Published in final edited form as:
Sociol Methodol. 2012 August ; 42(1): 206–256. doi:10.1177/0081175012461248.

Network Sampling with Memory: A proposal for more efficient sampling from social networks

Ted Mouw and **Ashton M. Verdery**

Abstract

Techniques for sampling from networks have grown into an important area of research across several fields. For sociologists, the possibility of sampling from a network is appealing for two reasons: (1) A network sample can yield substantively interesting data about network structures and social interactions, and (2) it is useful in situations where study populations are difficult or impossible to survey with traditional sampling approaches because of the lack of a sampling frame. Despite its appeal, methodological concerns about the precision and accuracy of network-based sampling methods remain. In particular, recent research has shown that sampling from a network using a random walk based approach such as Respondent Driven Sampling (RDS) can result in high design effects (DE)—the ratio of the sampling variance to the sampling variance of simple random sampling (SRS). A high design effect means that more cases must be collected to achieve the same level of precision as SRS. In this paper we propose an alternative strategy, Network Sampling with Memory (NSM), which collects network data from respondents in order to reduce design effects and, correspondingly, the number of interviews needed to achieve a given level of statistical power. NSM combines a "List" mode, where all individuals on the revealed network list are sampled with the same cumulative probability, with a "Search" mode, which gives priority to bridge nodes connecting the current sample to unexplored parts of the network. We test the relative efficiency of NSM compared to RDS and SRS on 162 school and university networks from Add Health and Facebook that range in size from 110 to 16,278 nodes. The results show that the average design effect for NSM on these 162 networks is 1.16, which is very close to the efficiency of a simple random sample (DE=1), and 98.5% lower than the average DE we observed for RDS.

1. Introduction

The prospect of sampling from social networks has long intrigued social scientists (cf. Goodman 1961). To fix terms, we define a sample from a social network as one where current respondents help recruit new respondents through either active (when respondents "drive" recruitment by bringing in new subjects) or passive means (when they facilitate contact between researchers and would-be subjects)[1]. Such an approach is a special case of a "link-tracing" sampling design which can be applied to sampling from any relational network, social or otherwise. For sociologists, the motivation to sample from social networks comes from three sources: a) a generally accepted belief that respondents, even stigmatized ones, are more likely to participate and respond truthfully in surveys when they are referred by acquaintances or friends rather than randomly contacted, b) an increasing desire to understand relational network structures and a growing body of theory suggesting

[28]The average finite sample bias was .0057 compared to .0036 for the Hybrid approaches with and without the 15% error rate with the truncated data (maximum degree of 20), compared to .297 and .0118 for RDS with the truncated and full network data.

[1]For simplicity of writing, we occasionally refer to sampling from a social network as "network sampling" in this paper. When we do this, we mean sampling from a social network and not the large body of work dedicated to making inferences about social network structures from more limited data (e.g., Potter Handcock et al. 2011).

the importance of social influence, and c) heightened interest in understanding stigmatized populations (such as those at risk of HIV) that are difficult to survey with traditional methods. Though it was long thought that samples from networks could not yield unbiased results, work in graph theory on the logic of random walks on graphs, which have been widely studied in the mathematics, physics, and computer science literatures (cf. Lovasz 1993), has shown that unbiased mean estimates are attainable. A basic finding is that, under certain conditions, if the random walk proceeds long enough, it will eventually settle into an equilibrium state where the sampling probability is proportional to the respondent's number of social ties (Lawler and Coyle 1999), which can be used as a weight to accurately estimate sample means and proportions (Volz and Heckathorn 2008). Overall, the potential appeal of link-tracing sampling methods should be growing in the current research climate as a result of declining survey participation rates and the challenge of contacting willing participants (Altrosic et al., 2001), because of increased interest in social network effects and peer influence (e.g., Burt 1984), and because of increased funding for studying marginalized groups, such as those at risk of HIV transmission.

Some of the best evidence for the effectiveness of link-tracing based sampling methods in human social networks comes from work on stigmatized, hidden, and hard to reach populations, which have historically been understudied because of the challenge of recruiting them. A recently developed random-walk based approach, “respondent driven sampling” (RDS) collects data on the number of ties each respondent has (Heckathorn 1997; Volz and Heckathorn 2008; Johnston 2008; cf. Gile and Handcock 2010 for a recent summary of this approach). Under certain conditions—in particular, a connected component with at least one triangle and undirected ties (i.e., everyone in the population is reachable through some chain of connections from everyone else, there is at least one group of three people who are all connected to each other, and all ties are reciprocated), a sufficiently long chain of referrals, sampling with replacement, and a large study population—the probability of being sampled is proportional to the number of ties that each respondent has. As a result, the inverse of the self-reported number of ties can be used as a sampling weight. Because of the attractiveness of its hypothetical sampling properties, the RDS approach has been widely used by researchers to study hidden populations such as intravenous drug users, undocumented immigrants, jazz musicians, etc. (cf. reviews by Abdul-Qadar et al., 2006; Ramirez-Valles et al., 2005).

Despite the popularity of RDS, recent research indicates that the accuracy of the approach is sensitive to the assumptions it makes about the social network of the underlying population (Goel and Salganick 2010; Gile and Handcock 2010; Lu et al. 2011). A fundamental problem with random walk based approaches in general, and RDS in particular, is that they have a higher sampling variance than simple random sampling. Whereas the sampling variance of a simple random sample is inversely related to sample size, the precision of RDS and random walks is a function of both sample size and the structure of the network: Bassetti and Diaconis (2006) and Goel and Salganick (2009) show that the variance of a random walk is mathematically related to the eigenvalue decomposition of the network—i.e., its underlying structure—and the degree to which the dependent variable maps onto the corresponding eigenvectors. In networks with high degrees of clustering, RDS has a tendency to get stuck in segregated components of the network, which can result in inaccurate samples. In certain networks, the variance of a RDS sample could be unacceptably high, but the researcher, with no knowledge of the network that he or she was sampling from, would be unable to distinguish between accurate or inaccurate results. Given this fundamental uncertainty, the ability to infer population-level prevalence is questionable. Indeed, Goel and Salganik (2010) find very high levels of sampling variance for RDS in simulated sampling using real network data.

In this paper, we propose a new approach, "Network Sampling with Memory" (NSM), which builds on recent developments in the mathematics and computer science literature on sampling large internet-based networks that modify the basic random walk approach by incorporating information on the local topography of the network, information on recently sampled cases, or multiple dependent random walks to improve sampling efficiency (Ribiero and Towsley 2010; Kharusi 2008; Ikeda et al. 2009; Alon et al 2008; Cooper, Frieze, and Radzik 2009; Berenbrink et al. 2010; Avin and Krishnanmanchari 2008). As we discuss in the paper, NSM improves the efficiency of sampling from a network by collecting network data from respondents as part of the survey, which is used to gradually reveal the list of population members. As the sample progresses, the list of people who have been nominated in the survey begins to resemble the full list of population members, which allows NSM to approximate the process of simple random sampling.

NSM samples from the network list by using two modes of sampling, a "List" mode and a "Search" mode. The List mode is very simple: the next person to be interviewed is selected by sampling with replacement from the list of all people who have been nominated. When a new person is added to the list (i.e., when someone is nominated for the first time), he or she is sampled at the current cumulative sampling rate of the sample to ensure that all individuals who have been nominated have the same probability of being sampled (otherwise, cases which were sampled first would be over-represented).

The Search mode attempts to speed up the process of exploring large or complicated networks by identifying respondents who appear to be "bridge nodes" to unexplored clusters of the network. We define bridge nodes as respondents who have a high proportion of friends that have not been nominated by anyone else in the sample. The Search mode gives priority to the unsampled friends of these bridge nodes, which pushes the sample towards new and unexplored regions of the network. NSM combines the Search and List modes together, using the Search mode first until the network has been explored, and then switching over to the List mode.

In the analysis section of the paper, we test the relative precision and accuracy of NSM compared to RDS and a random walk by conducting simulated sampling on 162 actual social networks, 62 school networks from Add Health and 100 large university networks from Facebook. The results indicate that NSM results in high levels of sampling precision (with sampling variances close to those of Simple Random Sampling), even on networks characterized by high degrees of clustering and homophily. Overall, we argue that NSM will prove useful for researchers in a variety of fields where collecting network data on social interaction is substantively interesting and where it is desirable to use the network data to improve the efficiency and precision of the sampling process.

2. Background

Referral Based Sampling of Hidden Populations

Hidden populations are an important area of study for understanding sociological topics such as the dynamics of disease transmission, immigration, the underground economy, or homelessness. They are also a prime area for referral based sampling approaches because they are demonstrably challenging to survey with other sampling approaches (Kendall et al., 2008). Recently, the use of referral based sampling methods in the study of hidden populations has been accompanied by important theoretical and statistical developments (Heckathorn 1997, 2002, 2007; Salganick and Heckathorn 2004; Magnani et al. 2005; McKnight et al. 2006). The key innovation, respondent driven sampling (RDS), builds upon the idea of taking a "random walk" on a network and has two properties which make it useful for sampling hidden populations. The first has to do with the recruitment process. In

RDS, respondents are paid for their participation in the study and incentivized to recruit others. Such incentives take the form of coupons, a fixed number of which are given to members of the current wave of respondents. New respondents must present a coupon obtained from a prior respondent to enter the study. The use of coupons allows the tracking of recruitment chains, payment of respondents for recruiting new members, and the maintenance of a high degree of confidentiality with minimal risk for participants and researchers (Yeka et al. 2006).

The second important property of RDS for sampling purposes is a statistical innovation. Respondents are asked to estimate how many people they know in the hidden population, and the inverse of each person's estimate is then used as a weight that allows researchers to discount which respondents were most likely to be sampled. It has been shown (Salganick and Heckathorn 2004; Volz and Heckathorn 2008; Goel and Salganick 2009) that such weighting is a mathematically appropriate technique for characterizing recruitment probabilities in a sample from a network if respondents are able to estimate their popularity with reasonable accuracy (cf. Neely 2009 for how errors in this estimation influence RDS estimates).

RDS has become quite popular: Johnston et al. (2006) cite 128 applications in countries outside of the United States and the Centers for Disease Control and Prevention have adopted the method on a large scale to understand the dynamics of HIV in the American injection drug user community (Abdul-Qadar et al., 2006). As a measure of the level of endorsement of RDS by funding agencies, a search of recent NIH grants indicates that at least $100 million in funding has been awarded to projects with “respondent driven sampling” (or “respondent-driven sampling”) in the keywords (NIH 2011). Overall, the recent popularity of RDS has made it the “gold-standard” against which competing proposals to understand hidden and hard to reach populations (e.g., time-location sampling, targeted sampling) are evaluated.

However, while methods to sample hidden and hard to reach populations have overcome many of the challenges facing referral based sampling, they have generated a new set of criticisms. Most critiques have focused on the unrealistic nature of the method's assumptions (Gile and Handcock 2010; Goel and Salganick 2009), though new approaches show that modifications to the method can make it less dependent on some of these (e.g., Gile 2011; Neely 2009). While these are important problems, the most damaging critique is that the method is highly ineffective, even when assumptions are met. The problem is that referral based approaches, while providing unbiased mean estimates, yield far more variable results than traditional survey sampling (Goel and Salganick 2009). This necessitates collecting larger sample sizes with referral based sampling than would be needed from random sampling to maintain the same level of statistical power. Perhaps the most important evidence of this effect comes from an innovative study which simulated RDS samples on real world social networks (Goel and Salganick 2010). There, the authors found that results from RDS were up to 70 times more variable than would be found had comparably sized simple random samples been taken (also see Lu et al. 2011). Such simulation evidence accords with other studies. For instance, repeated RDS samples of the same population have shown substantial variability in socially salient traits (Ma et al., 2007), and the confidence intervals of RDS samples from known populations do not always not contain the true population values and can often be too wide to be of use (Wejnert 2009; Wejnert and Heckathorn 2008; see discussion in Goel and Salganick 2010). Of course, a large number of surveys conducted by those with deep familiarity with the populations at hand have found that RDS results are often quite reasonable (cf., Malekinjad et al. 2008). However, as demonstrated by the simulation studies, there is the danger that any given RDS estimates may be far from the true population value.

Network Based Sampling in Non-Human Populations

Other fields have also been active in developing new network based sampling methods. As mentioned above, a large literature has focused on the properties of random walks on graphs. Bassetti and Diaconis (2006) derive the variance of a random walk based on a spectral decomposition of the network structure, and Goel and Salganick (2009) relate this to the variance of RDS estimates. More generally, this literature (cf. Lovasz 1993 for a review) is focused on the mixing time—the number of steps until the sampling probabilities converge--and cover time—the time needed until all nodes are sampled—of random walks on graphs. There is a direct connection between the accuracy of a sample derived from a random walk and the mixing and cover time because the more steps that are needed, on average, to reach the equilibrium sampling distribution or sample all of the cases, the less accurate the sample will be compared to simple random sampling. In computer science, a rapidly expanding literature focuses on practical aspects of improving the speed and accuracy of random walks to sample from the internet, but other examples include sampling from electrical networks and map exploration (e.g., Gasieniec and Radzik 2008).

Avin and Krishnanmanchari (2008) propose a modification to the random walk strategy where at each step, neighbors of the current node are queried and the one with the fewest visits from the random walk is selected as the next step. At each step, this pushes the random walk back to nodes whose neighbors are least likely to have been explored. In a related approach, Berenbrink et al., (2010) and Yu and Newman (2008) allow the random walk to explore the network in the vicinity of the current case before taking the next step. This collection of local information helps the walker know whether it is in a well-explored cluster or a new area. Alternatively, one can use multiple networks derived from the multiplex connections that link people to move between different parts of the network: Gjoka, Butts, Kurant and Markopoulou (2010) show that problems of sampling between segregated or disconnected components of large online social networks can be reduced by using multiple types of connections between users (i.e., group or event participation) even if there is no sample frame. If there is some means of obtaining a list of nodes in the network, Thompson (2006a), Ribeiro, Wang, and Towsely (2010), and Avrachenkov, Ribeiro, and Towsley (2010) propose an adaptive random walk, where a random walk is supplemented by periodic jumps to randomly selected parts of the network; a similar method was pioneered by Google to rank web-pages (Page and Brin 1988). Such approaches have been shown to reduce the number of cases needed to sample the network effectively.

A simple modification of the basic random walk is to provide a temporary restriction on nodes that can be visited. Kharusi (2008) proposes a random walk with memory—a "forward random walk" –based on preventing the walk from returning to the k most recently visited cases, and finds that cover times decrease dramatically compared to simple random walks in networks with high diameter (the distance between the most separated nodes of the network). Similarly, Alon et al., (2007) find an improvement in the mixing rate by not allowing the walk to go back to the most recently sampled case.

A slightly different approach uses information about a node's neighbors to tilt the selection probabilities in favor of visiting certain types of nodes. Ikeda et al., (2003, 2009) show that the cover time of a random walk can be reduced by modifying the transition matrix to oversample low-degree neighbors of the current node. Ikeda et al., (2009), for example, show that, under certain conditions, the cover time for a random walk can be reduced from $o(n^3)$ to $o(n^2)$ (where n is the number of nodes in the network) by using a biased transition matrix that weights the probability of selection of a node proportional to the inverse square root of its degree.

Finally, an alternative approach is to use multiple independent random walks rather than a single random walk. In human populations, this is roughly comparable to starting random walks from multiple seeds. The use of multiple random walks has been found to reduce cover times (Alon, et al., 2008; Cooper, Frieze, and Radzik 2009; Elasser and Sauerwald 2010), however it is still possible for each individual random walk to get stuck in sub-graphs of the overall network. Ribeiro and Towsley (2010) propose an approach based upon multiple dependent random walks, where the multiple walks share the same sampling process, and show that this approach has a faster mixing time than multiple independent walks and single random walks in simulations on several large social-media networks[2].

An important difference between this recent literature on improved random walk algorithms in large networks such as the internet, mobile networks, or electrical grids, and sampling from human populations is the cost of an interview. In the former, the cost of each interview —i.e. querying a node of the network—is small, often a matter of nano-seconds of computer time. Instead, the size of the underlying networks is the key issue, and computers' memory restrictions place limits on the collection and processing of topographic network information for large sample sizes. As a result, most of the articles discussed focus on how little extra information and memory is needed per case to improve the sampling efficiency. In contrast, in sampling from human populations, the cost in terms of time and effort in obtaining and conducting face-to-face or telephone interviews represents the key constraint, and as a result the overall sample size may be small. This is important because the efficacy of sampling from human social networks cannot rely on surveying a large number of people; indeed, the approaches used in computer science and elsewhere to reduce cover and mixing times seem like a good starting place for devising a new approach to sampling from social networks that performs well in finite samples.

In the next section, we build on this recent literature in computer science to propose a method for sampling from a network that uses the collection of social network data to retain a “memory” of the network and improve the efficiency of the sampling process. At the end of the paper, we argue, based on classic (e.g., Burt 1984, McCallister and Fischer 1978) and contemporary work on interviewing people about networks (e.g., Sandberg et al., 2008), that collecting data on a large number of respondents' alters is feasible in sampling from human populations, and that the gains in sampling accuracy and efficiency outweigh the added cost of the collection of network data.

3. Method

A key consideration of our approach is that we want a method that has desirable asymptotic and finite sample properties: not only do we want a method that is asymptotically unbiased (like random walk based approaches such as RDS when its assumptions are met), but we want the sampling variance and mean absolute bias to decline rapidly in the kinds of finite sample sizes actually collected in practice when sampling from human populations. In this section we outline our proposed solution, which consists of two different modes of sampling: a “List” mode that guarantees asymptotic unbiased estimates and a “Search” mode that is designed to work in tandem with the List mode to improve the precision of estimates in finite samples. In order to clarify the presentation of our method, Table 1 provides a glossary of terms used in this section of the paper. In addition, in Section B.2

[2]This discussion offers some intuition about RDS's efficacy as it is used in practice (when its assumptions are not always met [Gile and Handcock 2010]). The fact that RDS is conducted without replacement adds an element of memory to the random walk process, which, based on these results, would suggest that it may prove more efficient in without-replacement practice than with-replacement theory (although see Gile 2011). Additionally, the fact that RDS in practice starts from multiple seed individuals may also improve its efficiency as per Ribeiro and Towsley (2010).

below we provide an example of a sample drawn using the List mode to illustrate the basic mechanics that are involved.

First, imagine that we are sampling from a population, "A", which consists of N individuals ("nodes") who are connected through a social network. We assume that there is no sampling frame—i.e., there is no list of the N members of the population which would allow for simple random sampling. The basic intuition of our approach is quite simple: we use the sampling process to uncover the list of population members, and then sample from that list with replacement[3]. We call our approach "Network Sampling with Memory" (NSM) because we use name and demographic data on respondents' network members to "remember" the network; however, other types of data about network alters could be collected (e.g., Dombrowski et al. (2011).

The survey begins with an initial "seed" respondent.[4] Each step in the sampling process consists of an interview with a respondent which—in addition to the substantive survey questions—asks for a network roster of the respondent's friends or contacts who are members of population A. In each step, new network members are added to list L, which consists of all members of A who have been nominated by a respondent.[5] L also keeps track of the cumulative number of nominations each node has received. Let S indicate the current "step" of the sample (i.e., the number of interviews that have been completed), and let L_S indicate the number of people on the network roster L at step S.

List and Search sampling modes

NSM consists of two sampling modes, "List" and "Search". It is easiest to describe these sampling modes by beginning with what we call the "naïve list" mode of sampling from the network roster.

(3.A) Naïve list mode—Sample with replacement from the accumulated network roster L. The probability that a nominated node is sampled at step S is

$$p(sampled)=\frac{1}{L_S} \quad (1)$$

where L_s is the current size of the revealed network.

The problem with the naïve list mode is that nodes that are nominated early in the sample will be oversampled compared to nodes that are nominated later because the early nodes are eligible to be sampled for more steps. As early nodes are oversampled, this slows down the process of exploring the network.

(3.B) List mode and "even" sampling—To prevent the problem of oversampling nodes that are nominated early in the sample, the List mode ensures equal sampling probabilities by sampling new nodes—i.e., people who are added to L when they are nominated for the first time—at the same cumulative sampling rate as previously nominated nodes. This "even sampling" starts once the revealed network L reaches a certain size: Let L_s indicate the current size of the network after S interviews and let P1 equal the proportion of nominated nodes (L) that have been nominated once and only once so far in the sample.

[3]In our proposed approach, sampling with replacement is achieved by keeping track of how many times individuals are sampled and using this as a weight. No repeat interviews are actually conducted.
[4]Multiple seeds can be used; in general this is most effective if they are from different parts of the network.
[5]L also includes the initial seed respondent(s).

In the simulations that we conducted for this paper, the even sampling of new nodes begins when either:

a. $L_S \geq 200$, or

b. $S > 50$ and $P1 < .2$ [this ensures that the List mode is turned on for small networks]

Let S1 indicate the step at which even sampling begins. When S>S1, nodes that are nominated for the first time in Step S are subjected to the same sampling rates experienced by previously nominated nodes from Step S1 to S. A node that is interviewed in this way is called a "catch-up" interview, because it was interviewed while its cumulative sampling rate was "catching up" to the rest of the sample. In the next subsection, we describe this process more precisely:

(3.B.1) Even sampling and "Catch-up" Interviews: Let Q be a node that is nominated for the first time in Step S, and let U_Z be a uniform random number between 0 and 1, and let $Z = \{S1, S1+1, \ldots S\}$ indicate the steps from S1 to the current step S, excluding steps that were "catch-up" interviews themselves (see Steps 80-82 in Table 2 for an example). We "retroactively" sample node Q in Step Z if $U_z < \frac{1}{E_z}$, where E_Z is the number of nodes that were eligible to be sampled in Step Z (see the definition of the set of eligible nodes on the next page), repeating this process for all steps Z from S1 to S. The benefit of doing this is that it reduces the advantage that earlier nodes have of being oversampled compared to nodes that are added to L later in the sample.

The reason we choose $L_S \geq 200$ as the starting point for the sampling of new nodes is that if even sampling begins when the network size is too small the sampling rate of catch-up interviews quickly becomes very large. Based on the results of simulated sampling discussed below on networks ranging in size from 110 to 16,278 nodes, $L_S \geq 200$ worked very well, although other thresholds above or below 200 could also work. In general, for networks that are larger than the ones that we evaluate in this paper, L_S should be set higher to ensure that the ESR doesn't get too large too quickly.[6] Moreover, in very large networks, the Hybrid approach (which combines the Search and List modes, discussed below) works the best, as demonstrated by the results in Table 8 below.

In addition to ensuring that all nodes sampled after S1 are sampled at the same rate, we also want to equalize the cumulative sampling rate of nodes that were sampled before S1. After even sampling is turned on in step S1, we do this by excluding nodes that have high cumulative sampling probabilities until the rest of the sample catches up to them. Define each node's cumulative sampling rate as:

$$\text{"cumulative sampling rate"}(CSR_j) = \sum_{S=a+1}^{S=b} \frac{1}{E_S}, \text{if } j \in \text{eligible nodes}_S \quad (2)$$

Where a is the step that node j was first nominated, b is the current step, "j ∈ eligible nodes$_s$" indicates that node j was eligible to be sampled at Step S, and E_S is the number of nominated nodes eligible to be sampled in Step S. The sum of the step by step sampling rates since the initiation of even sampling is the "even sampling rate" (ESR) defined in

[6] One way to resolve this is to not turn on even sampling until P1 (the proportion of nominated nodes that have only 1 nomination) is less than 0.4, which ensures that large networks are explored before even sampling is turned on. See Equation 5 for a discussion of estimating the network size using P1.

Equation 3 (for an example of the calculation of the ESR see Table 2 below). We exclude "catch-up" interviews in the calculation of the ESR:

$$\text{“even sampling rate”}(ESR) = \sum_{S=S1}^{S=b} \frac{1}{E_S}, S \notin \text{“catch-up interview”} \quad (3)$$

In order to even out the sampling rate of early nodes, after Step S1 we temporarily exclude any node j from the set of nodes that are eligible to be sampled in Step S if $CSR_j > ESR$. These nodes are not eligible for sampling until the ESR increases. If there are fewer than 100 nodes where $CSR_j \leq ESR$ then the list of eligible nodes at Step S consists of the 100 nodes with the lowest CSR_j.

Definition of E_S: Based on Equations 2 and 3, we define E_S as the set of nodes eligible to be sampled by the List node in Step S, which consists of all nominated nodes that have $CSR_j \leq ESR$ unless the total number of nominated nodes with $CSR_j \leq ESR$ is less than 100, in which case E_S consists of the 100 nodes with the lowest CSR_j.

In summary, we list the sequence of steps involved in the List mode:

B1. Randomly select the next node to interview from L, sampling with replacement. If "even sampling" has been turned on ($S \geq S1$) , then only nodes $\in E_s$ are eligible to be sampled. Let W indicate the node that is selected in Step B1.

B2. Interview node W, and update the network roster L.

B3. If even sampling has been turned on, use the "catch-up" procedure described in Section 3.B.1 to see if any newly nominated nodes are randomly selected to be interviewed. If a new node is selected, this newly nominated node now becomes node W. Go to Step B2.

B4. Return to step B1 until the desired sample size has been reached.

(3.B.2) Example of sampling using the List mode: In order to illustrate how the List mode works in practice, Table 2 describes the sampling process for a single sample drawn from the 400-node test network, which is shown in Figure 2 and discussed in more detail below. Table 2 consists of 10 columns of information about each step of the sample. Column 1 shows the step number, and Column 10 shows the ID of the node that was interviewed in that step. The current size of the revealed network, L (Column 2), is updated after the interview and shown in the next step. For example, the sample begins with Step 0, which is the initial seed (Node # 92). Node #92 has 7 friends in the network, so the size of the revealed network as we begin Step 1 is 8 (the seed node + 7 friends), and the effective sampling rate for Step 1 is $\frac{1}{8} = 0.125$ (Column 5). By Step 5, a total of 25 different nodes have been nominated (L=25), and the sampling rate for Step 5 is 0.04.

Columns 6-9 show the cumulative sampling rate of four of the nodes in the network (ID #s 1,101,201, and 301). Node 1 is nominated for the first time by Node #23 in Step 4, and is eligible to be sampled for the first time in Step 5 (hence its cumulative sampling rate in Step 5 is 0.04). Node 301 is eligible to be sampled for the first time in Step 7. Note that because Nodes 1 and 301 were nominated early in the sample, they accumulate a higher CSR compared to Nodes 101 and 201. By Step 39, for example, the cumulative sampling rate for Node 1 is 0.323978, while Nodes 101 and 201 have yet to be nominated by any of the nodes that have been interviewed so far.

In Step 41, "even sampling" is turned on because the revealed network size is greater than 200. Now, all nodes that are nominated for the first time will be sampled at the "even sampling rate" (ESR) defined above in Equation 3 . After the start of even sampling, only nodes with a cumulative sampling rate (Columns 6-9) less than or equal to the ESR (Column 4) are eligible for sampling. In this example, Nodes 1 and 301 are temporarily excluded from sampling starting in Step 41. The only exception to this is right after the start of even sampling if there are fewer than 100 nodes with a CSR<ESR. In this case, the 100 nodes with the lowest CSR are eligible to be sampled. [7]

By Step 79, the number of revealed nodes is 357, and the ESR has risen to 0.205092, which means that newly nominated nodes have about a 20% chance of being interviewed. Based on random draws using the process described in Section 3.B.1, Steps 80-82 resulted in "catch-up" interviews where nodes that were nominated for the first time were randomly selected to be interviewed themselves. The benefit of this procedure, in addition to ensuring a gradual even sampling of the network, is that the sample gets pushed towards newly discovered parts of the network.

Finally, in Step 251 all of the nodes in the network have been nominated, and the cumulative sampling rate of all nodes is equal to the ESR, or, for nodes that were temporarily excluded from sampling eligibility, slightly less than the ESR. In Table 2, Nodes 101 and 201 have a CSR equal to the ESR (0.641112), while the two nodes that were nominated early in the sample are slightly lower (0.63988 for Node 1 and 0.640566 for Node 301).[8] The key point is that in Step 251 the entire network has been revealed by the survey and all nodes have been "evenly" sampled—just as if they had been sampled with simple random sampling—even though no sampling frame was available at the start of the sampling process. In the next section we discuss sampling weights, and later, in Section 5, we discuss diagnostics for convergence to simple random sampling when the overall size of the network is unknown.

(3.B.3) Nomination probability weights for the List mode: If the network has not been completely explored (if not all of the members of population A have been nominated) then high-degree nodes are more likely to have been nominated than low-degree nodes simply because they know more people. We estimate the nomination probability based on degree, assuming a random sampling of nodes. For example, in a 500 node network, the probability that node i of degree d_i will be nominated by a respondent in any singe interview is $\frac{d_i}{500}$. Similarly, the cumulative nomination probability for node i after S interviews is:

$$p\_\,\mathrm{nom}_i \cong 1 - \left(1 - \frac{d_i}{N}\right)^S, \quad (4)$$

where N is the number of nodes in the network.

Because N is unknown when we are sampling unknown networks, we estimate N using the capture recapture method (see Dombrowski et al. 2011 for a discussion of the capture-recapture method from social networks and a data collection strategy similar to that advocated here):

[7]This is because, at the start of even sampling, all nodes that have already been nominated will have CSR>ESR. Using the 100 nodes with the lowest CSR is a way to gradually "wean" the sample away from nodes with high CSR's.

[8]The reason for this is that after the start of even sampling, Nodes 1 and 301 were not eligible for sampling again until their CSR was less than or equal to the current ESR as described in Section 3.B.1. Thus they will be within the sampling rate of 1 step of the ESR.

$$\widehat{N}=\frac{L_s}{1-P1} \quad (5)$$

where L is the current number of nominated nodes, and P1 is the proportion of nominated nodes that have been nominated 1 time.[9] See Equations 9 and 10 below for the use of composite weights combining p_nom and the CSR to calculate the sample mean. Note that as the sample size S becomes large, the estimated probability of being nominated from Equation 4 will go to 1 for all nodes.

(3.C) Search Mode—As we will see in the results section, the List mode does a very good job by itself sampling from the Add Health and Facebook networks with a sample size of 500. However, in larger networks the precision of the samples decreases because it takes longer to explore the network.[10] In this section we add the "Search" mode, which pushes the sampling process toward unexplored parts of the network. Note that the search mode employed here is not the same thing as other searching modes commonly used in graph theory such as breadth- or depth-first search.[11] The Search mode is not designed to be used by itself, but only in combination with the List mode. We call the combination of the Search and List modes the "Hybrid" approach.

The intuition behind the search mode is based on the idea of a "bridge node" or a "structural hole" (Burt 1995) that connects two clusters of the network. Imagine that Node Y_1 is in a cluster of the network Y that has not been sampled, except for Node Y_1. As depicted in Figure 1, Node Y_1 links across a bridge to cluster X, which has been extensively sampled. Y1 's friends on the Y-side of the bridge have been nominated (by Y1)—so they appear on the network roster L—but they have not been sampled or nominated by any other respondent. These nodes represent the leading edge of an unexplored area of the network. The search mode is designed to sample from the friends of nodes like Y1 that are on the other side of a bridge to an undersampled cluster of the network.

Let d_j indicate the degree of Node j, which is the number of friends that node j has in the network. Let C_j indicate the number of j 's friends who are unsampled and have been nominated once and only once so far in the survey. For example, in Figure 1, the friends of Y1 on the Y side of the bridge (nodes Y8, Y5, and Y2) have been nominated only by Y1 whereas all nodes on the X side of the bridge that are unsampled (nodes X1, X3, and X8) have all been nominated at least two times. Define the proportion of j's friends that have been nominated 1 time as $p1_j=\frac{c_j}{d_j}$. $p1_j$ is calculated for all nodes in the network that have been sampled so far. In addition to $p1_j$ we calculate P1 (defined above as the overall proportion of nodes L that have only 1 nomination). Finally, we designate A1 as the minimum value of P1 that the Search mode will be used. When P1 gets low, this means that most of the nodes have been nominated by at least two respondents, which suggests that the network is fairly well explored and we can use the List mode without sacrificing much in the way of sampling efficiency.

The search mode consists of the following steps:

[9]In the calculation of P1, any node that has been sampled is counted as a node that has more than 1 nomination.
[10]This can be seen by looking at the effect of network size (log nodes) on the design effect for the List mode in Table 8 below.
[11]NSM's search mode skips to other parts of the network based on a probabilistic model of where the bridge nodes are most likely to be. In contrast, breadth-first search explores all of the nodes neighboring the currently sampled node before moving down the chain, while depth-first search moves down the chain of connections as fast as possible before visiting nodes that have already been nominated.

C1. (a) Calculate $p1_j$ for all sampled nodes, and (b) Calculate P1, the proportion of nominated nodes that have been nominated only 1 time.

C2. If P1, the overall proportion of nodes $\in L$ with 1 nomination, is greater than A1—the minimum threshold for the search mode--use the search mode. If $P1 \leq A1$, default to the List mode. This ensures that the List mode is used when much of the network has been explored. In the sampling simulations below, we set A1 to 0.4, but we tried other settings as well. Note that the reason we choose to switch to the List mode at a specific point (i.e. when $P1 \leq A1$), rather than always use the Search mode or "blend" the two modes together, is that the asymptotic properties of the Search mode will depend on the structure of the particular network being sampled. In contrast, the List mode converges to SRS as L approaches the true size of the network.

C3. We want to identify bridge nodes (such as Y1 in Figure 1) and then interview their unsampled friends who have only 1 nomination (i.e., Nodes Y2, Y5, and Y8 in Figure 1). To do this, we are going to identify the 5 nodes that are the most likely to be bridges, and then sample one of their unsampled network ties. In this step, we calculate the probability that a particular node is a bridge node. To begin with, we argue that bridge nodes are nodes that are the largest positive outliers in terms of their observed value of C_j–i.e., nodes that have unexpectedly large values of C_j given their degree and the current value of P1. First, we estimate the probability of that node j has C_j or more friends with only 1 nomination (conditional on node j being sampled), given d_j and P1, using the binomial distribution:

$$p(X \geq c_j | j \in L) \cong \sum_{i=c_j}^{d_j} \binom{d_j}{c_j} P1^{c_j} (1 - P1)^{d_j - c_j} \quad (6)$$

where, as noted above, P1 is the overall proportion of nodes $\in L$ with 1 nomination so far in the sample.

We estimate how "unexpected" the current value of C_j is for node j by multiplying the binomial tail probability in Equation 6 by the estimated probability that node j has been nominated by step S of the sample (from Equation 4 above)[12]:

$$p(X \geq c_j) \cong p\_\text{nom}_j \times p(X \geq c_j | j \in L) \quad (7)$$

Finally, we define the relative probability of node j being a bridge node as:

$$p(\text{bridge}) = 1 - p(X \geq c_j) \quad (7b)$$

We choose the 5 nodes with the highest values of $p(\text{bridge}_j)$ as the set of most likely bridge nodes.[13]

[12]A simpler method is just to choose among the nodes with the highest vales of $p1_j$, but this gives an advantage to nodes with low degree, as it is easier to obtain higher values of $p1_j$ just by chance. In addition, for simplicity we ignore the probability that j is sampled conditional on being nominated, as this will be the same for all nominated nodes after the initiation of even sampling, and the goal in Equation 7 is a relative measure of the likelihood of a bridge node.

[13]We also tried choosing among the top 2 and top 10 bridge nodes, and both of these alternatives seemed to work equally well. In addition, to ensure that these really are what we mean by "bridge nodes", only nodes with $p1_j \geq 0.4$ are eligible for consideration in Equation 7.

C4. Randomly select one of the top 5 bridge nodes from step C3, weighting the 5 nodes based on their values of $p(\text{bridge}_j)$ from Equation 7b. Let B indicate the selected bridge node.

C5. Let B1 indicate the set of B's friends that are unsampled and have 1 and only 1 nomination so far in the sample. Randomly select one node from B1 as the next interview.

C6. In order to incorporate the cumulative sampling rate from the Search mode with the even sampling aspect of the List mode, we calculate the sampling probability of each eligible friend F from any node j that was among the top 5 bridge nodes in Step C4. For the eligible friends of one of the top 5 bridge nodes j:

$$p(\text{node F sampled with Search}) = \frac{1}{c_j} \frac{p(bridge_j)}{\sum_{k=1}^{5} p(bridge_k)} \quad (8)$$

where C_j is how many eligible friends node j has, and $\frac{p(bridge_j)}{\sum_{k=1}^{5} p(bridge_k)}$ is the probability that node j is chosen from among the top 5 possible bridge nodes. This sampling rate in Equation 8 is then added to the CSR for node F (see Equation 2 above in Section 3.B) each time that node F is eligible to be sampled via the Search mode. As discussed above in Section 3.B, the List mode ensures the uniform sampling of all nominated nodes as the sample progresses by temporarily excluding any node that has a higher cumulative sampling rate than the even sampling rate defined in Equation 3. As a result, any oversampling of bridge nodes at the beginning of a sample will gradually disappear once the network is explored and the sampling process shifts to the List mode (i.e. after P1<A1 in Step C2).

C7. After calculating the Search mode sampling rates for the eligible nodes in Step C6, we update the network rosters and return to Step C1 of the Search mode, using the Search mode when $P1 > A1$ and the List mode when $P1 \leq A1$ in Step C2. Repeat this process until the desired sample size is reached (see Section 5 for a discussion of an ideal stopping point based on convergence to simple random sampling).

The Hybrid approach: What we are calling the "Hybrid" approach consists of the combination of the List and Search modes. The sample starts out in Search mode, but it defaults to the List mode when the network has been reasonably well explored, as indicated by the threshold A1 in steps C2 and C7 above.

Sample Mean: For all of the NSM variants (naïve List, List, and Hybrid), the sample mean from a single sample of S cases is calculated as the weighted mean $\bar{m}$:

$$\widehat{m} = \sum_{s=1}^{s=S} Y_s \cdot \left(\frac{w_s}{\sum_{k=1}^{k=S} w_k} \right), \quad (9)$$

where s refers to the node that was sampled in the s^{th} step of the sample and the case specific weight, w_s is defined as:

$$w_s = \frac{1}{CSR_s \cdot (p\_\ nom_s)}; \quad (10)$$

and CSR and p_nom are defined above (see Equations 2, 4, and 9)

4. Simulated Sampling

In this section, we test the performance of NSM (hybrid, list, and naïve list), RDS, and random walks (RW) along a variety of dimensions by using simulated sampling on a test network and 162 real social networks. After these simulated tests, Section 5 considers the question of convergence to simple random sampling, and Section 6 addresses the broader question of the feasibility of the method. Before we begin, we note that the complete computer code needed to replicate our findings (in Stata's Mata language) with the test network described below, as well as examples of alternative approaches, is available on the first author's web page (www.-----.-------).

In our simulated sampling, we use RDS and RW in order to provide a meaningful comparison to evaluate the efficiency of NSM. For RW, we start with a single, randomly selected seed, and all subsequent cases are drawn randomly from among the friends of the current case, sampling with replacement. RDS is conducted similar to RW, with the next respondent selected as a referral from among the friends of the current respondent. In these simulations, the difference between RW and RDS is that RDS allows for multiple referrals from each respondent[14]. We follow the simulated sampling approach of Goel and Salganick (2010) the probabilities of referring 0, 1, 2, or 3 new respondents in RDS are $\frac{1}{3}, \frac{1}{6}, \frac{1}{6}, \text{and } \frac{1}{3}$ respectively.

For each network, we use the different methods to collect 500 samples of 500 cases, starting at a randomly selected node and sampling with replacement. Though, in practice, the RDS methodology does not sample with replacement, we simulate our RDS samples with replacement because RDS has been shown to be biased when conducted without replacement as the proportion of the population sampled gets large (Gile and Handcock 2010)[15]; in this sense, sampling with replacement is necessary to allow us to observe the asymptotic properties of RDS. Further, other prominent simulation work evaluating RDS has also been conducted with replacement (Goel and Salganick 2010). To maintain consistency with those results and to provide a fair comparison between our RDS, RW and NSM results we therefore conduct all of the simulations with replacement.

Test statistics

We compute 4 different test statistics to evaluate the competing methods: average bias (Bias), mean absolute bias (MAB), standard deviation (SD), and the design effect (DE). First, define the bias (or error) of a single sample i (of size S cases) collected from network j as:

[14]We weight both the RW and RDS results by their inverse degree. For the binary outcomes we consider here (Y), we calculate the estimated proportion of Y=1 in our random walks as $\widehat{p}_{rw}(Y=1) = \frac{\sum_j^s y_j/d_j}{\sum_j^s 1/d_j}$, where y_j is the value of variable y for individual j (either 0 or 1), d_j is individual j's degree, and S is the number of sampled cases. Some have called a random walk weighted in this fashion a reweighted random walk (cf. Gjoka et al. 2010). This estimator is the RDS2 estimator in current use, so we also apply it to our RDS samples.

[15]RW approaches have not been implemented in the field to our knowledge, but, even if they were, it is a challenge to imagine that they would be conducted with replacement in a human population.

$$bias_{ij} = \hat{m}_{ij} - m_j \quad (11)$$

where m_j is the population mean of the dependent variable in network j, and m_{ij} is the estimate of m_j in sample i. The average bias in network j (for a particular method), is:

$$\text{average bias}_j = \frac{\sum_{i=1}^{N} \text{bias}_{ij}}{N} \quad (12)$$

where N is the number of different samples collected.

Because the average bias for any particular network could be positive or negative, we want to take the absolute value of the average bias of each network when summarizing it across different networks. Define the summary measure of average bias as:

$$\text{average bias} = \frac{1}{J} \sum_{j=1}^{J} \left| \frac{\sum_{i=1}^{N} bias_{ij}}{N} \right| \quad (13)$$

where J is the number of different networks in the study.

The mean absolute bias—the measure of the average magnitude of the error for each sample —is:

$$\text{MAB}_j = \frac{\sum_{i=1}^{N} |\text{bias}_{ij}|}{N} \quad (14)$$

Note that the difference between the average bias and the MAB as defined in Equations 13 and 14 is where the absolute value is taken—before or after averaging the bias across the N samples collected for each network. The variance of the bias, which we refer to as the "sampling variance" throughout the paper (Groves et. al. 2009:58) is:

$$\text{sampling variance} = \sigma_j^2 = \frac{\sum_{i=1}^{N} (\hat{m}_{ij} - m_j)^2}{N}, (\text{and } SD_j = \sigma_j) . \quad (15)$$

(and $SD_j = \sigma_j$).

In order to compare the sampling variance to simple random sampling, we also calculate the "design effect" which is:

$$DE_j = \frac{\sigma_j^2}{\sigma_{SRSj}^2} \quad (16)$$

where σ_{SRSj}^2 is the sampling variance of simple random sampling with a sample size of S cases.

Test network results

Before we sample from the Add Health and Facebook networks, we first try the methods using an artificial test network. The reason we do this is that this network is very easy to visualize, would be simple to reproduce, and it illustrates key differences in the performance of the competing methods. The test network consists of 400 nodes divided into 4 clusters of 100 nodes each. Each cluster is characterized by a different degree (i.e., the number of friends that each person has). In Figure 2, all of the nodes have 1 tie to someone in a different cluster, with the remainder of their social ties within their cluster. The four clusters A (bottom right cluster in Figure 2) , B (bottom left), C (center), and D (top) have degree 7,12, 17, and 22 respectively. This is a highly connected network, since every node has a tie to someone else in a different cluster.

To evaluate our simulated sampling from the test network, we examine the overall average degree in the network, which is 14.5, as the dependent variable. Because degree is clustered within the network, this should pose a problem for RW and RDS based approaches. Likewise, because the NSM, RDS, and RW approaches all rely on degree as either a means of exploring the network (NSM) or as a post-sample weight, average degree is a measure that we suspect will be challenging for all three approaches. For the test network, we collected 1000 samples of 1000 cases for each method, sampling with replacement.

Figure 3 shows the results for each method with sample sizes varying from 50-1000 cases, and Table 3 presents the same results in table format for sample sizes of 250, 500, and 1,000 cases. Because of space constraints in Figure 3, the naïve list results are only displayed in Table 3. Starting with the first row of Figure 3, we see that while all of the methods are asymptotically unbiased, the NSM Hybrid and List approaches get to zero much more rapidly than RDS or RW. A similar advantage is evident for the MAB, SD, and DE. The magnitude of the advantage is clear: the Hybrid and List approaches do better on MAB, SD, and DE with a sample size of around 75 than the RDS and RW approaches do with a sample size of 1,000. In practical terms, this means that generalizable inferences with tighter confidence intervals can be made from far fewer respondents collected via NSM.

Table 3 includes results for the naïve list approach. Recall that the key difference between the List and naïve list approaches is that the former chooses cases via a weighting scheme to ensure that nodes sampled early in the process are not over-represented. While the naïve list approach is very intuitive, it does not perform as well as the List approach, with a DE of 2.066 (with a sample size of 1,000). On the other hand, it still outperforms RDS and RW, which have DE's of 181 and 21.1 respectively.

The Hybrid and List approaches do well because once the network has been “discovered” – i.e., as the list L of nominated nodes approaches the population size N—then these approaches are identical to simple random sampling. RDS and RW have difficulty with this network because although it is highly connected, it is easy for them to get stuck in one of the clusters because they are wandering “blindly” with no accumulated network information other than a list of sampled cases. For example, for each node in cluster D (the bottom right hand side cluster), there are 19 within cluster ties and 1 tie to someone out of the cluster. As a result, if RDS or RW are currently interviewing in cluster D, they have only a 5% chance of exiting to a different cluster.

While Figure 3 shows that the SD of RDS and RW decline as the sample size increases, it does so at a relatively slow rate compared to SRS. As a result, the DE for these two methods does not go down; it is relatively constant for RW and goes up for RDS—which indicates that RDS loses ground relative to SRS in this network as the sample size increases. In

contrast, the Hybrid and List approaches have DE's that rapidly approach 1 in Figure 3, indicating increasing parity with simple random sampling.

Overview of the Add Health and Facebook networks

Table 4 shows descriptive statistics for the Add Health and Facebook networks. In all of the networks discussed here, we use only the largest connected component. This is a common and necessary approach in evaluating link-tracing samples via simulation because there are no links between components and correspondingly no way to transition from one to the other; indeed, a key assumption of RDS is that the results are only generalizable to the "giant component" (Volz and Heckathorn 2008). One of the datasets we use is the network data from the National Longitudinal Study of Adolescent Health. These are some of the most commonly studied network data in the world and are perfect for our evaluations because they contain numerous instantiations of the same network survey conducted at the same time in different places. The differences between networks allows us to explore how characteristics of those networks relate to biases and inefficiencies in the various methods we test. In the Add Health networks we have treated all nominations as symmetrical (i.e., we forced non-reciprocated nominations to be symmetric). Goel and Salganick (2010) considered the sensitivity of RDS results to this assumption in the Add Health data by looking only at ties that were actually reciprocated in the original data and found that the design effects were considerably worse than in the case where all ties were treated as symmetric.[16]

In the Add Health networks, the dependent variable is proportion white students in the school. Goel and Salganick found that race had the highest design effects among the variables they tested in the Add Health data, which, we argue, makes it a good variable to use to evaluate our competing methodologies. We limit our sample to the 62 schools where the proportion white was between 0.2 and 0.8[17]. In Table 4A, we see that the size of the Add Health networks ranges from 110 to 1,610 nodes. Because the respondents were asked to list their top 5 male and female friends, the average degree is low in these networks compared to the Facebook networks we discuss below; the global average degree is 8.48, and it ranges from a low of 4.28 to a high of 11.99. As a measure of the level of intergroup friendship taking group size into account, homophily is defined as the 1 minus the ratio of the number of cross-group friends divided by the expected number of cross group friends under random assignment.[18] On average, the level of white/non-white homophily averages 0.562, and goes from a low of -0.021 to 0.816.

A final descriptive measure is "Y-mean degree difference", which is the percentage difference in the average degree by the dependent variable (white/non-white in the case of the Add Health networks). This variable is a measure of whether one group has, on average, more friends in the network than the other group. Gjoka et al. (2010) find that average degree is particularly challenging to estimate with link tracing approaches. We calculate it as the absolute value of the difference in average degree among the groups divided by the overall average degree:

[16]Goel and Salgnaik considered how design effects varied across high-schools and high-school/middle-school combinations and found no substantive differences. As such, we do not distinguish these types of schools in our analysis; however, we do examine how biases vary across different observed parameters of these networks.

[17]We decided to omit schools with super-majorities by race because these cases were extremely problematic for RDS. This is in line with Wejnert's (2009) findings that very small or very large proportions pose particular challenges for RDS.

[18]By "random assignment" we mean the number of cross-group friends if each person's friends were assigned randomly to different groups. For example, in a school that is 80% white and 20% non-white, 20% of white students' friends would be non-white and 80% of non-white students' friends would be white.

$$\text{y-mean degree difference} = 100 \frac{\left| \bar{d}_{Y=1} - \bar{d}_{Y=0} \right|}{\bar{d}} \quad (17)$$

where $\bar{d}_{Y=A}$ represents the mean degree for group A of the dependent variable. The reason for including this variable is that we want to see whether having an imbalance in average degree across the dependent variable affects the results. The test network in Figure 2 is an extreme example of a network where average degree differs sharply across the different groups (i.e., clusters A, B, C, and D).

Add Health results

We start our discussion of the Add Health results by considering one of the smallest networks, network #112, which consists of 210 nodes and is 54.2% white. Of the 1,638 social ties in the network, 250 of them are cross race ties, indicating that there is considerable cross-race interaction even though friendship is racially segregated.

Figure 4 shows a picture of this network, with white students as black nodes. A key observation to make from this picture is that while the total number of intergroup ties is large (250), they are not evenly distributed across students. This is a contrast with the test network, where every node had 1 tie to a different cluster. This is important because it means that RDS and RW—which have no memory of the overall network structure—can get stuck in the interior of each of the two large clusters defined by race (white/non-white) in Figure 4—with a limited chance of transitioning to the other cluster.

Figure 5 and Table 5 show the results for sampling from network 112 with 500 samples of 500 cases each. The basic pattern of the results is similar to the test network discussed above. While all of the approaches appear to be asymptotically unbiased, the Hybrid and List approaches show the most rapid approach towards zero average bias. Once again, the key result is the dramatic difference in the design effects. The Hybrid and List approaches have DE's of 1 after around 200 cases (see Figure 5 panel B, row 2), while RDS and RW don't get anywhere close to 1 after 500 cases. The naïve list approach does not do as well as the List approach, with a DE of 4.697 in samples of 500 cases.

Columns 1 and 2 of Table 6 show the overall results from sampling the 62 Add Health networks. Column 1 shows the average (and median) DE and bias with the school's proportion white as the dependent variable, and Column 2 shows the results with mean degree as the dependent variable. The overall finding in Column 1 is that the NSM Hybrid and List approaches do very well on these networks with proportion white as the dependent variable, with average DE's of 1.110 and 1.236 respectively. In contrast, RDS has an average DE of 66.883 (median DE: 47.766), and RW has an average DE of 12.698. A key finding in Table 6 is that in these relatively small networks, the List approach does very well. Because the List approach is so intuitive and simple (see section 3B), we believe that it may have considerable appeal as an alternative to RDS when sampling from a network when the precision of the estimates is a concern and it is possible to collect basic network data. In Column 2 we see a similar pattern of results with mean as the dependent variable; the Hybrid and List approaches have the lowest average DEs (1.069 and 1.132 respectively), while the average DE for RDS is considerably higher (25.206).[19]

[19]In addition, we note that the average finite-sample bias in column 2 of Table 6 is comparable to the average bias reported in Column 1, provided we take the fact that the global mean of the dependent in Column 2 variable (mean degree in the school) is about 10 times larger than the global mean of proportion white.

Facebook networks

In order to test NSM on larger networks, we use data from Facebook, an online social network launched in February 2004. The data we use here were collected in 2005, when Facebook membership was only available to those affiliated with a university (i.e., those with a valid '.edu' email address at schools which had been registered with the company). The dataset we use contains a census of all users and the connections among them at 100 university networks that were part of Facebook in September 2005 (an overview of this data and access information is given in Traud et al., [2010a, 2010b])[20].

The bounded nature of student friendships within universities, in addition to replications over 100 social settings is an important part of why this dataset is appropriate for this analysis. Sociologists have used Facebook for a variety of research questions (Lewis et al., 2008; Wimmer and Lewis 2010). The most important feature of these data for our purposes is that they are reasonable representations of offline friendships within universities (Haythornthwaite 2005; Boyd and Ellison 2008; Lampe et al. 2007; Clouston et al. 2009), though there are some nuances to this (Ellison et al. 2006; Ellison et al., 2006). At that time, users were also relatively open about sharing information online (Gross and Acquisti 2005), which allows us access to several self-reported attributes of respondents.

In our sampling tests using the Facebook data, the proportion of first year students (freshman) in the network is our dependent variable. In preparing the data, we deleted cases that were missing information on the respondent's year in school. In addition, because the average degree of these networks was so large, we decided to test the Hybrid and List approaches by limiting the number of friends that each respondent could list to a maximum of 20, a number that is below the average degree we found in the field test of collecting network data we conducted that is discussed below in Section 6 and which could reasonably be collected in a survey. For NSM, more network data—in terms of the average number of friends listed on respondents' network rosters—is always better, because it means that the network is revealed faster (i.e., $L \square N$ more quickly, see section 3.B above). Thus, by limiting the number of friends each respondent can list, we are making it *harder* for NSM to sample from these networks. For the RDS and RW approaches, we will run the samples twice, once with a maximum of 20 friends per respondent, and once with all of the friends. In the simulated samples with the networks truncated at 20 friends, before the sample begins we randomly select 20 friends per respondent if the respondent has more than 20 friends. These 20 friends are randomly re-chosen for each replication of the sample.[21]

Table 4B presents basic descriptive statistics on the 100 Facebook networks. The level of friendship homophily between freshman/non-freshman is high in this data, averaging 0.786 and ranging from 0.378 to 0.9. The number of nodes averages 4,635 with a maximum of 16,278. The mean degree for the full data is 81.14, which is the reason we choose to limit the maximum degree for the Hybrid and List approaches to 20.

Facebook results

We begin by discussing results for the largest Facebook university network (“network 100”), which has 16,280 nodes. This network is 30.3% freshman, and has a homophily level of 0.683. The average degree is 106.4 (19.77 in the network truncated at degree 20). Figure 6 shows the results for RDS and RW using the full networks (because both RDS and RW do

[20]In the year after these data were captured, Facebook expanded to allow access to non-university students.

[21]In order to ensure that the truncated networks are connected, we arrange the nodes sequentially by class year (freshman/non-freshman) in a single line running from 1 to N and assign each node ties to their immediate neighbor in front and behind them. These are included in each instantiation of the network and are count towards the friendship totals for each node. The un-truncated networks are highly connected.

better on the full networks), and both the truncated and full networks are presented in Table 7.

The important finding in network 100 is that the Hybrid approach continues to do very well, with an average bias of 0.0003 after 500 steps, and a MAB of 0.01661. To put the MAB result into perspective, the first row of Figure 6 shows that Hybrid does better with a sample size of about 75 than either RDS or RW do with 500. In addition, while the List approach performed as well as the Hybrid approach on the smaller Add Health networks, this is not the case in this large Facebook network. Although the second row of Figure 6 shows that the DE of List is declining, it does not do so as rapidly as the Hybrid approach, resulting in a DE of 4.379 after 500 steps (cases), compared to 1.079 for the Hybrid.

Column 3 of Table 6 presents the overall results for the Facebook networks. The results confirm the basic discussion from Network 100 above: Hybrid and List have much lower DE's than RDS or RW. In addition, the DE of the List approach is slightly higher on the larger Facebook networks compared to the Add Health networks, and the Naïve List approach doesn't do much better than RW. In contrast to the List approach, the DE of Hybrid stays low (1.198).

As a summary measure of the overall results combining the Add Health (with proportion white as the dependent variable) and the Facebook networks together, we note that the average DE on these 162 networks using the Hybrid approach and List approaches was 1.16 and 1.87 respectively. These DEs are close to what would be expected in SRS (DE=1), and perform significantly better compared to the DEs of 77.4 for RDS and 12.3 for RW.[22] On all 162 of the networks both the NSM Hybrid and List approaches had a lower DE than RDS or RW.

Finally, in order to evaluate the sensitivity of the observed DE and average bias to the network characteristics from Table 4, in Tables 8 and 9 we combine the results of the Add Health and Facebook networks and run a regression analysis.[23] In Table 8, the dependent variable is the DE, and in Table 9 the dependent variable is the average bias calculated using Equation 13 above. The bottom two rows of each table present the mean and standard deviation of the dependent variable as a reminder that these vary considerably across the different methods. All the models include a dummy variable indicating the source of the data (1=Add Health, 0=Facebook). As discussed above in reference to Table 4, the variable “Y-mean degree difference” is a measure of whether the mean degree varies across categories of the dependent variable (see Equation 17 above). We use the natural log of the number of nodes to measure network size.

There are three key findings in the analysis of the DE in Table 8. First, for the Hybrid approach, none of the variables have a significant effect (at the p=.05 level): in other words, the method does not appear to be sensitive, with samples of 500 cases, to either the level of homophily or the size of the network. Overall, the R-squared for the Hybrid approach is 0.076, meaning that little of the variation in design effects is explained by these features of the networks. Second, the measure of network size, ln(nodes), has a significant positive effect on the DE for the List approach, which means that the precision of the estimates declines as the network size increases. Note that the effect is not overwhelmingly large, as going from 500 nodes [ln(nodes)=6.21] to 10,000 nodes [ln(nodes)=9.21] would increase the predicted DE by 0.964×3=2.892. Moreover, note that the problem of network size for the List approach can be solved by sampling a larger number of cases to reduce the sampling

[22]Using the full network data, the average DE for RDS and RW were 78.7 and 16.65 respectively.
[23]We use the results for proportion white as the dependent variable for the Add Health data.

variance. However, a comparison of the results for the Hybrid and List models indicates the reason why the Hybrid approach is superior—by using the search mode to explore the network, it increases the efficiency of the sampling process. Next, note that the naïve list approach is very sensitive to network size, with a coefficient of 12.14 for Ln(nodes) compared to 0.964 for the List approach.

The third key finding of Table 8 is that RDS, RW, and the naïve list approaches are very sensitive to the level of homophily in the network. The coefficient on homophily for RDS is very large and significant (b=284.1, s.e. = 15.16). In contrast, homophily has no effect on the Hybrid and List approaches.

Table 9 presents results for the analysis of the average bias across the 162 networks of the combined results. All 5 of these approaches are asymptotically unbiased—in the sense that as the sample size goes to infinity, the average bias will go to zero. As discussed above, for the Hybrid, List, and naïve list approaches, the asymptotic unbiased behavior derives from the fact that they are sampling from the accumulated list of network members L. However, in these finite samples of 500 cases, all 5 methods exhibit a small amount of bias (as indicated by the second row from the bottom in Table 9). As indicated in Table 9, the size of the network affects the magnitude of the finite sample bias for the Hybrid, List, and naïve list approaches, while homophily and average degree are significant predictors of the bias for RDS, and homophily is for RW. Notably, average degree is not associated with bias in the NSM approaches.[24]

5. Sampling diagnostics and convergence to Simple Random Sampling

Earlier in the paper, we made the claim that NSM converges to SRS as the sample size increases and all individuals in the population have been nominated on the network rosters (i..e, as $L_s \to N$). In this section, we state the assumptions that are behind this claim and show how the data from an NSM sample might be used to indicate when convergence to SRS has been achieved. In order to be clear about what we are claiming, by "converge" we mean that once NSM has "converged" to SRS, the properties and characteristics of samples obtained by NSM would be indistinguishable from samples obtained from SRS with the same sample size. Though our work in this regard is preliminary, it is our belief that such a diagnostic would provide an immensely useful tool for researchers in practice, allowing them to understand and quantify the statistical validity of a sample collected via NSM as well as to facilitate comparisons across samples.

Requirements for convergence to SRS:

(5.1) The network is connected. Starting from any node, all individuals can be reached by following a path of connections in the network.

(5.2) Respondents are willing to provide network data, and the quality of the network data (i.e. partial name information, basic demographics, last 3 digits of telephone numbers, etc.) is sufficient to uniquely identify network members.

(5.3) The sample size reaches a point where the number of nominated nodes in the network, L, is equal to the estimated size of the network using the capture-recapture method ($\hat{N}$ in Equation 5).

[24]Overall, when evaluating these results, note that we have chosen two dependent variables (% nonwhite and % first-year students) where there are high levels of homophily and clustering. The DE for RDS would be lower if we had chosen networks with lower levels of clustering, or different dependent variables. In practice, of course, the researcher will not know how clustered the network of the target population is, and the goal here has been to test NSM on networks that are difficult for network-based sampling algorithms.

(5.4) "Even" sampling has been used (see Section 3.B), and all nominated nodes have the same cumulative sampling probability.[25] See Step 251 of Table 2 for an example.

Figures 7 and 8 show how convergence to SRS can be identified using Facebook network #10 as an example, using NSM's Hybrid approach with 500 replications and samples of 1,000 cases for each replication. Facebook network #10 has 1,092 nodes. Figure 7 shows the average bias, the proportion of cases that are sampled using the Search mode, and P1 (the proportion of nominated nodes with only 1 nomination) by step (i.e. the number of cases in the sample). Because the Search mode is only used when P1>0.4, i.e., when more than 40% of the nominated nodes have only been nominated once, the use of the Search mode drops to zero as the average value of P1 drops below 0.4 around Step 100. In addition, the value of P1 declines towards zero as the sample size increases, indicating that the network is being completely explored around step 600. In addition, the average bias declines very rapidly and by step 300 it is very small—indicating that the sample estimate has become very accurate at this point.

Figure 8 shows the estimated network size $\hat{N}$ and number of nominated nodes by step (L_s) in this network, along with the observed 95% confidence intervals for both of these measures. It is easy to see in Figure 8 that they are both converging to the true value of the population size, 1,092. When $\hat{N}$ then all the nodes in the network have been revealed, and requirement 5.3 has been satisfied. Provided requirements 5.1, 5.2, and 5.4 are also satisfied, then NSM is statistically equivalent to SRS. As discussed above, the basic intuition why this is true is that the even sampling component of the List mode (Section 3B) ensures that all people who are nominated in the survey are subject to the same cumulative sampling rate. Once everyone in population A has been nominated ($L_S=N$), then we have a representative sample of S cases from population A.

The benefit of Figure 8 is that it illustrates how the data collected in an NSM survey can be used as a diagnostic tool to evaluate how close the sample is to converging to SRS. One way to define an adequate stopping rule for an NSM sample would be to collect a sample size S* such that $\hat{N}=L_{S^*}$.

In Figure 7, it is clear that for this particular network, a high level of accuracy is reached around Step 300 as the average bias is very close to zero. In Figure 8, we see that L and $\hat{N}$ are very close by Step 300 (they differ by 7%), although complete convergence between L and $\hat{N}$ isn't reached until around Step 600. If we define the degree of convergence between the number of nominated cases L and the estimated population size $\hat{N}$ as:

$$q=100\times\frac{L}{\hat{N}}, \quad (18)$$

then it may be the case that the marginal return in accuracy of going from q=0.93 to q=0.999 in general is very small. In future research, we plan on refining the definition of the optimal stopping point for an NSM sample based on different values of q, allowing for a reasonable threshold of convergence between $\hat{N}$ and L by testing the marginal gains in precision for a range of values of q over the 162 networks evaluated here.[26] As the purpose of this section is simply to illustrate that such diagnostics are possible with NSM, we leave further refinement of these issues to future work.

[25]Note that the use of sampling weights (Equation 9 and Equation 10 above) adjusts for unequal sampling rates among nodes prior to convergence to SRS.

[26]In general, we also note that we observed a very high level of accuracy for the predicted population size $\hat{G}$ using the capture-recapture method on all of the 162 networks that we tested.

6. Is it feasible?

We envision NSM being used by two types of research projects. First, network-based sampling methods, such as RDS, are attractive when the target population is rare or hidden and there is no sampling frame available for the researcher. Second, sometimes the goal of the survey is to actually collect data about the network itself, and NSM allows the researcher to collect an accurate sample while following the links connecting people in a network together. In this section we discuss the feasibility of collecting the network data needed to run NSM. There are three questions that must be answered to make that decision: (1) First, how many people can we expect individuals to nominate? (2) Second, can we protect confidentiality but at the same time match alters who have been nominated by multiple individuals? (3) Finally, how sensitive is the method to coding errors and other logistical problems? We consider these first two issues in depth using our experience surveying a hidden population and evidence from the literature. We then present some preliminary work on the third question.

Collecting network data

In the summer of 2010, we conducted a pilot study to test the feasibility of collecting network data on a difficult to reach population. Along with several other colleagues, we collected data on the migration network connecting a Mexican community from a medium sized town in the state of Guanajuato to migrants from that hometown currently living in North Carolina and Houston, Texas. The resulting survey, the 2010 Network Survey of Immigration and Transnationalism (NSIT) had a sample size of 150 in North Carolina, 52 in Houston, and 407 in Guanajuato. We started the data collection with a snowball sample in North Carolina and Houston, and then followed links back to Guanajuato.

We asked respondents to nominate network alters who were eligible to be in the survey and to whom they could refer us. In order to protect the privacy of respondents, we collected data on only the first four letters of the first name and last name of the respondent's network members, along with key social and demographic information that we could use for identification: nickname, gender, age, occupation, and number of children living in the household. In order to identify unique individuals in the resulting network data, we wrote a matching program in Stata that tolerates slight differences in the demographic and name variables in determining whether two network nominations (from different interviews) represent the same person. We used the Levenshtein edit distance (the number of edits needed to match two strings, cf. Reif [2010]) to allow for reporting and coding errors in the first name, last name, and nickname.

The North Carolina and Houston samples asked for up to 10 friends and 5 family members currently living in the destination community in the U.S., up to 6 total family and friends currently living in the origin city in Mexico, and up to 5 returned migrants currently living in the origin. Of the 150 respondents in N.C., the average number of network members was 21.2.[27] The network questions were placed at the beginning of the survey, and took an average of 10 minutes to complete. All of the interviews, aside from pretests, were conducted by community members. Of the 8,538 nominations of friends and family members in the survey overall, all but 19 reported a first name, 98.96% reported a last name, and 94.7% reported age.

In addition to the data we collected in the 2010 NSIT, there is other evidence that similar amounts of network data could realistically be collected on social surveys. A reasonable

[27]In the Mexico sample, we asked for data on up to 6 friends and 6 relatives currently living in Guanajuato, and up to 6 friends/family living in North Carolina and Houston, Texas. The overall average degree was 14.13.

question is how many alters a given ego might nominate because the larger the number of nominations a person can give, the quicker the Search mode will be able to explore the network. In general, it is clear that over a period of time, people come into contact with far more individuals than they are aware of—Boissevain (1974) and McCarty et al.(2000) estimate contact networks of between 1500 and 2000 people—but may be only be able to remember and identify a fraction of these. Dunbar (1998) reports that people typically can only remember about 150 individuals. In a recent example of extensive network data collection, Sandberg et al. (2008) found that residents of rural Senegal named an average of 21.4 unique network alters (the interquartile range was 17-25) when asked 15 name generating questions with open ended numbers of potential responses. Similarly, in a classic work, McCallister and Fischer (1978) found that most respondents in the Northern California Community Survey were able to name between 10 and 30 alters and the type of relation to them in 20 minutes of interview time (a pilot survey elicited an average of 20.3 names and the total survey an average of 18.5). Overall, based on the results from the 2010 NSIT and other surveys that have obtained between 10-30 nominations per respondent, we argue that it is possible to collect enough network data to improve the accuracy of network sampling. Of course, it might not be possible for all target populations, if the actual networks are sparse, or if participants are reluctant to provide (or it is deemed unsafe to ask them to provide) information about network ties.

Can we protect confidentiality in stigmatized populations?

Our work with the NSIT provides some evidence that it is possible to match alters using a small set of demographic characteristics that are unlikely to put respondents of stigmatized populations at risk of deductive disclosure. There are other ways to achieve this being explored in the literature. Dombrowski et al. (2011) devised a method to estimate the size of hidden populations using the capture-recapture method within a respondent driven sample. They achieved this by collecting data on respondents' appearance (approximate height, weight, hair color, eye color, gender and race/ethnicity) as well as a transformation of the last three digits of their telephone number. These digits were encoded as either being even or odd, and greater than or equal to five. Such a six-bit code is highly likely to be unique within a given population, but at the same time is nearly impossible to trace back to an individual person. They also solicited this information about five of each respondent's friends that were eligible for inclusion in the study. By matching such data on the individuals at-risk of being surveyed (by being friends with individuals who were surveyed and being eligible members of the population) to those who were actually interviewed, the authors were able to apply capture-recapture techniques and estimate the size of the total network.

Of course, not all populations of interest have mobile phones where they can look up their acquaintances numbers, and there is the potential that looking up such information may add considerable time to the survey. However, other encoding schemes could be devised. For instance, one could apply the same logic to the first three letters of individuals first and last names (e.g., is it a vowel or consonant, is it at the beginning, in the middle or at the end of the alphabet). Since we believe that the issues of respondent risk need to be weighed on a case by case basis, we do not have specific recommendations as to what is the best means of collecting such data, but we note that very promising approaches are being tested. Based on our experience with the NSIT and the evidence in the literature, we are confident that appropriate data collection strategies could be designed to maximize respondent anonymity while at the same time making network matching techniques feasible.

Test of the effect of network coding errors

How robust are the results to the possibility of coding errors in the identification of network members? First, a long literature has documented problematic evidence of recall bias and

under-enumeration in network measurement through self-reports (e.g., Eagle et al., 2009; Bignami-Van Assche and Watkins 2004; Bell et al., 2000; Cascairo et al., 1999; Freeman et al., 1987; Hammer 1984; Bernard and Killworth 1977; Coleman et al., 1966). It is realistic, certainly, to anticipate a certain degree of coding error in identifying network members based on survey data. Nonetheless, the combination of partial name data and key demographic variables may result in a fairly high level of accuracy. In the 2010 NSIT, we looked over the data for the first 1000 nominations with the help of community members after running the matching program on first name, last name, nickname, age, gender, origin community, and number of children, and our assessment was that the program was working just as accurately as someone with a detailed knowledge of the community could do in assigning unique identifiers to the nominations. This degree of matching is especially heartening given the relative name-similarities in the population and the cultural preference for nicknames.

Other attempts at matching network nominations to individuals in different populations have yielded similar results. Sandberg et al., (2008) tested the feasibility of augmenting a large data collection project ($N \approx 30{,}000$) with a social network survey that collected detailed relationship data from a sample of respondents. The innovation was that network alters would be linked to existing data records on them, circumventing the need to rely on ego reports of alter characteristics and greatly reducing the costs of network data collection. In field-testing this approach, they found that approximately 88% of nominated alters could be uniquely identified with just 5 pieces of information that could reasonably be collected from respondents. This evidence further underscores the feasibility of our proposed approach.

We tested the impact of coding errors on the efficiency of NSM by used a modified algorithm that creates errors in the coding of the network data with a certain probability. Network ties that are coded in error are assigned to someone else in the currently revealed network (a sampled or unsampled, but nominated, node). We tested the impact of errors on a sample of 44 of our large university Facebook networks discussed above, using a 15% rate of error. Using the Hybrid method with a 15% error rate, the average DE (see Equation 16 below) for these 44 networks with 500 replications of a sample size of 500 was 1.27, compared to a 1.23 average DE for the Hybrid method with no errors in the coding of the network data. For the purposes of comparison, the average DE for RDS with no errors in coding on these three networks was 65.66. 28 Overall, these results indicate that a moderate level of error in the coding of the network data has only a minimal affect the performance of NSM. At the same time, high levels of coding error would certainly reduce the efficiency of NSM, and it is crucial that careful attention is paid to collecting partial name and demographic identifiers that will facilitate a high level of accuracy in the coding of the network data.

Before NSM is implemented in the field, other logistical considerations should also be taken into account. These issues include developing a better understanding of how non-response may bias results, developing protocols to ensure respondent protection, and testing the method's efficacy when list-based sampling is implemented only a small number of times, which would reduce the burden of recontacting respondents. For example, in future research, we plan on testing a variant of NSM that minimizes or eliminates the need for recontact by entering the respondent's network information directly into a computer or a smartphone during the interview, updating the network, and asking for a referral at the end of the interview itself. Based on the updated network data, we could use the Search mode to select who we ask for a referral to among the respondent's network roster based on which node is most likely to lead to new or underexplored parts of the network. This approach would reduce the cost of conducting an NSM survey while still utilizing the accumulated network data to make the sampling process more efficient than a conventional random walk.

7. Discussion and Conclusion

Respondent driven sampling (RDS) is a random walk based sampling approach that has become popular as a method of collecting data from hidden or rare populations. However, the sampling variance of RDS depends on the structure of the underlying social network (Bassetti and Diaconis 2006; Goel and Salganick 2009), which is unknown to a researcher using RDS data that consists of only the links that were followed by the sample. Recent tests using real networks indicate that the sampling variance of RDS samples may be unacceptably high (Goel and Salganick 2010; Lu et al. 2011; McCreesh et al. 2012).

In this paper, we propose an alternative approach, Network Sampling with Memory (NSM), which collects network rosters as part of the survey, and uses the revealed list of network members to improve the efficiency of the sampling process. NSM samples from the list of network members using two sampling modes, List and Search. As discussed above, the basic innovation of the List mode is quite simple: rather than take a randomly selected link from the current case as in RDS or a random walk, in the List mode the next node to be sampled is chosen by sampling with replacement from the complete list of all people who have been nominated in the survey. While RDS and random walks (RW) can get stuck in segregated islands of the network—which increases the sampling variance—the List mode does not get stuck because it draws the next case from the set of all individuals who have appeared on the network rosters. In addition, the List mode ensures that all people on the network list are sampled at the same rate by sampling newly discovered network members at the same cumulative sampling rate of the sample as a whole. As the survey progresses and all the members of the population are gradually added to the network list, the List mode becomes identical to the process of simple random sampling.

As discussed above in Section 3.C, the Search mode is designed to speed up the process of exploring the network by giving priority to “bridge nodes” that connect the edges of the currently explored network to underexplored areas of the network. The “Hybrid” approach that we test in the simulated sampling combines the List and Search modes, using the Search mode initially, and then defaulting to the List mode once the network has been explored.

We test the relative performance of NSM versus RDS, RW, and simple random sampling by conducting simulated sampling on 162 observed social networks from Add Health and Facebook. The advantage of testing the performance of the methods on so many different networks is that we can be more confident that the results are not due to the idiosyncrasies of any particular network. The networks ranged in size from 110 to 16,278 nodes, with different levels of homophily and mean degree. Moreover, because these are real social networks, they will exhibit the kinds of complex network structures that artificial test networks may not have. Overall, as discussed above, our results indicated that NSM's Hybrid and List approaches had lower design effects (DE) than RDS or RW on all 162 of the Add Health and Facebook networks, based on drawing 500 samples of 500 cases each from each network. In particular, the Hybrid approach resulted in a 98.5% reduction in the average DE compared to RDS (1.16 versus 77.4).

Although in all of our tests NSM is a more efficient method of sampling from a social network than RDS, there is an added cost, which is the time and effort needed to collect network data as well as re-contact respondents in order to get referrals to previously nominated cases[29]. As discussed above in Section 6, we conducted a pilot study on collecting network data from a hidden population—immigrants in North Carolina from a

[29]Recontacting participants may not be feasible in some hidden populations. See the final paragraph of Section 6 for a discussion of a future variant of NSM that would eliminate recontact with respondents.

specific origin community in Mexico—and averaged 21.4 links per respondent in North Carolina, obtaining almost universal responses on the first four letters of the first and last names and the ages of the members of respondents' networks. This indicates that it is feasible to collect the level of network data needed to achieve the gains in accuracy we present in this paper. At the same time, is important to emphasize that we do not anticipate that NSM will be applicable for every situation. Nonetheless, researchers are devising innovative ways to collect this type of data in even highly stigmatized populations (Dombrowski et al. 2011). Indeed, Dombrowski et al. demonstrate that individuals' network alters could be uniquely identified for the purposes of capture-recapture methods while at the same time maintaining respondent confidentiality in exactly the type of population RDS was designed to be implemented in.

Even if the added difficulty of collecting network data means that NSM is not universally applicable for all types of hidden or difficult to reach populations, we believe there are two reasons why it represents an important addition to the literature on sampling from networks. First, if researchers can collect network data as part of their survey, our results indicate large potential gains in sampling efficiency: simply put, it takes far fewer cases for NSM to reach a given level of precision compared to RDS. Recognition of the tradeoff between the cost and precision of the survey is important even if a research project elects to use RDS or a conventional link-tracing survey design. Second, on a theoretical level, NSM provides a clear example of how gains in accuracy can result from a simple modification of the sampling algorithm compared to a random walk based approach. We seek to bridge the recent literature in mathematics and computer science on sampling large internet-based networks with the social science literature on sampling hidden or rare populations based on face-to-face or telephone interviews. While there is a fundamental difference in the cost of conducting an interview (trivial for an internet-based network and high for a face-to-face interview), there is likely to be considerable synergy and convergence between these two literatures in the future. In this light, we see NSM as a first step in demonstrating how collecting network data can improve the accuracy of sampling from networks.

In particular, we anticipate several avenues of future research. First, it is possible that with computer assisted interviews—i.e., inputting the network data directly during the interview —the interviewer might be able to solicit referrals to newly identified nodes of the network at the conclusion of each interview, thereby reducing the frequency with which respondents would need to be re-contacted to get a referral to the next person to interview. Second, we plan to explore whether it is possible to use resampling methods to infer the sampling variance of the resulting NSM sample. The idea would be to predict the ties from unsampled but nominated nodes using a statistical model of the network, and then repeatedly resample the data from the predicted network to calculate bootstrap standard errors. Third, we intend to assess how the accumulated network data could be used to develop in an optimal stopping rule regarding the sample size required to reach a certain degree of sampling precision.

Overall, in the current paper we have presented an alternative method for sampling from network-based populations which is both transparent in terms of its underlying mechanism and, in all of our tests, considerably more accurate than the methods currently in use. Indeed, in the results presented in Figures 3, 5 and 6 we found that a roughly equivalent level of sampling precision was achieved with 75 cases sampled using our proposed method as could be found with 500 to 1,000 cases using the currently dominant approach. Further, we have offered first steps toward quantifying the extent of NSM's results' convergence to simple random sampling, diagnostics which are currently unavailable or inadequate (cf. Neely 2009; Wejnert 2009) in the network sampling literature to date. Though our proposed method would constitute an increased logistical burden for the researcher as well as other potential tradeoffs, we believe it presents considerable promise. Rather than an end point,

we view this as a call for future advances in improving the efficiency and precision of sampling from networks using survey data on the revealed network of the target population.

Acknowledgments

We wish to thank Peter Mucha, Mason Porter and Adam D'Angelo for their generous provision of access to the Facebook data we use. We also thank Amanda Traud for helpful discussions concerning the structure of that data.

References

Abdul-Quadar AS, Heckathorn DD, Sabin K, Saidel T. Implementation and analysis of respondent-driven sampling: lessons learned from the field. Journal of urban health: bulletin of the New York Academy of Medicine. 2006; 83(7):i1–i5. [PubMed: 17058119]

Alon, N.; Avin, C.; Koucky, M.; Kozma, G.; Lotker, Z.; Tuttle, MR. Many random walks are faster than one. Proceedings of the twentieth annual symposium on parallelism in algorithms and architectures; New York, NY. Association for Computing Machinery; 2008.

Alon N, Benjamini I, Lubetzky E, Sodin S. Nonbacktracking random walks mix faster. Communications in Contemporary Mathematics. 2007; 9:585–603.

Atrostic BK, Bates Nancy, Burt Geraldine, Silberstein Adriana. Nonresponse in U.S. Government Household Surveys: Consistent Measures, Recent Trends and New Insights. Journal of Official Statistics. 2001; 17(2):209–26.

Avin C, Krishnamachari B. The Power of Choice in Random Walks: An Empirical Study. Computer Networks: Special Issue on Wireless Performance. 2008; 52(1):44–60.

Avrachenkov, K.; Ribeiro, B.; Towsley, D. Technical Report no 7394. Sophia Antipolis, France: Institut national de recherche en informatique et en automatique (INRIA); 2010. Improving Random Walk Estimation Accuracy with Uniform Restarts.

Bassetti F, Diaconis P. Examples comparing importance sampling and the Metropolis algorithm. Illinois Journal of Mathematics. 2006; 50(1-4):67–91.

Bell DC, Montoya ID, Atkinson JS. Partner concordance in reports of joint risk behaviors. Journal of Acquired Immune Deficiency Syndromes. 2000; 25:173–181. [PubMed: 11103048]

Berebrink, P.; Cooper, C.; Elsasser, R.; Radzik, T.; Sauerwald, T. Speeding up random walks with neighborhood exploration. Proceedings of the Twenty-First Annual ACM-SIAM Symposium on Discrete Algorithms; Philadelphia, PA. Society for Industrial and Applied Mathematics; 2010.

Bernard HR, Killworth PD. Informant accuracy in social network data II. Human Communications Research. 1977; 4:3–18.

Bignami-Van Assche, S.; Watkins, SC. Husband–wife disagreement in rural Malawi: a longitudinal analysis. Paper presented at the Annual Meetings of the Population Association of America; April; Boston. 2004.

Boissevain, J. Friends of friends. Oxford, UK: Blackwell Publishing; 1974.

Boyd DM, Ellison N. Social network sites: Definition, history, and scholarship. Journal of Computer-Mediated Communication. 2008; 13(1):210–230.

Brin, S.; Page, L. The Anatomy of a Large-Scale Hypertextual Web Search Engine. Seventh International World-Wide Web Conference (WWW 1998); April 14-18, 1998; Brisbane, Australia. 1998.

Burt, Ronald S. Structural Holes: The Structure of Competition. Cambridge MA; Harvard University Press: 1995.

Burt RS. Network items and the General Social Survey. Social Networks. 1984; 6:293–339.

Cascairo T, Carley KM, Krackhardt D. Positive affectivity and accuracy in social network perception. Motivation and Emotion. 1999; 23:285–306.

Clouston SP, Verdery AM, Amin S, Gauthier GR. The structure of undergraduate association networks: a quantitative ethnography. Connections. 2009; 29(2):18–31.

Coleman, JS.; Katz, E.; Menzel, H. Medical Innovation: A Diffusion Study. Indianapolis, IN: Bobbs-Merrill; 1966.

Cooper C, Frieze A, Radzik T. Multiple random walks and interacting particle systems. Lecture Notes in Computer Science. 2009; 5556:399–410.

Dombrowski, K.; Khan, B.; Wendel, T.; McLean, K.; Curtis, R.; Drucker, E. Working paper. City University of New York; Estimating the size of the methamphetamine-using population in New York City using network sampling techniques.

Dunbar RIM. The social brain hypothesis. Evolutionary Anthropology: Issues, News, and Reviews. 1998; 6(5):178–190.

Eagle N, Pentland AS, Lazer D. Inferring friendship network structure by using mobile phone data. Proceedings of the National Academy of Sciences. 2009; 106(36):15274–15278.

Elasser, R.; Sauerwald, T. Tight bounds for the cover time of multiple random walks. Theoretical Computer Science. 2010. http://dx.doi.org/10.1016/j.tcs.2010.08.010

Ellison N, Steinfield C, Lampe C. Spatially bounded online social networks and social capital. International Communication Association. 2006:1–37.

Freeman LC, Romney AK, Freeman SC. Cognitive structure and informant accuracy. American Anthropologist. 1987; 89:310–325.

Gasieniec L, Radzik T. Memory Efficient Anonymous Graph Exploration. Graph-Theoretic Concepts in Computer Science, Proceedings of the 34th International Workshop - WG 2008, Revised Papers, Lecture Notes in Computer Science. 2008; 5344:14–29.

Gile KJ, Handcock MS. Respondent-driven sampling: an assessment of current methodology. Sociological Methodology. 2010 no. 10.1111/j.1467-9531.2010.01223.x

Gile, Krista J. Improved Inference for Respondent-Driven Sampling Data with Application to HIV Prevalence Estimation. To appear, Journal of the American Statistical Association. 2011

Gjoka M, Kurant M, Butts C, Markopoulou A. A walk in Facebook: a case study of unbiased sampling of Facebook. IEEE INFOCOM 2010. 2010

Goel, Sharad; Salganik, Matthew J. Respondent-driven sampling as Markov chain Monte Carlo. Statistics in Medicine. 2009; 28:2202–2229. [PubMed: 19572381]

Goel, Sharad; Salganik, Matthew J. Assessing respondent-driven sampling. Proceedings of the National Academy of Sciences. 2010; 107:6743–6747.

Goodman LA. Snowball sampling. Annals of Mathematical Statistics. 1961; 32:148–170.

Gross, R.; Acquisti, A. Proceedings of WPES'05. ACM; Alexandria, VA: 2005. Information revelation and privacy in online social networks; p. 71-80.

Groves, RM.; Fowler, FJ., Jr; Couper, MP.; Lepkowski, JM.; Singer, E.; Tourangeau, R. Survey Methodology. New York: John Wiley; 2009.

Hammer M. Explorations into the meaning of social network interview data. Social Networks. 1984; 6:341–371.

Haythornthwaite C. Social networks and Internet connectivity effects. Information, Communication, & Society. 2005; 8(2):125–147.

Heckathorn DD. Respondent-driven sampling: a new approach to the study of hidden populations. Social Problems. 1997; 44:174–199.

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

Heckathorn DD. Extensions of respondent-driven sampling: analyzing continuous variables and controlling for differential recruitment. Sociological Methodology. 2007; 49:11–34.

Ikeda S, Kubo I, Yamashita M. The hitting and cover times of random walks on finite graphs using local degree information. Theoretical Computer Science. 2009; 410:94–100.

Ikeda S, Kubo I, Okumoto N, Yamashita M. Impact of local topological informaiton on random walks on finite graphs. Lecture Notes in Computer Science. 2003; 2719:1054–1067.

Johnston LG, Sabin K, Hien MT, Huong PT. Assessmnt of respondent-driven sampling for recruiting female sex workers in two Vietnamese cities: Reaching the unseen sex worker. Journal of Urban Health. 2006; 83:16–28.

Johnston, LG. Behavioural Surveillance: Introduction to Respondent Driven Sampling (Participant Manual). Centers for Disease Control and Prevention: Atlanta, GA; 2008. http://globalhealthsciences.ucsf.edu/PPHG/surveillance/other_modules.html

Kendall D, Kerr LRF, Gondim RC, Warneck GL, Macena RHM, Pontes MK, Johnston LG, Sabin K, McFarland W. An empirical comparison of respondent-driven sampling, time location sampling, and snowball sampling for behavioral surveillance in men who have sex with men, Fortaleza, Brazil. Aids and Behavior. 2008; 12(1):97–104.

Kharusi, FA. Technical Report 08-1. Department of Computer Science; King's College, London: 2008. Experiments with forward random walks.

Lampe, C.; Ellison, N.; Steinfeld, C. Proceedings of Conference on Human Factors in Computing Systems. New York: ACM Press; 2007. A familiar Face(book): Profile elements as signals in an online social network; p. 435-444.

Lawler, Gregory F.; Coyle, Lester N. Lectures on Contemporary Probability. American Mathematical Society. 1999

Lewis, Kevin; Kaufman, Jason; Gonzalez, Marco; Wimmer, Andreas; Nicholas, Christakis. Tastes, Ties, and Time: A New Social Network Dataset Using Facebook.com. Social Networks. 2008; 30:330–342.

Lovasz L. Random walks on graphs: a survey. Combinatorics. 1993; 2:1–46.

Lu X, Bengtsson L, Britton T, Camitz M, Kim BJ, Thorson A, Liljeros F. The sensitivity of respondent-driven sampling. Journal of the Royal Statistical Society: Series A Statistics in Society. 2011

Ma X, Zhang Q, He X, Yue H, Chen S, Raymond HF, Li Y, Xu M, Du H, McFarland W. Trends in prevalence of HIV, syphilis, hepatitis C, hepatitis B, and sexual risk behavior among men who have sex with men: results of 3 consecutive respondent-driven sampling surveys in Beijing, 2004-2006. Journal of Acquired Immune Deficiency Syndrome. 2007; 45(5):581–587.

Magnani R, Sabin K, Saidel T, Heckathorn DD. Review of sampling hard-to-reach and hidden populations for HIV surveillance. AIDS. 2005; 19:S67–S72. [PubMed: 15930843]

Malekinejad M, Johnston LG, Kendall C, Kerr LRFS, Rifkin MR, Rutherford GW. Using respondent-driven sampling methodology for HIV biological and behavioral surveillance in international settings: a systematic review. Aids and Behavior. 2008; 12:S105–S130. [PubMed: 18561018]

Mayer A, Puller SL. The old boy (and girl) network: social network formation on university campuses. Journal of Public Economics. 2008; 92:329–347.

McCallister L, Fischer CS. A procedure for surveying personal networks. Sociological Methods and Research. 1978; 7:131–148.

McCarty C, Killworth PD, Bernard HR, Johnsen EC, Shelley GA. Comparing two methods for estimating network size. Human Organization. 2000; 60(1):28–39.

McCreesh N, Frost S, Seeley J, Seelev J, Katongole J, Tarsh MN, Ndunguse R, Jichi F, Lunel NL, Maher D, Johnston LG, Sonnenberg P, Copas AJ, Hayes RJ, White RG. Evaluation of respondent-driven sampling. Epidemiology. 2012; 23:138–147. [PubMed: 22157309]

McKnight C, Jarles DD, Bramson H, Tower L, Abdul-Quader AS, Nemeth C, Heckathorn DD. Respondent-driven sampling in a study of drug users in New York City: Notes from the field. Journal of Urban Health. 2006; 83:54–59.

National Institutes of Health [NIH]. Project Reporter. 2011. Online web-page accessed on 12/15/2011 from http://projectreporternihgov/reportercfm. Text search for “Respondent driven sampling” OR “respondent-driven sampling” in all fiscal years/

Neely, William Whipple. Ph D Dissertation (Statistics). University of Wisconsin-Madison; 2009. Statistical Theory for Respondent Driven Sampling.

Potter, Gail E.; Handcock, Mark S.; Longini, Ira M., Jr; Halloran, M Elizabeth. Estimating within-household contact networks from egocentric data. Annals of Applied Statistics. 2011; 5(3):1816–1838. [PubMed: 22427793]

Ramirez-Valles J, Heckathorn DD, Vazquez R, Diaz RM, Campbell RT. The fit between theory and data in Respondent-Driven Sampling: Response to Heimer. AIDS and Behavior. 2005; 9:409–414.

Reif, Julian. STRGROUP: Stata module to match strings based on their Levenshtein edit distance. Statistical Software Components S457151, Boston College Department of Economics. 2010 revised 14 Aug 2010.

Ribeiro B, Towsley D. Estimating and Sampling Graphs with Multidimensional Random Walks. ACM SIGCOMM Internet Measurement Conference. 2010

Ribeiro B, Wang P, Towsley D. On Estimating Degree Distributions of Directed Graphs through Sampling. University of Massachusetts CMPSCI Technical Report UM-CS-2010-046. 2010

Salganik, Matthew J.; Heckathorn, Douglas D. Sampling and estimation in hidden populations using respondent-driven sampling. Sociological Methodology. 2004; 34:193–239.

Sandberg, J.; Rytina, S.; Lalou, R.; Delaunay, V. Social networks across the lifecourse and the development of the Niakhar Social Networks Survey Instrument; Presentation at the Annual Meetings of the Population Association of America; New Orleans, LA. 2008.

Thompson SK. Targeted random walk designs. Survey Methodology. 2006a; 32:11–24.

Thompson SK. Adaptive web sampling. Biometrics. 2006b

Traud AL, Kelsic ED, Mucha PJ, Porter MA. Comparing Community Structure to Characteristics in Online Collegiate Social Networks. SIAM Review arXiv:0809 0690. 2010a

Traud, AL.; Mucha, PM.; Porter, MA. Working Paper. University of North Carolina at Chapel Hill, Department of Mathematics; 2010b. Social Structure of Facebook Networks.

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

Wejnert C. An empirical test of respondent-driven sampling: point estimates, variance, degree measures, and out-of-equilibrium data. Sociological Methodology. 2009; 39(1):73–116. [PubMed: 20161130]

Wejnert, Cyprian; Heckathorn, Douglas D. Web-Based Network Sampling. Sociological Methods and Research. 2008; 37(1):105–134.

Wimmer, Andreas; Lewis, Kevin. Forthcoming. “Beyond and Below Racial Homophily: ERG Models of a Friendship Network Documented on Facebook”. American Journal of Sociology.

Yeka, William; Geraldine, Maibani-Michie; Prybylski, Dimitri; Donn, Colby. Application of Respondent Driven Sampling to Collect Baseline Data on FSWs and MSM for HIV Risk Reduction Interventions in Two Urban Centres in Papua New Guinea. Journal of Urban Health. 2006; 83(1):60–72.

Yu, I.; Newman, R. Working Paper. Dept of CISE, University of Florida; Gainesville, Florida: 2008. A topology-aware random walk.

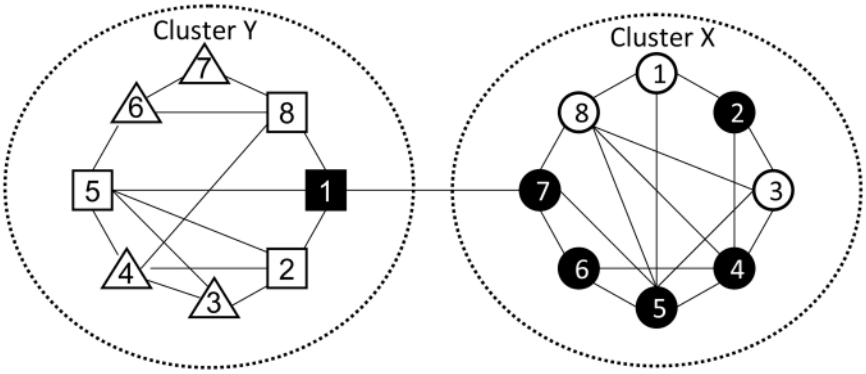

Figure 1. Illustrative network with two clusters

Notes: Hollow nodes are unsampled, dark nodes are sampled. Circles indicate nodes nominated 2+ times, squares indicate nodes nominated 1 time, triangles indicate nodes nominated 0 times.

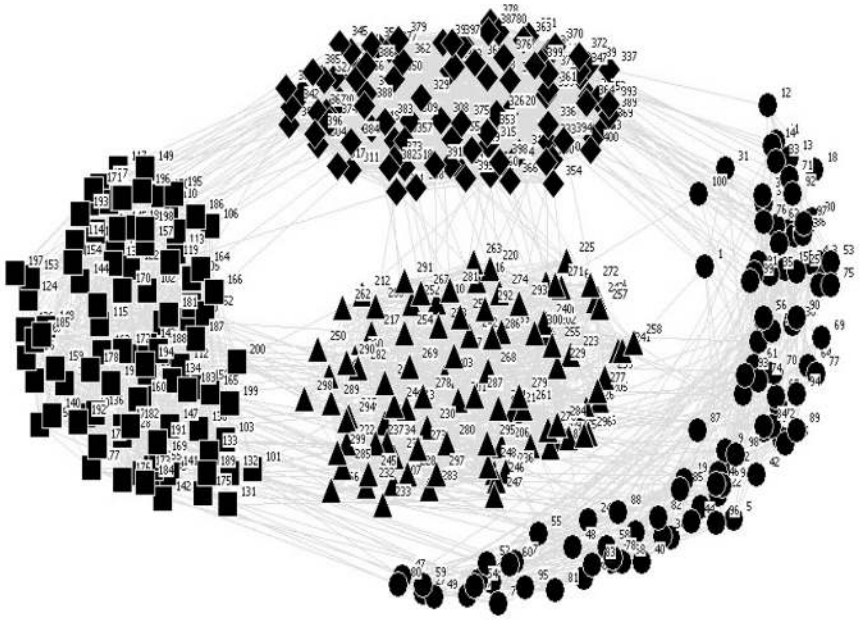

Figure 2.
400-Node Test Network.

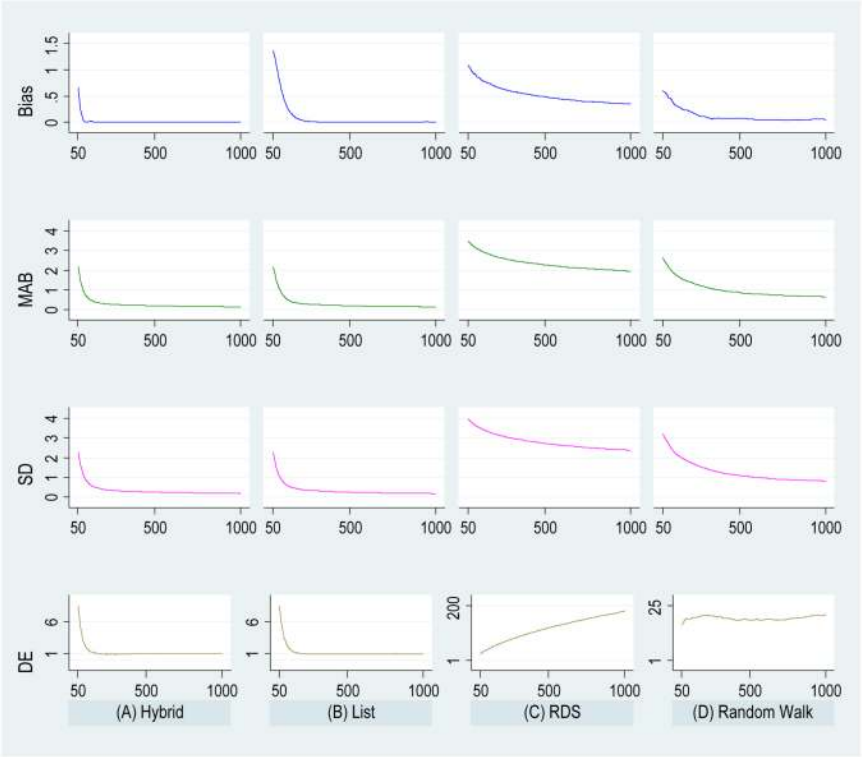

Figure 3. Test Network sampling results, (400 nodes, dependent variable: average degree)

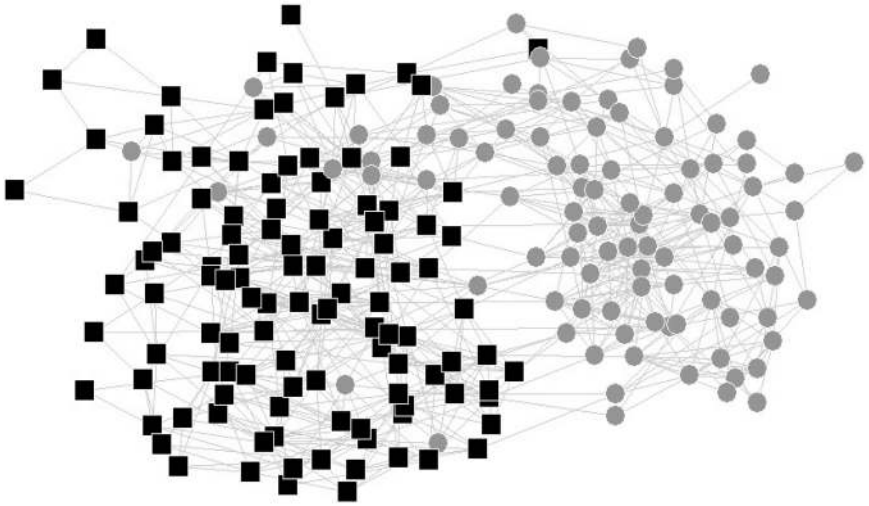

Figure 4. Add Health Network # 112
Notes: Nodes colored by student race. [Black squares = White , Gray circles = Non-white]

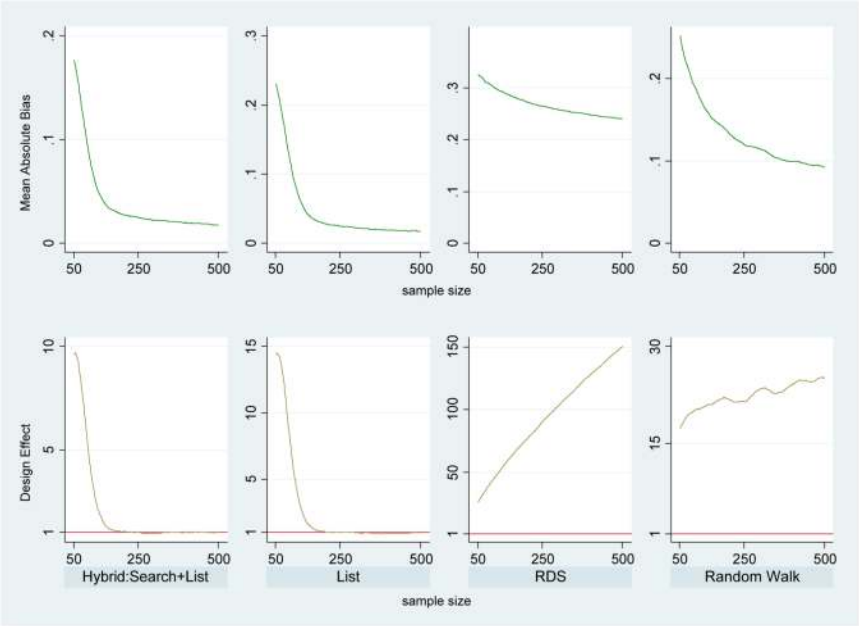

Figure 5. Sampling results from ADH network 112

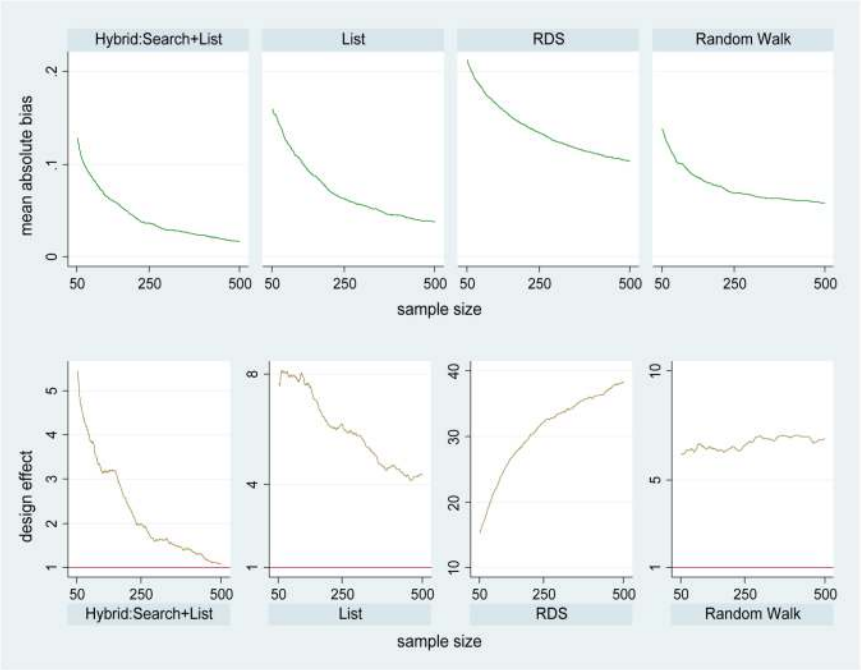

Figure 6. Sampling results the largest Facebook university network (16,280 nodes, dependent variable: proportion freshman)

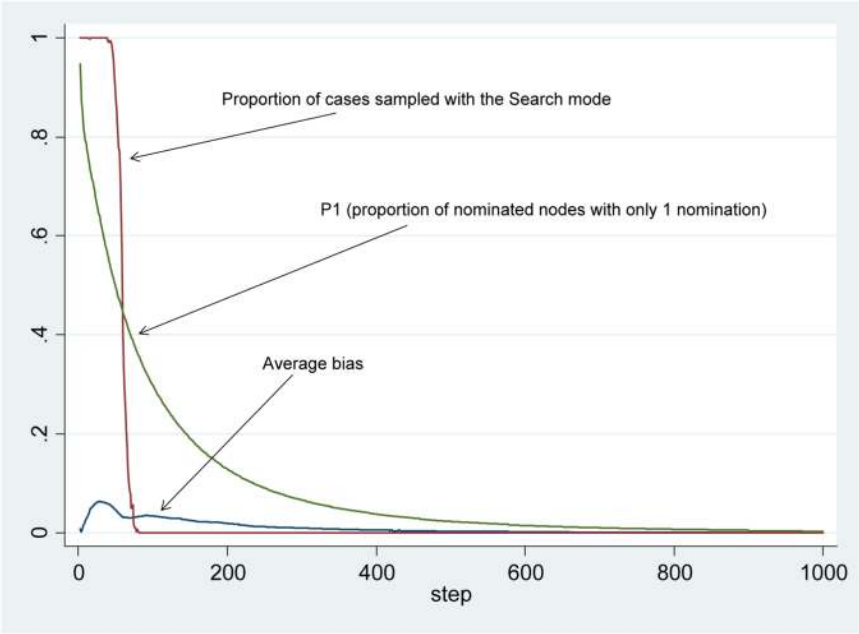

Figure 7. Diagnostic Properties for Facebook network #10 by sample size (step)

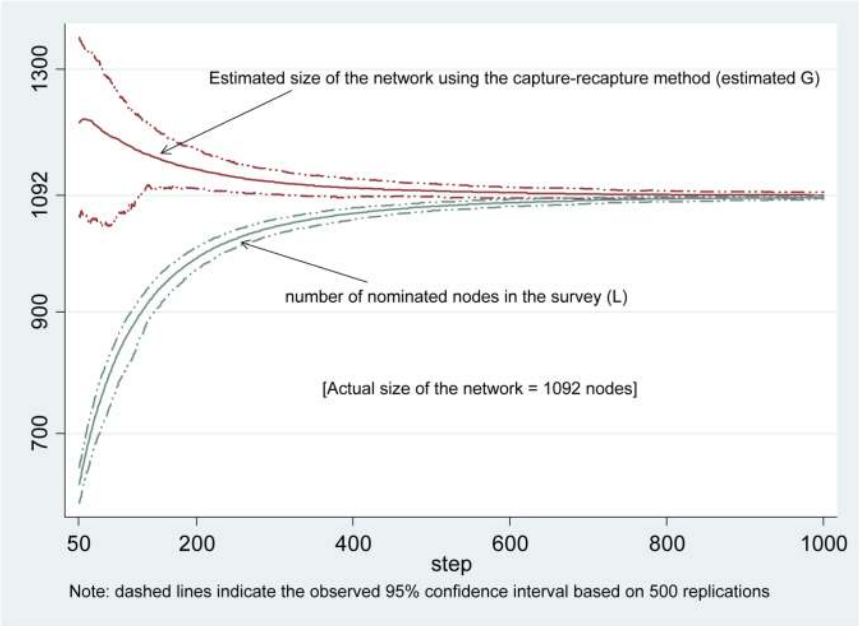

Figure 8. Number of nominated nodes and estimated network size by sample size (step), Facebook network

Table 1

Glossary of terms used

Term	Definition
List mode	The process of evenly sampling from the list L of nodes that have been nominated (see section 3B).
Even sampling	Sampling all nominated nodes (all nodes $\in L$) at the same rate. Newly nominate nodes are sampled at the current cumulative sampling rate of the existing sample (see section 3B).
Search mode	Using bridge nodes to sample unexplored clusters of the network (see section 3C).
Hybrid approach	The combination of the List and Search modes, see section 3C.
A	The population that is being sampled.
G	Number of nodes (people) in the network connecting population A.
$\widehat{G}=\frac{L_s}{1-P1}$	The estimated size of the network using the capture-recapture method. See Equation 5.
S	The current "step" of the sample—the number of interviews that have been completed
L	The list of nodes/people that have been nominated by respondents on the network rosters.
L_S	The number of nodes that have been nominated by step S of the sample
E_S	The number of nodes eligible for sampling via the List mode in step S of the sample (see section 3.B.1).
P1	The proportion of nominated nodes $\in L$ that have been nominated once and only once in the sample.
S1	The step that even sampling begins for the List mode. See section 3B.
CSR_j	The cumulative sampling rate for node j. See Equation 2
ESR	The even sampling rate. The cumulative sampling rate since the start of even sampling. See Equation 3.
d_i	The degree of node i. This is the number of friends that node I has in population A.
p_nom_i	The probability that node i has been nominated after S steps.
Bridge node	A node connecting different clusters of the network
c_j	The number of j's friends who have been nominated once and only once in the sample.
$P1_j=\frac{C_j}{d_j}$	The proportion of j's friends who have been nominated 1 time in the sample.
A1	The minimum level of P1 for the use of the Search mode. If P1>A1 use the Search mode, if $P1 \leq A1$ use the List mode. A1=0.4 is used in the sampling simulations.
Average bias	The degree to which the sample means differs from the population mean, averaged across all replications. See Equations 7 and 8.
Mean absolute bias (MAB)	The expected magnitude of the bias on any single replication, see Equation 9.
Sampling variance	How much variance there is in the sample estimates, see Equation 10.
Design effect (DE)	The ratio of the sampling variance of the method to the sampling variance of simple random sampling, see Equation 11.

Table 2

Example of sampling from the test network: step by step sampling rates.

					Cumulative sampling rate of Nodes (CSR). (Note: the inverse of this rate is used as a sampling weight, see section).				
Column #: 1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
"Step"	***revealed* Network size (L)**	**# of Elig. nodes**	**Even sampling Rate (ESR)**	**Sampling rate for step**	**Node 1**	**Node 101**	**Node 201**	**Node 301**	**ID of interviewed node**
0	0	0	0	NA	0	0	0	0	92 "seed"
1	8	8	0	0.125	0	0	0	0	25
2	14	14	0	0.071429	0	0	0	0	76
3	20	20	0	0.05	0	0	0	0	21
4	22	22	0	0.045	0	0	0	0	23
5	25	25	0	0.04	0.04	0	0	0	14
After 5 steps, the revealed network size is 25 [i.e., 25 different nodes have been nominated] Node 1 is nominated for the first time in step 5, and Node 301 is nominated for the first time in step 7									
6	46	46	0	0.021739	0.061739	0	0	0	378
7	63	63	0	0.015873	0.077612	0	0	0.015873	323
8	68	68	0	0.014706	0.092318	0	0	0.030579	56
9	73	73	0	0.013699	0.106017	0	0	0.044278	15
After 39 steps, Nodes 1 and 301 have large cumulative sampling rates. They will be temporarily excluded from the pool of eligible nodes once even sampling is turned on, until their cumulative sampling rate is less than or equal to the even sampling rate.									
39	194	194	0	0.005155	0.323978	0	0	0.262239	322
40	199	199	0	0.005025	0.329004	0	0	0.267264	81
41	203	100	0.01	0.01	0.329004	0	0	0.267264	79
"Even sampling" is turned on in step 41 because the network size > 200. Now, all newly nominated nodes are sampled at the current even sampling rate. Node 101 is nominated for the first time in step 43.									
42	211	100	0.02	0.01	0.329004	0	0	0.267264	138
43	213	100	0.03	0.01	0.329004	0.03	0	0.267264	177
44	218	100	0.04	0.01	0.329004	0.04	0	0.267264	122
After 44 steps, the even sampling rate is .04. Nodes 1 and 301 are still excluded from sampling eligibility because their cumulative sampling rate is greater than the cumulative even sampling rate.									
79	357	270	0.205092	0.003704	0.329004	0.205092	0	0.267264	199
80	358	NA	0.205092		0.329004	0.205092	0	0.267264	294

					Cumulative sampling rate of Nodes (CSR). (Note: the inverse of this rate is used as a sampling weight, see section).				
Column #: 1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
"Step"	***revealed* Network size (L)**	**# of Elig. nodes**	**Even sampling Rate (ESR)**	**Sampling rate for step**	**Node 1**	**Node 101**	**Node 201**	**Node 301**	**ID of interviewed node**
81	364	NA	0.205092		0.329004	0.205092	0	0.267264	287
82	367	NA	0.205092		0.329004	0.205092	0.205092	0.267264	217
Node 201 is nominated for the first time in step 82. Steps 80-82 are "catch-up" interviews of newly nominated nodes . In Step 251, all 400 nodes have been nominated, and the sample is approximately equivalent to a simple random sample									
248	399	399	0.633599	0.002506	0.632375	0.633599	0.633599	0.633053	101
249	399	399	0.636105	0.002506	0.634881	0.636105	0.636105	0.63556	389
250	399	399	0.638612	0.002506	0.637388	0.638612	0.638612	0.638066	300
251	400	400	0.641112	0.0025	0.639888	0.641112	0.641112	0.640566	275

Sociol Methodol. Author manuscript; available in PMC 2013 October 22.

Table 3

Test network sampling results (400 nodes, dependent variable: average degree)

		Method				
Variable	**Steps**	**Hybrid**	**List**	**Naive List**	**RDS**	**Random Walk**
Average Bias	250	.00036	.00472	.1472	.1278	.02825
	500	.000084	.000014	.07651	.09775	.01458
	1000	5.00e-06	.001203	.03781	.06936	.01095
Mean Absolute Bias	250	.05281	.05526	.1831	.5254	.2555
	500	.04001	.03938	.09757	.4568	.173
	1000	.02756	.02808	.05165	.3933	.1304
Design Effect	250	.8986	.9615	5.49	76.95	20.72
	500	1.015	.962	3.221	119.6	18.8
	1000	.9963	.9845	2.066	181	21.1
Standard Deviation	250	.06711	.06942	.1659	.6211	.3223
	500	.05045	.0491	.08985	.5476	.2171
	1000	.03533	.03512	.05088	.4762	.1626

Sociol Methodol. Author manuscript; available in PMC 2013 October 22.

Table 4

Table 4A Descriptive statistics for Add Health networks [N=62]

Variable	Mean	25th percentile	median	75th percentile	Minimum	Maximum
Proportion white	0.562	0.424	0.632	0.718	0.215	0.794
Homophily	0.380	0.221	0.358	0.509	–0.021	0.816
Nodes (number of people)	494.8	247	395	586	110	1610
Edges (number of ties)	3,441.5	1,638	2,622	4,276	608	12,794
Mean degree	8.48	7.43	8.46	9.78	4.28	11.99
Y-mean degree difference	10.36	4.40	7.90	13.62	0.02	42.61

Table 4B Descriptive statistics for Facebook networks [N=100]

Variable	Mean	25th percentile	median	75th percentile	Minimum	Maximum
Proportion freshman	0.282	0.231	0.281	0.336	0.141	0.462
Homophily	0.786	0.745	0.810	0.846	0.378	0.900
Nodes (number of people)	4,635.8	2,186	3,694.5	6,638	331	16,278
Edges (number of ties)	238,916.3	83,517	207,308	339,513	6,672	997,614
Mean degree	81.14	66.71	79.95	93.95	31.99	156.06
Y-mean degree difference	1.25	0.29	0.54	1.29	0.01	11.33

Table 5

Sampling results from Add Health network 112

		Method				
Variable	**Steps**	**Hybrid**	**List**	**Naive List**	**RDS**	**Random Walk**
Average Bias	250	.001193	.001793	.004337	.005243	.006462
	400	.001797	.0006971	.003917	.007063	.002586
	500	.0009851	.0001629	.003245	.006052	.002788
Mean Absolute Bias	250	.02489	.02484	.07168	.2647	.1201
	400	.01988	.01888	.04704	.2479	.09969
	500	.01767	.01759	.03981	.2405	.09269
Design Effect	250	1.002	.990	7.554	89.73	21.47
	400	1.002	.9243	5.285	127.4	24.14
	500	.9961	.9852	4.697	150.2	25.16
Standard Deviation	250	.03161	.03142	.0868	.2992	.1463
	400	.02499	.024	.0574	.2818	.1227
	500	.02229	.02217	.0484	.2737	.112

Table 6

Overall Add Health and Facebook Results.

	Network Source and Dependent Variable		
Column #	**1**	**2**	**3**
	Add Health[++]	**Add Health**[++]	**Facebook**[++]
Dependent variable	**Proportion white**	**Average degree**	**Proportion freshman**
Average Design effects			
NSM: hybrid	1.110	1.069	1.198
NSM: list	1.236	1.132	2.267
Naive list	5.605	2.152	22.85
RDS [**]			83.931
RDS full data [*]	66.883	25.206	86.135
RW [**]			12.031
RW full data [*]	12.698	6.652	19.107
Median Design Effects			
NSM: hybrid	1.055		1.198
NSM: list	1.093		2.267
Naive list	3.317		22.85
RDS [**]			83.931
RDS full data [*]	47.766		86.135
RW [**]			12.031
RW full data [*]	9.273		19.107
Average bias			
NSM: hybrid	0.0019	0.0328	0.0048
NSM: list	0.0021	0.0424	0.0117
Naive list	0.0168	0.3172	0.0483
RDS [**]			0.1363
RDS full data [*]	0.0091	0.0867	0.018
RW [**]			0.1047
RW full data [*]	0.0032	0.0228	0.0049
Median Average. Bias			
NSM: hybrid	0.0011		0.0037
NSM: list	0.0013		0.0091
Naive list	0.0116		0.0470
RDS [**]			0.1060
RDS full data [*]	0.005		0.0164
RW [**]			0.0691
RW full data [*]	0.0022		0.0044

Notes:

[+]The Add Health networks are run on the complete network data for each school.

[++]Because the average number of ties in the Facebook data was so high, the maximum number of ties was truncated at 20 for the NSM hybrid, list, and naïve list methods. This makes it harder for these methods. See the text for details.

[**]indicates that RDS and RW were run on the same truncated data (maximum degree of 20) as the NSM approaches for the Facebook data.

[*]indicates that RDS and RW were run on the full data, with no limit on maximum degree. See text for details.

Table 7

Sampling results the largest Facebook university network (16,280 nodes, dependent variable: proportion freshman)

		Truncated network data (maximum degree = 20)***				Full network data	
Variable	**Steps**	**Hybrid**	**List**	**RDS**	**Random Walk**	**RDS**	**Random Walk**
Average Bias	250	.0199	.03452	.4246	.04032	.03265	.03811
	400	.006125	.02586	.4388	.04299	.03769	.03649
	500	.0003046	.02095	.445	.0448	.03948	.0359
Mean Absolute Bias	250	.03642	.06275	.4247	.06897	.134	.07709
	400	.02322	.04476	.4388	.06162	.1118	.06553
	500	.01661	.0378	.445	.05838	.1035	.06044
Design Effect	250	1.997	6.182	26.00	6.575	32.12	9.75
	400	1.422	4.616	28.65	6.992	35.82	10.48
	500	1.079	4.379	29.63	6.893	38.34	10.61
Standard Deviation	250	.04115	.07231	.1483	.07455	.1651	.09082
	400	.02745	.0494	.1231	.06078	.1378	.07443
	500	.02138	.04303	.1120	.05397	.1275	0.0670

Notes:

*** The truncated network data limits the number of friends each respondent can nominate to 20. This is designed to make it harder for the NSM Hybrid and List approaches (see the text for details).

Table 8

OLS regression results for the Design Effect, combined Add Health and Facebook networks

		Method			
	Hybrid	**List**	**Naive List**	**RDS**	**Random Walk**
VARIABLES					
Homophily	0.158	−0.136	28.36***	284.1***	40.41***
	(0.144)	(0.338)	(5.563)	(15.16)	(2.757)
Add Health network	−0.182	0.471	16.67***	71.37***	6.933**
	(0.114)	(0.268)	(4.427)	(11.99)	(2.181)
Ln(nodes)	−0.00576	0.964***	12.14***	−8.917**	−2.030***
	(0.0278)	(0.0654)	(1.252)	(2.930)	(0.533)
Average Degree	−0.00111	−0.00451	−0.0159	−0.0950	0.0436
	(0.00121)	(0.00285)	(0.0494)	(0.127)	(0.0232)
Y-Mean degree difference	0.00714	0.0130	0.00201	−0.139	0.207**
	(0.00376)	(0.00886)	(0.144)	(0.397)	(0.0722)
Constant	1.202***	-5.133***	−94.72***	−56.61*	0.0898
	(0.236)	(0.555)	(9.751)	(24.88)	(4.525)
Observations	162	162	162	162	162
R-squared	0.076	0.698	0.658	0.718	0.687
Mean design effect	1.16	1.85	15.48	78.76	16.65
S.D. design effect	0.25	1.05	15.90	48.54	8.38

Notes: Standard errors in parentheses

*** p<0.001,

** p<0.01,

* p<0.05

Sociol Methodol. Author manuscript; available in PMC 2013 October 22.

Table 9

OLS regression results for the average bias, combined Add Health and Facebook networks**

	Method				
	Hybrid	**List**	**Naive List**	**RDS**	**Random Walk**
VARIABLES					
Homophily	0.000602	0.00208	0.0149	0.0139 *	0.00761 ***
	(0.00208)	(0.00387)	(0.00815)	(0.00632)	(0.00177)
Add Health network	0.00192	0.00429	−0.00768	−0.00190	−0.000757
	(0.00164)	(0.00306)	(0.00648)	(0.00500)	(0.00140)
Ln(nodes)	0.00169* ***	0.00637* ***	0.0115 ***	−0.00180	−0.00127 ***
	(0.000401)	(0.000748)	(0.00183)	(0.00122)	(0.000342)
Average Degree	1.89e-05	5.38e-06	9.86e-05	0.000169 **	1.44e-05
	(1.75e-05)	(3.25e-05)	(7.23e-05)	(5.31e-05)	(1.49e-05)
Y-Mean degree difference	4.51e-05	9.72e-05	0.00130 ***	0.000767 ***	4.71e-05
	(5.44e-05)	(0.000101)	(0.000211)	(0.000165)	(4.64e-05)
Constant	−0.0110 **	−0.0424 ***	−0.0645 ***	0.00715	0.00806 **
	(0.00341)	(0.00635)	(0.0143)	(0.0104)	(0.00291)
Observations	162	162	162	162	162
R-squared	0.241	0.512	0.669	0.303	0.238
Mean average bias	0.0037	0.0080	0.0348	0.0146	0.0043
S.D. average bias	0.0041	0.0094	0.0236	0.0128	0.0034

Notes:

$$\text{average bias} = \frac{1}{J}\sum_{j=1}^{J}\left|\frac{\sum_{i=1}^{500} bias_{ij}}{500}\right|$$

**"average bias" refers to the absolute value the average bias for 500 replications on all the networks, see the text for explanation. where $bias_{ij}$ is the bias (error) on the ith sample from the jth network.

Standard errors in parentheses

*** p<0.001,

** p<0.01,

* p<0.05