The Sensitivity of Respondent-driven Sampling Method

Xin Lu[a,b,*], Linus Bengtsson[b], Tom Britton[c], Martin Camitz[d], Beom Jun Kim[e], Anna Thorson[b], Fredrik Liljeros[a]

[a] *Department of Sociology, Stockholm University, Stockholm, Sweden*
[b] *Department of Public Health Sciences, Karolinska Institute, Stockholm, Sweden*
[c] *Department of Mathematics, Stockholm University, Stockholm, Sweden*
[d] *Department of Medical Epidemiology and Biostatistics, Karolinska Institute, Stockholm, Sweden*
[e] *Department of Physics, Sungkyunkwan University, Suwon, Korea.*

Abstract

Researchers in many scientific fields make inferences from individuals to larger groups. For many groups however, there is no list of members from which to take a random sample. Respondent-driven sampling (RDS) is a relatively new sampling methodology that circumvents this difficulty by using the social networks of the groups under study. The RDS method has been shown to provide unbiased estimates of population proportions given certain conditions. The method is now widely used in the study of HIV-related high-risk populations globally. In this paper, we test the RDS methodology by simulating RDS studies on the social networks of a large LGBT web community. The robustness of the RDS method is tested by violating, one by one, the conditions under which the method provides unbiased estimates. Results reveal that the risk of bias is large if networks are directed, or respondents choose to invite persons based on characteristics that are correlated with the study outcomes. If these two problems are absent, the RDS method shows strong resistance to low response rates and certain errors in the participants' reporting of their network sizes. Other issues that might affect the RDS estimates, such as the method for choosing initial participants, the maximum number of recruitments per participant, sampling with or without replacement and variations in network structures, are also simulated and discussed.

Keywords: directed network, hidden population, network, respondent-driven sampling, RDS, sensitivity

1. Introduction

Hidden or hard-to-reach populations, such as injecting drug users, men who have sex with men and sex workers, are generally difficult to access because of their strong privacy concerns and a lack of a well-defined sampling frame from which a random sample can be drawn (Heckathorn, 1997). Sampling frames are also lacking for many groups without strong privacy concerns, such as for example jazz musicians (Heckathorn and Jeffri, 2001). Methods for obtaining information about such groups have involved contacting especially knowledgeable persons within the group (key informant sampling) (Deaux and Callaghan, 1985), drawing a sample of participants from locations where group members are known to pass (targeted sampling) (Watters and Biernacki, 1989), or asking members of a group to give the researchers contact details of others in the same group (snowball sampling) (Erickson, 1979). However, these methods all introduce a considerable selection bias, which impairs generalization of findings from the sample to the studied population (Heckathorn, 1997; Magnani et al., 2005).

Respondent-driven sampling (RDS) is a method developed to overcome the challenges of selection bias when sampling hidden populations (Heckathorn, 1997, 2002; Salganik and Heckathorn, 2004; Salganik, 2006; Volz and Heckathorn, 2008). An RDS study starts out by purposively selecting some participants who are members of the study population (usually 5 to 15). These persons are called "seeds". The seeds are given a number of recruitment "coupons" (usually 3) to distribute to friends and acquaintances within the study population. If those friends who

[*] Address for correspondence: Xin Lu, Department of Sociology, Stockholm University, SE-106 91, Stockholm, Sweden.
Email address: lu.xin@sociology.su.se (Xin Lu)

receive a coupon decide to participate, they are in turn given the same number of coupons to invite further participants. Participants are rewarded for their personal participation in the study as well as for each peer they invite and who also participates. The invitation coupon contains a serial number that enables the researchers to follow the recruitment chains in the sample. If the recruitment chains are sufficiently long, the sample composition stabilizes and becomes independent of the characteristics of the seeds. What' s more, each participant is asked for the number of persons he or she knows within the study population, known as his/her "personal network size" or "degree". The degree of a participant is important to collect as participants with large degrees are oversampled and participants with small degrees are undersampled. Knowing the degree of each participant thus allows adjustment for this bias.

When the sample has been collected, the proportion of persons with the characteristic A in the population can be estimated by the *RDSII* estimator (Volz and Heckathorn, 2008):

$$\hat{P}_A = \sum_{i \in A \cap S} d_i^{-1} \Big/ \sum_{i \in S} d_i^{-1} \tag{1}$$

Where d_i is the degree of individual i, and S the set of sampled individuals.

Volz and Heckathorn (2008) proved that the *RDSII* estimator provides asymptotically unbiased estimates if the following assumptions are fulfilled:

(i) Reciprocity: individuals in the studied population maintain and recruit peers through reciprocal relationships, that is, the network within which recruitment happens, is undirected;

(ii) Connectedness: each individual in the studied population has a chance to be invited to participate, that is, the network forms a single component;

(iii) Sampling is with replacement: individuals are allowed to be recruited into the sample more than once;

(iv) Degree: respondents can accurately report their degree in the network;

(v) Random recruitment: peer-recruitment is a random selection from the respondents' personal network;

(vi) Each respondent recruits a single peer, that is, the number of recruitment coupons is one.

The ability of making generalizable estimates together with a feasible field implementation have contributed to a rapid increase in RDS studies conducted globally in recent years: to date, well over 100 studies in over 30 countries have been performed (Johnston et al., 2008; Malekinejad et al., 2008; Wejnert and Heckathorn, 2008).

However, the assumptions underlying the RDS estimator are not easily met in real life. First, most social networks contain directed edges, or edges which do not have the same strength in each direction. Second, to prevent participants from colluding to recruit each other back and forth to gain rewards, empirical RDS applications sample without replacement, meaning that respondents can only participate once. Third, it is difficult for respondents to report their degree accurately. Fourth, participants usually pass their coupons to peers with whom they have a close rather than a more distant relationship, which is not a random selection. Fifth, to avoid recruitment chains stopping too early, researchers most often use three coupons rather than one.

While rationality of those assumptions have been questioned and argued in the literature (Heckathorn, 1997, 2002; Salganik and Heckathorn, 2004; Salganik, 2006; Volz and Heckathorn, 2008; Heimer, 2005; Goel and Salganik, 2009), it is difficult to assess the reliability of RDS since the study population is usually unknown. Previous studies have mainly used artificially constructed networks and have often been linked to the introduction of new estimators, hence they have focused on different aspects and did not cover the whole scope of possible violations. The most recent and comprehensive study was made by Gile and Handcock (2009). Using artificially networks, which were constructed from pilot data from the CDC surveillance program (Abdul-Quader et al., 2006), they simulated RDS with respect to the bias induced by the violations of assumptions iii, v and vi. The population size they used was quite small (1000) and the fractions of sample sizes are relatively large, from 50% to 95% (500 up to 950). They addressed the possibility of reduction of bias by discarding early waves and found a potential bias caused by preferential selection of peers and sampling without replacement. However, the number of seeds, waves and coupons were fixed, and they didn't discuss other assumptions that might affect the RDS estimates, such as directedness of networks, recruitment failure, and degree reporting error amongst others.

Based on the increasing use of RDS within research, together with plausible real life difficulties in completely fulfilling the theoretical assumptions, we identified a need for systematically testing the robustness of the RDS method when sampling diverges from the basic assumptions in the analytical proof. In this study, we simulate RDS studies within a real-life social network, which is constructed by data extracted from a lesbian, gay, bisexual, transgender (LGBT) web community. By violating the RDS assumptions one by one, we seek to analyze the resistance of the estimator to bias and to evaluate how real life deviations from theoretical requirements will affect the estimates. We use the *RDSII* estimator for all RDS estimates in this article as this estimator has improved analytical powers compared to earlier RDS estimators and provides equivalent estimates when data-smoothing is used (Heckathorn, 2007; Volz and Heckathorn, 2008; Gile and Handcock, 2009).

Four measurements are used throughout this paper: the average estimate (AE), defined by the mean of the *RDSII* estimates, $AE_j = \sum_{i=1}^{m} est_{ij}/m$, where est_{ij} is the estimate of *RDSII* at the i^{th} simulation when sample size is j; the bias, defined by the absolute difference between AE and the true population, $Bias_j = |AE_j - P^*|$; the standard deviation (SD) of estimates for a given sample size j, SD_j; and finally the mean absolute error (MAE) of estimates for a given sample size j, $MAE_j = \sum_{i=1}^{m} |est_{ij} - P^*|/m$.

The rest of this paper is organized as follows: in section 2, we give a brief description of our data and networks; in section 3, we describe the results of simulating RDS studies in the networks when all of the assumptions are satisfied; in section 4, we test the effects of violating the assumptions one by one; and in section 5, we summarize and draw our conclusions.

2. The MSM Network

Data collection. "Qruiser"(www.qx.se), is the Nordic region's largest and most active web community for homosexual, bisexual, transgender, and queer persons. Contacts between members on the web site are maintained mainly by a "favorite list", on which each member can add any other member without approval from that member. Members can attend clubs (web pages about specific topics) and send messages to each other (Rybski et al., 2009).

We collected information on personal profiles as well as on all messages sent within the web-community from Dec 15, 2005, to Jan 18, 2006. During the 63 days of this data collection period, 12,590,911 messages were recorded and there were 184,819 distinct members registered on the web site.

Network Formation. Based on the membership profiles, we extracted a network that contained only members characterizing themselves as homosexual males. We define an outgoing edge to be formed if a member has another member on his favorite list. An edge is called reciprocal if a connected pair of members have both an ingoing and an outgoing edge between each other. If a pair does not have both an ingoing and an outgoing edge between each other, it is called irreciprocal. To avoid the inclusion of inactive persons, members were required to have sent at least one message to any other person on the site during the data collection period.

For our research purpose, only members of the Giant Connected Component (GCC), defined by the largest component connected with only reciprocal edges, were kept as nodes (16,082 nodes). If we keep only the reciprocal edges in that GCC, we obtain an undirected network ($G1$) with an average degree of 6.74. If we keep both reciprocal and irreciprocal edges, we obtain a directed network ($G2$) with an average degree of 17.2. Note that the definition of the GCC ensures that all nodes have a chance to be recruited with RDS sampling in both $G1$ and $G2$. Degree distributions for both $G1$ and $G2$ are plotted in Figure 1. The distributions are very skewed, e.g., half of the members in $G2$ have no more than 10 outgoing edges, while a small proportion of members have a large number of outgoing edges.

Homophily. An important issue for chain referral sampling is the homophily of edge formation, which is the probability for participants of connecting with friends that are similar to themselves rather than connecting randomly (Rapoport, 1980; Morris and Kretzschmar, 1995; McPherson et al., 2001; Heckathorn, 2002). The homophily of different groups in our network are shown in Table 1. The homophily with respect to *age* and *county* are fairly large, indicating a fair part of the edges being formed between members of the same *age* or between members living in the same *county*. Taking *county* within the undirected network, $G1$ as an example, members who live in Stockholm formed edges with members who also lived in Stockholm 50% of the time, while they formed edges randomly among all cities (including Stockholm) the remaining 50% of the time. The homophily for persons living Stockholm is thus 0.5. Homophilies for *civil status* and *profession* were very small, indicating that edges were formed as if members, regarding *civil status* and *profession*, chose randomly among other members.

Network Variation. To avoid misleading conclusions resulting from the effects of network structure and edge density in our simulations on the undirected network ($G1$), we created two variants of $G1$: the first type of networks ($G1_{add}$) was obtained by randomly adding reciprocal edges with properties proportional to $G1$ until the average degree

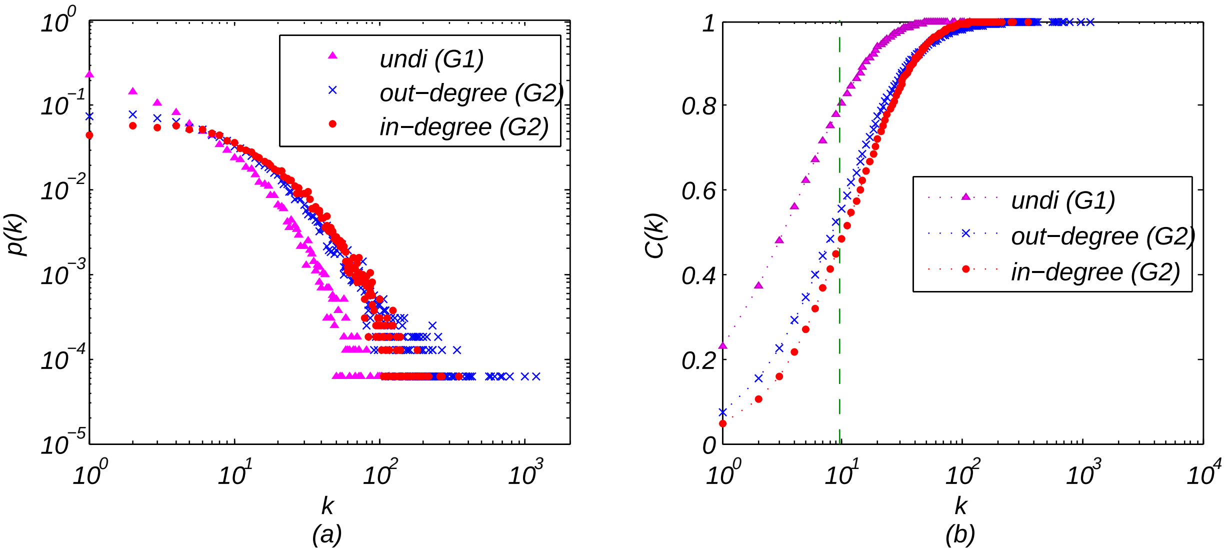

Figure 1: (a) Degree distribution (b) Cumulative degree distribution

Table 1: Proportions (P^*) and homophilies (H) of groups in the undirected ($G1$) and directed ($G2$) networks

	Age		**County**		**Civil Status**		**Profession**	
	Before 1980	others	Stockholm	others	Single	others	Employed	others
P^*	77.77%	22.23%	38.79%	61.21%	40.39%	59.61%	38.19%	61.81%
$HG1$	0.40	0.37	0.50	0.40	0.05	0.08	0.13	-0.05
$HG2$	0.23	0.34	0.50	0.28	0.03	0.07	0.06	0.02

was increased by 20, for each property, separately. The second type of networks ($G1_{rand}$) was obtained by randomly rewiring each pair of reciprocal edges to another node with the same property as the former one. After the procedures above, we obtained four denser networks and four randomized networks, all with the homophily unchanged for each property, respectively. The changes in degree distributions for both $G1_{add}$ and $G1_{rand}$ can be found in Appendix A.

Moreover, to test the effects of preferential recruitment in RDS (section 4.4), we weighted each reciprocal edge in $G1$ in two ways: by the maximum number of sent messages in any one direction, and by the minimum number of sent messages in any one direction. For example, if node A sent 10 messages to node B and received 5 back from B, the weight on edge $e_{A,B}$ ($e_{B,A}$) would be 10 for the maximum-weighted network ($G1_{max}$) and 5 for the minimum-weighted network ($G1_{min}$). In these two weighted networks, respondents were supposed to recruit peers with probability proportional to the edge weights.

We now proceed to describe the results of simulated RDS samplings under variable circumstances in the networks described above. We compare the true population proportions of the four variables in Table 1 (two with high homophily and two with low) with the RDS estimates given by the simulated samplings. All simulations are repeated 10,000 times unless otherwise stated.

3. RDS on the Undirected Network

We first ran simulations on the undirected network, $G1$, to see whether the *RDSII* estimator worked well when all the earlier stated assumptions (i-vi) were satisfied. We started each simulation with a single randomly selected seed and we restricted the number of coupons to one, that is, each recruiter could only recruit one other person (assumption vi). All respondents were selected randomly from the recruiters' personal network (assumption v), and nodes could be selected multiple times (sampling with replacement, assumption iii). Since all participants' degrees were assumed to be known by the participants themselves, and $G1$ is a single connected component with only reciprocal edges, assumptions i, ii and iv were also satisfied. We kept recruiting participants in the simulation until the sample size reached 10,000 participants. The average estimation, standard deviation, and mean absolute error are shown in Figure 2. Even though our network is sparse, the *RDSII* estimates converged to the true population proportions (P^*) very quickly. The bias in the average estimate for *age*, *county*, *civil status*, and *profession* is shown below in Figure 2. Appendix B shows clearer figures for sample sizes of less than 1000.

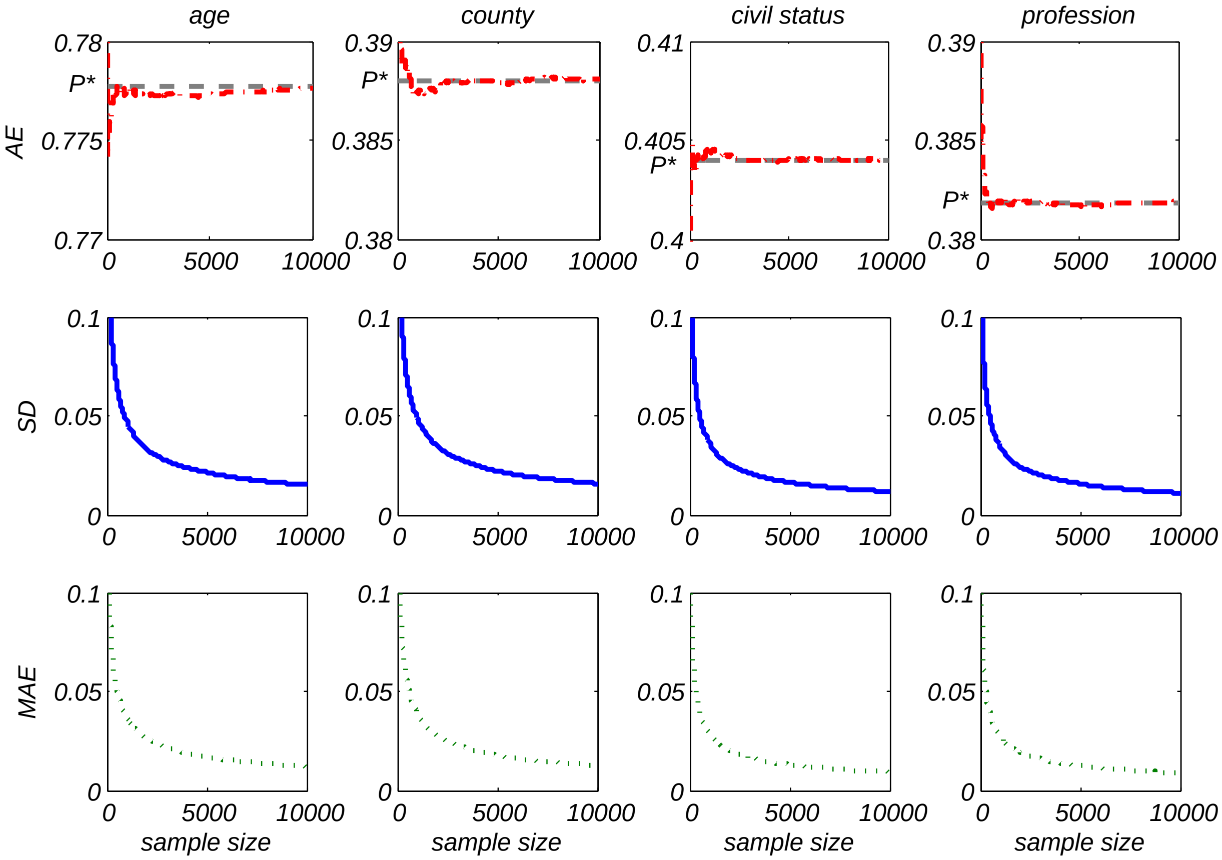

Figure 2: *RDSII* estimations on the undirected network ($G1$). The average estimates approached the true proportions very fast. When the sample size was 500 the Bias was only 0.0002, 0.0009, 0.00002, 0.0002 for *age*, *county*, *civil status* and *profession*, respectively.

The SD and MAE decreased to 2% when the sample size approached 10,000. However, it is rarely possible to recruit this many respondents in a real life RDS study. Reported sample sizes of real life RDS studies are virtually all less than 1000. For our network, we can see that the SD was around 5%, and the MAE was around 4% when the sample sizes were between 500 and 1000 participants.

We should note that our network is far from ideal: first, it is sparse compared to reported studies since the average degree is low, only 6.74; second, the degree distribution is skewed with almost 40% of network members having a degree no higher than 2; third, there is a high homophily with regard to the variables *age* and *county*. Despite these difficulties, the high precision and rapid convergence of the RDS estimates in Figure 2 reveal that the RDS estimator is asymptotically unbiased and works well on undirected networks.

4. Violations of Assumptions

4.1. RDS on directed networks

If a directed network forms a Giant Strongly Connected Component (GSCC) (Schwarte et al., 2002) in which every node can be reached by any other, and assumptions iii to iv are satisfied, we can model the RDS sampling as a Markov process, which has the equilibrium:

$$[x_1, x_2, \dots, x_N] \times \begin{bmatrix} 0 & e_{12}/d_1^o & \cdots & e_{1N}/d_1^o \\ e_{21}/d_2^o & 0 & \cdots & e_{2N}/d_2^o \\ \vdots & \vdots & \ddots & \vdots \\ e_{N1}/d_N^o & e_{N2}/d_N^o & \cdots & 0 \end{bmatrix} = [x_1, x_2, \dots, x_N] \tag{2}$$

Abbreviated as:

$$X^T \times A = X^T \tag{3}$$

Where X is the equilibrium of the Markov Process, $e_{ij} = 1$ if there is an edge from i to j, otherwise $e_{ij} = 0$, and d_i^o is the out-degree of i.

It can be easily verified that

$$x_i = d_i / \sum_{j=1}^{N} d_j \tag{4}$$

is the solution for Eq. (3) if the out-degree and in-degree are equal for all nodes. Actually, Eq. (4) is the underlying equation from which the *RDSII* estimator is derived (Volz and Heckathorn, 2008; Goel and Salganik, 2009).

However, the equilibrium can hardly be constructed for general directed networks. We can rewrite Eq. (3)as:

$$A^T \times X = X \tag{5}$$

This means that X should be an eigenvector of eigenvalue 1 for A^T. (The existence of eigenvalue 1 for A^T can be easily proved as the all-one vector is the eigenvector for A (Woess, 1994; Brin and Page, 1998; Page et al., 1999).

Let $V = [v_1, v_2, \dots, v_N]$ be the normalized eigenvector of A^T for eigenvalue 1, then an RDS sample $\{s_1, s_2, \dots, s_n\}$, can be weighted by

$$\hat{P}_A = \frac{\sum_{s_i \in A} \frac{1}{v_{s_i}}}{\sum_{s_j} \frac{1}{v_{s_j}}} \tag{6}$$

to estimate the proportion of individuals in group A in the population. We denote Eq. (6) as '*eig*' to weight the RDS samples in a directed network. Note that we can hardly know the value V from an RDS sample.

Both *RDSII* and *eig* estimations on the directed network ($G2$) are presented in Figure 3. Not surprisingly, the *RDSII* estimates were biased for all groups. For *age* and *county*, these biases were as high as 6%, while *civil status* and *profession* performed better at 0.5%, and 2.2%, respectively. However, the *eig* estimates weighted by Eq. (6) agreed well with the true population proportions.

The standard deviations (SD) were similar for all four groups (and very similar to the SD of the undirected networks). However, the MAE of *RDSII* in the directed network was much higher than that of the undirected networks for *age* and *county*, 7%-8%, indicating that if the network under study is partly directed, the use of *RDSII* estimations could result in relatively large errors. We can see that the difference for *civil status* and *profession* were very small, telling us that directedness of edges probably has little effect on *RDSII* for groups with low homophily.

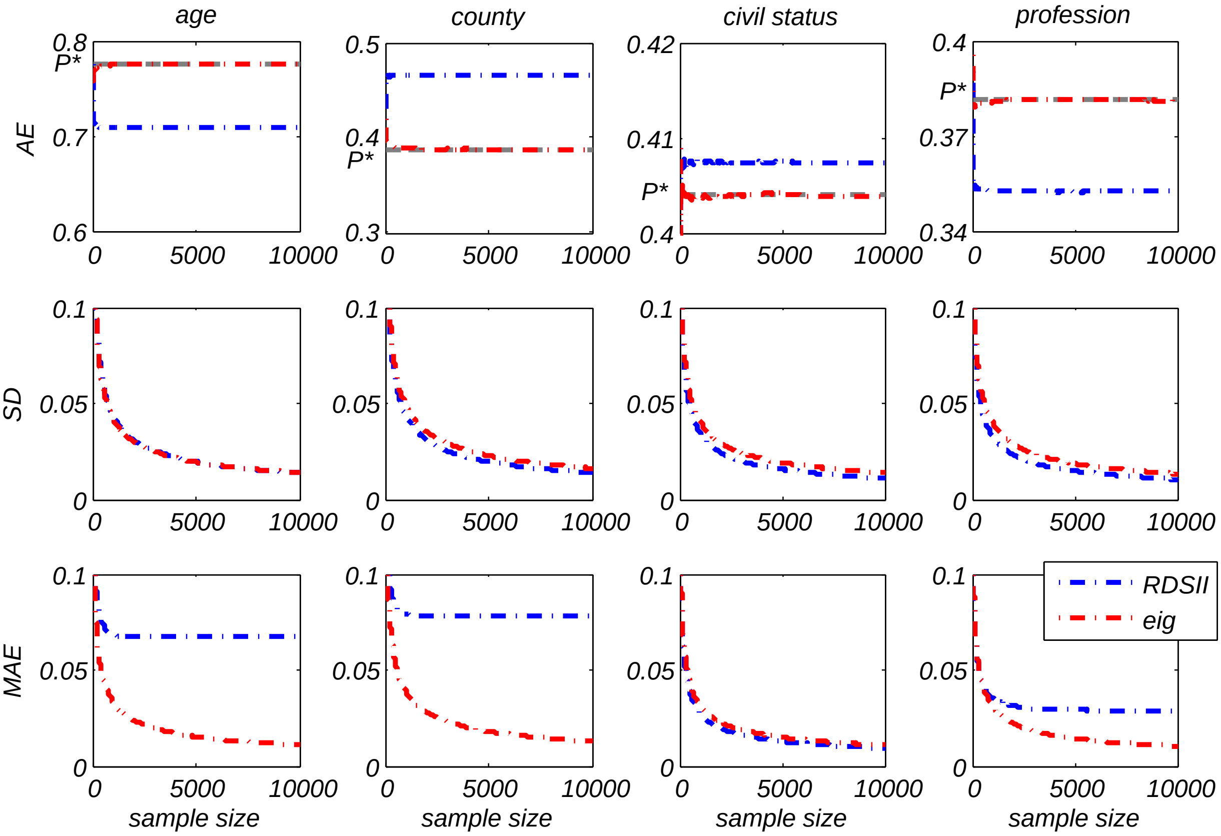

Figure 3: Estimations on the directed network ($G2$). Number of seeds=1, coupons=1, with replacement. Blue lines stand for the *RDSII* estimates and red lines for estimates weighted by eigenvectors.

4.2. Sampling without replacement

It is generally believed that sampling without replacement (SWOR) creates negligible bias compared to sampling with replacement (SWR) in RDS when the sample size is small relative to the population (Heckathorn, 1997, 2002; Volz and Heckathorn, 2008). We tested this proposition in our undirected network. What's more, to increase the generalizability of our results, we also compared the replacement effect on $G1_{add}$ and $G1_{rand}$. Results are shown in Figure 4. We can see from the figures that the *RDSII* estimations for SWR were asymptotically unbiased on all networks, while the estimations for SWOR were biased in different directions in the different networks.

The hypothesis above seemed to hold when sample sizes were small, that is, between 500 and 1000 (see enlarged figure in Appendix C. The maximum differences in the average estimates between SWR and SWOR were all within 1%. For *county*, *civil status* and *profession*, the average estimates of SWOR are even slightly closer to the true population than SWR. The SWOR always has a smaller SD and MAE than SWR and this is especially apparent in $G1$. Simulations not included in this paper indicate that networks with skewed degree distribution result in larger variances than those with a Poisson distribution and as networks gets denser, the variance becomes smaller.

4.3. Rejecting invitations and forgetting peers

While a great majority of participants in RDS studies report a social network size larger than three, far from all distributed coupons result in study participations (Johnston et al., 2008; Malekinejad et al., 2008). This could be seen as a violation of assumption iv, which states that participants can accurately report their personal network size (assumed to reflect chances of being invited or of the number or peers who have a chance to be invited by the person) and assumption vi, which states that all participants use their one coupon to make one successful recruitment. We can note that the latter assumption includes the dual assumption that each coupon generates one further participant and

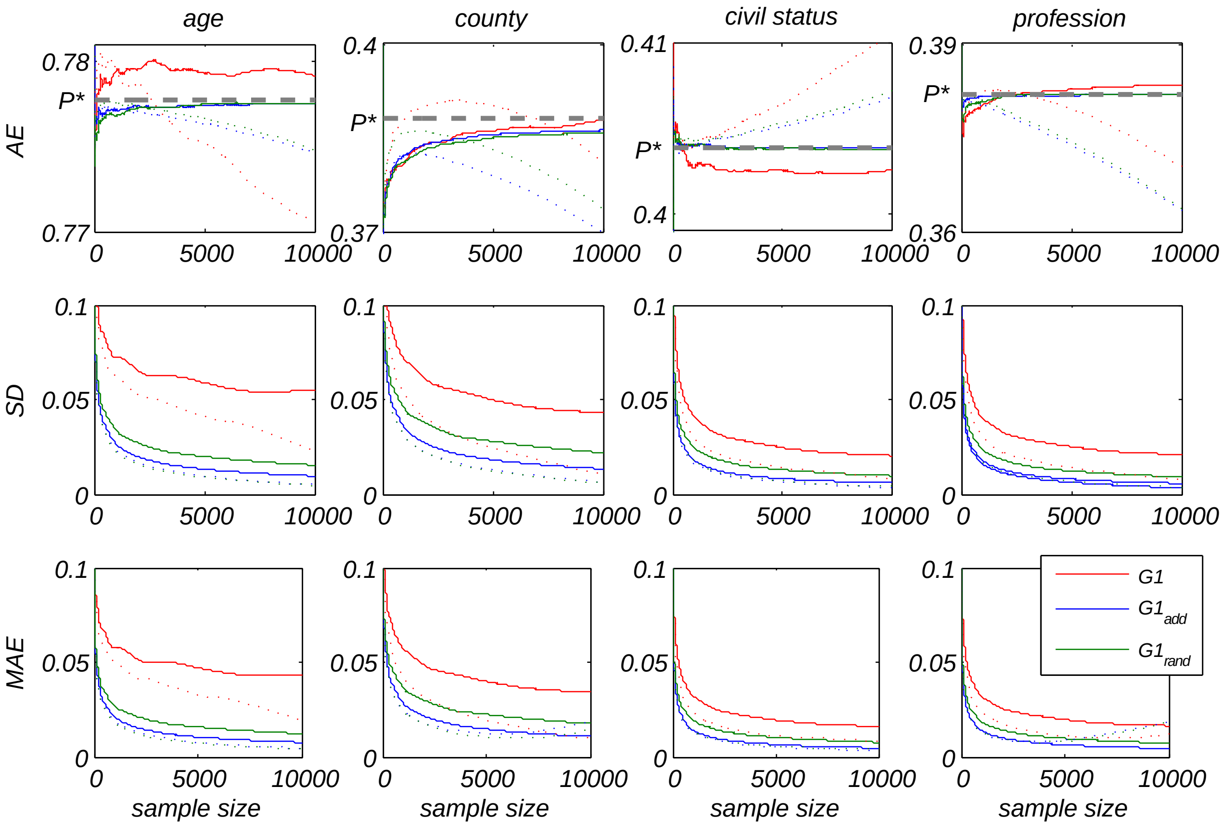

Figure 4: Effects of network structures and replacement. Number of seeds=10, coupons=3. Seeds were randomly selected at the beginning of each simulation. Solid lines represent sampling with replacement and dashed lines represent sampling without replacement.

also that each participant receives only one coupon. In our simulations however, three coupons were used, see below. We modeled the effects of deviating from these ideal assumptions by letting each invited member have a probability of rejecting invitations. For each invited member, the probability of rejecting an invitation is called p_r. A rejected coupon was discarded and not reused for recruiting a new member.

In addition recruiters could potentially have difficulties remembering all of their friends when considering whom to invite and when estimating the sizes of their personal networks. We modeled this by letting the recruiters ignore some of his/her edges when inviting members from his/her personal network. In the simulations, each edge of a recruiter was given a probability of being ignored. We called this probability p_i. An ignored edge had a zero probability of being selected and was not included in the network size of the participant when calculating the *RDSII* estimates.

Additionally, we set the number of seeds to 10 and coupons to 3 to make sure that the process could recruit a sample when p_r and p_i became large. Simulation results for the $G1$, $G1_{add}$, and $G1_{rand}$ are displayed as surfaces in Figure 5 and in Appendix D1.

When recruitment takes place with “rejecting” and “ignoring” in the original sparse network (Figure 5), and in the $G1_{rand}$ networks (Appendix D1), which are also sparse, the bias is small to moderate (up to 0.03). Simulation results on the edge-added networks reveal that in these networks, changes in p_r and p_i do not affect the bias Appendix D1. Effects on MAE are small in all networks (below 0.01).

In Figure 5 it is also interesting to observe that while increases in p_r and p_i increased the bias, the MAE actually decreased. The small differences seen when varying p_r and p_i, as well as the observed decreases in SD and MAE with increasing p_r and p_i, all imply a strong resistance of *RDSII* against these recruitment errors, as long as the recruitment chains are able to continue and generate the target sample size.

It should be noted that these simulations do not test all types of violations of assumption iv and vi. If we for

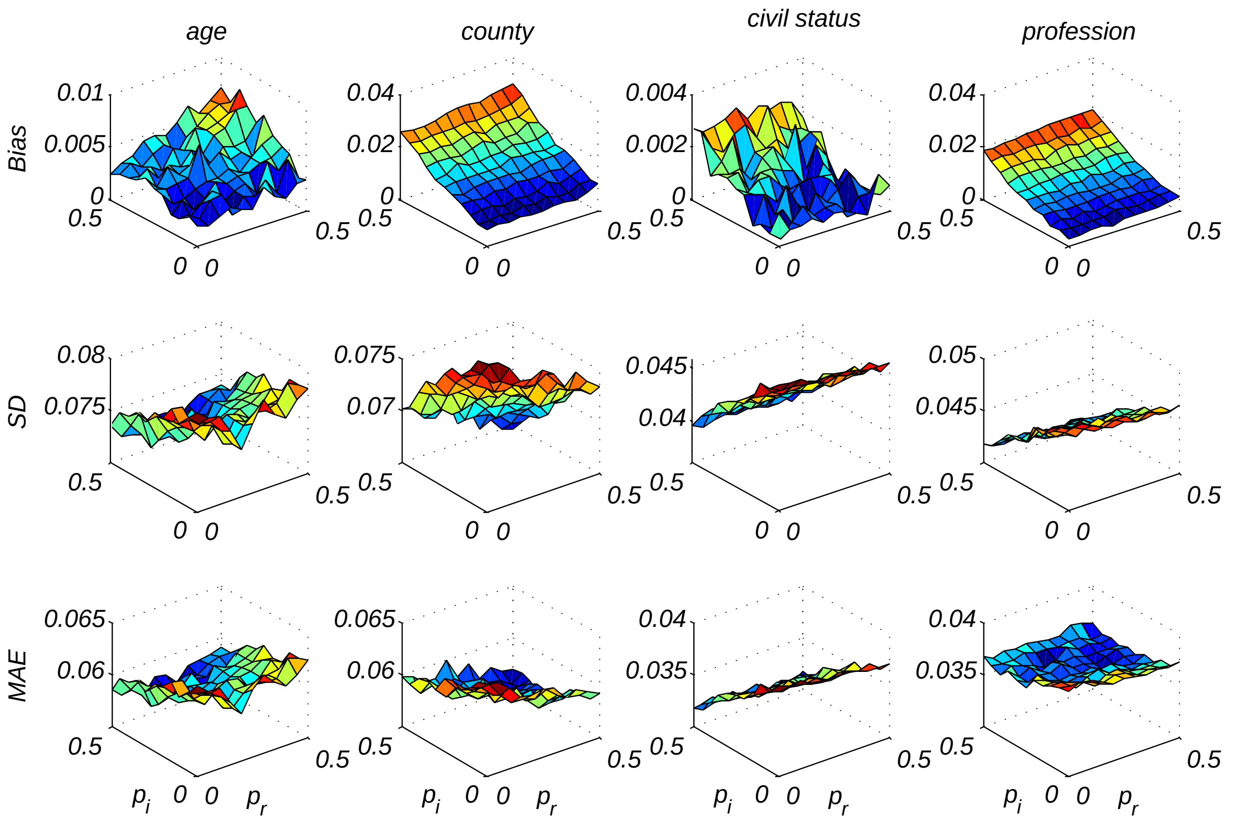

Figure 5: Sampling with ignore and reject probabilities in the original, undirected network ($G1$). Simulation was repeated 10,000 times for each combination, number of seeds=10, coupon=3, with replacement. Seeds were randomly selected at the beginning of each simulation. Sample size is 500.

example set p_i and p_r as dependent on any of our four outcome variables, errors could be much larger than described above. To evaluate these effects, we performed further simulations in which both the ignore and the reject probabilities differed, depending on group membership.

Let p_i be the probability that a member in the group of interest will be ignored by his friends when these friends are given the possibility to recruit, and let p_r be the probability that a coupon will be rejected by a member in the group of interest. Similarly, let p'_i and p'_r and be the corresponding ignore and reject probabilities for members in groups of noninterest. Surfaces for RDS with unequal recruiting probabilities are presented in Appendix D2. We can see that when the ignore or reject probabilities depended on the characteristics of the members, the RDS estimates gave large bias and error. Take the first simulation in Appendix D2, for example: when members born before 1980 rejected half of the invitations given to them and the members born after 1980 did not reject any invitations (p_i,and p'_i both set to 0), the bias was over 0.3 for *age*.

When $p_i = p'_i$, the Bias and MAE are small as long as $p_r = p'_r$ and vice versa. As both the ignore and reject actions will reduce the inclusion probabilities of group members, they have similar effects on the RDS estimates and can compensate for each other. For example, when the fixed ignore probabilities were $p_i = 0.1$, $p'_i = 0.3$, the minimum Bias and MAE were when $p_r > p'_r$. For all the simulations, the values of MAE are almost the same as those of the Bias, revealing that when the studied groups hold different ignore or reject probabilities, the RDS estimates will virtually always be too high or too low. Although differences between groups in p_i can be compensated for by inverse differences between the groups in p_r, it can be hypothesized that such a combination of probabilities would be unusual in real life. As participants in RDS studies are rewarded for successful recruitments, a rational and self-interested participant would seek to ignore contacts whom he/she considers unlikely to accept an invitation. This would mean that groups with high p_r would also have a high p_i. Unfortunately, such combinations of p_i and p_r

always give rise to the largest bias and MAE.

4.4. *Preferential recruitment*

According to Eq. (2), it is easy to infer that the *RDSII* estimator would be biased when recruitment is a non-random selection among the edges of each node, as Eq. (4) is no longer the equilibrium.

A plausible non-random recruitment scenario would be that contacts with whom the recruiter interacts more frequently have a higher probability of being invited than those with whom the recruiter interacts only seldom. Simulation results for RDS on $G1_{max}$, in which respondents were supposed to recruit peers with a probability proportional to the edge weights, are presented in Figure 6. We can see that the *RDSII* estimations were no longer unbiased for all groups. The biases were around 1%, 2%, 4%, and 3% for the four groups, respectively. When we compare the SD and MAE of preferential recruitment (PR) with uniform recruitment (UR), the preferential recruitment had larger SD and MAE values for all groups, indicating that if the respondents prefer to distribute the coupons to their closer friends, larger *RDSII* estimation errors might occur. Again when we use the *eig* estimations on the weighted network, they agree well with the true population.

Simulations on $G1_{min}$ reveal similar outcomes, see Appendix E.

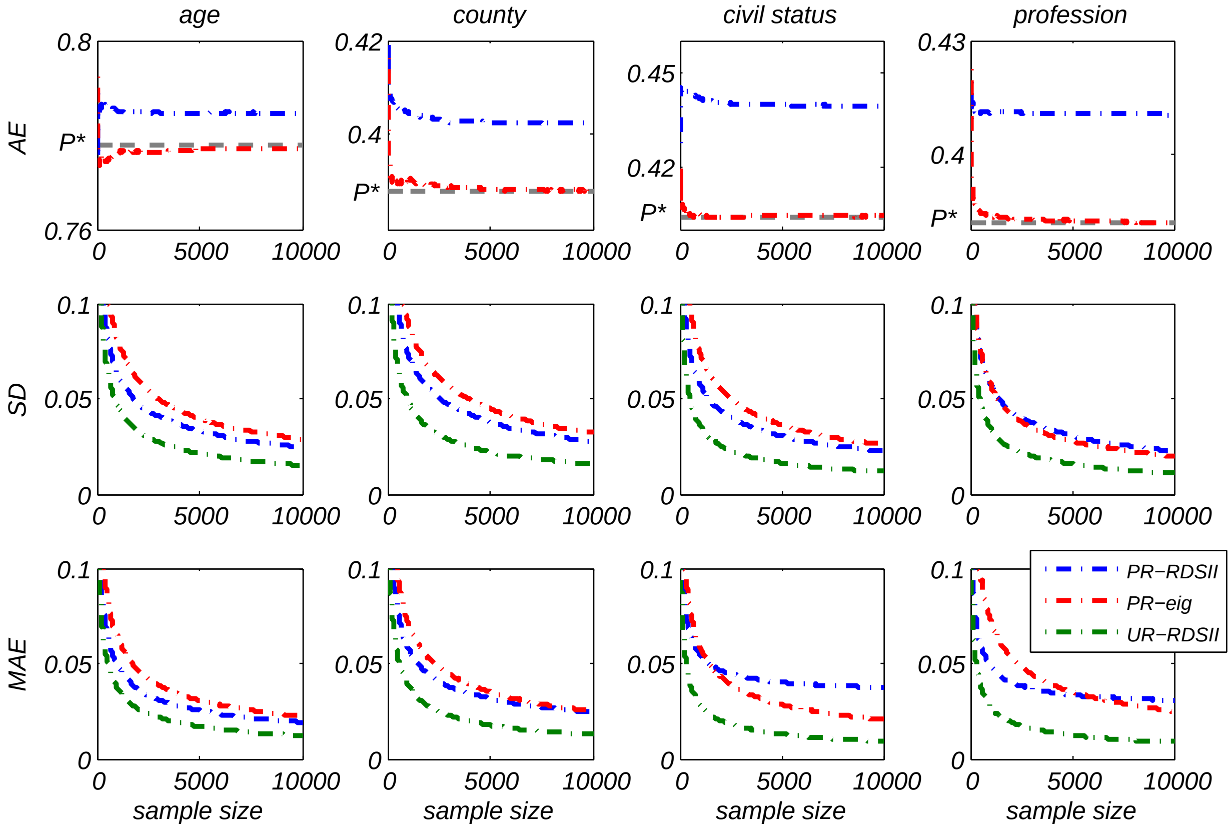

Figure 6: RDS on $G1_{max}$ with preferential recruitment. Number of seeds=1, coupons=1, sampling with replacement. Seeds were randomly selected at the beginning of each simulation. The blue lines represent estimations by *RDSII* and red lines represent estimations by eigenvector. Green lines are the *RDSII* estimations for recruitment with uniform probability. Dashed gray lines indicate the true population values.

4.5. *Effect of seeds and coupons*

Determination of the number of coupons per participant and the number and characteristics of the seeds are among the first problems encountered by researchers when preparing an RDS study. To evaluate the effects of variations in these parameters, we increased the number of coupons and seeds, with the seed(s) being selected either with uniform probability (type 1) or with probability proportional to the nodes' degree (type 2). Results are shown in Figure 7. The difference between selection types for both SD and MAE were minute, and we therefore do not show SD and MAE for different selection types separately.

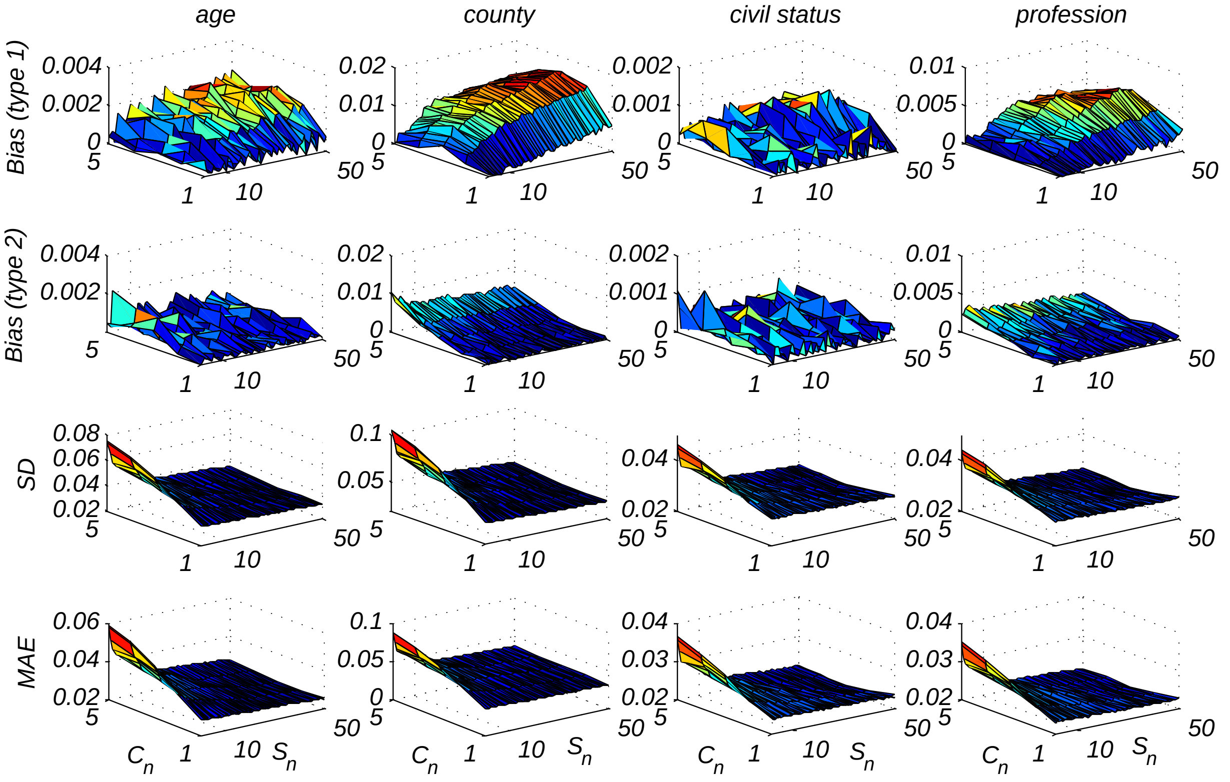

Figure 7: Effects of varying the number of seeds and coupons in $G1$ when sample size was 500, with replacement. Simulation repeated 10,000 times for each combination. C_n stands for the number of coupons and S_n for the number of seeds.

The first impression of this test is that an RDS study started by a type 2 selection seems to have a somewhat smaller bias than an RDS study started by a type 1 selection. The difference is slightly larger for *age* and *county*, which have larger homophilies. The number of seeds and coupons had small effects on the average *RDSII* estimates, but had an obvious effect on the SD and MAE: both the SD and MAE increased when the samplings used more coupons. This is probably because a study with a small number of coupons needs longer chains to reach the same sample size. Longer chains, in turn, are more likely to break out of homogenous sub-networks and therefore become more representative of the overall population. Simulations on $G1_{add}$ and $G1_{rand}$ point to the same conclusions and are shown in Appendix F. Real-life RDS studies cannot however use too few coupons or seeds as they will fail in generating long recruitment chains which are long enough.

5. Conclusions

Real social networks and the recruiting behavior of people in those networks can hardly meet all of the assumptions underlying the RDS estimators. Therefore it is crucial to know how much estimates are affected by deviations from these assumptions. This paper describes to the best of our knowledge, the first study simulating RDS studies within an actual hidden population and which in doing so can compare true population values to estimates obtained under various deviations from the *RDSII* assumptions.

Our simulations show that when all the assumptions underlying the *RDSII* estimator are fulfilled, results are excellent and asymptotically biased.

Further we show that the estimates adjusted by the eigenvector of eigenvalue 1 for the transition matrix are always unbiased for all groups and networks. This is further the reason why the *RDSII* estimator is biased when the network is directed or when recruitment is a non-random selection from the participants' social networks. We should note that currently this eigenvector cannot be inferred from the empirical data in an RDS study where only the information of recruitment chain and personal network sizes are known. However, our findings show that given knowledge about this eigenvector it is possible with RDS to make unbiased estimates even on directed networks.

It turns out that when sample sizes were relatively small, sampling without replacement, which is used in practice, actually had a slightly lower standard deviation and mean absolute error than simulations with replacement.

The *RDSII* estimator shows strong resistance to recruitment error when the probabilities that individuals will ignore contacts and reject invitations are independent of the individuals' characteristics. On the other hand, if these probabilities are dependent on the individuals' characteristics and if these characteristics are correlated with the outcome characteristic one wishes to estimate, the bias and MAE could become very large. As participants in RDS studies are rewarded for successful recruitments, rational and self-interested participants could be expected to ignore contacts who are considered less likely to accept invitations. Simulations show that such a combination of a group having a high probability of rejecting invitations and a high probability of being ignored can give rise to very large bias and error. We suggest that RDS studies should routinely compare participants reported network composition with actual recruitments to provide further empirical evidence on this issue from a wide variety of contexts.

Besides testing the violation of assumptions, we also analyzed some of the effects of network structure and homophily on the *RDSII* estimator. The results are consistent with previous studies (Volz and Heckathorn, 2008; Gile and Handcock, 2009; Goel and Salganik, 2009): networks which were sparse and had a skewed degree distribution had larger error and bias, and estimations of groups with small homophily performed better than estimations among groups with high homophily.

The deviations from the assumptions that we have simulated in this paper can be modeled in different ways, which affects the conclusions. We have opted for deviations that we consider relevant to RDS studies in different contexts, but the simulations still represent subjective choices and do not cover all situations relevant to real life RDS studies. Moreover, while we have tried to make results more generalizable by varying the properties of the original network, the network characteristics actually picked for simulations do not reflect all types of networks, which could impact on the interpretations of the results. For further studies it would be valuable to look at the effects of combinations of violations of the assumption in order to let the simulations better approximate reality.

Acknowledgement

This work is funded in part by the Riksbankens jubileumsfond (dnr: P2008-0674). B.J.K. acknowledges the support from the Korea Science and Engineering Foundation (Grant No. R01-2007-000-20084-0), and X.L. would like to thank China Scholarship Council (Grant No. 2008611091).

Appendix

App-A Degree distribution for edge-added and edge randomized networks

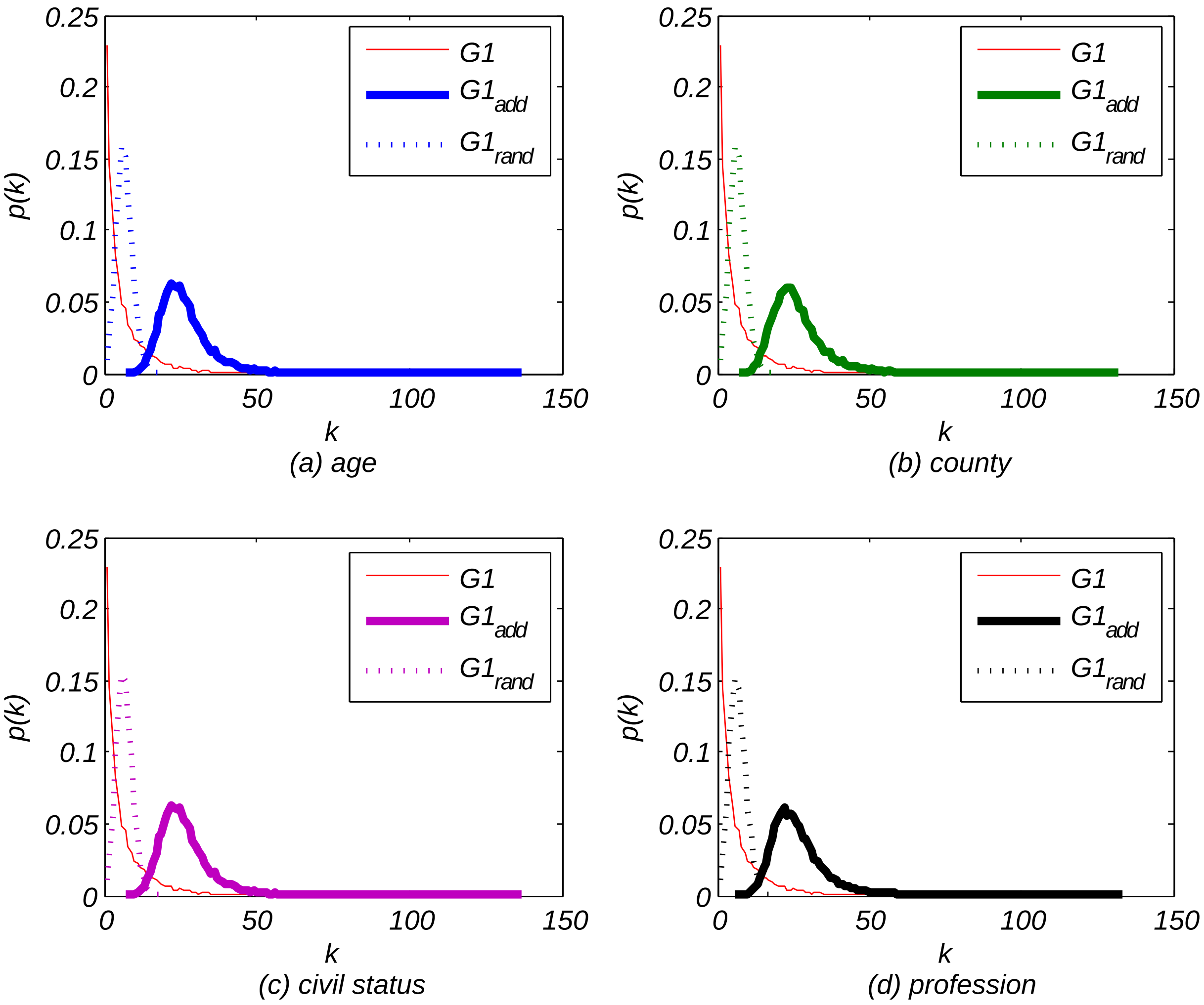

Figure 8: Degree distribution for edge-added and edge-randomized networks. Red solid lines stand for the original undirected network ($G1$), bold lines for networks whose average degrees were increased by 20 ($G1_{add}$), and dashed lines for networks whose edges were randomly rewired ($G1_{rand}$)

App-B RDS on undirected networks: small sample sizes

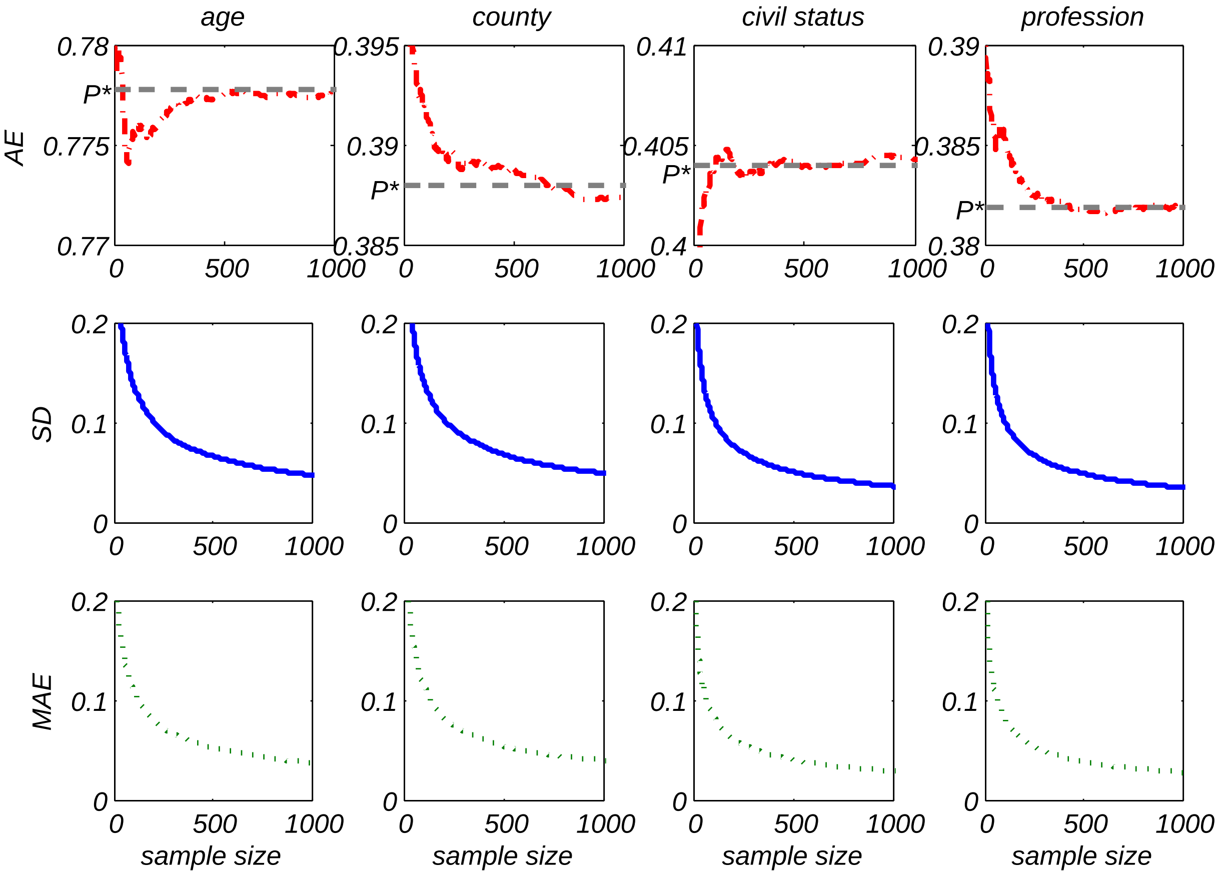

Figure 9: *RDSII* estimations on the undirected network ($G1$). The average estimates approach the true proportions very fast. When the sample size was 500 the Bias was only 0.0002, 0.0009, 0.00002, 0.0002 for *age*, *county*, *civil status* and *profession*, respectively.

App-C Sampling without replacement

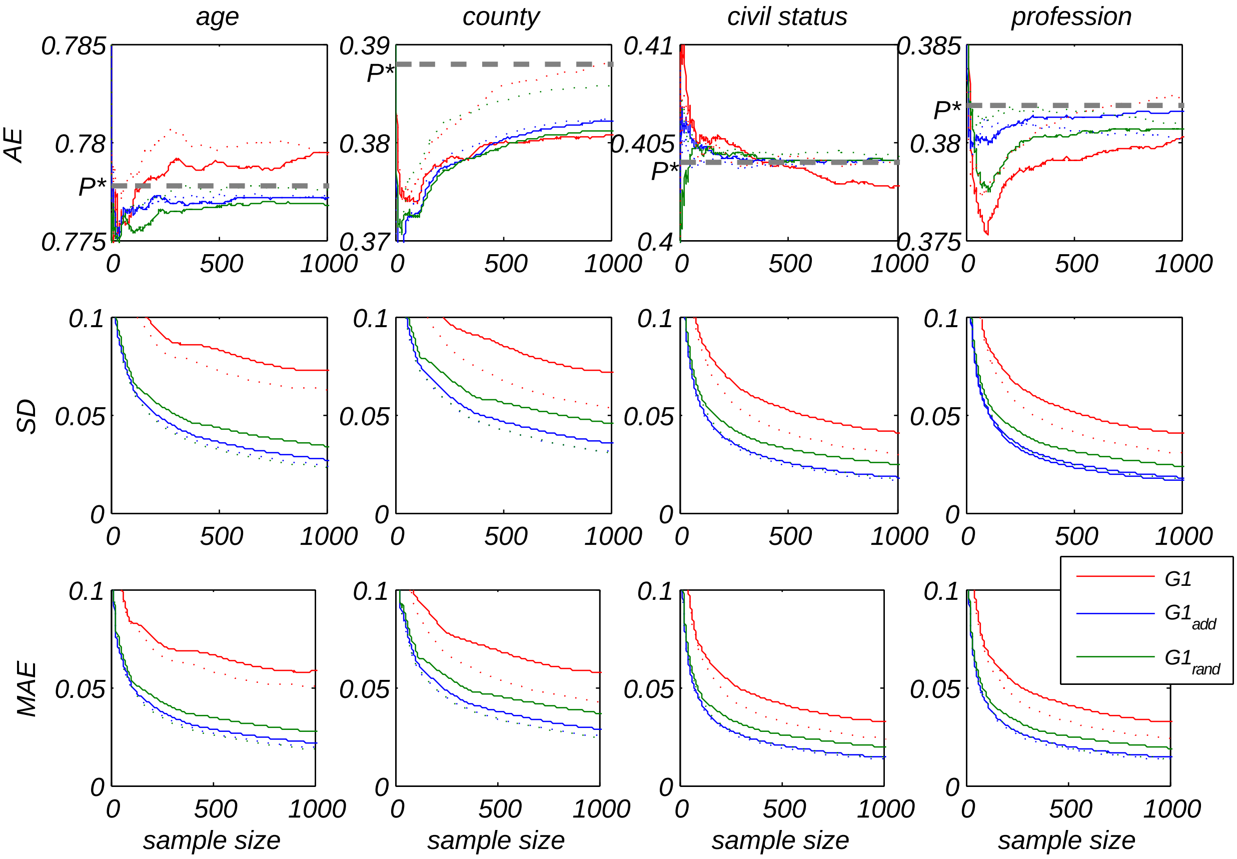

Figure 10: Effects of network structure and replacement. Number of seeds=10, coupons=3. Seeds were randomly selected at the beginning of each simulation. Solid lines represent sampling with replacement and dashed lines represent sampling without replacement.

App-D Sampling with ignore and reject probabilities

All simulations were repeated 10,000 times for each combination of probabilities, number of seeds=10, coupon=3, with replacement. Seeds were randomly selected at the beginning of each simulation. Estimates were calculated when sample sizes reached 500.

App-D1 RDS with ignore and reject probabilities independent of the individuals' characteristics

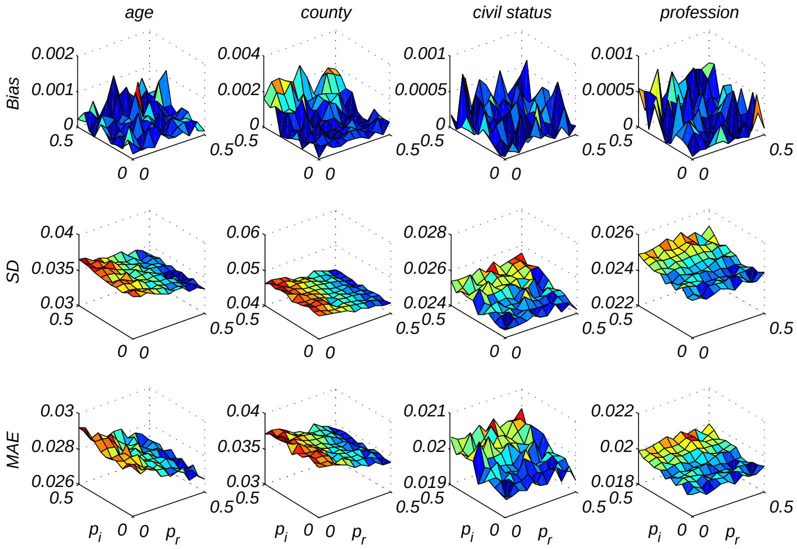

Figure 11: Sampling with ignore and reject probabilities in the edge-added networks ($G1_{add}$).

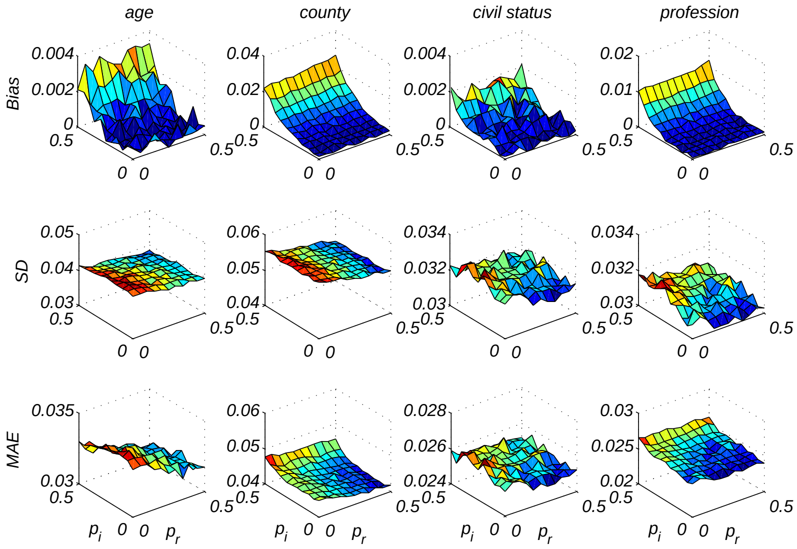

Figure 12: Sampling with ignore and reject probabilities in the edge-randomized networks ($G1_{rand}$).

App-D2 RDS with ignore and reject probabilities dependent on the individuals' characteristics

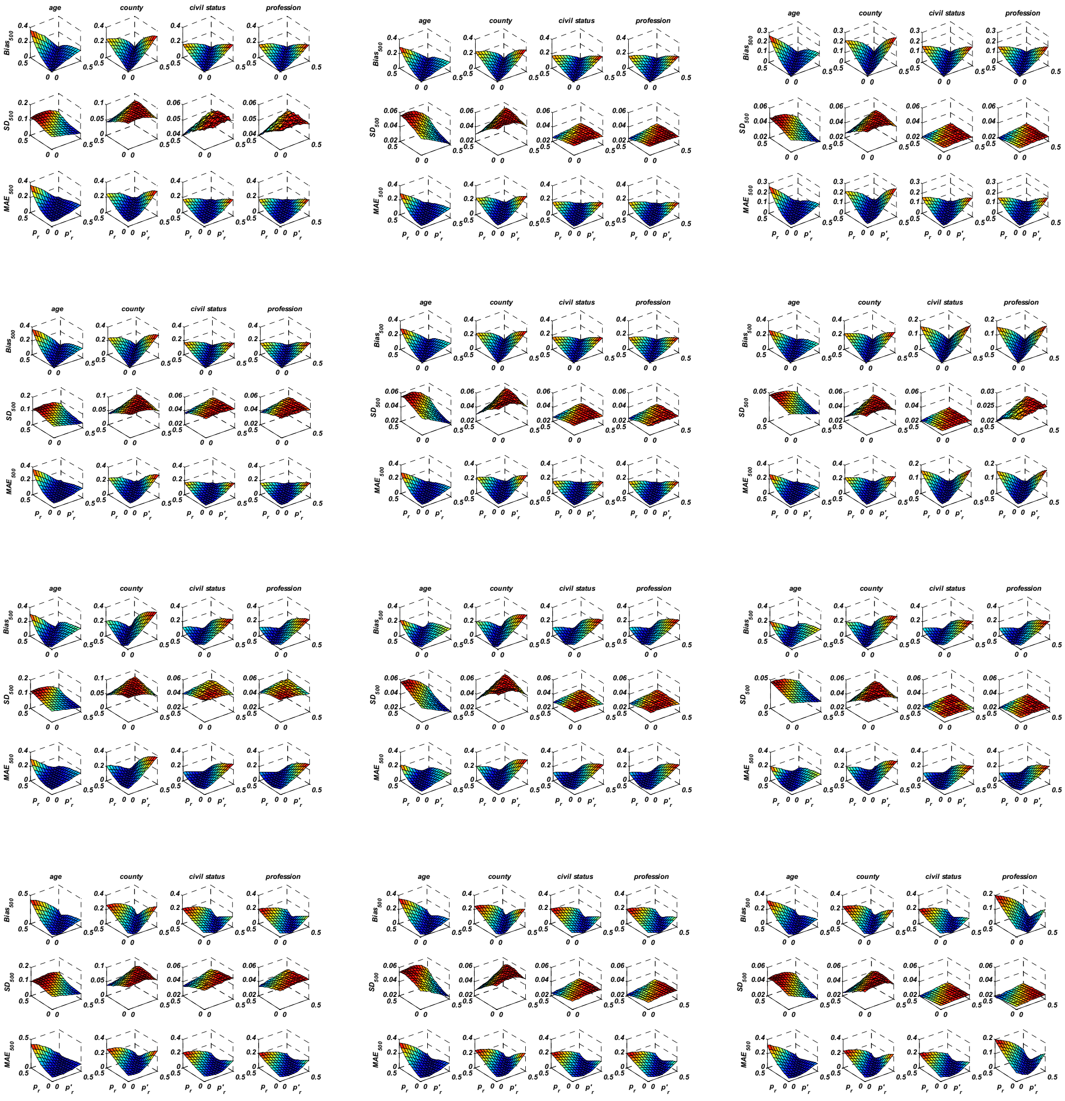

Figure 13: RDS with ignore and reject probabilities. From top to bottom: $p_i = 0$, $p'_i = 0$; $p_i = 0.2$, $p'_i = 0.2$; $p_i = 0.1$, $p'_i = 0.3$; $p_i = 0.3$, $p'_i = 0.1$. (Left: $G1$, Middle: $G1_{rand}$, Right: $G1_{add}$)

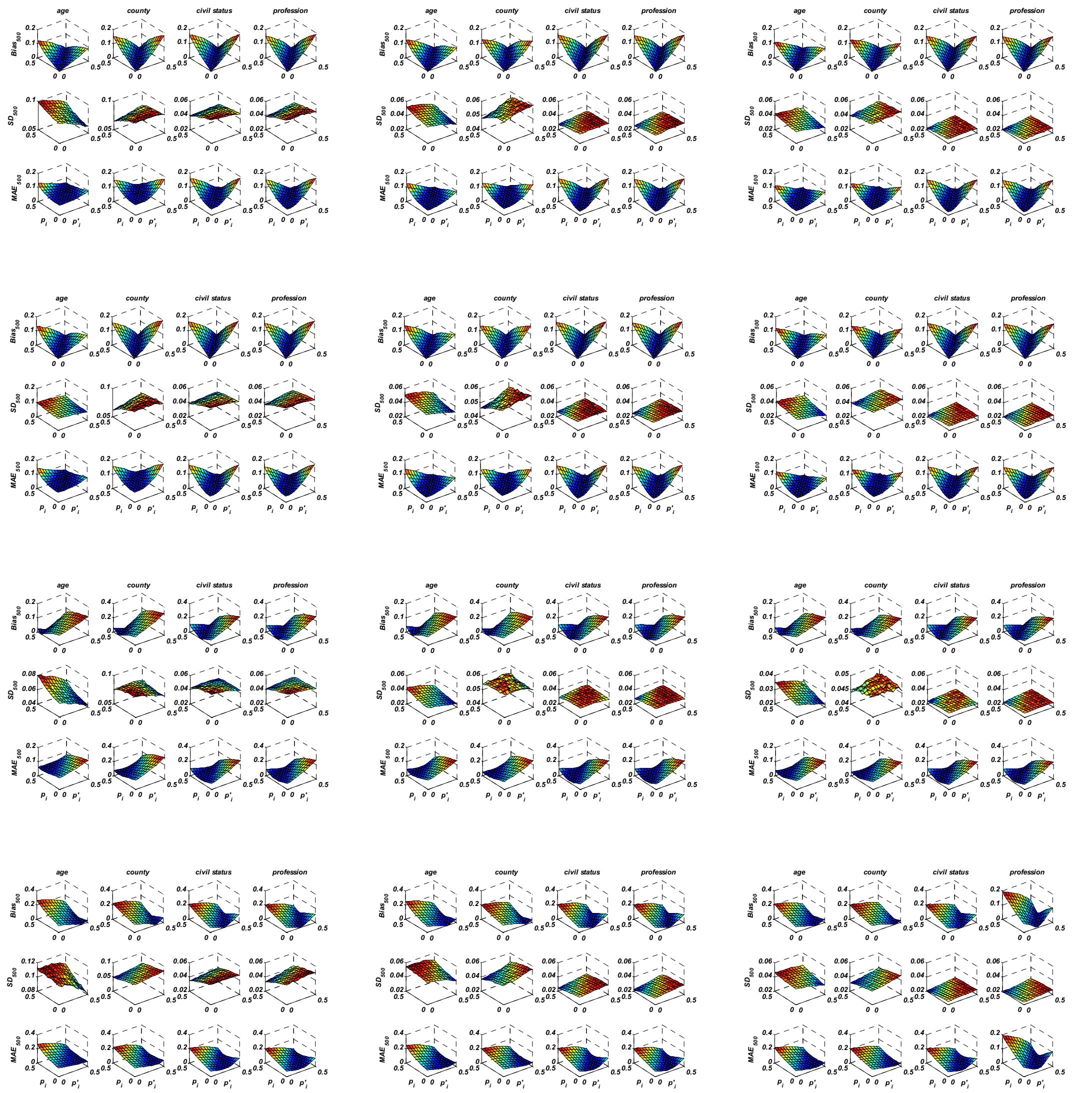

Figure 14: RDS with ignore and reject probabilities. From top to bottom: $p_r = 0$, $p'_r = 0$; $p_r = 0.2$, $p'_r = 0.2$; $p_r = 0.1$, $p'_r = 0.3$; $p_r = 0.3$, $p'_r = 0.1$. (Left: $G1$, Middle: $G1_{rand}$, Right: $G1_{add}$)

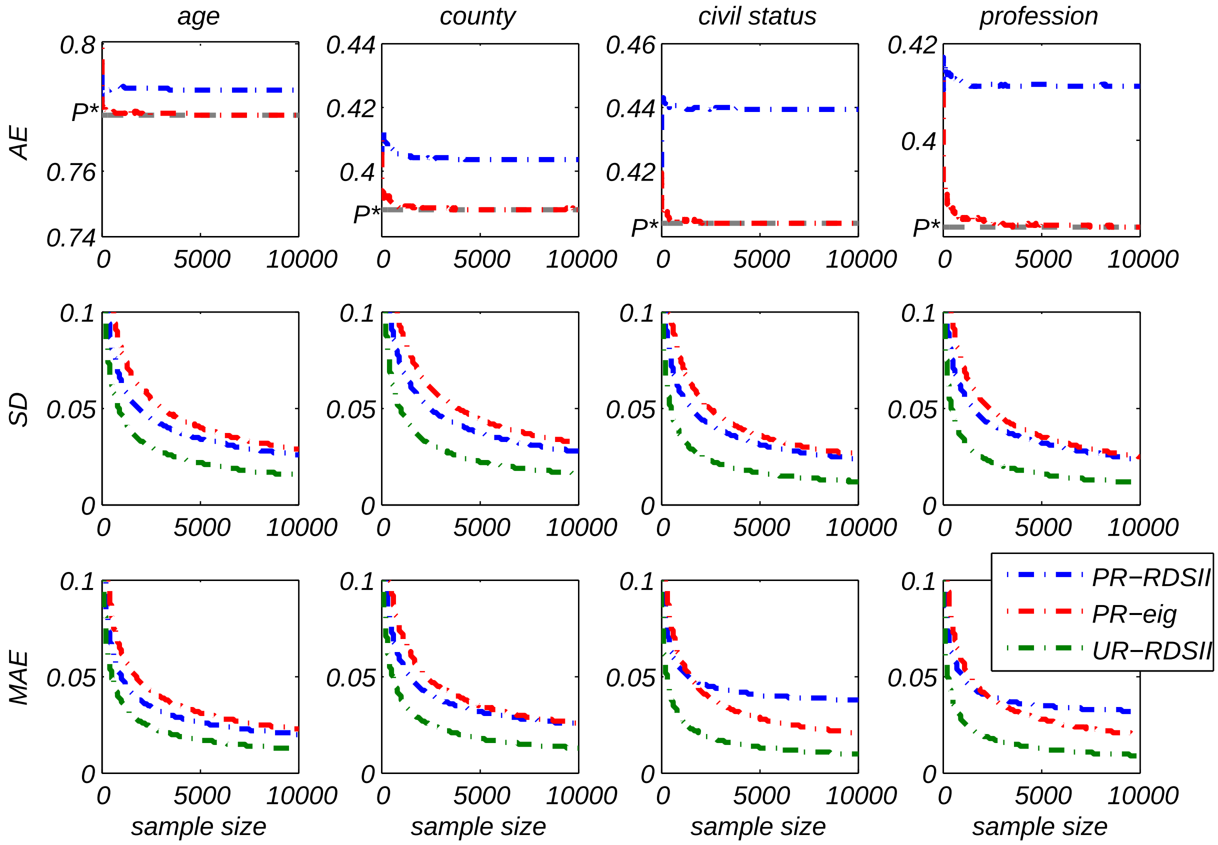

Figure 15: RDS on $G1_{min}$ with preferential recruitment. Number of seeds=1, coupons=1, sampling with replacement. Seeds were randomly selected at the beginning of each simulation. The blue lines represent estimations by *RDSII* and red lines represent estimations by eigenvector. Green lines are the *RDSII* estimations for recruitment with uniform probability. Dashed gray lines indicate the true population values.

App-F Effect of seeds and coupons

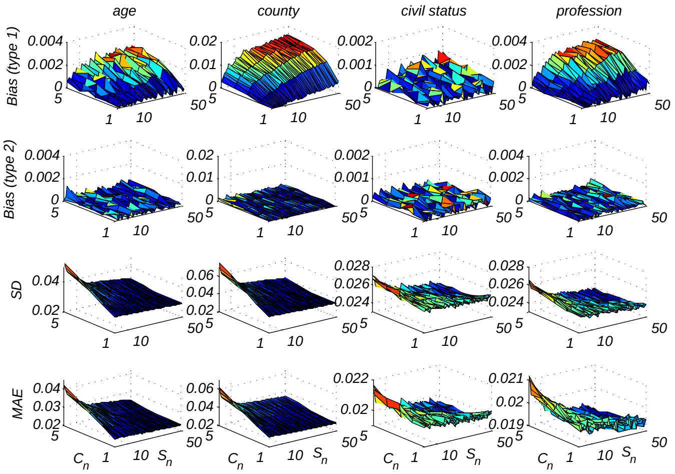

Figure 16: Effects of varying the number of seeds and coupons in $G1_{add}$ when sample size was 500, with replacement. Simulation repeated 10,000 times for each combination. C_n stands for the number of coupons and S_n for the number of seeds.

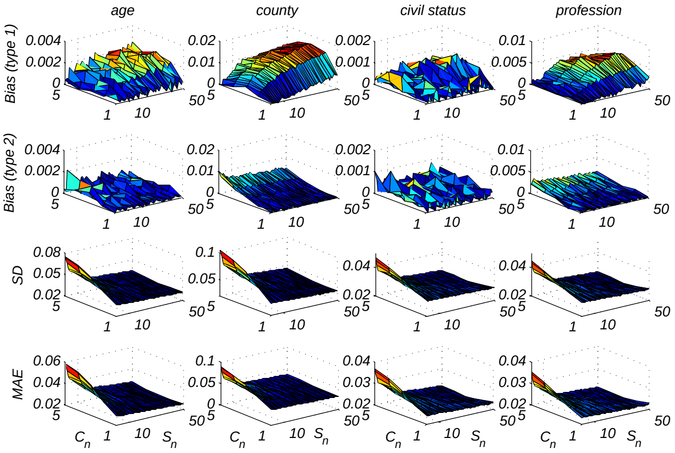

Figure 17: Effects of varying the number of seeds and coupons in $G1_{rand}$ when sample size was 500, with replacement. Simulation repeated 10,000 times for each combination. C_n stands for the number of coupons and S_n for the number of seeds.

References

Abdul-Quader, A., Heckathorn, D., McKnight, C., Bramson, H., Nemeth, C., Sabin, K., Gallagher, K., Des Jarlais, D., 2006. Effectiveness of respondent-driven sampling for recruiting drug users in new york city: Findings from a pilot study. Journal of Urban Health 83 (3), 459–476, 10.1007/s11524-006-9052-7.

Brin, S., Page, L., 1998. The anatomy of a large-scale hypertextual web search engine. In: Computer Networks and ISDN Systems. pp. 107–117.

Deaux, E., Callaghan, J., 1985. Key informant versus self-report estimates of health behavior. Evaluation Rev 9, 365–368.

Erickson, B. H., 1979. Some problems of inference from chain data. Sociological Methodology 10, 276–302.

Gile, K. J., Handcock, M. S., 2009. Respondent-driven sampling: An assessment of current methodology.
URL http://www.citebase.org/abstract?id=oai:arXiv.org:0904.1855

Goel, S., Salganik, M. J., 2009. Respondent-driven sampling as markov chain monte carlo. Statistics in Medicine 28 (17), 2202–2229.

Heckathorn, D. D., 1997. Respondent-driven sampling: A new approach to the study of hidden populations. Social Problems 44 (2), 174–199.

Heckathorn, D. D., 2002. Respondent-driven sampling ii: Deriving valid population estimates from chain-referral samples of hidden populations. Social Problems 49 (1), 11–34.

Heckathorn, D. D., 2007. Extensions of respondent-driven sampling: analyzing continuous variables and controlling for differential recruitment. In: Sociological Methodology 2007, Vol 37. Vol. 37 of Sociological Methodology. Blackwell Publishing, Oxford, pp. 151–208.

Heckathorn, D. D., Jeffri, J., 2001. Finding the beat: Using respondent-driven sampling to study jazz musicians. Poetics 28 (4), 307–329.

Heimer, R., 2005. Critical issues and further questions about respondent-driven sampling: Comment on ramirez-valles, et al . (2005). Aids and Behavior 9 (4), 403–408, 10.1007/s10461-005-9030-1.

Johnston, L. G., Malekinejad, M., Kendall, C., Iuppa, I. M., Rutherford, G. W., 2008. Implementation challenges to using respondent-driven sampling methodology for hiv biological and behavioral surveillance: Field experiences in international settings. Aids and Behavior 12 (4), S131–S141.

Magnani, R., Sabin, K., Saidel, T., Heckathorn, D., 2005. Review of sampling hard-to-reach and hidden populations for hiv surveillance. Aids 19, S67–S72.

Malekinejad, M., Johnston, L. G., Kendall, C., Kerr, L., Rifkin, M. R., Rutherford, G. W., 2008. Using respondent-driven sampling methodology for hiv biological and behavioral surveillance in international settings: A systematic review. Aids and Behavior 12 (4), S105–S130.

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

Morris, M., Kretzschmar, M., 1995. Concurrent partnerships and transmission dynamic in networks. Social Networks 17 (3-4), 299–318.

Page, L., Brin, S., Motwani, R., Winograd, T., 1999. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab.

Rapoport, A., 1980. A probabilistic approach to networks. Social Networks 2, 1–18.

Rybski, D., Buldyrev, S. V., Havlin, S., Liljeros, F., Makse, H. A., 2009. Scaling laws of human interaction activity. Proceedings of the National Academy of Sciences of the United States of America 106 (31), 12640–12645.

Salganik, M. J., 2006. Variance estimation, design effects, and sample size calculations for respondent-driven sampling. Journal of Urban Health-Bulletin of the New York Academy of Medicine 83 (6), I98–I112.

Salganik, M. J., Heckathorn, D. D., 2004. Sampling and estimation in hidden populations using respondent-driven sampling. In: Sociological Methodology, Vol 34. Vol. 34 of Sociological Methodology. Wiley-V C H Verlag Gmbh, Weinheim, pp. 193–239.

Schwarte, N., Cohen, R., Ben-Avraham, D., Barabasi, A. L., Havlin, S., 2002. Percolation in directed scale-free networks. Phys. Rev. E 66 (1), 015104.

Volz, E., Heckathorn, D. D., 2008. Probability based estimation theory for respondent driven sampling. Journal of Official Statistics 24 (1), 79–97.

Watters, J. K., Biernacki, P., 1989. Targeted sampling: Options for the study of hidden populations. Social Problems 36 (4), 416–430.

Wejnert, C., Heckathorn, D. D., 2008. Web-based network sampling - efficiency and efficacy of respondent-driven sampling for online research. Sociological Methods and Research 37 (1), 105–134.

Woess, W., 1994. Random-walks on infinite-graphs and groups: A survey on selected topics. Bulletin of the London Mathematical Society 26, 1–60.