Institute for Nursing Research (1R21NR10961). We thank members of the Symposium on Respondent-Driven Sampling, Department of Mathematics, Stockholm University for helpful comments and advice.

One essential feature of the sampling method includes keeping track of who has recruited whom. This is important because affiliation patterns (e.g., members of a racial/ethnic group tending to recruit members of the same racial/ethnic group) affect the composition of the sample. A second essential feature of the sampling method is asking each respondent how many members of the target population they know as acquaintances, friends, or closer than friends. The people in the target population known to an individual node defines the node's network neighborhood, and the number of nodes in the neighborhood defines the node's degree. Nodes of larger degree tend to be oversampled because they have a larger number of edges that serve as peer-recruitment paths.

The RDS estimators employ information about a respondent's degree and their affiliation patterns to correct for sources of bias inherent in chain-referral sampling. Specifically, in the computation of the RDS estimator, the estimated size of each of the population's subgroups is inflated or deflated based on whether the subgroup was judged to be under- or oversampled. In sum, the RDS estimator functions some what like a corrective lens that compensates for network-based sources of bias in the sampling process.

The advantage of RDS is that it provides a means for drawing probability samples of populations which cannot be effectively sampled using traditional population survey methods because they lack a sampling frame, and because these populations have social networks that are hard for outsiders to penetrate due to stigma or privacy concerns.

RDS studies have focused both on populations of relevance to public health, such as injection drug users (IDUs), men who have sex with men (MSM), and commercial sex workers (CSW); on populations of relevance to arts and culture such as jazz musicians (Heckathorn and Jeffri 2001) and visual artists (Jeffri et al. 2011); on hard-to-reach or rare general populations such as low-wage workers (Bernhardt et al. 2012) and Canadian urban aboriginals (Smylie et al. 2011); and on criminological populations such as underage sex trafficking victims (Curtis et al. 2008). A 2009 survey (Malekinejad et al. 2008) analyzed the results of 128 studies drawn from more than 28 countries. RDS has been employed in studies funded by agencies including, the Centers for Disease Control and Prevention (CDC), CDC/Global AIDS, Gates India, the United States Agency for International Development (USAID), the National Science Foundation (NSF), and National Institutes of Health (NIH) institutes including the National Institute on Drug Abuse (NIDA), the National Institute of Mental Health (NIMH), the National Institute on Child Health and Human Development (NICHD) and the National Institute of Nursing Research (NINR).

The popularity of RDS derives in large part from a proof (Salganik and Heckathorn 2004) showing that when the assumptions of the method are satisfied, population estimates are asymptotically unbiased. This means that bias is only on the order of $1/n$, where n is the sample size, so bias is trivial in samples of significant size. A subsequent paper (Heckathorn 2007) reduced by one the number of assumptions required by the method, so the proof of lack of bias depends on four conditions: (1) the network connecting the population is dense enough to form a single component; (2) recruiters know one another, as acquaintances, friends, or those closer than friends, so their relationships are reciprocal; (3) respondents recruit as though they are selecting randomly from their neighborhoods; and (4) sampling occurs with replacement. In this article, unless otherwise specified, we use the Heckathorn 2007 estimator because it requires fewer assumptions than other

estimators, and it controls for a form of bias ignored by other RDS estimators, differential recruitment by degree. (For a comparison of estimators, see Heckathorn 2011.)

The first three assumptions can be approximated given a suitable choice of population to study and of research design. Given its reliance on network-based recruitment, RDS is a method suited only for studying populations with relatively dense networks. Generally, this involves populations united by a contact pattern, that is, network connections created by virtue of membership in the population. For example, drug users form ties when purchasing and using drugs and jazz musicians form ties when performing in an ensemble. When membership in a population does not create contact patterns, as is the case for tax evaders, the population is not suitable for study using RDS or any other network-based method. Hence, assumption one determines the conditions under which RDS is a suitable sampling methodology.

The second assumption, that respondents know one another well enough for their relationships to be reciprocal, can be satisfied by appropriate research design. This involves making recruitment rights both scarce and valuable. Generally this is done through a combination of rewards for peer recruitment and quotas to limit the number of peers who can be recruited. Quotas are implemented by giving each respondent, that is, each potential recruiter, a limited number of recruitment coupons. Each coupon has a unique serial number which allows the recruiter to whom the coupon was given to be linked to the recruit who brings it into the study site. When recruitment rights are scarce and valuable, few respondents waste them on strangers who may fail to take advantage of the recruitment opportunity. Consequently, recruitment by strangers is generally infrequent—less than a few percent (Iguchi et al. 2009)—and these recruitments can be identified by asking respondents about their relationship to their recruiter, and then deleting these cases to produce a data set consisting exclusively of respondents who know one another, as acquaintances, friends, or closer than friends. The rationale for the reciprocity assumption is that for any individual, those who consider him or her as acquaintances or friends (i.e., the individual's "in-degree"), also tend to be viewed by the individual as an acquaintance or friend (i.e., the individual's "out-degree"). Because in-degree and out-degree are equivalent in networks where all ties are reciprocal, we refer merely to "degree".

The third assumption, that respondents recruit randomly from their neighborhoods, can be best approximated when opportunities to be interviewed are easily and safely available to all members of the neighborhood because respondents have no incentive to selectively favor or exclude any particular neighbors (Heckathorn 2007, 163–164). Here it is important to note that individuals are not assumed to recruit randomly from the target population, for this target population is generally far larger than any node's neighborhood. Furthermore, the composition of nodes' neighborhoods varies greatly, because those similar in race/ethnicity, education, income, religion, and other factors tend to affiliate, so neighborhoods are often relatively homogeneous, a factor termed homophily. Hence, when a node recruits randomly from its neighborhood, this does not mean that it is sampling randomly from the target population. Support for this assumption of random recruitment from neighborhoods has been found in several studies (e.g., Heckathorn et al. 2002; Wang et al. 2005); however, the conditions under which it holds or is violated warrant further study.

The final assumption, that sampling occurs with replacement, has a unique status because it is invariably counterfactual irrespective of choice of population or research

design; respondents can only be interviewed once, and hence replacement is excluded. Sampling with replacement is not feasible in practice because RDS survey respondents are generally compensated for their participation and allowing a single individual to participate multiple times could lead to strategic recruiting behavior by participants. Methods have been developed to reduce subject duplication through a database that records scars, tattoos, and biomarkers (Heckathorn et al. 2001). These methods are necessary in RDS sampling because most RDS studies require subject anonymity due to the sensitive nature of questions in these studies: some RDS studies have asked about sexual and drug use history (Iguchi et al. 2009), while others have asked employees to report their employers' violations of workplace laws (Bernard et al. 2010). Even if sampling with replacement did not create perverse recruitment incentives, it would be inefficient and costly to interview a person each time they were recruited. Compared to a hypothetical implementation of with-replacement RDS, without-replacement RDS yields data from a greater number of unique individuals.

If respondent-driven sampling were conducted with replacement and in accordance with the other assumptions, the RDS process would be a simple random walk on the network conforming to a Markov Process. Previous studies (Volz and Heckathorn 2008) have assumed that, for very small sampling fractions, the sampling-with-replacement assumption is valid; in our analysis, we put these statements to the test and, more broadly, explore the potential for bias resulting from this counterfactual assumption.

The general expectation in statistical analyses is that a sampling-with-replacement assumption, which is equivalent to the assumption of an infinite population size (Gile 2011, 32), becomes more problematic as the sampling fraction increases. This may suggest that the replacement assumption does not significantly bias RDS studies with small sampling fractions. For example, in the CDC's National HIV Behavioral Surveillance Injection Drug User (NHBS-IDU) study, the sampling fraction for the 23 study sites had a median of 2.3% and a range of 0.6% to 8% (Lansky et al. 2009). Similarly, in the NEA-funded study of jazz musicians, the sampling fractions were 0.8% and 1.6% in New York and San Francisco, respectively (Heckathorn and Jeffri 2001). However, given the unique nature of RDS sampling, especially the interdependence of observations owing to the tendency of respondents to recruit others like themselves, there can be no guarantee that this rule of thumb is valid.

Furthermore, there are inevitably studies in which the research design calls for a large sample from a small population. In such cases, the sampling fraction is large. The largest sampling fraction with which this research team has been directly involved was a study of IDUs in a small Connecticut town with a population of 59,000 of whom slightly more than 1,000 were injectors. The sampling fraction was estimated at a rather substantial 37% via capture-recapture using a combination of RDS and police statistics (Heckathorn et al. 2002).

In this article, we extend work by Gile (2011) and Gile and Handcock (2010) in analyzing bias introduced by violation of the sampling-with-replacement assumption, as well as overall bias of RDS estimates. We employ formal proofs complemented with simulation to explore the relationship between bias and sampling fraction, including factors that affect this relationship, such as network density, homophily and the degree distribution in the population. Our results suggest that under certain circumstances,

RDS estimates are surprisingly stable even in the presence of high (i.e., 50% or greater) sampling fractions.

Previous work, such as the Successive Sampling Estimator (Gile 2011), has explored solutions to reducing the bias of the Volz-Heckathorn (VH) estimator (Volz and Heckathorn 2008) for high sampling fractions. Such work is a useful exploration of the boundary conditions for RDS studies; however, a major limitation is that it focuses exclusively on sampling fractions of 50% to 95%, and does not explore the magnitude of bias in low and moderate sampling fractions. Sampling fractions approaching 100% are more akin to census-level studies, where the population proportions can be calculated directly, than to random samples requiring population estimators. In the limit of 100% sampling fraction (census), a simple sample proportion gives the correct prevalence value without needing an estimator. Therefore, while the Successive Sampling Estimator performs well at fractions over 50%, it is important to note that for some of the range of sampling fractions the Gile 2011 study covers, census-level measures, such as simple sample proportion, are often more appropriate for parameter estimation than RDS estimators such as VH or successive sampling. A variance-estimation method appropriate for RDS is still required as simple random sample variance estimates will be much too small. Note that Gile (2011) compares the Successive Sampling Estimator to the VH estimator. As long as all groups recruit equally effectively (as is the case in Gile's simulations and in this article's simulation), the VH and Heckathorn (2007) estimators produce identical point estimates and may be directly compared (Volz and Heckathorn 2008; Heckathorn 2007). In the discussion section of this study below, we address in more detail the differences between the Successive Sampling Estimator, the Heckathorn (2007) estimator and the simple sample proportion.

What we explore in this article is a range of sampling fractions not examined in previous studies, which were limited to a range from 50% to 95% (Gile 2011). Our key finding is that RDS estimates within the parameter ranges we examine (i.e., sampling fractions between 5% and 80%) are surprisingly unbiased even in samples with high (i.e., 50% or greater) sampling fractions. We did not examine sampling fractions in excess of 80%, because few field studies reach such a high proportion of the population. Gile (2011) has already examined this upper range and demonstrated that very large sampling fractions result in significant VH estimator bias. We found that expected bias is less than five percentage points away from the true population proportion across the ranges of sampling fraction, homophily, and degree distribution examined in our simulations.

We also find that violation of the sampling-with-replacement assumption is not a major contributing factor to bias in RDS simulations for sampling fractions under 40%; other factors, such as relative mean degree of the target group, tend to affect bias more than sampling without replacement in these conditions. In particular, both the mean and the 95% confidence intervals for sampling with and without replacement are essentially identical (within 1–2% of each other) for sampling fractions up to 20%.

Finally, we analyze the sampling variability of the RDS method, which we define as the width of the 95% confidence interval of the sampling distribution over a set of simulated RDS samples, using the Heckathorn (2007) estimator to calculate sample prevalence. We find that the sampling variability of RDS decreases with increasing

sampling fraction for sampling fractions under 40%. In combination, these findings suggest an optimal range of one to 20% of sampling fractions for RDS studies using the Heckathorn (2007) estimator.

The rest of this article is organized as follows: We first present a formal model of an RDS sampling process, with a few additional assumptions for simplicity. We use this model to derive several results about the extent of bias due to sampling without replacement in different network structures, for different values of sampling fraction and (for a two-group system) homophily and relative group size. Next, we confirm our formal results with simulations, and extend them further by offering simulations that go beyond the scope of our proofs. Finally, we offer a brief discussion of the analysis with implications for future RDS studies in the field, and conclude with directions for future work in this area.

2. Formal Model

For reference, we begin this section with a glossary of terms and notation used throughout the rest of the article. Notation is presented in Table 1.

Network Terms:

Social Network. The social network is the set of individuals in a population (nodes) and a set of connections (friendships, acquaintances, etc.) between these individuals (edges).

Node. In a social network, a person corresponds to a particular node in the network.

Edge. Edges connect nodes in a social network and represent the relationships between the individuals represented by the nodes. Relationships (i.e., friendship) can be undirected, such that an edge from node A to node B implies an edge from B to A, or directed. The social networks we consider are assumed to have undirected edges.

Path. Sequence of consecutive edges connecting two nodes in a network. For example, if we have a network of three individuals A,B,C with edges between A and B and B and C, then a path exists between A and C.

Neighborhood/Neighbors. For any node *ego* in a network, the nodes directly connected to ego via an edge define ego's neighborhood. Nodes connected by an edge are considered neighbors.

Degree. For any node *ego* in a network, ego's degree is the number of nodes in ego's neighborhood.

Degree Distribution. For a collection of nodes in the graph, usually the entire population or a sample, the degree distribution is a probability distribution $P(k)$ where $P(k)$ is the fraction of nodes in the collection with degree equal to k. Degree distributions are necessarily discrete, but it is common to approximate the shape of $P(k)$ with a continuous probability distribution. Common models of degree distribution include the uniform distribution, in which every node has the same degree, the power-law degree distribution, and the Poisson degree distribution.

Table 1. Notation

Symbol	Meaning
r	An individual recruit in an RDS sample, or participant in RDS study
r_i	The ith recruit in an RDS sample, or participant in RDS study, in order of recruitment
R_j	jth wave of recruiters for an RDS study. If $j = 0$, this is the set of "seed nodes" for the RDS study
R	The full collection of recruiters and recruits in an RDS sample, ordered by recruitment
NR	The number of recruiters in an RDS study
P	The larger population from which the RDS sample is drawn
N	The total number of potential recruits in an RDS study
G	The graph of potential recruits in an RDS study
$d(i)$	Degree of node i (representing recruit r_i) in G
$\mu_d(A)$	Mean degree of a set A of nodes (recruits) in G
L	The node-level variable of interest in an RDS study, for example HIV Status
$L(r)$	Value of L for recruit r, for example, "HIV-positive"
S	The set of all distinct values of L present in R, for example, {"HIV-positive", "HIV-negative", "unknown"}
s_k	s is an enumeration of S and k is an index variable for the elements of s. $k \in \{1, \ldots, \lvert S\rvert\}$
$p(s_c, s_d)$	Transition probability between two states of the Markov Process model of RDS, where individual states correspond to elements of S
$\pi(s_k)$	The proportion of recruits in R with value of $L = s_k$, a.k.a. the sample proportion of recruits with value $L = s_k$
σ	The proportion of the population in the sample. $\sigma = \lvert\{r: r \in R\}\rvert / N$
$\lfloor \sigma N \rfloor$	Sample size for a particular RDS sample
ω	Repeated sampling event where participant r_i in an RDS study attempts to recruit participant r_j but discovers that r_j has already been recruited
ρ_ω	Density of repeated sampling events in a particular RDS study or simulation

Poisson Degree Distribution. A degree distribution where the fraction of nodes with degree k follows the Poisson probability distribution, that is $(k) = \frac{\lambda^k e^{-\lambda}}{k!}$, where λ is equal to the mean node degree.

Power-Law Degree Distribution. A degree distribution where the fraction of nodes with a degree k follows a Power Law; $(k) \sim k^{-\gamma}$, where γ is a shape parameter.

Homophily (group-level attribute). The homophily index is the mathematical value capturing the extent to which individuals in a particular group are connected to nodes within the group rather than nodes in other groups. A homophily value of zero means the proportion of ties between members of a group is consistent with the proportion of within-group ties that would be expected if ties were formed at random. A positive homophily value indicates the presence of an in-group affiliation bias and a value of 1 indicates that the group is entirely isolated from other groups—all ties from the group are to other members of the group.

Tree. Named for the characteristic branching shape, a tree is a type of network where any two nodes are connected by exactly one path (ignoring the directionality of the edges if the edges are directed).

RDS Terms:

Recall that respondent-driven sampling starts with a set of seed nodes selected by some nonrandom process. Each seed node selects k random nodes from its neighborhood to generate a collection of new participants in a process called *recruitment*. In recruitment, the nodes selecting new participants are referred to as *recruiters* and the selected nodes are referred to as *recruits*. The seeds constitute sample wave 0 and the collection of nodes selected by the seeds constitutes sample wave 1. The nodes in wave 1 recruit wave 2 and so on with each new wave of recruits serving as recruiters for the next wave. Except for the seeds and the final wave of recruits (who do not recruit in turn), each participant serves as both a recruiter and as a recruit.

Seed. A member of the initial wave of recruiters in an RDS sample.

Recruit. A member of the second through last wave of recruiters in an RDS sample.

Recruitment Event. RDS is a type of chain-referral sample where individuals already in the sample attempt to recruit new participants from their own neighborhoods. We call each of these recruitment attempts a recruitment event. The recruitment event may succeed or fail. A node in the sample generates a recruitment event each time it tries to recruit a neighbor, so a pair of nodes may generate duplicate recruitment events when sampling with replacement. A successful recruitment by a node in sample wave w results in a new member in sample wave $w + 1$. Respondent-driven sampling only captures successful recruitment events.

Sample Chain. A sequence of edges and nodes that reflects a series of recruitments. Every recruit in an RDS sample can be traced back through a series of recruiter/recruit relationships to a seed node. When sampling is without replacement, the chain is unique and the set of nodes and edges from all the chains that begin from a particular seed define a directed tree graph with the seed as the root node.

Group. The set of individuals with the same value for a particular categorical variable (L). Membership in a group does not imply relationships to other members of the same group and ties between nodes do not imply membership in the same group. The population can be partitioned into groups based on a variable of interest (L) with the number of groups equal to the number of unique levels of (L). If the variable of interest were race, then the groups would correspond to unique race groups. If the variable of interest were HIV status, then the groups might correspond to "Positive", "Negative" and "Don't Know." The point of RDS analysis is to produce an estimate of the relative size of each group in the population.

Recruitment relationship. With the exception of the sample seeds, each individual recruited into the sample is recruited by someone else already in the sample. Each successful recruitment event generates a recruitment relationship in the sample between the recruiter

and the recruit. For a given individual-level categorical attribute (L), the recruitment relationships in the sample can be labeled by the value of L for the recruiter and the recruit. Recall that each individual in the population has an attribute corresponding to a value for categorical variable L so each sample member $r \in R$ can be labeled by L(r). For any particular categorical variable L, recruitment events in the sample can be labeled by the tuple (L(recruiter), L(recruit)). When L is a binary variable with levels 0 and 1, then there are four possible labels for a recruiter $\rightarrow$ recruit relationship: (0,0), (0,1), (1,0), (1,1).

Markov Process. A stochastic process whose next state depends only on its present state. RDS recruitment is a Markov Process because the future state of the process (future recruits) is dependent only on the current state (current recruits) via the recruitment mechanism.

Irreducible Markov Process. A Markov Process for which it is possible to get to any state from any state. RDS Recruitment is not an irreducible Markov Process because individuals cannot be recruited more than once, so once the process has gotten to a particular state (a particular set of recruits) it cannot return to that state.

Notation:

We use $|X|$ to indicate the number of elements in collection X. For collections that may contain duplicates, the size of the collection is the number of elements in the collection including duplicate entries.

We use set builder notation to describe sets, as in $S = \{L(r)\colon r \in R\}$, by which we mean S is the set of unique values produced by applying the function $L(r)$ to each element in the collection R. We also need to construct collections that may contain duplicate elements. For these collections, we adopt an angle bracket notation as in $R_{\text{values}} = \langle L(r)\colon r \in R\rangle$ by which we mean the collection R_{values} is the sequence $(L(r_i))_{i=1}^{|R|}$ where r is an enumeration of the collection R in order of recruitment. Though S and R_{values} have the same number of distinct values, the number of elements in R_{values} is equal to the number of elements in R, while the number of elements in S is the number of distinct values of the categorical variable L. Stated more simply, S is a collection of the unique values in R_{values}.

The formal foundation for RDS as a Markov Process has been well explored in previous articles, such as Heckathorn (1997) and Volz and Heckathorn (2008). Previous work (Heckathorn 1997; Volz and Heckathorn 2008) showed that RDS can be modeled as a Markov Process. We do not repeat the analysis here for space reasons, but summarize it.

Assume that we wish to estimate the population composition in terms of some individual-level categorical variable L. The first step is to gather an RDS sample. Consider an arbitrarily selected seed set R_0 of initial recruiters for an RDS study. These recruiters are embedded in a larger population P, members of which are connected in a social network. For simplicity, we posit that recruitment into the RDS study can occur through any network connection—that is, if two individuals are network neighbors, it is possible for one to recruit the other. It is possible to define possible recruitment paths differently, but the choice of definition does not affect our formal analysis, so we use this simple one. Each member of collection R_0 then recruits a random subcollection of its neighbors.

The elements of these subcollections form the collection R_1 of recruits (which is also a subcollection of the network neighbors of R_0). The attributes of the individuals in R_1 are measured, including node degree, the value for variable L. The process then repeats, with R_1 recruiting a subcollection of the neighbors of R_1 to generate the collection R_2 and R_j generating the collection R_{j+1} until the RDS sampling process is stopped. At the end of the study, the full collection of recruiters and recruits R is an object of analysis.

The RDS sample is a collection of nodes R, the measured attributes of nodes in R and the recruitment relationships between recruiter and recruit. In order to estimate the transition probabilities for the underlying Markov Process, we need to construct a list of all the successful recruitment events captured by the sample. For each sample wave $j > 0$ each recruit $r_i \in R_j$ can be paired with at least one recruiter $r_h \in R_{j-1}$ who recruited r_i into the RDS sample, generating a recruitment event. When sampling without replacement, there is exactly one recruiter for each recruit. When sampling with replacement, a particular individual may appear multiple times within a single wave and may also appear in later waves. This is not a problem because there is exactly one successful recruitment event for each appearance of an individual in the sample. The collection of recruitment pairings between R_j and R_{j-1} for all waves j where $j > 0$ constitutes the list of recruitment events in the sample.

Each individual in the population has some value for the categorical attribute L so each sample member $r \in R$ can be labeled by $L(\mathrm{r})$. The recruitment events in the sample can also be labeled by the sequence $\langle L(\text{recruiter}), L(\text{recruit})\rangle$.

Consider an RDS sampling process that starts from a single seed and where each recruiter generates at most one recruitment event. In this case, each sample wave $j : j > 0$ is generated by one recruitment event; R_j has exactly one element r_i and for any $r \in R_j : j > 0$, we can observe the node label of the ith recruit $L(r_i)$ and the label of the kth recruitment event $\langle L(r_{k-1}), L(r_k)\rangle$. The sequence of node labels produced by the sampling process can be modeled as a Markov Process on the distinct values of L.

Let S be the set of distinct values of L so $S = \{L(r)\colon\ r \in R\}$. We can model the recruitment process as a Markov Process, where the states $s_k \in S$, $k = 1, 2 \ldots |S|$ correspond to distinct values of L and the transition from states s_c to s_d represent a recruitment event with label $\langle s_c, s_d\rangle$. Thus the underlying Markov Process that models *RDS* is between different values of an individual variable L, not between individual nodes. The work in Volz and Heckathorn (2008) models RDS as a Markov Process between individual nodes; however, the result of their analysis is identical to the Heckathorn (1997) analysis: that is, so long as the underlying Markov Process is irreducible, a stationary equilibrium exists where the state of the Markov Process (the current node) is independent of the starting state (the seed node). At this point, the steady state distribution of the Markov Process modeling the RDS recruitment process is an unbiased estimator of the population. Similarly, Volz and Heckathorn's analysis requires the assumption of sampling with replacement to be met for the irreducibility condition to be satisfied. Accordingly, the analysis in the rest of the article would be substantively the same were we to choose Volz and Heckathorn's (2008) model instead of the Heckathorn (1997) model. We focus on the Heckathorn (1997) model because it is much more straightforward to formulate bias due to sampling without replacement in terms of transitions between groups than it is to formulate the same in terms of transitions between nodes. More discussion follows in the Appendix.

We can also conceive of this process as occurring between groups *A, B,*. . . where an individual group contains all recruits in *R* who have the variable value s_c. In either case, previous work on RDS shows that as long as the Markov Process is irreducible, a condition that holds if a number of assumptions, including sampling with replacement, are satisfied, it will reach equilibrium. After the Markov Process reaches equilibrium, the state of the system is independent of the starting state, and recruits sampled after equilibrium will be independent of the seeds. A Markov chain is irreducible if it is possible to move from every state to every other state in a finite number of steps—that is, there can be no groups or sets of groups with homophily of 1. When these conditions hold, the mix of recruits will be independent of the seeds. In this case, if all individuals have equal degree, the RDS sample is representative of the underlying population.

The Markov Process model of RDS is also critical for estimating population composition by the levels of *L*. Heckathorn (1997) shows that it is possible to use the transition probabilities $p(s_c, s_d)$ between states in the Markov Process to construct a system of simultaneous equations that will yield the sample proportions $\pi(s_c)$ of recruits with key variable value equal to s_c. Furthermore, as the same work showed, one can estimate the transition probabilities $p(s_c, s_d)$ using the frequency of individuals in the collection $\langle r\text{: } L(r) = s_c\rangle$ recruiting individuals in the collection $\langle r\text{: } L(r) = s_d\rangle$ as follows:

$$\hat{p}_{RDS}(s_c, s_d) = \frac{\left|Rec\left(\langle r\text{: } L(r) = s_c\rangle, \langle r\text{: } L(r) = s_d\rangle\right)\right|}{\left|Rec\left(\langle r\text{: } L(r) = s_c\rangle, R\right)\right|} \tag{1}$$

where *Rec*(*A,B*) is the collection of all recruitment events where an individual in *A* recruits another individual in *B*. Later, Salganik and Heckathorn (2004) showed that if sampling occurs with replacement (any individual can be recruited any number of times), the RDS estimates of the Markov Process transition probabilities are unbiased, so as sample size increases, $\hat{p}_{RDS}(s_c, s_d)$ approaches the equilibrium transition probabilities $p_{MC}(s_c, s_d)$ of the underlying Markov Process. When R is the result of a Markov Process, $\hat{p}_{RDS}(s_c, s_d)$ is exactly the same as the maximum-likelihood estimate for the transition probabilities $\hat{p}_{MC}(s_c, s_d)$.

$$\hat{p}_{MC}(s_c, s_d) = \frac{\left|D\left(\langle r\text{: } L(r) = s_c\rangle, \langle r\text{: } L(r) = s_d\rangle\right)\right|}{\left|D\left(\langle r\text{: } L(r) = s_c\rangle, R\right)\right|} \tag{2}$$

where *D*(*A,B*) is the set of network connections between individuals in collection *A* and individuals in collection *B*.

However, if sampling occurs without replacement, the Markov Process model of RDS must be called into question. Since every individual may be recruited at most once, and there are a finite number of individuals, the underlying Markov Process is no longer irreducible, and thus does not have a stationary equilibrium distribution.

We cannot use the reducible Markov Process for sampling without replacement to calculate transition probabilities $p(s_c, s_d)$ and sample proportions $\pi(s_c)$. However, we can still use the irreducible Markov Process for sampling with replacement, and calculate transition probabilities and sample proportions for that process, as long as RDS chains are sufficiently similar to those that would be produced under sampling with replacement. To the extent that there is a difference between actual RDS chains and chain-referral samples

with replacement, $\hat{p}_{RDS}(s_c, s_d)$ will not be an unbiased estimate of the true transition probabilities $p(s_c, s_d)$.

We note that $\hat{p}_{MC}(s_c, s_d)$ is the transition probability that any (not a particular) individual with key variable value s_c recruits any other (not a particular) individual with key variable value s_d. Therefore, it is a measure of transitions between groups of individuals and depends on the number of edges between these groups—in this sense, $\hat{p}_{MC}(s_c, s_d)$ depends on the network. When estimating $\hat{p}_{MC}(s_c, s_d)$ in the course of RDS analysis, researchers typically do not have access to the underlying network structure, so they estimate it via $\hat{p}_{RDS}(s_c, s_d)$, which is calculated based on the number of recruitments by individuals with key variable value s_c of individuals with key variable value s_d.

3. Bias in Sampling Without Replacement

We have shown that the underlying cause of bias due to sampling without replacement in RDS studies is the difference between RDS chains and those that would be created under chain-referral sampling with replacement. The next question is the magnitude and direction of that bias.

Let us begin by making an observation: sampling without replacement produces a bias in transition probability estimates when a participant in an RDS study with value $L(s_c)$ attempts to recruit another individual, say with value $L(s_d)$, but finds that individual has already been recruited. Bias occurs in this case because the equilibrium transition probabilities are based on the number of network connections between individuals with $L(s_c)$ and $L(s_d)$, but the recruitment failure prevents one of those connections from being included in the estimate counts. We make this observation formal by defining a *repeated sampling event*:

Definition 3.1 *A repeated sampling event ω is an event where participant r_i in an RDS study attempts to recruit participant r_j but discovers that r_j has already been recruited.*

Equivalently, we can think of a repeated sampling event as introducing a difference between an RDS chain and a with-replacement chain-referral sample on the sample population, with the same seeds. Now we can describe the bias due to sampling without replacement in terms of a frequency of discrete events.

Before we proceed with the rest of the analysis, it is important to point out that for the vast majority of RDS studies, we cannot measure the frequency of repeated sampling events directly, since most RDS studies do not ask recruiters how many peers they attempted to recruit into the study, nor how many of those peers refused because they had already participated. However, we can make general observations about the frequency of repeated sampling events in RDS, and, based on these observations, demonstrate analytically the dynamics of this frequency for different values of sampling (fraction, homophily, and so on).

We begin by observing that the occurrence of repeated sampling events is determined by exactly three factors:

- Network structure
- Sampling fraction
- Probability of any RDS recruitment chain following a particular edge in the network

For example, consider an undirected tree network with the RDS seed as the root. Then, regardless of sampling fraction or the probability of following any particular edge in the network, the only possible repeated sampling events are those where a recruit directly attempts to recruit her recruiter. Assuming these "backtracking" events do not occur (as we do below), no repeated sampling events occur, and the bias from sampling without replacement is zero. Conversely, consider a population where for some reason every recruit will attempt to recruit her recruiter. Then, regardless of (nonzero) sampling fraction and network structure, there will be some repeated sampling events, and the bias from sampling without replacement is nonzero.

Neither of these scenarios is likely to occur in an empirical RDS study; however, they are useful for two reasons. First, they provide us with theoretical bounds for the space of repeated sampling events. Second, they do resemble some empirical RDS scenarios. For example, networks of novice drug users in NY have been shown to resemble a star shape, with several recreational users connected to no one but a central active supplier (Wallace 1991).

In the following analysis, we will investigate first the density of repeated sampling events, and then the effect of this density on bias. We first investigate the effect of sampling fraction and network structure on the occurrence of these events, and then move on to the last factor, the probability of RDS recruitment along particular edges in the network.

Density of Repeated Sampling Events

The key factor in measuring and accounting for bias in sampling without replacement is the density of repeated sampling events, which we will call ρ_ω. We begin with a definition of ρ_ω:

Definition 3.2 *The density ρ_ω of repeated sampling events in a particular RDS study or simulation is the frequency of repeated sampling events divided by the total number of recruitment events in the study or simulation.*

A formal analysis of this quantity yields surprising observations about the correspondence between ρ_ω and particular network structures. We can use these observations to infer backwards from our understanding of network structure in hidden populations to expected levels of bias due to repeated sampling events. We begin this analysis by modifying the original assumptions about the RDS process outlined in Section 1, so as to incorporate sampling without replacement. Below, we present the new set of assumptions about RDS necessary for our analysis:

Assumption 3.3 The target population's network must be dense enough for the population to form a single component, so every node is reachable from every other node.

Assumption 3.4 Recruiters know one another, as acquaintances, friends, or those closer than friends, so their relationships are reciprocal and recruitment can occur in either direction along the tie.

Assumption 3.5 Respondents recruit as though they are selecting randomly from their neighborhoods.

Assumption 3.6 Recruits cannot attempt to recruit their recruiters. In other words, the random walk that generates the sample does not backtrack.

Assumption 3.7 The RDS recruitment process is asynchronous, that is, at no point is a potential recruit simultaneously approached by two or more recruiters.

Assumption 3.8 The RDS process begins with one seed.

Assumption 3.9 Respondents attempt to recruit a constant number k of their neighbors, or all of their neighbors, whichever is smaller. This number includes failed attempts to recruit due to repeated sampling events. *Attempt* here means that a recruiter will try to recruit some individual unless she has already been recruited, in which case the recruiter tries to recruit another individual in their network neighborhood and so on, until the recruiter has tried to recruit k individuals. The inclusion of failed attempts in k may mean that the RDS process stops when no recruiter has a legal recruit. In formal analysis, we are not concerned with the termination of specific RDS chains unless it happens deterministically, which is not the case for sampling fractions $< 100\%$. In simulations, chain termination is a concern, which we address by introducing additional seeds, one at a time (see the simulation section below).

Assumptions 3.3 and 3.4 pertain to the structure of the graph and are therefore scope conditions.

Assumptions 3.5, 3.6 and 3.7 pertain to the nature of the recruitment process, specifying a nonbacktracking random walk. Lee et al. (2012) showed that a nonbacktracking random walk retains the Markov property and will have the same stationary distribution as a simple random walk. Even though backtracking would create a repeated sampling event, Lee et al. show that eliminating backtracking does not change or bias the estimation of transition probabilities or the stationary distribution. We focus only on repeated sampling events that might produce bias estimates by excluding backtracking from our analysis.

Assumptions 3.8 and 3.9 are less reflective of empirical RDS studies, and we relax them in a simulation framework in Section 4.

In the following analysis, we focus on the graph of *potential recruits G,* in contrast to the recruitment graph of relationships between *actual* recruits. G is meant to represent the wider community, from which the seeds and the recruits are drawn. For example, in a study of jazz musicians in New York City, G is the graph of all jazz musicians in New York City. A graph consists of nodes and edges; in this example, the nodes are the jazz musicians in New York City and the edges are connections (friendships, professional relationships, etc.) between jazz musicians, along which recruitment may occur. Since the target sampling fraction, σ, is a rational number and the sample size has to be an integer (number of individuals), we define sample size as the greatest integer less than or equal to the sample fraction multiplied by the population size or $\lfloor \sigma N \rfloor$.

We now restate our observation about tree structures and the absence of repeated sampling events as a formal lemma:

Lemma 3.10 *Given assumptions 3.3–3.9, sampling without replacement cannot occur only in populations where the structure of relationships among members of the population is an undirected tree.*

Proof: Consider graph G, which is not an undirected tree graph. An undirected tree is a type of network where the edges are undirected and any two nodes are connected by exactly one path. If G is not an undirected tree, then there is at least one cycle in G that is a sequence of connected nodes (seed and recruits) $r_1 \ldots r_l$ where l is the length of the cycle and r_l has an edge to r_1. Then it is possible for r_1 to be a seed, and recruit r_2, who recruits r_3 and so on until r_l is recruited. Then it is possible for r_l to attempt to recruit r_1 at which point a repeated sampling event will occur.

Similarly, consider a graph G' that is an undirected tree graph. Consider some seed r_1. Then for any potential recruiter r consider the set of potential recruits PS. A repeated sampling event can occur only if some $p \in PS$ has already been recruited. But that means that a path exists from a seed r_1 to p that does not go through r. Since a path already exists from r_1 to p that does go through r, this means that a cycle must exist in G', which contradicts the claim that G' is an undirected tree. □

Lemma 3.10 shows that specific network structures imply particular levels of bias due to sampling without replacement. However, this does not mean there is a deterministic relationship between network structure and level of bias due to sampling without replacement, as we show with the following negative result:

Lemma 3.11 *Given assumptions 3.3–3.9, and a particular chain of recruitments, it is possible that this chain could have arisen without any repeated sampling events regardless of the underlying structure of relationships between the recruits.*

Proof: Given any connected graph G of potential recruits, we can remove edges from G until no cycles exist but all the nodes are still connected. This is the minimum spanning tree of G. Let the number of coupons for an RDS study on this network be greater than the degree of any node in the minimum spanning tree. Under these conditions, an RDS process can start at any node in G and end by recruiting all potential recruits avoiding any repeated sampling events simply by following the minimum spanning tree. □

However, Lemma 3.11 does not preclude estimation of bias due to sampling without replacement from network structure. We may not be able to calculate an exact amount of bias for a particular network structure analytically, but we can nevertheless define bounds for this type of bias and formalize its relationship to key variables such as the sampling fraction. In particular, we outline and then prove a number of theorems about the relationship between network structure and density ρ_ω of repeated sampling events. We begin with a theorem for making formal statements about the density of repeated sampling events for any network structure. This theorem will serve as a framework for proving statements about specific network structures.

Theorem 3.12 *Given assumptions 3.3–3.9, a graph G with N nodes and one or more cycles, a further assumption that each recruiter attempts to recruit exactly k neighbors, and an RDS process with sampling fraction σ, the density ρ_ω of repeated sampling events is:*

$$\rho_\omega = \frac{\sum_{i=1}^{NR} f(r_i)}{NR} \tag{3}$$

where $f(r_i)$ is a function for recruiter r expressing the fraction of her k recruits that have already been recruited, and NR is the number of recruiters in the RDS sample.

Proof: The density of repeated sampling events is the ratio of repeated sampling events (RSE) to the total number of recruitment events (RE). Symbolically, let us represent it as:

$$\rho_\omega = \frac{RSE}{RE} \tag{4}$$

As per the statement of the theorem, we make a further assumption that recruiters make exactly k recruitment attempts—in other words, that Assumption 3.9 holds and furthermore every individual has degree at least k. This assumption greatly simplifies the analysis, and we relax it in the simulation section. With this assumption, we can rewrite ρ_ω as:

$$\rho_\omega = \frac{RSE}{kNR} \tag{5}$$

where NR is the number of recruiters in the RDS sample. For the numerator, let $f(r_i)$ be a function for recruiter r_i expressing the fraction of her k recruits that have already been recruited. Then the numerator is:

$$RSE = \sum_{i=1}^{NR} kf(r_i) \tag{6}$$

Substituting in RSE, taking the constant k out of the sum and canceling, we get:

$$\rho_\omega = \frac{\sum_{i=1}^{NR} f(r_i)}{NR} \tag{7}$$

□

Having shown a general relationship between ρ_ω and sampling fraction σ for some graph G, we proceed to show specific instances of this relationship on Poisson Random Graphs, Small-World Graphs and Preferential Attachment Graphs.

A Poisson Random Graph is a graph where all ties are randomly targeted and the nodes have a Poisson degree distribution. We use the Erdős-Rényi version of a Poisson Random Graph (Erdős and Rényi 1959).

A Small-World Graph is a graph whose nodes are embedded in a regular lattice, but a fraction of the edges between these nodes are randomly rewired, creating enough shortcuts in the graph to lead to a small graph diameter (a small world). Nodes in Small-World Graphs have a regular degree distribution, that is, all nodes have identical degree. Watts and Strogatz (1998) describe the construction and properties of Small-World Graphs.

A Preferential Attachment Graph is a graph where nodes connect to others preferentially based on their degree. Nodes in Preferential Attachment Graphs have a power-law degree distribution. Preferential Attachment Graphs are described in Barabasi and Albert (1999).

Theorem 3.13 *Given 3.3–3.9, a Poisson Random Graph G with N nodes and one or more cycles, and an RDS process with sampling fraction σ, the density ρ_ω is bounded by the following inequality:*

$$\frac{\sigma}{2(k+1)} - \frac{1}{2N} \leq \rho_\omega \leq \frac{\sigma}{2} - \frac{1}{2N} \tag{8}$$

Proof: For a Poisson Random Graph, all ties are randomly targeted, so the probability of a tie targeting an already-recruited node is given by ν/N where ν is the current number of recruits. For recruiter r_i, $\nu = i - 1$, so we can rewrite Equation 3 as follows:

$$\rho_\omega = \frac{\sum_{i=1}^{NR} \frac{i-1}{N}}{NR} \tag{9}$$

or

$$\rho_\omega = \frac{(NR-1)}{2N} \tag{10}$$

Now we have the equation strictly in terms of NR the number of recruiters and N the population size. We can bound the number of recruiters by the following argument: In the simulation design of RDS, and also in RDS empirical studies, if an individual fails to recruit k recruits, a new recruiter is added. As we discuss above, we can assume for the purposes of this section that every individual has at least degree k, so the only way a recruiter fails to recruit k recruits is through a repeated sampling event, when the recruiter tries to recruit some individual who has already been recruited. So the minimum number of recruiters occurs when no repeated sampling events occur. In this case, every recruiter recruits exactly k recruits. We know the total number of participants in the sample (recruiters + recruits) is $\lfloor \sigma N \rfloor$, so we can derive the lower bound for the number of recruiters by solving:

$$NR + kNR \geq \lfloor \sigma N \rfloor \tag{11}$$

or

$$NR \geq \frac{\lfloor \sigma N \rfloor}{k+1} \tag{12}$$

Now let us consider the maximum number of recruiters. This occurs when all sampling events are repeated sampling events; when all current recruiter attempts lead to repeated sampling events in an empirical RDS study, a new recruiter is added to the sample. Thus, in this case, a new recruiter is added to the sample after every recruiter makes all of their recruitment attempts. This extremely rare situation would occur if all the initial seeds in an RDS study tried to recruit each other and only each other, and every subsequently added recruiter tried to recruit only from among the seeds. In this case, the number of recruiters is

just the sample size, so the other side of the inequality is:

$$NR \leq \lfloor \sigma N \rfloor \tag{13}$$

Using this inequality, we can put bounds on ρ_ω as follows:

$$\frac{\frac{\lfloor \sigma N \rfloor}{k+1} - 1}{2N} \leq \rho_\omega \leq \frac{\lfloor \sigma N \rfloor - 1}{2N} \tag{14}$$

or

$$\frac{\sigma}{2(k+1)} - \frac{1}{2N} \leq \rho_\omega \leq \frac{\sigma}{2} - \frac{1}{2N} \tag{15}$$

□

Equation 14 shows that ρ_ω increases linearly in the sampling fraction (all other terms are constant for a given RDS sample). Note that for large populations, $\frac{1}{2N}$ is negligible, so the bounds on ρ_ω are:

$$\frac{\sigma}{2(k+1)} \leq \rho_\omega \leq \frac{\sigma}{2} \tag{16}$$

Theorem 3.14 *Given assumptions 3.3–3.9, a Small-World Graph G with N nodes and one or more cycles and rewiring probability p, and an RDS process with sampling fraction σ, the density ρ_ω of repeated sampling events is bounded by the following inequality:*

$$\begin{aligned} &\frac{p(1+1-c_1-p+c_1p)}{2(k+1)}\sigma + \frac{p(1+1-c_1-p+c_1p)}{2N} - p\frac{1}{N} + c_1 - pc_1 \\ &\leq \rho_\omega \leq \\ &\frac{p(1+1-c_1-p+c_1p)}{2}\sigma + \frac{p(1+1-c_1-p+c_1p)}{2N} - p\frac{1}{N} + c_1 - pc_1 \end{aligned} \tag{17}$$

Proof: For a Small-World Graph with rewiring probability p, a fraction p of all ties are randomly targeted, while the rest are embedded within a regular lattice. Given some recruiter r_i making k recruitment attempts, kp of those attempts will be reaching random targets in the network, while $k(1-p)$ of those attempts will be reaching lattice neighbors. Accordingly, $f(r_i)$ will be an interpolation between p and $1-p$.

First, let us examine what happens in the case of lattice neighbors. Some fraction c of these will already have been recruited, in one of two ways: either they were recruited by their own lattice neighbors, or they were recruited through random ties. Let us call the fraction of neighbors recruited by their own lattice neighbors c_1, and the fraction recruited through random ties c_2.

The quantity c_1 is independent of sampling fraction. To see why, consider the example of a ring lattice. Since we are only looking at individuals recruited by lattice neighbors, the recruitment set on this network will resemble a line that grows at both ends. Each new recruit r_i appears at the end of the line and always has the same neighborhood composition

with respect to recruited versus nonrecruited individuals: half are already recruited (the half of r_i 's neighbors that are closer to the seed) whereas the other half are not already recruited (the exception being when all individuals are recruited and the ends of the line connect, but that lone case will not affect our estimations).

The quantity c_2 is the probability that some lattice neighbor j of r_i has already been recruited by another node k via a randomly rewired tie. For each such j, there are approximately i potential recruiters, and each recruiter can recruit j if it has a rewired tie (probability p) and it points to j (since rewired ties are random, the probability is uniform at $1/N$).

To calculate the quantity c, let us consider the processes that generate c_1 and c_2 as probabilistic events C_1 and C_2. In the equation below, $||$ is the logical OR notation. Then:

$$\mathrm{c} = (1-p)\big(C_1||C_2\big) \tag{18}$$

$$= (1-p)(1-(1-P(C_1))(1-P(C_2))) \tag{19}$$

$$= (1-p)\Big(1-(1-c_1)\big(1-p/N\big)^i\Big) \tag{20}$$

$$\approx (1-p)\big(1-(1-c_1)\big(1-pi/N\big)\big) \tag{21}$$

Note that Equation 21 is an approximation, based on a derivation by Tillé (2006). This approximation holds for $pi \ll N$, meaning that as long as $pi \ll N$, the left-hand side is almost exactly equal to the right-hand side. The quantity i is bounded by the number of recruiters, NR. Thus, $pi \ll N$ so long as:

$$pNR \ll N \tag{22}$$

The quantity NR itself is bounded by the sampling fraction σ, such that $NR \leq \lfloor \sigma N \rfloor$. Accordingly, the inequality holds as long as:

$$p\lfloor \sigma N \rfloor \ll N \tag{23}$$

or,

$$p\sigma \ll 1 \tag{24}$$

In the simulation section of the article we use $p = 0.2$, $0 < \sigma < 1$ so $p\sigma$ ranges between 0 and 0.2, which is significantly smaller than 1.

Next, consider the pk attempts that reach network neighbors through rewired ties. These ties are rewired at random, so the targets of those ties will be random nodes in the network. As in Theorem 3.13, the probability that any attempt reaches a node that has already been recruited is $(i-1)/N$ for the ith recruit. Now we are finally ready to write down ρ_ω.

$$\rho_\omega = \frac{\sum_{i=1}^{NR}\left(p\dfrac{i-1}{N}+(1-p)\left(1-(1-c_1)\left(1-p\dfrac{i}{N}\right)\right)\right)}{NR} \tag{25}$$

or

$$\rho_\omega = \frac{\frac{p(2-c_1-p+c_1p)}{N}\sum_{i=1}^{NR}(i) - p\frac{NR}{N} + c_1NR - pc_1NR}{NR} \tag{26}$$

or

$$\rho_\omega = \frac{p(2-c_1-p+c_1p)}{N}\frac{(NR+1)}{2} - p\frac{1}{N} + c_1 - pc_1 \tag{27}$$

As we showed above, the number of recruiters is between $\frac{\lfloor \sigma N \rfloor}{k+1}$ and $\lfloor \sigma N \rfloor$, so we can write:

$$\begin{aligned} &\frac{p\big(2-c_1-p+c_1p\big)}{N}\frac{\left(\frac{\lfloor \sigma N \rfloor}{k+1}+1\right)}{2} - \frac{p}{N} + c_1 - pc_1 \\ &\leq \rho_\omega \leq \\ &\frac{p\big(2-c_1-p+c_1p\big)}{N}\frac{\big(\lfloor \sigma N \rfloor+1\big)}{2} - \frac{p}{N} + c_1 - pc_1 \end{aligned} \tag{28}$$

or

$$\begin{aligned} &\frac{p(2-c_1-p+c_1p)}{2(k+1)}\sigma + \frac{p(2-c_1-p+c_1p)}{2N} - \frac{p}{N} + c_1 - pc_1 \\ &\leq \rho_\omega \leq \\ &\frac{p(2-c_1-p+c_1p)}{2}\sigma + \frac{p(2-c_1-p+c_1p)}{2N} - \frac{p}{N} + c_1 - pc_1 \end{aligned} \tag{29}$$

□

This is a much more complex form than Equation 14, but again, all the terms except for σ are constants for a particular RDS sample, so again ρ_ω increases linearly in σ.

Furthermore, recall that $p < 1$ and $c_1 < 1$. So, $p(2 - c_1 - p + c_1p) < 2$ and for large populations:

$$\frac{p(2-c_1-p+c_1p)}{2N} \approx 0 \tag{30}$$

or

$$\frac{p}{2N} \approx 0 \tag{31}$$

Then, for large populations, the bounds are:

$$\frac{p(2-c_1-p+c_1p)}{2(k+1)}\sigma + c_1 - pc_1 \leq \rho_\omega \leq \frac{p(2-c_1-p+c_1p)}{2}\sigma + c_1 - pc_1 \tag{32}$$

Theorem 3.15 *Given Assumptions 3.3–3.9, a Preferential Attachment Graph G with N nodes, one or more cycles, a degree distribution approximated by $P(x) \approx x^{-\alpha}$ and rewiring probability p, and an RDS process with sampling fraction σ, the density ρ_ω of repeated sampling events is a nonlinear function of σ that is sublinear for small values of σ and approaches linearity in σ for larger values of the sampling fraction.*

Proof: In a Preferential Attachment Graph, ties are not targeted randomly, but according to the degree of the target, with higher-degree nodes more likely to be tie targets. For recruit r_i, $f(r_i)$ thus depends not only on the number of individuals already recruited, but also on their sum degree. Specifically:

$$f(r_i) = \frac{SRF(i)}{SN} \tag{33}$$

where $SRF(i)$ is the sum degree over all nodes already recruited whereas SN is the sum degree over all nodes in the graph. In the beginning of the recruitment process, the recruited sample is more likely to contain network hubs than low-degree nodes, as all nodes (including the seed) are preferentially more likely to have ties to higher-degree nodes than to lower-degree nodes. However, the number of such hubs is very small, so the recruited sample quickly exhausts them and moves on towards lower-degree nodes. As this happens, the average degree over all recruits approaches the average degree over all nodes in the graph. When the average degree over all recruits is approximately equal to the average degree over all N nodes:

$$\frac{SRF(i)}{i} \approx \frac{SN}{N} \tag{34}$$

we have:

$$f(r_i) = \frac{SRF(i)}{SN} \tag{35}$$

$$= \frac{\frac{SRF(i)i}{i}}{\frac{N\,SN}{N}} \tag{36}$$

$$= \frac{i}{N}\frac{\frac{SRF(i)}{i}}{\frac{SN}{N}} \tag{37}$$

$$\approx \frac{i}{N} \tag{38}$$

This is almost the same expression as $f(r_i)$ for a Poisson Random Graph, where ρ_ω is linear in σ. Accordingly, in the limit of large sampling fraction, ρ_ω approaches a linear function of σ. However, for small sampling fractions, the sample may never reach this stage. In that case, the average degree over all recruits is much bigger than the average degree over all N nodes:

$$\frac{SRF(i)}{i} \gg \frac{SN}{N} \tag{39}$$

In this case, $f(r_i)$ is much bigger than $f(r_i)$ for a Poisson Random Graph. Therefore, ρ_ω values grow more quickly than for a Poisson Random Graph for small sampling fractions, but then grow ever slower as sampling fraction increases, approaching a linear growth rate. Thus, the second derivative of ρ_ω is initially negative, and it grows sublinearly for small sampling fractions. □

To give some idea of the range of sampling fractions, over which ρ_ω grows sublinearly, we consider the probability of high-degree nodes being picked in the sample, which also gives us the expected point at which these high-degree nodes are exhausted. Let us focus on nodes with above-average degree—so long as these nodes are picked, the average degree over the recruit set remains higher than the average degree over all nodes. The probability $p(n > \mu_d)$ of any one tie targeting a node n with above-average degree is given by:

$$p(n > \mu_d) = \frac{\sum_j d(j) > \mu_d}{SN} \tag{40}$$

$$\approx \frac{\int_{\mu_d}^{M} xP(x)dx}{\int_{m}^{M} xP(x)dx} \tag{41}$$

where $d(j)$ is the degree of node j, μ_d is the mean degree, M the max degree and m the min degree of G, x is degree, and $P(x)$ is the degree distribution of G. The approximation in Equation 41 is a smoothing out of Equation 40, since the degree distribution of G ranges only over discrete values of x. As we note in the theorem statement, $P(x) \approx x^{-\alpha}$. Therefore, Equation 41 evaluates to:

$$p(n > \mu_d) \approx \frac{M^{2-\alpha} - {\mu_d}^{2-\alpha}}{M^{2-\alpha} - m^{2-\alpha}} \tag{42}$$

where α is the best-fit exponent of the degree distribution of G. For $\alpha > 2$, μ_d is well-defined and equal to:

$$\mu_d = m\frac{\alpha - 1}{\alpha - 2} \tag{43}$$

So we can rewrite above as:

$$p(n > \mu_d) \approx \frac{M^{2-\alpha} - \left(m\frac{\alpha - 1}{\alpha - 2}\right)^{2-\alpha}}{M^{2-\alpha} - m^{2-\alpha}} \tag{44}$$

Note that for $\alpha > 2$, $M^{-2\alpha}$ is very close to 0. We can use that to reapproximate $p(n > \mu_d)$ as:

$$p(n > \mu_d) \approx \left(\frac{\alpha - 1}{\alpha - 2}\right)^{2-\alpha} \tag{45}$$

This function decreases superlinearly in α, but between $\alpha = 2$ and $\alpha = 3$ (the range for Preferential Attachment Graphs), it varies between .8 and .5. Now consider the fraction of nodes that have above-average degree, $P(N > \mu_d)$, which is derived from the cumulative

degree distribution of *G,* which is the integral of the partial degree distribution of G, $P(x)$:

$$P(N > \mu_d) \approx \frac{\int_{\mu_d}^{M} P(x)dx}{\int_{m}^{M} P(x)dx} \approx \left(\frac{\mu_d}{m}\right)^{-\alpha+1} \tag{46}$$

$$= \left(\frac{\alpha - 1}{\alpha - 2}\right)^{1-\alpha} \tag{47}$$

$$= p(x > \mu_d) \cdot \frac{\alpha - 2}{\alpha - 1} \tag{48}$$

In other words, given exponent α of 2.1, about 80% of the ties will be targeting nodes with above-average degree, whereas only about 10% of the nodes will have above-average degree. This disparity suggests that a sublinear relationship will exist between ρ_ω and σ, given α of 2.1 and sampling fractions much lower than ten percent. We can establish similar relationships for other values of α and σ, but note that as α increases, fewer and fewer of the early ties will point to above-average degree nodes. A complex nonmonotonic relationship therefore exists between the power-law exponent and the relationship between sampling fraction and density of repeated sampling events.

Bias Due to Repeated Sampling Events

We formalize the relationship between the density of repeated sampling events ρ_ω and the bias due to sampling without replacement. This bias can be expressed as the difference between the equilibrium transition probabilities for a Markov Process modeling a chain-referral sample with replacement and the estimated transition probabilities between different groups in an empirical RDS sample:

$$Bias_{SWOR} = \sum_{s_i} \sum_{s_j} \left| p_{MC}(s_i, s_j) - \hat{p}_{RDS}(s_i, s_j) \right| \tag{49}$$

where s_i, s_j are different values of the key variable L analyzed in the course of an RDS study as described in Section 2, and $\hat{p}_{RDS}(s_i, s_j)$ and $p_{MC}(s_i, s_j)$ are defined in Equations 1 and 2, respectively. In the case where no repeated sampling events are possible (e.g., on an undirected tree as described in Lemma 3.10), this bias tends asymptotically to 0 in the sampling fraction σ:

$$\lim_{\sigma \to 1} (Bias_{SWOR}) = 0 \tag{50}$$

In the case where repeated sampling events are possible, each event initially introduces a small amount of bias. Recall that $Rec(A,B)$ is the collection of all recruitment events where an individual in A recruits another individual in B, $D(A,B)$ is the set of network connections between individuals in collection A and individuals in collection B and the definitions of $\hat{p}_{RDS}$ and $\hat{p}_{MC}$ given in Equations 1 and 2. Consider a single repeated sampling event in the sampling process where a recruiter r_1, $L(r_1) = s_1$ tries to recruit another individual r_2, $L(r_2) = s_2$, but r_2 has already been recruited.

In the limit of $\sigma \rightarrow 1$, the number of recruitments from s_1 nodes to s_2 approaches the number of edges between s_1 and s_2 minus the single failed recruitment.

$$\begin{aligned} &\left|Rec\left(\langle r_i : L(r_i) = s_1\rangle, \left\langle r_j : L(r_j) = s_2\right\rangle\right)\right| \\ &\quad \rightarrow \left|D\left(\langle r_i : L(r_i) = s_1\rangle, \left\langle r_j : L(r_j) = s_2\right\rangle\right)\right| - 1 \end{aligned} \tag{51}$$

At the same time, the number of recruitments from s_1 nodes to any other node approaches the number of edges between s_1 and all other nodes minus the single failed recruitment:

$$\left|Rec\left(\langle r_i : L(r_i) = s_1\rangle, R\right)\right| \rightarrow \left|D\left(\langle r_i : L(r_i) = s_1\rangle, R\right)\right| - 1 \tag{52}$$

Let $a = \left|D\left(\langle r_i : L(r_i) = s_1\rangle, \left\langle r_j : L(r_j) = s_2\right\rangle\right)\right|$ and $b = \left|D\left(\langle r_i : L(r_i) = s_1\rangle, R\right)\right|$ so

$$\hat{p}_{RDS}(s_i, s_j) \rightarrow \frac{a-1}{b-1} \tag{53}$$

With a and b as defined above, $\hat{p}_{MC}(s_i, s_j) = \frac{a}{b}$ and so $Bias_{SWOR} \rightarrow \frac{a}{b} - \frac{a-1}{b-1}$. In general, if the density of repeated sampling events ρ_ω is uniform across all nodes and each node makes the same number of recruitment attempts, then, in the limit of $\sigma \rightarrow 1$:

$$\hat{p}_{RDS}(s_i, s_j) \rightarrow p_{MC}(s_i, s_j)(1 - \rho_\omega) \tag{54}$$

for all s_i, s_j and so $Bias_{SWOR} \rightarrow \sum_{s_i} \sum_{s_j} \left|p_{MC}(s_i, s_j)(\rho_\omega)\right|$. Note that the assumption that each node makes the same number of recruitment attempts is not entirely unrealistic in empirical RDS studies, since the number of coupons per recruiter is usually capped at a small value. The other assumption, that repeated sampling event density is uniform across all nodes, is less realistic, but useful for generalized results across network structures. In the simulation section of this article, we relax the uniform density assumption.

Bias Due to Degree Differential and Group Size

The case explored in Theorem 3.15 suggests that the degree of recruits can play an important role in creating bias due to sampling without replacement. In this section, we consider a scenario wherein two groups of recruiters are present, one with a drastically higher degree than the other, and explore the implications for density ρ_ω of repeated sampling events. We also consider the effect of group size on bias, that is, the case where two groups are present in the target population, but one group has many more members than the other. In this section, we make a further simplifying assumption for RDS samples: when a target population is divided into two groups, all members of a group have the same degree. This is a strong assumption, but it helps illustrate the fundamental effect of degree differential and group size on bias. We relax this assumption in the simulation section below.

The networks we construct in this section have the property that the probability of an edge targeting a particular node $= td(i)$ where $t > 0$ is a constant and $d(i)$ is the target node's degree. This mode of network construction allows us to examine a simplified version of preferential attachment tie formation behavior found in many empirical networks. Given our earlier assumption that all nodes in one group have the same

degree, we cannot construct actual preferential attachment networks; again, we relax this constraint in the simulation section below, where we examine true preferential attachment networks.

First, we explore only the effect of degree differential. Consider a population of potential recruits that consists of two equal-sized groups A and B, embedded in a network as described above. Furthermore, we have:

$$\mu_d(A) = c\mu_d(B) \tag{55}$$

where $\mu_d(A)$ is the mean degree of group A and $\mu_d(B)$ is the mean degree of group B, and c is a constant. Given a uniform distribution of recruits between network neighbors as per Assumption 3.5, the probability of a recruit coming from group A is c times the probability of a recruit coming from group B. Then, given r recruits, under sampling with replacement, $rc/(c + 1)$ of them will come from group A and $r/(c + 1)$ of them will come from group B. However, the sample is collected without replacement, so the sampled proportions of recruits from group A and B will differ from the with-replacement condition.

We illustrate the difference in sampled proportions of recruits from A and B under conditions of sampling without replacement with the following toy scenario: consider a population of 100 individuals, 50 of which are in group A and the other 50 in B. The sampling fraction is 70 percent and the average degree of group A is six times the average degree of group B.

Under conditions of sampling with replacement, 70 individuals are recruited (some multiple times), 60 of those individuals are from group A and ten from group B. However, under conditions of sampling without replacement, all we know is there are 60 recruitment attempts targeted at group A and ten at group B. Some of these attempts may lead to repeated sampling events, where an individual in group A has already been recruited, and others do not. An estimate of the density of repeated sampling events ρ_ω in these scenarios depends on calculating the expected fraction of recruitment attempts that end up as repeated sampling events.

We begin with a simple observation: consider the situation that, under sampling without replacement, the first 50 recruitment attempts targeted at group A each target a distinct recruit. Then we know the last ten attempts targeted at group A automatically lead to a repeated sampling event. From this situation, we observe that at some point a group may become *exhausted,* after which point all recruitment attempts targeting that group automatically lead to repeated sampling events. After this point, the density ρ_ω of repeated sampling events becomes an interpolation between 1 (the rate for events that target group A) and whatever the rate was previously; this interpolation rapidly converges to 1 as sampling fraction increases. We present a formal proof of this observation below.

We now investigate the case when A is not exhausted. In this case, recruitment attempts targeted at A lead to a repeated sampling event with a probability that rises in the number of individuals already recruited from A. That probability is zero if no recruits have yet come from A and approaches unity as A approaches exhaustion.

Finally, consider the effect of group size, which is very simple: the smaller the size of A, the more quickly it approaches exhaustion. In other words, the smaller A is, the earlier

the onset of the exhausted regime, during which ρ_ω converges rapidly to one as sampling fraction increases.

We now combine these observations into a formal argument. We begin by proving a lemma that gives a formal expression for the expected number of distinct recruits from some group U given m recruitment attempts targeting U. This lemma is necessary to calculate the exhaustion point for groups, and relies on the assumption we make at the beginning of this section: all individuals in a particular group have the same degree. We then use the lemma to prove a "master equation" theorem that combines sampling fraction, degree differential and group size into one expression for ρ_ω.

Lemma 3.16 *Given a group U of size N that has not yet been exhausted, such that all individuals in the group have the same degree, the expected number of distinct recruits from U given m recruitment attempts targeting U is bounded by the inequality:*

$$m - \frac{m(m-1)}{2N} < DR < m \tag{56}$$

Proof: The quantity DR is equivalent to the number of distinct elements NDE after sampling m elements with replacement from a set of N elements with uniform selection probabilities, which is given in Tillé (2006):

$$NDE = N - \frac{(N-1)^m N!}{N^m (N-1)!} \tag{57}$$

or

$$NDE = N\left(1 - \left[\frac{N-1}{N}\right]^m\right) \tag{58}$$

Focusing on the exponential term in Expression 58, we have:

$$\left[\frac{N-1}{N}\right]^m = \left[1 - \frac{1}{N}\right]^m = \left[1 + \frac{-1}{N}\right]^m = \tag{59}$$

by binomial expansion:

$$\begin{aligned} &= \sum_{k=0}^{m} \binom{m}{k} \left[\frac{-1}{N}\right]^k = \binom{m}{0} 1 + \binom{m}{1} \frac{-1}{N} + \binom{m}{2} \frac{1}{N^2} \\ &\quad + \ldots + \binom{m}{m} \left[\frac{-1}{N}\right]^m \end{aligned} \tag{60}$$

This series has the property that, for any $k \leq m$, $m < N$ the $k+1$st element is smaller in magnitude and opposite in sign to the kth element. The sign opposition comes from the -1 in the power term of the series. The magnitude difference comes from the fact that the $k+1$st element is $O([m/N]^k)$, which decreases in k since $m < N$.

This property implies that the first few terms will dominate the series. In particular, we can establish bounds of the series with the second and third partial sums: $1 - m/N$ and $1 - \frac{m}{N} + \frac{m(m-1)}{2N^2}$. Every subsequent term will alternatively drive the series closer to

$1 - m/N$ and to $1 - \frac{m}{N} + \frac{m(m-1)}{2N^2}$, by an ever-decreasing degree, so the final sum will always stay within those bounds. Accordingly, we can approximate the inner term as follows:

$$1 - \frac{m}{N} + \frac{m(m-1)}{2N^2} > \left[\frac{N-1}{N}\right]^m > 1 - \frac{m}{N} \tag{61}$$

We can now rewrite Equation 56 as:

$$N\left(1 - 1 + \frac{m}{N} - \frac{m(m-1)}{2N^2}\right) < NDE < N\left(1 - \left[1 - \frac{m}{N}\right]\right) \tag{62}$$

or

$$m - \frac{m(m-1)}{2N} < NDE < m \tag{63}$$

□

What does Equation 63 tell us? Instead of targeting m distinct nodes, m recruitment attempts target some slightly smaller number $m-\epsilon$ nodes. In other words, $m-\epsilon$ ties target distinct nodes in the network, and the remaining ϵ ties are redundant, that is, lead to repeated sampling events.

Operationally, we can approximate *DR* by setting *NDE* to its lower bound (by the argument above, *NDE* will be much closer to its lower bound than to its upper bound):

$$\epsilon = \frac{m(m-1)}{2N^2} \tag{64}$$

Then:

$$DR = NDE \approx m - \frac{m(m-1)}{2N^2} \tag{65}$$

We now follow with a definition of group exhaustion and then the "master equation" theorem. In Theorem 3.18, we use big O notation written as ($f = O(g)$), where f and g are non-negative functions. This notation indicates that f is asymptotically upper bounded by g, in other words, that there exists an integer n_0 and a constant $c > 0$, such that for all integers $n > n_0$, $f(n) \leq cg(n)$. In this particular case, we claim that ρ_ω is asymptotically bounded by another expression, either σ if neither group is exhausted or an expression that tends asymptotically to 1 as $\sigma \rightarrow 1$ if one of the groups is exhausted. In the latter case, Theorem 3.18 shows that ρ_ω is always less than some function, and that function itself is always less than 1.0 but approaches it quickly as σ increases.

Definition 3.17 *A group of potential recruits is said to be exhausted if, in the course of an RDS recruitment process, every individual in that group is recruited.*

Theorem 3.18 *Given Assumptions 3.3–3.9, and a population of N potential recruits split into two groups A and B, with α individuals in A and β individuals in B, such that every individual in A has degree d(A) and every individual in B has degree d(B), and the probability of an edge targeting a particular node = kd(i) where $k > 0$ is a constant and d(i) the node's degree, at sampling fraction σ the frequency of repeated sampling events,*

ρ_ω, is approximated by:

$$\rho_\omega = O(\sigma) \textit{ if neither group is exhausted} \tag{66}$$

$$\rho_\omega = O\left(1 - \frac{(1-k)PER(g)}{TR(g)}\right) \textit{ if some group } g \textit{ is exhausted}, \tag{67}$$

$$\textit{which tends asymptotically towards } 1 \textit{ as } \sigma \to 1$$

where $0 < k < 1$ is some constant, PER(g) is the number of recruitment attempts made targeting group g before g is exhausted, and TR(g) is the total number of recruitment attempts made targeting group g. Further, we can express TR(g) as:

$$TR(\mathrm{g}) = \frac{d(g)}{\Sigma_{g' \epsilon \{A,B\}} d(g')} \sigma N \tag{68}$$

And PER(g) as:

$$PER(\mathrm{g}) = s(g) + \epsilon \tag{69}$$

where $s(A) = \alpha$, $s(B) = \beta$ and ϵ is some small positive constant.

Proof: The proof follows from the observations before Lemma 3.16. For each of the two groups A and B in the target population, one of two cases is possible: either the group is exhausted, or it is not. If neither group is exhausted, then we only have to consider repeated sampling events due to the same recruit being targeted multiple times by chance. Since we assume every recruit within a group has the same degree, the probability of a repeated sampling event for some group is determined entirely by the number of individuals already recruited from that group. We can then use the same reasoning as in Theorem 3.13 to show that, when no group is exhausted, ρ_ω is dominated by the sum of two linear functions of s, which is itself a linear function of σ.

Now consider the case when some group g is exhausted. In this case, ρ_ω is an interpolation between the density of repeated sampling events prior to exhaustion, and 1, which is the density after exhaustion, and $\rho_{\omega NE}$, which is the density from any nonexhausted groups. Given TR total recruitment attempts made targeting g and PER of those attempts made prior to exhaustion, we can express this interpolation as:

$$\rho_\omega = O\left(\rho_{\omega NE} + \frac{kPER(g)}{TR(g)} + \frac{TR(g) - PER(g)}{TR(g)}\right) \tag{70}$$

or

$$\rho_\omega = O\left(\rho_{\omega NE} + 1 - \frac{(1-k)PER(g)}{TR(g)}\right) \tag{71}$$

where k is some constant between 0 and 1 representing the linear dependence between number of recruitment attempts and density of repeated sampling events prior to exhaustion. This asymptotic bound equation consists of two parts: $\rho_{\omega NE}$, which grows linearly in sampling fraction, and $1 - (1-k)PER(g)/TR(g)$. The growth of the second part depends entirely on $TR(g)$, since $PER(g)$ remains constant once g is exhausted and k is a constant. We now derive an expression for $TR(g)$. Since the network is a preferential

attachment network, the number of recruitment attempts targeting group g is simply the total number of recruitment attempts times the ratio of the group degree to the sum degree:

$$TR(\mathrm{g}) = \frac{d(g)}{\Sigma_{g' \epsilon \{A,B\}} d(g')} \sigma N \tag{72}$$

Plugging Equation 72 into Equation 71, we have:

$$\rho_\omega = O\left(\rho_{\omega NE} + 1 - \frac{(1-k)PER(g)(d(A)+d(B))}{\sigma d(g) N}\right) \tag{73}$$

Equation 73 is dominated by the nonlinear term $1 - (1-k)PER(g)(d(A)+d(B))/(\sigma d(g)N)$ so we can rewrite it as

$$\rho_\omega = O\left(1 - \frac{(1-k)PER(g)(d(A)+d(B))}{\sigma d(g) N}\right) \tag{74}$$

Note that this quantity increases asymptotically towards 1 as $\sigma \rightarrow 1$ since all the terms except σ are constants and σ is in the denominator of the fraction.

The only remaining piece of the proof is to derive an expression for $PER(g)$. $PER(g)$ is the number of recruitment attempts made targeting group g before the group is exhausted. Note that this is not simply the number of individuals in g, as we showed Lemma 3.16, making m recruitment attempts generally yields fewer than m distinct recruits. In order to calculate this value more precisely, we need to solve Approximation 65 for m given $DR = s(g)$, that is:

$$s(g) \approx m - \frac{m(m-1)}{2s(g)^2} \tag{75}$$

This equation is highly nonlinear but it is easy to estimate an approximate solution as $m \approx s(g) + \epsilon$ for some small ϵ (too small to significantly affect ρ_ω). Plugging in that estimate, we have:

$$s(g) \approx s(g) + \epsilon - \frac{1}{2} - \frac{(2\epsilon - 1)}{2s(g)} - \frac{\epsilon(\epsilon - 1)}{2s(g)^2} \tag{76}$$

For $s(g) \gg \epsilon$, the right-hand side is bigger than the left-hand side, and we have a sufficient condition—enough recruitment attempts have been targeted at g to exhaust it. So, we can approximate *PER(g)* as:

$$m \approx s(g) + \epsilon \tag{77}$$

□

4. Simulations

We employed simulation to explore scenarios not addressed by the analytic results. The analytic results do not specifically account for multiple seeds or recruitment processes with one versus multiple coupons given to recruiters, nor do they account for homophily,

so simulating the chain-recruitment process provides additional insight into how the presence of multiple seeds and the number of coupons available to the recruiters might impact the nonreplacement bias. In these simulations, the targeted subpopulation comprised 20% of the nodes. We simulated respondent-driven sampling and calculated the Heckathorn 2007 RDS estimates for the target subgroup.

Networks

The simulated chain-recruitment samples were generated from Watts–Strogatz Small-World Networks, Barabási-Albert preferential attachment networks and Erdős-Rényi random graphs as implemented in the Python package NetworkX (Hagberg et al. 2008). These networks have uniform, power-law, and Poisson degree distributions, respectively. We selected network parameters to maintain a mean degree of about eight for standard graphs and 16 for the higher-density graphs. Networks were of orders 500 and 5000 nodes. We present results for the 5000-node networks because we were unable to obtain meaningful results for low sampling fraction on the smaller networks. For example, a five-percent sample of a 500-node network would yield only 25 individuals. At this very small scale, RDS samples are extremely idiosyncratic and dependent on the seeds, so the variance between individual RDS samples is too large to produce a consistent pattern of results. Holding the number of nodes constant while manipulating sample fraction through sample size most closely models the tradeoffs field researchers must consider when implementing RDS. We recognize that allowing sample size to covary with sampling fraction can make the results harder to interpret, so we conducted a second set of simulations, following Gile (2011), holding sample size constant while varying the number of nodes in the network to manipulate the sampling fraction. This second approach covaries network density with sampling fraction. We report results that are consistent between the two approaches and note a few differences in the results section. Furthermore, we investigated large networks of up to 20,000 nodes and found that the results for these large networks did not differ substantially from the results reported below.

To manipulate homophily in the population, we created two separate networks. One network contained the entire target subgroup and the other network contained the rest of the population. These networks were then merged using a random double-edge swap that preserves the degree of each node (Maslov and Sneppen 2002). Different levels of rewiring result in different numbers of crosscutting ties, and thus create different levels of homophily. The target subgroup population proportion was held constant at $p = .20$. The networks ranged between zero and .75 homophily, where zero corresponds to the homophily score when ties in the network are completely at random and the maximum score of 1 corresponds to a group with no ties to other groups. Details of the calculation and rationale are described in Heckathorn (2002).

To explore the effect of high and low relative mean degree for the target population, we reallocated edges from one subgroup to the other while maintaining overall mean degree and similar overall edge count. Due to the interdependence of network parameters and the particular network generation algorithms used, it was impossible to reallocate edges in these networks without also causing some variation on other parameters. We chose to hold the node count constant while matching mean degree within groups across each network

type as closely as possible. We permitted slight variation in the number of total edges as long as this variation was less than one extra edge per node. For the low mean degree networks, the mean degree for the nontarget group was 1.5 times higher than the mean degree for the target group. In the high mean degree networks, the mean target group degree was 1.5 times the mean nontarget group degree. We collected simulated samples from each of the three levels of mean degree differences for each of the three network types. Table 2 shows the variation in the total number of edges for different levels of mean degree difference.

As documented in Gile (2011) and Gile and Handcock (2010), bias in RDS estimate increases with larger differences in mean degrees among groups. In other words, estimates of variables that are strongly related to degree will exhibit much larger bias than estimates of variables that have no relationship to degree. In order to select an appropriate degree ratio for our simulations, we estimated degree ratios for five public health RDS studies: two RDS surveys of Latino MSM (Ramirez-Valles et al. 2005) and three U.S. sites of the SATH-CAP studies (Iguchi et al. 2009). We examined 14 variables for each of the Latino MSM studies, resulting in 37 degree ratios per study; we examined 19 variables for each of the SATH-CAP studies, resulting in 48 degree ratios per study. We found that only three of the 218 degree ratios examined were greater than 2. Furthermore, we found that only SATH-CAP's RTI site has more than 25% of its ratios greater than 1.5. Over all studies and variables, 50% of ratios were less than 1.2, 88% of the ratios were less than 1.5, and 97.7% were less than 2. We conclude from this analysis that a significant majority of network size ratios will be less than 1.5 in public health RDS studies, and that virtually all network size ratios will be less than 2. Therefore, our simulations employ a ratio of 1.5 to be representative of the majority of ratios a typical RDS public health study will observe.

Simulation Parameters

We simulated with- and without-replacement RDS using both branching and nonbranching referral processes. Each simulation started from six randomly selected seed nodes and each seed was granted a recruitment quota of c. These seeds form the first wave of recruiters. Each recruiter recruited up to c of their available neighbors and each of these new recruits was allocated c successful recruitments for the next recruiting wave. In the without-replacement samples, nodes were considered available for recruiting only if they were not already in the sample. Each recruiter selected their recruits at random from their available neighbors until either c new recruits were generated or all available neighbors were recruited. Recruitment was asynchronous and the order of execution of recruitment privileges was randomized by node.

Table 2. Differences in Group Mean Degree for Target and Nontarget Groups

Target Group Mean	Target	Non-Target	Total Edges
Lower	12	18	42000
Same	16	16	40000
Higher	24	16	44000

The recruitment quota was one for nonbranching chains ($c = 1$) or three for branching chains ($c = 3$). We considered other quotas, including $c = 2$, $c = 4$ and c chosen uniformly at random between 0 and 4, but chose to focus on the distinction between branching and nonbranching samples as all branching samples were substantially similar and recruitment quotas are not the main focus of this article.

Each sample began with six seeds selected uniformly at random from the nodes in the network and chain recruitment continued until the target sample size was reached. In a typical RDS field study, where sampling is without replacement, the sample size is the number of people interviewed and equals the number of interviews and the number of unique nodes visited. When simulating with-replacement RDS, the sample size corresponds to the number of node visits (interviews) rather than the number of unique nodes visited. In simulated with-replacement samples the number of nonduplicate nodes in the sample ranged from about 60 to 99% of the sample size. As expected, the percentage of unique nodes visited decreased as sampling size increased. The relationship is linear with an average of 90% unique visits for sampling fractions of 0.05 and an average of 65% for sampling fractions of 0.8. The pattern was substantially similar on all three network types.

We explored sampling fractions of 0.05, 0.1, 0.2, 0.4, 0.6 and 0.8. Sample chains constructed without replacement occasionally terminate early when all nodes near the recruiting nodes have already been recruited, and the probability of early chain termination increases with the target sample fraction. To reach the target sample sizes, we adopted the following procedure: if a simulated RDS sample has no "productive" recruitment chains, but has not yet reached target sample size, add a single new seed to the sample, chosen uniformly at random from the set of nodes not yet recruited. We applied this procedure iteratively to all simulated samples that had failed to reach their target sample size, until the total number of recruits + seeds in these samples was equal to the target sample size.

Sample Filtering

The calculation of Heckathorn (2007) RDS estimates requires cross recruitment among the different subgroups in the population. When networks contain very few intergroup ties (i.e., have extreme levels of homophily), RDS samples drawn from these networks often fail to capture any of these critical ties. We excluded samples with fewer than four intergroup recruitments in each direction.

Though most samples could be collected from the initial seed cluster, some samples required the addition of new seeds as described above. While many of these samples succeeded after the addition of a few new seeds, some samples required the addition of many seeds to reach the target sampling fraction. Since seeds are recruited with an unknown probability in real-world RDS studies, their sample inclusion probability cannot be known. Instead, the network size for seeds is treated as missing data and must be imputed. In effect, this means that samples with a large proportion of seed nodes have a large proportion of missing data. When the number of seeds becomes too large, RDS estimation is inappropriate, so samples with more than five percent seeds were excluded from the analysis.

The proportion of samples excluded by these criteria depends on sampling method, sample size, and network structure. For instance, small samples from highly homophilous

networks are less likely to capture cross-group ties. In less dense networks with nonbranching samples, samples were much more likely to dead end, thus requiring additional seeds. For the non-constant sample size simulations two to twelve percent of samples at sampling fraction level .05 were excluded and most of the excluded samples were drawn from the most homophilous networks. The exclusion rate rapidly decreases with sampling fraction, dropping below one percent by 20% sampling fraction. We attribute this to the small sample sizes at low sampling fractions. The constant sample size simulations ($n = 500$) had a more constant rejection rate, typically below five percent.

The two criteria for sample exclusion—fewer than four intergroup recruitments in each direction and more than five percent seeds—are based on easy-to-recognize sample characteristics that are observable by researchers conducting RDS studies in the field and reflect the importance of cross-group ties to the RDS estimation process.

Simulation Results

We now present results of simulations that confirm and extend our analytical results. The parameter space we are examining is extremely high dimensional, including sampling fraction, network structure, relative mean degree of the target group, and homophily. In this section, we first present the results about ρ_ω, which confirm our analytical results, and then describe the bias of simulated RDS samples, which extend our analytical results.

We begin by looking at ρ_ω as a function of sampling fraction for without-replacement RDS samples across different network structures and mean degree values for the target group. Figures 1, 2, and 3 correspond to Poisson Random Graphs, Small-World Graphs, and Preferential Attachment Graphs, respectively. In each figure, degree differential changes from left to right, with a target group having a smaller mean degree than the rest of the network on the left pane; same mean degree as the rest of the network in the center pane; and a greater mean degree than the rest of the network in the right pane. See Table 2 for the absolute degree values of the target and nontarget groups in both cases. Finally, within each pane, we have sampling fraction on the x-axis and frequency of repeated sampling events on the y-axis. The multiple lines for each pane correspond to the nonbranching ($c = 1$) and branching ($c = 3$) cases and the width of the 95% confidence intervals for these cases, respectively.

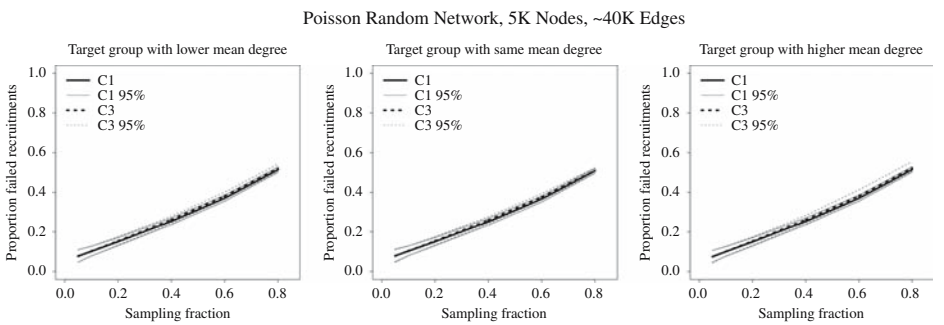

Fig. 1. Recruitment failures for Poisson Random Graphs

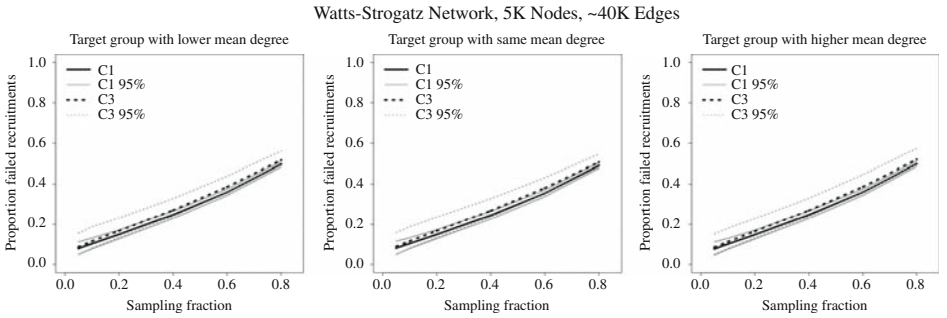

Fig. 2. Recruitment failures for Small-World Graphs

In Figure 1, we can see that in all three panes the relationship between ρ_ω and σ is linear, as predicted by Theorem 3.13. The absence of any asymptotic behavior suggests that neither group is exhausted in the course of sampling, so the results correspond to Equation 66 in Theorem 3.18.

In Figure 2, we can see that in all three panes, the relationship between ρ_ω and σ is linear, as predicted by Theorem 3.14. The absence of any asymptotic behavior suggests that neither group is exhausted in the course of sampling, so the results correspond with Equation 66 in Theorem 3.18.

In Figure 3, we can see that in all three panes, the relationship between ρ_ω and σ is slightly sublinear when $\sigma < .2$, and then quickly approaches linearity, as predicted by Theorem 3.15. The absence of any asymptotic behavior suggests that neither group is exhausted in the course of sampling, so the results correspond with Equation 66 in Theorem 3.18.

Note that in all Figures 1–3 there is no significant difference between the branching and the nonbranching cases, which confirms the independence of our analytic results in Theorems 3.13–3.15 and 3.18 on the number of respondents recruited by each recruiter, represented by the parameter k in Theorem 3.12.

We present a set of results that show the effect of homophily on ρ_ω. Our analysis does not account for the effect of homophily on the frequency of repeated sampling events, so the simulations serve as a useful counterpart for analyzing recruitment failures in high-homophily regimes. In our work, we use the definition of homophily presented in Heckathorn (2002).

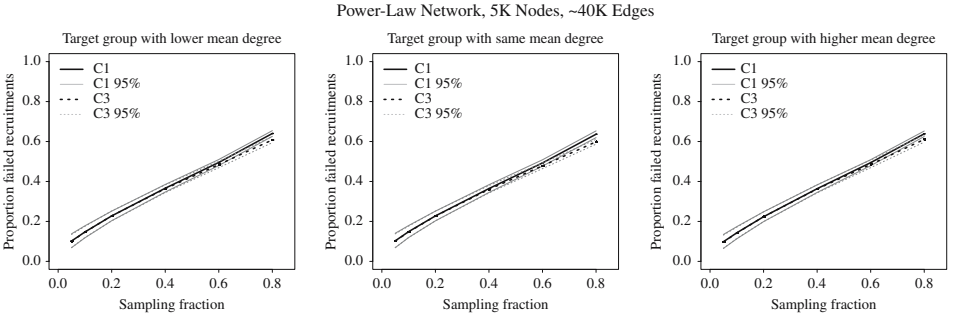

Fig. 3. Recruitment failures for Preferential Attachment Graphs

We find no significant difference in mean or the variance of the probability of repeated sampling events, ρ_ω, across the full range of homophily values we explore (0 to .75) for Poisson Random Graphs or Preferential Attachment Graphs, even when varying the target group mean degree and the branching value. For Poisson Random Graphs, ρ_ω ranges between 0.22 and 0.25 across the full range of homophily, and for Preferential Attachment Graphs, ρ_ω ranges between 0.35 and 0.36 across the full range of homophily.

Figure 4 shows the relationship between ρ_ω and homophily, averaged across all values of sampling fraction, for the Small-World Graph. In this figure, as in Figures 1–3, the degree ratio changes from left to right, with a target group having a smaller mean degree than the rest of the network on the left pane; the same mean degree as the rest of the network in the center pane; and a greater mean degree than the rest of the network in the right pane.

Figure 4 shows that for Small-World Graphs, homophily has an apparent effect on ρ_ω. Initially, the branching and nonbranching cases start out with the same level of ρ_ω, but as homophily increases, ρ_ω increases superlinearly for the branching case. The increase is likely due to the high level of clustering in Small-World Networks. More clustered networks have more within-group collisions even with a homophily of 0, and an increase in homophily will only exacerbate the within-group collisions for these networks. In contrast, Poisson Random Graphs and Preferential Attachment Graphs feature low levels of clustering, and the effect of homophily on ρ_ω is negligible ($< 5\%$) across the range of homophily values up to 0.7. Overall, the effect of homophily on nonreplacement bias is negligible compared to the effect of sampling fraction. High homophily does impact the probability of capturing cross-group ties in an RDS sample and is an important consideration in RDS survey design, but does not appear to contribute to the bias from sampling without replacement.

We now present simulation results that show the effect of sampling fraction on overall RDS bias. Sampling without replacement is only one factor that could affect the bias of an RDS estimate. For instance, in a larger sample, we would find more repeated sampling events, which may lead to increased bias, but also less dependence on initial conditions (seeds), which may lead to decreased bias. Therefore, we investigate both overall RDS bias and the part of it that is attributable to sampling without replacement. We run two parallel sets of simulations: one sampling without replacement, as above, and a second on the same network but sampling with replacement, so individuals can participate in a study more than once. For clarity of visualization, we here focus exclusively on the branching ($c = 3$) RDS samples. We discuss nonbranching samples at the end of this section.

The figures below show both mean bias and sampling variability, which we define as the range of the 95% confidence interval of the sampling distribution over the set of samples. Figure 5 shows the mean bias and sampling variability for the Poisson Random Graph. The three panels correspond to three levels of differential degree, with the target group having smaller degree on the left pane, equal degree in the center pane, and higher degree on the right pane. We use a legend to differentiate between samples drawn with (WR) and without replacement (WOR).

There are three important observations to make based on these graphs: first, while the expected bias in samples drawn with replacement is zero, the expected bias in samples drawn without replacement remains small (less than five percentage points away from the

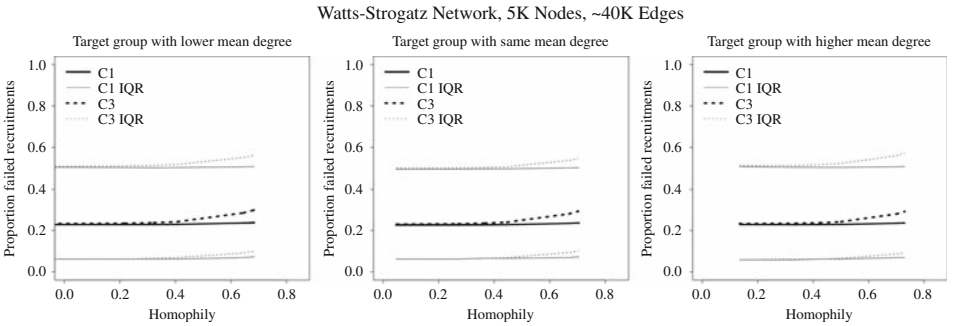

Fig. 4. Recruitment failures by homophily for Watts-Strogatz Network

true value), even up to sampling fractions of 80%. Therefore, we can say that *the effect of violating the sampling-with-replacement assumption on the expected bias of RDS estimates is five or fewer percentage points across the parameter range explored in this study*. Second, the mean bias and sampling variability for samples drawn without replacement are nearly identical to mean bias and sampling variability for samples drawn with replacement up to a sampling fraction of about 20%. Therefore, we can say that *sampling without replacement is not a major source of estimator bias in RDS studies for sampling fractions under 20%.* Finally, also across all three graph structures, we see a nonlinear relationship between the variability of samples drawn with versus without replacement. The variability of samples drawn with replacement shrinks to zero in the limit of 100% sampling fraction (not shown in graphs). The variability of samples drawn without replacement decreases and then increases: it follows the behavior of samples drawn with replacement up to sampling fraction ~40%, and then increases rapidly, except when the target group has higher mean degree. The reasons for the pattern will be explored in a future paper.

We plot the mean bias and sampling variability for two other types of networks, Watts-Strogatz and Power Law, in the figure in the supplemental data. The results for these types of networks are substantially similar to the results for Poisson Random Graphs as shown in Figure 5.

Nonbranching samples represent an idealized process that is not representative of empirical RDS studies, but the simulations indicate mean bias for nonbranching

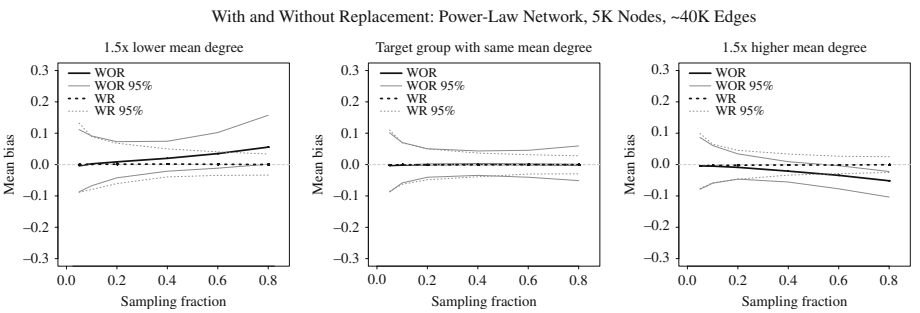

Fig. 5. Mean bias by sampling fraction on a Poisson Random Network structure for samples drawn with (WR) and without replacement (WOR)

without-replacement samples is very similar to the branching without-replacement samples (not shown in graphs). The relationship between sampling fraction and the sampling variability of nonbranching without-replacement samples is similar to that of nonbranching with-replacement samples (not shown in graphs).

Branching samples are less prone to chain exhaustion than nonbranching samples on the lower density graphs with mean degree of eight. Branching samples were able to reach their target sample size from the initial six seeds but nonbranching chains often became stuck, requiring many additional seeds to reach the target sample size on all network types. To reach the highest levels of sampling fraction, hundreds of additional seeds were required. Our sample filtering capped the allowable additional seeds at five percent of the overall sample. On the more dense mean degree sixteen networks, nonbranching samples are able grow as effectively as branching samples. Though not the main focus of this article, the observed difference in robustness between branching and nonbranching samples highlights the practical necessity of recruitment quotas greater than one for some network structures.

5. Discussion

The results above show the effect of sampling fraction, degree distribution, degree differential and group size on bias in RDS studies, both specifically due to sampling without replacement (in Figures 1 − 4), and overall bias from the true prevalence figures (in Figure 5). We can make six general observations based on the results:

The density of repeated sampling events, ρ_ω, increases monotonically in sampling fraction across network structures and degree differential. This increase is predicted by our analysis, and is generally linear for Poisson Random Graphs and Small-World Graphs, and slightly superlinear for Preferential Attachment Graphs. This increase suggests that bias due to sampling without replacement increases steadily with sampling fraction.

Homophily has a small effect on ρ_ω outside of Small-World Networks, which have a high level of clustering. Even for Small-World Networks, ρ_ω increases by less than ten percent from homophily $= 0$ to homophily $= .7$. Theoretical analyses (Heckathorn 2002, 28) have shown that the standard error of an RDS estimate increases exponentially with increases in homophily, so RDS is not a suitable sampling method for networks with homophily above .7.

Overall bias remains small across the range of simulated parameters, less than five percentage points for the highest sampling fractions. This suggests that, at high sampling fractions, increased bias due to sampling without replacement is counteracted by other factors (such as a larger sample size which usually results in a more diverse set of recruits). Note that bias is essentially zero when the target group has the same mean degree as the rest of the network, so many variables will exhibit minimal bias regardless of sampling fraction. As discussed above, approximately 85% of variables in RDS public health studies have degree ratios ≤ 1.5 and will therefore exhibit bias no greater than five percentage points, and the remaining 15% may exhibit more extreme levels of bias. Researchers should note the potential for more significant amounts of bias when extreme degree ratios are observed in an RDS study.

Both overall bias and sampling variability for samples drawn without replacement are essentially identical to the respective quantities for samples drawn with replacement for sampling fractions up to 20%. The implication of this result is that for sampling fractions of up to 20%, violations of the sampling-with-replacement assumption inherent in the Heckathorn 2007 RDS estimator can be considered negligible. This range includes RDS studies conducted in large cities, such as the CDC NHBS study of IDUs in 23 large US cities noted above, where the median sampling fraction was 2.3%, and studies of jazz musicians in New York City and San Francisco, with a maximum sampling fraction between the cities of 1.6%. For sampling fractions in the 20% to 40% range, the sampling-with-replacement assumption is a modest source of bias, with a magnitude of no more than two percentage points across the range of simulated parameters. Such cases may arise when RDS is employed in small towns, as in the case from Connecticut described above. Here, results should be interpreted with the potential for small amounts of bias in mind, especially for variables that have a strong relationship with degree.

The simulated 95% confidence interval is nonlinear for the branching cases drawn without replacement. The 95% confidence intervals for these cases shrink for sampling fractions of up to 40% and then diverge, so at very high sampling fractions, the RDS estimates for without-replacement branching cases for target groups with lower and equal mean degree have very wide 95% confidence intervals.

Our results help map the bias and recruitment failure spaces of the Heckathorn (2007) RDS estimates. Just as previous work (Heckathorn 2010) mapped design effect space, so we describe a parameter space where the Heckathorn (2007) RDS estimates produce different levels of bias. The bias is nearly constant in some parameters (homophily) and highly nonlinear in others (sampling fraction), with minimal bias in the parameter ranges of homophily $<.7$, and sampling fraction below 20%.

Our work has two simple implications for empirical RDS studies employing the Heckathorn (2007) estimator: in order to minimize bias, keep sampling fraction below 20%, and avoid very high ($>.7$) homophily networks. Our study points to the importance of presurvey ethnographic and field research to detect high levels of homophily and the use of empirical research such as capture-recapture methods to calculate the sampling fraction. We show that for small sampling fractions, RDS produces low levels of bias, but we do not rule out the possibility of using alternative estimators for very large sampling fractions.

6. Limitations

The principal goal of this article is to examine the effect of violation of the sampling-with-replacement assumption of RDS on bias of RDS population proportion estimates, specifically the effect of this violation for studies with high sampling fractions. The methodology of this article is a mix of analytic and simulation approaches. In this article, we have chosen to focus on our goal and methodology rather than explore many possible implications for RDS bias (and many possible methods); as a result, our analysis has several limitations.

First, we *only* explore the violation of one RDS assumption—sampling without replacement—and not the violation of other assumptions. Specifically, we do not explore violations of the assumption of random recruitment, which we repeat here:

Assumption 3.5 Respondents recruit as though they are selecting randomly from their neighborhoods.

This assumption states only that respondents recruit as though they are selecting randomly—not that respondents employ a truly random process when choosing recruits. There are complicated or nonsystematic ways that may achieve unbiased results without being truly random. At the same time, this assumption indicates that studies where recruitment has a particular bias (e.g., all respondents recruit people of the same gender) will produce biased estimates. This assumption is certainly worth examining, but is far beyond the scope of this article. We are interested in exploring this assumption in future work.

Second, our analysis focuses on targets of recruitment, not sources. We do not consider, for example, whether a respondent who tries to make multiple recruitments but fails (because they have already been recruited) will get discouraged and stop trying to recruit others. These questions are certainly worth investigating, but they are less analytically tractable than the issue of multiple recruitment attempts. For example, we have derived results about multiple recruitment attempts from assumptions about the underlying network structure and sampling fraction. Common network structures have been studied extensively in the literature analyzing social networks. Sampling fractions have been determined in meta-analyses of RDS studies (Wejnert et al. 2012). In contrast, in order to derive results about respondent behavior we would have to start with assumptions about the enthusiasm levels and so on of participants in RDS studies, which to our knowledge have not been explored in the RDS literature. Consequently, we think that a follow-up, empirical study of RDS participants would be better suited to analyzing RDS assumptions with respect to recruiter behavior.

Third, our analytic and simulation results examine the bias in the estimate of population proportions and not the sampling variance of the estimate. In our simulations, the difference between with- and without-replacement sampling variance increases with sample size and with-replacement sampling has lower sampling variance. This contrasts with Lu et al. (2012) who found less sampling variance in simulated without-replacement sampling on an empirical online social network. These different findings suggest that the relationship between sampling fraction and sampling variance may depend on currently unidentified elements of network structure. Sampling variance is clearly an important consideration that should be addressed in future work.

Fourth, we do not analytically link the proportion of repeated sampling events to the magnitude of the bias. The proportion of repeated sampling events is a way to quantify the divergence between with- and without-replacement sampling. We demonstrate both analytically and in simulations (in Figures 1–3) that the density of repeated sampling events (proxied by the proportion of failed recruitments in the simulation section) increases smoothly with the sampling fraction. The important analytic result is that this quantity grows in a bounded and predictable way as sampling fraction increases. Unfortunately, there are many possible sources of bias in RDS studies, including recruiter activity, recruiter preference for certain recruits, and failed recruitments. Most of these sources of bias are not analytically tractable. We are able to demonstrate via simulation that the increase in the proportion of repeated sampling events is associated with an increase in the bias from sampling with replacement. Our simulation results demonstrate a clear positive and near-linear relationship between the proportion of repeated sampling

events and sampling fraction and a clear positive relationship between sampling fraction and the magnitude of estimate bias when the subgroups in the target population have different mean degree.

Finally, we do not study empirical social networks of RDS participants. Such networks are extremely hard to collect: most RDS studies ask participants about the number, but not the exact identity, of their friends. This lack of specificity is crucial for privacy reasons; at the same time, it leads to a lack of knowledge about the empirical networks surveyed via RDS. There is no prima facie reason to assume the networks typically targeted by RDS surveys are structured in a fundamentally different way than fully mapped social networks; at the same time, empirical networks may have unusual features that lead to bias in RDS studies. For example, one individual may serve as the only broker between two otherwise physically and socially separated groups, such as one dealer connecting two groups of drug users in different neighborhoods. Such "chokepoints" would be extremely problematic for RDS studies, and yet for the abovementioned reason there are no empirical studies to our knowledge that investigate the frequency of these network structures in RDS studies. We hope that in future work we can study the question of whether additional, noninvasive questions during RDS studies can help researchers identify chokepoints or other unusual network structures in the field without violating participant privacy. Ultimately, we would like to come up with recommendations for researchers to dynamically adjust their sampling strategy when they encounter an unusual network structure, so as to minimize bias in the resulting RDS sample.

7. Conclusion

Our analysis has described a large parameter space of possible conditions for RDS studies, and the levels of bias across this space. We have shown that for a wide range of parameter values, mean bias remains extremely low, and the biased estimate is not significantly different from the true value under conditions which reasonably correspond to empirical RDS studies. We have also shown that higher levels of bias due to sampling without replacement do not necessarily correspond to higher levels of overall bias; on average, sampling without replacement is neither the only nor the dominant factor affecting RDS estimates.

Our results suggest that bias is negligible for sampling fractions up to 20%, a case which fits most studies of hidden populations in large urban settings, for example, the abovementioned studies where sampling fractions range from 0.6% to 8%. In the 20–40% range, the magnitude of bias depends on other parameters, especially whether the variable of interest is correlated with degree. Bias may be as much as two percent if the degree ratio is 1.5; if the variable is independent of degree, bias is again negligible. This case fits studies in small towns or sparsely populated rural areas, or studies in large cities with very large sample sizes. In the 40% to 80% range, the biases of the RDS estimator and the raw sample proportion tend to be in opposite directions, so an estimate in between these two will be less biased. Gile's (2011) successive sampling approach is a principled method to mediate between the RDS estimator and the raw sample proportion as a function of sampling fraction. Finally, at very high sampling fractions, the sample is best treated as a census rather than a sample, so no statistical estimation process is required. This fits

studies conducted either in very sparsely populated areas, or studies that are sufficient to saturate the target population.

An implication of these suggestions is that population-size estimation should be incorporated into all RDS studies; otherwise the most appropriate mode of analysis cannot be identified. A further implication is that population estimation should involve quantitative procedures such as capture-recapture or network scale-up, because estimates even from knowledgeable key informants can be wrong by more than an order of magnitude (Heckathorn et al. 2002).

This article introduced a new theoretical concept for analyzing bias in RDS analyses, a *repeated sampling event*, in which a peer-recruitment attempt fails because the respondent has already participated in the study. Analysis of the density of these events provides a new conceptual tool for analyzing bias in RDS analysis. It also has implications for research design. For example, in sparse networks where branching is precluded (i.e., respondents can recruit only a single peer), recruitment chains tend to die out quickly, so attaining a desired sample size may involve employing very large number of seeds; in extreme cases, more than half the sample may be seeds. Because RDS seeds contribute less information to the sample than a peer-recruited participant but cost the survey the same in terms of time and participation incentives, the efficiency of the project may be severely compromised. Furthermore, if a significant proportion of the sample is composed of seeds, RDS weights are not appropriate for calculating point estimates. However, if branching is permitted (e.g., allowing each respondent to recruit three peers), the number of seeds required to attain a specific sample size is dramatically reduced, thereby increasing the efficiency of the study. Hence, though we confirmed previous findings that branching can increase a study's design effects under some conditions (Goel and Salganik 2009), or have trivial effects under others (Heckathorn 2002), we also identify a compensatory benefit from building branching into a research design: in sparse networks, it can greatly increase a study's efficiency.

A final implication of the study is to demonstrate the importance of exploring a large parameter space when quantifying bias in RDS studies, for studies which explore only limited regions may produce misleading results, especially if the region investigated fails to encompass the full range of RDS studies reported in the literature. We explore sampling fractions between five percent and 80%, homophily between 0 and .7, and various network structures and degree distributions to produce results that can provide practical insight about the impact of nonreplacement bias on RDS estimates.

The supplemental data is available at: www.dx.doi.org/10.1515/JOS-2016-0002

Appendix: Relating This Analysis to RDS Work Previously Published in JOS

Implications of Violation of Sampling With Replacement for Volz-Heckathorn Estimator.

The Volz-Heckathorn estimator relies on a model of chain-referral sample as a random walk on a network. In the case of sampling with replacement, this model is accurate, and the random walk is a Markov Process, which in equilibrium occupies a node with probability proportional to degree (Salganik and Heckathorn 2004). However, in the case of sampling without replacement, the model is inaccurate: instead of being a random walk

on a network, the RDS sample is a self-avoiding walk (SAW) on a finite network. Since the network is finite and the SAW may not by definition visit a node more than once, in equilibrium the probability of it occupying any node approaches 0. In this case, we can no longer apply the arguments in Salganik and Heckathorn (2004), but must propose another model of RDS as a self-avoiding random walk.

By definition, an RDS sample is a finite-size chain-referral sample, in which no individual may be recruited more than once, drawn from a larger (but still finite) network. Thus, we can formalize RDS as a length $|R|$ SAW on a network of size $|P|$. This SAW corresponds to a reducible Markov Process MP^{WOR} on the set of nodes in R, where each state is a node and transitions between states correspond to recruitments. Note that for this reducible Markov Process, no state may have more than one incoming transition from another state. The reducible Markov Process has a number of differences to the irreducible Markov Process MP^{WR}, which models sampling with replacement. However, both processes may be encoded as transition matrices, and we can compare the transition matrices to measure the extent of bias due to sampling with replacement.

We can construct an incidence matrix M for the network, where for any individuals i, j in the larger population P, $R_{ij} = 1$ if i and j are connected, 0 otherwise. This matrix gives the equilibrium transition probabilities for MP^{WR}. Indeed, we can construct a *transition matrix* M^{WR} where for any individuals i, j in P, $M_{ij}^{WR} = 1$ if i recruited j into the chain-referral sample with replacement modeled by MP^{WR}, 0 otherwise. This matrix will approximate M in the sense that, for some node i, the larger i's in-degree ($\sum_i M_{ij}$), the more likely and more frequently will i be recruited in the chain-referral sample ($\sum_i M_{ij}^{WR}$). Similarly, we can construct a transition matrix M^{WOR} where for any individuals i, j in P, $M_{ij}^{WOR} = 1$ if i recruits j to participate in the RDS study, and 0 otherwise.

Note that if no individual is ever recruited more than once in the course of the chain-referral process modeled by MP^{WR}, then $M^{WOR} = M^{WR}$. However, even a single repeated sampling event can introduce a chain of differences between the two transition matrices—for example, let A be a with-replacement sample wherein i recruits j who recruits k who recruits i who recruits l. Let B be a repeated sampling sample wherein i recruits j who recruits k, and then the sample ends. The corresponding transition matrices would look as follows:

A

	i	j	k	l
i	0	1	0	1
j	0	0	1	0
k	1	0	0	0
l	0	0	0	0

B

	i	j	k	l
i	0	1	0	0
j	0	0	1	0
k	0	0	0	0
l	0	0	0	0

Still, by definition any atomic (cell-level) differences between M^{WOR} and M^{WR} are due entirely to repeated sampling events, and so we can operationalize the bias due to sampling with replacement as the difference between the two matrices.

8. References

Barabási, A.L. and R. Albert. 1999. "Emergence of Scaling in Random Networks." *Science* 286: 509–512. Doi: http://dx.doi.org/10.1126/science.286.5439.509.

Bernard, H.R., T. Hallett, A. Iovita, E.C. Johnsen, R. Lyerla, C. McCarty, M. Mahy, M.J. Salganik, T. Saliuk, O. Scutelniciuc, G.A. Shelley, P. Sirinirund, S. Weir, and D.F. Stroup. 2010. "Counting Hard-to-Count Populations: the Network Scale-Up Method for Public Health." *Sexually Transmitted Infections* 86 (suppl. II): ii11–ii15. Doi: http://dx.doi.org/10.1136/sti.2010.044446.

Bernhardt, A., M.W. Spiller, and N. Theodore. 2012. "Employers Gone Rogue: Explaining Industry Variation in Violations of Workplace Laws." *Industrial and Labor Relations Review*. Available at: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2013376 (accessed January 2016).

Curtis, R., K. Terry, M. Dank, K. Dombrowski, and B. Khan. 2008. *The Commercial Sexual Exploitation of Children in New York City, Volume One: The CSEC Population in New York City: Size, Characteristics, and Needs*, Final report submitted to the National Institute of Justice. New York: Center for Court Innovation and John Jay College of Criminal Justice. Available at: https://www.ncjrs.gov/pdffiles1/nij/grants/225083.pdf (accessed January 2016).

Erdős, P. and A. Rényi. 1959. "On Random Graphs." *Publ. Math* 6: 290–297.

Gile, K.J. 2011. "Improved Inference for Respondent-Driven Sampling Data with Application to HIV Prevalence Estimation." *Journal of the American Statistical Association* 106: 135–146. Doi: http://dx.doi.org/10.1198/jasa.2011.ap09475.

Gile, K.J. and M.S. Handcock. 2010. "Respondent-Driven Sampling: An Assessment of Current Methodology." *Sociological Methodology* 40: 285–327. Doi: http://dx.doi.org/10.1111/j.1467-9531.2010.01223.x.

Goel, S. and M.J. Salganik. 2009. "Respondent-Driven Sampling as Markov Chain Monte Carlo." *Statistics in Medicine* 28: 2202–2229. Doi: http://dx.doi.org/10.1002/sim.3613.

Hagberg, A.A., D.A. Schult, and P.J. Swart. 2008. "Exploring Network Structure, Dynamics, and Function Using NetworkX." In Proceedings of the 7th Python in Science Conference (SciPy2008), Pasadena, CA, August 2008. Edited by G. Varoquaux, T. Vaught, and J. Millman. 11–15.

Heckathorn, D. 1997. "Respondent-Driven Sampling: a New Approach to the Study of Hidden Populations." *Social Problems* 44: 174–199. Doi: http://dx.doi.org/10.2307/3096941.

Heckathorn, D.D. 2002. "Respondent-Driven Sampling II: Deriving Valid Population Estimates from Chain-Referral Samples of Hidden Populations." *Social Problems* 49: 11–34. Doi: http://dx.doi.org/10.1525/sp. 2002.49.1.11.

Heckathorn, D.D. 2007. "Extensions of Respondent-Driven Sampling: Analyzing Continuous Variables and Controlling for Differential Recruitment." *Sociological Methodology* 37: 151–207. Doi: http://dx.doi.org/10.1111/j.1467-9531.2007.00188.x.

Heckathorn, D.D. 2010. "Sampling Elusive and Hard-to-Reach Populations: The Scope and Limits of Respondent-Driven Sampling." Presented at the Duke Population Research Institute Workshop "Challenging Samples: Networks and Surveys in Demographic and Health Research" on May 7, 2010.

Heckathorn, D.D. 2011. "Snow ball versus Respondent-driven Samping." *Sociological Methodology* 41: 355–366. Doi: http://dx.doi.org/10.1111/j.1467-9531.2011.01244.x.

Heckathorn, D.D. and J. Jeffri. 2001. "Finding the Beat: Using Respondent-Driven Sampling to Study Jazz Musicians." *Poetics* 28: 307–329. Doi: http://dx.doi.org/10.1016/S0304-422X(01)80006-1.

Heckathorn, D.D., R.S. Broadhead, and B. Sergeyev. 2001. "A Methodology for Reducing Respondent Duplication and Impersonation in Samples of Hidden Populations." *Journal of Drug Issues* 31: 543–564.

Heckathorn, D.D., S. Semaan, R.S. Broadhead, and J.J. Hughes. 2002. "Extensions of Respondent-Driven Sampling: A New Approach to the Study of Injection Drug Users Ages 18–25." *AIDS and Behavior* 6: 55–67. Doi: http://dx.doi.org/10.1023/A:1014528612685.

Iguchi, M.Y., A.J. Ober, S.H. Berry, T. Fain, D.D. Heckathorn, P.M. Gorbach, R. Heimer, A. Kozlov, L.J. Ouellet, S. Shoptaw, and W.A. Zule. 2009. "Simultaneous Recruitment of Drug Users and Men Who Have Sex with Men in the United States and Russia Using Respondent-Driven Sampling: Sampling Methods and Implications." *Journal of Urban Health* 86 (Suppl. 1): 5–31. Doi: http://dx.doi.org/10.1007/s11524-009-9365-4.

Jeffri, J., D.D. Heckathorn, and M.W. Spiller. 2011. "Painting Your Life: a Study of Aging Visual Artists in New York City." *Poetics* 39: 19–43. Doi: http://dx.doi.org/10.1016/j.poetic.2010.11.001.

Lansky, A., A. Drake, and H.T. Pham. 2009. *HIV-Associated Behaviors Among Injecting-Drug Users – 23 Cities, United States, May 2005-February 2006*. Morbidity and Mortality Weekly Report April 10, 2009, 58: 329–332. Available at: www.cdc.gov/mmwr/pdf/wk/mm5813.pdf (accessed January 2016).

Lee, C.-H., X. Xu, and D.Y. Eun. 2012. "Beyond Random Walk and Metropolis-Hastings Samplers." In Proceedings of the 12th ACM SIGMETRICS/PERFORMANCE Joint International Conference on Measurement and Modeling of Computer Systems – SIGMETRICS'12, 319. ACM Press. Doi: http://dx.doi.org/10.1145/2254756.2254795.

Lu, X., L. Bengtsson, T. Britton, M. Camitz, B.J. Kim, A. Thorson, and F. Liljeros. 2012. "The Sensitivity of Respondent-Driven Sampling." *Journal of the Royal Statistical Society. Series A: Statistics in Society* 175: 191–216. Doi: http://dx.doi.org/10.1111/j.1467-985X.2011.00711.x.

Malekinejad, M., L.G. Johnston, C. Kendall, L.R.F.S. Kerr, M.R. Rifkin, and G.W. Rutherford. 2008. "Using Respondent-Driven Sampling Methodology for HIV Biological and Behavioral Surveillance in International Settings: A Systematic Review." *AIDS and Behavior* 12(suppl. 4): 105–130. Doi: http://dx.doi.org/10.1007/s10461-088-9421-1.

Maslov, S. and K. Sneppen. 2002. "Specificity and Stability in Topology of Protein Networks." *Science* 296: 910–913. Doi: http://dx.doi.org/10.1126/science.1065103.

Ramirez-Valles, J., D.D. Heckathorn, R. Vázquez, R.M. Diaz, and R.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. Doi: http://dx.doi.org/10.1007/s10461-005-9012-3.

Salganik, M.J. and D.D. Heckathorn. 2004. "Sampling and Estimation in Hidden Populations Using Respondent-Driven Sampling." *Sociological Methodology* 34: 193–239. Doi: http://dx.doi.org/10.1111/j.0081-1750.2004.00152.x.

Smylie, J., M. Firestone, L. Cochran, C. Prince, S. Maracle, M. Morley, S. Mayo, M.W. Spiller, and B. McPherson. 2011. *Our Health Counts Urban Aboriginal Health Database Research Project – Community Report First Nations Adults and Children, City of Hamilton*. Hamilton: De Dwa Da Dehs Neys Aboriginal Health Centre. Available from: http://www.stmichaelshospital.com/crich/wp-content/uploads/ourhealth-counts-report.pdf (accessed January 2016).

Tillé, Y. 2006. *Sampling Algorithms*. New York: Springer.

Volz, E. and D.D. Heckathorn. 2008. "Probability Based Estimation Theory for Respondent Driven Sampling." *Journal of Official Statistics* 24: 79–97.

Wallace, R. 1991. "Social Disintegration and the Spread of AIDS: Thresholds for Propagation Along 'Sociogeographic' Networks." *Social Science & Medicine* 33: 1155–1162. Doi: http://dx.doi.org/10.1016/0277-9536(91)90231-Z.

Wang, J., 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. Doi: http://dx.doi.org/10.1016/j.drugalcdep. 2004.10.011.

Watts, D.J. and S.H. Strogatz. 1998. "Collective Dynamics of 'Small-World' Networks." *Nature* 393(6684): 440–442. Doi: http://dx.doi.org/10.1038/30918.

Wejnert, C., H. Pham, N. Krishna, B. Le, and E. DiNenno. 2012. "Estimating Design Effect and Calculating Sample Size for Respondent-Driven Sampling Studies of Injection Drug Users in the United States." *AIDS and Behavior* 16(4): 797–806. Doi: http://dx.doi.org/10.1007/s10461-012-0147-8.

Received August 2014
Revised July 2015
Accepted September 2015