Mathematical Statistics
Stockholm University

Multiple seed structure and disconnected networks in respondent-driven sampling

Jens Malmros
Luis E.C. Rocha

Research Report 2016:5

ISSN 1650-0377

Postal address:

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

Internet:

http://www.math.su.se

Mathematical Statistics
Stockholm University
Research Report **2016:5**,
http://www.math.su.se

Multiple seed structure and disconnected networks in respondent-driven sampling

Jens Malmros*† Luis E.C. Rocha‡

March 2016

Abstract

Respondent-driven sampling (RDS) is a link-tracing sampling method that is especially suitable for sampling hidden populations. RDS combines an efficient snowball-type sampling scheme with inferential procedures that yield unbiased population estimates under some assumptions about the sampling procedure and population structure. Several seed individuals are typically used to initiate RDS recruitment. However, standard RDS estimation theory assume that all sampled individuals originate from only one seed. We present an estimator, based on a random walk with teleportation, which accounts for the multiple seed structure of RDS. The new estimator can also be used on populations with disconnected social networks. We numerically evaluate our estimator by simulations on artificial and real networks. Our estimator outperforms previous estimators, especially when the proportion of seeds in the sample is large. We recommend our new estimator to be used in RDS studies, in particular when the number of seeds is large or the social network of the population is disconnected.

Key words: Respondent-driven sampling; Seeds; Disconnected network; Random walk with teleportation.

*Department of Mathematics, Stockholm university, SE-106 91 Stockholm, Sweden
†Corresponding author: jensm@math.su.se
‡Karolinska Institutet, Stockholm, Sweden; Université de Namur, Belgium

Multiple seed structure and disconnected networks in respondent-driven sampling

Jens Malmros

Stockholm University, Sweden

Luis E.C. Rocha

Karolinska Institutet, Stockholm, Sweden

Université de Namur, Belgium

1. Introduction

Some human populations are difficult to survey for various reasons, for example, if no sampling frame for the population exists and the population is small relative to the general population, if members of the population are difficult to identify or unwilling to disclose themselves, or if individuals in the population are reluctant to participate in surveys. Examples of such *hidden* or *hard-to-survey* populations (Schwartländer et al., 2001; Tourangeau et al., 2014) include several groups that are subject to marginalisation or stigmatisation, e.g., injecting drug users, homosexual men, sex workers, illegal immigrants, and the homeless (Beyrer et al., 2012; Faugier and Sargeant, 1997; Sudman et al., 1988). Because of their characteristics, hidden populations can often not be satisfactorily investigated using standard sampling procedures and thus alternative sampling and estimation techniques must be considered (Magnani et al., 2005; Barros et al., 2015). A reasonably efficient and cost-effective way to sample from hidden populations is to utilise link-tracing techniques (Thompson, 1990; Thompson and Frank, 2000; Thompson, 2012). In such procedures, the population is assumed to be connected by a social network and previously sampled individuals are engaged in the recruitment of their social contacts to the sample. While

link-tracing procedures have been considered relatively efficient in collecting sufficiently sized samples from hidden populations, the samples obtained have often been viewed as convenience samples not suitable for inference because of the substantial bias that occurs from the selection procedure (Erickson, 1979).

A relatively recent and increasingly popular link-tracing methodology is *respondent-driven sampling* (RDS) (Heckathorn, 1997). The method is essentially an extension of snowball sampling (Biernacki and Waldorf, 1981) for which inferential procedures facilitating unbiased population estimates have been developed (Salganik and Heckathorn, 2004; Volz and Heckathorn, 2008). An RDS study begins with the formation of an initial group of individuals, the *seeds*, which are typically recruited among known population members. The seeds are provided with coupons, typically between three to five, which are to be distributed to their peers in the population of interest. An individual that has received a coupon is eligible for participation in the study upon presenting the coupon at the study site. After taking part in the study, sampled population members (i.e., respondents) are also given coupons which are to be distributed to those of their peers which have not yet participated in the study. This is repeated with subsequently sampled individuals until the desired sample size has been reached or until recruitment ceases by itself, in which case often additional seeds are recruited among not yet sampled members of the target population in order to re-initiate recruitment to the sample (Malekinejad et al., 2008). There are typically incentives given to individuals both for their own participation as well as for the participation of those to whom they have given coupons.

The most commonly used RDS estimator, the Volz-Heckathorn (V-H) estimator (Volz and Heckathorn, 2008), assumes that the RDS recruitment process can be approximated by a simple random walk on the social network of the population and also makes some assumptions about the structure of the social network. For example, it is assumed that sampling occurs with replacement, that respondents recruit randomly from their social contacts, and that the social relations in the population are mutual. It is also assumed that respondents

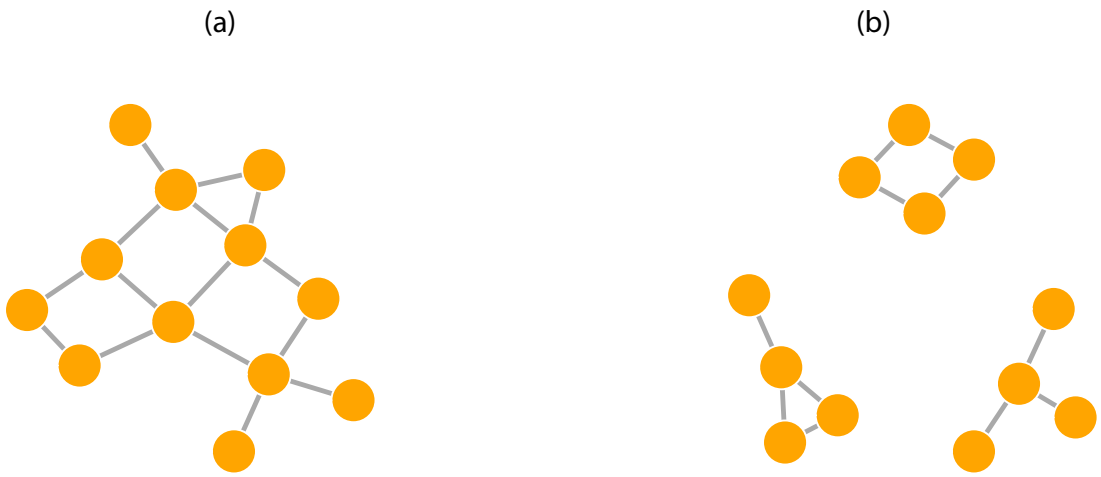

Figure 1. Schematic illustration of two social networks where the circles represent the vertices and the lines represent the edges. In (a), the network is connected in one component. In (b), the network is disconnected with three connected components.

accurately report their number of social contacts, or, in network lingo, their *degree*. In the general with-replacement design-based sampling framework, we can form an asymptotically unbiased Hansen-Hurwitz type ratio estimator (Hansen and Hurwitz, 1943) of the mean of a population trait y from a sample S as

$$\hat{\mu} = \frac{\sum_{u \in S} \frac{y_u}{p_u}}{\sum_{u \in S} \frac{1}{p_u}}, \tag{1}$$

where y_u and p_u are the values of y and the draw-wise selection probability for a sampled individual u, respectively. For the V-H estimator, $\hat{\mu}_{\mathrm{V-H}}$, p_u is replaced by d_u, since population members are assumed to be sampled with probability proportional to their degree from the random walk in stationarity in this case. V-H estimates have been shown to be sensitive to situations where the relatively strong assumptions on the recruitment process and the network structure do not hold (Gile and Handcock, 2010; Goel and Salganik, 2010; Lu et al., 2012; McCreesh et al., 2012; Tomas and Gile, 2011; Wejnert, 2009). In part due to these results, other RDS estimators have been developed (Gile, 2011; Lu et al., 2013; Lu, 2013; Gile and Handcock, 2015).

When the social network of the population is *connected*, the social relationships between individuals bind together all members of the population in one

connected *component* (Figure 1 a). Otherwise, the network is *disconnected* and has several connected components (Figure 1 b). It is in the nature of the simple random walk that it can not reach all population members in populations with disconnected social networks. This means that the V-H estimator is not applicable to such populations; a shortcoming that is shared with all proposed RDS estimators available in the literature. Additionally, the structural properties of the network may be such that a link-tracing sampling procedure is contained within parts of the network, something which is likely to affect RDS estimates even though the network is connected (Burt et al., 2010; McCreesh et al., 2011; Salganik, 2006). This may e.g. be the result of community structure within the network, where groups of individuals are more closely connected among each other than with individuals in different groups (Rocha et al., 2016) or so-called bottlenecks, where a single individual is the link between otherwise disconnected parts of the network (Johnston et al., 2013). Most RDS studies start with several seeds (e.g., 10) that each initiates a recruitment tree of its own. However, the simple random walk approximation assumes that only one recruitment chain can be used to describe the whole sample. This discrepancy becomes larger as the proportion of seeds in the sample increases, which may, e.g., be the result of additional seeds being recruited to the study. This is not uncommon; for example, in (Malekinejad et al., 2008), 43% of the reviewed empirical RDS studies with available data reported the use of additional seeds and in several studies, the sample consisted of more than 10% seeds. In some cases, the proportion of seeds in the sample may be as large as 18% (Stein et al., 2014) or even 30% (Strömdahl et al., 2015).

In this paper, we extend RDS estimation to account for the multiple seed structure and populations with disconnected social networks. We use a *random walk with teleportation* (RWWT) to model the RDS recruitment process (Brin and Page, 1998). The RWWT may, in each time step, go to a randomly chosen social contact of the currently visited individual, like the simple random walk used in the V-H estimator, or jump to a randomly chosen individual in the

population. Hence, the RWWT may visit parts of the social network that are not connected to each other and explore them separately, which is not possible in the simple random walk. Moreover, the set of individuals visited by the RWWT will be made up of multiple chains of neighbouring population members, each originating from a randomly selected individual (among all individuals). Each of these chains may be viewed as an approximation to the recruitment tree originating from one seed. Hence, the RWWT is able to account for the multiple seed structure of RDS under the assumption that seeds are selected uniformly; this assumption is discussed in Section 5. To describe the social structure of our target population, we use the so-called *configuration model* (Molloy and Reed, 1995, 1998), a random graph model for which the degree distribution of the resulting graph may be specified, to fit the network of interest.

The rest of the paper is organised as follows. In Subsections 2.1 and 2.2 we formally define the configuration model and the RWWT, respectively. In Section 3, we present our calculations for the stationary distribution of the process (Subsection 3.1) and how to estimate it (Subsection 3.2). We evaluate our estimator by simulations for varying proportion of seeds in the sample and for populations with disconnected networks in Section 4. Our findings are then discussed in Section 5.

2. Preliminaries

We first introduce some network terminology which is used in the following. Formally, a social network is composed of a set of *vertices* V that represents the actors (e.g., individuals) and a set of edges E which represents the relations that connect the actors together (see Figure 1). The network can be represented by its *adjacency matrix* $A = \{a_{uv}\}$, where u and v belong to the set of vertices. We consider *undirected* networks only, i.e., networks where all relations are mutual. Then, $a_{uv} = a_{vu} = 1$ if there is an edge between two vertices u and v and $a_{uv} = a_{vu} = 0$ otherwise. We say that two vertices u and v are *neighbours* if there is an edge between u and v. As previously mentioned, the degree d_u of a

vertex u is the number of contacts of u, where $d_u = \sum_v a_{uv} = \sum_v a_{vu}$.

2.1. Configuration model

Assume that we have a set of n vertices. For the results in the rest of the paper, we consider the infinite population limit $n \to \infty$. Let D be a random variable, defined on the non-negative integers, that represents the degree distribution, i.e., the distribution of vertex degrees. To construct the network, we assign to each vertex a number of stubs or half-edges, independently drawn from D. Then, we randomly form pairs of all the stubs. If the number of stubs is uneven, we discard one stub, which does not affect our results in the limit of infinite population size. This construction procedure may generate self-loops and multiedges, i.e., edges that connects a vertex to itself and several edges between the same two vertices; the proportion of these is however small when $\mathbb{E}(D^2) < \infty$. In particular, the probability that the generated graph is simple, i.e., that it contains no self-loops or multiedges, is bounded away from 0 if $\mathbb{E}(D^2) < \infty$ (e.g., Britton et al., 2007, Lemma 5.3). Hence, we condition on the generated graphs being simple under the assumption that the second moment of D is finite in the following. We denote networks generated from this model by $G(V, E)$, where V is the set of vertices and E is the set of edges.

2.2. Random walk with teleportation

A RWWT in discrete time $\{Z_t; t = 0, 1, 2, \ldots\}$, taking place on a network $G(V, E)$, is a Markov process with state space given by the vertex set of the network. In each step, the walker traverses to a randomly chosen neighbour of the last visited vertex with probability $c \in [0, 1]$ or moves to a uniformly chosen vertex $v \in V$ with probability $1 - c$. Let the transition probability between two vertices u and v be p_{uv}. The transition matrix $P = \{p_{uv}\}$, $u, v \in V$, of $\{Z_t\}$ is then given by

$$P = cAD^{-1} + (1 - c)\frac{1}{n}\bar{1}^T\bar{1},$$

where A is the adjacency matrix of G, D is a diagonal matrix with the degree sequence of vertices in G at its diagonal, and $\bar{1}$ is the column vector of ones.

3. Theory

3.1. Stationary distribution

Assume that we have a configuration model network $G(V, E)$ of size n, where n is assumed to be large. Let the degree distribution of G be given by the random variable D. Further assume that we have a RWWT $\{X_t; t = 0, 1, 2, \ldots\}$ taking place on this network. We assume that the structure of G is unknown in $\{X_t\}$ but that the degrees of visited vertices are known. Let $v \in V$ be an arbitrarily chosen fixed vertex with known degree d_v. We are interested in the limiting probability that X_t is at v.

Assume that the random walk visits vertex $u \neq v$ at time s. In what follows, we write $u \leftrightarrow v$ if u and v are neighbours and $u \nleftrightarrow v$ otherwise. The probability that v is visited at time $s + 1$ is then given by

$$\begin{aligned} p_{uv} &= \mathbb{P}(X_{s+1} = v | X_s = u) \\ &= \mathbb{P}(X_{s+1} = v | u \leftrightarrow v, X_s = u)\, \mathbb{P}(u \leftrightarrow v) \\ &+ \mathbb{P}(X_{s+1} = v | u \nleftrightarrow v, X_s = u)\, \mathbb{P}(u \nleftrightarrow v). \end{aligned}$$

First, we consider the case where u and v are neighbours. Let J denote the event that the random walk makes a random jump at s. We have

$$\begin{aligned} \mathbb{P}(X_{s+1} = v | u \leftrightarrow v, X_s = u) &= \mathbb{P}(X_{s+1} = v | u \leftrightarrow v, X_s = u, J)\, \mathbb{P}(J) \\ &+ \mathbb{P}(X_{s+1} = v | u \leftrightarrow v, X_s = u, J^{\complement})\, \mathbb{P}(J^{\complement}). \end{aligned}$$

By the definition of $\{X_t\}$, $\mathbb{P}(X_{s+1} = v | u \leftrightarrow v, X_s = u, J) = 1/n$ and $\mathbb{P}(J) = 1-c$. If the random walk does not jump at s, it will only visit v at $s+1$ if it traverses along the edge between u and v, which happens with probability $1/d_u$. Hence,

$$\mathbb{P}(X_{s+1} = v | u \leftrightarrow v, X_s = u) = \frac{1}{n}(1 - c) + \frac{1}{d_u} c.$$

If u and v are not neighbours, v may only be visited at $s+1$ through a random jump. Hence,

$$\begin{aligned}\mathbb{P}(X_{s+1} = v | u \nleftrightarrow v, X_s = u) &= \mathbb{P}(X_{s+1} = v | u \nleftrightarrow v, X_s = u, J)\,\mathbb{P}(J) \\ &= \frac{1}{n}(1-c).\end{aligned}$$

By construction, $\mathbb{P}(u \leftrightarrow v) = d_u d_v/(2|E| - 1)$, where we may approximate $2|E| - 1$ by $2|E| = n\,\mathbb{E}(D)$ for large n. From these results, we have for $u \neq v$ that

$$\begin{aligned}p_{uv} &\approx \left(\frac{1}{n}(1-c) + \frac{1}{d_u}c\right)\frac{d_u d_v}{n\,\mathbb{E}(D)} + \frac{1}{n}(1-c)\left(1 - \frac{d_u d_v}{n\,\mathbb{E}(D)}\right) \\ &= c\frac{d_v}{n\,\mathbb{E}(D)} + (1-c)\frac{1}{n} \\ &= \frac{1}{n}\left(c\frac{d_v}{\mathbb{E}(D)} + 1 - c\right).\end{aligned}$$

If we the random walk visits v at time s, then it may only visit v again at $s+1$ by a random jump; hence,

$$p_{vv} = \frac{1}{n}(1-c).$$

Define

$$\pi_u = \frac{1}{n}\left(c\frac{d_u}{\mathbb{E}(D)} + 1 - c\right)$$

for all $u \in V$. We have

$$\begin{aligned}\sum_{u\in V}\pi_u p_{uv} &= \sum_{u\in V; u\neq v}\pi_u p_{uv} + \pi_v p_{vv} \\ &= \frac{1}{n}\left(c\frac{d_v}{\mathbb{E}(D)} + 1 - c\right)(1-\pi_v) + \frac{1}{n}(1-c)\pi_v \\ &= \frac{1}{n}\left(c\frac{d_v}{\mathbb{E}(D)} + 1 - c\right) - \frac{cd_v}{n\,\mathbb{E}(D)}\pi_v \\ &\approx \pi_v,\end{aligned}$$

where the approximation comes from that $cd_v/(n\,\mathbb{E}(D))\pi_v$ is $O(1/n^2)$. Because v was arbitrarily chosen,

$$\pi_v = \frac{1}{n}\left(c\frac{d_v}{\mathbb{E}(D)} + 1 - c\right) \propto c\frac{d_v}{\mathbb{E}(D)} + 1 - c, \quad v \in V \tag{2}$$

gives the stationary distribution of the RWWT on the configuration model network.

Note that, because the transition probabilities and the stationary distribution are very similar, it is close to redundant to make the assumption that the process is in stationarity when we later consider sampling from this process; this was also noted in Gile (2011) for the simple random walk on the configuration model network. We will however do so for a stringent exposition. Also note that it has been shown that the RWWT has the same stationary distribution on Chung-Lu random graphs and Erdős-Rényi random graphs (Kadavankandy et al., 2015). In general however, there exists no closed expression for the stationary distribution of the RWWT on an undirected network (Grolmusz, 2015). A generalisation of the RWWT is to let the probability to be visited when a jump has occurred to be different between different vertices, which in general will introduce a dependence on n in Eq. (2). If however the probability that a vertex is visited when a jump has occurred is proportional to its degree, all individuals are sampled with probability proportional to degree and the V-H estimator is recovered.

3.2. Estimation of c and $\mathbb{E}(D)$

Under the assumption that we obtain our sample by sampling with replacement from the RWWT in stationarity, we can use the stationary distribution from Eq. (2) as the draw-wise selection probabilities in the estimator in Eq. (1). However, in order to do so, we need to estimate the unknown parameter c and $\mathbb{E}(D)$. From here on we assume that S is a sample of size n_S from an RDS study with m seeds in which, for each sampled individual u, the property of interest y_u and the degree d_u is recorded. Under the assumptions of Subsection 3.1, we may view this sample as the outcome of a RWWT on a configuration model network which has jumped at m occasions during the collection of the sample. The jump probability $1 - c$ can then be estimated by the proportion of seeds in the sample m/n_S and we get an estimator $\hat{c}$ of c as

$$\hat{c} = 1 - \frac{m}{n_S}. \tag{3}$$

In order to estimate $\mathbb{E}(D)$, we consider a partition of the sample S in two parts: S_J which consists of those individuals that were sampled as the result of a jump by the random walk and S_{RW} which consists of those individuals that were sampled as the result of an edge traversal. Because the inclusion of an individual in either partition of S are independent of the composition of the other partition under our assumptions, S_J and S_{RW} are independent. The sample partitions are easily identified from the RDS sample; S_J comprises the seeds and $S_{RW} = S \setminus S_j$. The sizes of S_J and S_{RW} are given by m and $n_S - m$, respectively. We will proceed by deriving two estimators $\widehat{\mathbb{E}(D)}_J$ and $\widehat{\mathbb{E}(D)}_{RW}$ of the expected degree from sampled individuals in S_J and S_{RW}, respectively. The individuals in S_J are sampled randomly with replacement and hence an estimator of $\mathbb{E}(D)$ is (Särndal et al., 1992, ch. 2.9)

$$\widehat{\mathbb{E}(D)}_J = \frac{\sum_{u \in S_J} d_u}{m}. \tag{4}$$

The variance of $\widehat{\mathbb{E}(D)}_J$ is estimated by

$$\widehat{Var}\left(\widehat{\mathbb{E}(D)}_J\right) = \frac{s_J^2}{m}, \tag{5}$$

where s_J^2 is the sample variance of the degrees of individuals in S_J. Because the individuals in S_{RW} are sampled by edge traversal in the random walk, their draw-wise selection probabilities are proportional to their degree. We have that an asymptotically unbiased estimator of the expected degree can be derived from the ratio of two Hansen-Hurwitz estimators (Salganik and Heckathorn, 2004) as

$$\widehat{\mathbb{E}(D)}_{RW} = \frac{n_S - m}{\sum_{u \in S_{RW}} 1/d_u}. \tag{6}$$

We obtain an approximative estimator of $Var\left(\widehat{\mathbb{E}(D)}_{RW}\right)$ by applying the Delta method and substituting population quantities with their sample counterparts:

$$\widehat{Var}\left(\widehat{\mathbb{E}(D)}_{RW}\right) \approx \left(\frac{1}{\left(\overline{d^{-1}}\right)^4}\right) \frac{s_{d^{-1}}^2}{n_S - m}, \tag{7}$$

where $\overline{d^{-1}}$ and $s_{d^{-1}}^2$ are the sample mean and variance of the inverse degrees of individuals in S_{RW}, respectively. Then, we combine these estimators in a

composite estimator $\widehat{\mathbb{E}(D)}$ (Schaible, 1978) of the expected degree:

$$\widehat{\mathbb{E}(D)} = w\widehat{\mathbb{E}(D)}_J + (1-w)\widehat{\mathbb{E}(D)}_{RW}, \tag{8}$$

where $0 \leq w \leq 1$. We want to choose w such that the variance of $\widehat{\mathbb{E}(D)}$ is minimized. Because S_J and S_{RW} are independent samples, the variance of $\widehat{\mathbb{E}(D)}$ is a weighted sum of the variances of $\widehat{\mathbb{E}(D)}_J$ and $\widehat{\mathbb{E}(D)}_{RW}$. Taking the variance and differentiating in Eq. (8) yields that the minimal variance is obtained when $w = w^*$, where

$$w^* = \frac{Var\left(\widehat{\mathbb{E}(D)}_{RW}\right)}{Var\left(\widehat{\mathbb{E}(D)}_J\right) + Var\left(\widehat{\mathbb{E}(D)}_{RW}\right)}. \tag{9}$$

We obtain an estimate $\hat{w}^*$ by substituting the estimates from Eqs. (5) and (7) into Eq. (9). To find estimates $\{\hat{\pi}_u; u \in V\}$ of the stationary distribution of the RWWT on the configuration model network we may then substitute the estimates given by Eqs. (3), (8), and (9) into Eq. (2); we have

$$\hat{\pi}_u \propto \hat{c}\frac{d_u}{\widehat{\mathbb{E}(D)}} + 1 - \hat{c}, u \in S. \tag{10}$$

The estimated stationary distribution can then be substituted into Eq. (1) to obtain an estimator $\hat{\mu}_{\mathrm{T}}$ of population properties as

$$\hat{\mu}_{\mathrm{T}} = \frac{\sum_{u\in S}\frac{y_u}{\hat{\pi}_u}}{\sum_{u\in S}\frac{1}{\hat{\pi}_u}}. \tag{11}$$

From Eqs. (2), (10), and (11), we recover, as limiting cases for $\hat{\mu}_{\mathrm{T}}$, the sample mean when $c \to 0$, i.e., when the draw-wise selection probabilities all are similar, and the V-H estimator when $c \to 1$.

4. Numerical simulations

Our estimator extends the V-H estimator in two respects: i) it accounts for the multiple seed structure of RDS and ii) it is valid for disconnected networks. We focus on these properties of the estimator in our evaluation and compare with the limiting cases given by the V-H estimator and the sample mean. We do not consider estimators that require population parameters that are traditionally

not collected or estimable within the RDS sample (Gile, 2011; Lu et al., 2013; Lu, 2013; Gile and Handcock, 2015).

To test the performance of our estimator, we simulate the RDS process in a population represented by a random network. We generate the network with $N = 10000$ individuals (i.e. the size of the target population) using the configuration model in which the degree distribution D is given by $\mathbb{P}(D = d) = p_d$. We consider two degree distributions: a power-law with exponential cutoff, for which

$$p_d = \frac{\lambda^{1-\alpha}}{\Gamma(1-\alpha, \lambda d_{\min})} d^{-\alpha} \exp(-\lambda d),$$

and a log-normal, for which

$$p_d = \frac{1}{d\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln\ d - \theta)^2}{2\sigma^2}\right).$$

These distributions are chosen because they reproduce the degree heterogeneity observed in social networks (e.g., Amaral et al., 2000). We choose the parameters $d_{\min} = 3$, $\alpha = 2.5$, and $\lambda = 0.00001$ for the power-law, and $\theta = 2.0$ and $\sigma = 0.5$ for the log-normal, such that the average degrees become 7.47 ± 0.30 and 7.87 ± 0.05 ($\pm$ represents the standard deviation over 100 samples of the network with a given degree distribution), respectively. Let y be a hypothetical trait taking values 0 or 1 (e.g. being healthy or infected with a disease). We select 15% of the population, starting with the individual with the largest degree and proceeding in decreasing order of degree, to assign the value $y = 1$. The remaining individuals in the population are assigned $y = 0$. To reduce degree-trait correlations, we go through all vertices and with probability 0.2, we uniformly select a second vertex and swap states (e.g. infected $\rightarrow$ non-infected). This procedure conserves the total number of infected individuals in the population.

We start the RDS process with m seeds uniformly chosen within the target population. All seeds start recruitment at the same time. At each time step, an individual invites 3 peers. We assume that all invited peers participate in the experiment. An individual that has already participated in the study may not be invited again at a later time. Recruitment thus stops if the desired sample

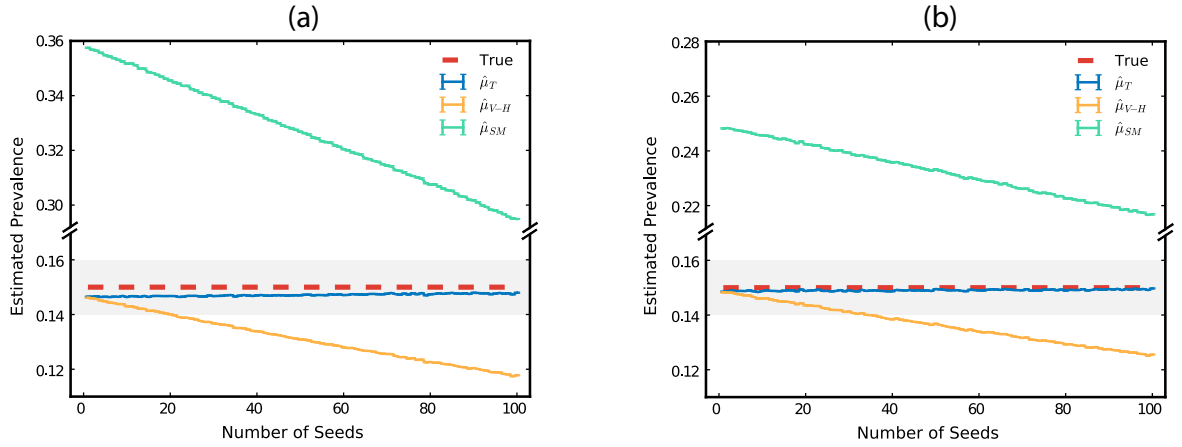

Figure 2. Comparative performance of the estimators in a network with a single connected component. Mean of the estimated prevalence and respective standard error (vertical bars, generally smaller than the width of the curves) for varying number of initial seeds (a) Power-law with exponential cutoff and (b) Log-normal degree distributions. We repeat the simulations 100 times for each of the 100 random network samples, therefore, the average and standard error are calculated over 10000 realizations for each number of seeds. Note that the vertical axis is broken in both (a) and (b).

size $n_S = 300$ is achieved or no more recruitments occur. Figure 2 shows the performance of our estimator $\hat{\mu}_{\mathrm{T}}$ in comparison to V-H ($\hat{\mu}_{\text{V-H}} = \frac{\sum_{u\in S} y_u d_u^{-1}}{\sum_{u\in S} d_u^{-1}}$) and the sample mean ($\hat{\mu}_{\mathrm{SM}} = \frac{1}{n_S}\sum_{u\in S} y_u$) for two configurations of networks with power-law with exponential cutoff (Fig. 2 a) and log-normal (Fig. 2 b) degree distributions. For all estimators, we include the seeds in the sample. Comparatively, our estimator has the best performance irrespective of the number of seeds or degree distribution, slightly underestimating the true prevalence. The V-H estimator increasingly underestimates the true prevalence for increasing number of seeds but performs similarly to our estimator for low number of seeds ($m \lesssim 5$). The sample mean, on the other hand, substantially over-estimates the prevalence as expected but improves performance for increasing number of seeds.

We now make an experiment on a social network with two connected components. We first divide the population into two groups of 5000 vertices each.

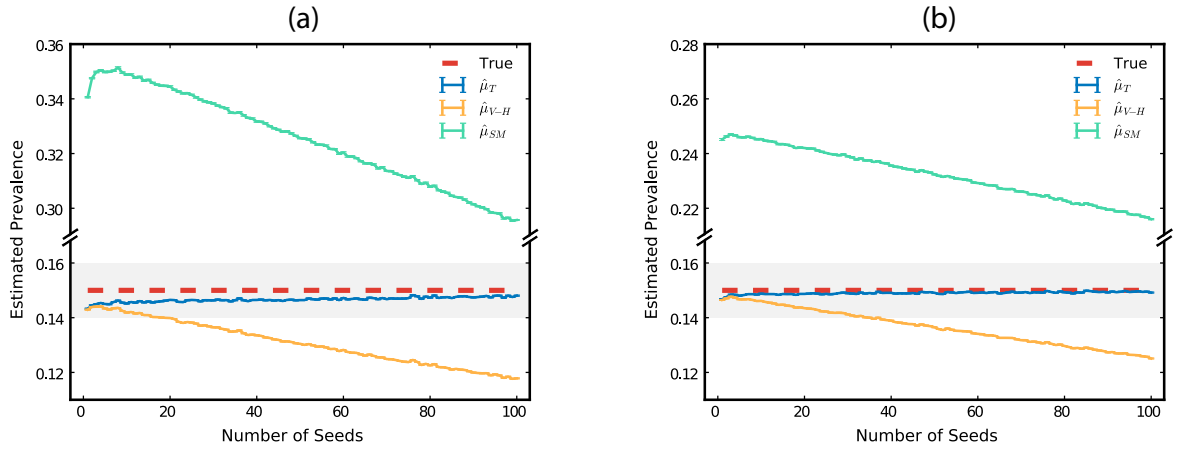

Figure 3. Comparative performance of the estimators in a network with two connected components. Mean of the estimated prevalence and respective standard error (vertical bars, generally smaller than the width of the curves) for varying number of initial seeds (a) Power-law with exponential cutoff and (b) Log-normal degree distributions. We repeat the simulations 100 times for each of the 100 random network samples, therefore, the average and standard error are calculated over 10000 realizations for each number of seeds. Note that the vertical axis is broken in both (a) and (b).

We then generate stubs for each vertex in the same way as before but only uniformly connect vertices belonging to the same group. The trait y is distributed according to the degree of the vertices, as done for the single component case. The power-law with exponential cutoff now has mean degree 7.43 ± 0.29 and the log-normal 7.87 ± 0.05. Although the V-H estimator are not designed for such disconnected networks, in practice one does not know if the social network is connected and thus simply apply the estimator on the collected data. Our estimator however can be safely applied in such settings without restrictions. We nevertheless compare the performance of all three estimators by repeating the previous experiment on this disconnected network. Figure 3 shows that our estimator generally outperforms V-H and the sample mean. Nevertheless, our estimator slightly underestimates the true prevalence if few seeds are used. The mismatch is larger for the power-law with exponential cutoff network (Fig. 3 a) in comparison to the log-normal case (Fig. 3 b).

We now compare the estimators in a realistic setting, in which the network

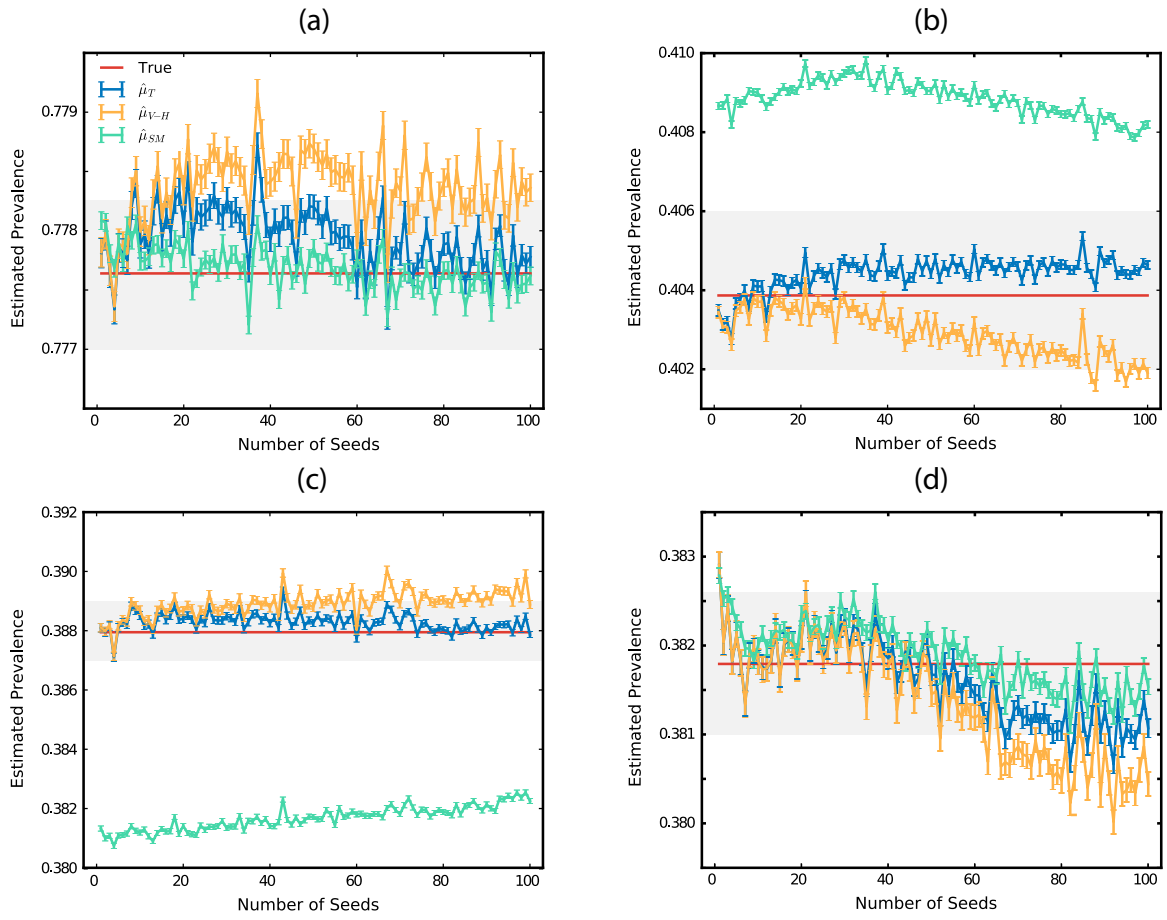

Figure 4. Comparative performance of the estimators in a social network with given prevalence of four different traits. Mean of the estimated prevalence and respective standard error (vertical bars) for varying number of initial seeds (a) Age, (b) Civil Status, (c) County, and (d) Profession. We repeat the simulations 10000 times for different starting conditions, therefore, the average and standard error are calculated over 100000 realizations for each number of seeds.

structure and the prevalence of various individual traits are known. We use an online social network targeting homosexual, bisexual, transgender, and queer persons called Qruiser (www.qx.se); this network was previously analysed in Rybski et al. (2009) and used to evaluate RDS in Lu et al. (2012). The network is connected and contains 16,082 individuals which identify themselves as homosexual males and has 108,334 social ties. The average degree is 13.47. For each individual, 4 dichotomous properties have been extracted from his user profile: age (born before 1980/others), civil status (single/others), county (live in Stockholm/others), and profession (employed/others). Here we also target a sample size of 300 and an individual may invite 3 peers. Figure 4 b,c show that our estimator outperforms the other two estimators for detecting the civil status and the county of living, respectively. In these two cases, the correction given by our estimator becomes visible as the number of seeds increases. For the age and profession (Figures 4 a,d), on the other hand, our estimator performs similarly to the sample mean but better than the V-H estimator. We see that even in those situations in which V-H performs well, some improvement is obtained by using our estimator.

5. Discussion

In this work, we present a novel RDS estimator that utilises a RWWT approximation of the RDS recruitment process. The new estimator is able to account for the multiple seed structure of RDS not considered by the usual simple random walk approximation of RDS. It is also valid for populations with disconnected social networks and does not require information that is traditionally not collected in an RDS sample. To test the performance of our estimator against the V-H estimator and the sample mean, we simulate RDS experiments on theoretical networks with a given prevalence of an hypothetical binary variable y. The results show that our estimator generally outperforms the V-H estimator and the sample mean irrespective of the number of seeds. We also performed simulations on a real online social network. In this more complex situation,

our estimator overall performs better than the V-H estimator and the sample mean, although the improvement with respect to the number of seeds is not as large as for the generated networks. In our experiments on configuration model networks, the variable y is preferentially distributed in high-degree individuals. In this scenario, both our and the V-H estimators underestimate whereas the sample mean substantially over-estimates the true prevalence. The difference between our estimator and the V-H estimator gets larger for increasing number of seeds, but our estimator performs substantially better. This is expected since the component of the estimator accounting for the assumed simple random sampling of the seeds gets more relevant and thus the performance of V-H decreases significantly with increasing number of seeds. Since it is not uncommon that the seeds correspond to $5-10\%$ of the sample size in empirical studies (Malekinejad et al., 2008), or even larger proportions in certain studies, our results show that one may expect substantial biases in the estimates given by the V-H estimator. This bias, generated by the seeds, becomes small or negligible when our estimator is used; additionally, we conclude that the situation with additional seeds is not a major problem for RDS if the corrected estimator is adopted.

In actual RDS practice, seeds are not likely to be selected randomly. Rather, because the seeds are typically chosen among population members known to researchers, the seeds will form a convenience sample, the dependence on which the usual RDS assumption of convergence to equilibrium is meant to handle. However, little or no information exists on the composition of the seeds with respect to sampled properties and network structure in most RDS studies. Nevertheless, uniform sampling is generally a reasonable first approximation. It is often recommended that the seeds are selected such as to reflect the composition of the population (WHO, 2013). The ambition to select a diversified seed sample may result in seeds being selected from the parts of the network that are separate from each other, or that have weak connections. Hence, this ambition may aid in the actual network of coupon distribution not being connected, in which case our estimator is to be preferred.

If the seeds are removed from the sample, the individuals that remain were sampled with probability proportional to degree. Hence, if we estimate population properties without the seeds, we recover the V-H estimator despite that the sample is assumed to come from a RWWT. This implies that the assumption of a connected network is superfluous for the V-H estimator if the seeds are not used for estimation purposes or if the seeds are assumed to be selected with probability proportional to degree. In theory, if the sample is assumed to come from a RWWT, we would need to assume that the social network of the population is a configuration model network in order to use the V-H estimator. However, in the practical estimation process, this becomes a technicality, and we argue that it should not be necessary to make this assumption. The results for other random graph models mentioned in Subsection 3.1 further supports this argument.

Following our results, we recommend the use of our estimator: i) if the proportion of seeds in the sample is more than 5%, either from the initial seeds or from additional seeds that joined along the experiment; ii) if the social network is expected to be disconnected or with weak ties between groups of individuals (e.g. segregated or highly clustered groups inside the target population). Finally, our estimator requires a few more steps for calculation than the well-known V-H estimator. We thus provide a step-by-step guide on how to implement the estimation procedure in the Appendix. Note that no new information is necessary to use our estimator but the number of seeds and degree of the sampled individuals as available in typical RDS studies.

Acknowledgements

Jens Malmros is supported by The Swedish Research Council, project no. 621-2012-3868. LECR is a Chargé de recherche of the Fonds de la Recherche Scientifique - FNRS. The authors would like to thank Fredrik Liljeros and Xin Lu for use of the Qruiser data and Tom Britton for helpful discussions.

Appendix: Estimation procedure implementation

Let S be a respondent-driven sampling (RDS) sample of size n_S from an RDS study with m seeds in which each sampled individual u in S is surveyed for a variable y_u and has degree d_u. We proceed as follows to obtain estimates $\hat{\mu}_{\mathrm{T}}$ of the mean of y.

(a) Calculate an estimate $\hat{c}$ from Eq. (3).

(b) Split S into two samples: S_J which consists of the m seeds and S_{RW} which consists of the rest of the sample.

(c) Calculate the following estimates:

  (i) $\widehat{\mathbb{E}(D)}_J$ as the sample mean of the degrees of individuals in S_J from Eq. (4).

  (ii) $\widehat{\mathbb{E}(D)}_{RW}$ as the harmonic mean of the degrees of individuals in S_{RW} from Eq. (6)

  (iii) $\widehat{Var}\left(\widehat{\mathbb{E}(D)}_J\right)$ from Eq. (5). $s_J^2 = (1/(m-1))\sum_{u\in S_J}(d_u - \bar{d}_J)^2$, where $\bar{d}$ is the mean degree of individuals in S_J.

  (iv) $\widehat{Var}\left(\widehat{\mathbb{E}(D)}_{RW}\right)$ from Eq. (7). $s_{d^{-1}}^2 = 1/(n_S - m - 1)\sum_{u\in S_{RW}}(1/d_u - \overline{d^{-1}})^2$, where $\overline{d^{-1}}$ is the mean of $1/d_u$, for sampled u in S_{RW}.

  (v) $\hat{w}^*$ from Eq. (9) with substituted estimates $\widehat{Var}\left(\widehat{\mathbb{E}(D)}_J\right)$ and $\widehat{Var}\left(\widehat{\mathbb{E}(D)}_{RW}\right)$.

  (vi) $\widehat{\mathbb{E}(D)}$ from Eq. (8) where we put $w = \hat{w}^*$.

(d) Estimate the draw-wise selection probability $\hat{\pi}_u$ for every sampled individual $u \in S$ from Eq. (10).

(e) Estimate the mean of y with the estimator $\hat{\mu}_{\mathrm{T}}$ from Eq. (11).

References

Amaral, L. A. N., Scala, A., Barthelemy, M. and Stanley, H. E. (2000) Classes of small-world networks. *Proc. Natl. Acad. Sci. USA*, **97**, 11149–11152.

Barros, A. B., Dias, S. F. and Martins, M. R. O. (2015) Hard-to-reach popula-

tions of men who have sex with men and sex workers: a systematic review on sampling methods. *Sys. Rev.*, **4**, 1.

Beyrer, C., Baral, S. D., van Griensven, F., Goodreau, S. M., Chariyalertsak, S., Wirtz, A. L. and Brookmeyer, R. (2012) Global epidemiology of {HIV} infection in men who have sex with men. *Lancet*, **380**, 367 – 377.

Biernacki, P. and Waldorf, D. (1981) Snowball sampling: Problems and techniques of chain referral sampling. *Sociol. Method Res.*, **10**, 141–163.

Brin, S. and Page, L. (1998) The anatomy of a large-scale hypertextual web search engine. In *Seventh International World-Wide Web Conference (WWW 1998)*, 107–117.

Britton, T., Janson, S. and Martin-Löf, A. (2007) Graphs with specified degree distributions, simple epidemics, and local vaccination strategies. *Adv. Appl. Prob.*, **39**, 922–948.

Burt, R. D., Hagan, H., Sabin, K. and Thiede, H. (2010) Evaluating respondent-driven sampling in a major metropolitan area: Comparing injection drug users in the 2005 seattle area national hiv behavioral surveillance system survey with participants in the raven and kiwi studies. *Ann. Epidemiol.*, **20**, 159–167.

Erickson, B. H. (1979) Some problems of inference from chain data. *Sociol. Methodol.*, **10**, 276–302.

Faugier, J. and Sargeant, M. (1997) Sampling hard to reach populations. *J. Adv. Nurs.*, **26**, 790 – 797.

Gile, K. J. (2011) Improved inference for respondent-driven sampling data with application to hiv prevalence estimation. *J. Am. Stat. Assoc.*, **106**, 135–146.

Gile, K. J. and Handcock, M. S. (2010) Respondent-driven sampling: An assessment of current methodology. *Sociol. Methodol.*, **40**, 285–327.

— (2015) Network model-assisted inference from respondent-driven sampling data. *J. R. Stat. Soc. A*, **178**, 619–639.

Goel, S. and Salganik, M. J. (2010) Assessing respondent-driven sampling. *Proc. Natl. Acad. Sci. USA*, **107**, 6743–6747.

Grolmusz, V. (2015) A note on the PageRank of undirected graphs. *Inform. Process. Lett.*, **115**, 633–634.

Hansen, M. H. and Hurwitz, W. N. (1943) On the theory of sampling from finite populations. *Ann Math Stat*, **14**, 333–362.

Heckathorn, D. D. (1997) Respondent-driven sampling: a new approach to the study of hidden populations. *Soc. Probl.*, **44**, 174–199.

Johnston, L., Chen, Y.-H., Silva-Santisteban, A. and Raymond, H. (2013) An empirical examination of respondent driven sampling design effects among HIV risk groups from studies conducted around the world. *AIDS Behav*, **17**, 2202 – 2210.

Kadavankandy, A., Avrachenkov, K., Prokhorenkova, L. O. and Raigorodskii, A. M. (2015) Pagerank in undirected random graphs. *CoRR*, **abs/1511.04925**.

Lu, X. (2013) Linked Ego Networks: Improving estimate reliability and validity with respondent-driven sampling. *Soc. Networks*, **35**, 669–685.

Lu, X., Bengtsson, L., Britton, T., Camitz, M., Kim, B. J., Thorson, A. and Liljeros, F. (2012) The sensitivity of respondent-driven sampling. *J. R. Stat. Soc. A*, **175**, 191–216.

Lu, X., Malmros, J., Liljeros, F. and Britton, T. (2013) Respondent-driven sampling on directed networks. *Electron. J. Stat.*, **7**, 292–322.

Magnani, R., Sabin, K., Saidel, T. and Heckathorn, D. (2005) Review of sampling hard-to-reach and hidden populations for HIV surveillance. *AIDS*, **19**, 67–72.

Malekinejad, M., Johnston, L. G., Kendall, C., Franco Sansigolo Kerr, L. R., Rifkin, M. R. and Rutherford, G. W. (2008) Using respondent-driven sampling

methodology for HIV biological and behavioral surveillance in international settings: A systematic review. *AIDS Behav*, **12**, 105–130.

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

McCreesh, N., Johnston, L. G., Copas, A., Sonnenberg, P., Seeley, J., Hayes, R. J., Frost, S. D. and White, R. G. (2011) Evaluation of the role of location and distance in recruitment in respondent-driven sampling. *Int. J. Health Geogr.*, **10**, 1.

Molloy, M. and Reed, B. (1995) A critical-point for random graphs with a given degree sequence. *Random Struct. Algor.*, **6**, 161–179.

— (1998) The size of the giant component of a random graph with a given degree sequence. *Comb. Probab. Comput.*, **7**, 295–305.

Rocha, L. E., Thorson, A. E., Lambiotte, R. and Liljeros, F. (2016) Respondent-driven sampling bias induced by community structure and response rates in social networks. *J. R. Stat. Soc. A*.

Salganik, M. J. (2006) Variance estimation, design effects, and sample size calculations for respondent-driven sampling. *J. Urban Health*, **83**, 98–112.

Salganik, M. J. and Heckathorn, D. D. (2004) Sampling and estimation in hidden populations using respondent-driven sampling. *Sociol. Methodol.*, **34**, 193–240.

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

Schaible, W. L. (1978) Choosing weights for composite estimators for small area statistics. In *Proceedings of the Section on Survery Research Methods*, 741–746. American Statistical Association.

Schwartländer, B., Ghys, P. D., Pisani, E., Kiessling, S., Lazzari, S., Caraël, M. and Kaldor, J. M. (2001) HIV surveillance in hard-to-reach populations. *AIDS*, **15**, S1–S3.

Stein, M. L., van Steenbergen, J. E., Chanyasanha, C., Tipayamongkholgul, M., Buskens, V., van der Heijden, P. G. M., Sabaiwan, W., Bengtsson, L., Lu, X., Thorson, A. E. and Kretzschmar, M. E. E. (2014) Online Respondent-Driven Sampling for Studying Contact Patterns Relevant for the Spread of Close-Contact Pathogens: A Pilot Study in Thailand. *PLoS ONE*, **9**, e85256.

Strömdahl, S., Lu, X., Bengtsson, L., Liljeros, F. and Thorson, A. (2015) Implementation of web-based respondent driven sampling among men who have sex with men in sweden. *PLoS ONE*, **10**, e0138599.

Sudman, S., Sirken, M. G. and Cowan, C. D. (1988) Sampling rare and elusive populations. *Science*, **240**, 991–996.

Thompson, S. K. (1990) Adaptive cluster sampling. *J. Am. Stat. Assoc.*, **85**, 1050–1059.

— (2012) *Sampling*. Hoboken: Wiley.

Thompson, S. K. and Frank, O. (2000) Model-based estimation with link-tracing sampling designs. *Surv. Methodol.*, 87–98.

Tomas, A. and Gile, K. J. (2011) The effect of differential recruitment, non-response and non-recruitment on estimators for respondent-driven sampling. *Electron. J. Stat.*, **5**, 899–934.

Tourangeau, R., Edwards, B., Johnson, T. P., Bates, N. and Wolter, K. M. (2014) *Hard-to-survey Populations*. Cambridge University Press.

Volz, E. and Heckathorn, D. D. (2008) Probability based estimation theory for respondent driven sampling. *J. Off. Stat.*, **24**, 79–97.

Wejnert, C. (2009) An empirical test of respondent-driven sampling: Point estimates, variance, degree measures, and out-of-equilibrium data. *Sociol. Methodol.*, **39**, 73–116.

WHO (2013) Introduction to HIV/AIDS and sexually transmitted infection surveillance: module 4: introduction to respondent-driven sampling. In *EMRO Meeting Reports*. World Health Organization.