Noname manuscript No.
(will be inserted by the editor)

Surveying hard-to-reach groups through sampled respondents in a social network

A comparison of two survey strategies

Tyler H. McCormick · Ran He · Eric Kolaczyk · Tian Zheng

Received: date / Accepted: date

Abstract The sampling frame in most social science surveys misses members of certain groups, such as the homeless or individuals living with HIV. These groups are known as *hard-to-reach* groups. One strategy for learning about these groups, or subpopulations, involves reaching hard-to-reach group members through their social network. In this paper we compare the efficiency of two common methods for subpopulation size estimation using data from standard surveys. These designs are examples of *mental link tracing* designs. These designs begin with a randomly sampled set of network members (nodes) and then reach other nodes indirectly through questions asked to the sampled nodes. Mental link tracing designs cost significantly less than traditional link tracing designs, yet introduce additional sources of potential bias. We examine the influence of one such source of bias using simulation studies. We then

Tyler McCormick is partially supported by NIAID grant R01 HD54511. This work was partially completed while McCormick was supported by a Google PhD Fellowship in Statistics. The research of Tian Zheng is, in parts, supported by NSF grants DMS-0714669 and SES-1023176, NIH grant R01 GM070789, and a 2010 Google research award. Eric Kolaczyk is supported by ONR award N000140910654.

Tyler H. McCormick
Department of Statistics
University of Washington
Box 354320
Seattle, WA 98195
Tel.: 206-221-6981
Fax: 206-221-6873
E-mail: tylermc@u.washington.edu

Tian Zheng
Department of Statistics
Columbia University
1255 Amsterdam Ave MC 4690
New York, NY 10027
Tel.: 212-851-2134
Fax: 212-851-2163
E-mail: tzheng@stat.columbia.edu

demonstrate our findings using data from the General Social Survey collected in 2004 and 2006. Additionally, we provide survey design suggestions for future surveys incorporating such designs.

Keywords Aggregated Relational Data · Egocentric nominations · Hard-to-reach groups · Mental link tracing design · Sampling · Social network

1 Introduction

Standard surveys often miss members of certain groups, known as *hard-to-reach groups*. Members of these groups may be physically difficult to reach using standard recruitment techniques (homeless individuals are unlikely to be reached using random-digit dialing, for example). In other cases, members of some groups may be reluctant to self-identify because of social pressure or stigma (Shelley et al, 2006). A third group of individuals is difficult to reach because of issues with both access and reporting (commercial sex workers, for example). Despite the difficulty reaching these groups, information about hard-to-reach groups is often important for public health and epidemiological monitoring and evaluation.

Even basic information about these groups, such as the group size, is typically unknown. *Link Tracing* designs are one approach to counting members of hard-to-reach groups. These designs recruit respondents directly from other respondents' networks (see Salganik and Heckathorn (2004), for example), making the sampling mechanism similar to a stochastic process on the social network (Goel and Salganik, 2009). Link tracing designs affords researchers face-to-face contact with members of hard-to-reach groups, facilitating exhaustive interviews and even genetic or medical testing. The price for an entrée to these groups is high, however, as the sampling mechanism requires physically locating the nominated respondents' network members. Estimates from link tracing designs are also biased because of the network structure captured during selection, with much statistical research devoted to re-weight observations from link tracing designs to have properties resembling a simple-random-sample. This bias is an issue for estimating the size of a hard-to-reach group and makes link tracing designs unsuitable for measuring information about the general population. Recent statistical advances for one such design, Respondent-driven Sampling, are presented in work such as Handcock and Gile (2010).

Other approaches to reaching members of these populations through their social network involve accessing respondents' social networks indirectly. In contrast to designs presented in Handcock and Gile (2010), these *mental link tracing* designs use respondents selected through standard surveys (random digit dialing telephone surveys, for example) and ask respondents questions about actors in their social network. Mental link tracing designs are related to designs used in health statistics known as multiplicity sampling (see Sirken (1970), for example). In contrast to traditional link tracing designs, these methods do not require reaching members of the hard-to-reach groups directly. Instead,

they access hard-to-reach groups indirectly through the social networks of respondents on standard surveys. Mental link tracing designs never afford direct access to members of hard-to-reach populations, making the level of detail achievable though physically tracing a respondent's network impossible with indirectly observed data. Unlike link tracing designs, however, these methods require no special sampling techniques and are easily incorporated into standard surveys. Indirectly observed network data are, therefore, feasible for a broader range of researchers across the social sciences, public health, and epidemiology to implement with significantly lower cost than link tracing. Recent work with this data demonstrates that features of network structure, such as homophily (the tendency for actors to form relationships with similar others), are distinguishable even after the aggregation described above McCormick et al (2010).

In this paper we compare the efficiency of two common methods for subpopulation size estimation using data from standard surveys. First, *Aggregated Relational Data (ARD)* asks respondents how many individuals they know in a particular group of interest. Researchers view the number known in a group of interest as a proportion of the respondent's network (which requires estimating the respondent's total network size) and then "scale-up" from the total proportion of respondents' networks to the size of the group of interest in the overall population. *Egocentric nominations* involve first asking a respondent to nominate a pre-chosen number of members from their network. An enumerator then goes one-by-one through the list of nominated individuals and asks detailed questions. To obtain and estimate the total size of a particular group in the population, the total proportion of the nominated individuals across respondents is scaled to the size of the total population. A key feature of both mental link tracing designs and traditional link tracing designs is the confounding of the sampling mechanism with the underlying social network. In both cases there are two distinct, but not independent, processes: (i) tie formation and (ii) nomination. For our purposes we assume tie formation has already occurred. We still cannot ignore this process, however, since the set of potential alters a respondent could nominate is limited to the people with whom the respondent has ties. We focus on three types of error which can cause bias in mental link tracing estimates. First, *barrier effects* are a potential source of bias for both estimates. Barrier effects occur when there are departures from random mixing (the propensity for a tie between two actors depends only on their degree) in the underlying network. With barrier effects, some individuals systematically know more (or fewer) members of a specific subpopulation than would be expected under random mixing. Barrier effects are often the result of homophily. For example, people tend to know others of similar age and gender (McPherson et al, 2001). While barrier effects come about because of the tie formation process in the network, the other two sources of error we consider arise as part of the nomination process. A second source of error, *calibration bias*, occurs when respondents have difficulty recalling accurately the number of members of a group they know. Calibration bias typically is more severe for larger groups. Calibration bias is particularly

influential in ARD. The third source of error is *preferential nomination bias*, which typically manifests in egocentric nominations. Preferential nomination bias occurs when a respondent is required to nominate a subset of the people they know in a group. Under our local model we assume that the respondent decides which alters to nominate by choosing randomly. This is unlikely to be the case, however, and may lead respondents to nominate a subset of alters which are not representative of their the overall set of individuals they know in that group.

In evaluating these methods, we find that the two sampling strategies have complimentary strengths. In the absence of the sources of bias described above, ARD is consistently preferable since using egocentric nominations produces a smaller set of (indirectly) reached alters. Using simulation, however, we find that the performance of ARD depends heavily on the level of calibration bias and barrier effects. ARD was, in fact, more susceptible to barrier effects than egocentric nominations. Thus, ARD requires more statistical modeling to overcome barrier effects, but reaches more alters than data collected using egocentric nominations.

We begin by describing the two commonly used sampling schemes in more detail in Section 2. We then, in Section 3, compare the performance of these methods using three examples: a simulation study, data from a large online social network, and data collected from the General Social Survey. We give design recommendations for future surveys and provide a discussion in Section 4.

2 Two sampling methods

In this section we present two commonly used mental link tracing designs. Both of these designs begin with a (non-network) sample of respondents. These respondents then answer questions about members of their social network who are not directly observed. A key distinction between these methods and link tracing designs is that network structure does not drive the recruitment of survey respondents. Rather, the network structure impacts recruitment at the second stage from each of many independent starting points. In the remainder of this section we describe in detail two types of network data often collected on standard surveys.

2.1 Aggregated Relational Data

In Aggregated Relational Data (ARD), respondents answer questions of the form "How many X's do you know?" for a group, X. Defining *know* defines the relationship that forms the network of interest. We can make this relationship diffuse by using a broader definition of know, more rigorous to capture a set of more intimate acquaintances, or use a different relationship entirely, such as *trust* to measure yet another network. The 2006 General Social Survey (GSS),

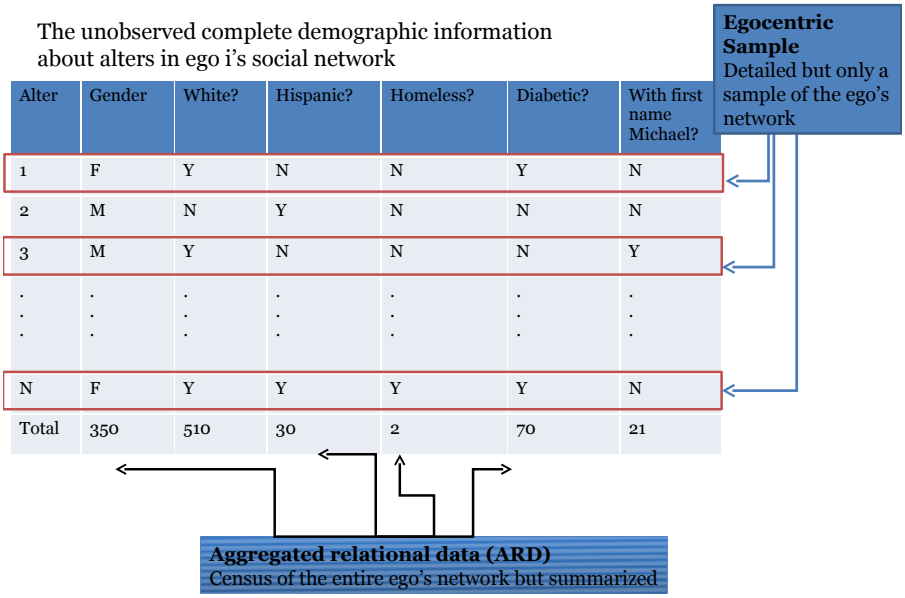

Fig. 1 A graphical representation of ARD and egocentric nominations.

which we analyze later, uses ARD questions using two relationships, knowing and trusting. Knowing is defined in the following manner:

> I'm going to ask you some questions about all the people that you are acquainted with (meaning that you know their name and would stop and talk at least for a moment if you ran into the person on the street or in a shopping mall). Again, please answer the question as best you can.

Given this network, "How many X's do you know?" data are a type of network sample. If respondents could recall perfectly from their network and had full knowledge of all of the group memberships of all alters, then these data would be "equivalent" to asking a respondent if they know each member of a particular group of alters. If every Michael in the US population were standing in a room, for example, we could imagine asking the respondent if he/she has a tie with each person in the room. Rather than reporting these ties individually as in the complete network case, however, our data consist of only the total number of links the respondent has with Michaels. The features of this design are illustrated in Figure 2 where the respondent does not report information about any particular alter but instead gives the total number of alters known in each of the columns.

The estimator typically used with ARD for the proportion of individuals in a population that belong to a hard-to-reach group is, for a sample of size n,

$$\hat{R}_{ARD} = \frac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} \hat{d}_i} \tag{1}$$

where x_i is the number of people respondent i knows in the population of interest and d_i is the personal network size (degree) of person i.

McCormick et al (2010) show that each response can be viewed as a binomial random variable with the number of trials being the number of alters in the groups of interest and the probability being the ego's degree over the total population size. Since the degree (or the ego's network size) itself needs to be estimated, the variance of the subpopulation size estimator depends on the variance of the the degree estimator.

The variance of the subpopulation size estimator also depends on the mean degree of the sample, with higher average cluster sizes resulting in lower variance. In the case of ARD, degree is related to the definition of "know." A more stringent definition of know (trusting the alter with a loan of a large sum of money, for example) will result in a lower average cluster size and a broader definition will produce larger clusters. Since the underlying population remains the same, using a broader definition of know will allow even more respondents to be reached and mitigate the impact of the clustering. In practice, respondents must accurately nominate members of the group of interest even as the number of alters they are asked to consider increases.

2.1.1 Calibration bias

ARD asks respondents to perform a complicated psychological exercise, which introduces potential sources of bias. One such source of bias comes from respondents having difficulty recalling accurately the members of their network who belong to a particular category. One way to conceptualize this bias would be as respondent recalling inaccurately from their true personal network. This phenomenon is difficult to quantify at the level of the individual respondent, however. Instead, we conceive of calibration bias as a respondent recalling accurately from a subset of their total personal network, their *recalled network*. The level of misspecification between the true and recalled networks can then be estimated at the population level based on discrepancies between estimates using ARD and information on population sizes for populations with sizes available from official sources.

Since the bias comes about from a mis-calibration between the ego's actual network and their recalled network, we refer to this bias as *calibration bias*. Previous work (Zheng et al, 2006; Killworth et al, 2003) has noted that respondents underrecall the number of people they know in large subpopulations (e.g., people named Michael) and overrecall the number of people they know in small subpopulations (e.g., people who committed suicide).

2.2 Egocentric nominations

In contrast to ARD, which measure a (potentially) large subset of the ties in the network indirectly, egocentric nominations measure a relatively small number of ties in greater detail. These data collect a specific subset of ties sent by the *ego* (respondent) and a small number of recipients, or *alters*. Typically, respondents are asked to nominate a number of relations. For each person they nominate the interviewer then asks follow-up questions about each alter.

The 2004 GSS includes questions which ask "From time to time, most people discuss important matters with other people. Looking back over the last six months–who are the people with whom you discussed matters important to you? Just tell me their first names or initials." Egocentric nominations are also in practice related to cluster sampling. The egocentric nominations use a sample of the ssus, rather than a census. In most cases this is not a random sample, but rather is related to the strength of the tie of interest (a common prompt asks for a respondent's "three best friends," for example). This truncation introduces an additional source of between-cluster variation.

We can again view the standard estimator for egocentric nominations as a ratio estimator taking the form

$$\hat{R}_{ego} = \frac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} n_i^{ego}}.$$

Like the ARD estimates, the variance and bias of the egocentric nomination estimates depend on the number of respondents in the survey. Unlike ARD, where respondents do a census of their local network, egocentric nominations have additional variability from the process by which respondents choose which subset of potential alters to nominate. This process may be an additional source of bias which is not reflected in these computations, as described subsequently.

2.2.1 Preferential nomination bias

To this point, we assumed that the respondents selected among potential alters in a simple random sample. In practice, respondents likely select based on a myriad of factors which may induce bias known as *preferential recall bias*. If asked about their friends, for example, respondents may answer in order starting with their "best" friends. These individuals may be different from the others the respondent considers friends, inducing bias. Alternatively, the respondent may answer based on the most recent contact. In both cases, these issues will only impact the validity of the estimates if there is an association between being a member of the group of interest and the sampling mechanism (if living with HIV prevents one from becoming friends with some alters, for example).

3 Comparing efficiency in subpopulation size estimation

In this section we present three numerical studies which demonstrate the performance of the two mental link tracing designs. In the first two examples, we perform simulation experiments on two networks. The first example is a simulation study using a network simulated using the algorithm of Blitzstein and Diaconis (2006). Using this network, we are able to manipulate various aspects of the sampling mechanism and evaluate the effect on the resulting estimates. The next two examples involve actual data from large networks. As a second example, we use a network of friendships between Slashdot[1] users. Using an existing, large network provides insights into the performance of the method under realistic circumstances. The third example uses actual ARD and Egocentric nomination questions from the 2004 and 2006 General Social Surveys. Unlike the other two examples, the underlying network is not known in this example and that ARD and egocentric nomination questions were fielded to actual respondents, rather than being simulated from complete network data. Since these are actual data, the various sources of bias previously described cannot be intentionally manipulated.

3.1 Simulated network

We first present results using a simulated network. To simulate the network, we used the sequential algorithm of Blitzstein and Diaconis (2006). This approach involves choosing a vertex in each step and adding the edge such that the residual degree sequence is still graphical. The algorithm starts with the degree sequence, d, and add edges until the degree sequence is reduced to 0. We generated a degree sequence following a lognormal distribution.

We began by simulating a network of size five-thousand. We then performed the following:

1. Simulate hard-to-reach group membership with probability p.
2. Sample n nodes using a SRS.
3. (Under some set-ups) simulate recall issues and/or assortative mixing.
4. Poll the vertices connected to each selected node and compute $\hat{N}_k^{ego}$ and $\hat{N}_k^{ARD}$.
5. Repeat 300 times for each specification at various values of n and N_k.

In our simulation experiments, we use two different criteria to evaluate efficiency. One is coefficient of variation of the root mean squared error (RMSE), which is defined as the relative RMSE

$$RRMSE = \frac{\sqrt{MSE(\hat{p})}}{\bar{\hat{p}}} = \frac{\sqrt{\frac{1}{m}\sum_{l=1}^{m}(\hat{p}_l - \bar{\hat{p}})^2}}{\bar{\hat{p}}},$$

[1] Slashdot is a technology news blog. Slashdot recently introduced a feature, known as Slashdot Zoo, which allows users to connect to one another as friends

where m is the times of iteration. The smaller value of RRMSE, the more efficient the method is.

Another criteria of efficiency is effective sample size n_{eff}, which is defined as the size of simple random sample needed such that the variance of estimator $\hat{p}$ is the same as that of sampling method we use. Suppose that using simple random sample method, we learn that x respondents belong to group X. In our simulation experiments, x simply follows a Binomial distribution, that is, $x \sim \text{Binomial}(n_{\text{eff}}, p)$, where p is the true infection probability. Therefore, the estimator of p using simple random sample method (SRS), $\hat{p}_{SRS} = \frac{x}{n_{\text{eff}}}$ and the variance of this estimator is $Var(\hat{p}_{SRS}) = \frac{p(1-p)}{n_{\text{eff}}}$. Therefore, the equation

$$\frac{p(1-p)}{n_{\text{eff}}} = MSE(\hat{p}),$$

implies $n_{\text{eff}} = \frac{p(1-p)}{MSE(\hat{p})}$. The larger value of n_{eff}, the more efficient the method is.

3.1.1 Calibration bias

A fundamental difference between the two designs is the number of alters a respondent must consider when answering the question. In an egocentric design, respondents only consider small sets of initially nominated alters. In ARD, however, respondents mentally poll their entire network to count the number of alters belonging to each population of interest, making calibration bias a common issue with ARD. In this section, we present results of a simulation study which further explores the impact of calibration bias. For these simulations, we assume membership in the hard-to-reach group is distributed randomly through the population and the propensity to form ties is simply inversely proportion to the respondents' degree. This simulation set-up is simplistic with respect to network structure. This simplicity affords the opportunity to isolate the impact of calibration bias on the two estimates. and confirm the results presented in the previous sections if the assumptions made by the ratio estimators are valid.

To simulate calibration bias, we used a calibration curve developed by McCormick and Zheng (2007). As mentioned in the previous section, we adopt the interpretation interpretation of calibration bias where respondents answer alternative questions based on their own restrictive definition of know which defines the distinction between their actual and recalled personal networks. Therefore, the size of the group of interest in the recalled network, $N_k^{'}$, is smaller than the subpopulation size of people in the respondent's actual personal network, N_k. The following calibration curve interpolates between the actual and recalled networks, $p_k = \frac{N_k}{N}$ and $p_k^{'} = \frac{N_k^{'}}{N}$:

$$p_k^{'} = \begin{cases} p_k \left[\frac{b}{p_k} \exp\left(\frac{1}{a}\left(1 - \left[\frac{b}{p_k}\right]^a\right)\right)\right]^{1/2} & if\ p_k \geq b \\ p_k & if\ p_k < b, \end{cases} \tag{2}$$

where $a > 0$ and $0 < b < 1$ are constants. The derivation of this curve is detailed in McCormick and Zheng (2007), though the main features of the curve are determined by the two parameters, a and b. The curve assumes near perfect recall for small alter groups begins accounting for calibration bias at a particular group size, b. The severity of the correction increases as the size of the alter group increases with the rate of interest controlled by the second parameter a. For the simulations in this section, we evaluated the performance of the estimators at a variety of different levels of calibration bias and show an illustrative subset of those results here.

We performed the simulation experiment described above for hypothetical hard-to-reach groups ranging from around .02% of the total population to about .2%. We also evaluated a number of different sample sizes. Figures 3.1.1 and 3.1.1 are from the same simulation experiments and display the resulting performance of the estimators in terms of RRMSE and effective sample size. When there is no recall issue, ARD consistently outperforms the egocentric

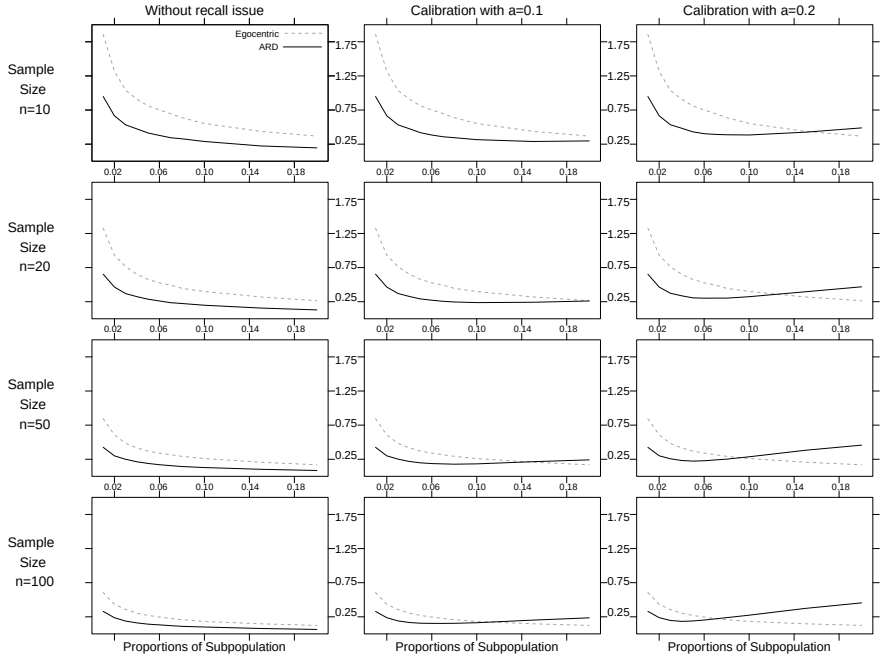

Fig. 2 A comparison of the performance of two different sampling methods. The gray dashed line is the RRMSE of estimator using egocentric sampling method with egocentric size 3. The black solid line represents for the RRMSE of estimator based on ARD method.

nominations. Taking recall error into consideration, egocentric method seems to be better when subpopulation size is large while ARD method performs better when subpopulation size is small. Additionally, based on effective sam-

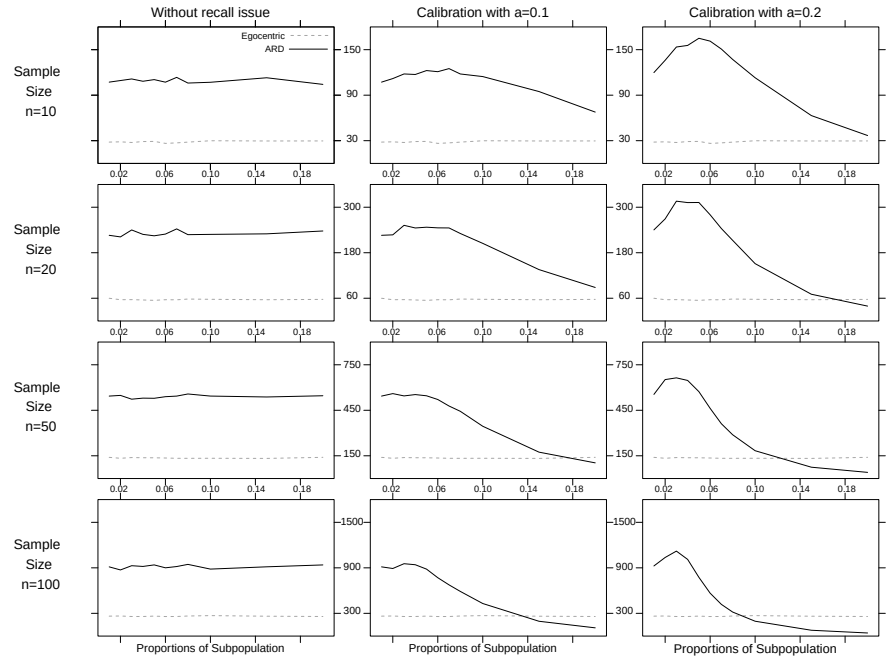

Fig. 3 A comparison of the performance of two different sampling methods. The gray dashed line is the n_{eff} of estimator using egocentric sampling method with egocentric size 3. The black solid line represents for the n_{eff} of estimator based on ARD method.

ple size criteria, the line for egocentric method is flat, which implies that the performance of efficiency of it does not depend on the subpopulation size.

3.1.2 Homophily

The second set of simulations incorporate calibration bias, but also add additional aspects of network structure. In the simulations in the previous section, the propensity to form ties is simply the inverse of the degree of the ego. To simulate homophily, we simulated two groups in the population and inflate the propensity for tie formation for within-group interactions. Figure 3.1.2 repeats the simulation experiment from the previous section with homophily being part of tie formation. We ran the simulations for a variety of levels of homophily and present the two most extreme (with the lesser extreme being the no-homophily case presented in the previous section) as an illustration.

Figure 3.1.2 displays the results for the simulations for a network with high homphily. Comparing Figure 3.1.2 to Figure 3.1.1, we see a similar overall pattern in the performance. The distinction, however, is that in Figure 3.1.2 both designs perform worse than in Figure 3.1.1. Homophily, especially to such a degree, causes clustering in the network. This clustering increases the discrepancy between samples that include or do not include members of a certain

cluster, thus increasing the variance of the estimates. Since this increase in variance comes from the structure of the underlying network, not as a manifestation of the sampling mechanism we see an effect in all of the simulations regardless of the presence of calibration bias.

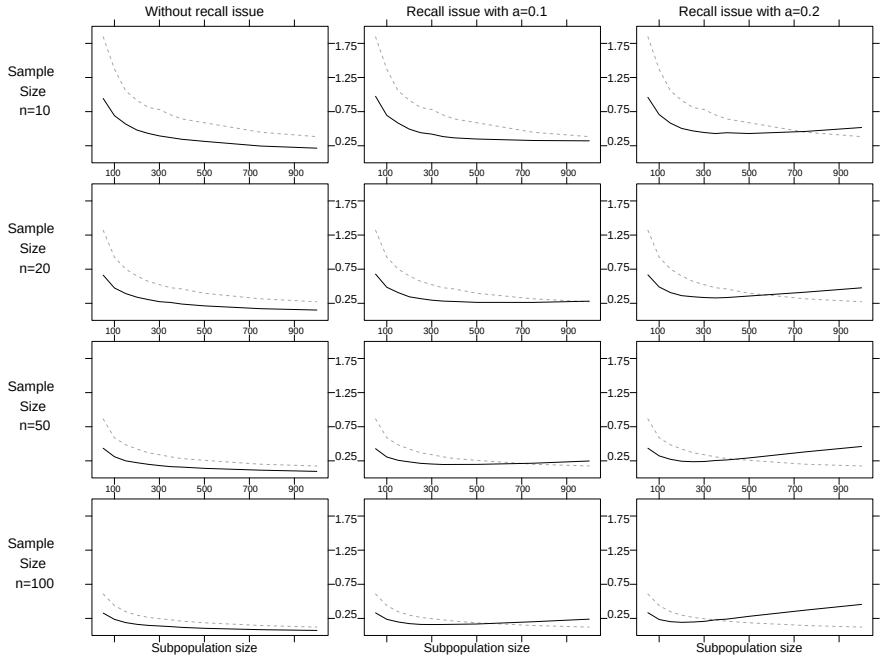

Fig. 4 A comparison of the performance of two different sampling methods in simulations with high homophily. The gray dashed line is the RRMSE of estimator using egocentric sampling method with egocentric size 3. The black solid line represents for the RRMSE of estimator based on ARD method.

3.2 Slashdot Zoo

In the previous section, we used carefully controlled simulations to evaluate particular aspects of the performance of estimators based on these two designs. Regardless of the sophistication of a simulation, however, we cannot replicate the complexity of an actual network. In this section we begin with a large network and simulate our two sampling designs using the edges and structure already present in the network. Our goal is to explore the variability of the two designs' estimators when there is an association between the structure of the network and membership in the group of interest.

We use a snapshot of the website Slashdot, which is a technology news website which features user and editor-evaluated technology news. Slashdot

recently introduced a feature known as Slashdot Zoo, which is a social networking component to the site that allows users to list one another as friends. These data are from a snapshot of Slashdot Zoo obtained in February 2009 (Leskovek et al, 2009). The network contains approximately 82, 000 nodes and about 948, 000 edges.

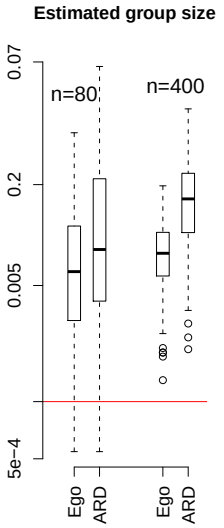

Fig. 5 Boxplots showing performance using data from Slashdot Zoo. Each box represents 300 samples from the Slashdot Zoo data of size either 80 (about .1% of the population) or 400 (about .5% of thepopulation). For each sample, we computed the size of the hard-to-reach population using both egocentric and ARD designs. The red line represents the hard-to-reach group size (.1% of the population).

As previously mentioned, many hard-to-reach groups are structurally removed from the majority of society. This network property leads to highly clustered communities made up almost entirely of members of the hard-to-reach group and very few individuals who do not belong to the group. To simulate this association on the Slashdot network, we first need to find the individuals who have highly interconnected personal networks. We accomplish this by, for every member of the network, first selecting the subgraph that consists of the ego's personal network. We then consider the edges between members of the ego's personal network (excluding the ego) and use these edges to compute the average within-personal network degree of the nodes connected to the ego. We selected the nodes with the top .1% average within personal network degree and assign these nodes to be members of the hard-to-reach group. The simulation then begins as in the previous section by selecting a simple random sample of respondents and computing ratio estimators using

both egocentric nominations and ARD. Figure 3.2 displays the results of the

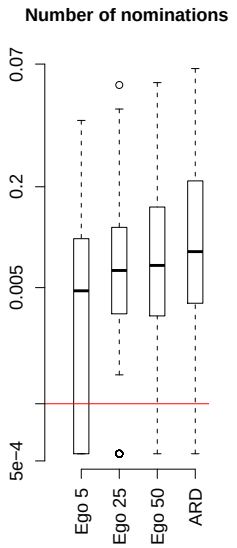

Fig. 6 Boxplots showing performance using data from Slashdot Zoo. Each box represents results using 300 simulated samples of size 80 from the Slashdot Zoo network. WIth each sample, we computed ARD estimates and egocentric nominations using either 5, 25, or 50 possible alters. The decrease in variability as the nomination size increases from 5 to 25 and then increase as it becomes closer to ARD indicates the interaction between increasing (hypothetical) sample size and network structure. The red line represents the hard-to-reach group size (.1% of the population).

simulation using two different sample sizes. Each box represents 300 samples from the Slashdot Zoo data of size either 80 (about .1% of the population) or 400 (about .5% of thepopulation). The overall variability in the estimates decreases with increasing sample size, as expected. The variability appears slightly higher in ARD for smaller samples and is about equal for both designs in larger samples. The bias (as measured informally by the median in the boxplot) appears to increase with larger samples. The variance in these estimates is largely driven by the very high variance in the number of individuals known in the hard-to-reach group. Respondents who know members of the hard-to-reach group likely know many, with it not being uncommon for some respondents to have as much as three-fourths of their personal network made up of members of the hard-to-reach group. In both case, the ratio estimators only use the aggregate number reported across all of the sampled respondents and cannot account for this variation. Thus, increasing the sample size does reduce variation by including more hypothetical alters, but also increases the likelihood of mentally recruiting a network member who is extremely highly connected to the group of interest, causing bias.

Figure 3.2 cements this point by narrowing the distance between egocentric and ARD designs. Each bar in this figure represents a different number of possible nominations for the egocentric design. Again, recall that these simulations do not account for preferential nomination bias. Moving from five to twent-five nominations reduces the variance (though the larger number of estimates which are 0 drives much of the variability) by increasing the number of mentally recruited individuals. As the number of nominations increases towards the individuals' personal network sizes (ARD), then the bias again increases as the chances of recruiting highly connected individuals increases.

3.3 The 2004 and 2006 General Social Surveys

The General Social Survey is one of the flagship surveys at the National Opinion Research Center (NORC). It started in 1972 and completed its 26th round in 2006 and, with the exception of the U.S. Census, is the most frequently analyzed source of information in the social sciences. The GSS does not use a network-based sampling design. Both egocentric nomination and ARD questions have appeared on the GSS. Along with special topics modules, the GSS contains a standard set of demographic and attitudinal questions.

The 2004 GSS asks egocentric nomination questions while the 2006 GSS asks ARD. Along with asking respondents how many people they know, the 2006 data also ask respondents how many they trust. The definition of trust is operationalized (DiPrete et al, 2011) and includes as one of its components discussing important matters. The definition of trust used in the 2006 GSS is, therefore, more similar to the prompt used to collect egocentric nominations in the 2004 data than the definition of know. Since these questions were asked only two years apart they provide an opportunity to compare the performance of the two methods described here using data which have actually been used by other authors to estimate population sizes. We compare the performance of the two methods for estimating the racial breakdown of the US population. The 2006 GSS asks respondents how many people they trust in four categories (White, African American, Hispanic, and Asian) while the 2004 GSS asks the race of each hypothetical alter (White, African American, Hispanic, Asian, or Other). The racial breakdown of the population is a convenient example since the sizes of the populations vary considerably but are reliably and widely available from other sources. Questions asking about race also have smaller transmission errors than other groups typically reached through mental link tracing designs. These data were not designed to estimate the racial breakdown of the US population, though using them for this purpose provides insights into the benefits and shortcomings of this type of data. For the ARD estimators, we use the total number of individuals reported in all of the categories as the degree.

Figure 3.3 compares the results from the two estimators to commonly accepted demographic estimates. The egocentric nominations perform well overall with most estimates being within ten percent of the widely accepted esti-

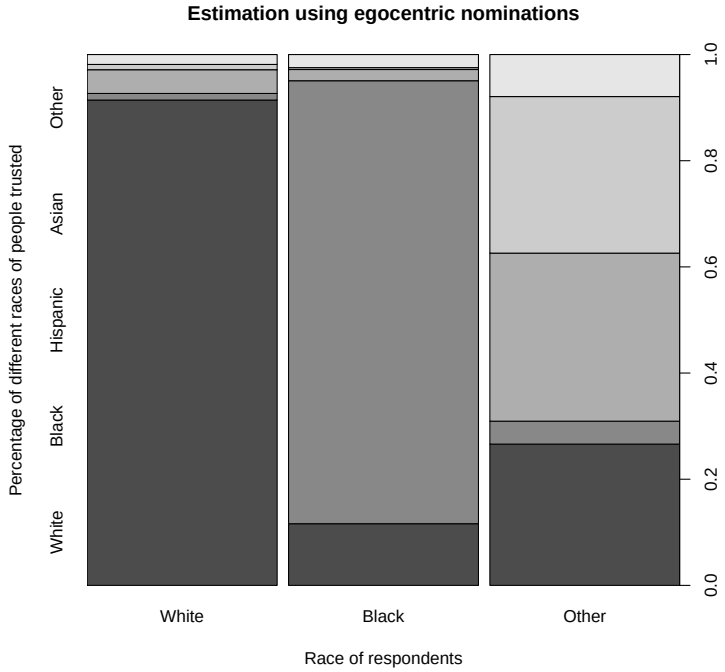

Fig. 7 A comparison of the performance of two different sampling methods using data from the 2006 and 2004 General Social Survey.

mates. The estimates using ARD are less compelling in this example. The poor performance of ARD leads to insights regarding the data collection mechanism. These data were collected using intervals (a respondent reports knowing 6-10 African Americans, for example). Using intervals is one way to alleviate the burden on respondents of asking about larger categories. In this case, however, using intervals obfuscates the distinction between the number known by a respondent among different races. The largest interval was "11 or more" and the other large intervals chosen spanned three to five known. It is not possible, therefore, to distinguish between a respondent who trusts 6 African Americans and 10 Caucasians one who trusts 10 African Americans and 6 Caucasians. Since even the smallest racial groups asked were still large, there were many responses in the larger intervals. The need to estimate respondent degree only adds to the issues caused by the intervals. A respondent who answers "6-10" on each of the responses, for example, may have a total degree of 24 or 40. Figures 3.3 and 3.3 presents the results using the ratio estimates described in the previous sections. In these two figures the results are separated by the race of the respondent and the race of the population of interest, demonstrating the impact of this obfuscation. Each race of respondent in Figure 3.3 overestimates that number of individuals in the population in their racial group.

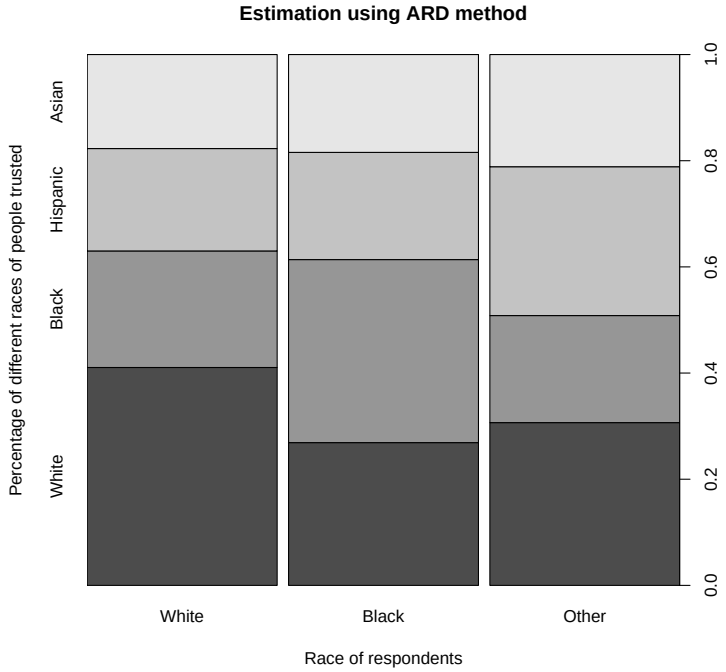

Fig. 8 A comparison of the performance of two different sampling methods using data from the 2006 and 2004 General Social Survey.

This occurs as a result of preferential attachment within a race but is balanced as the final estimate is the aggregate total for the population. In Figure 3.3, however, the within-race overestimation is significantly reduced. Recall issues undoubtedly also exacerbate the issues created by the intervals and again impact both the number reported in the group of interest and the estimated degree.

The exception to the reasonable performance of the egocentric estimator is the estimated number of Asian individuals in the population. The egocentric estimator gives the percent of Asians in the population at about 2%, while the accepted population estimate is 4%. There are various potential reasons for this underestimate which cannot be distinguished using this data. One possible explanation involves an underlying assumption of both methods about the networks of the hypothetical respondents. If, for example, the networks of Asian Americans were overall smaller than the networks of other Americans, we would expect both the egocentric and ARD estimators to underestimate the proportion of individuals in that group in the population.

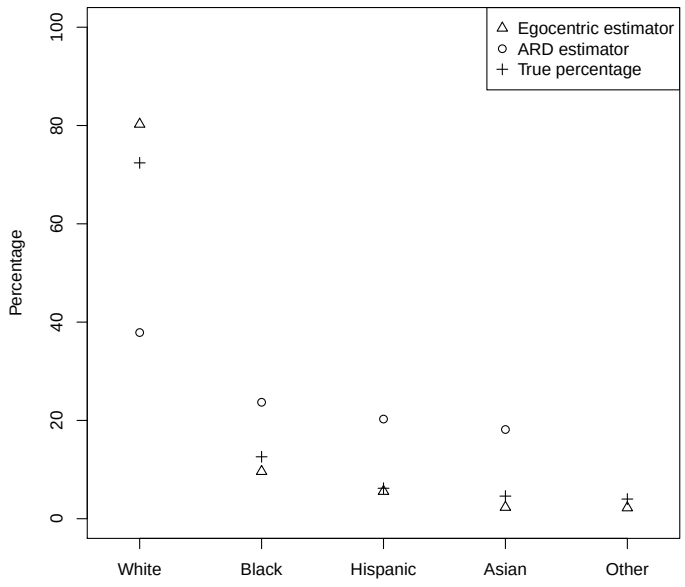

Fig. 9 A comparison of the performance of two different sampling methods using data from the 2006 and 2004 General Social Survey.

4 Discussion and survey design recommendations

We compare two commonly used network-based methods for estimating the sizes of populations which are difficult to reach using standard surveys. Using the simulated and actual data experiments from the previous sections, we now describe general guidelines which we hope will aid practitioners in deciding how to incorporate mental link tracing designs to reach traditionally hard-to-reach groups. First, as previously noted, both methods are susceptible to bias due to particular types of network structure. We have discussed explicitly the impact of assortative mixing. When certain demographic information is available about some groups, McCormick et al (2010) propose statistical models to address the impact of assortative mixing in degree estimation and could be used as a basis for future work with population size estimation.

Table 1 presents a summary of the advantages and limitations of these methods. A major factor in deciding which of the methods to use is the type of information researchers desire. Researchers interested in detailed information about each of the hypothetical alters is better suited using egocentric nominations, for example, since this method gives the opportunity to ask multiple questions about the same hypothetical alter. ARD in contrast, does not

examine any of a respondent's links directly and thus would not provide the desired information in this context. Similarly, as noted in Table 1 if direct contact with alters is necessary (for blood or DNA samples, for example), neither method would be suitable. Combining this knowledge with insights gleaned

Table 1 Comparison of ARD and egocentric nominations.

Method	Advantage	Limitation
ARD	-"Reaches" more respondents. -Easily ask about multiple groups.	-Need to estimate degree. -Susceptible to recall issues.
Both	-Savings of time and resources. -Reach hard-to-reach groups through network.	-Non-random mixing. -No direct contact with group members.
Egocentric	-Detailed information about alters. -Less susceptible to recall issues.	-Preferential recall bias.

from simulation, we also suggest that researchers consider the approximate size of the group of interest when selecting a survey mechanism. ARD would be most beneficial in cases where the group of interest is small. As noted in Table 1, ARD reach the most hypothetical respondents. If a survey asks 1, 500 respondents ARD questions and each respondent has an average degree of 750, for example, then the survey (indirectly) reaches $1,500 \times 750 \approx 1.13$million respondents. ARD is known to be influenced by recall issues, however, so the benefits to reaching additional population members decreases as the bias introduced by recall issues increases. For larger groups, therefore, egocentric nominations display better performance.

Researchers may also alter the way that egocentric or ARD questions are posed to respondents to mitigate the influence of the limitations of each of the methods. For ARD, for example, recall issues are potentially quite influential for larger populations. As noted in Section 3.3, one approach to dealing with recall issues is to give respondents intervals for responses. An alternative approach is to partition the group of interest into multiple, smaller groups (into age categories, for example). This strategy scaffolds respondents to think systematically about the individuals they know in the group of interest and lessens the chances of recall errors. This strategy does, however, take additional time.

For egocentric nominations, preferential nomination could cause bias if the subset of individuals the respondent nominates are more or less likely to belong to the group of interest. To mitigate this bias, researchers could consider a strategy which introduces randomness into the nominations process. Enumerators could ask respondents to, for example, nominate only the hypothetical alters who have a birthday in a particular month or have a name which begins with a certain letter. The choice of how to perform the randomization would, of course, require care and be context specific.

To this point, we have considered preferential nomination bias and calibration bias as though they are caused by two different underlying phenomena. The two processes may be related, however. Specifically, both preferential nom-

ination and calibration bias can be thought of as manifestations of the way that respondents go about dredging their personal networks. One may imagine a respondents' propensity to nominate/count an alter as some latent process (which will undoubtedly change as the ego's relationship with the alter changes and based on chance through things such as time since last interaction) by which respondents dredge their network. Under both designs, there is a truncation mechanism that excludes personal network members with lower values of the process. In the case of ARD, this truncation varies based on respondents' ability and/or desire to dredge farther into his or her network. For egocentric nominations, however, the truncation occurs at the same number of respondents for each ego. This does not, however, mean that the level of bias is necessarily equal across all of the egos since the variability between latent processes across individuals is likely quite large.

Both methods also likely suffer from an additional source of bias through transmission errors. Transmission errors occur when the respondent knows someone in a specific subpopulation but is not aware that the person is actually in that subpopulation. This type of error is particularly salient with disease transmission (a person might know someone who is diabetic, for example, but may not know the alter's medical status). These transmission errors likely vary from group to group depending on the sensitivity and visibility of the information. These errors are extremely difficult to quantify, because very little is known about how much information respondents have about the people they know (Laumann 1969; Killworth et al. 2006; Shelley et al. 2006). Recent work by Salganik et al (2011) provides some insights which could be used for future statistical modeling.

The work in this paper also raises questions about the nature of realistic simulations for network sampling. In our simulations, we chose to allow the underlying network to remain fixed and introduce variability through the selection of nodes. An alternative approach would be to simulate a new network at each realization. This approach would then speak to the performance of the estimators under variations of both network topology and sampling.

An additional point of discussion is the context-specific nature of any judgement about the quality of an estimator. If the goal is estimating the number of individuals living with HIV in a particular region to determine the number of medications to order, for example, estimates which produce errors on the order of tens of thousands of potential patients may be undesirable. This may be the case even if the error rate for the proposed estimator is significantly lower than those used previously.

The methods presented with the Slashdot Zoo data also provide insights. Our novel method for assigning hard-to-reach group membership ensures an association between membership in the hard-to-reach group and network structure. The method also makes an association between the network effects and the size of the hard-to-reach group. Large hard-to-reach groups must have smaller average within personal network degree. The Slashdot data in general also has a very high number of hubs which, not surprisingly, also tend to have very high mean within personal network degree. Though these hubs

work to simulate the clustering that is likely seen in many hard-to-reach populations, they also have a very large number of incoming ties, which is likely not the case in a hard-to-reach population. This feature could be specific to an online network where the hubs with many incoming ties serve a different function than they would in a network in the physical world. In Slashdot, these nodes could represent, for example, major news outlets. Many individuals would likely wish to have a tie to the news source, though they may not have much other direct communication. An above average degree distribution of individuals in the hard-to-reach groups is precisely the phenomenon which we suggested (hypothetically) to explain the under-estimation of Asians in the GSS data. The increasing bias that comes with higher sample size in these simulations studies should not be taken as evidence for smaller samples. Rather, this finding demonstrates the repercussions of network structure when increasing the hypothetical sample size and motivates further statistical work in methods for data from standard surveys which accommodate (or even exploit) more complicated forms of network structure.

5 Acknowledgements

The authors gratefully acknowledge the support of the SAMSI Complex Networks Program.

References

Blitzstein J, Diaconis P (2006) A sequential importance sampling algorithm for generating random graphs with prescribed degrees. preprint pp 1–35

DiPrete TA, Gelman A, McCormick T, Teitler J, Zheng T (2011) Segregation in Social Networks Based on Acquaintanceship and Trust. The American Journal of Sociology 116:1234–1283

Goel S, Salganik M (2009) Respondent-driven sampling as Markov chain Monte Carlo. Statistics in Medicine 28(17):2202–2229

Handcock MS, Gile KJ (2010) Modeling social networks from sampled data. The Annals of Applied Statistics 4(1):5–25

Killworth PD, McCarty C, Bernard HR, Johnsen EC, Domini J, Shelley GA (2003) Two Interpretations of Reports of Knowledge of Subpopulation Sizes. Social Networks 25:141–160

Leskovek J, lang kj, Dasgupta A, Mahoney MW (2009) Community Structure in Large Networks: Natural Cluster Sizes and the Absence of Large Well-Defined Clusters. Internet Mathematics 6:29–123

Lohr S (1999) Sampling: Design and Analysis. Duxbury Press

McCormick T, Salganik MJ, Zheng T (2010) How many people do you know?: Efficiently estimating personal network size. Journal of the American Statistical Association 105:59–70

McCormick TH, Zheng T (2007) Adjusting for Recall Bias in "How Many X's Do You Know?" Surveys. In: Proceedings of the Joint Statistical Meetings

McPherson M, Smith-Lovin L, Cook JM (2001) Birds of a feather: homophily in social networks. Annual Review of Sociology 27:415–444

Salganik MJ, Heckathorn DD (2004) Sampling and estimation in hidden populations using respondent-driven sampling. Sociological Methodology 34:193–239

Salganik MJ, Mello MB, Adbo AH, Bertoni N, Fatzio D, Bastos FI (2011) The game of contacts: Estimating the social visibility of groups. Social Networks 33:70–78

Shelley GE, Killworth PD, Bernard HR, McCarty C, Johnsen EC, Rice RE (2006) Who knows your HIV status II?: Information propogation within social networks of seropositive people. Human Organization 65(4):430–444

Sirken, MG (1970) Household Surveys with Multiplicity. Journal of the American Statistical Association 65(329):257-266

Zheng T, Salganik MJ, Gelman A (2006) How many people do you know in prison?: Using overdispersion in count data to estimate social structure. Journal of the American Statistical Association 101:409–423