DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 32, No. 2, 2016, pp. 433–459, http://dx.doi.org/10.1515/JOS-2016-0023

Random Walks on Directed Networks: Inference and Respondent-Driven Sampling

Jens Malmros[1], *Naoki Masuda*[2], *and Tom Britton*[3]

Respondent-driven sampling (RDS) is often used to estimate population properties (e.g., sexual risk behavior) in hard-to-reach populations. In RDS, already sampled individuals recruit population members to the sample from their social contacts in an efficient snowball-like sampling procedure. By assuming a Markov model for the recruitment of individuals, asymptotically unbiased estimates of population characteristics can be obtained. Current RDS estimation methodology assumes that the social network is undirected, that is, all edges are reciprocal. However, empirical social networks in general also include a substantial number of nonreciprocal edges. In this article, we develop an estimation method for RDS in populations connected by social networks that include reciprocal and nonreciprocal edges. We derive estimators of the selection probabilities of individuals as a function of the number of outgoing edges of sampled individuals. The proposed estimators are evaluated on artificial and empirical networks and are shown to generally perform better than existing estimators. This is the case in particular when the fraction of directed edges in the network is large.

Key words: Hidden population; social network; nonreciprocal relationship; Markov model.

1. Introduction

Hidden or hard-to-reach populations include several groups of importance to public health research, for example, men who have sex with men (MSM), sex workers (SWs), and injecting drug users (IDUs). A hidden population is typically characterized by i) strong privacy concerns due to illicit or stigmatized behavior, and ii) there is no sampling frame, that is, the size and composition of the population are unknown (Heckathorn 1997). Therefore, it is in general difficult for survey researchers to access hidden populations and draw valid conclusions from sampled data. Several methods have been used to sample from hidden populations, for example, key informant sampling (Deaux and Callaghan 1985), venue-based sampling (Muhib et al. 2001), and snowball sampling (Erickson 1979). However, because of the substantial selection bias inherent in these methods, the

[1] Department of Mathematics, Stockholm University, SE-106 91 Stockholm, Sweden. Email: jensm@math.su.se (corresponding author)
[2] Department of Mathematical Informatics, The University of Tokyo, 7-3-1 Hongo, Bunkyo, Tokyo 113-8656, Japan. Email: naoki.masuda@bristol.ac.uk
[3] Department of Engineering Mathematics, University of Bristol, Merchant Venturers Building, Woodland Road, Clifton, Bristol BS8 1UB, United Kingdom. Email: tomb@math.su.se

Acknowledgments: J. M. was supported by Grant No. 2009-5759 from the Swedish Research Council. N. M. was supported by Grants-in-Aid for Scientific Research (No. 23681033) from MEXT, Japan, the Nakajima Foundation, and the Aihara Project, the FIRST program from JSPS, initiated by CSTP, Japan. T. B. is grateful to Riksbankens Jubileumsfond (contract P12-0705:1) for financial support. The authors would like to thank Prof. Fredrik Liljeros and Dr. Xin Lu, Department of Sociology, Stockholm University, for use of the Qruiser dataset.

samples obtained have been considered only for convenience purposes (Magnani et al. 2005). Respondent-driven sampling (RDS) is a more recent sampling methodology for hidden populations (Heckathorn 1997; Salganik and Heckathorn 2004; Volz and Heckathorn 2008). RDS combines an improved link-tracing sampling mechanism, similar to snowball sampling, with a mathematical model that is able to produce asymptotically unbiased estimates of population characteristics given that some assumptions about the sampling procedure are fulfilled. Because of these advantages, RDS has become the primary choice for the study of hidden populations. Some recent examples of RDS studies includes MSM in Panama (Hakre et al. 2014), Dar es Salaam, Tanzania (Bui et al. 2014), SWs in Shiraz, Iran (Kazerooni et al. 2013), Kampala, Uganda (Schwitters et al. 2014), IDUs throughout India (Solomon et al. 2015), methamphetamine users in Cape Town, South Africa (Hobkirk et al. 2015), unauthorized migrant workers in San Diego (Zhang et al. 2014), and low-wage workers in US cities (Bernhardt et al. 2013).

In RDS, the social network of the population is used both in the sampling procedure and for inference. Before we describe RDS in more detail, we will introduce some concepts from social network theory (for a comprehensive reference on social network theory, see Wasserman and Faust 1994). Formally, a *social network* is a (finite) set of actors, for example, individuals, couples, or organizations, that are connected through some type of relation, for example, friendship, kinship, or professional agreements. In graph-theoretical terms, the actors are referred to as *vertices* and their relations as *edges*. The relation between two actors can be reciprocal, that is, the relation is mutual, or it can be nonreciprocal. For example, an individual may choose another individual as a friend. If the other individual in turn chooses the first individual as a friend, the relation is reciprocal, and if that individual does not choose the first one, the relation is nonreciprocal. A reciprocal edge is called an *undirected edge* and a nonreciprocal edge is called a *directed edge*. A network in which the directions of edges are ignored is referred to as an *undirected network*. A network in which the directions of edges are meaningful is referred to as a *directed network*. Note that a directed network may include nonreciprocal and reciprocal edges. In Figure 1, we see three nonreciprocal edges.

One might also consider individual properties of the vertices in the network. The *neighbors* of a vertex are the set of vertices to which it connects by an edge. In Figure 1, the neighbors of vertex v includes all vertices except one to the lower left. The *degree* of a vertex refers to the number of neighbors it has in an undirected network. If we ignore the directions of edges in Figure 1, vertex v has degree four. If the network is directed, one

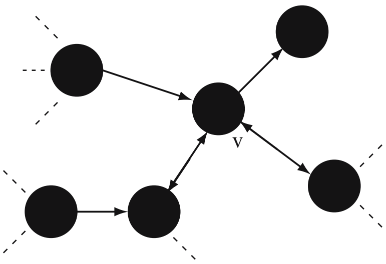


Fig. 1. Sample illustration of a part of a directed social network with six vertices and thirten edges. Vertex v has an undirected degree of two, an incoming degree of one, and an outgoing degree of one.

must consider the directions of edges. A directed edge is either *incoming* to or *outgoing* from a vertex. Because a directed network also may include reciprocal edges, one can identify three types of edges for a vertex w in a directed network, and hence three different degrees: the *undirected degree*, which refers to the number of vertices w connects to by an undirected edge, the *incoming degree*, which refers to the number of vertices w connects to by an incoming edge, and the *outgoing degree*, which refers to the number of vertices w connects to by an outgoing edge. In Figure 1, vertex v has an undirected degree of two, an incoming degree of one, and an outgoing degree of one. The *out-degree* of a vertex is obtained by adding the undirected degree and the outgoing degree. Similarly, the *in-degree* is obtained by adding the undirected degree and the incoming degree. The distribution of vertex degrees in the whole network is called the *degree distribution*. In an undirected network, this distribution is given by the random variable D. In a directed network, the distributions are given by $D^{(\text{un})}$, $D^{(\text{in})}$, and $D^{(\text{out})}$ for the undirected, incoming and outgoing degree, respectively. Formally, these random variables are the degrees of a vertex drawn uniformly at random from the set of all vertices in the network.

An RDS study begins with the selection of a seed group of individuals from the population. The seeds are typically chosen among population members who are well-known to researchers and that supposedly have a large number of contacts. Each seed is provided with a fixed number of coupons, typically between three to five, which are to be distributed among each seed's neighbors. The coupons effectively act as tickets for participation in the study, and each neighbor who has received a coupon is allowed to enter the study upon presenting the coupon at the study site. Those who have received a coupon and joined the study (i.e., respondents) are also provided with coupons to be distributed to their neighbors who have not yet obtained a coupon. This procedure is then repeated until the desired sample size has been reached. The sampling procedure ensures that the identities of participating individuals are not revealed, but because the coupons are numbered, it is possible to obtain the pattern of recruitment throughout the population. Rewards are given to a respondent for his or her participation and for the participation of his or her coupon recipients. This results in social pressure on coupon recipients, which is believed to facilitate effective recruitment. For each respondent, the properties of interest (e.g., HIV status and number of recent sexual encounters) are recorded. Respondents are also asked to provide the number of people they know in the population; this corresponds to the degree in an undirected network and the out-degree in a directed network.

Suppose that we are interested in estimating the proportion of individuals in a population of unknown size N with a specific trait A (e.g., HIV status), denoted p_A. Assume that we have obtained a sample s from an RDS study on this population. In order to estimate p_A from s, we assume that the RDS recruitment process behaves like a random walk on the social network of the population. To this end, it is assumed that (i) respondents recruit peers from their social contacts with equal probability, (ii) each recruitment consists of only one peer, (iii) sampling is done with replacement, (iv) the degree of respondents is reported without error, (v) the social network of the population is undirected, and (vi) the population forms a connected network. Assumption (vi) essentially means that any vertex in the network can be reached from any other vertex in the network, that is, regardless of where the sampling procedure starts, it is possible to sample all members of the population. If the recruitment process has reached equilibrium, we may then estimate p_A by

$$\hat{p}_A^{VH} = \frac{\sum_{i \in s} 1_i(A)/d_i}{\sum_{i \in s} 1/d_i}, \tag{1}$$

where $1_i(A)$ equals one if i has trait A and zero otherwise and d_i is the degree of vertex i (Volz and Heckathorn 2008). In general, when the random walk is in equilibrium and has a known stationary distribution $\{\pi_i;\ i = 1, \ldots, N\}$, we obtain an unequal probability estimator for p_A as

$$\hat{p}_A = \frac{\sum_{i \in s} 1_i(A)/\pi_i}{\sum_{i \in s} 1/\pi_i}. \tag{2}$$

In an undirected network, the stationary distribution is proportional to degree, that is, $\pi_i \propto d_i$ (Doyle and Snell 1984; Lovász 1993). Hence, the estimator in Equation (1) is obtained by using this fact to replace π_i with d_i in Equation (2). Note that the estimators in Eqs. (1) and (2) are the ratio of two Horwitz-Thompson estimators, of the population total and the population size, respectively, from which asymptotically unbiased estimates can be obtained (Särndal et al. 1992, ch. 5.6). This follows because we sample from the random walk model in equilibrium. In practice, Assumptions (i)-(vi) put RDS recruitment in the framework of an irreducible Markov chain for which equilibrium will be approached asymptotically. Although asymptotic equilibrium will not be reached in an RDS study, the recruitment process may come to an approximate equilibrium, and the use of the estimator in Equation (1) can be motivated. Hence, the Markov model obtained from Assumptions (i)-(vi) facilitates the transition from a convenience sample of seeds to a probability sample for which unequal probability estimation procedures can be used.

In most RDS studies, it is not likely that Assumptions (i)-(vi) will hold simultaneously. In this case, the random walk model of the recruitment process will at best be an approximation to the true process and the estimator in Equation (1) may be subject to substantial bias and variance. In recent years, much RDS research has focused on the sensitivity of RDS estimators to violations of Assumptions (i)-(vi). For example, in Gile and Handcock (2010) it was shown that the violation of Assumption (iii) from large sample fractions ($>50\%$ of the population) may result in large bias, and in Tomas and Gile (2011) it was shown that bias can be large when Assumption (i) is violated by differential recruitment, that is, the tendency of individuals to preferentially recruit neighbors with certain properties. In Lu et al. (2012), it was found that bias can be substantial if the social network of the population is directed (violation of Assumption (v)), or if recruitment is correlated with study variables (violation of Assumption (i)). Moreover, RDS has been empirically evaluated in, for example, Goel and Salganik (2010), where simulations on empirical networks showed that variance in RDS estimates can be five to ten times larger than in estimates from simple random sampling, and in McCreesh et al. (2012), where it was shown that only $50\%-74\%$ of 95% RDS confidence intervals (using bootstrap variance estimates) covered the true population values in an RDS study on a known population of male households in rural Uganda. Several attempts have been made to find new estimators for RDS. In Gile (2011), a successive-sampling estimator that utilizes prior

Table 1. Proportion of directed edges in social networks

Real-life social networks	Prop. dir.	**Online social networks**	Prop. dir.
High-tech managers (Wasserman and Faust 1994)	0.71	Google+ (Oct 2011) (Gong and Xu 2014)	0.62
Dining partners (Moreno 1960)	0.76	Flickr (May 2007) (Gong and Xu 2014)	0.55
Radio amateurs (Killworth and Bernard 1976)	0.59	LiveJournal (Dec 2006) (Mislove et al. 2007)	0.26
Dutch college students (Van De Bunt, Van Duijn, and Snijders 1999)	0.19	Twitter (June 2009) (Kwak et al. 2010)	0.78
Campus hall residents (Freeman, Webster, and Kirke 1998)	0.38	University e-mail (Newman, Forrest, and Balthrop 2002)	0.77
Jazz musicians (Gleiser and Danon 2003)	0.52	Enron e-mail (Boldi and Vigna 2004) (Boldi et al. 2011)	0.85

information on the population size is derived and in Gile and Handcock (2015), an estimator utilizing a superpopulation model for the social network is presented. In Lu et al. (2013), an estimator for RDS on directed social networks utilizing prior information on the in-degrees of groups of population members is presented. Lu (2013) gives an estimator that uses additional information on the composition of sampled individuals' contacts.

Current RDS estimation procedures (except Lu et al. 2013) assume that the social network of the population is undirected (cf. Assumption (v)). However, real social networks are directed in general and often include a considerable number of nonreciprocal edges. Examples of real-life social networks and social networks from online communities, including e-mail social networks, and their fraction of nonreciprocal edges among the total number of edges are shown in Table 1. In real-life social networks, such as those listed in Table 1, network data are often gathered by asking individuals to list, for example, all or some of their friends (Marsden 1990). Then, if an individual i lists j as his or her friend, but j does not list i, there will be a directed edge from i to j. For example, in the network of Dutch college students in Table 1, students were asked to list all their friends among the other residents (Van De Bunt et al. 1999), and in the dining partners' network, individuals were asked to name their two most preferred choices of dining partners. In online social networks, directed edges typically occur because an individual can add another member of the social network to his or her friend list without that member adding him or her, and in the e-mail networks, edges are formed from an individual to another if the latter is present in the former's address book (Newman et al. 2002) or if a message has been sent from the former to the latter (Boldi and Vigna 2004; Boldi et al. 2011).

The presence of directed edges may induce substantial bias and variance in the estimator in Equation (1) and other RDS estimators. For example, in their evaluation of RDS by simulations on an empirical network, Lu et al. (2012) found that the presence of directed edges caused bias as high as 0.06 in estimates from Equation (1); this can be compared with the bias of less than 0.01 induced by violation of the sampling with replacement assumption in the same study. In Lu et al. (2013), simulations on generated networks for which the proportion of directed edges was controlled showed that even a small proportion of directed edges can introduce bias in the estimator in Equation (1) and that the bias can be large (≈ 0.075) when the proportion of directed edges increases. There is also evidence of recruitment taking place by nonreciprocal relations in empirical RDS studies. For example, in an RDS study of IDUs in Sydney, Australia, 29% of participants described their relationship with their recruiter as "Not very close" (Paquette et al. 2011), and in an RDS study of IDUs in Tijuana, Mexico, 62% characterized their relationship to their recruiter as "friend" (Abramovitz et al. 2009). In an RDS study of MSM in Chicago, 13% said that they were "Not at all close" to their recruiter, and 17% characterized the relationship as "other" (instead of friend/acquaintance/partner/relative/coworker) (Phillips et al. 2014), and in an RDS study of an aboriginal community in Labrador, Canada, 80% of those recruited indicated that their recruiter was a "close relative", "distant relative", "close friend", or "friend" (Dombrowski et al. 2013).

The purpose of this article is to develop an estimator for p_A that does not require prior information on population properties for RDS in populations with directed social networks. To estimate p_A without bias from an RDS sample in such cases, we need to accurately calculate Equation (2). Because the RDS estimation method assumes a random walk behavior of the recruitment process, a random walk framework for directed networks is a key component of this expansion. This is no trivial task, because the random walk behaves very differently in undirected and directed networks. In particular, the stationary distribution of the random walk is simply proportional to the degree of the vertex in undirected networks, whereas it is affected by the entire network structure in directed networks (Donato et al. 2004; Langville and Meyer 2006; Masuda and Ohtsuki 2009). We aim to develop such a framework through which we can find estimators for the stationary distribution $\{\pi_i\}$ of the random walk on a directed network to be used in Equation (2) to estimate p_A. We will do this in several steps. Initially, we assume that we observe both the undirected degree, the incoming degree, and the outgoing degree of all vertices that are sampled. We consider the probability of returning to the same vertex after two steps in the random walk and use renewal theory to find an estimator for $\{\pi_i\}$. Then, we consider this estimation procedure in the more realistic situation when we only observe the out-degrees of sampled individuals. First, we derive results for the situation in which the expectations of the degree distributions are known. Then, we drop this assumption and by assuming a model for the social network of the population, we can estimate the unknown expectations. This gives our final estimator. All estimators are then evaluated and compared to existing RDS estimators by means of simulations.

2. Random Walks on Directed Networks

We consider a directed, strongly connected network G with N vertices. The assumption that the network is strongly connected is the equivalent of Assumption (vi) for directed networks, and means that it is possible to go from any vertex v to any other vertex w and then back (Newman 2010). Let $e_{ij} = 1$ if there is a directed edge from i to j and zero otherwise. An undirected edge exists between i and j if and only if $e_{ij} = e_{ji} = 1$. We denote the number of undirected, incoming, and outgoing edges at vertex i by $d_i^{(\mathrm{un})}$, $d_i^{(\mathrm{in})}$, and $d_i^{(\mathrm{out})}$, respectively. The degree distributions are given by the corresponding random variables $D^{(\mathrm{un})}$, $D^{(\mathrm{in})}$, and $D^{(\mathrm{out})}$. For an undirected network, we obtain $d_i^{(\mathrm{in})} = d_i^{(\mathrm{out})} = 0$, and refer to the degree of vertex i as $d_i = d_i^{(\mathrm{un})}$. Otherwise, the degree of vertex i refers to the triplet $\left(d_i^{(\mathrm{un})}, d_i^{(\mathrm{in})}, d_i^{(\mathrm{out})}\right)$. Then, $d_i^{(\mathrm{un})} + d_i^{(\mathrm{in})}$ and $d_i^{(\mathrm{un})} + d_i^{(\mathrm{out})}$ is the in-degree and out-degree of vertex i, respectively. In this notation, vertex v in Figure 1 has $d_v^{(\mathrm{un})} = 2$, $d_v^{(\mathrm{in})} = 1$, and $d_v^{(\mathrm{out})} = 1$. If the network in Figure 1 was undirected, we would obtain $d_v = 4$. It should be noted that, during the random walk, we may observe for example the out-degree $d_i^{(\mathrm{un})} + d_i^{(\mathrm{out})}$, but not the $d_i^{(\mathrm{un})}$ and $d_i^{(\mathrm{out})}$ values separately.

Consider the simple random walk $X = \{X(t); t = 0, 1, \ldots\}$ with state space $S = \{1, \ldots, N\}$ on G such that the walker staying at vertex i moves to any of the $d_i^{(\mathrm{un})} + d_i^{(\mathrm{out})}$ neighbors reached by an undirected or outgoing edge with equal probability. We denote the stationary distribution of X by $\{\pi_i; i = 1, \ldots, N\}$, where $\pi_i = \lim_{t\to\infty} P(X(t) = i)$. The stationary distribution exists if the network is aperiodic, that is, the walker will not return periodically to the same vertex repeatedly during the walk. If we sample from the random walk in equilibrium, we refer to $\{\pi_i\}$ as the *selection probabilities* of the vertices in G.

For an arbitrary network, we obtain

$$\pi_i = \sum_{j=1}^{N} \frac{e_{ji}}{\sum_{\ell=1}^{N} e_{j\ell}} \pi_j = \sum_{j=1}^{N} \frac{e_{ji}}{d_j^{(\mathrm{un})} + d_j^{(\mathrm{out})}} \pi_j, i = 1, \ldots, N, \tag{3}$$

where the stationary distribution is fully defined by $\sum_{i=1}^{N} \pi_i = 1$. In undirected networks, we obtain $\pi_i = d_i / \sum_{j=1}^{N} d_j$. In contrast, there is no analytical closed-form solution for $\{\pi_i\}$ in directed networks. If a directed network has little assortativity (i.e., degree correlation between adjacent vertices), $\{\pi_i\}$ is often accurately estimated by the normalized in-degree (Fortunato et al. 2008; Ghoshal and Barabási 2011) because

$$\pi_i \approx \sum_{j=1}^{N} \frac{e_{ji}}{d_j^{(\mathrm{un})} + d_j^{(\mathrm{out})}} \bar{\pi} \propto \sum_{j=1}^{N} e_{ji} = d_i^{(\mathrm{in})} + d_i^{(\mathrm{un})}, \tag{4}$$

where $\bar{\pi}$ is the average selection probability. Equation (4) depends only on the in-degree of vertices, that is, it provides a local description of the global solution to Equation (3). However, the estimate given by (4) is often inaccurate in general directed networks (Donato et al. 2004; Masuda and Ohtsuki 2009). Moreover, since it is much easier for individuals to assess their out-degree, that is, how many people they know, than their in-degree, that is, by how many people they are known, it is common to observe only the out-degree. In this case, Equation (4) cannot be used with an RDS sample.

3. Estimating Selection Probabilities

We now derive estimators of the selection probabilities for the random walk on directed networks. We first derive an estimation scheme when the full degree $\left(d_i^{(\mathrm{un})}, d_i^{(\mathrm{in})}, d_i^{(\mathrm{out})}\right)$ is observed for all the vertices i visited by the random walk. Then, we restrict this estimation to the situation in which only the out-degree $d_i^{\mathrm{un}} + d_i^{\mathrm{out}}$ of the visited vertices is observed. In both situations it is assumed that the degrees are observed without error. Note that the random walk allows a vertex to be visited multiple times, whereas it is typically not allowed to be sampled several times in an RDS study.

3.1. Estimating Selection Probabilities from Full Degrees

In order to estimate $\{\pi_i\}$, we assume that $X(t_0) = i$, where t_0 is sufficiently large for the stationary distribution to be reached. We evaluate the frequency with which $X(t)$ visits i in the subsequent times. If $X(t)$ leaves i through an undirected edge $e_{i\cdot}^{(\mathrm{un})}$, where $e_{i\cdot}^{(\mathrm{un})}$ is one of the $d_i^{(\mathrm{un})}$ undirected edges owned by i, $X(t)$ may return to i after two steps using the same edge and repeat the same type of returns m times in total, perhaps using different undirected edges $e_{i\cdot}^{(\mathrm{un})}$. Then, $X(t_0) = X(t_0 + 2) = \cdots = X(t_0 + 2m) = i$ and $X(t_0 + 2m + 2) = k$ for some $k \neq i$.

If $X(t_0 + 2) = i$, the walk first moves from i through an undirected edge to vertex j at $t = t_0 + 1$ and returns to i through the same edge at $t = t_0 + 2$. The probability of this event is given by $d_i^{(\mathrm{un})} / \left(d_i^{(\mathrm{un})} + d_i^{(\mathrm{out})}\right) \cdot 1/\left(d_j^{(\mathrm{un})} + d_j^{(\mathrm{out})}\right)$. Because the out-degree of vertex j, that is, $d_j^{(\mathrm{un})} + d_j^{(\mathrm{out})}$, is unknown, we approximate $1/\left(d_j^{(\mathrm{un})} + d_j^{(\mathrm{out})}\right)$ by $E(1/(\tilde{D}^{(\mathrm{un})} + D^{(\mathrm{out})}))$. Here $\tilde{D}^{(\mathrm{un})}$ denotes the undirected degree distribution under the condition that the vertex is reached by following an undirected edge. This yields a *size-biased* distribution for the undirected degree, given by $P(\tilde{D}^{(\mathrm{un})} = d) \propto dP(D^{(\mathrm{un})} = d)$ (Newman 2010). It is also possible to estimate $1/\left(d_j^{(\mathrm{un})} + d_j^{(\mathrm{out})}\right)$ by $1/E(\tilde{D}^{(\mathrm{un})} + D^{(\mathrm{out})})$, which however proved to have very little effect in our simulations, and if any, a slightly worse one. Thus, we estimate the probability of returning to vertex i after two steps by

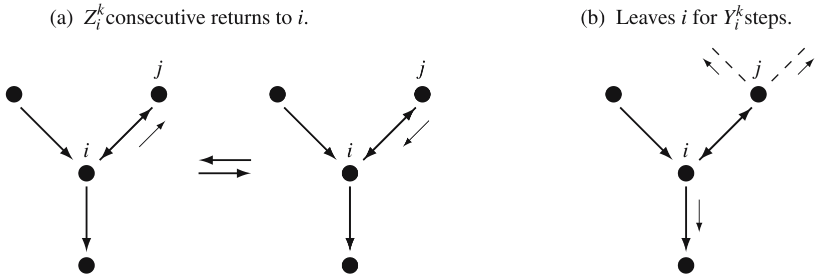
$$p_i^{(\mathrm{ret})} = \frac{d_i^{(\mathrm{un})}}{d_i^{(\mathrm{un})} + d_i^{(\mathrm{out})}} E\left(\frac{1}{\tilde{D}^{(\mathrm{un})} + D^{(\mathrm{out})}}\right). \tag{5}$$

When $t \geq t_0 + 2m + 3$, we use Equation (4) to estimate the probability of visiting vertex i at any time as being proportional to $d_i^{(\mathrm{un})} + d_i^{(\mathrm{in})}$, that is,

$$p_i^{(\mathrm{vis})} = \frac{d_i^{(\mathrm{un})} + d_i^{(\mathrm{in})}}{N(E(D^{(\mathrm{un})}) + E(D^{(\mathrm{in})}))}. \tag{6}$$

Under these estimates, the number of returns after two steps to vertex i, counting the starting point $X(t_0) = i$ as the first return to i, is geometrically distributed with expected value $1/\left(1 - p_i^{(\mathrm{ret})}\right)$, and the number of steps starting from $t = t_0 + 2m + 2$, including this step, and ending at the time immediately before visiting i with probability $p_i^{(\mathrm{vis})}$ is geometrically distributed with the expected value $1/p_i^{(\mathrm{vis})}$.

We then have a renewal process, that is, a process which repeatedly regenerates at random times such that the intervals between them are of independent and identically distributed lengths. These random times are called renewals. We denote our process $\left\{R_i^n; n \geq 1, R_i^0 = 0\right\}$, with the nth renewal occurring at the random time $R_i^n = \sum_{k=1}^n \left(2Z_i^k + Y_i^k\right)$, where $Z_i^k \sim Ge\left(1 - p_i^{(\mathrm{ret})}\right)$ and $Y_i^k \sim Ge\left(p_i^{(\mathrm{vis})}\right)$. In Figure 2, the behavior of the process during the kth renewal period is shown schematically. Figure 2(a) shows the behavior of the walk when it makes consecutive returns to i. During this time, the walker always leaves i through an undirected edge, which is not necessarily the same edge each time (left part of Figure 2(a)), and returns after two time steps to i via the same edge (right part of Figure 2(a)). This is repeated such that the walker makes in total Z_i^k consecutive returns to i. The duration of this is $2Z_i^k$ time steps. Figure 2(b) shows the behavior of the walk when it leaves i and does not return after two time steps. This occurs


(a) Z_i^kconsecutive returns to i.

(b) Leaves i for Y_i^ksteps.

Fig. 2. Schematic illustration of the kth renewal period.

if either the walker leaves i by an outgoing edge, through which it is impossible to return directly to i, or if the walker leaves i by an undirected edge but does not return to i through this edge in the next time step. When the walker has left i, the time until its return to i is Y_i^k time steps. The average time step between consecutive renewal events is equal to $2E\left(Z_i^k\right) + E\left(Y_i^k\right)$. The average number of visits to i between two renewal events, with the visit to i at $t = t_0$ included, is equal to $E\left(Z_i^k\right)$. Therefore, from renewal theory (see e.g., Feller 1950), we obtain an estimate of π_i as

$$\pi_i \approx \frac{E\left(Z_i^k\right)}{2E\left(Z_i^k\right) + E\left(Y_i^k\right)} = \frac{\dfrac{1}{1 - p_i^{(\mathrm{ret})}}}{2\dfrac{1}{1 - p_i^{(\mathrm{ret})}} + \dfrac{1}{p_i^{(\mathrm{vis})}}} = \frac{p_i^{(\mathrm{vis})}}{2p_i^{(\mathrm{vis})} + 1 - p_i^{(\mathrm{ret})}}. \tag{7}$$

Because $p_i^{(\mathrm{ret})} = O(1)$ and $p_i^{(\mathrm{vis})} = O(1/N)$, removing higher-order terms in Equation (7) yields

$$\hat{\pi}_i \approx \frac{p_i^{(\mathrm{vis})}}{1 - p_i^{(\mathrm{ret})}} \propto \frac{d_i^{(\mathrm{un})} + d_i^{(\mathrm{in})}}{1 - \dfrac{d_i^{(\mathrm{un})}}{d_i^{(\mathrm{un})} + d_i^{(\mathrm{out})}} E\left(\dfrac{1}{\tilde{D}^{(\mathrm{un})} + D^{(\mathrm{out})}}\right)}. \tag{8}$$

The proportionality constant is given by imposing that $\sum_{i=1}^{N} \hat{\pi}_i = 1$. If the network is undirected, we obtain $\hat{\pi}_i \propto d_i^{(\mathrm{un})}$, such that $\hat{\pi}_i$ coincides with the exact solution used in Equation (1). If the network is without reciprocal edges, the estimator is proportional to incoming degree $d_i^{(\mathrm{in})}$.

3.2. Estimating Selection Probabilities from Out-Degrees

A common situation in RDS is that only the out-degrees (i.e., $d_i^{(\mathrm{un})} + d_i^{(\mathrm{out})}$) of respondents are observed. Then, the estimator of the selection probabilities given by Equation (8) cannot be used directly. To cope with this situation, we estimate the number of undirected, incoming, and outgoing edges from the observed out-degrees and substitute the estimated $\left(\hat{d}_i^{(\mathrm{un})}, \hat{d}_i^{(\mathrm{in})}, \hat{d}_i^{(\mathrm{out})}\right)$ in Equation (8).

Assume that we have observed the out-degree $d_i^{(\mathrm{un})} + d_i^{(\mathrm{out})}$ of vertex i. We estimate $d_i^{(\mathrm{un})}$ and $d_i^{(\mathrm{out})}$ by their expected proportions of the observed out-degree, and the incoming degree by its expectation, as follows:

$$\begin{cases} \hat{d}_i^{(\mathrm{un})} = \dfrac{E(D^{(\mathrm{un})})}{E(D^{(\mathrm{un})}) + E(D^{(\mathrm{out})})}\left(d_i^{(\mathrm{un})} + d_i^{(\mathrm{out})}\right), \\ \hat{d}_i^{(\mathrm{out})} = \dfrac{E(D^{(\mathrm{out})})}{E(D^{(\mathrm{un})}) + E(D^{(\mathrm{out})})}\left(d_i^{(\mathrm{un})} + d_i^{(\mathrm{out})}\right), \\ \hat{d}_i^{(\mathrm{in})} = E(D^{(\mathrm{in})}). \end{cases} \tag{9}$$

The expectations used in Equation (9) rely on the assumption that we have a random sample from the network, which is not true in this case. We have evaluated the case of a size-biased distribution for incoming and/or undirected degrees; however, our numerical results suggest that this makes little difference, and if any, increases the bias of selection probability estimators. Therefore, we stay with the estimators given by Equation (9).

When $\left(\hat{d}_i^{(\mathrm{un})}, \hat{d}_i^{(\mathrm{in})}, \hat{d}_i^{(\mathrm{out})}\right)$ is substituted in Equation (8) in place of $\left(d_i^{(\mathrm{un})}, d_i^{(\mathrm{in})}, d_i^{(\mathrm{out})}\right)$, the term $\hat{d}_i^{(\mathrm{un})} / \left(\hat{d}_i^{(\mathrm{un})} + \hat{d}_i^{(\mathrm{out})}\right)$ in the denominator is constant. Therefore, the estimator is proportional to $\hat{d}_i^{(\mathrm{un})} + \hat{d}_i^{(\mathrm{in})}$ and hence equivalent to Equation (4) calculated with the estimated degrees.

3.3. Estimating Expectations of Degree Distributions

The degree estimators in Equation (9) rely on $E(D^{(\mathrm{un})})$, $E(D^{(\mathrm{in})})$, and $E(D^{(\mathrm{out})})$, which are not estimable from a typical RDS sample, where only the out-degrees $d_i^{(\mathrm{un})} + d_i^{(\mathrm{out})}$ of respondents are observed. In order to extend the estimation procedure to handle these unknown expectations, we assume a model for the network by which they can be estimated.

Specifically, it is assumed that the observed network is a realization of a directed equivalent of the simple $G(N, p = \lambda/(N-1))$ random graph (Erdős and Rényi 1960). This graph has N vertices and hence $\binom{N}{2}$ pairs of vertices. Given parameters $\alpha \in [0, 1]$ and $\lambda \in [0, N-1]$, each pair of vertices independently forms an edge with probability $\lambda/(N-1)$, which is undirected with probability $(1-\alpha)$ and directed with probability α. When the edge is directed, the direction is selected with equal probability. Because each vertex may connect to each of the other $N-1$ vertices, it follows that λ is the expected total degree of a vertex. We also have that α is the fraction of directed edges as $N \longrightarrow \infty$.

Because edges are formed independently of each other, vertex degrees are binomially distributed. Hence, if N is large, $D^{(\mathrm{un})}$, $D^{(\mathrm{in})}$, and $D^{(\mathrm{out})}$ approximately follow independent Poisson distributions with parameters $(1-\alpha)\lambda$, $\alpha\lambda/2$, and $\alpha\lambda/2$, respectively. It follows that the out-degree $D^{(\mathrm{un})} + D^{(\mathrm{out})}$ and the in-degree $D^{(\mathrm{un})} + D^{(\mathrm{in})}$ are both Poisson distributed with parameter $(2-\alpha)\lambda/2$. Consequently, to estimate the unknown expectations, it is enough to estimate α and λ, and substitute the estimated $\hat{\alpha}$ and $\hat{\lambda}$ in the expectations of the (Poissonian) degree distributions.

To find an estimator of α, we again consider the random walk $X = \{X(t)\}$ on the network. Assume that $e_{ij} = 1$, $X(t_0) = i$, and $X(t_0 + 1) = j$, for a large t_0. If $X(t_0 + 2) = i$, the edge between i and j is undirected, that is, $e_{ij} = e_{ji} = 1$, and the random walk leaves vertex j via e_{ji}. The probability that the edge is undirected is set to $(1-\alpha)/(1-\alpha/2)$, that is, the probability that an edge selected uniformly at random among all undirected and incoming edges is undirected. This will only approximately hold for the random walk, but simulations show that it is a reasonable approximation. If there is an undirected edge between i and j (i.e., $e_{ji} = 1$), the random walk leaves j via e_{ji} with probability $1/\left(d_j^{(\mathrm{un})} + d_j^{(\mathrm{out})}\right)$. Thus, the random walk revisits vertex i at $t_0 + 2$ under the directed E-R random-graph model with probability

$$p_j^{(\mathrm{rev})} = \frac{1-\alpha}{1-\alpha/2} \cdot \frac{1}{d_j^{(\mathrm{un})} + d_j^{(\mathrm{out})}}. \tag{10}$$

Let M be the number of revisits, as described above, during l consecutive steps, where l is typically equal to the sample size. Then, we have $M = \sum_{k=2}^{l} M_k$, where $M_k = 1$ if a revisit occurs in step k and $M_k = 0$ otherwise. M_k is Bernoulli distributed, $M_k \sim \mathrm{Be}\left(p_{j_{k-1}}^{(\mathrm{rev})}\right)$, where j_{k-1} is the vertex visited in step $k-1$. We obtain the expected number of revisits as

$$E(M) = \frac{1-\alpha}{1-\alpha/2} \sum_{k=1}^{l-1} \frac{1}{d_{j_k}^{(\mathrm{un})} + d_{j_k}^{(\mathrm{out})}}. \tag{11}$$

If m is the observed number of revisits, we set $m = E(M)$ in Equation (11) to obtain the moment estimator

$$\hat{\alpha} = \frac{m - \sum_{k=1}^{l-1} \left(d_{j_k}^{(\mathrm{un})} + d_{j_k}^{(\mathrm{out})}\right)^{-1}}{m/2 - \sum_{k=1}^{l-1} \left(d_{j_k}^{(\mathrm{un})} + d_{j_k}^{(\mathrm{out})}\right)^{-1}}. \tag{12}$$

If the estimated $\hat{\alpha} < 0$, we force $\hat{\alpha} = 0$.

Given $\hat{\alpha}$, we estimate λ as follows. If $\alpha = 0$, the network contains only undirected edges, and the observed out-degree equals the observed undirected degree, which has a size-biased distribution with $E(\tilde{D}^{(\mathrm{un})}) = \lambda + 1$. If $\alpha = 1$, the network has only directed edges, and the expected observed out-degree is well approximated by the expected number of outgoing edges, $\lambda/2$. By linearly interpolating the expected observed out-degree between $\alpha = 0$ and $\alpha = 1$, and substituting it with the mean sample out-degree $\bar{u}$, we obtain $\bar{u} = \lambda/2 + (1 - \alpha)(1 + \lambda/2)$, which yields an estimator of λ as

$$\hat{\lambda} = \frac{\bar{u} + \hat{\alpha} - 1}{1 - \hat{\alpha}/2}. \tag{13}$$

Using $\hat{\alpha}$ and $\hat{\lambda}$, we can estimate the expectations of the degree distributions under the random-graph model. For example, $E(D^{(\mathrm{un})})$ is estimated by $(1 - \hat{\alpha})\hat{\lambda}$. By substituting these estimated expectations in Eqs. (8) and (9), we obtain an estimator of the selection probability of vertex i as

$$\hat{\pi}_i \propto \hat{d}_i^{(\mathrm{un})} + \hat{d}_i^{(\mathrm{in})} = \frac{1-\hat{\alpha}}{1-\hat{\alpha}/2} \left(d_i^{(\mathrm{un})} + d_i^{(\mathrm{out})}\right) + \frac{\hat{\alpha}\hat{\lambda}}{2}. \tag{14}$$

When $\alpha = 0$ is assumed known and used in place of $\hat{\alpha}$, the estimator in Equation (14) is equivalent to that used in Equation (1). When $\hat{\alpha} = \alpha = 1$, the estimator is proportional to one, and thus equivalent to the sample mean. This reflects the fact that, if $\alpha = 1$, the network has no undirected edges, and the out-degree is equal to the outgoing degree, which does not provide any information on the selection probability of a vertex in this case.

It should be noted that the construction of the directed Erdős-Rényi graphs results in vertices having the same out-degree and in-degree on average, which is not likely to occur in actual populations where RDS is used. This makes estimation of in-degree using only the observed out-degree feasible, and might possibly favor the performance of the estimator in Equation (14) for networks generated by this model.

4. Simulation Setup

We numerically examine the accuracy of our estimation schemes on directed Erdős-Rényi graphs, a model of directed power-law networks (i.e., networks with a power-law degree distribution), and an online MSM social network. We evaluate both the estimated selection probabilities and corresponding estimates of p_A. As described in Section 1, real-life directed social networks show a varying fraction of directed edges, corresponding to a diversity of α values. Therefore, α is varied in the model networks. We also vary λ and other network parameters. We study the performance of our estimators when the full degree is observed and when only the out-degree is observed, and compare them with existing estimators. We do not consider RDS estimators that are not based on the random walk framework because they fall outside the scope of this study.

4.1. Network Models and Empirical Network

The first model network that we use is a variant of the Erdős-Rényi graph with a mixture of undirected and directed edges, as described in Section 3. We generate the networks with $\alpha \in \{0.25, 0.5, 0.75\}$ and $\lambda \in \{5, 10, 15\}$. We then extract the largest strongly connected component of the generated network, which has $O(N)$ vertices for all combinations of α and λ.

The directed Erdős-Rényi networks have Poisson degree distributions with quickly decaying tails. In fact, many empirical networks have heavy-tailed degree distributions as represented by the power law (Newman 2010). In other words, there are typically small numbers of vertices whose degrees are huge, and a majority of vertices have small degrees. To mimic heavy-tailed degree distributions, we also use a variant of a power-law network model (Goh et al. 2001; Chung and Lu 2002; Chung et al. 2003). The original algorithm for generating undirected power-law networks presented in Goh et al. (2001) is as follows.

We fix the number of vertices N and expected degree $E(D)$. Then, we set the weight of vertex i $(1 \leq i \leq N)$ to be $w_i = i^{-\tau}$. As specified in the following, w_i represents the extent to which vertex i attracts edges; a large w_i value yields a large degree. Parameter $0 \leq \tau \leq 1$ controls the power-law exponent of the degree distribution. If $\tau = 0$, all w_i are equal such that each vertex is statistically the same. In this case, the degree distribution produced by the following algorithm will not be heavy-tailed. When $\tau > 0$, a vertex with small i possesses large w_i and will in fact have a very large degree. Then, we select a pair of vertices i and j $(1 \leq i \neq j \leq N)$ with probability proportional to $w_i w_j$. If the two vertices are not yet connected, we connect them by an undirected edge. We repeat the procedure until the network has $E(D)N/2$ edges such that the expected degree is equal to $E(D)$. The expected degree of vertex i is proportional to w_i, and the degree distribution is given by $p(d) \propto d^{-\gamma}$, where $\gamma = 1 + \frac{1}{\tau}$ (Goh et al. 2001).

To generate a power-law network in which undirected and directed edges are mixed with a desired fraction, we extend the algorithm as follows. First, we specify the expected undirected degree $E(D^{(\mathrm{un})})$ and generate an undirected network. Second, we define $w_i^{\mathrm{in}} = (\sigma^{\mathrm{in}}(i))^{-\tau^{\mathrm{in}}}$ $(1 \le i \le N)$, where σ^{in} is a realization of the random permutation on $1, \ldots, N$. Parameter τ^{in} specifies the power-law exponent of the incoming degree distribution. Similar to the undirected case, a vertex with a small $\sigma^{\mathrm{in}}(i)$ value will have a large in-degree. Similarly, we set $w_i^{\mathrm{out}} = (\sigma^{\mathrm{out}}(i))^{-\tau^{\mathrm{out}}}$ $(1 \le i \le N)$, where σ^{out} is another realization of the random permutation on $1, \ldots, N$. Third, we select a pair of vertices with probability proportional to $w_i^{\mathrm{in}} w_j^{\mathrm{out}}$. If $i \ne j$ and there is no directed edge from j to i yet, we place a directed edge from j to i. We repeat the procedure until a total of $E(D^{(in)})N/2$ edges are placed. It should be noted that $E(D^{(\mathrm{in})}) = E(D^{(\mathrm{out})})$. The incoming degree distribution is given by $p(d^{\mathrm{in}}) \propto (d^{\mathrm{in}})^{-\gamma^{\mathrm{in}}}$, where $\gamma^{\mathrm{in}} = 1 + \frac{1}{\tau^{\mathrm{in}}}$, and similar for the outgoing degree distribution. Finally, we superpose the obtained undirected network and directed network to make a single graph. If the combined graph is not strongly connected, we discard it and start over until a strongly connected network is generated. By construction, a network constructed from this model is devoid of degree correlation.

In both network models, we vary the probability of a vertex being assigned property A as proportional to six different combinations of its degree: in-degree, out-degree, undirected degree, incoming degree, outgoing degree, and directed degree, that is, the sum of incoming and outgoing degree. Formally, if $P(\text{vertex } i \text{ has } A) \propto g\left(d_i^{(\mathrm{un})}, d_i^{(\mathrm{in})}, d_i^{(\mathrm{out})}\right)$, we let g be equal to $\left(d_i^{(\mathrm{un})} + d_i^{(\mathrm{in})}\right)$, $\left(d_i^{(\mathrm{un})} + d_i^{(\mathrm{out})}\right)$, $d_i^{(\mathrm{un})}$, $d_i^{(\mathrm{in})}$, $d_i^{(\mathrm{out})}$, and $\left(d_i^{(\mathrm{in})} + d_i^{(\mathrm{out})}\right)$, respectively. We refer to these as different ways to allocate property A. We also examined the case in which we assigned the property uniformly over all vertices. However, because the performance of the different estimators is similar in this case, we do not show the results in the following. For all allocations of A, the property is assigned in such a way that the expected proportion of vertices being assigned A is equal to some fixed value p. Because A is stochastically assigned, the actual proportion p_A of vertices with A will vary between realized allocations.

We also evaluate our estimators on an online MSM social network, extracted during Dec 2005-Jan 2006 from www.qruiser.com, which is the Nordic region's largest community for lesbian, gay, bisexual, transgender, and queer persons (Rybski et al. 2009). In this network, an edge represents that at least one message has been sent between the two vertices connected by that edge. A directed edge occurs if messages have only been sent in one direction between two vertices. The data set considered here was first described in Lu et al. (2012) and represents a subpopulation of the original data set consisting of 16,082 male homosexual members in a directed social network that is made up of one strongly connected component. This network represents the social structure of a hidden population and makes it possible to evaluate the effect of the presence of nonreciprocal edges in RDS. It has previously been used to evaluate the performance of RDS estimators under different violations of Assumptions (i)-(vi) in Lu et al. (2012) and in directed social networks in Lu et al. (2013). The data set also includes users' profiles, which are seldom available. From these, we obtain four dichotomous individual properties: age (born before 1980 or not), county (live in Stockholm or not), civil status (married or unmarried), and profession (employed or unemployed). This makes it possible to evaluate the performance of RDS estimators of population proportions on this network. The fraction of directed edges in the network is equal to $\alpha = 0.76$. The in-degree and out-degree distributions are skewed, and the mean number of edges λ is equal to 27.74 (Lu et al. 2012). Preferably, RDS would be evaluated on a network which is known to depict that on which the recruitment process in RDS takes place. Such network data is rare, however, and in its absence, the considered network is a good option for RDS evaluation.

4.2. *Evaluation of Estimators*

We compared the performance of our estimators of the selection probabilities with three other estimators. We refer to our estimator $\{\hat{\pi}_i\}$ obtained from Equation (8) as $\left\{\hat{\pi}_i^{(\mathrm{ren})}\right\}$, where *ren* stands for renewal. This estimator is compared to $\left\{\hat{\pi}_i^{(\mathrm{uni})}\right\}$, which assigns a uniform probability $\hat{\pi}_i^{(\mathrm{uni})} = 1/N$ for all i, $\left\{\hat{\pi}_i^{(\mathrm{outdeg})}\right\}$, for which $\hat{\pi}_i^{(\mathrm{outdeg})} \propto d_i^{(\mathrm{un})} + d_i^{(\mathrm{out})}$, that is, proportional to out-degree (Equation 1), and $\left\{\hat{\pi}_i^{(\mathrm{indeg})}\right\}$, where $\hat{\pi}_i^{(\mathrm{indeg})} \propto d_i^{(\mathrm{un})} + d_i^{(\mathrm{in})}$, that is, proportional to in-degree (Equation 4). Note that if the network is undirected, $\left\{\hat{\pi}_i^{(\mathrm{outdeg})}\right\}$ and $\left\{\hat{\pi}_i^{(\mathrm{indeg})}\right\}$ are equal. However, typically in RDS the out-degree is observed and $\left\{\hat{\pi}_i^{(\mathrm{outdeg})}\right\}$, which is used in the current RDS estimator in Equation (1), is the estimator that should be considered in the undirected case. In the following, we suppress the $\{\}$ notation.

To assess the performance of an estimator we calculated its estimated selection probabilities $\hat{\pi}_i$ and the true stationary distribution π_i for all the vertices in the given network. Then, we calculated the *total variation distance* defined by

$$D_{TV} = \frac{1}{2}\sum_{i=1}^{N} |\hat{\pi}_i - \pi_i| \tag{15}$$

(Levin et al. 2009). The stationary distribution π_i was obtained using the power method, which is an iterative method that works as follows (Langville and Meyer 2006). Starting from an arbitrary nonzero vector of size N, in each iteration the resulting vector is multiplied with the matrix $\{e_{ij}\}$, where $e_{ij} = 1$ if there is a directed edge from i to j. Then, under some conditions that hold for the networks used in this study, the resulting vector converges to the stationary distribution, which is the eigenvector corresponding to the largest eigenvalue of $\{e_{ij}\}$. We use an accuracy of 10^{-10} in terms of the total variation distance for the two distributions given in the successive two steps of the power iteration.

For $\hat{\pi}^{(\mathrm{ren})}$, we considered three variants depending on the information available from observed degree and knowledge of the expectations of the degree distributions. When the full degree $\left(d_i^{(\mathrm{un})}, d_i^{(\mathrm{in})}, d_i^{(\mathrm{out})}\right)$ was observed, we used Equation (8) to calculate $\hat{\pi}^{(\mathrm{ren})}$, where $E(1/(\tilde{D}^{(\mathrm{un})} + D^{(\mathrm{out})}))$ is estimated by the mean of the inverse sample out-degrees. We denote the corresponding estimator with $\hat{\pi}_{\mathrm{f.d.}}^{(\mathrm{ren})}$, where *f.d.* stands for "full degree". When only the out-degree was observed and the expectations of the degree distributions were known, we used Equation (9). This case is only evaluated for the directed Erdős-Rényi graphs, and the corresponding estimator is denoted by $\hat{\pi}_{\alpha,\lambda}^{(\mathrm{ren})}$. If only the out-degree was observed and the expectations of the degree distributions were unknown, we used Equations (12), (13), and (14), and the estimator is denoted $\hat{\pi}^{(\mathrm{ren})}$.

We obtained a sample of size n_s from each generated network by means of a random walk starting from a randomly selected vertex. In the random walk, we collected the degree of visited vertices and observed whether they had property A or not. We estimated the population proportion p_A from the sample by replacing π in Equation (2) by either $\hat{\pi}^{(\text{uni})}$, $\hat{\pi}^{(\text{outdeg})}$, $\hat{\pi}^{(\text{indeg})}$, or any of the variants of $\hat{\pi}^{(\text{ren})}$, yielding estimates $\hat{p}_A^{(\text{uni})}$, $\hat{p}_A^{(\text{outdeg})}$, $\hat{p}_A^{(\text{indeg})}$, or $\hat{p}_A^{(\text{ren})}$, respectively. Note that $\hat{p}_A^{(\text{uni})}$ yields the sample proportion suggested as an estimator for RDS in Heckathorn (1997), $\hat{p}_A^{(\text{outdeg})}$ yields the RDS estimator from Volz and Heckathorn (2008), where the direction of edges is ignored, and $\hat{p}_A^{(\text{indeg})}$ gives the RDS estimator for directed networks from Lu et al. (2013).

5. Numerical Results

5.1. Directed Erdős-Rényi Graphs

In Table 2, we show the mean of the total variation distance D_{TV} between the true stationary distribution and $\hat{\pi}^{(\text{uni})}$, $\hat{\pi}^{(\text{outdeg})}$, $\hat{\pi}^{(\text{indeg})}$, and $\hat{\pi}^{(\text{ren})}_{\text{f.d.}}$, calculated on the basis of 1,000 realizations of the largest strongly connected component of the directed random graph having $N = 1{,}000$ vertices. Because the standard deviation of D_{TV} is similar

Table 2. Mean and average standard deviation (s.d.) of D_{TV} for the directed random graph when $\left(d_i^{(un)}, d_i^{(in)}, d_i^{(out)}\right)$ is observed and moments of the degree distributions are known. The lowest D_{TV} value is marked in boldface. We set $N = 1{,}000$

(a) $\alpha = 0.1$

λ	$\hat{\pi}^{(\text{uni})}$	$\hat{\pi}^{(\text{outdeg})}$	$\hat{\pi}^{(\text{indeg})}$	$\hat{\pi}^{(\text{ren})}_{\text{f.d}}$	s.d.
5	0.185	0.074	0.042	**0.041**	0.004
10	0.131	0.045	0.017	**0.016**	0.002
15	0.106	0.036	**0.010**	**0.010**	0.001

(b) $\alpha = 0.25$

	$\hat{\pi}^{(\text{uni})}$	$\hat{\pi}^{(\text{outdeg})}$	$\hat{\pi}^{(\text{indeg})}$	$\hat{\pi}^{(\text{ren})}_{\text{f.d}}$	s.d.
	0.203	0.134	0.077	**0.075**	0.005
	0.140	0.081	0.031	**0.030**	0.002
	0.112	0.063	**0.019**	**0.019**	0.002

(c) $\alpha = 0.5$

λ	$\hat{\pi}^{(\text{uni})}$	$\hat{\pi}^{(\text{outdeg})}$	$\hat{\pi}^{(\text{indeg})}$	$\hat{\pi}^{(\text{ren})}_{\text{f.d}}$	s.d.
5	0.247	0.225	0.138	**0.133**	0.009
10	0.160	0.136	0.056	**0.055**	0.004
15	0.126	0.105	0.034	**0.033**	0.002

(d) $\alpha = 0.75$

	$\hat{\pi}^{(\text{uni})}$	$\hat{\pi}^{(\text{outdeg})}$	$\hat{\pi}^{(\text{indeg})}$	$\hat{\pi}^{(\text{ren})}_{\text{f.d}}$	s.d.
	0.303	0.319	0.207	**0.201**	0.014
	0.188	0.201	0.090	**0.088**	0.005
	0.144	0.156	**0.055**	**0.055**	0.003

between the estimators, we show an average over the four estimators. The sample size n_s used in $\hat{\pi}^{(\text{ren})}_{\text{f.d.}}$ is 500. We also tried $n_s = 200$, which gave similar results. The D_{TV} value of $\hat{\pi}^{(\text{indeg})}$ and $\hat{\pi}^{(\text{ren})}_{\text{f.d.}}$ is much smaller than that of $\hat{\pi}^{(\text{uni})}$ and $\hat{\pi}^{(\text{outdeg})}$ for all values of α and λ. Furthermore, $\hat{\pi}^{(\text{ren})}_{\text{f.d}}$ always gives smaller D_{TV} than $\pi^{(\text{indeg})}$ although the two values are similar for many combinations of the parameters.

In Table 3, we show the mean and average s.d. of D_{TV} when the out-degree, that is, $d_i^{(\text{un})} + d_i^{(\text{out})}$, is observed but the individual $d_i^{(\text{un})}$ and $d_i^{(\text{out})}$ values are not. The assumptions underlying the network generation are the same as those for Table 2, and $n_s = 500$. We consider two cases. In the first case, the expectations of the degree distributions are known, and we use the estimator $\hat{\pi}^{(\text{ren})}_{\alpha,\lambda}$. In the second case, they are not known, and we use $\hat{\pi}^{(\text{ren})}$. Results for $\hat{\pi}^{(\text{indeg})}$ are not shown in Table 3 because in-degree is not observed. Table 3 indicates that D_{TV} for $\hat{\pi}^{(\text{ren})}$ is smaller than that for $\hat{\pi}^{(\text{uni})}$ and $\hat{\pi}^{(\text{outdeg})}$ when α is 0.5 and 0.75. When $\alpha = 0.75$, $\hat{\pi}^{(\text{outdeg})}$ yields the largest D_{TV}. For $\alpha = 0.1$ and 0.25, $\hat{\pi}^{(\text{ren})}$ and $\hat{\pi}^{(\text{outdeg})}$ yield similar results. For all parameter values $\hat{\pi}^{(\text{ren})}_{\alpha,\lambda}$ slightly outperforms $\hat{\pi}^{(\text{ren})}$. We tried $n_s = 200$ (not shown), which gave similar s.d. for $\hat{\pi}^{(\text{ren})}_{\alpha,\lambda}$, and similarly for $\hat{\pi}^{(\text{ren})}$, except for $\alpha = 0.1$, where, for example, $\lambda = 15$ yielded the s.d. values of 0.0039 and 0.0073 for $n_s = 500$ and $n_s = 200$, respectively.

Table 3. Mean and average s.d. of D_{TV} for the directed random graph when $d_i^{(un)} + d_i^{(out)}$ is observed. We set $N = 1{,}000$

(a) $\alpha = 0.1$

λ	$\hat{\pi}^{(\text{uni})}$	$\hat{\pi}^{(\text{outdeg})}$	$\hat{\pi}^{(\text{ren})}_{\alpha,\lambda}$	$\hat{\pi}^{(\text{ren})}$	s.d.
5	0.185	**0.074**	**0.074**	0.075	0.004
10	0.131	**0.045**	**0.045**	0.047	0.003
15	0.106	0.036	**0.035**	0.037	0.002

(b) $\alpha = 0.25$

	$\hat{\pi}^{(\text{uni})}$	$\hat{\pi}^{(\text{outdeg})}$	$\hat{\pi}^{(\text{ren})}_{\alpha,\lambda}$	$\hat{\pi}^{(\text{ren})}$	s.d.
	0.203	0.135	**0.132**	0.133	0.006
	0.140	0.081	**0.079**	0.080	0.003
	0.112	0.063	**0.061**	0.063	0.002

(c) $\alpha = 0.5$

λ	$\hat{\pi}^{(\text{uni})}$	$\hat{\pi}^{(\text{outdeg})}$	$\hat{\pi}^{(\text{ren})}_{\alpha,\lambda}$	$\hat{\pi}^{(\text{ren})}$	s.d.
5	0.246	0.225	**0.214**	0.215	0.010
10	0.160	0.136	**0.127**	0.128	0.004
15	0.125	0.105	**0.098**	0.099	0.003

(d) $\alpha = 0.75$

	$\hat{\pi}^{(\text{uni})}$	$\hat{\pi}^{(\text{outdeg})}$	$\hat{\pi}^{(\text{ren})}_{\alpha,\lambda}$	$\hat{\pi}^{(\text{ren})}$	s.d.
	0.303	0.318	**0.294**	0.295	0.014
	0.188	0.201	**0.177**	0.178	0.006
	0.144	0.156	**0.135**	0.135	0.004

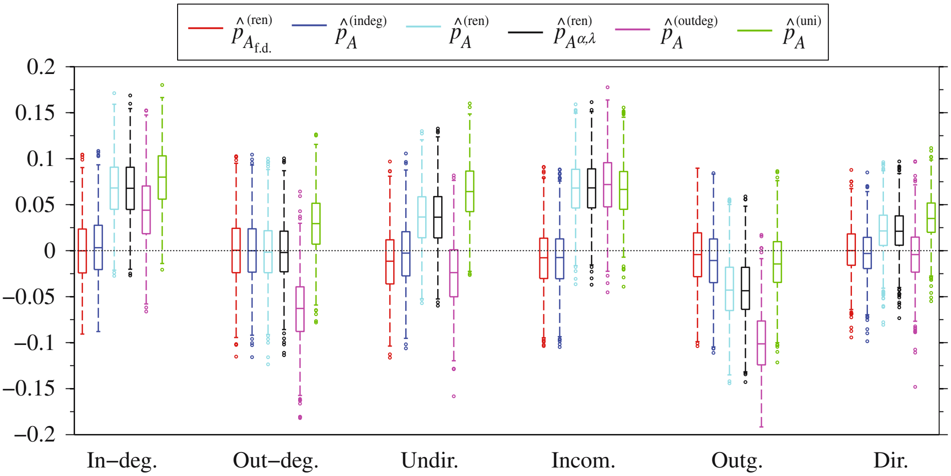


Fig. 3. Deviations of estimated $\hat{p}_A$ from the true value in the directed Erdős-Rényi graphs with $N = 1{,}000$, $\alpha = 0.75$, $\lambda = 10$, $p = 0.5$, and $n_s = 500$. Each group of boxplots corresponds to $\hat{p}^{(ren)}_{A_{f.d.}}$, $\hat{p}^{(indeg)}_{A}$, $\hat{p}^{(ren)}_{A}$, $\hat{p}^{(ren)}_{A_{\alpha,\lambda}}$, $\hat{p}^{(outdeg)}_{A}$, and $\hat{p}^{(uni)}_{A}$ for one allocation of the individual property A. The abbreviations for the allocations corresponds to the function g, that is, In-deg. equals $\left(d_i^{(un)} + d_i^{(in)}\right)$, Out-deg. $\left(d_i^{(un)} + d_i^{(out)}\right)$, Undir. $d_i^{(un)}$, Incom. $d_i^{(in)}$, Outg. $d_i^{(out)}$, and Dir. $\left(d_i^{(in)} + d_i^{(out)}\right)$.

To compare estimated p_A, we generated 1,000 networks for each combination of the parameters $\alpha \in \{0.25,\ 0.5,\ 0.75\}$ and $\lambda = 10$. On each of these networks we in turn allocate the property A in each of the six ways described in Section 1. The probability of a vertex having A is denoted by $p \in \{0.2,\ 0.5\}$. For each network and allocation, we simulate a random walk with length $n_s \in \{200,\ 500\}$ and calculate the differences between the estimated and the actual proportions of the population with property A. In Figure 3, results for $\alpha = 0.75$, $p = 0.5$, and $n_s = 500$ are shown. The six groups of boxplots correspond to the six different ways of allocating A (see Section 1). The six boxplots in each group correspond to $\hat{p}^{(\mathrm{ren})}_{A_{\mathrm{f.d.}}}$, $\hat{p}^{(\mathrm{indeg})}_{A}$, $\hat{p}^{(\mathrm{ren})}_{A}$, $\hat{p}^{(\mathrm{ren})}_{A_{\alpha,\lambda}}$, $\hat{p}^{(\mathrm{outdeg})}_{A}$, and $\hat{p}^{(\mathrm{uni})}_{A}$, respectively.

We see that the bias of $\hat{p}^{(\mathrm{ren})}_{A_{\mathrm{f.d}}}$ and $\hat{p}^{(\mathrm{indeg})}_{A}$ is small for all allocations, as is to be expected. For the estimators utilizing the out-degree, $\hat{p}^{(\mathrm{ren})}_{A}$, $\hat{p}^{(\mathrm{ren})}_{A_{\alpha,\lambda}}$, and $\hat{p}^{(\mathrm{outdeg})}_{A}$, Figure 3 indicates that the choice of how to allocate A has a significant impact on the performance of the estimators. When A is allocated proportional to the out-degree (Out-deg. in Figure 3), $\hat{p}^{(\mathrm{ren})}_{A}$ and $\hat{p}^{(\mathrm{ren})}_{A_{\alpha,\lambda}}$ yield the most accurate result, and when A is allocated proportional to the number of directed edges (Dir. in Figure 3), $\hat{p}^{(\mathrm{outdeg})}_{A}$ is most accurate. This is true for almost all parameter combinations. In general, the bias and variance increase with both α and p for all estimators, and a small n_s results in an increased variance, as is to be expected. In the supplemental data, these findings are further illustrated by numerical results with (α, p, s) equal to (0.5, 0.2, 500), (0.25, 0.5, 500), and (0.75, 0.5, 200). The supplemental file is available at: http://dx.doi.org/jos-2016-0023

5.2. *Networks With Power-Law Degree Distributions*

To generate power-law networks, we set the expected total number of edges for each vertex to 16, while we set the expected number of undirected and directed edges equal to ($E(D^{\ (\mathrm{un})}$),

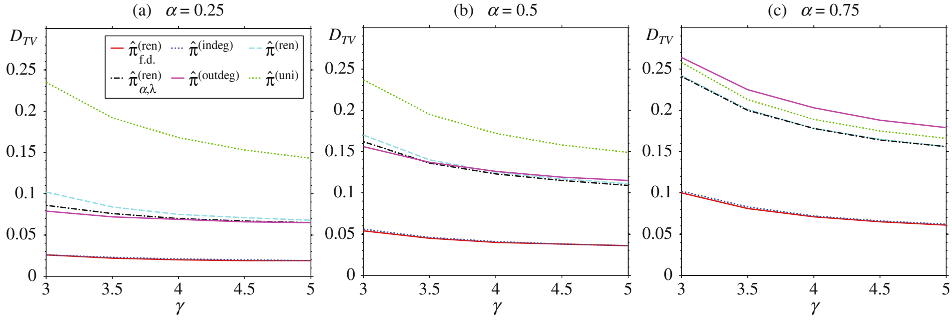


Fig. 4. Average D_{TV} between the true stationary distribution and $\hat{\pi}^{(ren)}_{f.d.}$, $\hat{\pi}^{(indeg)}$, $\hat{\pi}^{(ren)}$, $\hat{\pi}^{(ren)}_{\alpha,\lambda}$, $\hat{\pi}^{(outdeg)}$, and $\hat{\pi}^{(uni)}$ in the power-law networks with N = 1,000, α equal to a) 0.25, b) 0.5, and c) 0.75, and $n_s = 500$.

$E(D^{(in)} + D^{(out)})) = (12, 4)$, (8, 8), and (4, 12). The three cases yield $\alpha = 0.25$, 0.5, and 0.75, respectively. For each combination of the parameters, we generate 1,000 networks of size $N = 1{,}000$ and calculate the mean of the D_{TV}. We also calculate the s.d., which is of magnitude 10^{-3} and therefore not shown. The sample size n_s is set to 200 and 500.

The average D_{TV} values for $\hat{\pi}^{(ren)}_{f.d.}$, $\hat{\pi}^{(indeg)}$, $\hat{\pi}^{(ren)}$, $\hat{\pi}^{(ren)}_{\alpha,\lambda}$, $\hat{\pi}^{(outdeg)}$, and $\hat{\pi}^{(uni)}$ are shown in Figure 4 for various α and γ values. Figure 4 suggests that $\hat{\pi}^{(ren)}_{f.d}$ and $\hat{\pi}^{(indeg)}$ are the most accurate among the four estimators, with $\hat{\pi}^{(ren)}_{f.d}$ being slightly better. When $\alpha = 0.25$ and 0.5, $\hat{\pi}^{(ren)}_{\alpha,\lambda}$ has a lower mean D_{TV} than $\hat{\pi}^{(ren)}$, but this difference is not seen when $\alpha = 0.75$. $\hat{\pi}^{(outdeg)}$ performs better than $\hat{\pi}^{(ren)}$ for all values of γ when $\alpha = 0.25$, and the opposite result holds true when $\alpha = 0.75$.

In Figure 5, the results for $\hat{p}^{(ren)}_{A_{f.d.}}$, $\hat{p}^{(indeg)}_A$, $\hat{p}^{(ren)}_A$, $\hat{p}^{(ren)}_{A_{\alpha,\lambda}}$, $\hat{p}^{(outdeg)}_A$, and $\hat{p}^{(uni)}_A$ when $\gamma = 3$, $E(D^{(un)}) = 4$, $E(D^{(in)} + D^{(out)}) = 12$, $p = 0.2$, and $n_s = 500$ are shown. The figure indicates that $\hat{p}^{(ren)}_{A_{f.d.}}$ and $\hat{p}^{(indeg)}_A$ have small bias across different allocations of A. In contrast, the magnitude of the bias of $\hat{p}^{(ren)}_A$, $\hat{p}^{(ren)}_{A_{\alpha,\lambda}}$, and $\hat{p}^{(outdeg)}_A$ depends on the allocation type; $\hat{p}^{(ren)}_A$ has the smallest bias when A is allocated proportional to the undirected degree, and $\hat{p}^{(ren)}_{A_{\alpha,\lambda}}$

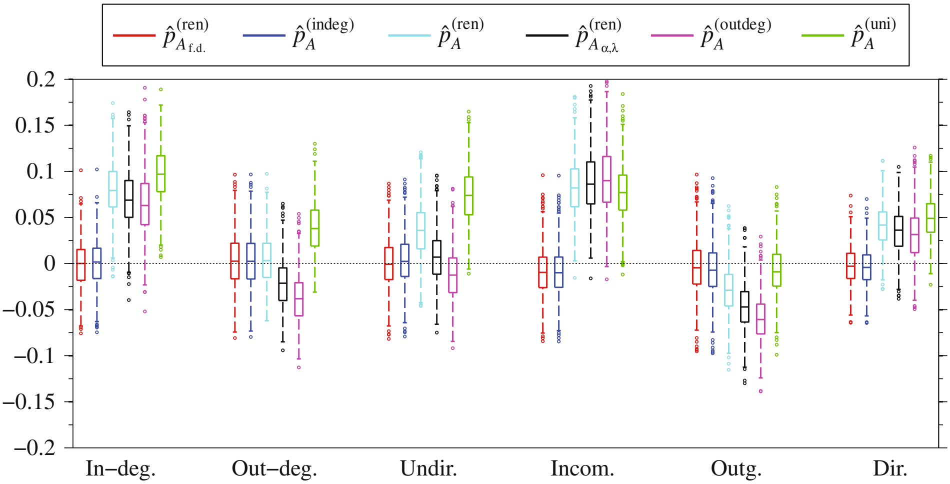


Fig. 5. Deviations of estimated p_A from the true population proportion in the power-law networks for $\gamma = 3$, $E(D^{(un)}) = 4$, $E(D^{(in)} + D^{(out)}) = 12$, $p = 0.2$, and $n_s = 500$. Each group of boxplots corresponds to $\hat{p}^{(ren)}_{A_{f.d.}}$, $\hat{p}^{(indeg)}_A$, $\hat{p}^{(ren)}_A$, $\hat{p}^{(ren)}_{A_{\alpha,\lambda}}$, $\hat{p}^{(outdeg)}_A$, and $\hat{p}^{(uni)}_A$, for one allocation of A.

Table 4. *D_{TV} between the true stationary distribution and $\hat{\pi}^{(ren)}_{f.d.}$, $\hat{\pi}^{(indeg)}$, $\hat{\pi}^{(ren)}$, $\hat{\pi}^{(outdeg)}$ and $\hat{\pi}^{(uni)}$. S.d. is shown in the second row, but only applies to $\hat{\pi}^{(ren)}_{f.d.}$ and $\hat{\pi}^{(ren)}$.*

$\hat{\pi}^{(\text{ren})}_{\text{f.d.}}$	$\hat{\pi}^{(\text{indeg})}$	$\hat{\pi}^{(\text{ren})}$	$\hat{\pi}^{(\text{outdeg})}$	$\hat{\pi}^{(\text{uni})}$
0.2198	0.2248	0.4057	0.4290	0.4484
0.0004	–	0.0048	–	

and $\hat{p}_A^{(\text{outdeg})}$ when A is allocated proportional to the out-degree. Their relative performance is hard to assess for other allocations. In general, a large fraction of directed edges, small γ, and large p increase bias and variance, and variance decreases with n_s. The supplemental data contains numerical results for $(\gamma, E(D^{(\text{un})}), E(D^{(\text{in})} + D^{(\text{out})}), p, s) = (4.5, 4, 12, 0.2, 500), (4.5, 4, 12, 0.5, 500), (4.5, 12, 4, 0.5, 500)$, and $(3, 4, 12, 0.2, 200)$ to further support these results.

5.3. *Online MSM Network*

For the Qruiser online MSM network, we first evaluate $\hat{\pi}^{(\text{ren})}_{\text{f.d.}}$, $\hat{\pi}^{(\text{indeg})}$, $\hat{\pi}^{(\text{ren})}$, $\hat{\pi}^{(\text{outdeg})}$ and $\hat{\pi}^{(\text{uni})}$. The results are shown in Table 4. Note that $\hat{\pi}^{(\text{ren})}_{\alpha,\lambda}$ is not evaluated because α and λ are not known beforehand. For $\hat{\pi}^{(\text{uni})}$, $\hat{\pi}^{(\text{outdeg})}$, and $\hat{\pi}^{(\text{indeg})}$, D_{TV} to the true selection probabilities is exactly calculated. For $\hat{\pi}^{(\text{ren})}_{\text{f.d.}}$ and $\hat{\pi}^{(\text{ren})}$, we show the mean and s.d. of D_{TV} on the basis of 1,000 samples of size 500. We see that $\hat{\pi}^{(\text{ren})}_{\text{f.d.}}$ has smaller D_{TV} than $\hat{\pi}^{(\text{indeg})}$, and that the mean D_{TV} of $\hat{\pi}^{(\text{ren})}$ is smaller than that of $\hat{\pi}^{(\text{uni})}$ and $\hat{\pi}^{(\text{outdeg})}$.

In Figure 6, we show estimates of the population proportions of the age, county, civil status, and profession properties. The true population proportions are shown by the dashed lines. The sample size is 500. Figure 6 indicates that $\hat{p}^{(\text{ren})}_{A_{\text{f.d.}}}$ performs best of all estimators. Among the estimators utilizing $d_i^{(\text{un})} + d_i^{(\text{out})}$, $\hat{p}_A^{(\text{ren})}$ has the smallest overall bias. Moreover, the variance of $\hat{p}_A^{(\text{ren})}$ is smaller than for $\hat{p}_A^{(\text{outdeg})}$ for all properties, in particular the civil status.

6. Conclusion and Discussion

We developed statistical procedures for the random walk on directed networks to account for the empirical fact that social networks generally include nonreciprocal edges. The proposed estimation procedures typically outperformed the considered existing methods that neglect directed edges in the scenarios investigated in the simulations. In the present study, the best accuracy of estimation was obtained when undirected, incoming, and outgoing degree are observed separately for sampled individuals. In this case, our estimator $\hat{\pi}^{(\text{ren})}_{\text{f.d.}}$ should be compared to $\hat{\pi}^{(\text{indeg})}$ when the expectations of the degree distributions are known. In Tables 2 and 4, and Figure 4, it is seen that $\hat{\pi}^{(\text{ren})}_{\text{f.d.}}$ performs slightly better than $\hat{\pi}^{(\text{indeg})}$ in all the studied situations. The corresponding estimated proportions given by $\hat{p}^{(\text{ren})}_{A_{\text{f.d.}}}$ and $\hat{p}_A^{(\text{indeg})}$ in Figures 3, 5, and 6 are very similar. In the more realistic scenario in which only the sum of undirected and outgoing edges of sampled individuals is known, all estimation procedures are less precise. In this situation, we compare our new estimator $\hat{\pi}^{(\text{ren})}$ with the estimator $\hat{\pi}^{(\text{outdeg})}$ that one would use if ignoring the direction of edges (Tables 3 and 4, and Figure 4). We also include $\hat{\pi}^{(\text{ren})}_{\alpha,\lambda}$ in the

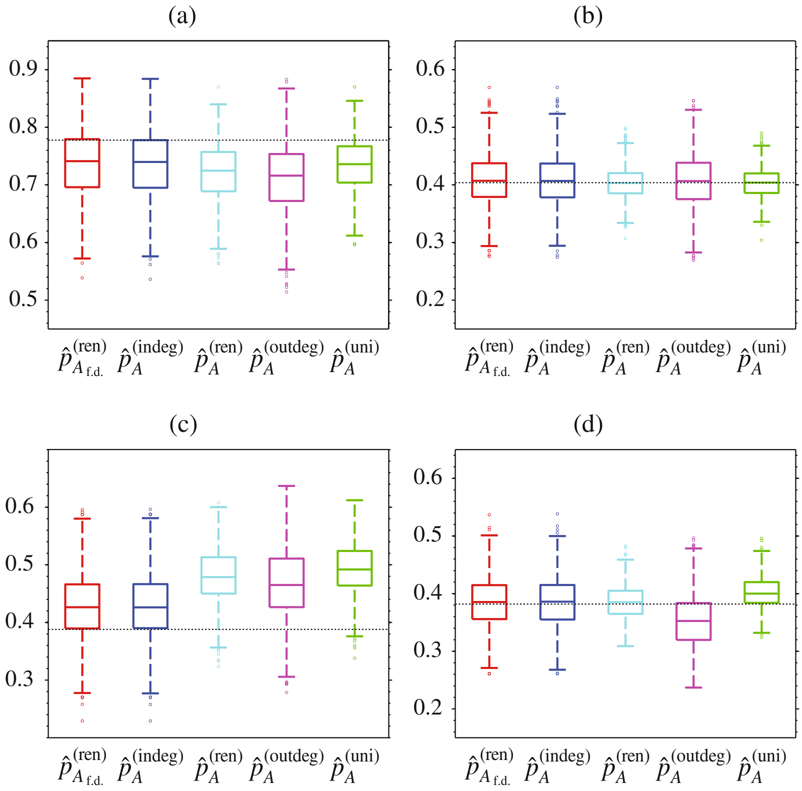


Fig. 6. Estimates of population proportions in the Qruiser network for a) age, b) civil status, c) county, and d) profession. Each figure shows $\hat{p}^{(ren)}_{A_{f.d.}}$, $\hat{p}^{(indeg)}_A$, $\hat{p}^{(ren)}_A$, $\hat{p}^{(outdeg)}_A$, and $\hat{p}^{(uni)}_A$. The true population proportions are shown by the dashed lines and are equal to 0.77, 0.40, 0.39, and 0.38 for age, civil status, county, and profession, respectively.

comparison for the generated networks, and it can be seen that the performance of $\hat{\pi}^{(\text{ren})}_{\alpha,\lambda}$ is only slightly better than that of $\hat{\pi}^{(\text{ren})}$. Because $\hat{\pi}^{(\text{ren})}$ will deviate further from $\hat{\pi}^{(\text{outdeg})}$ when $\hat{\alpha}$ increases, as seen in Equation (14), it outperforms $\hat{\pi}^{(\text{outdeg})}$ except when the fraction of directed edges α is small (0.1 in Table 3 and 0.25 in Figure 4). Our simulations showed that estimators of population proportions were highly sensitive to how the property of interest is allocated in the social network. For example, Figures 3 and 5 indicate that the results of the estimators $\hat{p}^{(\text{ren})}_A$, $\hat{p}^{(\text{ren})}_{A_{\alpha,\lambda}}$, and $\hat{p}^{(\text{outdeg})}_A$ depend strongly on the allocation of the property A. We believe that the question of how properties are distributed in empirical social networks is of interest to further study.

It is generally believed that recruitment does not happen over nonreciprocal edges in RDS, which is however refuted by the examples in Section 1. Furthermore, recruitment over nonreciprocal edges may occur on a relatively large scale in the presence of coupon selling, that is, when respondents trade coupons instead of randomly distributing them among their peers in order to increase their personal profit from study participation. Coupon selling is a side effect of the dual incentive system of RDS. It has been observed by, for example, Scott (2008) in an RDS study of IDUs in Chicago, where interviews with participants indicated that coupon selling was common and also that it had side effects such as increased risk exposure and violence among participants (see also Broadhead 2008; Ouellet 2008). For additional examples of coupon selling, see Johnston et al. (2008) and the references therein, where implications of the size of the incentives and the practical study setup on coupon selling are also discussed. In RDS studies where there is evidence of coupon selling, it might be difficult to obtain valid information on the occurrence of nonreciprocal recruitments, and then the possibility of such recruitments should be taken into account for estimation.

Information on the nonreciprocal edges in the network can be obtained from several sources. The fraction of directed edges, α, may be known for some social networks, and then we can estimate the total mean degree λ using only the mean sample out-degree in Equation (13). If α is not known, it may be estimated by utilizing additional information from an RDS sample. As previously discussed in Section 1, in the majority of RDS studies respondents quantify the nature of the relationship with their recruiter. Through this, the proportion of recruitments that occur over nonreciprocal edges (i.e., coupons passed from strangers) can be obtained and used as an estimate of α in Equation (13). In Gile, Johnston, and Salganik (2015), an alternative estimation procedure is given. This procedure utilizes several questions on respondents' degrees that serve to calculate the differences between the number of incoming and outgoing edges, which are then used to produce an estimate. However, as the authors point out, this procedure may be subject to large reporting errors. Additionally, an alternative to the standard formulation for assessing reciprocation is given in the same paper. It is also possible to estimate α through information on the number of revisits m used in Equation (12). This could be done by asking, for example, "Would you give a coupon to the person who gave you a coupon if he or she had not yet participated in the study?". This has been done in RDS studies (e.g., Bui et al. 2014), but the question may be cognitively difficult for respondents.

An alternative strategy would be to develop a sampling procedure that accounts for a directed social network of the population, that is, in which it is possible to determine whether an edge is undirected, incoming, or outgoing from a vertex, and then utilize this information for estimation. For example, in some RDS studies, the characteristics of neighbors of respondents have been collected (see Lu 2013 and the references therein). If such data were also to include, for example, the number of undirected, incoming, and outgoing edges of an individual, they could be useful in RDS estimation. As previously noted, however, it is difficult for respondents to provide such data. Alternatively, the sampling procedure could be adapted to the case of directed social networks by encouraging respondents to recruit people that are less known to them. Then, one could expect that recruitment takes place on nonreciprocal edges to a larger extent and possibly more easily identify and account for these recruitments in estimation. However, such a sampling scheme may reduce the ability of RDS to successfully penetrate the population, and may also suffer from difficulties in deciding on edge directions from sampled data.

In the present study, we considered RDS estimators that are based on the random walk framework for estimation. It could also be of interest to consider the RDS estimators of Gile (2011), Gile and Handcock (2015), Lu et al. (2013), and Lu (2013) mentioned in Section 1 for the situations studied in this article. The estimator of Gile (2011), while not adapted to the case of directed networks, is in a sense a combination of the $\hat{p}_A^{(\mathrm{outdeg})}$ and

$\hat{p}_A^{(\mathrm{uni})}$ estimators. Hence, it can be expected to perform better than our estimator in cases where a combination of these two estimators would be favorable (e.g., when A is allocated proportional to out-degree in Figure 5 in the supplemental data), given that prior information on the population size is available. The model-assisted approach of Gile and Handcock (2015) incorporates network structural properties through an exponential random-graph model (ERGM) (e.g., Robins et al. 2007) for the network. Hence, it might be less sensitive to the different allocations of the property A that were seen to have relatively large effects on the estimators considered in our simulations. Additionally, the ERGM should not be difficult to extend to the case of directed networks. The estimator in Lu et al. (2013) is similar to our estimators in that it is developed for directed networks and could be expected to perform similarly to $\hat{p}_A^{(\mathrm{indeg})}$ given that prior information on the ratio of average in-degrees of groups in the network is available. The estimator of Lu (2013) has performed well in a recent evaluation (Verdery et al. 2015) and it could be of interest to extend it to the case of directed networks. In future work, it would be of interest to make a comprehensive evaluation of the performance of the estimators presented in this article as well as other RDS estimators, both random walk-based and nonrandom walk-based, on simulated RDS samples and data from actual RDS studies.

The main focus of the present article was on accounting for directed edges in a social network when performing RDS. There are also other assumptions in existing estimation procedures (including the current one) worthy of relaxing. For example, the methods typically assume that participants choose coupon recipients uniformly at random among their neighbors in the social network. In reality, they probably are more likely to sample closely connected neighbors, which may bias estimators of selection probabilities. Extending the RDS methods by allowing weighted edges warrants future work. It should be noted that our methods allow the two weights on the same undirected edge in the opposite directions to be different, because our framework targets directed networks. Alternatively, it is also possible that some of the previously mentioned recently developed estimators could be extended to the case of directed weighted networks.

7. References

Abramovitz, D., E.M. Volz, S.A. Strathdee, T.L. Patterson, A. Vera, and S.D. Frost. 2009. "Using Respondent Driven Sampling in a Hidden Population at Risk of HIV Infection: Who Do HIV-Positive Recruiters Recruit?" *Sexually Transmitted Diseases* 36: 750–756. Doi: http://dx.doi.org/10.1097/OLQ.0b013e3181b0f311.

Bernhardt, A., M.W. Spiller, and D. Polson. 2013. "All Work and No Pay: Violations of Employment and Labor Laws in Chicago, Los Angeles and New York City." *Social Forces* 91: 725–746. Doi: http://dx.doi.org/10.1093/sf/sos193.

Boldi, P., M. Rosa, M. Santini, and S. Vigna. 2011. "Layered Label Propagation: A Multiresolution Coordinate-Free Ordering for Compressing Social Networks." In Proceedings of the 20th International Conference on World Wide Web. 587–596. Available at: dl.acm.org/citation.cfm?id=1963405. (accessed Feb 2014).

Boldi, P. and S. Vigna. 2004. "The Webgraph Framework I: Compression Techniques." In Proceedings of the 13th International Conference on World Wide Web. 595–602. Available at: dl.acm.org/citation.cfm?id=988672. (accessed Feb 2014).

Broadhead, R.S. 2008. “Notes on a Cautionary (Tall) Tale About Respondent-Driven Sampling: A Critique of Scott’s Ethnography.” *The International Journal of Drug Policy* 19: 235–237. Doi: http://dx.doi.org/10.1016/j.drugpo.2008.02.014.

Bui, T., J. Nyoni, M. Ross, J. Mbwambo, C. Markham, and S. McCurdy. 2014. “Sexual Motivation, Sexual Transactions and Sexual Risk Behaviors in Men Who Have Sex with Men in Dar es Salaam, Tanzania.” *AIDS and Behavior* 18: 2432–2441. Doi: http://dx.doi.org/10.1007/s10461-014-0808-x.

Chung, F. and L.Y. Lu. 2002. “The Average Distances in Random Graphs with Given Expected Degrees.” *Proceedings of the National Academy of Sciences of the United States of America* 99: 15879–15882. Doi: http://dx.doi.org/10.1073/pnas.252631999.

Chung, F., L.Y. Lu, and V. Vu. 2003. “Spectra of Random Graphs with Given Expected Degrees.” *Proceedings of the National Academy of Sciences of the United States of America* 100: 6313–6318. Doi: http://dx.doi.org/10.1073/pnas.0937490100.

Deaux, E. and J. Callaghan. 1985. “Key Informant Versus Self-Report Estimates of Health-Risk.” *Evaluation Review* 9: 365–368. Doi: http://dx.doi.org/10.1177/0193841X8500900308.

Dombrowski, K., B. Khan, J. Moses, E. Channell, and E. Misshula. 2013. “Assessing Respondent Driven Sampling for Network Studies in Ethnographic Contexts.” *Advances in Anthropology* 3: 1–9. Doi: http://dx.doi.org/10.4236/aa.2013.31001.

Donato, D., L. Laura, S. Leonardi, and S. Millozzi. 2004. “Large Scale Properties of the Webgraph.” *European Physical Journal B* 38: 239–243. Doi: http://dx.doi.org/10.1140/epjb/e2004-00056-6.

Doyle, P.G. and J.L. Snell. 1984. *Random Walks and Electric Networks*. The Mathematical Association of America: Washington.

Erdős, P. and A. Renyi. 1960. “On the Evolution of Random Graphs.” *Publications of the Mathematical Institute of the Hungarian Academy of Science* 5: 17–61.

Erickson, B.H. 1979. “Some Problems of Inference from Chain Data.” *Sociological Methodology* 10: 276–302. Doi: http://dx.doi.org/10.2307/270774.

Feller, W. 1950. *An Introduction to Probability Theory and Its Applications*, Vol. 1. New York: Wiley.

Fortunato, S., M. Boguñá, A. Flammini, and F. Menczer. 2008. “Approximating PageRank from In-Degree.” In *Algorithms and Models for the Web-Graph*, edited by W. Aiello, A. Broder, J. Janssen, and E. Milios, 59–71. Heidelberg: Springer.

Freeman, L.C., C.M. Webster, and D.M. Kirke. 1998. “Exploring Social Structure Using Dynamic Three-Dimensional Color Images.” *Social Networks* 20: 109–118. Doi: http://dx.doi.org/10.1016/S0378-8733(9700016-6).

Ghoshal, G. and A.L. Barabási. 2011. “Ranking Stability and Super-Stable Nodes in Complex Networks.” *Nature Communications* 2: 394. Doi: http://dx.doi.org/10.1038/ncomms1396.

Gile, K.J. 2011. “Improved Inference for Respondent-Driven Sampling Data with Application to HIV Prevalence Estimation.” *Journal of the Americal Statistical Association* 106: 135–146. Doi: http://dx.doi.org/10.1198/jasa.2011.ap09475.

Gile, K.J. and M.S. Handcock. 2010. “Respondent-Driven Sampling: An Assessment of Current Methodology.” *Sociological Methodology* 40: 285–327. Doi: http://dx.doi.org/10.1111/j.1467-9531.2010.01223.x.

Gile, K.J. and M.S. Handcock. 2015. "Network Model-Assisted Inference from Respondent-Driven Sampling Data." *Journal of the Royal Statistical Society A* 178: 619–639. Doi: http://dx.doi.org/10.1111/rssa.12091.

Gile, K.J., L.G. Johnston, and M.J. Salganik. 2015. "Diagnostics for Respondent-Driven Sampling." *Journal of the Royal Statistical Society A* 178: 241–269. Doi: http://dx.doi.org/10.1111/rssa.12059.

Gleiser, P. and L. Danon. 2003. "Community Structure in Jazz." *Advances in Complex Systems* 6: 565–573. Doi: http://dx.doi.org/10.1142/S0219525903001067.

Goel, S. and M.J. Salganik. 2010. "Assessing Respondent-Driven Sampling." *Proceedings of the National Academy of Sciences of the United States of America* 107: 6743–6747. Doi: http://dx.doi.org/10.1073/pnas.1000261107.

Goh, K.I., B. Kahng, and D. Kim. 2001. "Universal Behavior of Load Distribution in Scale-Free Networks." *Physical Review Letters* 87: 278701-4. Doi: http://dx.doi.org/10.1103/PhysRevLett.87.278701.

Gong, N.Z. and W. Xu. 2014. "Reciprocal Versus Parasocial Relationships in Online Social Networks." *Social Network Analysis and Mining* 4: 1–14. Doi: http://dx.doi.org/10.1007/s13278-014-0184-6.

Hakre, S., G. Arteaga, A. Núñez, N. Arambu, B. Aumakhan, M. Liu, S. Peel, J. Pascale, and P. Scott. 2014. "Prevalence of HIV, Syphilis, and Other Sexually Transmitted Infections among MSM from Three Cities in Panama." *Journal of the Urban Health* 91: 793–808. Doi: http://dx.doi.org/10.1007/s11524-014-9885-4.

Heckathorn, D.D. 1997. "Respondent-Driven Sampling: A New Approach to the Study of Hidden Populations." *Social Problems* 44: 174–199. Doi: http://dx.doi.org/10.2307/3096941.

Hobkirk, A.L., M.H. Watt, K.T. Green, J.C. Beckham, D. Skinner, and C.S. Meade. 2015. "Mediators of Interpersonal Violence and Drug Addiction Severity Among Methamphetamine Users in Cape Town, South Africa." *Addictive Behaviors* 42: 167–171. Doi: http://dx.doi.org/10.1016/j.addbeh.2014.11.030.

Johnston, L.G., M. Malekinejad, C. Kendall, I.M. Iuppa, and G.W. Rutherford. 2008. "Implementation Challenges to Using Respondent-Driven Sampling Methodology for HIV Biological and Behavioral Surveillance: Field Experiences in International Settings." *AIDS and Behavior* 12: 131–141. Doi: http://dx.doi.org/10.1007/s10461-008-9413-1.

Kazerooni, P.A., N. Motazedian, M. Motamedifar, M. Sayadi, M. Sabet, M.A. Lari, and K. Kamali. 2013. "The Prevalence of Human Immunodeficiency Virus and Sexually Transmitted Infections Among Female Sex Workers in Shiraz, South of Iran: By Respondent-Driven Sampling." *International Journal of STD and AIDS* 25: 155–161. Doi: http://dx.doi.org/10.1177/0956462413496227.

Killworth, P.D. and H.R. Bernard. 1976. "Informant Accuracy in Social Network Data." *Human Organization* 35: 269–286.

Kwak, H., C. Lee, H. Park, and S. Moon. 2010. "What is Twitter, a Social Network or a News Media?" In Proceedings of the 19th International Conference on World Wide Web. 591–600. Available at: dl.acm.org/citation.cfm?id=1772690 (accessed Feb 2014).

Langville, A.N., and C.D. Meyer. 2006. *Google's PageRank and Beyond*. Princeton: Princeton University Press.

Levin, D.A., Y. Peres, and E.L. Wilmer. 2009. *Markov Chains and Mixing Times*. Providence: American Mathematical Society.

Lovász, L. 1993. "Random Walks on Graphs: A Survey." *Bolyai Society Mathematical Studies* 2: 1–46.

Lu, X. 2013. "Linked Ego Networks: Improving Estimate Reliability and Validity with Respondent-Driven Sampling." *Social Networks* 35: 669–685. Doi: http://dx.doi.org/10.1016/j.socnet.2013.10.001.

Lu, X., L. Bengtsson, T. Britton, M. Camitz, B.J. Kim, A. Thorson, and F. Liljeros. 2012. "The Sensitivity of Respondent-Driven Sampling." *Journal of the Royal Statistical Society A* 175: 191–216. Doi: http://dx.doi.org/10.1111/j.1467-985X.2011.00711.x.

Lu, X., J. Malmros, F. Liljeros, and T. Britton. 2013. "Respondent-Driven Sampling on Directed Networks." *Electronic Journal of Statistics* 7: 292–322. Doi: http://dx.doi.org/10.1214/13-EJS772.

Magnani, R., K. Sabin, T. Saidel, and D. Heckathorn. 2005. "Review of Sampling Hard-to-Reach and Hidden Populations for HIV Surveillance." *AIDS* 19 (Supplement 2): 67–72. Doi: http://dx.doi.org/10.1097/01.aids.0000172879.20628.e1.

Marsden, P.V. 1990. "Network Data and Measurement." *Annual Review of Sociology* 16: 435–463. Doi: http://dx.doi.org/10.1146/annurev.so.16.080190.002251.

Masuda, N. and H. Ohtsuki. 2009. "Evolutionary Dynamics and Fixation Probabilities in Directed Networks." *New Journal of Physics* 11: 033012. Doi: http://dx.doi.org/10.1088/1367-2630/11/3/033012.

McCreesh, N., S.D.W. Frost, J. Seeley, J. Katongole, M.N. Tarsh, R. Ndunguse, F. Jichi, N.L. Lunel, D. Maher, L.G. Johnston, P. Sonnenberg, A.J. Copas, R.J. Hayes, and R.G. White. 2012. "Evaluation of Respondent-Driven Sampling." *Epidemiology* 23: 138–147. Doi: http://dx.doi.org/10.1097/EDE.0b013e31823ac17c.

Mislove, A., M. Marcon, K.P. Gummadi, P. Druschel, and B. Bhattacharjee. 2007. "Measurement and Analysis of Online Social Networks." In Proceedings of the 7th ACM SIGCOMM Conference on Internet measurement. 29–42. October 23–26, 2007 San Diego, CA, USA. Available at: dl.acm.org/citation.cfm?id=1298306 (accessed Feb 2014).

Moreno, J.L. 1960. *The Sociometry Reader*. New York: Free Press.

Muhib, F.B., L.S. Lin, A. Stueve, R.L. Miller, W.L. Ford, W.D. Johnson, and P.J. Smith, Community Intervention Trial for Youth Study Team. 2001. "A Venue-Based Method for Sampling Hard-to-Reach Populations." *Public Health Reports* 116 (Suppl. 1): 216–222.

Newman, M. 2010. *Networks: an Introduction*. Oxford: Oxford University Press.

Newman, M.E., S. Forrest, and J. Balthrop. 2002. "Email Networks and the Spread of Computer Viruses." *Physical Review E* 66: 035101. Doi: http://dx.doi.org/10.1103/PhysRevE.66.035101.

Ouellet, L.J. 2008. "Cautionary Comments on an Ethnographic Tale Gone Wrong." *International Journal of Drug Policy* 19: 238–240. Doi: http://dx.doi.org/10.1016/j.drugpo.2008.02.013.

Paquette, D.M., J. Bryant, and J.D. Wit. 2011. "Use of Respondent-Driven Sampling to Enhance Understanding of Injecting Networks: A Study of People Who Inject Drugs in

Sydney, Australia." *International Journal of Drug Policy* 22: 267–273. Doi: http://dx.doi.org/10.1016/j.drugpo.2011.03.007.

Phillips II, G., L.M. Kuhns, R. Garofalo, and B. Mustanski. 2014. "Do Recruitment Patterns of Young Men Who Have Sex With Men (YMSM) Recruited Through Respondent-Driven Sampling (RDS) Violate Assumptions?" *Journal of Epidemiology and Community Health* 68: 1207–1212. Doi: http://dx.doi.org/10.1136/jech-2014-204206.

Robins, G., P. Pattison, Y. Kalish, and D. Lusher. 2007. "An Introduction to Exponential Random Graph (p*) Models for Social Networks." *Social Networks* 29: 173–191. Doi: http://dx.doi.org/10.1016/j.socnet.2006.08.002.

Rybski, D., S.V. Buldyrev, S. Havlin, F. Liljeros, and H.A. Makse. 2009. "Scaling Laws of Human Interaction Activity." *Proceedings of the National Academy of Sciences of the United States of America* 106: 12640–12645. Doi: http://dx.doi.org/10.1073/pnas.0902667106.

Salganik, M.J. and D.D. Heckathorn. 2004. "Sampling and Estimation in Hidden Populations Using Respondent-Driven Sampling." *Sociological Methodology* 34: 193–240. Doi: http://dx.doi.org/10.1111/j.0081-1750.2004.00152.x.

Särndal, C.-E., B. Swensson, and J.H. Wretman. 1992. *Model Assisted Survey Sampling*. New York: Springer.

Schwitters, A., M. Swaminathan, D. Serwadda, M. Muyonga, R. Shiraishi, I. Benech, S. Mital, R. Bosa, G. Lubwama, and W. Hladik. 2012. "Prevalence of Rape and Client-Initiated Gender-Based Violence Among Female Sex Workers: Kampala, Uganda, 2012." *AIDS and Behavior* 19: 68–76. Doi: http://dx.doi.org/10.1007/s10461-014-0957-y.

Scott, G. 2008. "'They Got Their Program, and I Got Mine': A Cautionary Tale Concerning the Ethical Implications of Using Respondent-Driven Sampling to Study Injection Drug Users." *International Journal of Drug Policy* 19: 42–51. Doi: http://dx.doi.org/10.1016/j.drugpo.2007.11.014.

Solomon, S.S., S.H. Mehta, A.K. Srikrishnan, S. Solomon, A.M. McFall, O. Laeyendecker, D.D. Celentano, S.H. Iqbal, S. Anand, C.K. Vasudevan, S. Saravanan, G.M. Lucas, H.R. Kumar, M.S. Sulkowski, and T.C. Quinn. 2015. "Burden of Hepatitis C Virus Disease and Access to Hepatitis C Virus Services in People Who Inject Drugs in India: A Cross-Sectional Study." *The Lancet Infectious Diseases* 15: 36–45. Doi: http://dx.doi.org/10.1016/S1473-3099(14)71045-X.

Tomas, A. and K.J. Gile. 2011. "The Effect of Differential Recruitment, Non-Response and Non-Recruitment on Estimators for Respondent-Driven Sampling." *Electronic Journal of Statistics* 5: 899–934. Doi: http://dx.doi.org/10.1214/11-EJS630.

Van de Bunt, G., M. van Duijn, and T. Snijders. 1999. "Friendship Networks Through Time: An Actor-Oriented Dynamic Statistical Network Model." *Computational and Mathematical Organization Theory* 5: 167–192. Doi: http://dx.doi.org/10.1023/A:1009683123448.

Verdery, A.M., M.G. Merli, J. Moody, J.A. Smith, and J.C. Fisher. 2015. "Brief Report: Respondent-Driven Sampling Estimators Under Real and Theoretical Recruitment Conditions of Female Sex Workers in China." *Epidemiology* 26: 661–665. Doi: http://dx.doi.org/10.1097/EDE.0000000000000335.

Volz, E. and D.D. Heckathorn. 2008. "Probability Based Estimation Theory for Respondent Driven Sampling." *Journal of Official Statistics* 24: 79–97.

Wasserman, S. and K. Faust. 1994. *Social Network Analysis*. New York: Cambridge University Press.

Zhang, S.X., M.W. Spiller, B.K. Finch, and Y. Qin. 2014. "Estimating Labor Trafficking among Unauthorized Migrant Workers in San Diego." *The Annals of the American Academy of Political and Social Science* 653: 65–86. Doi: http://dx.doi.org/10.1177/0002716213519237.

Received February 2014
Revised October 2015
Accepted October 2015