Mathematical Statistics
Stockholm University

Respondent-driven Sampling on Directed Networks

Xin Lu
Jens Malmros
Fredrik Liljeros
Tom Britton

Research Report 2011:8

ISSN 1650-0377

Postal address:

Mathematical Statistics
Dept. of Mathematics
Stockholm University
SE-106 91 Stockholm
Sweden

Internet:

http://www.math.su.se/matstat

Mathematical Statistics
Stockholm University
Research Report **2011:8**,
http://www.math.su.se/matstat

Respondent-driven Sampling on Directed Networks

Xin Lu, Jens Malmros, Fredrik Liljeros and Tom Britton

Novenber 2011

Abstract

Respondent-driven sampling (RDS) is a commonly used substitute for random sampling when studying hidden populations, such as injective drug users or men who have sex with men, for which no sampling frame is known. The method works like a snowball sample but can, given that some assumptions are met, generate unbiased population estimates. One key assumption, not likely to be met, is that the acquaintance network in which the recruitment process takes place is undirected, meaning that all recruiters should have the potential to be recruited by the person they recruit. Here we investigate the potential bias of directedness by simulating RDS on real and artificial network structures. We show that directedness is likely to generate bias that cannot be compensated for unless the sampled individuals know how many potentially may have recruited them (i.e. their indegree), which is unlikely in most situations. We propose three indegree-based estimators for RDS on directed networks and show that they can be used together with prior information or in a sensitivity analysis, taking uncertainty of indegree properties of the network into account for the situation when only outdegrees are observed.

Respondent-driven Sampling on Directed Networks

Xin Lu[*]
Department of Public Health Sciences, karolinska Institutet
Department of Sociology, Stockholm University

Jens Malmors
Department of Mathematics, Stockholm University

Fredrik Liljeros
Department of Sociology, Stockholm University

Tom Britton
Department of Mathematics, Stockholm University

November 28, 2011

Abstract

Respondent-driven sampling (RDS) is a commonly used substitute for random sampling when studying hidden populations, such as injective drug users or men who have sex with men, for which no sampling frame is known. The method works like a snowball sample but can, given that some assumptions are met, generate unbiased population estimates. One key assumption, not likely to be met, is that the acquaintance network in which the recruitment process takes place is undirected, meaning that all recruiters should have the potential to be recruited by the person they recruit. Here we investigate the potential bias of directedness by simulating RDS on real and artificial network structures. We show that directedness is likely to generate bias that cannot be compensated for unless the sampled individuals know how many potentially may have recruited them (i.e. their indegree), which is unlikely in most situations. We propose three indegree-based estimators for RDS on directed networks and show that they can be used together with prior information or in a sensitivity analysis, taking uncertainty of indegree properties of the network into account for the situation when only outdegrees are observed.

---

[*] Corresponding author, lu.xin@sociology.su.se

Keywords: Respondent-driven Sampling | directed networks | HIV

Hidden populations (hard-to-reach populations), such as injecting drug users (IDU), men who have sex with men (MSM), and sex workers (SW) and their sexual partners, are generally considered as critical actors in the HIV epidemic [1–3]. Consequently, obtaining population characteristics and risk behaviors of these populations are critical for developing efficient disease control strategies. However, the lack of sampling frames for such populations makes traditional estimation methods based on random samples practically useless. Other methods have been proposed for such situations, for example key informant sampling [4], targeted/location sampling [5] and snowball sampling [6].

A more recent method is Respondent Driven Sampling (RDS), which was proposed to overcome difficulties when sampling hidden populations [7–9]. The RDS method starts with an initial selection of respondents, who are called "seeds". Each seed is given a number of "coupons" – tickets for participation in the study – to distribute to friends and acquaintances within the population of interest. When interviewed, a new respondent is in turn given coupons to distribute. Everyone is rewarded both for completing the interview, and for recruiting their peers into the study. If the recruitment chains are sufficiently long, the sample composition will stabilize and become independent of the seeds. Additionally, information about who recruits whom and each respondent's personal network size (degree) are recorded.

Suppose a RDS study is performed on a connected undirected network with the additional assumptions that sampling of peer recruitment is done with replacement and all participants randomly recruit a fixed number of peers from their neighbors. Then, the sampling probability of an individual v_i will be proportional to its degree when the sample reaches equilibrium. The population fraction p_A having a certain property A (e.g. p_A could denote the fraction among intravenous drug-users that are HIV-positive) can then be estimated by the weighted proportion of the sample fraction [9]:

$$\hat{p}_{outdeg} = \frac{\sum\limits_{v_i \in U \cap A} {d_i}^{-1}}{\sum\limits_{v_i \in U} {d_i}^{-1}}, \tag{1}$$

where the sample population U has been divided into two disjoint subsets A and $B = A^C$ depending on the reported properties of respondents and d_i denotes the degree of individual i in the sample. This estimator has an equivalent expression which uses the harmonic means of the degrees for the two sub-groups in the sample: $\hat{\hat{D}}_A = \frac{n_A}{\sum\limits_{v_i \in U \cap A} {d_i}^{-1}}$ and $\hat{\hat{D}}_B = \frac{n_B}{\sum\limits_{v_i \in U \cap B} {d_i}^{-1}}$ (n_A and n_B denote the number of A- and B-individuals in the sample respectively). Using these quantities, an equivalent expression for $\hat{p}_A$ is given by [8, 11]:

$$\hat{p}_{outdeg} = \frac{\hat{s}_{BA}\hat{\hat{D}}_B}{\hat{s}_{AB}\hat{\hat{D}}_A + \hat{s}_{BA}\hat{\hat{D}}_B}, \tag{2}$$

where $\hat{s}_{BA}$ denotes the sample fraction of all B-respondents naming A-peers and similarly $\hat{s}_{AB}$ denotes the sample fraction of all A-respondents naming B-peers.

The ability to produce unbiased population estimates and a feasible field implementation have contributed to a rapid increase in RDS studies conducted globally in recent years [2, 10]. There has also been an increase in studies evaluating the performance of RDS estimators as well as in developing new estimators [12–15]. Previous studies are mostly based on the assumption that the underlying network is undirected, i.e., all relationships through which recruitments could take place are reciprocal. However, it is well-known that social networks, such as friendship networks, are generally directed to various extents. For example, in the study conducted by Scott and Dana [16], only 6,669 out of 12,931 “best friend” nominations were found to be reciprocal, and in the study conducted by Wallace [17, 18], an average of 55.0 reciprocal nominations per respondent were found while the mean degree was 94.8. It has been shown that current RDS estimators may generate relatively large biases and errors if the studied networks are directed [19], indicating that estimates from previous RDS studies should be interpreted and generalized with caution.

This study aims to derive new estimators allowing for networks to be (partially) directed. We develop estimators utilizing data with various levels of detail. To compare the performance of estimators, and to assess the influence of structural network properties, we use simulated data, as well as a real online MSM social network, to generate networks with varying directedness, degree correlation, indegree-outdegree correlation and homophily. Additionally, we propose a sensitivity analysis method to generate estimate intervals induced by intervals of unobserved network properties.

1 RDS estimation on directed networks

We now present a few estimators for the community fraction p_A having a certain property A. The different estimators rely on different amounts of information and different assumptions about the network on which RDS is performed. We let G denote our (partially) directed network and let $e_{ij} = 1$ if there is a directed edge from i to j and $e_{ij} = 0$ otherwise. A reciprocal edge between i and j is hence reflected by $e_{ij} = e_{ji} = 1$. We assume that G is strongly connected, i.e., that there is a directed path between any pair of nodes – otherwise we of course have no chance to estimate p_A well since it may then be impossible to reach certain parts of the community with RDS. Finally, we let N denote the community size, most often an unknown quantity in hidden or hard-to-reach populations.

1.1 Knowing the complete network structure

RDS is simply a random walk on G with transition matrix $R = \{a_{ij} = e_{ij}/d_i^{out}, 1 \leqslant i, j \leqslant N\}$, where d_i^{out} is the outdegree of node i. This process has a unique equilibrium distribution $\pi = [\pi_1 \cdots \pi_N]$ satisfying $R^T \pi^T = \pi^T$, indicating that π is the eigenvector corresponding to eigenvalue 1 for R^T. If G is completely known, as currently assumed, the stationary distribution π can be computed numerically (albeit time-consuming if the community is large) and this can then be used to obtain the Hansen-Hurwitz estimator where observations are weighted by the inverse of the sampling probability [19]:

$$\hat{p}_{comp} = \frac{\sum\limits_{v_i \in U \cap A} {\pi_i}^{-1}}{\sum\limits_{v_j \in U} {\pi_j}^{-1}}. \tag{3}$$

Admittedly, knowing the complete network information is a very unrealistic scenario, and we present it here as it will serve as a gold standard compared to other estimators based on more realistically available information.

1.2 Observing the indegree in the RDS sample

We now relax the unrealistic assumption that the complete network structure is observed and instead assume knowledge only about the individuals sampled in RDS. More specifically we here assume that, besides observing whether the sampled individuals have property A or not, we also observe their *indegree*.

For RDS on *undirected* networks, the (stationary) sampling probability of a node is proportional to its degree. In a directed network one could perhaps expect that the corresponding result was true if we simply replace "degree" by "indegree", since it makes sense that a node having many other nodes "pointing" to it should have a higher chance of being sampled. This is not true in general, however, if degree is not correlated between adjacent nodes in the network, i.e., the in/out-degree of the two nodes connected by a directed link are independent to each other, then the sampling probability for a node is proportional to its indegree under mean field approximation (see [20]):

$$\bar{\pi}(d_i^{in}) \sim \frac{d_i^{in}}{N\bar{D}}, \tag{4}$$

where $\bar{D}$ is the average degree of the network. (4) implies that for networks with no degree correlation, the RDS sample can be weighted by respondents' indegrees to estimate population proportions:

$$\hat{p}_{indeg} = \frac{\sum\limits_{v_i \in U \cap A} (d_i^{in})^{-1}}{\sum\limits_{v_j \in U} (d_j^{in})^{-1}}. \tag{5}$$

This estimator will in general not be consistent since most social networks have positive degree correlations [21–23]. However, for many networks it performs well as we will see in the simulations.

1.3 Observing the indegree and outdegree in the RDS sample

Combining indegree with other network data collected in the current RDS implementations, i.e., the outdegree and recruitment information, it is possible to derive another estimator. Recall that for any group in a network, the sum of indegree always equals the sum of outdegree, i.e.,

$$\begin{cases} N_A\bar{D}_A^{out}S_{AA} + N_B\bar{D}_B^{out}S_{BA} = N_A\bar{D}_A^{in} \\ N_A\bar{D}_A^{out}S_{AB} + N_B\bar{D}_B^{out}S_{BB} = N_B\bar{D}_B^{in} \end{cases}, \tag{6}$$

where $\{S_{XY}, X, Y \in \{A, B\}\}$ is the recruitment matrix representing the proportion of edges going from group X to group Y.

For simplicity, let $m^* = \frac{\bar{D}_A^{in}}{\bar{D}_B^{in}}$ and $w^* = \frac{\bar{D}_A^{out}}{\bar{D}_B^{out}}$ be the average indegree and outdegree ratio of the two groups of nodes in the network, and let $\phi = \frac{N_A}{N_B}$ be the relative group size proportion. Dividing the above equation gives a solution of ϕ:

$$\phi = \frac{w^*S_{AA} - m^*S_{BB}}{2m^*w^*S_{AB}} + \sqrt{\frac{S_{BA}}{m^*w^*S_{AB}} + (\frac{m^*S_{BB} - w^*S_{AA}}{2m^*w^*S_{AB}})^2}. \tag{7}$$

Then, if we can correctly estimate m^*, w^* and S, we obtain another estimator:

$$\hat{p}_{degsum} = \frac{\hat{\phi}}{1 + \hat{\phi}}, \tag{8}$$

in which we replace unknown population quantities in ϕ by their estimates from the RDS sample.

From the previous section, the average indegree ratio m^* can be estimated by the harmonic mean ratio of indegrees from the sample for networks with no degree correlation: $\hat{m}^* = \frac{n_A / \sum_{v_i \in U \cap A} (d_i^{in})^{-1}}{n_B / \sum_{v_i \in U \cap B} (d_i^{in})^{-1}}$. It is however generally not possible to consistently estimate w^* and S using only the average outdegree and observed recruitment matrix. The sample mean outdegree will be an unbiased estimator only if there is no dependence (correlation) between the indegree and outdegree of nodes, while the harmonic mean of outdegree is expected to have higher precision if the indegree-outdegree correlation is high. Consequently, we have included two degree sum-based estimators, $\hat{p}^M_{degsum}$ and $\hat{p}^H_{degsum}$, in which w^* is estimated by: $\hat{w}^* = \frac{\sum_{v_i \in U \cap A} d_i^{out} / n_A}{\sum_{v_i \in U \cap B} d_i^{out} / n_B}$ and $\hat{w}^* = \frac{n_A / \sum_{v_i \in U \cap A} (d_i^{out})^{-1}}{n_B / \sum_{v_i \in U \cap B} (d_i^{out})^{-1}}$, respectively (the superscripts M and H indicate whether the (arithmetic) mean or the harmonic mean of outdegree has been used to estimate w^*). We have also tried to adjust potential bias in the estimation of S by replacing individual inclusion probabilities with group inclusion probabilities (see Supportive Information (SI) for details), which however didn't improve the results and we therefore prefer to use the observed recruitment matrix from the sample to estimate S in (7).

The factor w^* was named the *activity ratio* in [13], since it quantifies how active nodes in different groups are in building their personal networks. Following this, we henceforth refer to m^* as the *attractivity ratio*, as it reflects how "attractive" nodes in different groups are, or to which group of nodes edges are inclined to form.

1.4 Sensitivity analysis when indegree is not known

Hardly ever is the indegree observed in RDS studies. Consequently, the use of $\hat{p}_{indeg}$ and $\hat{p}_{degsum}$ is limited in practice. For $\hat{p}^M_{degsum}$ and $\hat{p}^H_{degsum}$, if the indegree is not known, the estimate of average indegree ratio, $\hat{m}^*$, becomes an unknown parameter in (8). If prior information about m^* is available, these two estimators can still be used to provide valid estimates. Prior information may, for example, be obtained by expert opinions, or by using previous empirical results. What's more, even if there is little prior knowledge about the targeted population, we can, instead of providing a point estimate with fixed parameters, use a range of m values to generate an estimate interval for p_A. That is, if m^* is assumed to lie within a certain range, $[m_{min}, m_{max}]$, we get an interval of $\hat{p}_A$, $[\hat{p}_A(m_{min}), \hat{p}_A(m_{max})]$, by varying m in (8).

The above method can also be applied with $\hat{p}_{indeg}$, since we can rewrite (5) as:

$$\hat{p}_{indeg} = \sum_{v_i \in U \cap A} (d_i^{in})^{-1} / \sum_{v_j \in U} (d_j^{in})^{-1} = \frac{n_A/\hat{\hat{D}}_A^{in}}{n_A/\hat{\hat{D}}_A^{in} + n_B/\hat{\hat{D}}_B^{in}} = \frac{n_A/n_B}{n_A/n_B + \hat{\hat{D}}_A^{in}/\hat{\hat{D}}_B^{in}}.$$

Replacing $\hat{\hat{D}}_A^{in}/\hat{\hat{D}}_B^{in}$ with m, we have:

$$\hat{p}_{indeg}(m) = \frac{n_A/n_B}{n_A/n_B + m}. \tag{9}$$

Following this, we will perform a sensitivity analysis on $\hat{p}^M_{degsum}(m)$, $\hat{p}^H_{degsum}(m)$, and $\hat{p}_{indeg}(m)$. We will vary the ratio of average indegrees m to get an interval of the various estimators of p_A, $[\hat{p}_A(m_{min}), \hat{p}_A(m_{max})]$, with m lying in a certain range, $[m_{min}, m_{max}]$. By choosing an interval centered on a value of m based on prior information, we will get intervals of possible p_A values which more fully accounts for the situation when the network is directed, and provides valuable results on the sensitivity of estimators to correctness of indegree assumptions about the network.

2 Network Data and Study Design

2.1 Network parameters

To evaluate how variation in network structure can affect the precision of our estimators, we generate sets of networks with different structures by varying network parameters: *Directedness* (λ, the proportion of irreciprocal edges in the network); *Indegree correlation* (γ, quantified by the indegree-based assortativity as defined in [21]); *Indegree-outdegree correlation* (ρ, the Pearson correlation of indegree and outdegree); *Homophily* (h_A, the probability that nodes connect with neighbors that are similar to themselves with respect to the studied feature A rather than that they connect randomly [8, 24–26]). The activity ratio w^*, as well as the attractivity ratio m^*, are also used as network structure parameters in our assessment. For further explanations of these parameters see SI.

2.2 Network Data

We use both artificially generated families of networks (**Net1** and **Net2**), and an MSM online social network [19] in our evaluation. Net1 is generated starting from a random pure directed network, in which indegree and outdegree are uncorrelated ($\rho \approx 0$); then, the irreciprocal edges are rewired in a particular way that doesn't change the nodes' degree. Networks with different levels of directedness are generated (down to $\lambda = 0.2$), while the indegree-outdegree correlations remain unchanged. Then, nodes are assigned either property A or B in order to achieve different attractivity ratios $m^* \in [0.7,\ 1.4]$ (see Table 1). In Net2, a random undirected network is generated; then, to obtain directedness, reciprocal edges are randomly rewired in such a way, that for any network in Net2 with directedness λ, the indegree-outdegree correlation is $\rho \approx 1 - \lambda$. Different attractivity ratios are generated as for Net1. Finally, we further rewire edges with respect to the nodes' properties in order to achieve different levels of homophily: $h_A \in [0, 0.5]$ (see Table 1).

The anonymized MSM social network used in this study (previously analyzed in [19, 27]) comes from the Nordic region's largest and most active web community for homosexual, bisexual, transgender, and queer persons (www.qruiser.com). Contacts between members on the web site are maintained by a "favorites list", on which each member can add any other member without approval from that member. From this network, we obtain the giant strongly connected component of the friendship network of those members who identify themselves as homosexual males. 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 proportions of nodes having a specific value of these properties are listed in Table 1.

Based on the directed MSM network, we use a shuffling method, slightly different from what was described in [28], to generate networks with different levels of indegree correlation (**Net3**). Detailed information on the generation process of the above networks can be found in the supportive information SI.

RDS processes are simulated on these networks, and estimates of population properties are calculated using the proposed new estimators $\hat{p}_{indeg}$, $\hat{p}^{M}_{degsum}$, $\hat{p}^{H}_{degsum}$, along with the simple sample mean $\hat{p}$ and $\hat{p}_{outdeg}$ ([1]). In each simulation, seeds are uniformly selected and coupons are randomly distributed to the recruiters' neighbors. Both the sampling without replacement and sampling with replacement schemes are used, separately. We choose sample sizes up to 500 for Net1 and Net2, and 1000 for the MSM network and Net3. All simulations are repeated 10,000 times.

3 Results

3.1 Performance of $\hat{p}_{indeg}$ and $\hat{p}_{degsum}$

Net1 and Net2. Simulations were first performed on the artificial networks Net1 and Net2, and the root mean square error (RMSE) of $\hat{p}$, $\hat{p}_{outdeg}$, $\hat{p}_{indeg}$ and $\hat{p}^{M}_{degsum}$

are shown in Fig. 1 (the estimates of $\hat{p}^H_{degsum}$ are very similar to $\hat{p}^M_{degsum}$ and thus omitted here).

We see that the estimator $\hat{p}_{outdeg}$ is no longer unbiased when the networks are directed, i.e., $\lambda > 0$, and that the RMSE increases with increasing directedness and $|m^* - 1|$. For networks without indegree-outdegree correlation (Net1, see Fig. 1(a)), $\hat{p}_{outdeg}$ produces the same level of error as the sample mean $\hat{p}$; for networks with indegree-outdegree correlation (Net2, see Fig. 1(b), (c)), $\hat{p}_{outdeg}$ produces less error than that of $\hat{p}$, but larger than that of $\hat{p}_{indeg}$ and $\hat{p}^M_{degsum}$, indicating that the use of current RDS estimators for networks with unknown directedness could generate large errors.

By Contrast, utilizing the sampled individuals' indegree information, both $\hat{p}_{indeg}$ and $\hat{p}^M_{degsum}$ generate small errors, and show a more stable behavior of the RMSE than $\hat{p}_{outdeg}$. For simulations of RDS on Net1 and Net2 with one seed, one coupon and sampling with replacement, the conclusions are similar (see SI.fig5).

The MSM network. The same setup as above has been used for the MSM network simulations. The estimators are displayed as box plots in the left panel of Fig. 2. In each box, the central line is the median, the dot is the mean, the edges of the box are the 25th (q_1) and 75th (q_3) percentiles. Estimates $1.5(q_3 - q_1)$ away from the box are shown as outliers beyond the whiskers.

For variables with large homophily and attractivity ratios which significantly differ from 1 (age, county), $\hat{p}_{outdeg}$ has a large bias. For example, its estimates of the proportion of MSM members who live in Stockholm are on average 5.7 percentage units higher than the true value, and for age, civil status and profession, the sample mean, $\hat{p}$, has even less bias than $\hat{p}_{outdeg}$. For variables with limited homophily and difference in mean indegree (civil status, profession), the differences between $\hat{p}_{indeg}$, $\hat{p}^M_{degsum}$ and $\hat{p}^H_{degsum}$ are negligible.

As the indegree correlation is low in the MSM network, $\gamma = 0.03$, we can assume that the no indegree correlation assumption for $\hat{p}_{indeg}$ and $\hat{p}_{degsum}$ is met. They both have very low bias for all four variables; the differences between average estimates and the true population proportions are within 0.5 percentage units, indicating that for networks with small indegree correlation, the indegree can be a good approximation of the inclusion probabilities in the RDS process. Since the indegree and outdegree of nodes are positively correlated ($\rho = 0.39$), the use of the harmonic mean for estimation of the average outdegree in $\hat{p}^H_{degsum}$ provides estimates with slightly better properties than those that use the simple mean ($\hat{p}^M_{degsum}$).

Note that in Fig. 2 we use 10 seeds and 3 coupons to mimic the real practice of RDS; such a setting makes the number of waves needed to reach a sample size of 1000 no more than 5 (sampling with replacement). However, when we use only one seed and coupon in the simulation, much longer waves are needed and consequently, the inclusion probabilities of nodes might be more affected by other structural properties of the network such as transitivity, clusters, etc. These effects are reflected in SI.fig6, where the proposed new estimators still outperforms $\hat{p}_{outdeg}$, and $\hat{p}^{M}_{degsum}$ and $\hat{p}^{H}_{degsum}$ have the smallest bias and error for all variables, and $\hat{p}_{indeg}$ generates limited bias for age and county.

In summary, simulations on the MSM network reveal that $\hat{p}_{indeg}$, $\hat{p}^{M}_{degsum}$ and $\hat{p}^{H}_{degsum}$ generate the smallest bias and error among the estimators. When the sample size is reached by short recruitment waves, $\hat{p}_{indeg}$ has the best performance, while on the other side, $\hat{p}^{M}_{degsum}$ and $\hat{p}^{H}_{degsum}$ generate less bias and error when the recruitment chains are long.

Net3. The above studies reveal that the indegree-based estimators work well on directed networks with low indegree correlation. To test how the violation of this assumption would affect the precision of these estimators we use RDS simulations on indegree correlated networks, Net3. The left panel of Fig. 3 shows the simulation results on Net3 with $\gamma = 0.4$.

Surprisingly, no obvious difference is found with the results seen for the MSM network. The average estimates of $\hat{p}$ and $\hat{p}_{outdeg}$ are almost identical to those shown in Fig. 2, and the differences between $\hat{p}_{indeg}$, $\hat{p}^{M}_{degsum}$ and $\hat{p}^{H}_{degsum}$ are also small. A closer look at the results also shows that the standard errors of all estimators are smaller than before. Simulations with one seed and coupon, and sampling with replacement (see SI.fig7) also lead to the same results. This reveals that, for networks generated in this paper, the indegree-based estimators are quite robust to changes in indegree correlations.

3.2 Sensitivity analysis with $\hat{p}_{indeg}(m)$ and $\hat{p}_{degsum}(m)$

Currently, most RDS practices can only provide a point estimated population property from the sample (based on $\hat{p}_{outdeg}$). It is thus important for researchers and policy makers to know how sensitive this point estimate is to variations in the ideal assumptions. According to (5) and (8), we can see how robust $\hat{p}_{indeg}$ and $\hat{p}_{degsum}$ are to assumptions about unobservable indegree properties of the network, by varying m within a range of plausible values.

Net1 and Net2. In order to perform sensitivity analysis, we let the attractivity ratio m used in $\hat{p}_{indeg}(m)$ and $\hat{p}^{M}_{degsum}(m)$ be fixed at different values ($m = [0.7,\ 1.4]$), and check how the RMSE changes along the domain of combinations of directedness and attractivity ratio (see Fig. 4). The RMSE of $\hat{p}^{M}_{degsum}(m)$ is almost the same as for $\hat{p}_{indeg}(m)$ in Net1 and is thus not shown in Fig. 4(a).

Obviously, the RMSE of $\hat{p}_{indeg}(m)$ and $\hat{p}^{M}_{degsum}(m)$ become larger when the m value deviates far from the true population ratio m^{*}, and minimal when m approaches m^{*}, telling us that with appropriate prior information about the studied population, it is possible to estimate population properties with very small error using the proposed estimators.

It's also worth noting that the RMSE does not increase with increased directedness for networks with the same m^* value, indicating that the RMSE of $\hat{p}_{indeg}(m)$ or $\hat{p}^M_{degsum}(m)$ would be the same as long as $|m - m^*|$ is the same, regardless of directedness of networks.

When homophily is high and indegree and outdegree are correlated, further deviation of m from m^* produce larger errors in $\hat{p}^M_{degsum}(m)$ than $\hat{p}_{indeg}(m)$, as shown in Fig. 4(c).

The MSM network and Net3. The right panels in Fig. 2 and Fig. 3 show the sensitivity analysis of $\hat{p}_{indeg}(m)$ and $\hat{p}^M_{degsum}(m)$ with m varying from 0.7 to 1.4, for the MSM network and Net3, respectively.

From (9) it is clear that $\hat{p}_{indeg}(m)$ decreases in m, and that its curve of average estimates with varying m values intersects with the true population line closely at m^* on both networks. For high homophily variables such as age and county, $\hat{p}_{indeg}(m)$ has less bias than $\hat{p}^M_{degsum}(m)$ when m is far from m^*, similar to the comparison made on Net2. On the other hand, $\hat{p}_{indeg}(m)$ and $\hat{p}^M_{degsum}(m)$ are almost the same within the interval of m for variables with low homophily, i.e., civil status and profession.

The magnitude of changes of average estimates are however not dependent on the homophily, e.g., in the sensitivity analysis on the MSM network, when m is varied from 0.7 to 1.4, the change in average estimates of $\hat{p}_{indeg}(m)$ are 12 percentage units for age, and 17 percentage units for civil status, respectively.

Sensitivity analysis of simulations with one seed and coupon, sampling with replacement, are also produced on all networks and can be found in SI.fig8; the conclusions are similar to those above.

4 Conclusion and Discussion

Despite the widely acknowledged evidence for the existence of directedness among social networks, the effect of directedness on RDS estimates has seldom been evaluated. This could be problematic since all previous reported RDS estimates rely on the assumption that the studied networks are purely reciprocal, the violation of which will result in unknown bias. To address this situation, we have proposed several estimators for RDS on directed networks, assuming different available information about nodes in the sample.

Given respondents' indegree, $\hat{p}_{indeg}$, $\hat{p}^M_{degsum}$, and $\hat{p}^H_{degsum}$ all outperform the current widely used $\hat{p}_{outdeg}$ for networks with certain amounts of directedness. The proposed estimators show strong resistance to variations in directedness, indegree correlation, indegree-outdegree correlation, homophily and attractivity ratio. When the recruitment chains are short in the sample, $\hat{p}_{indeg}$ works slightly better, while $\hat{p}^M_{degsum}$ and $\hat{p}^H_{degsum}$ perform better when the recruitment chains are long. When the indegree-outdegree correlation is positive (the common situation), $\hat{p}^H_{degsum}$ further provides less bias and error than $\hat{p}^M_{degsum}$. Whenever respondents' indegrees are known, our proposed estimators are expected to produce estimates with less bias and error in directed networks.

Since the indegree cannot usually be collected with RDS, we propose a sensitivity analysis method in order to take into account uncertainties about unobservable indegree properties of the network. The method is based on assumptions about one single parameter: the attractivity ratio m being the harmonic indegree mean ratio of the two groups in the network (having property A or not). Uncertainty assumptions about m induces estimate intervals of p_A, the community fraction having property A. Prior information about m may possibly be utilized and may be obtained by expert opinions, or by using previous empirical studies related to the studied population. Additionally, since the indegree-outdegree correlation is positive in most social networks [16–18], the activity ratio $\hat{w}^*$ observed from the sample may be an indicator of where to vary m from. For example, the difference between m^* and w^* in the MSM network, is 0.27, 0.17, 0.02, and 0.05 for age, county, civil status and profession, respectively [19]. As an illustration on how to implement the sensitivity analysis method in RDS practice when the indegree is not collected, we take data given in [29] to make the proposed sensitivity analysis using $\hat{p}_{indeg}(m)$ and $\hat{p}^H_{degsum}(m)$ (Note here we use the harmonic means of outdegrees in $\hat{p}^H_{degsum}(m)$ instead of $\hat{p}^M_{degsum}(m)$, owing to the limit of data provided). A sample of 618 drug users in New York City were collected using RDS with eight seeds. We use the same data to produce estimates on the proportion of males and injectors among drug users in New York City. The activity ratio ($\hat{w}^*$, weighted) for males is 0.99, indicating that there is little difference of the size of personal networks with respect to gender. However, the activity ratio for injectors is 1.58, indicating that injecting drug users know 58% more drug users than those who don't inject drugs. We thus vary the m values around the observed activity ratio to perform the sensitivity analysis; the length of the test interval is arbitrarily set to 1. In Fig. 5, we can see that when $m = \hat{w}^*$, the $\hat{p}_{indeg}(m)$ and $\hat{p}^H_{degsum}(m)$ estimates are equal to the estimates given by $\hat{p}_{outdeg}$. When the network is assumed directed and $m \in [0.5, 1.5]$, the estimated proportion of male drug users will vary from 0.88 to 0.66 for $\hat{p}_{indeg}(m)$ and 0.87 to 0.68 for $\hat{p}^H_{degsum}(m)$, respectively. The proportion of injecting drug users, varies from 0.45 to 0.62 for $\hat{p}_{indeg}(m)$ and 0.41 to 0.64 for $\hat{p}^H_{degsum}(m)$ when $m \in [1, 2]$. It is most likely possible to reduce the length of the intervals of p_A values, at least with prior information on the population; this is also subject to future research.

In conclusion, we can see that the sensitivity analysis provides means for RDS practitioners to understand the robustness of sample inference to the violation of certain assumptions: that the network may be partially directed, and that the degree data collected from respondents may contain reporting error. Clearly, for a studied network with unknown directedness, it is better to report an interval of estimates based on a range of m values, since it provides a more detailed image of the situation and more reasonable advice on how to understand the studied population and make policy. We henceforth recommend the new estimator-based sensitivity analysis method to be used in future RDS studies.

Acknowledgment

The authors thank Prof. Kimmo Eriksson for helpful discussions. This work is funded in part by the Riksbankens Jubileumsfond (dnr: P2008-0674), and X.L. would like to thank China Scholarship Council (Grant No. 2008611091). Thanks are due to Sida for their support to RDS development work in Vietnam.

References

[1] U. J. P. on HIV/AIDS, Global report: Unaids report on the global aids epidemic 2010, tech. rep., 2010.

[2] M. Malekinejad, L. G. Johnston, C. Kendall, L. Kerr, M. R. Rifkin, and G.W. Rutherford, Using respondent-driven sampling methodology for hiv biological and behavioral surveillance in international settings: A systematic review, *Aids and Behavior*, vol. 12, no. 4, pp. S105-S130, 2008.

[3] S. Goel and M. J. Salganik, Respondent-driven sampling as markov chain monte carlo, *Statistics in Medicine*, vol. 28, no. 17, pp. 2202-2229, 2009.

[4] E. Deaux and J. Callaghan, Key informant versus self-report estimates of health behavior, *Evaluation Rev*, vol. 9, pp. 365-368, 1985.

[5] J. K. Watters and P. Biernacki, Targeted sampling: Options for the study of hidden populations, *Social Problems*, vol. 36, no. 4, pp. 416-430, 1989.

[6] B. H. Erickson, Some problems of inference from chain data, *Sociological Methodology*, vol. 10, pp. 276-302, 1979.

[7] D. D. Heckathorn, Respondent-driven sampling: A new approach to the study of hidden populations, *Social Problems*, vol. 44, no. 2, pp. 174-199, 1997.

[8] D. D. Heckathorn, Respondent-driven sampling ii: Deriving valid population estimates from chain-referral samples of hidden populations, *Social Problems*, vol. 49, no. 1, pp. 11-34, 2002.

[9] E. Volz and D. D. Heckathorn, Probability based estimation theory for respondent driven sampling, *Journal of Official Statistics*, vol. 24, no. 1, pp. 79-97, 2008.

[10] L. G. Johnston, M. Malekinejad, C. Kendall, I. M. Iuppa, and G. W. Rutherford, Implementation challenges to using respondent-driven sampling methodology for hiv biological and behavioral surveillance: Field experiences in international settings, *Aids and Behavior*, vol. 12, no. 4, pp. S131-S141, 2008.

[11] M. J. Salganik and D. D. Heckathorn, Sampling and estimation in hidden populations using respondent-driven sampling, *Sociological Methodology*, vol. 34, pp. 193-239, Weinheim: Wiley-V C H Verlag Gmbh, 2004.

[12] S. Goel and M. J. Salganik, Assessing respondent-driven sampling, *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, no. 15, pp. 6743-6747, 2010.

[13] K. J. Gile and M. S. Handcock, Respondent-driven sampling: An assessment of current methodology, *Sociological Methodology*, pp. no. doi: 10.1111/j.1467-9531.2010.01223.x, 2010.

[14] A. Tomas and K. J. Gile, The effect of differential recruitment, non-response and non-recruitment on estimators for respondent-driven sampling, *http://arxiv.org/abs/1012.4122*, 2010.

[15] K. J. Gile, Improved inference for respondent-driven sampling data with application to hiv prevalence estimation, *Journal of the American Statistical Association*, vol. 106, no. 493, pp. 135-146, 2011.

[16] S. J. South and D. L. Haynie, Friendship networks of mobile adolescents, *Social Forces*, vol. 83, no. 1, pp. 315-350, 2004.

[17] W. L. Wallace, Student culture: Social structure and continuity in a liberal arts college. Chicago: Aldine Publishing Company, 1966.

[18] S. L. Feld and W. C. Carter, Detecting measurement bias in respondent reports of personal networks, *Social Networks*, vol. 24, no. 4, pp. 365-383, 2002.

[19] X. Lu, L. Bengtsson, T. Britton, M. Camitz, B. J. Kim, A. Thorson, and F. Liljeros, The sensitivity of respondent-driven sampling, *Journal of the Royal Statistical Society: Series A*, early view accessed online, http://dx.doi.org/10.1111/j.1467-985X.2011.00711.x, 2011.

[20] S. Fortunato, M. Boguñá, A. Flammini, and F. Menczer, Algorithms and models for the web-graph, ch. Approximating PageRank from In-Degree, pp. 59-71, Berlin, Heidelberg: Springer-Verlag, 2008.

[21] M. E. J. Newman, Assortative mixing in networks, *Phys. Rev. Lett.*, vol. 89, p. 208701, Oct 2002.

[22] M. E. J. Newman, Mixing patterns in networks, *Phys. Rev. E*, vol. 67, p. 026126, Feb 2003.

[23] J. G. Foster, D. V. Foster, P. Grassberger, and M. Paczuski, Edge direction and the structure of networks, *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, pp. 10815-10820, JUN 15 2010.

[24] M. Morris and M. Kretzschmar, Concurrent partnerships and transmission dynamic in networks, *Social Networks*, vol. 17, no. 3-4, pp. 299-318, 1995.

[25] A. Rapoport, A probabilistic approach to networks, *Social Networks*, vol. 2, pp. 1-18, 1980.

[26] M. McPherson, L. Smith-Lovin, and J. M. Cook, Birds of a feather: Homophily in social networks, *Annual Review of Sociology*, vol. 27, pp. 415-444, 2001.

[27] D. Rybski, S. Buldyrev, S. Havlin, F. Liljeros, and H. Makse, Scaling laws of human interaction activity, *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, no. 31, pp. 12640-12645, 2009.

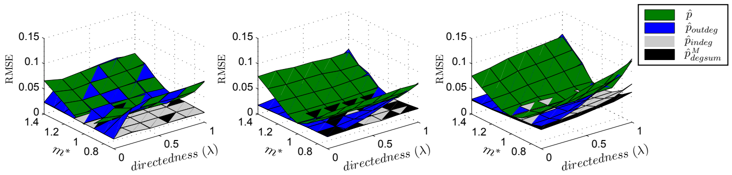
[28] R. Xulvi-Brunet and I. M. Sokolov, Reshuffling scale-free networks: From random to assortative, *Physical Review E*, vol. 70, no. 6, p. 066102, 2004.

[29] A. S. Abdul-Quader, D. D. Heckathorn, C. McKnight, H. Bramson, C. Nemeth, K. Sabin, K. Gallagher, and D. C. Des Jarlais, Effectiveness of respondent-driven sampling for recruiting drug users in new york city: Findings from a pilot study, *Journal of Urban Health-Bulletin of the New York Academy of Medicine*, vol. 83, no. 3, pp. 459-476, 2006.

Table 1: Basic statistics of Net1, Net2, Net3 and the MSM network

	Network size (N)	Average degree ($\bar{D}$)	Directed-ness (λ)	indegree corre-lation (γ)	indegree-outdegree correlation (ρ)		Homophily (h)	Attractivity ratio (m^*)	P
Net1	$10,000$	10	$[0,1]$	$[-0.09,0.01]$	≈ 0		$[-0.30,0.22]$	$[0.7,1.4]$	70%
Net2	$10,000$	10	$[0,1]$	$[-0.03,0.14]$	$\approx 1-\lambda$		$[0,0.5]$	$[0.7,1.4]$	30%
MSM Network	$16,082$	17.2	0.61	0.03	0.39	*age*	0.23	1.22	77%
						ct	0.50	1.15	39%
						cs	0.03	0.98	40%
						pf	0.06	1.10	38%
Net3	$--$[*]	$--$	$[0.61,0.91]$	$[0,0.4]$	$--$		$--$	$--$	$--$

[*] Same as the MSM network



(a) (b) (c)

Figure 1: Root Mean Square Error of RDS estimators on Net1 and Net2. (a) Net1; (b) Net2 with homophily $h_A = 0$; (c) Net2 with homophily $h_A = 0.4$. Sampling without replacement, number of seeds=10, coupons=3, sample size=500.

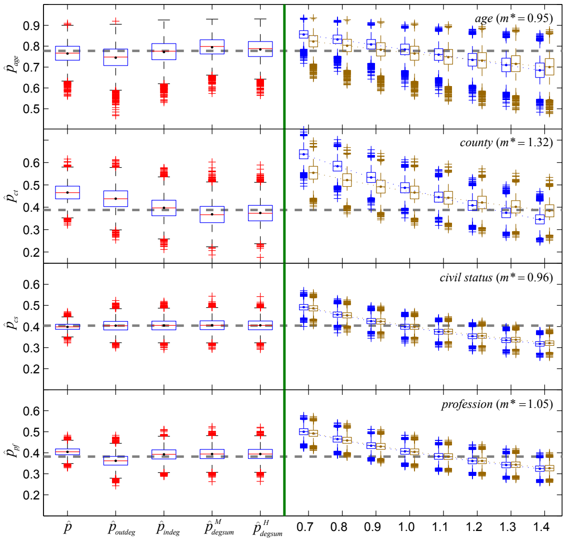


Figure 2: RDS on MSM network. The right panel shows sensitivity analysis of $\hat{p}_{indeg}(m)$ (brown) and $\hat{p}^{M}_{degsum}(m)$ (blue) with m varying from 0.7 to 1.4, plots are horizontally shifted a few points to avoid overlapping. Sampling with replacement, number of seeds=10, number of coupons=3, sample size=1000.

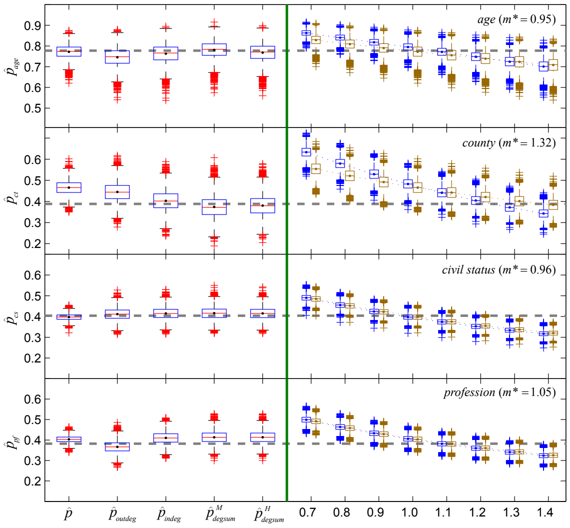


Figure 3: RDS on Net3 with indegree correlation $\gamma = 0.4$. The right panel shows sensitivity analysis of $\hat{p}_{indeg}(m)$ (brown) and $\hat{p}^M_{degsum}(m)$ (blue) with m varying from 0.7 to 1.4, plots are horizontally shifted a few points to avoid overlapping. Sampling without replacement, number of seeds=10, number of coupons=3, sample size=1000.

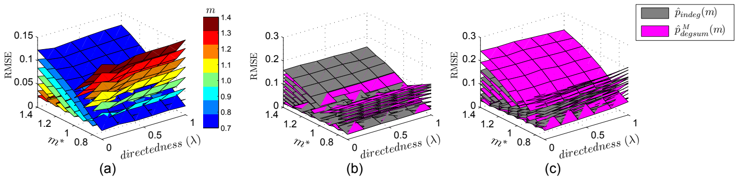


Figure 4: Sensitivity analysis of $\hat{p}_{indeg}(m)$ and $\hat{p}^M_{degsum}(m)$ on Net1 and Net2 with tested m values. (a) Net1, $\hat{p}^M_{degsum}(m)$ not shown as it is similar to $\hat{p}_{indeg}(m)$; (b) Net2 with homophily $h_A = 0$; (c) Net2 with homophily $h_A = 0.4$. Sampling without replacement, number of seeds=10, coupons=3, sample size=500.

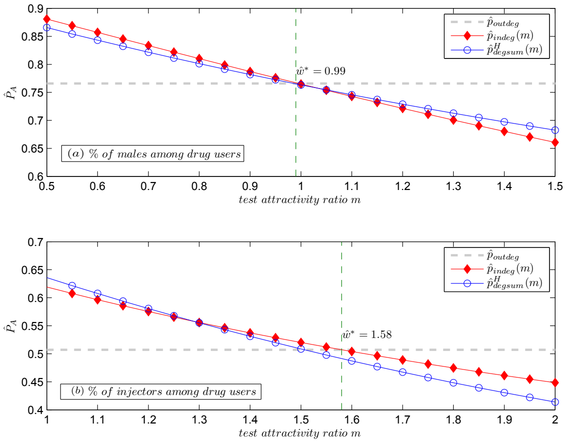


Figure 5: Sensitivity analysis of RDS estimates for proportion of (a) males and (b) injectors among drug users in New York City.