Social Networks 35 (2013) 669–685

Contents lists available at ScienceDirect

Social Networks

journal homepage: www.elsevier.com/locate/socnet

Linked Ego Networks: Improving estimate reliability and validity with respondent-driven sampling

Xin Lu [a,b,c,*]

[a] College of Information System and Management, National University of Defense Technology, Changsha, China
[b] Department of Sociology, Stockholm University, Stockholm, Sweden
[c] Department of Public Health Sciences, Karolinska Institutet, Stockholm, Sweden

ARTICLE INFO

Keywords:
Ego networks
Respondent-driven sampling
HIV
Reporting error
Differential recruitment

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 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 precision of estimates for population characteristics is greatly improved. The proposed estimator shows superior advantages over traditional RDS estimators, and most importantly, the method exhibits strong robustness to the recruitment preference of respondents and degree reporting error, which commonly happen in RDS practice and may generate large estimate biases and errors for traditional RDS estimators. 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.

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 such situations, such as key informant sampling (Deaux and Callaghan, 1985), targeted/location sampling (Watters and Biernacki, 1989), and snowball sampling (Erickson, 1979). However, these methods all introduce a considerable selection bias, which impairs generalization of the findings from the sample to the studied population (Heckathorn, 1997; Magnani et al., 2005). 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). RDS uses a link-tracing network sampling design and provides, given fulfillment of a limited number of assumptions, asymptotically unbiased population estimates as well as a feasible implementation, making it the state-of-the-art sampling method for studying *hard-to-access* populations (Johnston et al., 2008; Wejnert, 2009; Lansky et al., 2007; Kogan et al., 2011; Wejnert and Heckathorn, 2008).

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 (Heckathorn, 1997). 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 (Heckathorn, 2002; Salganik and Heckathorn, 2004).

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 (Salganik and Heckathorn, 2004; Heckathorn, 2007; Volz and Heckathorn, 2008). Following this, the probability for each node to be included in the RDS sample is proportional to its degree. Specifically, for

* Correspondence to: College of Information System and Management, National University of Defense Technology, 410073 Changsha, China. Tel.: +46 0739606393.
E-mail addresses: xin.lu.84@gmail.com, xin.lu@ki.se

0378-8733/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.socnet.2013.10.001

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 Salganik and Heckathorn (2004) and Volz and Heckathorn (2008):

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

or

$$\hat{P}_A^{RDSII} = \frac{\sum_{v_i \in A \cap U} d_i^{-1}}{\sum_{v_i \in U} d_i^{-1}}, \tag{2}$$

where $\hat{\bar{D}}_A = n_A/(\sum_{v_i \in A \bigcap U} d_i^{-1})$ and $\hat{\bar{D}}_B = n_B/(\sum_{v_i \in B \cap U} d_i^{-1})$ are the estimated average degrees for individuals of group A and B in the population. Both estimators give asymptotically unbiased estimates when the above assumptions are fulfilled (Salganik and Heckathorn, 2004; Volz and Heckathorn, 2008).

The methodology of RDS is nicely designed; however, the assumptions underlying the RDS estimators are rarely met in practice (Wejnert, 2009; Tomas and Gile, 2011; Goel and Salganik, 2010; Bengtsson et al., 2012). 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. (2012), 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 simulating RDS process on an empirical MSM network as well as artificial networks and comparing RDS estimates 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 non-reciprocal relationships. When the sample size is relatively large (>50% of the population), similar results were also found by Gile and Handcock (2010), 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 (Salganik, 2006). In a study by Goel and Salganik (2010) 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 (Salganik, 2006). Moreover, McCreesh et al. (2012) 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 (2011) 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. (2013) proposed new estimators for RDS on directed networks, with known in degree 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 *RDSI*, *RDSII* estimators, and the estimators newly developed by Gile (2011), Gile and Handcock (2011) and Lu et al. (2013) utilize the same information collected by standard RDS practice, that is, the recruitment matrix S, and the degree and studied properties of each respondent in the sample. 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. (2008), 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, information about supporting, drug using, and sexual behaviors between respondents and their network members was collected (Li et al., 2011). One of the most thorough RDS studies utilizing ego network information was done by Rudolph et al. (2011), 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 (Britton and Trapman, 2012; Everett and Borgatti, 2005) in the way that each "ego" is connected with (recruited by) its recruiter. For example, in a partial chain of RDS as illustrated in Fig. 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 $U = \{v_1, v_2, \ldots, v_n\}$, 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)–(iv) fulfilled, the probability that each node will be included in the sample, $Pr(v_i)$, will be proportional to its degree (Salganik and Heckathorn, 2004; Heckathorn, 2007; Volz and Heckathorn, 2008):

$$Pr(v_i) \propto \frac{d_i}{\sum_{k=1}^{N} d_k}, \tag{3}$$

where N is the size of the population of interest.

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

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

that is, each link has the same probability of being selected via the RDS process. Consequently, the observed recruitment matrix S is a

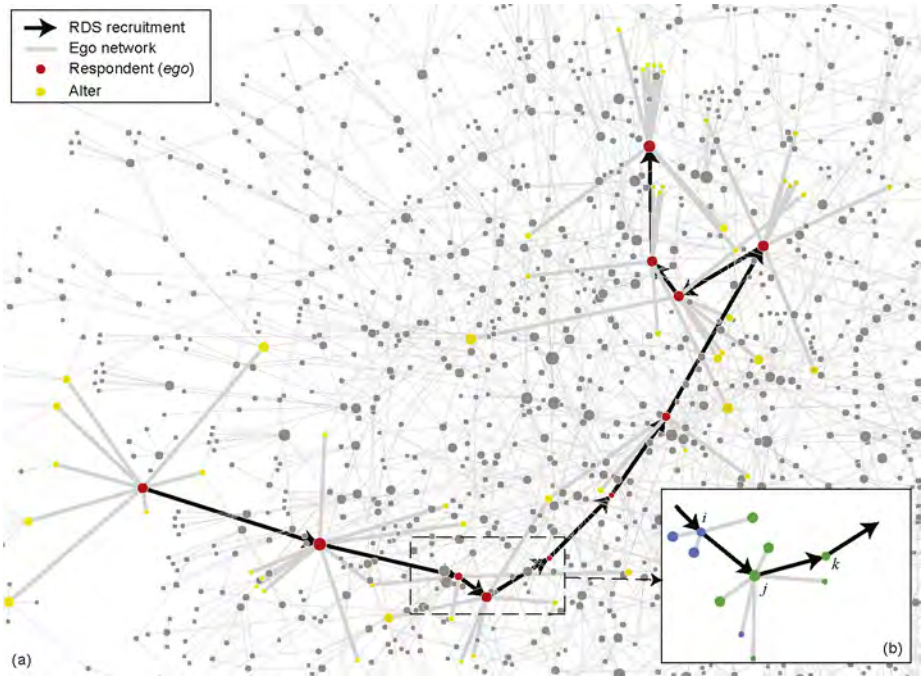

Fig. 1. A RDS chain with egocentric data. (a) RDS on a network. Red nodes are those that participated in the RDS survey, and yellow nodes are ego network composition inferred by participants; (b) a partial RDS chain with color representing properties of nodes. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

random sample for the cross-group links of the network (Salganik and Heckathorn, 2004).

The above are general inferences from a typical RDS process. Up to now, we can turn our attention to the egocentric data source. Let $Pr(e^{ego}_{i\to j})$ 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:

$$Pr(e^{ego}_{i\to j}) = Pr(v_i) \propto \frac{d_i}{\sum_{k=1}^{N} d_k eq2}. \tag{5}$$

Consequently, to estimate the proportion of type $e_{X\to Y}(X, Y \in \{A, B\})$ links in the population, s^*_{XY}, we can weight the observed number of type $e^{ego}_{X\to Y}$ links by their inclusion probability to construct a *generalized Hansen–Hurwitz estimator* (Hansen and Hurwitz, 1943):

$$\hat{s}^{ego}_{XY} = \frac{\hat{N}^{ego}_{XY}}{\hat{N}^{ego}_{XA} + \hat{N}^{ego}_{XB}} = \frac{\sum_{v_i\in X\bigcap U} n^Y_i/d_i}{\sum_{v_j\in X\bigcap U} n^A_j/d_j + \sum_{v_j\in X\bigcap U} n^B_j/d_j}, \tag{6}$$

where $\hat{N}^{ego}_{XY} = \sum_{v_i\in X\cap U} n^Y_i/d_i$ 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:

$$\sum_{v_j\in X\bigcap U} \frac{n^A_j}{d_j} + \sum_{v_j\in X\bigcap U} \frac{n^B_j}{d_j} = \sum_{v_j\in X\bigcap U} \left(\frac{n^A_j + n^B_j}{d_j}\right)$$

$$= \sum_{v_j\in X\bigcap U} \left(\frac{d_j}{d_j}\right) = n_X, \tag{7}$$

we have:

$$\hat{s}^{ego}_{XY} = \frac{1}{n_X} \cdot \sum_{v_i\in X\cap U} \frac{n^Y_i}{d_i}. \tag{8}$$

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

Using $\hat{s}^{ego}_{XY}$ 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:

$$N_A \overline{D}^*_A s^*_{AB} = N_B \overline{D}^*_B s^*_{BA}. \tag{9}$$

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

If we let $\hat{s}^{ego}_{XY}$ be the estimator of s^*_{XY} and let $\hat{\overline{D}}_X = n_X/(\sum_{v_i\in X\cap U} d_i^{-1})$ be the estimator of $\overline{D}^*_X$ $(X, Y \in \{A, B\})$, then P^*_A can be estimated by:

$$\hat{P}^{RDSI^{ego}}_A = \frac{\hat{s}^{ego}_{BA}\hat{\overline{D}}_B}{\hat{s}^{ego}_{AB}\hat{\overline{D}}_A + \hat{s}^{ego}_{BA}\hat{\overline{D}}_B}. \tag{10}$$

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

First, the sample size for inferring $\hat{s}^{ego}_{XY}$, 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 (Kogan et al., 2011; Tomas and Gile, 2011; de Mello et al., 2008), which leads to unknown bias and error for the representativeness of s_{XY}. $\hat{s}^{ego}_{XY}$, 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, $\hat{s}^{ego}_{XY}$ can greatly reduce estimate bias and error for such violation of assumption.

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
County (*ct*)	38.8	0.30	0.50	1.22
Civil status (*cs*)	40.4	0.57	0.05	0.97
Profession (*pf*)	38.2	0.54	0.13	1.21

Note also that the implementation of $RDSI^{ego}$ 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 determine 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 Lu et al. (2012, 2013) and Rybski et al. (2009), 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 $N = 16,082$, 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.

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, $w = \overline{D}_A^*/\overline{D}_B^*$. Previous studies have found that homophily and activity ratio are two critical factors that may affect the performance of RDS estimators (Gile and Handcock, 2010). 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 (Toivonen et al., 2009; Kumpula et al., 2007). These networks are configured with population size $N = 10,000$, 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): Like 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 Gile and Handcock (2010), 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 between the two simulation settings 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 (Kogan et al., 2011). In a study of MSM in Campinas City, Brazil (de Mello et al., 2008), participants were reported most often to recruit close peers or peers they believed practiced risky behaviors. In Gile and Handcock (2010), Lu et al. (2012), and Tomas and Gile (2011), 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 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 all peers among the personal network of the respondent have equal chance of receiving a coupon; and $p_A^{diff} = 1$ corresponds to differential recruitment, when peers of type A have one times more probability of receiving a coupon from the respondent, i.e., each peer of type A is twice as likely to be recruited as each peer of type B. We can see that $p_A^{diff} = 1$ models an extreme scenario of differential recruitment with 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 (McCreesh et al., 2012; Goel and Salganik, 2010), in addition to *RDSI* and $RDSI^{ego}$, we also include the raw sample composition in the analysis. The *RDSII* estimator in our simulations provides estimates with little difference to *RDSI* 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 (Gile, 2011; Lu et al., 2013; Gile and Handcock, 2011) are thus beyond the purpose of this study and are excluded from comparison.

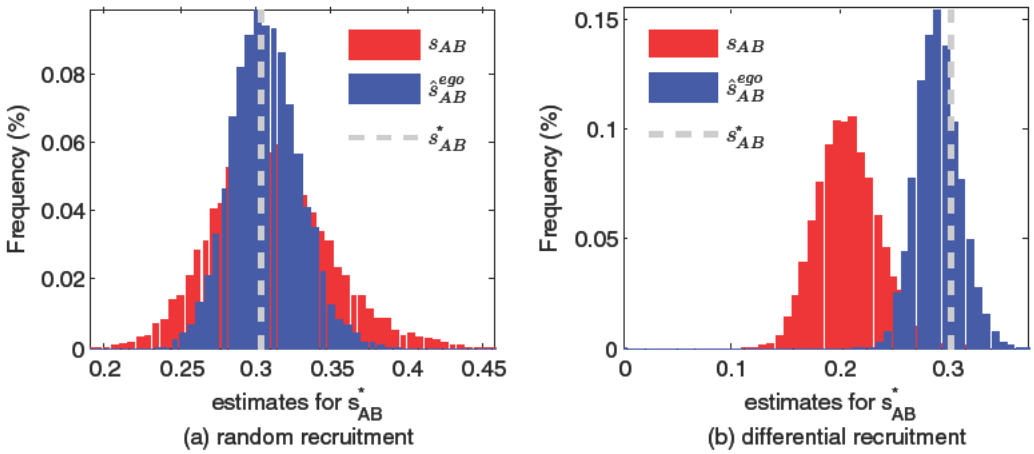

Fig. 2. Distribution of RDS estimates for $s^*_{AB}(ct)$. Dashed line shows population value $s^*_{AB}(ct)$. (a) Participants recruit respondents randomly among their friends, $p^{diff}_A = 0$; (b) participants are two times more likely to recruit friends of type *A* than friends of type *B*, $p^{diff}_A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

Four measurements are then carried out after the RDS simulations: the *Bias*, which is the absolute difference between the average estimate and population value, $|((\sum_{i=1}^{m} est_i)/m) - P^*_A|$ or $|((\sum_{i=1}^{m} est_i)/m) - s^*_{AB}|$, where est_i is the estimate from the ith simulation and m the number of simulation times; the *standard deviation* (SD) of estimates; the *root mean square error* (RMSE), $\sqrt{(\sum_{i=1}^{m}(est_i - P^*_A)^2)/m}$; and lastly, the *Percentage* an estimator outperforms the rest in all simulations: P^{best} = (*times the estimator gives closest estimate to* s^*_{AB} *or* P^*_A)$/m$.

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 *RDSI* and *RDSI*^{ego} lies in the estimation of the recruitment matrix *S*. As a first step, we therefore simulate the RDS process with random recruitment ($p^{diff}_A = 0$) and differential recruitment ($p^{diff}_A = 1$) 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, $\hat{s}^{ego}_{AB}$, 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 Fig. 2. Clearly, when the random recruitment assumption is fulfilled (Fig. 2(a)), both s_{AB} and $\hat{s}^{ego}_{AB}$ are unbiased. Estimates by $\hat{s}^{ego}_{AB}$ 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 $\hat{s}^{ego}_{AB}$ becomes more evident when RDS is implemented with differential recruitment. We can see from Fig. 2(b) that when peers who live in Stockholm are two times more likely to be recruited by their friends ($p^{diff}_A = 1$), the raw sample recruit proportion is largely undersampled (Bias = 0.09), while $\hat{s}^{ego}_{AB}$ 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, $\hat{s}^{ego}_{AB}$ is 70% times closer to s^*_{AB} than s_{AB}. Under differential recruitment, almost all $\hat{s}^{ego}_{AB}$ estimates ($P^{best} = 0.98$) outperform s_{AB}.

Simulation results for estimates of all variables are summarized in Table 2. The conclusions are similar to those above. $\hat{s}^{ego}_{AB}$ 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. $\hat{s}^{ego}_{AB}$, 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}, $\hat{s}^{ego}_{AB}$ produces estimates that are closer to the true population value s^*_{AB} 62–74% 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 $\hat{s}^{ego}_{AB}$ over s_{AB}.

4.1.2. Estimates of population compositions

The superiority of $\hat{s}^{ego}_{AB}$ over s_{AB} shown in the above section suggests that the *RDSI*^{ego} estimator should also give less bias and error than *RDSI*. To confirm this, we compare the simulation results of *RDSI*, and *RDSI*^{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 Fig. 3, where the median (middle line), the 25th and 75th percentiles (box) and outliers (whiskers) are shown. When $p^{diff}_A = 0$, there is on average an oversample of individuals who live in Stockholm (0.05) in the raw sample; however, if adjusted, *RDSI* and *RDSI*^{ego} all give unbiased estimates (Fig. 3(a)). When $p^{diff}_A = 1$, i.e., respondents are twice as likely to recruit friends from Stockholm rather than friends from other counties, the improvement in estimates by *RDSI*^{ego} 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 *RDSI*^{ego} is only 0.02 and RMSE is 0.06 (Fig. 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 $P^{best} = 0.93$.

The above conclusions are similar for all other variables (see Table 3): when $p^{diff}_A = 0$, both *RDSI* and *RDSI*^{ego} have little bias, while *RDSI*^{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

Table 2
Statistics of estimates for s^*_{AB} by s_{AB} and $\hat{s}^{ego}_{AB}$.

		Bias (standard deviation)		RMSE (P^{best})	
		s_{AB}	$\hat{s}^{ego}_{AB}$	s_{AB}	$\hat{s}^{ego}_{AB}$
Random recruitment					
Seed = 6 coupon = 2 SWOR	*age*	0.00 (0.03)	0.00[a] (0.03[a])	0.03 (0.37)	0.03[a] (0.63[a])
	ct	0.01 (0.04)	0.00[a] (0.02[a])	0.04 (0.30)	0.02[a] (0.70[a])
	cs	0.00 (0.04)	0.00[a] (0.02[a])	0.04 (0.26)	0.02[a] (0.74[a])
	pf	0.00 (0.04)	0.00[a] (0.02[a])	0.04 (0.26)	0.02[a] (0.74[a])
Differential recruitment					
Seed = 6 coupon = 2 SWOR	*age*	0.04 (0.03)	0.01[a] (0.03[a])	0.05 (0.16)	0.03[a] (0.84[a])
	ct	0.09 (0.03)	0.01[a] (0.02[a])	0.10 (0.02)	0.02[a] (0.98[a])
	cs	0.13 (0.04)	0.02[a] (0.02[a])	0.13 (0.00)	0.03[a] (1.0[a])
	pf	0.13 (0.03)	0.02[a] (0.02[a])	0.13 (0.00)	0.02[a] (1.0[a])

[a] Corresponding statistic is better than the other estimator.

Table 3
Statistics of estimates for P^*_A by sample mean, *RDSI* and $RDSI^{ego}$.

		Bias (standard deviation)			RMSE (P^{best})		
		Sample	*RDSI*	$RDSI^{ego}$	Sample	*RDSI*	$RDSI^{ego}$
Random recruitment							
Seed = 6 coupon = 2 SWOR	*age*	0.01 (0.06)	0.00[a] (0.07)	0.00 (0.06[a])	0.06 (0.37)	0.07 (0.23)	0.06[a] (0.40[a])
	ct	0.05 (0.05)	0.00 (0.06)	0.00[a] (0.05[a])	0.07 (0.28)	0.06 (0.29)	0.05[a] (0.43[a])
	cs	0.01 (0.03[a])	0.00 (0.04)	0.00[a] (0.03)	0.03[a] (0.51[a])	0.04 (0.17)	0.03 (0.32)
	pf	0.05 (0.03[a])	0.00[a] (0.04)	0.00 (0.03)	0.05 (0.20)	0.04 (0.32)	0.03[a] (0.48[a])
Differential recruitment							
Seed = 6 coupon = 2 SWOR	*age*	0.09 (0.05[a])	0.08 (0.06)	0.02[a] (0.07)	0.10 (0.10)	0.10 (0.12)	0.07[a] (0.79[a])
	ct	0.20 (0.06)	0.17(0.07)	0.02[a] (0.06[a])	0.21 (0.00)	0.18 (0.06)	0.06[a] (0.93[a])
	cs	0.12 (0.03[a])	0.13 (0.05)	0.02[a] (0.04)	0.13 (0.00)	0.14 (0.04)	0.04[a] (0.96[a])
	pf	0.18 (0.03[a])	0.13 (0.05)	0.02[a] (0.04)	0.18 (0.00)	0.14 (0.05)	0.04[a] (0.95[a])

[a] Corresponding statistic is better than other estimators.

estimators; $RDSI^{ego}$ always has larger P^{best} for all variables except *cs*, which has low homophily and a close to one activity ratio. *RDSI*, 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*. $RDSI^{ego}$ again becomes dominant when the sampling is done with differential recruitment. The bias ranges in [0.00, 0.02] and RMSE in [0.04, 0.07], while for sample composition and *RDSI* the bias and RMSE are much larger, [0.07, 0.20] and [0.09, 0.21], respectively.

To better understand the robustness of $RDSI^{ego}$ to differential recruitment, we simulate RDS processes on the MSM network with p^{diff}_A varying from 0 to 1. The average estimates for the four variables are shown in Fig. 4. While the bias of *RDSI* increases progressively with p^{diff}_A, $RDSI^{ego}$ 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 Fig. 5, where the biases of both *RDSI* and $RDSI^{ego}$ are shown for networks with different levels of homophily ($h_A \in [0, 0.5]$) and activity ratio $w \in [0.5, 2.5]$.

There is no clear trend on how the estimate bias will increase over homophily or activity ratio. The effects of homophily and activity ratio are mixed with impact of other network structural

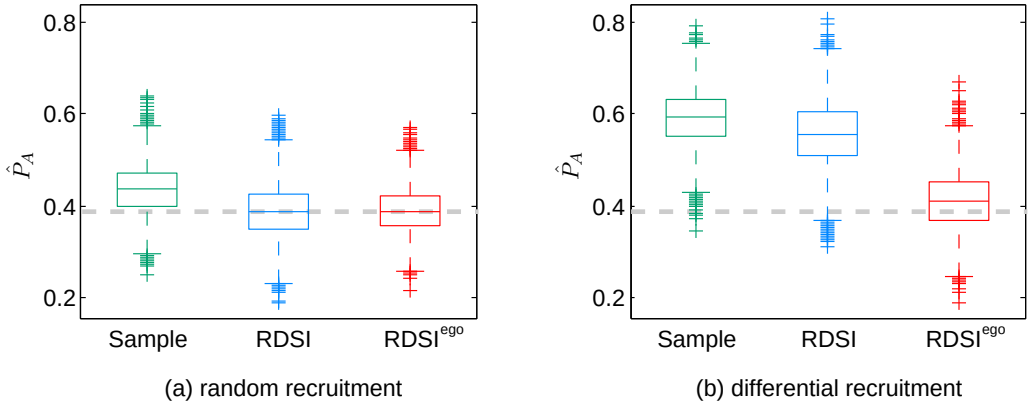

Fig. 3. RDS estimates for P^*_{ct}. Dashed line is of population value P^*_{ct}. (a) Participants recruit respondents randomly among their friends, $p^{diff}_A = 0$; (b) participants are two times more likely to recruit friends of type *A* rather than friends of type *B*, $p^{diff}_A = 1$.

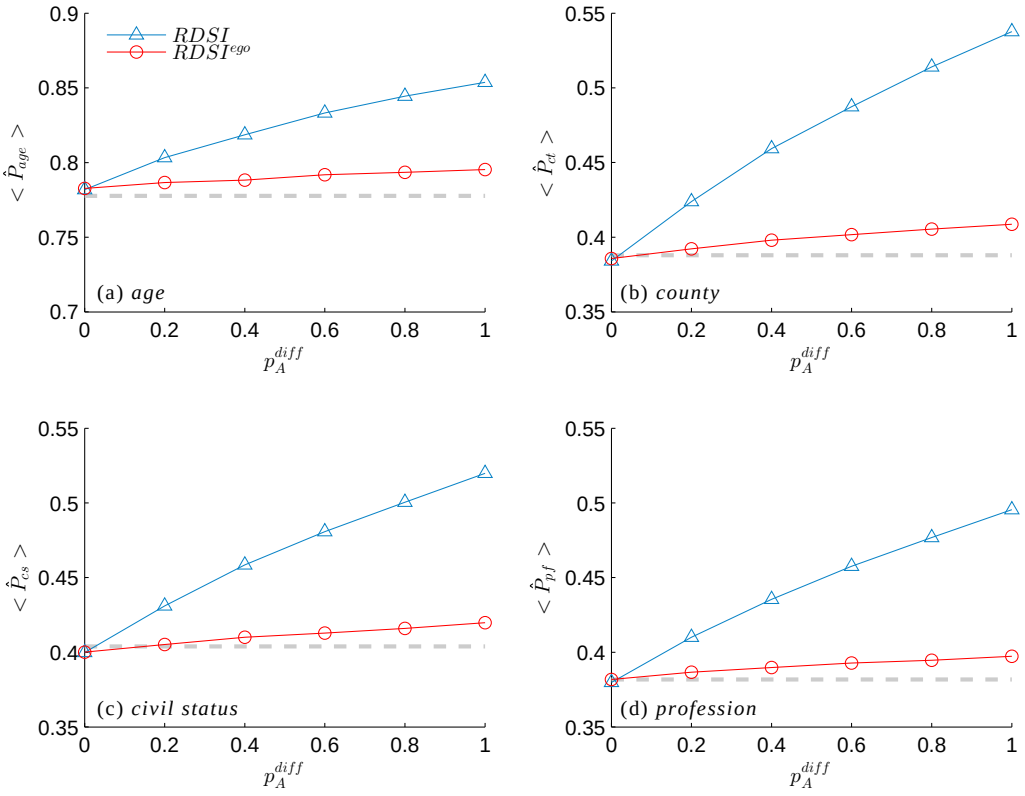

Fig. 4. Average estimates of *RDSI* and *RDSI*^{ego} on the MSM network with varying level of differential recruitment.

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

4.2. Degree reporting error

With the superior performance observed from the above section, we will now focus on *RDSI*^{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 (Salganik et al., 2011; Lu et al., 2013). This reporting error will not only affect the estimates of average degree, but further bias the estimate of the recruitment matrix in *RDSI*^{ego}, $\hat{s}_{XY}^{ego}(X, Y \in A, B)$.

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 Fig. 6 for the MSM network and Fig. 7 for KOSKK networks.

Surprisingly, on both the MSM network and KOSKK networks, even with 20% of all alters being miscounted, the biases of *RDSI*^{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 Lu et al. (2012), where the degree reporting error was modeled as unawareness of existing relationships. These findings imply 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 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 Figs. 13 and 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 $p_{A\mapsto B}^{error}$ and $p_{B\mapsto A}^{error}$ 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 Figs. 8 and 9.

Contrary to the robustness to degree reporting error, the *RDSI*^{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 exceeds 0.1 as long as $p_A^{diff} > 0.1$

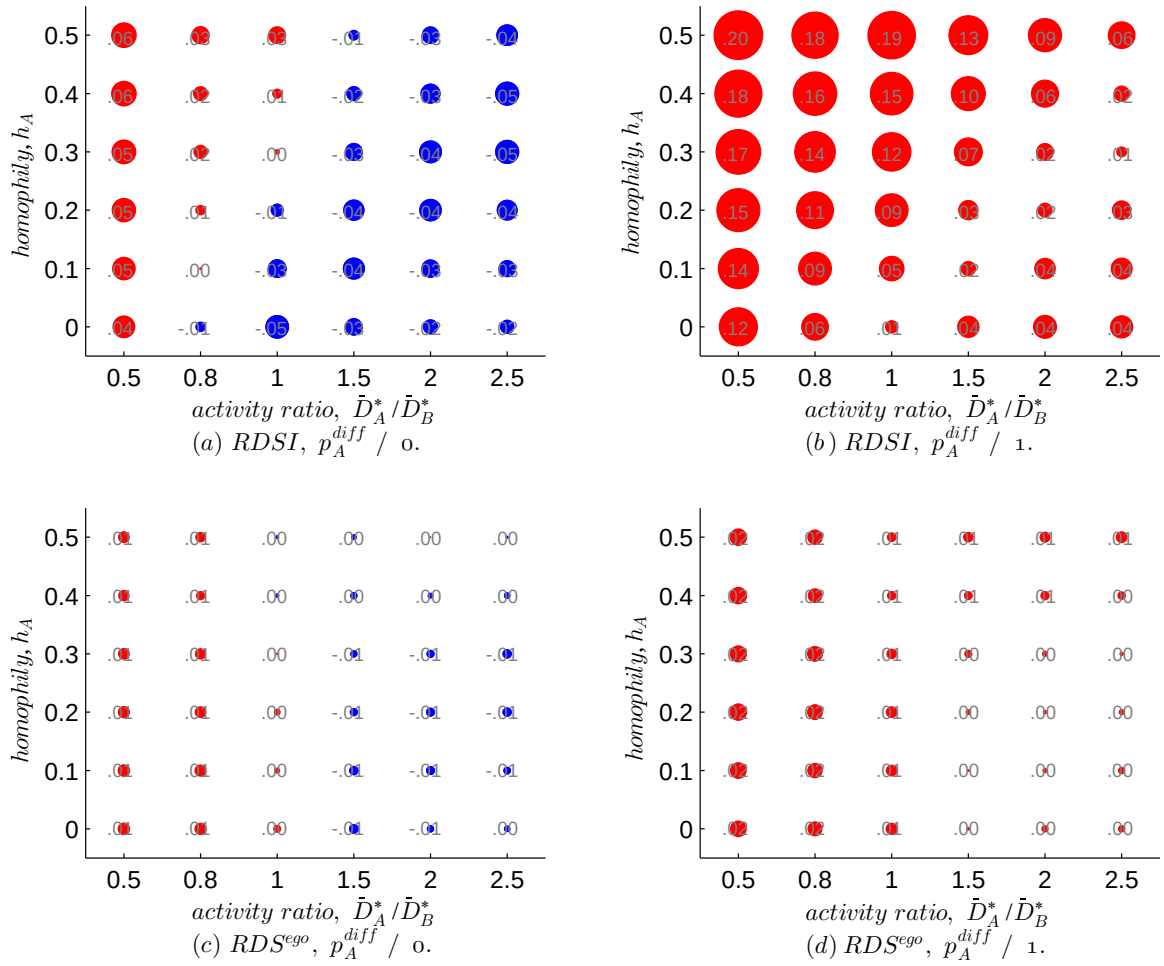

Fig. 5. Bias of *RDSI* and $RDSI^{ego}$ on KOSKK networks with random recruitment (a, c) and differential recruitment (b, d). The color of circles stands for the direction of bias: positive (red), negative (blue). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

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 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 ($w \leq 1$). 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 $w \leq 1$, 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 $RDSI^{ego}$ 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 reduction of bias is observed, see Appendix Figs. 15 and 16.

5. Conclusion and discussion

Ego network data has been collected for decades and exists largely in sociological surveys (Britton and Trapman, 2012; Everett and Borgatti, 2005; Handcock and Gile, 2010; Newman, 2003; Mizruchi and Marquis, 2006; Marsden, 2002; Hanneman and Riddle, 2005); 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, $RDSI^{ego}$, 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, $RDSI^{ego}$ shows strong robustness to differential recruitment, a 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 $RDSI^{ego}$ performs consistently well on networks with varying homophily, activity ratio, and community structures. The limitation of $RDSI^{ego}$ 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 or about 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

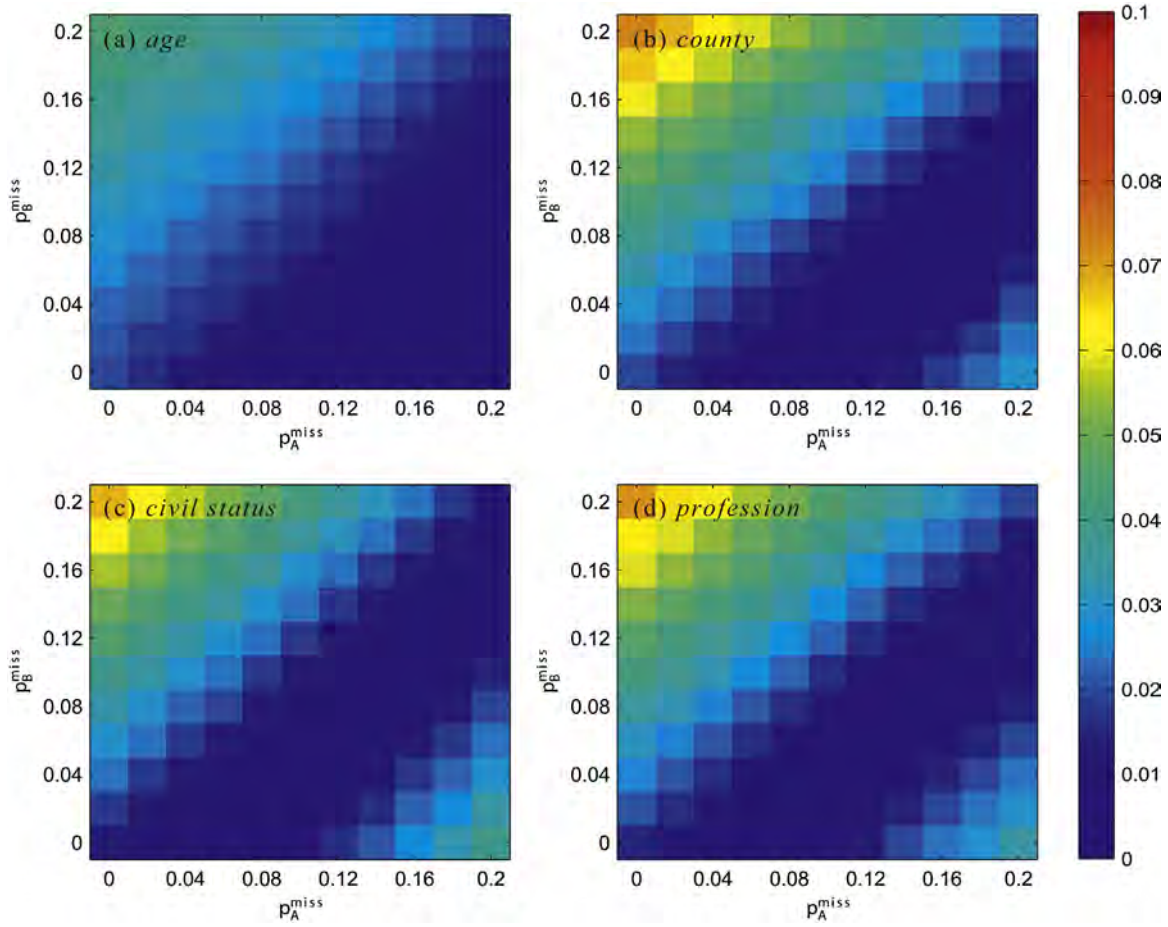

Fig. 6. Bias of $RDSI^{ego}$ on the MSM network with differential recruitment and *degree reporting error*.

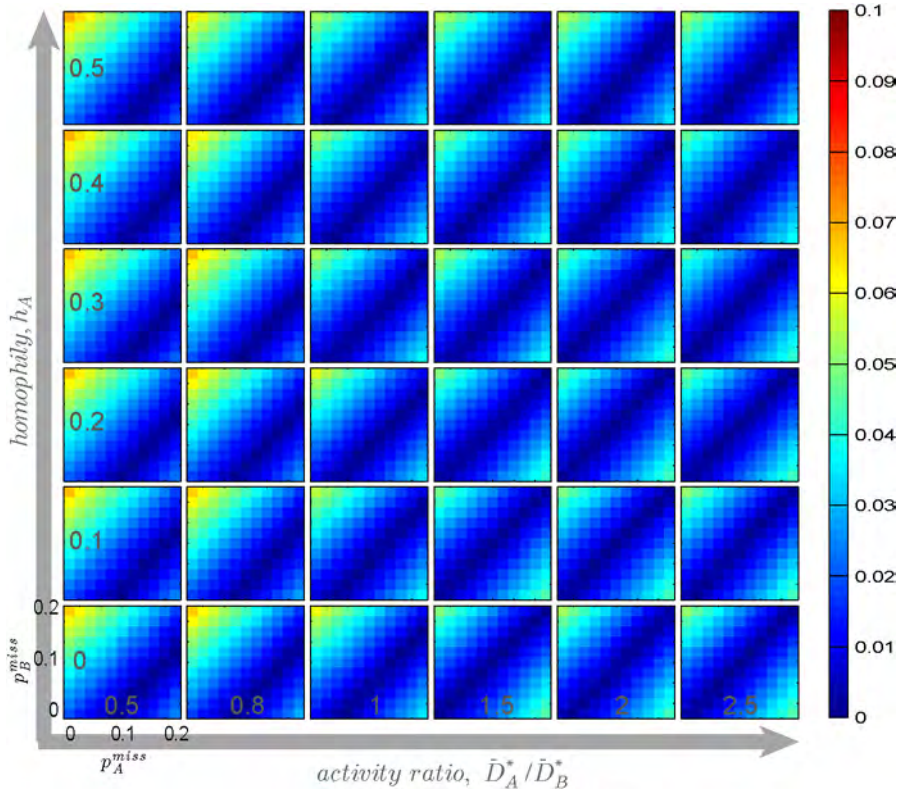

Fig. 7. Bias of $RDSI^{ego}$ on KOSKK network with differential recruitment and *degree reporting error*.

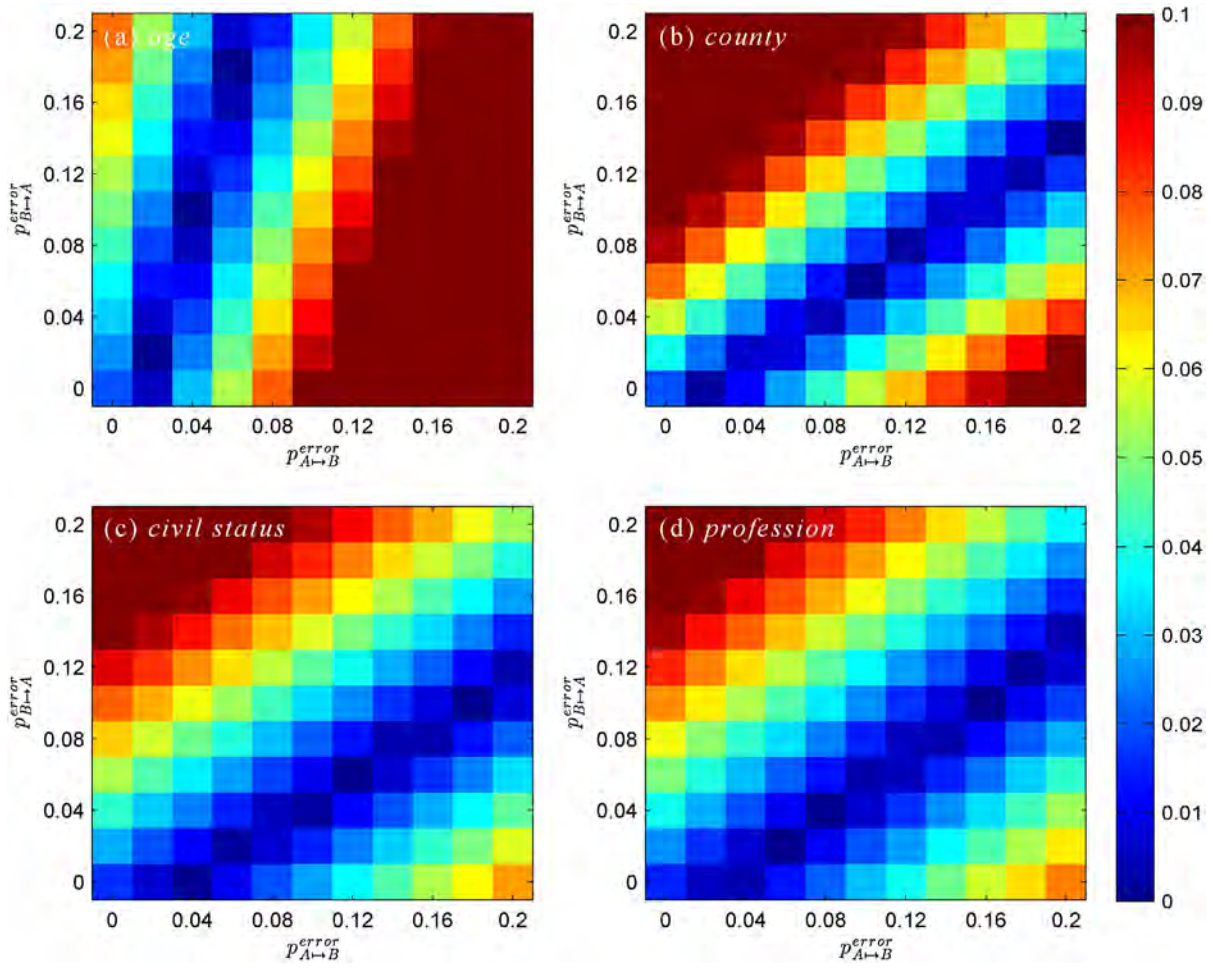

Fig. 8. Bias of $RDSI^{ego}$ on the MSM network with differential recruitment and *ego network reporting error*.

survey methods on how to design and collect ego network data has been extensively studied (Kogovšek and Ferligoj, 2005; Matzat and Snijders, 2010; Burt, 1984; Marin, 2004). Additionally, certain variables, e.g., drug use, may be highly sensitive in the general population but may not 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 shown that even with 20% of alters being unidentified, $RDSI^{ego}$ was still able to produce estimates with bias less than 0.05 most of the time. On the other hand, $RDSI^{ego}$ 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., $p^{error}_{A\mapsto B} \gg p^{error}_{B\mapsto A}$). 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 *RDSI*, which has been used in most RDS studies so far, fails to outperform the sample composition in many simulation settings. Second, we modified the traditional bootstrap method for constructing confidence intervals (CIs) with $RDSI^{ego}$ (see Appendix) and it shows that the modified procedure 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. However, even with the improved procedure, 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 (McCreesh et al., 2012; Goel and Salganik, 2010; Salganik, 2012). It is worth noting that the ego network data was incorporated in a newly developed estimator (*GH-estimator*) (Gile and Handcock, 2011). However, the *GH-estimator* requires the population size as a prerequisite to generate estimates. For RDS implementations where prior information on population data is extremely difficult to obtain, such as HIV/AIDS related high risk populations, the application of *GH-estimator* would be impaired.

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 (Gile and Handcock, 2010; Lu et al., 2012; Tomas

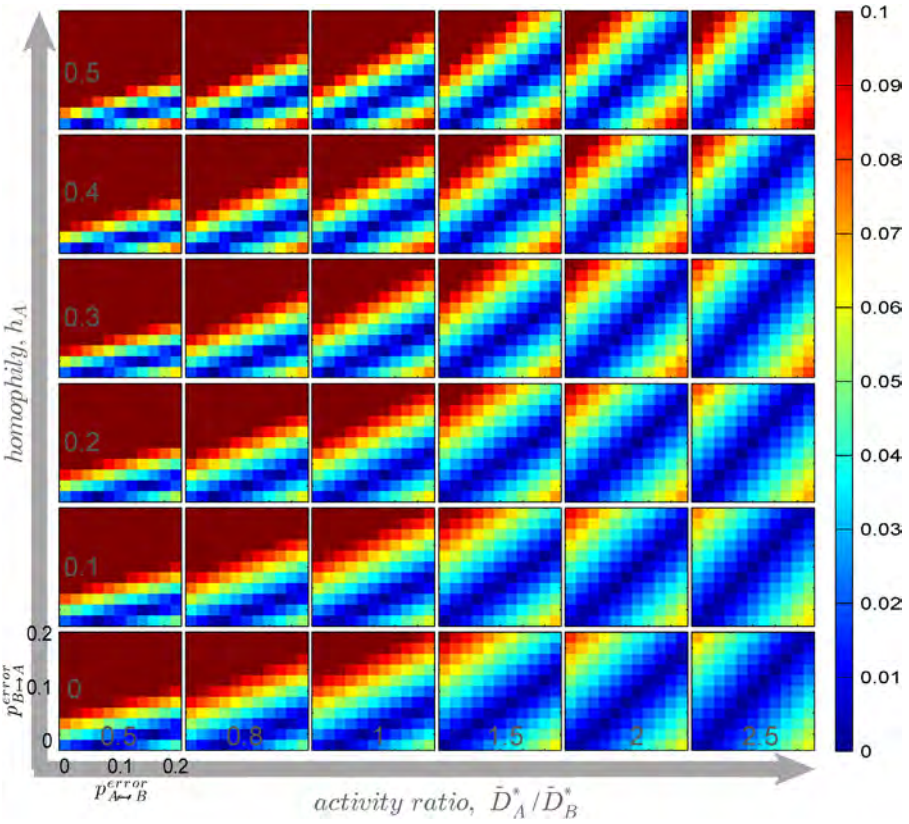

Fig. 9. Bias of $RDSI^{ego}$ on KOSKK network with differential recruitment and *ego network reporting error*.

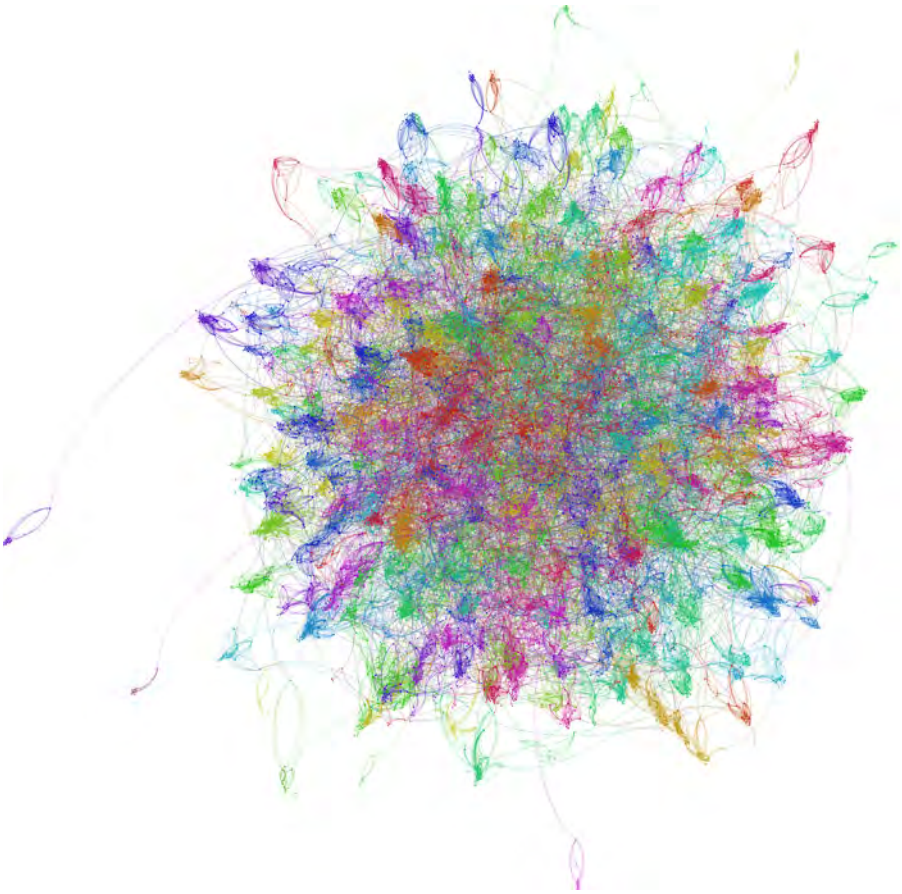

Fig. 10. Visualization of the KOSKK network generated with $\delta = 0.6$, $\overline{D}^* = 10$.

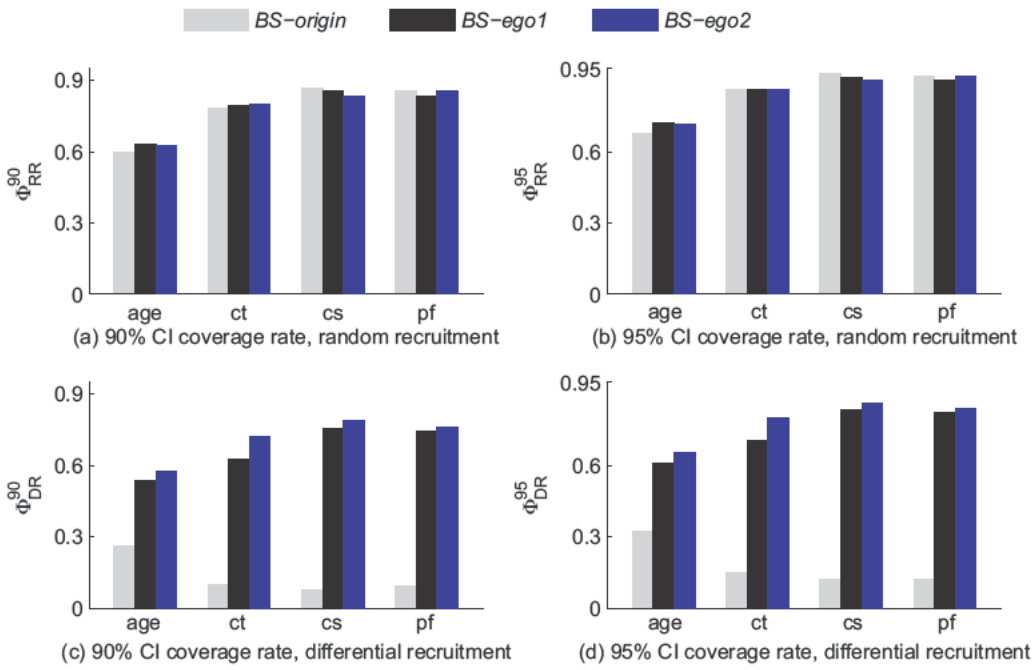

Fig. 11. Coverage rate of 90% and 95% confidence interval, by bootstrap procedure *BS-origin*, *BS-ego1*, and *BS-ego2* on the MSM network.

activity ratio, $\bar{D}^{\star}_{A} / \bar{D}^{\star}_{B}$

homophily, h_A	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]
0.2	.64 (.86) [.86]	.28 (.84) [.81]	.23 (.85) [.81]	.48 (.77) [.81]	.47 (.75) [.80]	.47 (.81) [.86]
0.1	.66 (.80) [.82]	.33 (.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]

(a) Φ^{95}_{RR}

activity ratio, $\bar{D}^{\star}_{A} / \bar{D}^{\star}_{B}$

homophily, h_A	0.5	0.8	1.0	1.5	2.0	2.5
0.5	.19 (.73) [.87]	.16 (.74) [.82]	.16 (.81) [.78]	.27 (.76) [.81]	.32 (.76) [.75]	.48 (.70) [.72]
0.4	.17 (.75) [.87]	.18 (.78) [.82]	.19 (.77) [.76]	.30 (.80) [.81]	.43 (.77) [.77]	.73 (.68) [.68]
0.3	.18 (.74) [.83]	.20 (.80) [.83]	.19 (.80) [.78]	.32 (.79) [.80]	.65 (.73) [.72]	.79 (.74) [.73]
0.2	.18 (.74) [.85]	.22 (.78) [.83]	.22 (.79) [.79]	.44 (.79) [.80]	.69 (.77) [.76]	.57 (.81) [.83]
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^{95}_{DR}

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

and Gile, 2011; Bengtsson and Thorson, 2010). The robustness of $RDSI^{ego}$ 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 (Gjoka et al., 2010; Lee et al., 2006; Yoon et al., 2007), such as sampling of internet contents from which the ego network data is more reliable and may be more efficiently retrieved.

Acknowledgement

The author would like to thank Professor Fredrik Liljeros and Dr. Linus Bengtsson for helpful discussions. This work has been

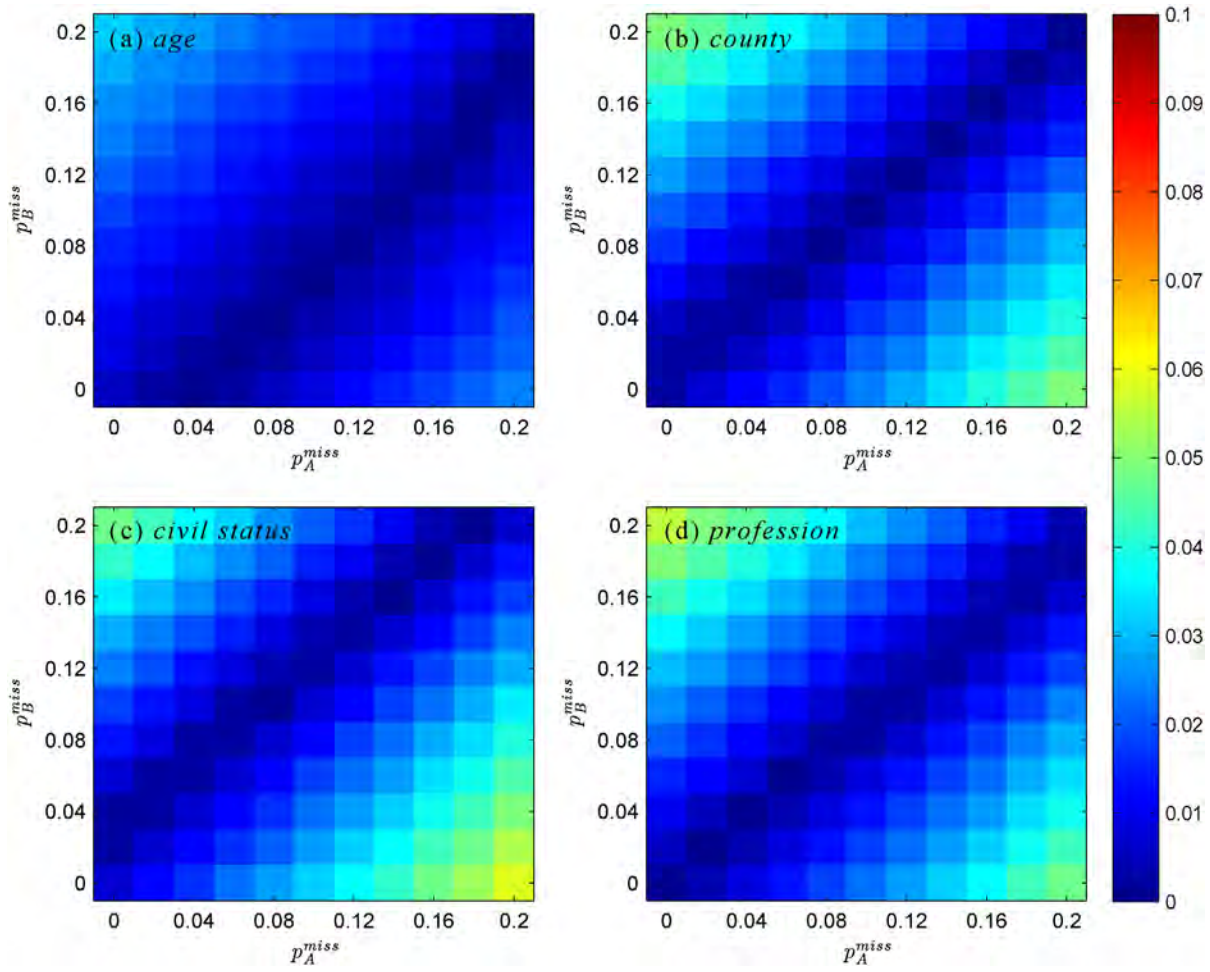

Fig. 13. Bias of $RDSI^{ego}$ on the MSM network with random recruitment and *degree reporting error*.

partially funded by Riksbankens Jubileumsfond (The Bank of Sweden Tercentenary Foundation).

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 features of social networks (Kumpula et al., 2007): (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 (Toivonen et al., 2009), 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 $N = 10,000$, $w_0 = 1$, $p_r = 0.0005$, $p_d = 0.001$, $\delta = 0.6$, 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 Fig. 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

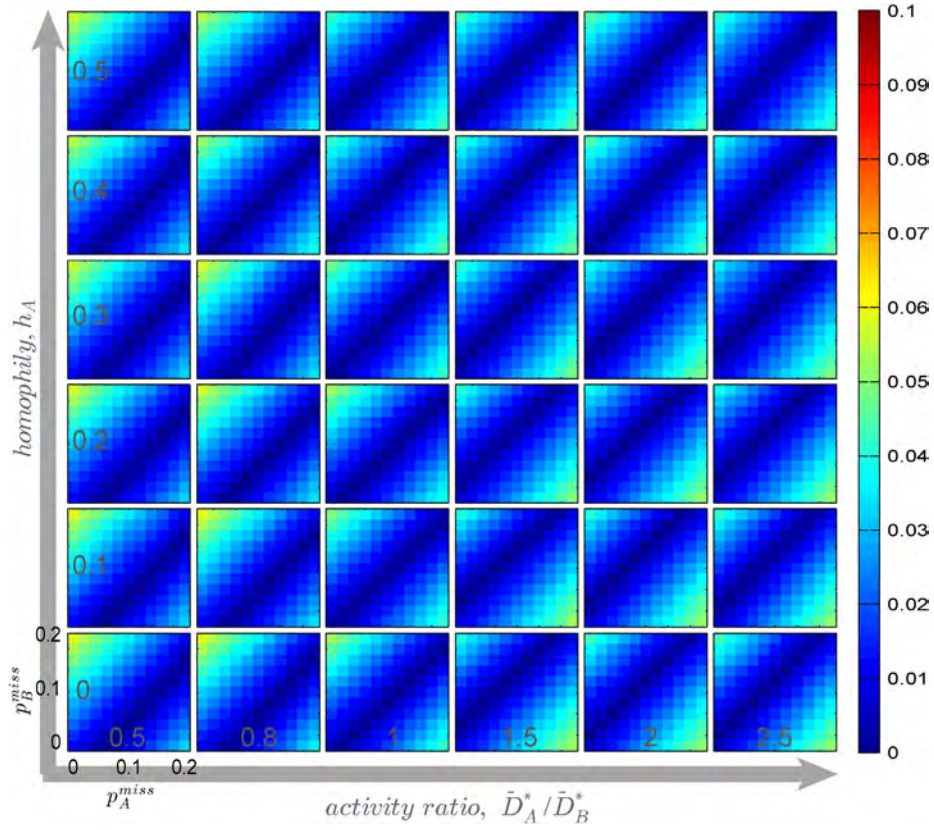

Fig. 14. Bias of $RDSI^{ego}$ on KOSKK network with random recruitment and *degree reporting error*.

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. Clearly, this procedure keeps the degrees for the selected nodes unchanged and will not alter the activity ratio of the network. The above process is repeated until $h_A = h_A^*$.

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 (Goel and Salganik, 2010; Heckathorn, 2002; Salganik, 2006). 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 (2006) and was implemented in the RDS data analysis software RDSAT (Volz et al., 2007). 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-ego1*: We implement the same resampling procedure as with *BS-origin*; however, when each replicated sample is collected, $RDSI^{ego}$ is used to calculate the RDS estimate, rather than *RDSI*;
(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 $\hat{s}_{BA}^{ego}$ and $1 - \hat{s}_{BA}^{ego}$, respectively. The above process is repeated until the same size as original sample is reached. $RDSI^{ego}$ 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 Salganik (2006), 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 Figs. 11 and 12.

On the MSM network, when sampling with random recruitment, we can see from Fig. 11(a) and (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

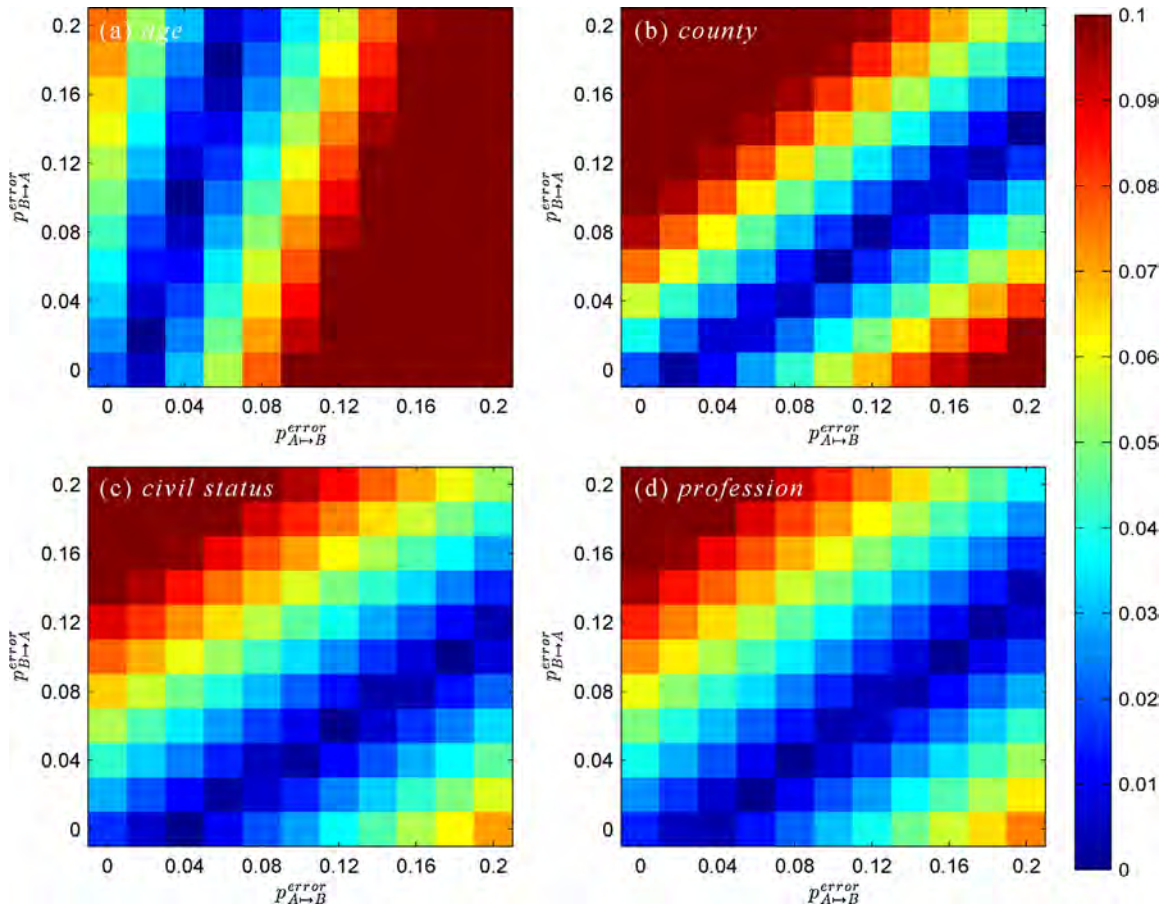

Fig. 15. Bias of $RDSI^{ego}$ on the MSM network with random recruitment and *ego network reporting error*.

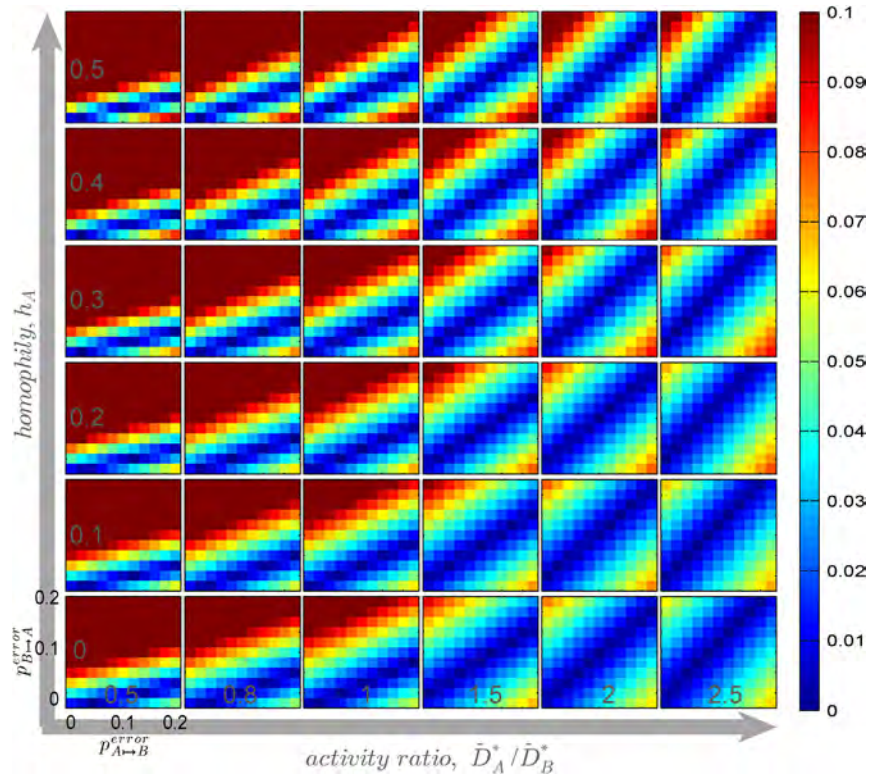

Fig. 16. Bias of $RDSI^{ego}$ on KOSKK network with random recruitment and *ego network reporting error*.

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 (Fig. 11(c) and (d)), the coverage rate of *BS-origin* becomes extremely small and practically useless. This is because the *RDSI* 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 $RDSI^{ego}$-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.

Appendix C. Supporting figures

Figs. 13–16.

References

Bengtsson, L., Lu, X., Nguyen, Q.C., Camitz, M., Hoang, N.L., Nguyen, T.A., Liljeros, F., Thorson, A., 2012. Implementation of web-based respondent-driven sampling among men who have sex with men in vietnam. PLoS ONE 7 (11), e49417.

Bengtsson, L., Thorson, A., 2010. Global HIV surveillance among MSM: is risk behavior seriously underestimated? AIDS 24 (15), 2301–2303.

Britton, T., Trapman, P., 2012. Inferring global network properties from egocentric data with applications to epidemics. arXiv:1201.2788v1.

Burt, R.S., 1984. Network items and the general social survey. Social Networks 6 (4), 293–339.

Deaux, E., Callaghan, J.W., 1985. Key informant versus self-report estimates of health-risk behavior. Evaluation Review 9 (3), 365–368.

de Mello, M., de Araujo Pinho, A., Chinaglia, M., Tun, W., Júnior, A.B., Ilário, M.C.F.J., Reis, P., Salles, R.C.S., Westman, S., Díaz, J., 2008. 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, Technical Report, Population Council.

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

Everett, M., Borgatti, S.P., 2005. Ego network betweenness. Social Networks 27 (1), 31–38.

Gile, K.J., 2011. Improved inference for respondent-driven sampling data with application to HIV prevalence estimation. Journal of the American Statistical Association 106 (493), 135–146.

Gile, K.J., Handcock, M.S., 2010. Respondent-driven sampling: an assessment of current methodology. Sociological Methodology 40, 285–327.

Gile, K.J., Handcock, M.S., 2011. Network model-assisted inference from respondent-driven sampling data. arXiv:1108.0298v1.

Gjoka, M., Kurant, M., Butts, C.T., Markopoulou, A., 2010. Walking in facebook: a case study of unbiased sampling of OSNs. In: INFOCOM, 2010 Proceedings IEEE, pp. 1–9.

Goel, S., Salganik, M.J., 2010. Assessing respondent-driven sampling. Proceedings of the National Academy of Sciences of the United States of America 107 (15), 6743–6747.

Handcock, M.S., Gile, K.J., 2010. Modeling social networks from sampled data. Annals of Applied Statistics 4 (1), 5–25.

Hanneman, R.A., Riddle, M., 2005. Introduction to Social Network Methods. University of California, Riverside, Riverside, CA.

Hansen, M.H., Hurwitz, W.N., 1943. On the theory of sampling from finite populations. Annals of Mathematical Statistics 14 (4), 333–362.

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. Sociological Methodology 37 (37), 151–208.

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.

Kogan, S.M., Wejnert, C., Chen, Y.F., Brody, G.H., Slater, L.M., 2011. Respondent-driven sampling with hard-to-reach emerging adults: an introduction and case study with rural African Americans. Journal of Adolescent Research 26 (1), 30–60.

Kogovšek, T., Ferligoj, A., 2005. The quality of measurement of personal support subnetworks. Quality & Quantity 38 (5), 517–532.

Kumpula, J., Onnela, J., Saramäki, J., Kaski, K., Kertész, J., 2007. Emergence of communities in weighted networks. Physical Review Letters 99, 228701.

Lansky, A., Abdul-Quader, A.S., Cribbin, M., Hall, T., Finlayson, T.J., Garfein, R.S., Lin, L.S., Sullivan, P.S., 2007. Developing an HIV behavioral surveillance system for injecting drug users: the national HIV behavioral surveillance system. Public Health Reports 122, 48–55.

Lee, S.H., Kim, P.-J., Jeong, H., 2006. Statistical properties of sampled networks. Physical Review E 73 (1), 016102.

Li, J., Liu, H.J., Li, J.H., Luo, J., Koram, N., Detels, R., 2011. Sexual transmissibility of HIV among opiate users with concurrent sexual partnerships: an egocentric network study in Yunnan, China. Addiction 106 (10), 1780–1787.

Lu, X., Bengtsson, L., Britton, T., Camitz, M., Kim, B.J., Thorson, A., Liljeros, F., 2012. The sensitivity of respondent-driven sampling. Journal of the Royal Statistical Society: Series A (Statistics in Society) 175 (1), 191–216.

Lu, X., Malmros, J., Liljeros, F., Britton, T., 2013. Respondent-driven sampling on directed networks. Electronic Journal of Statistics 7, 292–322.

Magnani, R., Sabin, K., Saidel, T., Heckathorn, D., 2005. Review of sampling hard-to-reach and hidden populations for HIV surveillance. AIDS 19 (Suppl. 2), S67–S72.

Marin, A., 2004. Are respondents more likely to list alters with certain characteristics? Implications for name generator data. Social Networks 26 (4), 289–307.

Marsden, P.V., 2002. Egocentric and sociocentric measures of network centrality. Social Networks 24 (4), 407–422.

Matzat, U., Snijders, C., 2010. Does the online collection of ego-centered network data reduce data quality? An experimental comparison. Social Networks 32 (2), 105–111.

McCreesh, N., Frost, S.D.W., Seeley, J., Katongole, J., Tarsh, M.N., Ndunguse, R., Jichi, F., Lunel, N.L., Maher, D., Johnston, L.G., Sonnenberg, P., Copas, A.J., Hayes, R.J., White, R.G., 2012. Evaluation of respondent-driven sampling. Epidemiology 23 (1), 138–147.

Mizruchi, M.S., Marquis, C., 2006. Egocentric, sociocentric, or dyadic? Identifying the appropriate level of analysis in the study of organizational networks. Social Networks 28 (3), 187–208.

Newman, M.E.J., 2003. Ego-centered networks and the ripple effect. Social Networks 25 (1), 83–95.

Rudolph, A.E., Latkin, C., Crawford, N.D., Jones, K.C., Fuller, C.M., 2011. Does respondent driven sampling alter the social network composition and health-seeking behaviors of illicit drug users followed prospectively? PLoS ONE 6 (5).

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., Mello, M., Abdo, A., Bertoni, N., Fazito, D., Bastos, F., 2011. The game of contacts: estimating the social visibility of groups. Social Networks 33 (1), 70–78.

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., 2012. Commentary: respondent-driven sampling in the real world. Epidemiology 23 (1), 148–150.

Salganik, M.J., Heckathorn, D.D., 2004. Sampling and estimation in hidden populations using respondent-driven sampling. Sociological Methodology, vol. 34. Wiley-VCH Verlag GmbH, Weinheim, pp. 193–239.

Toivonen, R., Kovanen, L., Kivelä, M., Onnela, J., Saramäki, J., Kaski, K., 2009. A comparative study of social network models: network evolution models and nodal attribute models. Social Networks 31 (4), 240–254.

Tomas, A., Gile, K.J., 2011. The effect of differential recruitment, non-response and non-recruitment on estimators for respondent-driven sampling. Electronic Journal of Statistics 5, 899–934.

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

Volz, E., Wejnert, C., Degani, I., Heckathorn, D.D., 2007. Respondent-driven sampling analysis tool (RDSAT) Version 5.6.

Watters, J.K., Biernacki, P., 1989. Targeted sampling: options for the study of hidden populations. Social Problems 36 (4), 416–430.
Wejnert, C., 2009. An empirical test of respondent-driven sampling: point estimates, variance, degree measures, and out-of-equilibrium data. Sociological Methodology 2009 39 (39), 73–116.
Wejnert, C., Heckathorn, D.D., 2008. Web-based network sampling – efficiency and efficacy of respondent-driven sampling for online research. Sociological Methods & Research 37 (1), 105–134.
Yoon, S., Lee, S., Yook, S.-H., Kim, Y., 2007. Statistical properties of sampled networks by random walks. Physical Review E 75 (4), 046114.