Linked Ego Networks: Improving Estimate Reliability and Validity with Respondent-driven Sampling

Xin Lu^{a,b,c,*}

^aDepartment of Public Health Sciences, Karolinska Institutet, Stockholm, Sweden ^bDepartment of Sociology, Stockholm University, Stockholm, Sweden ^cCollege of Information Systems and Management, National University of Defense Technology, Changsha, China

Abstract

Respondent-driven sampling (RDS) is currently widely used for the study of HIV/AIDS-related high risk populations. However, recent studies have shown that traditional RDS methods are likely to generate large variances and may be severely biased since the assumptions behind RDS are seldom fully met in real life. To improve estimation in RDS studies, we propose a new method to generate estimates with ego network data, which is collected by asking RDS respondents about the composition of their personal networks, such as "*what proportion of your friends are married?*". By simulations on an extracted real-world social network of gay men as well as on artificial networks with varying structural properties, we show that the new estimator, $RDS I^{ego}$ shows superior performance over traditional RDS estimators. Importantly, $RDS I^{ego}$ exhibits strong robustness to the preference of peer recruitment and variations in network structural properties, such as homophily, activity ratio, and community structure. While the biases of traditional RDS estimators can sometimes be as large as $10\% \sim 20\%$, biases of all $RDS I^{ego}$ estimates are well restrained to be less than 2%. The positive results henceforth encourage researchers to collect ego network data for variables of interests by RDS, for both hard-to-access populations and general populations when random sampling is not applicable. The limitation of $RDS I^{ego}$ is evaluated by simulating RDS assuming different level of reporting error.

Keywords: networks, ego networks, respondent-driven sampling, differential recruitment, reporting error

1. Introduction

In many forms of research, there is no list of all members for the studied population (i.e., a *sampling frame*) from which a random sample may be drawn and estimates about the population characteristics may be inferred based on the select probabilities of sample units. Non-probability sampling methods may be used for for such situations, such as key informant sampling [1], targeted/location sampling [2], and snowball sampling [3]. However, these methods all introduce a considerable selection bias, which impairs generalization of the findings from the sample to the studied population [4, 5]. Respondent-driven sampling (RDS) is an alternative method that is currently being used extensively in public health research for the study of *hard-to-access* populations, e.g., injecting drug users (IDUs), men who have sex with men (MSM) and sex workers (SWs). With a link-tracing network sampling design, the RDS method provides unbiased population estimates as well as a feasible implementation, making it the state-of-the-art sampling method for studying *hard-to-access* populations [6, 7, 8, 9, 10].

RDS starts with a number of pre-selected respondents who serve as "seeds". After an interview, the seeds are asked to distribute a certain number of coupons (usually 3) to their friends who are also within the studied population. Individuals with a valid coupon can then participate in the study and are provided the same number of coupons to distribute. The above recruitment process is repeated until the desired sample size is reached [4]. In a typical RDS, information about who recruits whom and the respondents' number of friends within the population (degree) are also recorded for the purpose of generating population estimates from the sample [11, 12].

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

Suppose a RDS study is conducted on a connected network with the additional assumptions that (i) network links are undirected, (ii) sampling of peer recruitment is done with replacement, (iii) each participant recruits one peer from his/her neighbors, and (iv) the peer recruitment is a random selection among all the participant's neighbors. Then the RDS process can be modeled as a Markov process, and the composition of the sample will stabilize and be independent of the properties of the seeds [12, 13, 14]. Following this, the probability for each node to be included in the RDS sample is proportional to its degree. Specifically, for a given sample $U = \{v_1, v_2, \ldots, v_n\}$, with n_A being the number of respondents in the sample with property A (e.g., HIV-positive) and $n_B = n - n_A$ being the rest. Let $\{d_1, d_2, \ldots, d_n\}$ be the respondents' degree and $S = \begin{bmatrix} s_{AA} & s_{AB} \\ s_{BA} & s_{BB} \end{bmatrix}$ be the recruitment matrix observed from the sample, where s_{XY} is the proportion of recruitments from group X to group Y (for the purpose of this paper, we consider a binary property such that each individual belongs either to group A or B). Then the proportion of individuals belonging to group A in the population, P_A^* , can be estimated by [12, 14]:

$$\hat{P}_A = \frac{s_{BA}\bar{D}_B}{s_{AB}\bar{D}_A + s_{BA}\bar{D}_B} (RDSI), \tag{1}$$

or

$$\hat{P}_A = \frac{\sum\limits_{v_i \in A \cap U} d_i^{-1}}{\sum\limits_{v_i \in U} d_i^{-1}} (RDS II),$$
(2)

where $\hat{D}_A = \frac{n_A}{\sum\limits_{v_i \in A \cap U} d_i^{-1}}$ and $\hat{D}_B = \frac{n_B}{\sum\limits_{v_i \in B \cap U} d_i^{-1}}$ are the estimated average degrees for individuals of group *A* and *B* in the population. Both estimators give asymptotically unbiased estimates. The estimation procedure above is also called a

re-weighted random walk (RWRW) in other fields [15].

The methodology of RDS is nicely designed; however, the assumptions underlying the RDS estimators are rarely met in practice [7, 16, 17]. For example, empirical RDS studies use more than one coupon and sampling is conducted without replacement, that is, each respondent is only allowed to participate once. A comprehensive evaluation has been made by Lu et al [18], where the effects of violation of assumptions (i)~(iv), as well as the effect of selection and number of seeds and coupons, were evaluated one by one, by simulated RDS process on an empirical MSM network as well as artificial networks with known population properties. They have shown that when the sample size is relatively small (< 10% of the population), RDS estimators have a strong resistance to violations of certain assumptions, such as low response rate and errors in self-reporting of degrees, and the like. On the other hand, large bias and variance may result from differential recruitments, or from networks with irreciprocal relationships. When the sample size is relatively large (> 50% of the population), similar results were also found by Gile and Handcock [19], where they focused on the sensitivity of RDS estimators to the selection of seeds, respondent behavior and violation of assumption (ii).

It was not until recently that researchers found the variance in RDS may have been severely underestimated [20]. In a study by Goel and Salganik [17] based on simulated RDS samples on empirical networks, they found that the RDS estimator typically generates five to ten times greater variance than simple random sampling [20]. Moreover, McCreesh et al [21] conducted a RDS study on male household heads in rural Uganda where the true population data was known, and they found that only one-third of RDS estimates outperformed the raw proportions in the RDS sample, and only 50%-74% of RDS 95% confidence intervals, calculated based on a bootstrap approach for RDS, included the true population proportion.

For the above reasons, there has been an increasing interest in developing new RDS estimators to improve the performance of RDS. For example, Gile [22] developed a successive-sampling-based estimator for RDS to adjust the assumption of sampling with replacement and demonstrated its superior performance when the size of the population is known. Lu et al [23] proposed a series of new estimators for RDS on directed networks, with known indegree difference between estimated groups. Both of the above estimators can be used as a sensitivity test when the required population parameters are not known.

Both the traditional *RDS I*, *RDS II* estimators, and the estimators newly developed by Gile et al [22, 24] and Lu et al [23] utilize the same information collected by standard RDS practice, that is, the recruitment matrix *S*, respondents'

degree, and sample property. There is however scope to improve estimates dramatically if data on the composition of respondents' ego networks can be put to use. Such data has already been collected for other purposes in many RDS studies. For example, in a RDS study of MSM in Campinas City, Brazil, by de Mello et al [25], respondents were asked to describe the percentage of certain characteristics among their friends/acquaintances, such as disclosure of sexual orientation to family, HIV status, and the like. In a RDS study of opiate users in Yunnan, China, various information about supporting, drug using, and sexual behaviors between respondents and their network members were collected [26]. One of the most thorough RDS studies utilizing ego network information was done by Rudolph et al [27], in which they asked the respondents to provide extensive characteristics for each alter within their personal networks such as demographic characteristics, history of incarceration, and drug injection and crack and heroin use.

Aiming to improve the RDS estimator, we will focus on how to integrate this additional information in the estimation process to generate improved population estimates. The rest of this paper is organized as follows. In Section 2, we develop a new estimator that integrates traditional RDS data with egocentric data; in Section 3, we describe network data used for simulation and study design; in Section 4, we evaluate the performance of the new estimator by simulated RDS processes under various settings; and in Section 5, we summarize and draw our conclusions.

2. RDSI^{ego}: estimator for RDS with egocentric data

The ego networks from a RDS sample differ from general egocentric data collected in many sociological surveys [28] in the way that each "ego" is connected with (recruited by) its recruiter. For example, in a partial chain of RDS as illustrated in Figure 1, participants v_i , v_j , v_k , are asked to provide personal network compositions and v_j and v_k are recruited by v_i , v_j , respectively.

For each respondent v_i in a RDS sample $U = \{v_1, v_2, \dots, v_n\}$, let n_i^A , n_i^B be the number of v_i 's friends with property A, B, respectively. We then start to show how to integrate the ego network information for estimating the proportion of individuals with property A in the population, P_A^* . Assuming that the RDS process is conducted on a connected, undirected network with assumptions (i)-(iii) fulfilled, the probability that each node will be included in the sample, $Pr(v_i)$, will be proportional to its degree [12, 13, 14]:

$$Pr(v_i) \sim \frac{d_i}{\sum_{i=1}^N d_i},\tag{3}$$

where *N* is the size of the population of interest.

Consequently, the probability that each link $e_{i \rightarrow j}$ will be selected to recruit a friend, $Pr(e_{i \rightarrow j})$, depends on $Pr(v_i)$. Under the random recruitment assumption, we have:

$$Pr(e_{i\to j}) = \Pr(v_i) \cdot \frac{1}{d_i} \sim \frac{1}{\sum_{i=1}^N d_j},\tag{4}$$

that is, each link has the same probability of being selected via the RDS process. Consequently, the observed recruitment matrix S is a random sample for the cross-group links of the network [12].

The above are general inferences from a typical RDS process. Up to now, we can turn our attention to the egocentric data source. Let $Pr(e_{i \to j}^{ego})$ be the probability that link $e_{i \to j}$ will be reported by "ego" v_i , since $e_{i \to j}$ is reported as long as v_i is included in the sample, then:

$$\Pr(e_{i \to j}^{ego}) = \Pr(v_i) \sim \frac{d_i}{\sum_{i=1}^N d_j}.$$
(5)

Consequently, to estimate the proportion of type $e_{X \to Y}$ ($X, Y \in \{A, B\}$) links in the population, s_{XY}^* , we can weigh the observed number of type $e_{X \to Y}^{ego}$ links by their inclusion probability to construct a *generalized Hansen-Hurwitz* estimator [29]:

$$\hat{s}_{XY}^{ego} = \frac{\hat{N}_{XY}^{ego}}{\hat{N}_{XA}^{ego} + \hat{N}_{XB}^{ego}} = \frac{\sum_{v_i \in X \cap U} \frac{n_i^i}{d_i}}{\sum_{v_j \in X \cap U} \frac{n_j^i}{d_j} + \sum_{v_j \in X \cap U} \frac{n_j^B}{d_j}},$$
(6)



Figure 1: A RDS chain with egocentric data. (a) RDS on a network. Red nodes are those that participated in the RDS survey, and yellow nodes are ego network composition inferred by participants; (b) A partial RDS chain with color representing properties of nodes.

where $\hat{N}_{XY}^{ego} = \sum_{v_i \in X \cap U} \frac{n_i^Y}{d_i}$ is the weighted number of type $e_{X \to Y}$ links reported in the sample's ego networks. Since the denominator in (6) can be rewritten as:

$$\sum_{v_j \in X \cap U} \frac{n_j^A}{d_j} + \sum_{v_j \in X \cap U} \frac{n_j^B}{d_j} = \sum_{v_j \in X \cap U} \left(\frac{n_j^A + n_j^B}{d_j} \right) = \sum_{v_j \in X \cap U} \left(\frac{d_j}{d_j} \right) = n_X, \tag{7}$$

we have:

$$\hat{s}_{XY}^{ego} = \frac{1}{n_X} \cdot \sum_{\nu_i \in X \cap U} \frac{n_i^Y}{d_i}.$$
(8)

Note that in (8), the recruitment links are also counted as reported ego-alter links, and taking Figure 1 as an example, $e_{i \rightarrow j}$ and $e_{j \rightarrow i}$ will be counted as *blue* \rightarrow *green* type ego-alter link and *green* \rightarrow *blue* type ego-alter link, separately.

Using \hat{s}_{XY}^{ego} from (8) as an alternative to S, which is used in the *RDSI* estimator, we can estimate P_A^* by the same equation as (1). For the sake of clarity, the procedure for deriving (1) is replicated as follows:

In an undirected network, the number of cross-group links from A to B should equal the number of links from B

to A:

$$N_A \bar{D}_A^* s_{AB}^* = N_B \bar{D}_B^* s_{BA}^*.$$
(9)

where $N_A = N - N_B$ is the number of individuals of group A in the population, and \bar{D}_A^* , \bar{D}_B^* are average degrees for the two groups.

If we let \hat{s}_{XY}^{ego} be the estimator of s_{XY}^* and let $\hat{D}_X = \frac{n_X}{\sum\limits_{v_i \in X \cap U} d_i^{-1}}$ be the estimator of \bar{D}_X^* ($X, Y \in \{A, B\}$), then P_A^* can be estimated by:

$$\hat{P}_{A} = \frac{\hat{s}_{BA}^{ego}\hat{D}_{B}}{\hat{s}_{AR}^{ego}\hat{D}_{A} + \hat{s}_{RA}^{ego}\hat{D}_{B}} (RDSI^{ego}).$$
(10)

In all, the *RDS I^{ego}* estimator uses the ego network data-based estimation of recruitment matrix, \hat{s}_{XY}^{ego} , instead of the observed *S* used in *RDS I*. There are at least two advantages to using \hat{s}_{XY}^{ego} rather than s_{XY} :

First, the sample size for inferring \hat{s}_{XY}^{ego} , is considerably larger than that for s_{XY} , reducing random error and making the estimates more reliable;

Second, in real RDS practice, respondents can hardly recruit their friends randomly [9, 16, 25], which leads to unknown bias and error for the representativeness of s_{XY} . \hat{s}_{XY}^{ego} , on the other hand, takes all of an ego's links into consideration, and consequently avoids this problem. Even the inclusion probability for a node may be shifted away from $Pr(v_i)$ when there are non-random recruitments; as we will see in section 4, \hat{s}_{XY}^{ego} can greatly reduce estimate bias and error for such violation of assumption.

Note also the the implementation of $RDS I^{ego}$ does not necessarily require each respondent *i* to list each of her/his alters' property: since degree is always collected in RDS, an estimated proportion of friends with a certain property A, r_i^A , would be enough to get to know the number of alters from group A, $r_i^A d_i$.

3. Simulation study design

3.1. Network data

In this paper we use both an anonymized empirical social network and simulated networks to evaluate the performance of the newly proposed estimator. The empirical network, previously analyzed in [18, 23, 30], comes from the Nordic region's largest and most active web community for homosexual, bisexual, transgender, and queer persons. Nodes of the network are website members who identify themselves as homosexual males, and links are friendship relations defined as two nodes adding each other on their "favorite list", based on which they maintain their contacts and send messages. Only nodes and links within the giant connected component are used for this study, yielding a network of size N = 16082, and average degree $\overline{D}^* = 6.74$. Four dichotomous properties from users' profiles have been studied: *age* (born before 1980), county (live in Stockholm, *ct*), civil status (married, *cs*), and profession (employed, *pf*). The population value of group proportion (P_A^*), cross-group link probability (s_{AB}^*), homophily, and activity ratio, are listed in Table 1.

Table 1: Basic statistics for variables in the MSM network

variable	$P_{A}^{*}(\%)$	S_{AB}^*	Homophily	Activity ratio
age	77.8	0.13	0.40	1.05
ct	38.8	0.30	0.50	1.22
CS	40.4	0.57	0.05	0.97
pf	38.2	0.54	0.13	1.21

Homophily, quantified as $h_A = 1 - s_{AB}^* / P_B^*$, is the probability that nodes connect with their friends who are similar to themselves rather than randomly. If the homophily of a property is 0, it means that all nodes are connected to their friends purely randomly, regardless of this property; if the homophily is 1, it means that all nodes with a particular property are connected to friends with the same property. *Activity ratio*, is the ratio of mean degree for group A to

group B, $w = \bar{D}_A^*/\bar{D}_B^*$. Previous studies have found that homophily and activity ratio are two critical factors that may affect the performance of RDS estimators [19]. Generally, the larger the homophily or difference between a group's mean degrees, the larger will be the bias and variance of the estimates. The various levels of homophily and activity ratio of the four variables in the MSM network provides a rich test base for RDS estimators. For example, the homophily for the county is 0.50, which means that members who live in Stockholm form links with members who also live in Stockholm 50% of the time, while they form links randomly among all cities (including Stockholm) the remaining 50% of the time. The civil status has a very low level of homophily, indicating that edges are formed as if randomly among other members, regardless of their marital status.

To systematically evaluate the effect of homophily and activity ratio on the performance of RDS estimators, we have also generated a set of simulated networks with $h_A \in [0, 0.5]$ and $w \in [0.5, 2.5]$ based on the KOSKK model, which is among the best social network models that can produce most realistic network structure with respect to degree distributions, assortativity, clustering spectra, geodesic path distributions, and community structure, and the like [31, 32]. These networks are configured with population size N = 10000, average degree $\overline{D}^* = 10$, and population value $P_A^* = 30\%$ (see Appendix for details).

3.2. Study design

Based on the MSM network and artificial KOSSK networks, RDS processes are then simulated and the sample proportions and estimates are compared with population value to evaluate the accuracy of different estimators. In particular, we consider the following aspects:

Sample size: we set the sample size to 500.

Sampling without replacement (SWOR): alike most empirical RDS studies, nodes are not allowed to be recruited again if they have already been in the sample.

Number of seeds and coupons: following [19], we consider two scenarios: 6 seeds with 2 coupons, contributing to 500 respondents from 6 waves, and 10 seeds with 3 coupons, contributing to 500 respondents from 4 waves. However, we do not find significant difference for both settings in simulations and thus choose to show results with 6 seeds and 2 coupons.

Random and differential recruitment: one of the assumptions that is most unlikely to be met in real life is that participants randomly recruit peers. For example, respondents may tend to recruit people who they think will benefit most from the RDS incentives [9]. In a study of MSM in Campinas City, Brazil [25], participants were reported most often to recruit close peers or peers they believed practiced risky behaviors. In [19, 18, 16], it has been shown that all current RDS estimators would generate bias when the outcome variables are related to the tendency of such non-random distribution of coupons among respondents' personal networks (differential recruitment). To test the robustness of the new estimator, we consider both scenarios. Let $p_A^{diff} \in [0, 1]$ be the probability that

To test the robustness of the new estimator, we consider both scenarios. Let $p_A^{aiff} \in [0, 1]$ be the probability that individuals from group *A* are p_A^{diff} times more likely to be recruited by both group *A* members and group *B* members, then $p_A^{diff} = 0$ corresponds to random recruitment, when coupons are randomly distributed to respondents' friends, and $p_A^{diff} = 1$ corresponds to the extreme case scenario that both group *A* members and group *B* members are twice as likely to recruit peers of type *A*, which would largely oversample both individuals from group *A* and the proportion of recruitment links toward group *A*, s_{AA} and s_{BA} .

Reporting error about degree and ego networks: the new estimator requires respondents to report ego network information, bringing a new challenge in RDS. We simulate reporting error in two stages of a RDS process: first, when a respondent reports his or her degree, any alters of type *A* or *B* will be missed and not reported with probability p_A^{miss} or p_B^{miss} , respectively; second, when the composition of an ego network is reported, any alters of type *A* will be misclassified as type *B* with probability $p_{A\mapsto B}^{error}$, and any alters of type *B* will be misclassified as type *A* with probability $p_{B\mapsto A}^{error}$ vice versa.

RDS estimators: since previous studies have suggested that sample composition may sometimes be an even better approximation of P_A^* than traditional RDS estimators [21, 17], in addition to *RDS I* and *RDS I^{ego}*, we also include the raw sample composition in the analysis. The *RDS II* estimator in our simulations provides estimates with little difference to *RDS I* and is thus not presented separately.

Since we are interested in generating feasible population estimates by information only collected within the RDS sample, the newly developed estimators that require known population parameters [22, 23, 24] are thus beyond the purpose of this study and are excluded from comparison.

Four measurements are then carried out after the RDS simulations: the *Bias*, which is the absolute difference between the average estimate and population value, $\left|\frac{\sum_{i=1}^{m} est_i}{m} - P_A^*\right|$ or $\left|\frac{\sum_{i=1}^{m} est_i}{m} - s_{AB}^*\right|$, where est_i is the estimate from the *i*th simulation and *m* the number of simulation times; the *Standard Deviation* (SD) of estimates; the *Root Mean* Square Error (RMSE), $\sqrt{\frac{\sum_{i=1}^{m} (est_i - P_A^*)^2}{m}}$; and lastly, the *Percentage* an estimator outperforms the rest in all simulations: $P^{best} = \frac{times the estimator gives closest estimate to <math>s_{AB}^*$ or P_A^* .

All simulations were repeated 10,000 times, and seeds were excluded from the calculation of estimates in this study.

4. Results

4.1. Random and differential recruitment

4.1.1. Estimates of network link types

The difference between *RDS I* and *RDS I^{ego}* lies in the estimation of the recruitment matrix *S*. As a first step, we therefor simulate the RDS process with random recruitment $(p_A^{diff} = 0)$ and differential recruitment $(p_A^{diff} = 1)$ and then estimate the proportion of type $e_{A\to B}$ links in the population, s_{AB}^* , by both the raw sample recruitment proportion, s_{AB} , and the proposed ego-network-based estimator, \hat{s}_{AB}^{ego} , for all four variables in the MSM network, *age*, *ct*, *cs* and *pf*, respectively.

An example of the simulation results for *ct* is presented in Figure 2. Clearly, when the random recruitment assumption is fulfilled (Figure 2(a)), both s_{AB} and \hat{s}_{AB}^{ego} are unbiased. Estimates by \hat{s}_{AB}^{ego} peak more closely to s_{AB}^* and have less variance than s_{AB} (SD = 0.02 compared to 0.04, see Table 2. The difference between s_{AB} and \hat{s}_{AB}^{ego} becomes more evident when RDS is implemented with differential recruitment. We can see from Figure 2(b) that when peers who live in Stockholm are two times more likely to be recruited by their friends, the raw sample recruit proportion is largely undersampled (Bias=0.09), while \hat{s}_{AB}^{ego} still provides robust estimates (Bias=0.01) with less variance (SD=0.02). If we compare the performance of estimates for each simulation under random recruitment, \hat{s}_{AB}^{ego} is 70% times closer to s_{AB}^* than s_{AB} . Under differential recruitment, almost all \hat{s}_{AB}^{ego} estimates ($P^{best} = 0.98$) outperform s_{AB} .



Figure 2: Distribution of RDS estimates for $s_{AB}^*(ct)$. Dashed line shows population value $s_{AB}^*(ct)$. (a) Participants recruit respondents randomly among their friends, $p_A^{diff} = 0$; (b) Participants are two times more likely to recruit friends of type A than friends of type B, $p_A^{diff} = 1$.

Simulation results for estimates of all variables are summarized in Table 2. The conclusions are similar to those above. \hat{s}_{AB}^{ego} gives less bias, SD, RMSE, and gives for most instances closer estimates, regardless of homophily and activity ratios. The precision of s_{AB} depends largely on the random recruitment assumption; the bias and RMSE of s_{AB} are a maximum of 0.01 and 0.04 for all simulation settings when peers are randomly recruited, while the maximum bias and RMSE all increase to 0.13 when differential recruitment happens. \hat{s}_{AB}^{ego} , on the other side, shows great robustness to violation of this assumption. The maximum bias and RMSE for all variables are less than 0.02 and 0.03, respectively.

Regarding P^{best} , \hat{s}_{AB}^{ego} produces estimates that are closer to the true population value s_{AB}^* 62% to 74% of the time when sampling is with random recruitment; when sampling with differential recruitment, P^{best} increases to 77%~100%, revealing the superior performance of \hat{s}_{AB}^{ego} over s_{AB} .

Table 2: Statistics of estimates for s_{AB}^* by s_{AB} and s_{AB}^*						
		Bias (stand	ard deviation)	RMSE (P ^{best})		
Random recruitment		SAB	\hat{s}_{AB}^{ego}	SAB	\hat{s}^{ego}_{AB}	
sood-6	age	.00 (.03)	.00*(.03*)	.03 (.37)	.03*(.63*)	
seeu=0	ct	.01 (.04)	.00*(.02*)	.04 (.30)	.02*(.70*)	
coupon=2	CS	.00 (.04)	.00*(.02*)	.04 (.26)	.02*(.74*)	
SWOR	pf	.00 (.04)	.00*(.02*)	.04 (.26)	.02*(.74*)	
Differential	recruitme	nt				
cood-6	age	.04 (.03)	.01*(.03*)	.05 (.16)	.03*(.84*)	
seeu=0	ct	.09 (.03)	.01*(.02*)	.10 (.02)	.02*(.98*)	
SWOP	CS	.13 (.04)	.02*(.02*)	.13 (.00)	.03*(1.0*)	
SWOR	pf	.13 (.03)	.02*(.02*)	.13 (.00)	.02*(1.0*)	

^{*} corresponding statistic is better than the other estimator.

4.1.2. Estimates of population compositions

The superiority of \hat{s}_{AB}^{ego} over s_{AB} shown in the above section suggests that the *RDS I*^{ego} estimator should also give less bias and error than *RDS I*. To confirm this, we compare the simulation results of *RDS I*, and *RDS I*^{ego} to estimate population proportions on both the MSM network and the KOSKK networks.

First, we take the estimates of P_A^* for ct as an example. The result is presented as boxplots in Figure 3, where the median (middle line), the 25th and 75th percentiles (box) and outliers (whiskers) are shown. When $p_A^{diff} = 0$, there is on average an oversample of individuals who live in Stockholm (0.05) in the raw sample; however, if adjusted, *RDS I* and *RDS I^{ego}* all give unbiased estimates (Figure 3(a)). When $p_A^{diff} = 1$, i.e., respondents are twice as likely to recruit friends from Stockholm rather than friends from other counties, the improvement in estimates by RDS Iego becomes much more significant. While the sample composition/RDSI has a bias of 0.20/0.17 and RMSE of 0.21/0.18, the bias for RDS I^{ego} is only 0.02 and RMSE is 0.06 (Figure 3(b)). Another notable finding is that the number of times an estimator provides the closest estimate is almost equal between sample composition and RDSI under random recruitment ($P^{best} = 0.28$ for sample composition and $P^{best} = 0.29$ for RDSI, see Table 3), implying that even when the RDS sample is collected under ideal conditions, the traditional adjusted population estimates may perform as poorly as the raw sample proportion. $RDSI^{ego}$, by contrast, produces estimates closest to P_A^* 43% of the time. For sampling with differential recruitment, $RDSI^{ego}$ is far superior to the other estimators, with $P^{\dot{b}est} = 0.93$.



Figure 3: RDS estimates for P_{ct}^* . Dashed line is of population value P_{ct}^* . (a) Participants recruit respondents randomly among their friends, $p_A^{diff} = 0$; (b) Participants are two times more likely to recruit friends of type *A* rather than friends of type *B*, $p_A^{diff} = 1$.

The above conclusions are similar for all other variables (see Table 3): when $p_A^{diff} = 0$, both *RDS I* and *RDS I*^{ego} have little bias, while *RDS I*^{ego} generates less SD and RMSE, and provides the closest estimates than the rest estimators 10% more often. It is interesting to compare P^{best} of sample composition with the rest of the estimators; *RDS I*^{ego} always has larger P^{best} for all variables except *cs*, which has low homophily and a close to one activity ratio. *RDS I*, by contrast, cannot consistently outperform the sample composition. It has almost the same probability of providing the closest estimate to P_A^* as the sample composition for *ct*, and is even less likely to be better when estimating *age* and *cs. RDS I*^{ego} again becomes dominant when the sample composition and *RDS I* the bias and RMSE are much larger, [0.07, 0.20] and [0.09, 0.21], respectively.

		Bias (s	tandard dev	riation)		F	RMSE (P ^{best}	[†])
Random recr	ruitment	sample	RDS I	RDS I ^{ego}	_	sample	RDS I	RDS I ^{ego}
sood-6	age	.01 (.06)	.00*(.07)	.00 (.06*)		.06 (.37)	.07 (.23)	.06*(.40*)
seed=0	ct	.05 (.05)	.00 (.06)	.00*(.05*)		.07 (.28)	.06 (.29)	.05*(.43*)
SWOP	CS	.01 (.03*)	.00 (.04)	.00*(.03)		.03*(.51*)	.04 (.17)	.03 (.32)
SWOK	pf	.05 (.03*)	.00*(.04)	.00 (.03)		.05 (.20)	.04 (.32)	.03*(.48*)
Differential 1	recruitme	nt						
cood-6	age	.09 (.05*)	.08 (.06)	.02*(.07)		.10 (.10)	.10 (.12)	.07*(.79*)
seed=0	ct	.20 (.06)	.17(.07)	.02*(.06*)		.21 (.00)	.18 (.06)	.06*(.93*)
SWOR	CS	.12 (.03*)	.13 (.05)	.02*(.04)		.13 (.00)	.14 (.04)	.04*(.96*)
	pf	.18 (.03*)	.13 (.05)	.02*(.04)		.18 (.00)	.14 (.05)	.04*(.95*)

Table 3: Statistics of estimates for P_A^* by sample mean, RDS I and RDS I^{ego}

* corresponding statistic is better than other estimators.

To better understand the robustness of $RDSI^{ego}$ to differential recruitment, we simulate RDS processes on the MSM network with p_A^{diff} varying from 0 to 1. The average estimates for the four variables are shown in Figure 4. While the bias of RDSI increases progressively with p_A^{diff} , $RDSI^{ego}$ shows a clear resistance over different levels of differential recruitment. Additionally, we can see that the magnitude of bias of RDSI does not depend solely on either the homophily or activity ratio, implying that, without the collection of ego network information, more sophisticated modifications are needed for RDSI to adapt differential recruitment.

The complexity of joint effect of homophily and activity ratio is more evident for RDS estimates on the KOSKK networks, as shown in Figure 5, where the biases of both *RDS I* and *RDS I*^{ego} are shown for networks with different levels of homophily ($h_A \in [0, 0.5]$) and activity ratio $w \in [0.5, 2.5]$.

Generally, bias increases with homophily and difference between average degrees, however, these effects are mixed with impact of other network structural properties, for example the community structure resulted by the KOSKK model, making networks with certain combinations of h_A and w least biased. *RDS I^{ego}* shows resistance over all these structural effects: when $p_A^{diff} = 0$, the bias for *RDS I* ranges from 0.00 to 0.06, while for *RDS I^{ego}*, this range is only [0.00, 0.01]; when $p_A^{diff} = 1$, the maximum bias for *RDS I* goes up to 0.20, while the maximum bias for *RDS I^{ego}* stays around 0.02.

4.2. Degree reporting error

With the superior performance observed from the above section, we will from this section focus on *RDS I^{ego}* and evaluate factors that may bring extra sources of biases.

The degree reporting error parameters p_A^{miss} and p_B^{miss} , capture the fact that in social network surveys, especially surveys targeting hidden populations, individuals in the target population may not be identified by their friends and would thus be miscounted when a respondent reports the personal network size [33, 23]. This reporting error will not only affect the estimates of average degree, but further bias the estimate of the recruitment matrix in $RDSI^{ego}$, $S_{XY}^{ego}(X, Y \in A, B)$.

We simulate RDS with degree reporting error $p_A^{miss} \in [0, 0.2]$ and $p_B^{miss} \in [0, 0.2]$, that is, a maximum of 20% friends with property A or B may be unidentified as the target population. To account for the absolute worst case scenario, differential recruitment ($p_A^{diff} = 1$) is also included in the simulation. Results are presented in Figure 6 for the MSM network and Figure 7 for KOSKK networks.



Figure 4: Average estimates of RDSI and RDSI^{ego} on the MSM network with varying level of differential recruitment.



Figure 5: Bias of RDS1 and RDS1^{ego} on KOSKK networks with random recruitment (a, c) and differential recruitment (b, d).



Figure 6: Bias of RDS Iego on the MSM network with differential recruitment and degree reporting error.

Surprisingly, on both the MSM network and KOSKK networks, even with 20% of all alters being miscounted, the biases of *RDS I^{ego}* range mostly within [0.00, 0.05] with a few exceptions. The worst case scenario occurs when 20% of all alters from one group are missed in the reported degree, while none from the other group is missed, with the maximum bias around 0.07. When miscounted alters are less than 10%, most configurations of $[p_A^{miss}, p_B^{miss}]$ produce biases less than 0.04.

We can also see a symmetric effect of p_A^{miss} and p_B^{miss} , the bias maintains on the same level as long as the two parameters change in the same direction. This effect was previously examined in [18], where the degree reporting error was modeled as unawareness of existing relationships. These findings implies that the magnitude of bias resulted by degree reporting error is much less than the error itself, since the increase of reporting error on one group can "compensate" reporting error on the other group; tolerable bias would be expected when the reporting error is limited.

It is worth noting that the biases analyzed here are outcomes of RDS simulations with "extreme" differential recruitment. We have also ran simulations with random recruitment ($p_A^{diff} = 0$), which generate similar patterns (e.g., the symmetric effect, where the maximum bias occurs) with smaller biases, see Appendix Figure 13 and Figure 14.

4.3. Ego network reporting error

Another reporting error related to the implementation of $RDSI^{ego}$, is that even when individuals fulfilling the sample inclusion criteria are correctly identified, their characteristics, especially for sensitive variables such as HIV status and sexual preference, may be incorrectly reported by their friends. By varying $p_{A \rightarrow B}^{error}$ and $p_{B \rightarrow A}^{error}$ from 0, when the composition of ego networks are accurately reported, to 0.2, when 20% of the alters are misclassified, we run simulations on the MSM networks and KOSKK networks, to evaluate the sensitivity of $RDSI^{ego}$ to the reporting error in ego network compositions. Similar to the previous section, we use differential recruitment and set $p_A^{diff} = 1$. Results are shown in Figure 8 and Figure 9.

Contrary to the robustness to degree reporting error, the *RDS* I^{ego} estimator is much more sensitive to the ego network reporting error on both the MSM network and KOSKK networks. On the MSM network, the bias readily excesses 0.1 as long as $p_A^{diff} > 0.1$ and $p_B^{diff} < 0.1$ for *age*, and $p_A^{diff} < 0.1$ and $p_B^{diff} > 0.1$ for *ct*. The biases for the



Figure 7: Bias of RDS I^{ego} on KOSKK network with differential recruitment and degree reporting error.

other two variables with less homophily are relatively smaller, as long as the misclassification error for alters of both groups is less than 10%.

Given $p_{A\mapsto B}^{error} \in [0, 0.2]$ and $p_{B\mapsto A}^{error} \in [0, 0.2]$, the ego network reporting error on KOSKK networks produces much larger bias for networks with low activity ratios ($w \le 1$). And the increase of $p_{B\mapsto A}^{error}$ is apparently more harmful than the increase of $p_{A\mapsto B}^{error}$. This effect is due to the fact that when $w \le 1$, a large amount of alters for respondents in the RDS sample are from group *B* (note also $P_B^* = 0.7$), a small probability of misclassifying *B* alters as *A* alters will result in a large absolute number of over-reported *A* alters in the end, making *RDS I^{ego}* generate estimates much higher than the true population value P_A^* . For this reason, variables with high activity ratios, on the other hand, are less sensitive to the network reporting error.

The above reasoning can also be verified with estimates for *age* on the MSM network, which has a relatively balanced activity ratio (w = 1.05), but a population proportion of 70%. Therefore, reporting error regarding the group with higher population proportion and activity ratio will result in substantial amount of misclassified alters in the ego networks and greatly affect the estimates.

Simulations with random recruitment have also been carried out, however the ego network reporting error seems to be the dominant factor driving estimate error for *RDS*^{ego}, no significant reduce of bias is observed, see Appendix Figure 15 and Figure 16.

5. Conclusion and discussion

Ego network data has been collected for decades and exists largely in sociological surveys [28, 34, 35, 36, 37, 38]; the RDS sampling mechanism further makes it possible to collect "*linked-ego network*" data. By combining RDS recruitment trees with ego networks, this study developed a new estimator, *RDS I^{ego}*, for RDS studies. Given that participants can accurately report the composition of their personal networks, this estimator has superior performance over traditional RDS estimators. Most importantly, *RDS I^{ego}* shows strong robustness to differential recruitment, a



Figure 8: Bias of RDS I^{ego} on the MSM network with differential recruitment and ego network reporting error.

violation of the RDS assumptions that may cause large bias and estimation error and is not under the control of the researchers. Evaluation studies on our simulated KOSKK networks also show that *RDS I^{ego}* performs consistently well on networks with varying homophily, activity ratio, and community structures.

The limitation of *RDS I^{ego}* is rooted in the need to collect ego network data. Many RDS studies are designed for use among hidden populations, who may be reluctant to share certain private information with their friends. Consequently, the proposed method is primarily suited for less sensitive variables, which the respondent can be expected to know about his contacts. Such information may for example include socio-demographic variables (e.g., gender, age groups, profession, marital status, etc.) for which survey methods on how to design and collect ego network data has been extensively studied [39, 40, 41, 42]. Additionally, certain variables, e.g. drug use, may be highly sensitive in the general population but may not be at all be so in an IDU population.

By modeling the difficulty in understanding of personal network composition as degree reporting error and ego network reporting error, which quantify the level of mutual knowledge about studied variables shared with friends, we have showed that even with 20% of alters being unidentified, *RDS I^{ego}* was still able to produce estimates with bias less than 0.05 most of the time. On the other hand, *RDS I^{ego}* is sensitive to the error of misclassifying alters. If 20% of alters from one group is mistakenly reported as belonging to the other group, estimate bias can exceed 0.1 when the probability of misclassifying members of one group is substantially larger than misclassification of members in the other group (e.g., $p_{A\mapsto B}^{error} \gg p_{B\mapsto A}^{error}$). Fortunately, the result shows that when the studied variables only related to a small proportion of alters, that is, if P_A^* is low and *w* is relatively small, the increase of error in misclassifying *A* as *B* members will have a small influence on the bias. Consequently, for many sensitive variables surveyed in RDS studies, if the reporting error of a low prevalence trait (e.g., HIV status) is mainly "false negatives", e.g., alters with HIV are reported as healthy friends since they are reluctant to reveal this information to their egos, estimates with small bias are still expected to be able to achieve.

There are other interesting findings from this study. First, the performance of *RDS I*, which has been used in most RDS studies so far, fails to outperform the sample composition in many simulation settings. Second, we propose in this paper a new bootstrap method for constructing confidence intervals (CIs) with *RDS I^{ego}* (see Appendix). Simulations in this paper and recent studies [21, 17, 43], has shown that the traditional bootstrapping method underestimates



Figure 9: Bias of RDS I^{ego} on KOSKK network with differential recruitment and ego network reporting error.

variance. However, the proposed bootstrap method in this paper is able to generate CIs that much better approximate the expected coverage rates and performs fairly consistent to variations of homophily, activity ratio and community structures of networks.

In summary, we have shown that, by combining the traditional RDS sampling design with collection of ego network data, population estimates can improve drastically. What's most important, since RDS is a chain-referral designed sampling strategy, once the sample is started from seeds, the distribution of coupons is largely out of the control of researchers, and non-random recruitment often occurs, which has been proved to generate large estimate bias and error [19, 18, 16, 44]. The robustness of $RDS I^{ego}$ to differential recruitment offers researchers the ability to largely reduce estimate error. Additionally, by comparing \hat{S}^{ego} with the observed raw sample recruitment matrix *S*, the severity of differential recruitment may be assessed. For future RDS studies, we encourage ego network questions to be integrated with traditional RDS questionnaires along with the improved bootstrap procedure. Due to the limitations inherent in the collection of sensitive variables from stigmatized group, the new method may be better suited to less sensitive variables. This new method is also applicable to sampling problems in other fields [15, 45, 46], such as sampling of internet contents from which the ego network data is more reliable and may be more efficiently retrieved.

Acknowledgements

The author would like to thank Professor Fredrik Liljeros and Dr. Linus Bengtsson for helpful discussions. This work has been funded by China Scholarship Council (Grant No. 2008611091) and Riksbankens Jubileumsfond (The Bank of Sweden Tercentenary Foundation).

Appendix

Appendix A: Generation process for KOSKK networks

As one of the dynamical network evolution models, the KOSKK model utilizes network link weights to generate networks with key common feathers of social networks [32]: (i) skewed degree distribution, (ii) assortative mixing, (iii) high average clustering coefficient, (iv) small average shortest path lengths, and (v) community structures. In a comprehensive comparative study [31], the KOSKK model was found to be one of the best social network models that can generate similar-to-real social network structures, among nodal attribute models, network evolution models as well as ERGM models.

In a KOSKK model, the network is initiated with *N* nodes and zero edges, and then evolved with three mechanisms:

(i) *Local attachment*. Select a node *i* randomly, and choose one of *i*'s neighbor *j* with probability $w_{ij}/\sum_j w_{ij}$, where w_{ij} is the weight on link e_{ij} . If *j* has another neighbor apart from *i*, choose one of them (node *k*) with probability $w_{jk}/\sum_k (w_{jk} - w_{ij})$. If there is no link between *i* and *k*, connect *k* to *i* with probability p_{Δ} and set $w_{ki} = w_0$. Increase link weight w_{ij}, w_{jk} , and w_{ki} (if was already present) by δ .

(ii) *Global attachment*. Connect *i* to a random node *l* with probability p_r (or with probability 1 if *i* has no connections) and set $w_{il} = w_0$.

(iii) Node deletion. Select a random node and with probability p_d remove all of its connections.

With larger δ , clearer community structures will be generated, as new links are created preferably through strong links. When p_d is fixed, the average degree is obtained by adjusting p_{Δ} for each δ . In our simulation, we set N = 10000, $w_0 = 1$, $p_r = 0.0005$, $p_d = 0.001$, $\delta = 0.6$, and the network average degree $\overline{D}^* = 10$. The process was ran 10^8 time steps to achieve stationary network characteristics. At the end of the process, a few nodes will be isolated due to the *node deletion* step, we simply randomly link these nodes to the giant connected component to make sure all nodes in the network are connected. As δ is relatively large, the obtained network shows a clear community structure, see Figure 10.

Based on the above network, we then start the configuration of homophily and activity ratio. Let *w* be the activity ratio of the current network and w^* be the activity ratio we want to obtain. At the beginning, 30% of the nodes are randomly selected and assigned with property *A*, the rest of nodes are then assigned with property *B*. If $w > w^*$, we randomly pick a node with property *A*, *i*, and a node with property *B*, *j*, if $d_i > d_j$, we then exchange the properties of the two nodes, i.e., *i* becomes a *B* node, and *j* becomes a *A* node. If $w < w^*$, we exchange the properties of *i*, *j* only when $d_i < d_j$. The above process is repeated until $w = w^*$.

For each of the network configured with w^* , we use a rewiring process to adjust the homophily. Recall that the homophily is depended on the number of cross group links as $h_A = 1 - s_{AB}^* / P_B^*$, smaller s_{AB} indicates high homophily. Let h_A be the homophily of the current network and h_A^* be the desired value, if $h_A > h_A^*$, we randomly pick two within group links $i \leftrightarrow j$, $k \leftrightarrow l$, with i, j belonging to group A, and k, l belonging to group B, and rewire them to $i \leftrightarrow k$, $j \leftrightarrow l$, to increase cross group links. Similarly, if $h_A < h_A^*$, we randomly pick two cross group links and rewire them to form two within group links. The above process is repeated until $h_A = h_A^*$.



Figure 10: Visualization of the KOSKK network generated with $\delta = 0.6$, $\bar{D}^* = 10$.

Appendix B: Confidence interval estimation

The precision of a sample estimate is usually enhanced by providing a confidence interval (CI), which gives a range within which the true population is expected to be found with some level of certainty. Due to the complex sample design of RDS, simple random sampling based CIs are generally narrower than expected [17, 11, 20]. Consequently, bootstrap methods are used to construct CIs around RDS estimates.

The current widely used bootstrap procedure for RDS (*BS-origin*) was proposed by Salganik [20, 47]. In this procedure, respondents are divided into two groups depending on the property of their recruiters, that is, those who are recruited by *A* nodes (A_{rec}), and those who are recruited by *B* nodes (B_{rec}). Then the bootstrap starts by a randomly chosen respondent. If the respondent has property *A*, then the next respondent is randomly picked from A_{rec} , otherwise from B_{rec} . Such a procedure is repeated with replacement until the original RDS sample size is reached, then the RDS estimate is calculated based on the replicated sample. When *R*-replicated samples are bootstrapped, the resulting middle 90%/95% estimates from the ordered *R* estimates are then used as the estimated CI.

We extend the *BS*-origin in two different ways:

(a) *BS*-*ego*1: we implement the same resampling procedure as with *BS*-*origin*; however, when each replicated sample is collected, *RDS I*^{*ego*} is used to calculate the RDS estimate, rather than *RDS I*;

(b) *BS*-ego2: we divide the sample into two groups depending on the property of the respondents, that is, those with property *A* (A_{set}) and those with property *B* (B_{set}). Then the bootstrap procedure is started with a randomly picked respondent. If the respondent has property *A*, then the probability of selecting the next respondent from A_{set} or B_{set} , is $1 - \hat{s}_{AB}^{ego}$ and \hat{s}_{AB}^{ego} , respectively. If the respondent has property *B*, then the probability of selecting the next respondent from A_{set} or B_{set} , is \hat{s}_{AB}^{ego} , and \hat{s}_{AB}^{ego} , respectively. The above process is repeated until the same size as original sample is reached. *RDS I^{ego}* is then used to calculate the RDS estimate for each replicated sample.

We expect that the modification in the bootstrap procedure of *BS-ego2* by introducing the ego network data based estimate \hat{s}_{AB}^{ego} and \hat{s}_{BA}^{ego} can improve the performance of estimated CIs when the RDS is done with differential recruitment.

Following [20], we use simulations on both the MSM network and KOSKK networks to compare the performance of *BS-origin*, *BS-ego1*, and *BS-ego2*. For each variable, 1000 RDS samples are collected, and for each of these 1000 samples we construct the 90% and 95% CIs based on 1000 replicate samples drawn by the above bootstrap procedures. The proportion of times the generated confidence interval contains the true population value P_A^* when sampling with random recruitment and differential recruitment (denoted as Φ_{RR}^{90} , Φ_{DR}^{90} , and Φ_{RR}^{95} , Φ_{DR}^{95}) is compared with different bootstrap methods and are presented in Figure 11 and Figure 12.

On the MSM network, when sampling with random recruitment, we can see from Figure 11(a), (b) that all three methods produce similar coverage rates for the tested variables. The coverage rate for *age* is significantly smaller than the desired value for both Φ_{RR}^{90} and Φ_{RR}^{95} , indicating that even under ideal conditions, the bootstrap-based CIs in RDS may be much narrower than expected. When the RDS is done with differential recruitment (Figure 11(c), (d)), the coverage rate of *BS*-origin becomes extremely small and practically useless. This is because the *RDS I* estimates are largely biased from the true population value when differential recruitment exists. The coverage rates of *BS*-ego1 and *BS*-ego2, on the other hand, are well above 50% for all the four variables and therefore outperform *BS*-origin in an absolute sense. In general, there is 5%~10% more coverage in Φ_{DR}^{90} and Φ_{DR}^{95} for *BS*-ego2 compared to *BS*-ego1, implying that the modified bootstrap procedure is more resistant to the violation of the random recruitment assumption in RDS.

BS-origin performs poorly on KOSKK networks for both sampling with random recruitment and sampling with differential recruitment, with a majority of 95% coverage rates under 50%. The *RDS I^{ego}*-based bootstrap methods, all produce coverage rates 20%~60% higher than *BS*-origin. When $p_A^{diff} = 0$, there is no significant difference between *BS*-ego1 and *BS*-ego2, however, when $p_A^{diff} = 1$, *BS*-ego2 is able to produce 8%~14% higher coverage rates than *BS*-ego1 in extreme cases (w = 0.5).

It is worth noting that, even *BS-ego2* shows superior performance over *BS-origin* and is robustness to variations in network structure properties evaluated in this study (e.g., homophily, activity ratio, and the like.), the bootstrapped CIs rarely approach required coverage rates. On KOSKK networks, it is common that the 95% coverage rates are 5%~20% lower than expected. Even the community structure in these networks may impede the performance of RDS estimates as well as the bootstrap methods, future work is needed to develop CI estimate methods with improved precision.



Figure 11: Coverage rate of 90% and 95% confidence interval, by bootstrap procedure BS-origin, BS-ego1, and BS-ego2 on the MSM network.

0.5 0.8 1.0 1.5 2.0	2.5
0.5 .63 (.85) [.87] .24 (.86) [.84] .18 (.80) [.79] .40 (.79) [.83] .69 (.82) [.81]	.65 (.73) [.74]
0.4 .67 (.87) [.88] .24 (.82) [.83] .20 (.81) [.78] .37 (.77) [.82] .64 (.76) [.78]	.46 (.68) [.68]
0.3 .65 (.85) [.85] .26 (.85) [.84] .21 (.82) [.77] .44 (.79) [.83] .51 (.71) [.73]	.34 (.69) [.74]
6 0.2 .64 (.86) [.86] .28 (.84) [.81] .23 (.85) [.81] .48 (.77) [.81] .47 (.75) [.80]	.47 (.81) [.86]
59 (.82) [.84] [.81] .28 (.83) [.81] .45 (.74) [.79] .59 (.82) [.84]	.65 (.88) [.90]
0.0 .65 (.82) [.83] .39 (.85) [.82] .33 (.85) [.83] .45 (.81) [.86] .65 (.86) [.88]	.79 (.88) [.91]

Φ^{95}_{RR}
itit

		activity ratio, \overline{D}_{A}^{*} / \overline{D}_{B}^{*}					
		0.5	0.8	1.0	1.5	2.0	2.5
ł	0.5	.19 (.73) [.87]	.16 (.74) [.82]	.16 (.81) [.78]	.27 (.76) [.81]	.32 (.76) [.75]	.48 (.70) [.72]
Ч.	0.4	.17 (.75) [.87]	.18 (.78) [.82]	.19 (.77) [.76]	.30 (.80) [.81]	.43 (.77) [.77]	.73 (.68) [.68]
hily	0.3	.18 (.74) [.83]	.20 (.80) [.83]	.19 (.80) [.78]	.32 (.79) [.80]	.65 (.73) [.72]	.79 (.74) [.73]
idoi	0.2	.18 (.74) [.85]	.22 (.78) [.83]	.22 (.79) [.79]	.44 (.79) [.80]	.69 (.77) [.76]	.57 (.81) [.83]
non	0.1	.20 (.74) [.83]	.28 (.77) [.78]	.25 (.80) [.79]	.53 (.79) [.79]	.52 (.84) [.83]	.45 (.87) [.88]
	0.0	.21 (.71) [.79]	.27 (.78) [.78]	.38 (.80) [.80]	.35 (.85) [.84]	.39 (.87) [.86]	.37 (.89) [.90]
		(b) Φ_{DR}^{95}					

(b) (\mathbb{D}_1^2
-------	------------------

Figure 12: Coverage rate of 95% confidence interval, by bootstrap procedure *BS-origin*, (*BS-ego1*), and [*BS-ego2*] on KOSKK networks. (a) Random recruitment, $p_A^{diff} = 0$; (b) Differential recruitment, $p_A^{diff} = 1$.

Appendix C: Supporting figures



Figure 13: Bias of RDS I^{ego} on the MSM network with random recruitment and degree reporting error.



Figure 14: Bias of RDS Iego on KOSKK network with random recruitment and degree reporting error.



Figure 15: Bias of RDS I^{ego} on the MSM network with random recruitment and ego network reporting error.



Figure 16: Bias of RDS I^{ego} on KOSKK network with random recruitment and ego network reporting error.

References

- [1] Deaux E, Callaghan JW. Key Informant Versus Self-Report Estimates of Health-Risk Behavior. Evaluation Review. 1985;9(3):365-368.
- [2] Watters JK, Biernacki P. Targeted Sampling: Options for the Study of Hidden Populations. Social Problems. 1989;36(4):416–430.
- [3] Erickson BH. Some problems of inference from chain data. Sociological Methodology. 1979;10:276–302.
- [4] Heckathorn DD. Respondent-driven sampling: A new approach to the study of hidden populations. Social Problems. 1997;44(2):174–199.
- [5] Magnani R, Sabin K, Saidel T, Heckathorn D. Review of sampling hard-to-reach and hidden populations for HIV surveillance. AIDS. 2005;19 Suppl 2:S67–72.
- [6] Johnston LG, Malekinejad M, Kendall C, Iuppa IM, Rutherford GW. Implementation challenges to using respondent-driven sampling methodology for HIV biological and behavioral surveillance: Field experiences in international settings. Aids and Behavior. 2008;12(4):S131–S141.
- [7] Wejnert C. An Empirical Test of Respondent-Driven Sampling: Point Estimates, Variance, Degree Measures, and out-of-Equilibrium Data. Sociological Methodology 2009, Vol 39. 2009;39:73–116.
- [8] Lansky A, Abdul-Quader AS, Cribbin M, Hall T, Finlayson TJ, Garfein RS, et al. Developing an HIV Behavioral Surveillance System for injecting drug users: The National HIV Behavioral Surveillance System. Public Health Reports. 2007;122:48–55.
- [9] Kogan SM, Wejnert C, Chen YF, Brody GH, Slater LM. Respondent-Driven Sampling With Hard-to-Reach Emerging Adults: An Introduction and Case Study With Rural African Americans. Journal of Adolescent Research. 2011;26(1):30–60.
- [10] Wejnert C, Heckathorn DD. Web-based network sampling Efficiency and efficacy of respondent-driven sampling for online research. Sociological Methods & Research. 2008;37(1):105–134.
- [11] Heckathorn DD. Respondent-driven sampling II: Deriving valid population estimates from chain-referral samples of hidden populations. Social Problems. 2002;49(1):11–34.
- [12] Salganik MJ, Heckathorn DD. In: Sampling and estimation in hidden populations using respondent-driven sampling. vol. 34 of Sociological Methodology. Weinheim: Wiley-V C H Verlag Gmbh; 2004. p. 193–239.
- [13] Heckathorn DD. Extensions of Respondent-Driven Sampling: Analyzing Continuous Variables and Controlling for Differential Recruitment. Sociological Methodology 2007, Vol 37. 2007;37:151–208.
- [14] Volz E, Heckathorn DD. Probability Based Estimation Theory for Respondent Driven Sampling. Journal of Official Statistics. 2008;24(1):79– 97.
- [15] Gjoka M, Kurant M, Butts CT, Markopoulou A. Walking in Facebook: A Case Study of Unbiased Sampling of OSNs. In: INFOCOM, 2010 Proceedings IEEE; 2010. p. 1–9.
- [16] Tomas A, Gile KJ. The effect of differential recruitment, non-response and non-recruitment on estimators for respondent-driven sampling. Electronic Journal of Statistics. 2011;5:899–934.
- [17] Goel S, Salganik MJ. Assessing respondent-driven sampling. Proceedings of the National Academy of Sciences of the United States of America. 2010;107(15):6743–6747.
- [18] Lu X, Bengtsson L, Britton T, Camitz M, Kim BJ, Thorson A, et al. The sensitivity of respondent-driven sampling. Journal of the Royal Statistical Society: Series A (Statistics in Society). 2012;175(1):191–216.
- [19] Gile KJ, Handcock MS. Respondent-Driven Sampling: An Assessment of Current Methodology. Sociological Methodology. 2010;40:285– 327.
- [20] Salganik MJ. Variance estimation, design effects, and sample size calculations for respondent-driven sampling. Journal of Urban Health-Bulletin of the New York Academy of Medicine. 2006;83(6):198–1112.
- [21] McCreesh N, Frost SDW, Seeley J, Katongole J, Tarsh MN, Ndunguse R, et al. Evaluation of Respondent-driven Sampling. Epidemiology. 2012;23(1):138–147.
- [22] Gile KJ. Improved Inference for Respondent-Driven Sampling Data With Application to HIV Prevalence Estimation. Journal of the American Statistical Association. 2011;106(493):135–146.
- [23] Lu X, Malmros J, Liljeros F, Britton T. Respondent-driven Sampling on Directed Networks. arXiv:12011927v1. 2012;.
- [24] Gile KJ, Handcock MS. Network Model-Assisted Inference from Respondent-Driven Sampling Data. arXiv:11080298v1. 2011;.
- [25] de Mello M, de Araujo Pinho A, Chinaglia M, Tun W, Júnior AB, Ilário MCFJ, et al. Assessment of risk factors for HIV infection among men who have sex with men in the Metropolitan Area Of Campinas City, Brazil, Using Respondent-Driven Sampling. Population Council; 2008.
- [26] Li J, Liu HJ, Li JH, Luo J, Koram N, Detels R. Sexual transmissibility of HIV among opiate users with concurrent sexual partnerships: an egocentric network study in Yunnan, China. Addiction. 2011;106(10):1780–1787.
- [27] Rudolph AE, Latkin C, Crawford ND, Jones KC, Fuller CM. Does Respondent Driven Sampling Alter the Social Network Composition and Health-Seeking Behaviors of Illicit Drug Users Followed Prospectively? Plos One. 2011;6(5).
- [28] Britton T, Trapman P. Inferring global network properties from egocentric data with applications to epidemics. arXiv:12012788v1. 2012;.
- [29] Hansen MH, Hurwitz WN. On the Theory of Sampling from Finite Populations. The Annals of Mathematical Statistics. 1943;14(4):333–362.
- [30] Rybski D, Buldyrev SV, Havlin S, Liljeros F, Makse HA. Scaling laws of human interaction activity. Proceedings of the National Academy of Sciences of the United States of America. 2009;106(31):12640–12645.
- [31] Toivonen R, Kovanen L, Kivel? M, Onnela JP, Saram?ki J, Kaski K. A comparative study of social network models: Network evolution models and nodal attribute models. Social Networks. 2009;31(4):240–254.
- [32] Kumpula JM, Onnela JP, Saram?ki J, Kaski K, Kertsz J. Emergence of communities in weighted networks. Physical review letters. 2007;99(22):228701.
- [33] Salganik MJ, Mello MB, Abdo AH, Bertoni N, Fazito D, Bastos FI. The game of contacts: estimating the social visibility of groups. Social Networks. 2011;33(1):70–78.
- [34] Handcock MS, Gile KJ. Modeling Social Networks from Sampled Data. Annals of Applied Statistics. 2010;4(1):5–25.
- [35] Newman MEJ. Ego-centered networks and the ripple effect. Social Networks. 2003;25(1):83–95.
- [36] Mizruchi MS, Marquis C. Egocentric, sociocentric, or dyadic? Identifying the appropriate level of analysis in the study of organizational networks. Social Networks. 2006;28(3):187–208.

- [37] Marsden PV. Egocentric and sociocentric measures of network centrality. Social Networks. 2002;24(4):407-422.
- [38] Hanneman RA, Riddle M. Introduction to social network methods. Riverside, CA: University of California, Riverside; 2005.
- [39] Kogovšek T, Ferligoj A. The quality of measurement of personal support subnetworks. Quality & Quantity. 2005;38(5):517-532.
- [40] Matzat U, Snijders C. Does the online collection of ego-centered network data reduce data quality? An experimental comparison. Social Networks. 2010;32(2):105–111.
- [41] Burt RS. Network items and the general social survey. Social Networks. 1984;6(4):293-339.
- [42] Marin A. Are respondents more likely to list alters with certain characteristics? Implications for name generator data. Social Networks. 2004;26(4):289–307.
- [43] Salganik MJ. Commentary: Respondent-driven Sampling in the Real World. Epidemiology. 2012;23(1):148-50.
- [44] Bengtsson L, Thorson A. Global HIV surveillance among MSM: is risk behavior seriously underestimated? AIDS. 2010;24(15):2301–2303.
- [45] Lee SH, Kim PJ, Jeong H. Statistical properties of sampled networks. Physical Review E. 2006;73(1):016102.
- [46] Yoon S, Lee S, Yook SH, Kim Y. Statistical properties of sampled networks by random walks. Physical Review E. 2007;75(4):046114.
- [47] Volz E, Wejnert C, Degani I, Heckathorn DD. Respondent-Driven Sampling Analysis Tool (RDSAT) Version 5.6. Cornell University; 2007.