Electronic Journal of Statistics
Vol. 14 (2020) 2678–2713
ISSN: 1935-7524
https://doi.org/10.1214/20-EJS1718

Correcting for differential recruitment in respondent-driven sampling data using ego-network information

Isabelle S. Beaudry[*,†,‡,§] **and Krista J. Gile**[¶]

Department of Statistics, Pontificia Universidad Católica de Chile, Santiago, Chile
Department of Mathematics and Statistics, University of Massachusetts Amherst, Amherst, Massachusetts, USA
e-mail: isabelle.beaudry@mat.uc.cl; gile@math.umass.edu

Abstract: Respondent-Driven sampling (RDS) is a sampling method devised to overcome challenges with sampling hard-to-reach human populations. The sampling starts with a limited number of individuals who are asked to recruit a small number of their contacts. Every surveyed individual is subsequently given the same opportunity to recruit additional members of the target population until a pre-established sample size is achieved. The recruitment process consequently implies that the survey respondents are responsible for deciding who enters the study. Most RDS prevalence estimators assume that participants select among their contacts completely at random. The main objective of this work is to correct the inference for departure from this assumption, such as systematic recruitment based on the characteristics of the individuals or based on the nature of relationships. To accomplish this, we introduce three forms of non-random recruitment, provide estimators for these recruitment behaviors and extend three estimators and their associated variance procedures. The proposed methodology is assessed through a simulation study capturing various sampling and network features. Finally, the proposed methods are applied to a public health setting.

Keywords and phrases: Hard-to-reach population sampling, non-sampling errors, network sampling, social networks.

Received February 2019.

Acknowledgements

The project described was supported by the Fondecyt grant 11171076, SES-1230081 from NSF, including support from the National Agricultural Statistics Service, by NSERC graduate scholarships and fellowships PGSD3, by grant number MH 89266 and DA 032059 from National Institutes of Health as well as by Millennium Science Initiative of the Ministry of Economy, Development, and Tourism, grant “Millenium Nucleus Center for the Discovery of Structures

[*] Pontificia Universidad Católica de Chile
[†] Millennium Nucleus Center for the Discovery of Structures in Complex Data
[‡] Centro de Riesgos y Seguros UC
[§] Instituto Milenio Fundamentos de los Datos, Chile
[¶] University of Massachusetts Amherst

in Complex Data" and "Instituto Milenio Fundamentos de los Datos, Chile". Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the Demographic and Behavioral Sciences (DBS) Branch or the National Science Foundation, NSERC nor NIH. The authors would also like to thank Corinne M. Mar for her contribution to the application and to the US Centers for Disease Control and Prevention for providing data.

1. Introduction

Respondent-driven sampling (RDS, Heckathorn (1997)) is a link-tracing network sampling method widely used to study hard-to-reach human populations connected by a network of social ties. Despite its wide adoption, statistical inference from RDS data continues to be subject to various sources of bias. Among them, is the bias induced by the individuals behaviors, such as participants preferentially recruiting subgroups of the population and/or individuals systematically electing not to participate in the study. Those behaviors are referred to as differential recruitment and they may be of particular interest due to the many documented instances in the context of RDS studies. The main contribution of this paper is to correct the inference for differential recruitment.

RDS draws its name from the recruiting method: previous respondents choose which of their network contacts will be recruited next by passing on uniquely-identified coupons. Researchers limit the branching of the sample, and therefore reduce the number of highly-dependent pairs in the sample, by limiting the number of coupons (Gile and Handcock, 2010; Rohe, 2015). This has proven a highly effective sampling method in many populations, and is widely used, especially in settings at high risk for HIV (Johnston et al., 2008; Malekinejad et al., 2008; Montealegre et al., 2013)

The primary object of inference from RDS data is population prevalence, possibly of a disease, a demographic characteristic, or a behavior. Such inference relies on many strong assumptions which are sometimes unrealistic in practice (Gile and Handcock, 2010). In particular, despite the growing empirical evidence (Frost et al., 2006; Iguchi et al., 2009; Liu et al., 2012; Mccreesh et al., 2012) that participants choose recruits in systematic ways, most inferential methods assume respondents make recruitments completely at random among their contacts in the target population. Sensitivity analysis performed with simulated and real data demonstrate that non-random recruitment potentially yields large biases in RDS estimators when the non-random recruitment patterns are associated with the object of inference (Frost et al., 2006; Gile and Handcock, 2010; Tomas and Gile, 2011; Lu et al., 2012; Verdery et al., 2015; Shi, Cameron and Heckathorn, 2019). There are also several available methods to diagnose non-random recruitment (Wejnert and Heckathorn, 2008; Liu et al., 2012; Yamanis et al., 2013; Gile, Johnston and Salganik, 2015; McLaughlin, 2016). The contribution of this work is to propose mathematical definitions for differential recruitment (DR), present a framework to estimate sampling probabilities in presence of DR and to correct prevalence estimators for the induced bias.

No estimator for RDS data dominates the others Tomas and Gile (2011); each has strengths and weaknesses. Therefore, rather than introducing a single proposed estimator, we propose extensions of three existing estimators: those introduced by Volz and Heckathorn (2008), Salganik and Heckathorn (2004), and Lu (2013). The estimator proposed by Lu (2013) is a modification of that of Salganik and Heckathorn (2004) which is partially robust to some forms of differential recruitment. Our work improves all three estimators in the case of three types of differential recruitment. Like Lu's estimator, our estimators require additional data beyond that typically collected in RDS: data on the composition of each respondent's local network contacts. It is this information that makes the recruitment biases identifiable in the context of social network homophily, or tendency for like to be connected to like (see Crawford et al. (2018) for a discussion of this non-identifiability). In addition, and unlike Lu's estimator, we consider differential recruitment based on both the outcome variable of interest, or based on another variable. This is especially important when the outcome of interest, such as HIV status, is not visible to a participant's contacts, making it impossible to collect local network composition from respondents (UNAIDS, 2014).

In the next Section, we introduce Respondent-Driven Sampling and the current estimators. The proposed prevalence estimators along with the parametrization of the three form of differential recruitment are discussed in Section 3. A comparison of their performance under various sampling conditions and network features is assessed in a simulation study presented in Section 4. Section 5 presents an application to a public health setting. Finally, we conclude with a discussion of the proposed methods in Section 6.

2. Respondent-driven sampling

An RDS sample begins with a set of *seeds* selected by the researchers, by some (strategic) convenience mechanism. Beginning with these seeds, each respondent is given a small number of uniquely-identified coupons to pass to contacts in the target population, making them eligible for enrollment. Respondents are compensated for their time completing the survey, and also given a small recruitment incentive for each successful recruitment. A key feature of data collection is that each respondent is asked to report her or his number of contacts in the target population. For the methods proposed in this paper, it is also necessary to collect data on the composition of the local contact network in terms of any variable which is expected to influence recruitment choices, as well as, for some estimators, on the outcome of interest.

2.1. Notation and sampling methodology

Consider a population of N individuals, also the *nodes* of the network, with labels $1, 2, ..., N$. These are connected by social ties represented by a sociomatrix

$Y \in \{0,1\}^{N\times N}$, such that $y_{ij} = 1$ if i connects to j, and $y_{ij} = 0$ otherwise. We assume an undirected network such that $y_{ij} = y_{ji} \,\forall\, i,j \in \{1,2,...,N\}$.

We use the vector $\mathbf{z} \in \{0,1\}^N$ to represent the outcome of interest. We refer to $\mathbf{z}$ as "infection status," $z_i = 1$ for infected persons in reference to the wide use of RDS in HIV applications, and for ease of discussion, although clearly it may represent any binary outcome of interest. For some of our proposed extensions, we also consider another binary nodal variable, denoted $\mathbf{x} \in \{0,1\}^N$, a non-outcome variable which may influence sampling decisions. Each node also has a *degree*, or number of contacts in the target population. We denote this $d_i = \sum_{j=1}^N y_{ij}$, and assume it is accurately reported by respondents.

Recall that our inferential goal is to estimate the prevalence of infection status. We denote the true value as $\mu = \frac{1}{N}\sum_{i=1}^N z_i$. We estimate μ using a sample of size n. The vector $\mathbf{S} \in \{0,1\}^N$ denotes sampling such that:

$$S_i = \begin{cases} 1 & \text{person } i \text{ has been sampled} \\ 0 & \text{otherwise} \end{cases} \qquad i \in \{1,2,...,N\}.$$

We sometimes refer to sets of nodes based on their variable values. We use calligraphic letters with superscripts to refer to sets of nodes with common values of a variable. For example $\mathcal{S}^1 = \{i : S_i = 1\}$. $\mathcal{S}^0$, $\mathcal{X}^1$, $\mathcal{X}^0$, $\mathcal{Z}^1$, and $\mathcal{Z}^0$ are defined similarly.

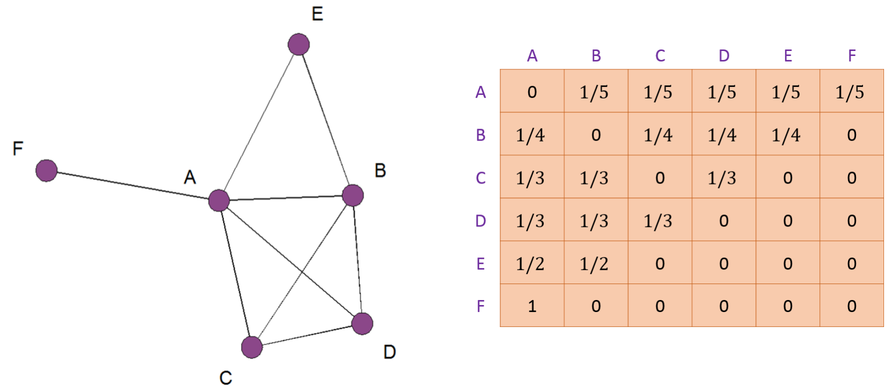
We also sometimes refer to sums. In particular, we sometimes sum degrees over categories of nodal covariates. These are denoted: $d_{i1}^z = \sum_{j=1}^N y_{ij}z_j$, $d_{i0}^z = \sum_{j=1}^N y_{ij}(1-z_j)$, $d_{i1}^x = \sum_{j=1}^N y_{ij}x_j$, $d_{i0}^x = \sum_{j=1}^N y_{ij}(1-x_j)$. We also refer to counts of sampled nodes in each category: $n_1^z = \sum_{i=1}^N S_i z_i$ and $n_0^z = \sum_{i=1}^N S_i(1-z_i)$.

2.2. Prevalence estimators

In this Section, we briefly describe three of the prevalence estimators we later extend in Section 3: the estimators developed by Volz and Heckathorn (2008), by Salganik and Heckathorn (2004) and by Lu (2013). They are respectively denoted $\hat{\mu}_{VH}$, $\hat{\mu}_{SH}$ and $\hat{\mu}_{Lu}$ and referred to as the VH, the SH and the Lu estimators.

All three estimators are based on an adapted version of the design-based estimator introduced by Hansen and Hurwitz (1943), the HH estimator. The HH estimator depends on p_i, that is, the probability that node i is selected at a given draw. The complexity of RDS and the lack of knowledge about the social network structure however prevent the exact determination of the sampling probabilities. For this reason, the described RDS estimators rely on a modified HH estimator, the HH-style estimator, where the sampling probabilities are instead estimated and denoted $\hat{p}_i$.

All RDS estimators discussed in this section assume that RDS may be well approximated by a discrete Markov chain (MC) on the state space of the network nodes to estimate the sampling probabilities $\hat{p}_i$. Conceptually, the transitions of



	A	B	C	D	E	F
A	0	1/5	1/5	1/5	1/5	1/5
B	1/4	0	1/4	1/4	1/4	0
C	1/3	1/3	0	1/3	0	0
D	1/3	1/3	1/3	0	0	0
E	1/2	1/2	0	0	0	0
F	1	0	0	0	0	0

(a) Small Network

(b) Transition matrix P

FIG 1. *Transition probability matrix (b) for a random walk on the nodes of the network depicted in (a) under random recruitment.*

the MC from one state (e.g. node i) to another (e.g. node j) represents the peer recruitment (e.g. j was recruited by i) as if nodes were constrained to recruit exactly one node by transition. Furthermore, even though individuals may only participate once in an RDS study, the MC representation allows sampling with replacement. The transitions of the MC are also assumed completely at random (Salganik and Heckathorn, 2004; Gile and Handcock, 2010). This is equivalent to presuming that participants recruit completely at random among all their contacts, or *alters*, in the target population. Finally, the recruitment process is assumed to occur on a single component network solely constituted of reciprocated ties. In summary, RDS is represented by a random walk (RW) on the nodes of a fully connected undirected network.

The described RW is mathematically fully specified by a transition probability matrix P, where p_{ij}, is the probability that node j is selected conditional on the process' current state i. This probability is equal to $p_{ij} = y_{ij}/d_i \ \forall \ i, j \in \{1, 2, ..., N\}$, since node i is constrained to recruit completely at random among its alters. Panel 1b of Figure 1 illustrates a simple case of a transition matrix P for the undirected network shown in panel 1a.

Under the presumed network structure, the with replacement MC is irreducible and all states are positive recurrent. The combination of these properties results in the existence of a unique stationary distribution denoted π (Ross, 2014). As demonstrated by Salganik and Heckathorn (2004), under random recruitment, the unique stationary distribution of this RW on the network nodes is given by:

$$\pi_i = \frac{d_i}{\sum_{i=1}^{N} d_i} \propto d_i \quad \text{for } \forall \ i \in \{1, 2, ..., N\}. \tag{2.1}$$

This stationary distribution may be interpreted as the proportion of time the process visits each state in the long run. The RDS estimators developed under this framework assume the sampling starts at stationarity. If this holds, participants' estimated sampling probabilities are proportional to their degree ($\hat{p}_i \propto d_i$).

The resulting estimated sampling probabilities are then used in the prevalence estimators proposed by Volz and Heckathorn (2008), Salganik and Heckathorn (2004) and Lu (2013). For instance, the VH estimator takes the form of a generalized HH-style estimator:

$$\hat{\mu}_{VH} = \frac{\frac{1}{n}\sum_{i=1}^{N} \dfrac{S_i z_i}{\hat{p}_i}}{\frac{1}{n}\sum_{i=1}^{N} \dfrac{S_i}{\hat{p}_i}} = \frac{\sum_{i=1}^{N} \dfrac{S_i z_i}{d_i}}{\sum_{i=1}^{N} \dfrac{S_i}{d_i}}. \tag{2.2}$$

As for the SH and Lu estimators, they have been derived based on the idea that, for an undirected network, the number of cross ties from infected to uninfected individuals should be equal to number of cross ties from uninfected to infected individuals. This observation leads to the following relation:

$$\mu = \frac{C_{01} D_0}{C_{01} D_0 + C_{10} D_1}, \tag{2.3}$$

where C_{10} (C_{01}) is the proportion of cross ties from infected (uninfected) to uninfected (infected) individuals and where D_1 (D_0) is the average degree of infected (uninfected) individuals. These quantities are not directly observed from the RDS sample. Therefore, the SH and Lu estimators replace these quantities with their estimated counterparts. The estimation of the average degree for each infection status for both the SH and Lu estimators is determined through a generalized HH-style estimator.

Although SH and Lu estimators rely on the same average degree estimators, they differ in the estimation of the proportion of cross ties. The SH methodology estimates the proportion of cross ties based on the observed unweighted recruitment patterns, whereas Lu's estimators for C_{10} and C_{01} are based on a generalized HH-style estimator. The latter rely on the degrees partitioned with respect to the outcome variable (d^z_{i0} and d^z_{i0}), which have not traditionally been collected in RDS surveys. However, incorporating this information has the potential of greatly reducing the DR bias (Lu, 2013; Verdery et al., 2015), provided this information is reported accurately. Expressions for the SH and Lu estimators may be found in equations (2.4) and (2.5). As pointed out by Beaudry, Gile and Mehta (2017), estimators of the SH form may be expressed as a function of the VH estimator:

$$\hat{\mu}_{SH} = \frac{\hat{\mu}_{VH}}{\hat{\mu}_{VH} + (1 - \hat{\mu}_{VH}) \left[\frac{n^z_1 r_{10} (r_{01} + r_{00})}{n^z_0 r_{01} (r_{10} + r_{11})} \right]} \tag{2.4}$$

$$\hat{\mu}_{Lu} = \frac{\hat{\mu}_{VH}}{\hat{\mu}_{VH} + (1 - \hat{\mu}_{VH}) \left[\frac{\sum_{i=1}^{N} S_i z_i d^z_{i0} / d_i}{\sum_{i=1}^{N} S_i (1 - z_i) d^z_{i1} / d_i} \right]}, \tag{2.5}$$

where $r_{uv} = \sum_{i,j} S_{ij} \mathbb{1}_{[z_i=u, z_j=v]}$, $u, v \in \{0, 1\}$ and $S_{ij} = 1$ if node i recruited node j and $S_{ij} = 0$ otherwise. In short, r_{uv} represents the observed number of recruitments from nodes with infection status "u" to nodes with status "v".

2.3. Variance estimators

In this section, we describe the bootstrap procedure introduced by Salganik (2006), referred to as the SH bootstrap, as well as the extension proposed by Lu (2013), to estimate the variance of RDS prevalence estimators.

In both cases, the bootstrap procedure consists of three steps. First, resamples of size n are drawn from the observed RDS samples. This step is repeated a large number of times, denoted B. Second, an RDS prevalence estimate is calculated for each of the B replicates. Third, a confidence interval is calculated based on the B prevalence estimates.

One of the main differences between the two bootstrap procedures lies in their resampling step, which is designed to preserve the transition probabilities of the estimated matrix P partitioned on the outcome variable. The bootstrap procedures are derived to be consistent with the respective estimated C_{01} and C_{10}. In the case of the SH bootstrap, the probability of sampling node j given that the RW process is at node i is equal to:

$$\Pr(R_t^{SH.boot} = j | R_{t-1}^{SH.boot} = i, z_k) = \begin{cases} 1/(r_{10} + r_{11}), & \text{if } z_i = z_k = 1 \\ 1/(r_{01} + r_{00}), & \text{if } z_i = z_k = 0 \\ 0, & \text{otherwise,} \end{cases} \tag{2.6}$$

where k refers to the individual who recruited node j in the actual RDS survey.

Lu (2013) proposed two bootstrap procedures. We consider the one which produces transition probabilities corresponding to the methodology employed in $\hat{\mu}_{Lu}$. The bootstrap procedure sequentially samples nodes in a RW fashion. The RW probabilities of transitions depend on the nodes membership to the sets $\mathcal{Z}^1$ and $\mathcal{Z}^0$. For instance, if the recruiting node is in $\mathcal{Z}^1$, then the probability of transitioning to a node in $\mathcal{Z}^0$ ($\mathcal{Z}^1$) is equal to $\hat{C}_{10}^{Lu}$ ($1-\hat{C}_{10}^{Lu}$). By construction, the bootstrap transition probabilities are consistent with the estimated proportion of ties between the infection groups in the network.

3. Differential recruitment

In this section, we extend the RDS estimators presented in Sections 2.2 and 2.3 to improve inference in the presence of differential recruitment (DR). We begin by operationalizing the concept of DR. Then, we specify new transition matrices reflecting three distinct forms of DR and derive the associated random walk stationary distributions. A maximum likelihood estimator is subsequently proposed to estimate the DR parameters required to estimate the sampling probabilities. Lastly, we present the extended versions of $\hat{\mu}_{VH}$, $\hat{\mu}_{SH}$ and $\hat{\mu}_{Lu}$, and corresponding measures of uncertainty.

3.1. Specification of differential recruitment

Under random recruitment, participants are assumed to recruit among their alters completely at random. Because recruitment is a social act, it is naive to assume this is always the case. Systematic violations of the random recruitment assumption are referred to as DR.

DR may arise in a variety of ways. Participants may favor the recruitment of individuals based on their characteristics (nodal attributes), or based on the nature of their relationship (tie attributes). The nodal characteristic inducing DR is represented by the indicator vector $\mathbf{x} \in \{0,1\}^N$ whereas the tie characteristic is represented by the symmetric indicator matrix $W \in \{0,1\}^{N \times N}$. Although we treat the binary case for the vector $\mathbf{x}$ and matrix W, these objects do not necessarily need to be binary. We provide some results for the general case where $\mathbf{x}$ is a categorical variable with $G \in \{2, 3, ...\}$ categories in the appendix.

Consistent with Tomas and Gile (2011), DR on the nodal attributes may be classified into two categories: within group and between group DR. Within group DR occurs when participants preferentially select alters similar to themselves, such as contacts of the same ethnic group, whereas between group DR results from all classes of respondents preferentially recruiting their contacts with a given characteristic. Gile, Johnston and Salganik (2015) find, for example, that respondents in four studies of injecting drug users in the Dominican Republic seem to systematically recruit their employed contacts more often than their unemployed contacts, perhaps due to the recruiters elevated confidence that these more reliable contacts would follow through in participating in the study. DR on the tie attribute is the result of participants preferably selecting individuals on the basis of their relationship with them. For example, Wang et al. (2005) found that 78.9% of respondents in an MDMA users study reported being recruited by a friend as opposed to 14.9% by an acquaintance and 3.4% by a relative. Participants' actual tie composition differing from those proportions would be evidence of recruitment based on tie characteristic. In this section, we address these three forms of DR.

The magnitude of differential recruitment behavior is quantified by parameter ϕ. In each case, this parameter represents the ratio of the probability of selecting a member of the target population with the nodal or tie preferred attribute to the probability of recruiting a member without it. For example, survey participants systematically recruiting males with a probability twice as high as other genders is denoted $\phi = 2$. Also, a recruitment regime completely at random implies that $\phi = 1$.

Definitions for the three parameters are presented in Table 1. The subscripts b, w, t indicate the form of DR, that is, between groups, within groups, and on tie attribute, respectively.

3.2. Sampling probabilities

To estimate the sampling probabilities, we derive the stationary distributions of the random walks with DR. We define in this Section the transition matrices

Table 1

Specification of three forms of DR under the RW scheme.
R_t indicates the state of the RW at step t.

DR Form	Parametrization
Between groups	$\phi_b = \dfrac{\Pr(R_t = j\vert\ R_{t-1} = i,\ y_{ij} = 1,\ x_j = 1)}{\Pr(R_t = j\vert\ R_{t-1} = i,\ y_{ij} = 1,\ x_j = 0)}$
Within groups	$\phi_w = \dfrac{\Pr(R_t = j\vert\ R_{t-1} = i,\ y_{ij} = 1,\ x_i = x_j)}{\Pr(R_t = j\vert\ R_{t-1} = i,\ y_{ij} = 1,\ x_i \neq x_j)}$
Tie	$\phi_t = \dfrac{\Pr(R_t = j\vert\ R_{t-1} = i,\ y_{ij} = 1,\ w_{ij} = 1)}{\Pr(R_t = j\vert\ R_{t-1} = i,\ y_{ij} = 1,\ w_{ij} = 0)}$

characterizing the three Markov chains and prove the existence and uniqueness of their stationary distributions contingent on some network features.

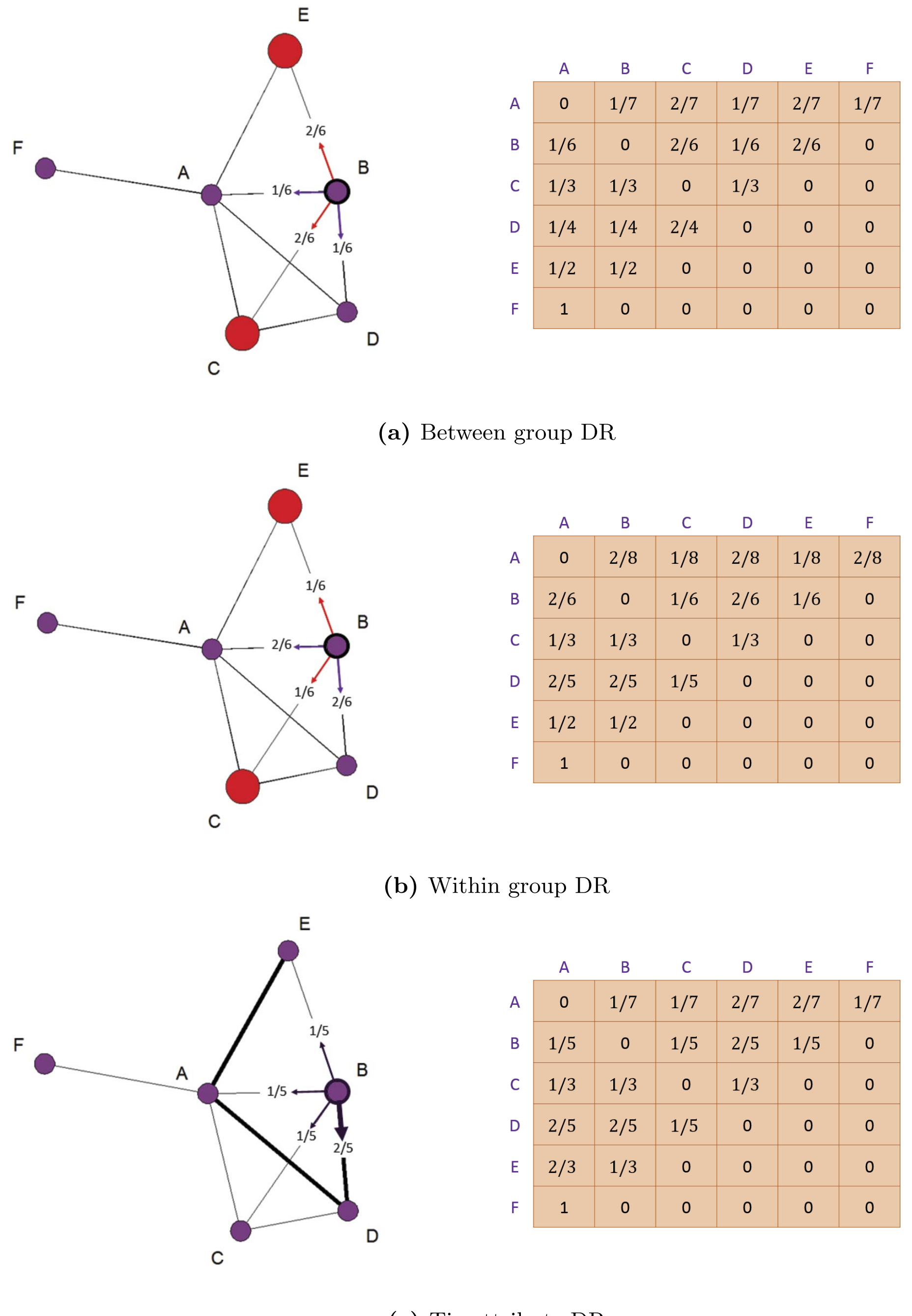
The transition matrices, P, specify the conditional probabilities of getting to any states given the previous state visited. For instance, p_{ij} is the probability of next selecting node j given that node i is the recruiting node.

Figure 2 shows simple examples of transition matrices for each of the three cases of DR of magnitude two ($\phi = 2$). Let us suppose that the size of the nodes in figure 2a and 2b represents a nodal attribute inducing DR. For instance, let the large nodes indicate that the individual resides in neighborhood N_1 as opposed to living in neighborhood N_2 depicted by the smaller size nodes. Under the previously introduced notation, $\mathbf{x} = \{0, 0, 1, 0, 1, 0\}$ where the nodes are arranged in alphabetical order. Figure 2a illustrates the case of between group DR so that all classes of participants favor the recruitment of nodes in N_1. This represents a hypothetical situation where every participant systematically favors the recruitment of their contacts living in the neighborhood where the study is conducted, for instance. As the left hand side of Figure 2a suggests, when the RW is in state B, the probability of selecting node C or E ($p_{BC} = p_{BE} = 2/6$) is twice as high as the probability of selecting node A or D ($p_{BA} = p_{BD} = 1/6$), that is, $\phi_b = 2$. The entire transition matrix P for this small network may be found in the right hand side of Figure 2a. More generally, P is given by:

$$p^b_{ij} = \frac{\phi_b^{x_j} y_{ij}}{\sum_{j'=1}^{N} \phi_b^{x_{j'}} y_{ij'}} \tag{3.1}$$

in the random walk (RW) with between group DR. For $\phi_b = 1$, that is, for a recruitment regime completely at random, $p^b_{ij} = \frac{y_{ij}}{\sum_{j'=1}^{N} y_{ij'}} = \frac{y_{ij}}{d_i}$ as expected. The transition probabilities for the general case where $\mathbf{x}$ has multiple categories is given in the appendix.

The derivation of within group DR transition probabilities is similar. However, instead of always favoring individuals residing in N_1, participants recruit more heavily alters living in the same neighborhood as themselves. Node B in Figure 2b for instance recruits A or D with a probability twice as large as the probability of selecting node C or E ($\phi_w = 2$). Consequently, we obtain the



	A	B	C	D	E	F
A	0	1/7	2/7	1/7	2/7	1/7
B	1/6	0	2/6	1/6	2/6	0
C	1/3	1/3	0	1/3	0	0
D	1/4	1/4	2/4	0	0	0
E	1/2	1/2	0	0	0	0
F	1	0	0	0	0	0

(a) Between group DR

	A	B	C	D	E	F
A	0	2/8	1/8	2/8	1/8	2/8
B	2/6	0	1/6	2/6	1/6	0
C	1/3	1/3	0	1/3	0	0
D	2/5	2/5	1/5	0	0	0
E	1/2	1/2	0	0	0	0
F	1	0	0	0	0	0

(b) Within group DR

	A	B	C	D	E	F
A	0	1/7	1/7	2/7	2/7	1/7
B	1/5	0	1/5	2/5	1/5	0
C	1/3	1/3	0	1/3	0	0
D	2/5	2/5	1/5	0	0	0
E	2/3	1/3	0	0	0	0
F	1	0	0	0	0	0

(c) Tie attribute DR

Fig 2. *Transition probability matrix (right) for a random walk on the nodes of the networks depicted on the left with three forms of DR of magnitude two ($\phi = 2$).*

following expression for the within group transition probability between node i and j:

$$p^w_{ij} = \frac{[\phi_w^{x_i} x_j + \phi_w^{1-x_i}(1-x_j)]y_{ij}}{\sum_{j'=1}^{N}[\phi_w^{x_i} x_{j'} + \phi_w^{1-x_i}(1-x_{j'})]y_{ij'}}. \tag{3.2}$$

An illustration for transition probabilities for tie attribute DR is provided in Figure 2c. Thicker ties in the plot on the left panel signify that the relationship type induces DR. Participants may exhibit the tendency to recruit more frequently close friends than acquaintances for example. According to this figure, only six entries in the underlying matrix of tie attributes W are equal to one, w_{AD}, w_{AE}, w_{BD}, and since W is assumed symmetric, the corresponding reciprocal relationships w_{DA}, w_{EA} and w_{DB}. Under this RW, B is twice as likely to select D over the other incident nodes. The entire matrix P for this example is provided in the right panel of Figure 2c, but the expression for any entry p^t_{ij} is given by:

$$p^t_{ij} = \frac{\phi_t^{w_{ij}} y_{ij}}{\sum_{j'=1}^{N} \phi_t^{w_{ij'}} y_{ij'}}. \tag{3.3}$$

We now consider the stationary distributions which are used as sampling weights in the extended versions of $\hat{\mu}_{VH}$, $\hat{\mu}_{SH}$ and $\hat{\mu}_{Lu}$. To ensure the Markov chains (MC) are irreducible, we strictly consider random walks on fully connected undirected networks where self ties are not permitted, a standard assumption for RDS, and assume $\phi > 0$. Finally, we assume a finite network to ensure the MC is positive recurrent. If those conditions are met, then there exists a unique stationary distribution for each of those stochastic processes (Ross, 2014).

Result 3.1. *Let R_t denote the state at step t of a MC on the nodes of a fully connected undirected network without self ties. Assume that this MC has the following transition probabilities:*

$$p^b_{ij} = \frac{\phi_b^{x_j} y_{ij}}{\sum_{j'=1}^{N} \phi_b^{x_{j'}} y_{ij'}},$$

where $\phi_b > 0$. Then the stationary distribution of this random walk is such that:

$$\pi^b_i \propto d^b_i = \phi_b^{x_i}(\phi_b d^x_{i1} + d^x_{i0}) \quad \textit{for } i \in \{1, 2, ..., N\}. \tag{3.4}$$

Result 3.2. *Let R_t denote the state at step t of a MC on the nodes of a fully connected undirected network without self ties. Assume that this MC has the following transition probabilities:*

$$p^w_{ij} = \frac{[\phi_w^{x_i} x_j + \phi_w^{1-x_i}(1-x_j)]y_{ij}}{\sum_{j'=1}^{N}[\phi_w^{x_i} x_{j'} + \phi_w^{1-x_i}(1-x_{j'})]y_{ij'}},$$

where $\phi_w > 0$. Then the stationary distribution of this random walk is:

$$\pi_i^w \propto d_i^w = \phi_w^{x_i} d_{i1}^x + \phi_w^{1-x_i} d_{i0}^x \tag{3.5}$$

for $i \in \{1, 2, ..., N\}$.

Result 3.3. *Let R_t denote the state at step t of a MC on the nodes of a fully connected undirected network without self ties. Assume that the MC has the following transition probabilities:*

$$p_{ij}^t = \frac{\phi_t^{w_{ij}} y_{ij}}{\sum_{j'=1}^N \phi_t^{w_{ij'}} y_{ij'}},$$

where $\phi_t > 0$. Then the stationary distribution of this random walk is:

$$\pi_i^t \propto d_i^t = \phi_t d_{i1}^w + d_{i0}^w \quad i \in \{1, 2, ..., N\}. \tag{3.6}$$

The proofs for Results 3.1, 3.2 and 3.3 as well as the stationary distribution for the general case of between-group DR may be found in the appendix.

The resulting stationary distributions all involve the ϕ parameters which are generally unknown since the sampling is driven by the respondents. However, these parameters may be estimated by maximizing the following likelihood functions:

$$L(\phi|R = r) \propto \prod_{i \in \mathcal{S}^1 \setminus S_0} \Pr(R_i = r_i | R_{i-1} = r_{i-1}, \phi_*) = \prod_{i \in \mathcal{S}^1 \setminus S_0} p^*_{r_{i-1} r_i}, \tag{3.7}$$

where R is an n-dimensional random vector specifying the state of the RW, S_0 is the set of seeds in the RDS study, and p^*_{ij} is the DR transition probabilities from node i to node j where $^* \in \{b, w, t\}$ identifies the form of DR.

The resulting estimate for ϕ's may be replaced in the stationary distributions so that the estimated stationary distributions for node $i \in \{1, 2, ..., N\}$ in equations (3.4), (3.5) and (3.6) respectively become proportional to:

$$\widehat{d_i^b} = \widehat{\phi}_b^{x_i} (\widehat{\phi}_b d_{i1}^x + d_{i0}^x), \tag{3.8}$$

$$\widehat{d_i^w} = \widehat{\phi}_w^{x_i} d_{i1}^x + \widehat{\phi}_w^{(1-x_i)} d_{i0}^x, \text{ and} \tag{3.9}$$

$$\widehat{d_i^t} = \widehat{\phi}_t d_{i1}^w + d_{i0}^w, \tag{3.10}$$

which are consistent estimators under the RW and network models assumed in this manuscript.

3.3. Extended prevalence estimators

In this section, we discuss the adjustments to the prevalence estimators to address DR. With the exception of the $\hat{\mu}_{SH}$, the extension consists in replacing the random recruitment sampling probabilities with the appropriate DR ones.

The extended prevalence estimator based on the VH estimator is:

$$\hat{\mu}^*_{VH.dr} = \frac{\sum_{i=1}^N S_i z_i/\widehat{d}_i^*}{\sum_{i=1}^N S_i/\widehat{d}_i^*}, \tag{3.11}$$

where $^* \in \{b, w, t\}$ identifies the form of DR. Correspondingly, every generalized HH-style estimator entering the $\hat{\mu}_{Lu}$ estimator, that is, $\hat{C}^{Lu}_{01}$, $\hat{C}^{Lu}_{10}$, $\hat{D}_0$ and $\hat{D}_1$, is modified in the same fashion, thus leading to:

$$\hat{\mu}^*_{Lu.dr} = \frac{\hat{\mu}^*_{VH.dr}}{\hat{\mu}^*_{VH.dr} + (1-\hat{\mu}^*_{VH.dr})\left[\frac{\sum_{i=1}^N S_i z_i d^z_{i0}/\widehat{d}_i^*}{\sum_{i=1}^N S_i(1-z_i) d^z_{i1}/\widehat{d}_i^*}\right]}. \tag{3.12}$$

Under non-random recruitment, the observed recruitment patterns fail to provide an unbiased estimate of the socio-matrix cross ties. The systematic over recruitment of infected individuals for instance, translates into larger r_{01} and r_{11} than expected at random and therefore, compromises the estimation of the cross tie proportions. Consequently, in addition to replacing the selection probabilities with the DR stationary distribution, modifications to the estimators $\hat{C}^{SH}_{01}$ and $\hat{C}^{SH}_{10}$ need also to be considered to adapt the SH methodology for DR. We propose the following estimators:

$$\hat{C}^{SH.b}_{01} = \frac{r_{01}}{r_{01} + \widehat{\phi}^z_b r_{00}}, \quad \hat{C}^{SH.b}_{10} = \frac{\widehat{\phi}^z_b r_{10}}{\widehat{\phi}^z_b r_{10} + r_{11}}; \tag{3.13}$$

$$\hat{C}^{SH.w}_{01} = \frac{\widehat{\phi}^z_w r_{01}}{\widehat{\phi}^z_w r_{01} + r_{00}}, \quad \hat{C}^{SH.w}_{10} = \frac{\widehat{\phi}^z_w r_{10}}{\widehat{\phi}^z_w r_{10} + r_{11}}; \tag{3.14}$$

$$\hat{C}^{SH.t}_{01} = \frac{r^w_{01} + r^{\bar{w}}_{01}\widehat{\phi}^z_t}{r^w_{01} + r^w_{00} + (r^{\bar{w}}_{01} + r^{\bar{w}}_{00})\widehat{\phi}^z_t}, \quad \hat{C}^{SH.t}_{10} = \frac{r^w_{10} + r^{\bar{w}}_{10}\widehat{\phi}^z_t}{r^w_{10} + r^w_{11} + (r^{\bar{w}}_{10} + r^{\bar{w}}_{11})\widehat{\phi}^z_t}; \tag{3.15}$$

where $\widehat{\phi}^z_*$ is an estimate of the relative probability of being recruited for infected individuals when compared to uninfected individuals and where:

$$r^w_{kl} = \sum_{i,j} S_{ij} \mathbb{1}_{[z_i=k, z_j=l]} w_{ij}, \quad r^{\bar{w}}_{kl} = \sum_{i,j} S_{ij} \mathbb{1}_{[z_i=k, z_j=l]} (1-w_{ij}) \tag{3.16}$$

for $k, l \in \{0, 1\}$ such that r^w_{kl} ($r^{\bar{w}}_{kl}$) represents the observed number of recruitments from nodes with infection status "k" to nodes with infection status "l" which are (not) preferentially recruited. The resulting extended SH estimator is:

$$\hat{\mu}^*_{SH.dr} = \frac{\hat{\mu}^*_{VH.dr}}{\hat{\mu}^*_{VH.dr} + (1-\hat{\mu}^*_{VH.dr})\left[\left(\frac{\hat{C}^{SH.*}_{10}}{\hat{C}^{SH.*}_{01}}\right)\frac{\sum_{i=1}^N S_i z_i d_i/\widehat{d}_i^*}{\sum_{i=1}^N S_i(1-z_i) d_i/\widehat{d}_i^*}\right]}. \tag{3.17}$$

The estimators shown in equations (3.11), (3.12), and (3.17) are consistent estimators for μ under the sampling and network assumptions discussed in this manuscript. A detailed discussion on the asymptotic properties of the estimators is provided in the appendix.

We note that, through the sampling probabilities, all extended prevalence estimators require ego-network composition on the DR characteristic, that is, d^x_{i1} and d^x_{i0}, $\forall\ i \in \mathcal{S}^1$. This information has not traditionally been collected in RDS surveys (Lu, 2013), however an increasing number of studies now include this information (Liu et al., 2009, 2012; Crawford et al., 2018). We also note that $\hat{\mu}^b_{VH.dr}$ is the only estimator not requiring the ego network composition on the outcome variable, that is, d^z_{i1} and d^z_{i0}, $\forall\ i \in \mathcal{S}^1$. This is especially relevant in the context of RDS as important outcome variables may not be visible to network contacts. Not relying on such information to estimate the prevalence while alleviating the effect of DR therefore constitutes a substantial contribution to the current RDS prevalence estimation methodology.

For simplicity purposes, sections 3.2 and 3.3 presented results for the specific case of binary DR variables. Estimators for the general case of between group DR along with a simulation study are discussed in the appendix. This illustrates that our methodology may be adapted to treat various specifications of DR by simply carefully defining the recruitment process and deriving the associated stationary distribution.

3.4. Uncertainty of the estimators

So far, we have discussed methodology to estimate the prevalence of an outcome variable $\mathbf{z}$ with RDS data when participants preferentially recruit individuals based on their characteristics or relationships. In this section we develop methodology to assess the uncertainty of the proposed estimators $\hat{\mu}^*_{VH.dr}$, $\hat{\mu}^*_{SH.dr}$ and of $\hat{\mu}^*_{Lu.dr}$. We focus on the between group DR. Specifications for the other forms of DR are included in the appendix. Our proposed variance estimators extend the bootstrap procedure proposed by Lu (2013) summarized in Section 2.3.

Under the random recruitment assumption, uniform edge sampling assures the proportion of cross recruitment from one infection group to the other is an unbiased estimator for the proportion of cross ties in Y, the network adjacency matrix. However, in the presence of DR, the observed proportion may be severely biased. Since the objective of the bootstrap procedure is to reflect the recruitment process of RDS as closely as possible through the RW representation, we emphasize that the transition probabilities in the variance estimator are designed to model the recruitment process and not necessarily the proportion of cross ties. Therefore, one of the fundamental differences between our proposed methodology and the SH and Lu's procedures is that the transition probabilities are driven by the DR characteristic and its associated parameter ϕ rather than by the outcome variable. More specifically, we propose the following uncertainty estimation algorithm:

1. **Step 1 – Re-sampling**: After selecting the first individual completely at random, the next individuals are selected sequentially with probability varying based on their membership in the sets $\mathcal{X}^1$ or $\mathcal{X}^0$ and also, based on the membership of the recruiting individual. Table 2 summarizes the transition probabilities for each of the four possible cases.

Table 2
Transition probabilities used in the bootstrap procedure under between group DR, where i is index of the recruiting node and $j \neq i$ is the index of any available node for recruitment.

	$j \in \mathcal{X}^0$	$j \in \mathcal{X}^1$
$i \in \mathcal{X}^0$	$\frac{1}{n_0^x}\left[1 - \frac{\hat{\phi}^b \sum_{i=1}^N S_i(1-x_i)d_{i1}^x/d_i^b}{n_0^x}\right]$	$\frac{1}{n_1^x}\left[\frac{\hat{\phi}^b \sum_{i=1}^N S_i(1-x_i)d_{i1}^x/d_i^b}{n_0^x}\right]$
$i \in \mathcal{X}^1$	$\frac{1}{n_0^x}\left[\frac{\hat{\phi}^b \sum_{i=1}^N S_i x_i d_{i0}^x/d_i^b}{n_1^x}\right]$	$\frac{1}{n_1^x}\left[1 - \frac{\hat{\phi}^b \sum_{i=1}^N S_i x_i d_{i0}^x/d_i^b}{n_1^x}\right]$

2. **Step 2 – Estimation**: Prevalence estimates are computed for all replicates. In the present case, the prevalence is determined based on the extended prevalence estimator for which the variability is estimated. For example, if the objective is to estimate the variability of $\hat{\mu}^b_{VH.dr}$, then the prevalence estimates are calculated with equation (3.11).
3. **Step 3 – Confidence Interval**: The first two steps are repeated a large number of times B. The standard deviation calculated from the obtained prevalence estimates is used to construct a studentized confidence interval.

We propose two variations of the bootstrap procedure. Both approaches begin by generating replicates (step 1) based on the transition probabilities calculated with the original $\hat{\phi}^b$, that is, $\hat{\phi}^b$ used in the DR prevalence estimate. For each of these replicates, a new $\hat{\phi}^b$ is calculated from the re-sampled data. The resulting set of $\hat{\phi}^b$'s is referred to as the bootstrapped $\hat{\phi}^b$'s and are used to determined the prevalence estimates in step 2. The only difference between the two variations of the procedures is that under the second version, step 1 is repeated twice. Once with the transition probabilities determined based on the original $\hat{\phi}^b$ and again, based on the bootstrapped $\hat{\phi}^b$'s.

The resulting bootstrap procedures are intended to capture the uncertainty pertaining to the sampling process assuming a random walk approximation to RDS. Also, by recalculating $\hat{\phi}^b$ for each replicate, we adjust for the variability of this parameter. However, neither variability due to a super-population model nor variability induced by other RDS-specific characteristics is reflected in these bootstrap estimators.

To estimate the variance of $\hat{\mu}^*_{SH.dr}$, we have also extended the original SH bootstrap discussed in Section 2.3. New transition probabilities are derived based on equations (3.13), (3.14) and (3.15) to capture the various forms of DR.

4. Simulation study

4.1. Simulation study design

The complexity of the RDS sampling method prevents an analytical assessment of the performance of the proposed prevalence estimators. Therefore, we present a simulation study to compare their performance under a variety of sampling conditions and network features. In this section, we present the design and results from the simulation study. The simulation study was performed with the statistical software `R` (R Core Team, 2018) and the packages `statnet` (Handcock et al., 2015) and `RDS` (Handcock, Fellows and Gile, 2015).

4.1.1. Network features

There is a vast body of literature studying the tendency of people to form ties with individuals with whom they share common attributes (Kandel, 1978; McPherson, Smith-Lovin and Cook, 2001; Currarini, Jackson and Pin, 2009). Therefore, it is critical for this simulation study to evaluate the sensitivity of the proposed methodology to this social behavior, known as homophily.

Exponential-family random graph models (ERGMs) (Frank and Strauss, 1986; Hunter, Goodreau and Handcock, 2008; Hunter and Handcock, 2006) provides the flexibility to incorporate this feature. This may be done by parametrizing the model so that the rate of ties among a certain group of individuals, η_{11}, differs from the rate of ties among members belonging to different groups, η_{10}. We use the following ERGM parametrization for undirected networks:

$$\Pr(Y = y|X = \mathbf{x}, H) = \frac{\exp\left\{H^T g(y, \mathbf{x})\right\}}{c(H|\mathbf{x})}, \tag{4.1}$$

where $H^T g(y, \mathbf{x})$ is equal to:

$$\eta \sum_{i,j} y_{ij} + \eta_{01} \sum_{i,j} y_{ij} x_i (1 - x_j) + \eta_{11} \sum_{i,j} y_{ij} x_i x_j, \tag{4.2}$$

and where

$$c(H|\mathbf{x}) = \sum_{s \in \mathcal{Y}(\mathbf{x})} \exp\left\{H^T g(s, \mathbf{x})\right\}, \tag{4.3}$$

where $\mathcal{Y}(\mathbf{x})$ is the space of all binary undirected networks of N nodes consistent with $\mathbf{x}$. This allows to control the overall propensity of forming a tie as well as the density of ties among alike members and across groups. In summary, this model formulation allows for homophily, which we define as:

$$\tau = \frac{\Pr(Y_{ij} = 1|X_i = X_j = 1)}{\Pr(Y_{ij} = 1|X_i \neq X_j)}. \tag{4.4}$$

The simulated networks were generated using $\tau = 1$ (no homophily) and $\tau = 5$ (elevated homophily) with respect to the DR variable $\mathbf{x}$. The rate of ties was also chosen so to produce an average degree of ten. The vector $\mathbf{x}$ contained 35% of individuals with the DR characteristic, $\mu_x = 0.35$, and the outcome variable was positive 20% of the time, $\mu_z = 0.20$. Infected individuals were selected among the individuals with the DR characteristics. A total of one thousand networks of size $N = 1000$ nodes were generated, for each of the values for τ using the R packages `statnet` (Handcock et al., 2015).

4.1.2. Sampling

The simulated RDS process in this study is intended to exhibit features approaching those of actual RDS studies. For instance, the nodes are sampled without replacement. Also, ten seeds initiate the sample instead of one as assumed by the estimators. Such seeds are selected according to the stationary distribution. Each sampled node subsequently recruits a maximum of two participants. A smaller number of recruits is allowed when there are less than two unsampled alters connected to the recruiting node. Nodes are presumed to recruit under one of the three recruitment regimes:

- recruitment completely at random ($\phi = 1$),
- moderate differential recruitment ($\phi = 2$), or
- elevated differential recruitment ($\phi = 4$)

with respect to the DR variable $\mathbf{x}$ or to the tie attribute matrix W. Nodes receiving an invitation to participate into the survey are presumed to systematically accept the invitation. The sampling process stops when the target sample size of two hundred is attained. One RDS sample is drawn from each network.

In summary, the six basic scenarios correspond to all possible permutations of the three levels of DR and the two levels of network homophily.

4.2. Results: Point estimates

Results from the simulation study for between group DR are presented in Figure 3 (and in Table 6 of the appendix). Figure 3 displays results for the six scenarios described in the previous section. The two levels of network homophily are shown on the horizontal panels, $\tau \in \{1, 5\}$ and the three levels of DR on $\mathbf{x}$ are shown on the vertical panels, $\phi \in \{1, 2, 4\}$. Estimates for the prevalence of $\mathbf{z}$ obtained from the six estimators are summarized by box plots which appear in the following order within each scenario: $\hat{\mu}_{VH}$, $\hat{\mu}^b_{VH.dr}$, $\hat{\mu}_{SH}$, $\hat{\mu}^b_{SH.dr}$ $\hat{\mu}_{Lu}$ and $\hat{\mu}^b_{Lu.dr}$. Estimators are grouped into three categories $\hat{\mu}_{VH}$, $\hat{\mu}_{SH}$, and $\hat{\mu}_{Lu}$ on the x-axis of each scenario and the box plot color within each category indicates the specific version of the estimator: original estimators (purple), and the extended estimators for DR proposed in this paper (yellow). The true population parameter μ is represented by the horizontal blue line.

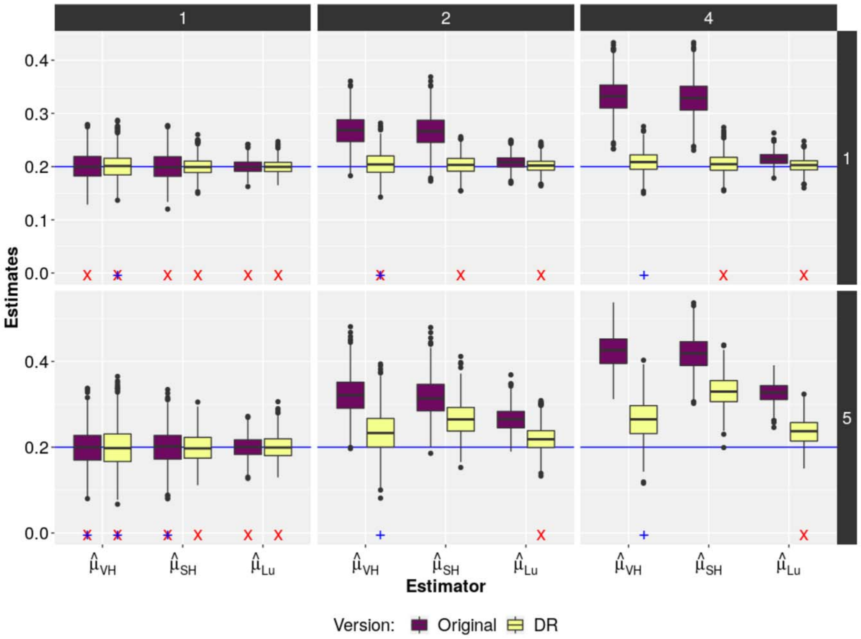


FIG 3. *Estimates produced with varying levels of network homophily, that is, $\tau \in \{1,5\}$, (horizontal panels) and between group DR on* $\mathbf{x}$*, that is, $\phi \in \{1,2,4\}$ (vertical panels). Estimators are presented in the following order: $\hat{\mu}_{VH}$, $\hat{\mu}^b_{VH.dr}$, $\hat{\mu}_{SH}$, $\hat{\mu}^b_{SH.dr}$ $\hat{\mu}_{Lu}$ and $\hat{\mu}^b_{Lu.dr}$. The blue horizontal line represents the true population prevalence for the variable* $\mathbf{z}$.

We first observe that all estimators have virtually no bias in the scenarios where no DR is simulated. In addition, in comparison with $\hat{\mu}_{VH}$ and $\hat{\mu}_{SH}$, Lu and the extended DR estimators display a reduced and similar variability. The reduction in the uncertainty is partly attributable to the fact that although ϕ is approximately equal to one on average, it slightly varies from this value in any particular simulated sample. These small departures from recruitment completely at random are corrected for in the extended estimators and therefore, produce estimates with smaller errors. For $\hat{\mu}_{VH.dr}$, the variability introduced by the network homophily offsets this effect. For Lu and the extended SH and Lu estimators, the decrease in variability is also explained by the improved estimation of the cross-recruitments.

Our simulation also corroborates the findings discussed in various studies that DR induces strong biases in $\hat{\mu}_{VH}$ and $\hat{\mu}_{SH}$ (Frost et al., 2006; Gile and Handcock, 2010; Tomas and Gile, 2011; Lu et al., 2012; Verdery et al., 2015; Shi, Cameron and Heckathorn, 2019). This holds even for a moderate value for ϕ. A between group DR of magnitude two, for instance, yields a relative bias of approximately 34% in the original VH and SH estimators in scenarios without homophily and a relative bias of 61% with homophily. As pointed out by Lu (2013) and Verdery et al. (2015), the estimator proposed by Lu, which incorporates information about ego-network composition, is far more robust to

DR than the original estimators under all assessed scenarios. Without network homophily, a relatively small bias remains when $\phi \neq 1$. The remaining bias is explained by the fact that the estimator relies on sampling probabilities which do not account for DR. Conversely, our simulation study suggests that, when the outcome variable differs from the DR characteristic, a significant relative bias (33% for $\phi = 2$) remains in presence of network homophily.

As observed in Figure 3 all discussed extended estimators reduce, in some cases substantially, the DR bias when compared to their original counterpart. To assess the performance of the extended estimators more formally, we tested how significantly different the mean square error of the estimators were using a pairwise Bonferroni procedure at a family-wise error rate of 5% for each of the six scenarios. Firstly, we assumes that only ego-network composition on the DR characteristic could reasonably be trusted (d_{i0}^x and d_{i1}^x). Under this assumption, only $\hat{\mu}_{VH}$, $\hat{\mu}_{VH.dr}^b$ and $\hat{\mu}_{SH}$ may be computed as $\hat{\mu}_{SH.dr}^b$ $\hat{\mu}_{Lu}$ and $\hat{\mu}_{Lu.dr}^b$ also require ego network composition on the outcome variable (d_{i0}^z and d_{i1}^z). The estimators included in the best set of estimators identified by this procedure are indicated by blue crosses in Figure 3. We observe that only $\hat{\mu}_{VH.dr}^b$ systematically appears in the best set of estimators. Secondly, we presumed that participants may also accurately report on the ego-network composition with respect to the outcome variable. The estimators included in the best set of estimators are labelled with the red X's. Those results suggest that only $\hat{\mu}_{Lu.dr}^b$ consistently appears in the best set of estimators.

These conclusions also hold when the outcome variable coincides with the DR characteristic ($\mathbf{z} = \mathbf{x}$), with the exception that Lu estimator is less sensitive to this class of DR. Results are presented in the appendix as well as results for the other two forms of DR.

The analysis presented above supposes that the outcome variable $\mathbf{z}$ is closely related to the DR characteristic. A simulation study has also been performed to assess the performance of the estimators when the variable inducing DR is unrelated to $\mathbf{z}$. The results indicated that the DR bias is smaller in instances where the variable $\mathbf{x}$ is not closely related to the outcome variable $\mathbf{z}$.

Additional scenarios under between group DR have also been simulated to heuristically evaluate the magnitude of seed selection bias, such as when seeds are selected strictly among infected individuals. We found that, except when there is no DR and the network is homophilous, the DR methodology still provides substantial bias reduction compared to the alternative estimators.

Finally, a simulation study where DR takes place on a variable with multiple groups is also presented in the appendix. We discuss results under scenarios under which the DR model is correctly specified as well as misspecified.

4.3. Results: Variance estimates

In this section, we assess the performance of the proposed bootstrap variance estimators described in Section 2.3 and Section 3.4 at various levels of between group DR and network homophily. We also evaluate the impact of

DR on the overall inference by comparing coverage rates of the 95% confidence intervals for the traditional RDS estimators and their extended versions.

The performance of the uncertainty estimators is evaluated by comparing the estimated standard deviation ($\hat{\sigma}$) to our best estimates of the true variability which consists of the standard deviation of the simulated prevalence estimates under each scenario (s's). Figure 4 presents the coverage rate of the 95% confidence intervals (upper plot) along with the difference between $\hat{\sigma}$ and s (lower plot). The results are shown for the nine bootstrap procedures under between group DR on $\mathbf{x}$ while making inference about the $\mathbf{z}$. These figures are organized in the same way as Figure 3, that is, the two horizontal panels display the results for the two levels of network homophily ($\tau \in \{1, 5\}$) and the vertical panels are divided according to the DR parameter $\phi \in \{1, 2, 4\}$. The estimators are presented in the following order within each scenario: $\hat{\sigma}(\hat{\mu}_{VH})$, $\hat{\sigma}(\hat{\mu}_{VH.drv1})$, $\hat{\sigma}(\hat{\mu}_{VH.drv2})$, $\hat{\sigma}(\hat{\mu}_{SH})$, $\hat{\sigma}(\hat{\mu}_{SH.drv1})$, $\hat{\sigma}(\hat{\mu}_{SH.drv2})$, $\hat{\sigma}(\hat{\mu}_{Lu})$, $\hat{\sigma}(\hat{\mu}_{Lu.drv1})$ and $\hat{\sigma}(\hat{\mu}_{Lu.drv2})$.

The first row of the upper plot shows that, without homophily, the 95% confidence intervals' coverage rates for the extended bootstrap procedures are either similar (no DR) or significantly better (DR > 1) than the original VH and SH or slightly better than Lu's variance estimators. This improvement is mostly attributable to the reduction in bias in the point estimation. However, in the case of the VH estimators, the higher coverage is also explained by the improved estimation of the variance as it may be seen in the bottom plot of the same figure. The results displayed in the second row of the two plots were produced with homophilous networks. Although the coverage rates are significantly lower than for non homophilous networks, we may conclude that, in the presence DR, the proposed inferential procedures provide higher coverage rates than the presented alternative methods. However, the poor coverage of the confidence intervals produced with $\hat{\sigma}(\hat{\mu}_{SH.drv1})$, and $\hat{\sigma}(\hat{\mu}_{SH.drv2})$, is mostly due to the remaining bias in the point estimates.

The comparison of the coverage rates further demonstrates that when the ego network composition on the DR characteristic is available, then inference may be improved by using the extended VH methodology. In the presence of strong network homophily however, it is preferable to also incorporate ego network data on the outcome variable through the Lu extended methodology.

The results presented in Figure 4 only suggest a slight outperformance of the bootstrap procedures where the estimated ϕ are resampled ($v2$) over the versions without this resampling step ($v1$).

The analysis presented in this section corresponds to the case in which the DR is related to the variable $\mathbf{x}$, but the inference is performed about the outcome variable $\mathbf{z}$. The appendix includes results for the case where the outcome variable and the DR characteristic coincide. The conclusions are similar, with the exceptions that the extended SH and Lu's methodologies perform better than the cases presented here whereas the results for the extended VH procedures deteriorate. Overall though, the proposed methods still outperform the studied alternatives in presence of DR.

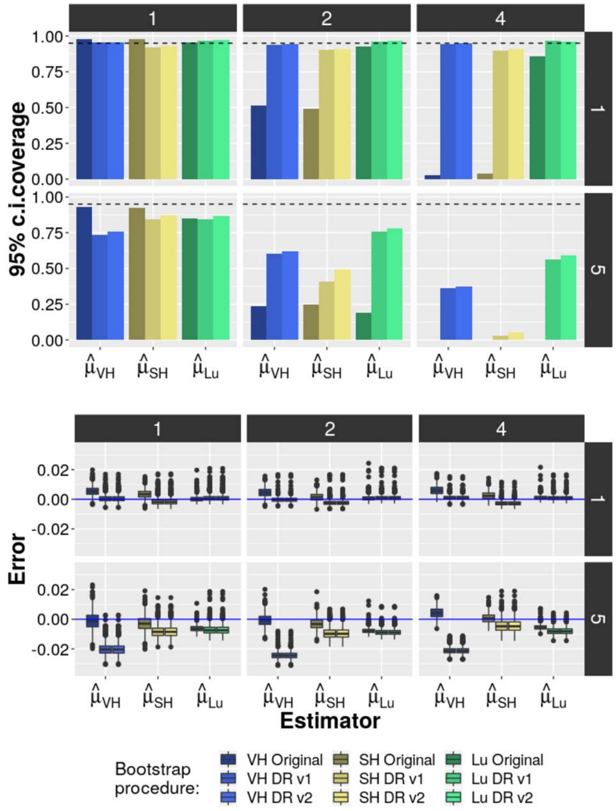


Fig 4. *95% confidence interval coverage rates, where the coverage rates are the percentage of the intervals including the true population proportion μ of 20%. The dashed line is set at 95%. Bias of the standard deviation estimates calculated as $\bar{\hat{\sigma}} - s$, where $\bar{\hat{\sigma}}$ is the average estimated standard deviation under a bootstrap methodology and s is the sample standard deviation.*

5. Application

The US Centers for Disease Control and Prevention (CDC) use RDS for behavioral surveillance of people who inject drugs (PWID) and high risk heterosexuals (HRH) in 25 US cities every 3 years (Gallagher et al., 2007). While the resulting data are highly sensitive human subjects data, and also do not include known true values to which to compare our results, we conduct our analyses on simulated datasets created to match the characteristics of fourteen CDC studies of PWID (Lansky et al., 2007). These simulated datasets are data-matched on:

- Estimated prevalence of and mean degree by one high-homophily binary variable denoted $\mathbf{z}$
- Estimated prevalence of and mean degree by one moderate-homophily binary variable denoted $\mathbf{x}$
- The joint prevalence (therefore relationship between) $\mathbf{z}$ and $\mathbf{x}$.

Simulated networks of size $N = 10,000$ nodes were generated as simulations from exponential random graph models (ERGMs) using the `statnet` `R` package (Handcock et al., 2015; R Core Team, 2018). These networks are ideal for applying the proposed methodology because they reflect realistic characteristics of two related variables which may influence sampling, in a setting in which the true prevalence of the outcome variable $\mathbf{z}$ is known. One hundred network replicates for each of the fourteen realistic networks were simulated. In addtion, RDS samples were drawn for each of the simulated networks starting with ten seeds. Everyone recruited up to 2 individuals, without replacement, until a sample size of 500 was attained. Recruitment was assumed to follow a between group differential recruitment scheme, where $\phi = 2$.

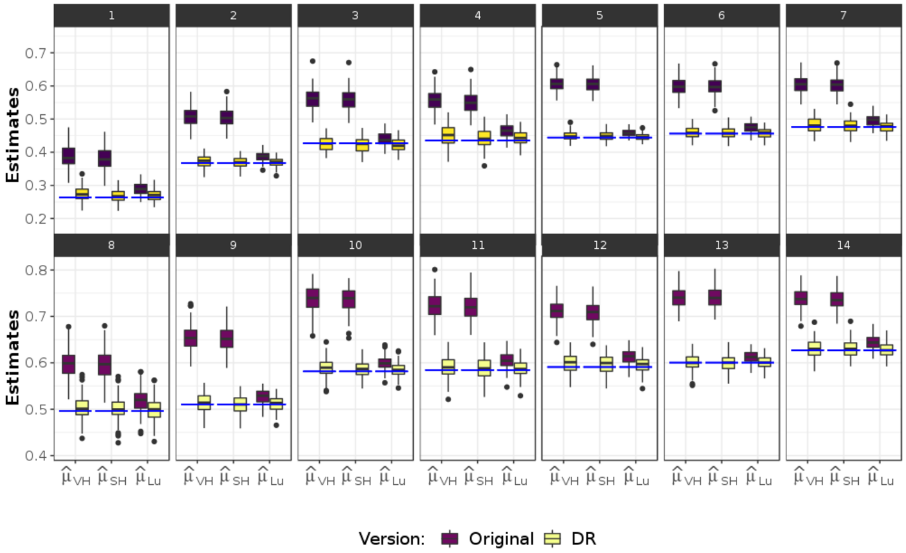


Fig 5. *Prevalence estimates for 14 populations simulated to reproduce realistic features of CDC data and where RDS is simulated with between group DR on* $\mathbf{x}$ *such that* $\phi = 2$. *Estimators are presented in the following order:* $\hat\mu_{VH}$, $\hat\mu^b_{VH.dr}$, $\hat\mu_{SH}$, $\hat\mu^b_{SH.dr}$ $\hat\mu_{Lu}$ *and* $\hat\mu^b_{Lu.dr}$. *The blue horizontal line represents the true population prevalence for the variable* $\mathbf{x}$.

The simulated distributions of the prevalence estimates for this application are shown in Figure 5. Every rectangle represents one of the fourteen simulated population conditions. The populations are ordered based on the prevalence of the outcome variable, which are represented by the blue horizontal lines. Within each area, the distribution of the six prevalence estimators discussed in this manuscript are displayed, that is, the three original estimators in purple paired with their respective extended version in yellow. For the purpose of the illustration, $\mathbf{z}$ is assumed to be equal to $\mathbf{x}$, that is, the DR occurs on the outcome variable. As it may be observed from Figure 5, while $\hat\mu_{Lu}$ is considerably more robust to DR than the VH and SH, all three versions of the DR estimators reduce the DR bias. The root mean square error (RMSE) is smaller for the extended estimators and is minimized with the $\hat{\mu}^b_{Lu.dr}$ under every scenario.

Figure 6 shows the coverage rates of the 95% confidence intervals constructed based on the original methodology and the two extensions of the bootstrap variance estimators. The results are consistent with those obtained in the simulation study. The highest coverage rates are produced from the Lu extended procedures. Also, similar to previous results, the original VH and SH procedures produce very low coverage rates, sometimes none of the intervals contains the true value even for a relatively small value of DR ($\phi = 2$).

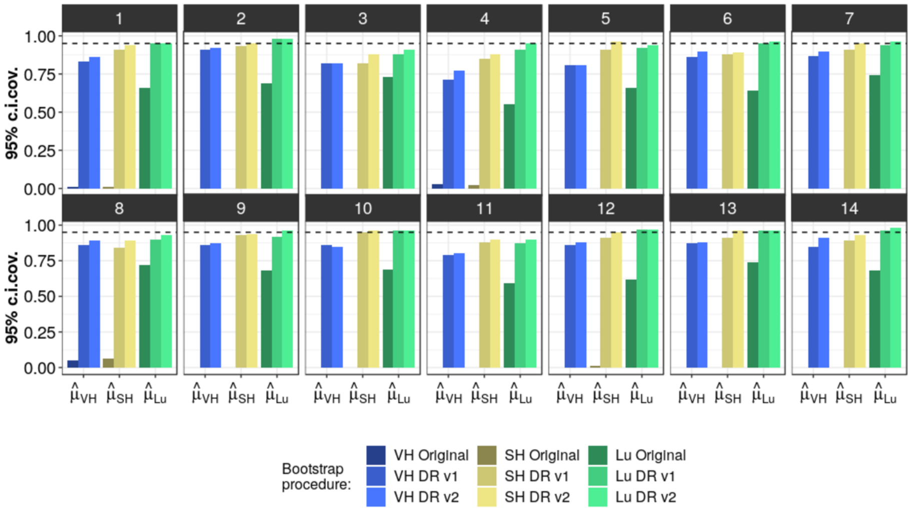


Fig 6. *95% confidence interval coverage rates, where the coverage rates are the percentage of the intervals including the true population proportion μ. The dashed line is set at 95%.*

Although not presented in this section, inference on an outcome variable $\mathbf{z}$ different than $\mathbf{x}$ has also been performed. The induced DR on $\mathbf{z}$ resulting from DR on $\mathbf{x}$ was however minimal thus suggesting a weak relation between the two variables. In most populations, the RMSE was also minimized by $\hat{\mu}^b_{Lu.dr}$. However, in some cases, $\hat{\mu}_{Lu}$ produced a similar but lower RMSE than $\hat{\mu}^b_{Lu.dr}$. As for the coverage rates of the 95% confidence intervals, Lu DR procedures appeared to outperform or perform similarly to other methods in the majority of the populations thus highlighting the advantage of utilizing both ego network information when available.

6. Discussion

Sampling hard-to-reach populations is a challenging problem. RDS has provided ways to circumvent some of the issues specific to those populations that make the use of traditional sampling methods impractical. However, the sampling process under RDS is out of the control of the researchers conducting the studies and therefore, this sampling method is highly susceptible to biases induced by participants' behaviors. The main contribution of this work is to introduce inferential methodologies correcting existing RDS prevalence estimators and their uncertainty estimators for biases induced by various forms of differential recruitment (DR), where DR is understood as a systematic departure from recruitment completely at random. The methodology addresses biases induced by DR arising from participants preferentially recruiting a sub-group of their contacts or from a systematic nonresponse of a sub-group of their contacts. However, it does not explicitly distinguish the source of the DR bias.

The estimators presented in this work suppose participants' sampling probabilities may be estimated from the stationary distribution of a random walk (RW) on the state space of the network nodes. The derivation of the stationary distribution for the original estimators assumes that participants recruit completely at random among their contacts in the target population. Our approach modifies this assumption and instead proposes three sampling schemes under which participants systematically recruit individuals based on one of their nodal characteristics or based on their type of relationship with them. By explicitly defining those sampling schemes we were able to derive the RW characterizing those behaviors and their associated stationary distributions. The revised estimators rely on the stationary distributions of the modified RW. Results from the simulation study show that this methodology greatly reduces biases induced by the various forms of DR. However, these methods require additional data about participants' ego-network compositions.

The comparison of the estimators' bias in our simulation study suggests that $\hat{\mu}^b_{Lu.dr}$ generally outperforms the discussed alternative estimators. However, this estimator requires participants to correctly report the ego-network composition on both the outcome variable and on the characteristic driving DR. In the context of RDS, it might be unrealistic to assume participants will be able to provide the outcome variable ego-network composition, as in the case of disease, this information is generally not visible to other members of the target population. Under such circumstances, it is recommendable to instead use $\hat{\mu}^b_{VH.dr}$ to estimate the prevalence, as this estimator does not require this information and produces smaller biases than $\hat{\mu}_{VH}$ and $\hat{\mu}_{SH}$ in presence of DR.

We have also proposed uncertainty estimators. The uncertainty is estimated through a bootstrap procedure capturing the variability associated with the RW sampling as well as with the estimation of the magnitude of the DR parameter. Results from the simulation study show that the variance estimators perform relatively well with non homophilous networks. Combined with the lower bias of the point estimate, this significantly improves the coverage rates of the 95% confidence intervals from the original version of the VH and SH estimators in presence of DR. The coverage rates under those assumptions are also greater than under Lu's methodology, but to a lesser extent. With homophilous networks however, the proposed bootstrap procedures tend to underestimate variability. This may be explained by the fact that those procedures do not reflect some of the RDS specific features. Although the underestimation of the variance affects the width of the 95% confidence intervals, the coverage rates for $\hat{\mu}^b_{VH.dr}$ and

$\hat{\mu}^b_{Lu.dr}$ are significantly better than those produced by the conventional estimators when with $\phi = 2$ or 4. We conclude that the proposed extended methods improve the inference in the presence of DR despite the underestimation of the variance.

We have applied the proposed methods to data simulated to match some features of data collected by the US Centers for Disease Control and Prevention (CDC) for behavioral surveillance among people who inject drugs (PWID). The findings from this analysis are similar to the conclusions obtained from the simulation study. In presence of DR, there is a significant reduction in the bias for the extended versions of the VH and SH estimators and a moderate reduction in the case of the extended version of the Lu estimator. The example also highlights that in cases where the outcome variable differs from the DR characteristic, the improvement is not as significant when the DR is weakly related to the outcome variable.

Although the methodology presented in this manuscript addresses specific forms of DR, we believe it provides a general framework where one could assess different forms of DR by deriving new stationary distributions for alternative random walks. For instance, one might consider treating nonuniform DR. Consequently, we believe that the proposed methodologies are promising and could significantly improve traditional estimators when participants do not recruit at random.

Appendix

A.1. Stationary distributions with DR

In this section, we prove *Results 3.1–3.3*, which correspond to the stationary distributions of the random walk with three forms of differential recruitment.

Proof Result 3.1.

By assumption, $p^b_{ij} = \dfrac{\phi_b^{x_j} y_{ij}}{\sum_{j'=1}^N \phi_b^{x_{j'}} y_{ij'}}$.

Therefore, we have that:

$$\begin{aligned}\sum_{i=1}^N \pi_i^b p_{ij}^b &= \sum_{i=1}^N \left[\frac{\phi_b^{x_i}(\phi_b d_{i1}^x + d_{i0}^x)}{K}\right]\left[\frac{\phi_b^{x_j} y_{ij}}{\sum_{j'=1}^N \phi_b^{x_{j'}} y_{ij'}}\right] \\ &= \frac{\phi_b^{x_j}}{K}\sum_{i=1}^N \phi_b^{x_i} y_{ji} = \frac{\phi_b^{x_j}(\phi_b d_{j1}^x + d_{j0}^x)}{K} = \pi_j^b,\end{aligned}$$

where K is a normalizing constant such that $\sum_{i=1}^N \pi_i^b = 1$. Therefore, π_i^b satisfies the global balance equations for all $i \in \{1, 2, ..., N\}$ and $\pi^b = \{\pi_1^b, \pi_2^b, ..., \pi_N^b\}$ is the stationary distribution for this RW. □

Proof Result 3.2.

By assumption, $p_{ij}^w = \dfrac{[\phi_w^{x_i} x_j + \phi_w^{1-x_i}(1-x_j)]y_{ij}}{\sum_{j'=1}^N [\phi_w^{x_i} x_{j'} + \phi_w^{1-x_i}(1-x_{j'})]y_{ij'}}$.

Therefore, we have that:

$$\begin{aligned}\sum_{i=1}^N \pi_i^w p_{ij}^w &= \sum_{i=1}^N \left[\frac{\phi_w^{x_i} d_{i1}^x + \phi_w^{1-x_i} d_{i0}^x}{K}\right]\left[\frac{[\phi_w^{x_i} x_j + \phi_w^{1-x_i}(1-x_j)]y_{ij}}{\sum_{j'=1}^N [\phi_w^{x_i} x_{j'} + \phi_w^{1-x_i}(1-x_{j'})]y_{ij'}}\right] \\ &= \frac{1}{K}\sum_{i=1}^N [\phi_w^{x_j} x_i + \phi_w^{1-x_j}(1-x_i)]y_{ji} = \frac{\phi_w^{x_j} d_{j1}^x + \phi_w^{1-x_j} d_{j0}^x}{K} = \pi_j^w,\end{aligned}$$

where K is a normalizing constant such that $\sum_{i=1}^N \pi_i^w = 1$. Therefore, π_i^w satisfies the global balance equations for all $i \in \{1, 2, ..., N\}$ and $\pi^w = \{\pi_1^w, \pi_2^w, ..., \pi_N^w\}$ is the stationary distribution for this RW. □

Proof Result 3.3.

By assumption, $p_{ij}^t = \dfrac{\phi_t^{w_{ij}} y_{ij}}{\sum_{j'=1}^N \phi_t^{w_{ij'}} y_{ij'}}$.

Therefore, we have that:

$$\begin{aligned}\sum_{i=1}^N \pi_i^t p_{ij}^t &= \sum_{i=1}^N \left[\frac{\phi_t d_{i1}^w + d_{i0}^w}{K}\right]\left[\frac{\phi_t^{w_{ij}} y_{ij}}{\sum_{j'=1}^N \phi_t^{w_{ij'}} y_{ij'}}\right] \\ &= \frac{1}{K}\sum_{i=1}^N \phi_t^{w_{ji}} y_{ji} = \frac{\phi_t d_{j1}^w + d_{j0}^w}{K} = \pi_j^t,\end{aligned}$$

where K is a normalizing constant such that $\sum_{i=1}^N \pi_i^t = 1$. Therefore, π_i^t satisfies the global balance equations for all $i \in \{1, 2, ..., N\}$ and $\pi^t = \{\pi_1^t, \pi_2^t, ..., \pi_N^t\}$ is the stationary distribution for this RW. □

A.2. Consistency of the prevalence estimators

This section discusses the consistency of the estimators $\hat{\mu}^*_{VH.dr}$, $\hat{\mu}^*_{Lu.dr}$ and $\hat{\mu}^*_{SH.dr}$ and the necessary conditions for them to be consistent. All derivations are based on the assumptions stated in the manuscript, that is, that the sampling is performed through a random walk over a fully connected undirected network. This asymptotic framework is consistent with Volz and Heckathorn (2008) and Goel and Salganik (2009), who assume a RW on the state space of the nodes of a fully connected network, but contrasts with the framework used by Li et al. (2017), who consider asymptotic unbiasedness based on RDS being represented as a tree indexed Markov process.

These proofs addresses consistency as the length of a Markov chain (MC) sample drawn from a fixed network increases. We therefore condition on the fixed network structure, and these proofs are agnostic to the network distribution from which the network was drawn. Finally, since we let the MC sample size go to infinity, the estimators $\hat{\mu}^*_{VH.dr}$, $\hat{\mu}^*_{Lu.dr}$ and $\hat{\mu}^*_{SH.dr}$ are expressed in a slightly different form than shown in the manuscript. In particular, we now assume that i represents the indexing of the MC sample order instead of the indexing of the nodes in the network.

Result A.1. *Let R_t denote the state at step t of a Markov chain (MC) on the nodes of a fully connected undirected network without self ties. Assume that this MC has transition probabilities defined in equations (3.1), (3.2) or (3.3), where $\phi_* > 0$. Also, suppose that ϕ_* is estimated by maximizing the likelihood function shown in equation (3.7) using n steps of the MC. Then, $\hat{\mu}^*_{VH.dr}$ is a consistent estimator for μ if $\sum_{i=1}^n 1/\widehat{d}^*_i$ does not tend to 0 as $n \to \infty$.*

Proof. By the properties of the maximum likelihood estimators, $\hat{\phi}_*$ is a consistent estimator for ϕ_*. As such, equations (3.8)–(3.10) are consistent estimators for equations (3.4)–(3.6), respectively. Therefore, the estimators for the sampling probabilities are consistent for the true sampling probabilities.

By Hastings (1970) and Chebyshev's inequality, since the estimators for the sampling probabilities are consistent, $\hat{\mu}^*_{VH.dr}$ is a consistent estimator for μ if its denominator, $\sum_{i=1}^n 1/\widehat{d}^*_i$, does not tend to zero for large n. □

Result A.2. *Let R_t denote the state at step t of a MC on the nodes of a fully connected undirected network without self ties. Assume that this MC has transition probabilities defined in equations (3.1), (3.2) or (3.3), where $\phi_* > 0$. Also, suppose that ϕ_* is estimated by maximizing the likelihood function shown in equation (3.7) using n steps of the MC. Then, $\hat{\mu}^*_{Lu.dr}$ is a consistent estimator for μ if $\sum_{i=1}^n 1/\widehat{d}^*_i$ and $\sum_{i=1}^n (1-z_i) d^z_{i1}/\widehat{d}^*_i$ do not tend to zero as $n \to \infty$.*

Proof. To show that $\hat{\mu}^*_{Lu.dr}$ is a consistent estimator for μ, it suffices to show that $c = \frac{\sum_{i=1}^n z_i d^z_{i0}/\widehat{d}^*_i}{\sum_{i=1}^n (1-z_i) d^z_{i1}/\widehat{d}^*_i}$ is a consistent estimator for 1. This is due to the fact that, as per equation (3.12) when $c \to 1$, then $\hat{\mu}^*_{Lu.dr} \to \hat{\mu}^*_{VH.dr}$ and as such, using Result A.1, $\hat{\mu}^*_{Lu.dr}$ is consistent for μ.

Since the estimators for the sampling probabilities are consistent for the true sampling probabilities, then:

- $\sum_{i=1}^n z_i d^z_{i0}/\widehat{d}^*_i$ is a consistent estimator for T_{10}; and
- $\sum_{i=1}^n (1-z_i) d^z_{i1}/\widehat{d}^*_i$ is a consistent estimator for T_{01},

where, T_{10} is the number of ties from nodes in $\mathcal{Z}^1$ to nodes in $\mathcal{Z}^0$ and T_{01} is the number of ties from nodes in $\mathcal{Z}^0$ to nodes in $\mathcal{Z}^1$. Since the network is assumed undirected, we have that $T_{10} = T_{01}$. We therefore conclude that c is a consistent estimator for 1. Consequently, $\hat{\mu}^*_{Lu.dr}$ is a consistent estimator for μ provided $\sum_{i=1}^n 1/\widehat{d}^*_i$ or $\sum_{i=1}^n (1-z_i) d^z_{i1}/\widehat{d}^*_i$ do not tend to zero for large n. □

We demonstrate the consistency of $\hat{\mu}^b_{SH.dr}$ in the specific case of between-group DR on the outcome variable $\mathbf{z}$. Similar derivations may be obtained for the other forms of DR.

Result A.3. *Let R_t denote the state at step t of a MC on the nodes of a fully connected undirected network without self ties. Assume that this MC has* *transition probabilities defined in equations (3.1), where $x_i = z_i$ and $\phi_b > 0$. Also, suppose that ϕ_b is estimated by maximizing the likelihood function shown in equation (3.7) using n steps of the MC. Then, $\hat{\mu}^b_{SH.dr}$ is a consistent estimator for μ if $\sum_{i=1}^n 1/\widehat{d_i^b}$ and $\sum_{i=1}^n (1-z_i)d_i/\widehat{d_i^b}$ do not tend to zero as $n \to \infty$.*

Proof. To show that $\hat{\mu}^b_{SH.dr}$ is a consistent estimator for μ, it suffices to show that

$$c = \left(\frac{\hat{C}^{SH.b}_{10}}{\hat{C}^{SH.b}_{01}}\right) \frac{\sum_{i=1}^n z_i d_i/\widehat{d_i^b}}{\sum_{i=1}^n (1-z_i)d_i/\widehat{d_i^b}}$$

is a consistent estimator for 1. This is due to the fact that, as per equation (3.17) when $c \to 1$, then $\hat{\mu}^b_{SH.dr} \to \hat{\mu}^b_{VH.dr}$ and as such, using Result A.1, $\hat{\mu}^b_{SH.dr}$ is consistent for μ.

Using similar arguments as in the ones used in the proof of Results A.1 and A.2, we note that

$$\frac{\sum_{i=1}^n z_i d_i/\widehat{d_i^b}}{\sum_{i=1}^n (1-z_i)d_i/\widehat{d_i^b}}$$

is a consistent estimator for $\frac{T_{1\cdot}}{T_{0\cdot}}$, provided that $\sum_{i=1}^n (1-z_i)d_i/\widehat{d_i^b}$ does not converge to 0 for large n, where $T_{1\cdot}$ and $T_{0\cdot}$ are the total number of ties for nodes in $\mathcal{Z}^1$ and $\mathcal{Z}^0$, respectively. Therefore, for c to be consistent for 1, we need to show that $\frac{\hat{C}^{SH.b}_{10}}{\hat{C}^{SH.b}_{01}}$ is a consistent estimator for $\frac{T_{0\cdot}}{T_{1\cdot}}$.

We first observe that $\hat{C}^{SH.b}_{10} = \frac{\widehat{\phi}^z_b r_{10}}{\widehat{\phi}^z_b r_{10} + r_{11}} = \frac{\widehat{\phi}^z_b s_{10}}{\widehat{\phi}^z_b s_{10} + s_{11}}$, where $s_{kl} = r_{kl}/(n-1)$ for $k, l \in \{0,1\}$ is the proportion of recruitments from nodes in $\mathcal{Z}^k$ to nodes in $\mathcal{Z}^l$ in the MC of n steps. Secondly, let us define the corresponding random variable $S_{kl,n} = \frac{1}{n-1}\sum_{g=1}^{n-1} \mathbb{1}_{[z_g = k \cap z_{g+1} = l]}$. Then, $E[S_{kl,n}] = \sum_{i\in\mathcal{Z}^k}\sum_{j\in\mathcal{Z}^l} p^b_{ij}\pi^b_i$, which, for fixed ϕ_b, leads to:

- $E[S_{01,n}] = \sum_{i\in\mathcal{Z}^0}\sum_{j\in\mathcal{Z}^1} p^b_{ij}\pi^b_i = \sum_{i\in\mathcal{Z}^0}\sum_{j\in\mathcal{Z}^1}\left[\frac{\phi_b y_{ij}}{\phi_b d^z_{i1} + d^z_{i0}}\right]\left[\frac{\phi_b d^z_{i1} + d^z_{i0}}{K}\right] = \frac{\phi_b T_{01}}{K}$,
- $E[S_{00,n}] = \sum_{i\in\mathcal{Z}^0}\sum_{j\in\mathcal{Z}^0} p^b_{ij}\pi^b_i = \frac{T_{00}}{K}$,
- $E[S_{10,n}] = \sum_{i\in\mathcal{Z}^1}\sum_{j\in\mathcal{Z}^0} p^b_{ij}\pi^b_i = \frac{\phi_b T_{10}}{K}$, and
- $E[S_{11,n}] = \sum_{i\in\mathcal{Z}^1}\sum_{j\in\mathcal{Z}^1} p^b_{ij}\pi^b_i = \frac{(\phi_b)^2 T_{11}}{K}$,

where K is the normalizing constant of the stationary distribution and where T_{kl} is the number of ties from nodes in $\mathcal{Z}^k$ to nodes in $\mathcal{Z}^l$ for $k, l \in \{0,1\}$. Also, the variance for these random variables is:

$$Var[S_{kl,n}] = \left[\frac{1}{n-1}\right]^2 \sum_{\mathbf{q}\in\Omega}\left[\sum_{t=1}^{n-1}\mathbb{1}_{[z_{q_t}=k, z_{q_{t+1}}=l]} - (n-1)E[S_{kl,n}]\right]^2 \pi_{q_1}\prod_{t=1}^{n-1} p^b_{q_t q_{t+1}}$$

where $\mathbf{q} = \{q_1, ..., q_n\}$ is a possible sequence of n states of the MC, Ω is the sample space of $\mathbf{q}$, and where $\pi_{q_1}\prod_{t=1}^{n-1} p^b_{q_t q_{t+1}}$ is the probability of observing the sequence of states $\mathbf{q}$. Consequently, the variance tends to zero as the sample size increases since $\left(\sum_{t=1}^{n-1}\mathbb{1}_{[z_{q_t}=k, z_{q_{t+1}}=l]} - (n-1)E[S_{kl,n}]\right)^2 < (n-1)^2$. By Chebychev's inequality and the asymptotic properties of the MLE, we conclude that s_{kl} is a consistent estimator for $E[S_{kl,n}]$, and therefore, $\hat{C}_{10}^{SH.b}$ is a consistent estimator for

$$\frac{\phi_b E[S_{10,n}]}{\phi_b E[S_{10,n}] + E[S_{11,n}]} = \frac{\phi_b \times \phi^b T_{10}}{\phi_b \times \phi_b T_{10} + (\phi_b)^2 T_{11}} = \frac{T_{10}}{T_{1\cdot}}.$$

Similarly, $\hat{C}_{01}^{SH.b}$ is a consistent estimator for $\frac{T_{01}}{T_{0\cdot}}$. We recall that under the assumption that the network is undirected, $T_{01} = T_{10}$ and therefore, this shows that $\left(\frac{\hat{C}_{10}^{SH.b}}{\hat{C}_{01}^{SH.b}}\right)$ is a consistent estimator for $\frac{T_{0\cdot}}{T_{1\cdot}}$, which also proves that c is consistent for 1. □

A.3. Multilevel between group DR

Suppose that the between group DR variable $\mathbf{x}$ is such that $\mathbf{x} \in \{1, ..., G\}^N$. Therefore, there are G possible values for x_i, for $i \in \{1, ..., N\}$. Without loss of generality, also suppose that the G-th group is the reference group for the DR, which we define as the group serving as the comparison to measure the magnitude of the DR:

$$\phi_g = \begin{cases} \dfrac{\Pr(R_t = i|\ R_{t-1} = j,\ y_{ij} = 1,\ x_i = g)}{\Pr(R_t = i|\ R_{t-1} = j,\ y_{ij} = 1,\ x_i = G)}, & g \in \{1, ..., G-1\} \\ 1, & g = G \end{cases}.$$

Then, assuming a RW on the nodes of a single component undirected network with transition probabilities:

$$p_{ij} = \frac{\phi_{x_j} y_{ij}}{\sum_{j'=1}^{N} \phi_{x_{j'}} y_{ij'}}, \tag{A.1}$$

it is possible to demonstrate that the stationary distribution of this RW is such that $\pi_i \propto \phi_{x_i} \sum_{g=1}^{G} d_i^g \phi_g$, where $d_i^g = \sum_{j=1}^{N} y_{ij} \mathbb{1}_{[x_j = g]}$. Since the ϕ_{x_i}'s are unknown, they are estimated by maximizing the following likelihood function:

$$L(\phi_1, ..., \phi_{G-1} | R = r) \propto \prod_{i \in \mathcal{S}^1 \setminus S_0} p_{r_{i-1} r_i}, \tag{A.2}$$

where $p_{r_{i-1} r_i}$ are the transition probabilities according to equation A.1 for each of the observed recruitment. Finally, the extended VH and Lu prevalence estimators remain as shown in equations 3.11 and 3.12, with $\hat{d}_i^*$ now proportional to $\hat{\phi}_{x_i} \sum_{g=1}^{G} d_i^g \hat{\phi}_g$ to take into account the multilevel between group DR.

We have performed an additional simulation study to evaluate the estimators under between group DR with multiple groups. In this simulation, we created a DR variable x with 3 groups such that: $\phi_1 = 4, \phi_2 = 2,$ and $\phi_3 = 1$. We assumed that 35% of the population belonged to the first group, 35% to the second group, and 30% to the last group. Also, we assumed the prevalence of the outcome of interest to be 20%, equally split between the first and second group. We applied the multilevel methodology for 3-group between group DR, first assuming that the model specification was correct (with g1) and then assuming that the researcher was unaware of the first group (without g1). Therefore, the model misspecification scenario ignores the groups with the highest level of DR. We calculated $\hat{\mu}_{VH}$, $\hat{\mu}_{VH.dr}$, $\hat{\mu}_{Lu}$ and $\hat{\mu}_{Lu.dr}$. The results are shown in Table 3. We note the good performance of $\hat{\mu}_{VH.dr}$ and $\hat{\mu}_{Lu.dr}$ under the scenario "with g1" when making inference for either the prevalence of $g2$ (35%) or the outcome variable (20%). Model misspecification does not seem to alter significantly the performance of $\hat{\mu}_{VH.dr}$ nor $\hat{\mu}_{Lu.dr}$ under the simulated scenario when estimating the prevalence of $g2$. However, we note that $\hat{\mu}_{VH.dr}$ is biased when estimating the prevalence of the outcome variable. Therefore, we would recommend using $\hat{\mu}_{Lu.dr}$ to minimize the impact of model misspecification.

TABLE 3
Model Misspecification.

Estimator	mean (truth =0.35)	sd	mean (truth =0.20)	sd
$\hat{\mu}_{VH}$	0.515	0.034	0.239	0.028
$\hat{\mu}_{VH.dr}$ (with g1)	0.360	0.023	0.202	0.024
$\hat{\mu}_{VH.dr}$ (without g1)	0.345	0.023	0.241	0.028
$\hat{\mu}_{Lu}$	0.369	0.016	0.204	0.013
$\hat{\mu}_{Lu.dr}$ (with g1)	0.353	0.017	0.200	0.022
$\hat{\mu}_{Lu.dr}$ (without g1)	0.348	0.021	0.204	0.022

A.4. Transition probabilities in uncertainty estimators

In this section, we present the transition probabilities used in our extension of Lu's bootstrap procedure for the cases of within group and tie differential recruitment. In both Table 4 and Table 5, i represents the index of the recruiting node and $j \neq i$ the index of any available node for recruitment.

TABLE 4
Within group DR bootstrap transition probabilities.

	$j \in \mathcal{X}^0$	$j \in \mathcal{X}^1$
$i \in \mathcal{X}^0$	$\frac{1}{n_0^x}\left[1 - \frac{\sum_{i=1}^N S_i(1-x_i)d_{i1}^x/d_i^w}{n_0^x}\right]$	$\frac{1}{n_1^x}\left[\frac{\sum_{i=1}^N S_i(1-x_i)d_{i1}^x/d_i^w}{n_0^x}\right]$
$i \in \mathcal{X}^1$	$\frac{1}{n_0^x}\left[\frac{\sum_{i=1}^N S_i x_i d_{i0}^x/d_i^w}{n_1^x}\right]$	$\frac{1}{n_1^x}\left[1 - \frac{\sum_{i=1}^N S_i x_i d_{i0}^x/d_i^w}{n_1^x}\right]$

TABLE 5
Tie DR bootstrap transition probabilities.

	$j \in \mathcal{X}^0$	$j \in \mathcal{X}^1$
$i \in \mathcal{X}^0$	$\frac{1}{n_0^x}\left[1 - \frac{\sum\sum_{i\neq j}(\hat{\phi}^t)^{w_{ij}} S_i(1-x_i)x_j y_{ij}/d_i^t}{n_0^x}\right]$	$\frac{1}{n_1^x}\left[\frac{\sum\sum_{i\neq j}(\hat{\phi}^t)^{w_{ij}} S_i(1-x_i)x_j y_{ij}/d_i^t}{n_0^x}\right]$
$i \in \mathcal{X}^1$	$\frac{1}{n_0^x}\left[\frac{\sum\sum_{i\neq j}(\hat{\phi}^t)^{w_{ij}} S_i x_i(1-x_j)y_{ij}/d_i^t}{n_1^x}\right]$	$\frac{1}{n_1^x}\left[1 - \frac{\sum\sum_{i\neq j}(\hat{\phi}^t)^{w_{ij}} S_i x_i(1-x_j)y_{ij}/d_i^t}{n_1^x}\right]$

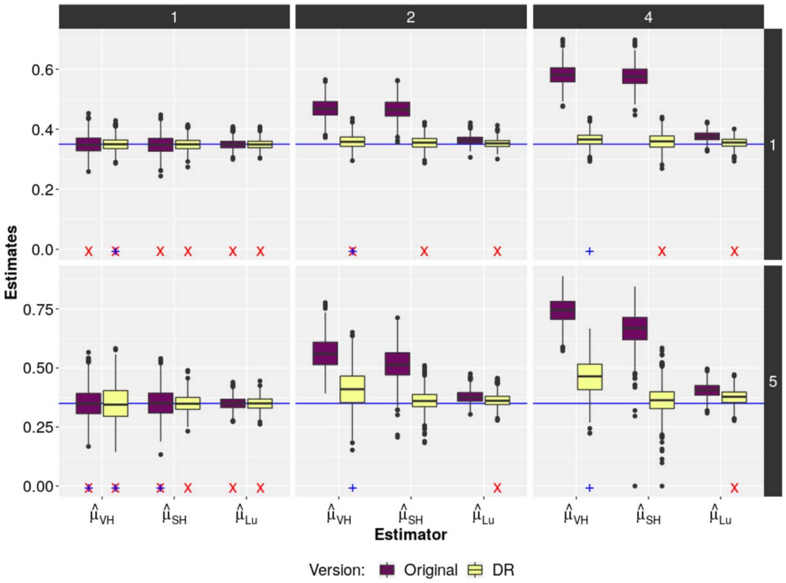


FIG 7. *Estimates produced with varying levels of network homophily, that is, $\tau \in \{1, 5\}$, (horizontal panels) and between group DR on $\mathbf{x}$, that is, $\phi \in \{1, 2, 4\}$ (vertical panels). Estimators are presented in the following order: $\hat{\mu}_{VH}$, $\hat{\mu}^b_{VH.dr}$, $\hat{\mu}_{SH}$, $\hat{\mu}^b_{SH.dr}$ $\hat{\mu}_{Lu}$ and $\hat{\mu}^b_{Lu.dr}$. The blue horizontal line represents the true population prevalence for the variable $\mathbf{x}$.*

TABLE 6

Mean, standard deviation (sd) and root-mean-square error (RMSE) of the estimates produced with varying levels of network homophily, $\tau \in \{1, 5\}$, and between group DR on $\mathbf{x}$, $\phi \in \{1, 2, 4\}$. The true population prevalence for the variable $\mathbf{z}$ is $\mu = 0.2$.

Estimator	ϕ	τ	mean	sd	RMSE	ϕ	τ	mean	sd	RMSE
$\hat{\mu}_{VH}$	1	1	0.201	0.026	0.835	1	5	0.201	0.042	1.342
$\hat{\mu}_{VH.dr}$	1	1	0.201	0.024	0.753	1	5	0.201	0.049	1.539
$\hat{\mu}_{SH}$	1	1	0.200	0.027	0.842	1	5	0.201	0.042	1.331
$\hat{\mu}_{SH.dr}$	1	1	0.200	0.016	0.519	1	5	0.199	0.035	1.103
$\hat{\mu}_{Lu}$	1	1	0.200	0.013	0.411	1	5	0.200	0.025	0.788
$\hat{\mu}_{Lu.dr}$	1	1	0.200	0.013	0.412	1	5	0.201	0.029	0.915
$\hat{\mu}_{VH}$	2	1	0.269	0.030	2.372	2	5	0.321	0.046	4.103
$\hat{\mu}_{VH.dr}$	2	1	0.205	0.022	0.718	2	5	0.235	0.050	1.927
$\hat{\mu}_{SH}$	2	1	0.267	0.032	2.336	2	5	0.316	0.046	3.934
$\hat{\mu}_{SH.dr}$	2	1	0.203	0.017	0.554	2	5	0.265	0.039	2.409
$\hat{\mu}_{Lu}$	2	1	0.208	0.013	0.479	2	5	0.265	0.028	2.235
$\hat{\mu}_{Lu.dr}$	2	1	0.202	0.013	0.405	2	5	0.219	0.029	1.114
$\hat{\mu}_{VH}$	4	1	0.332	0.031	4.288	4	5	0.425	0.040	7.222
$\hat{\mu}_{VH.dr}$	4	1	0.208	0.020	0.675	4	5	0.264	0.046	2.493
$\hat{\mu}_{SH}$	4	1	0.329	0.033	4.203	4	5	0.418	0.042	7.009
$\hat{\mu}_{SH.dr}$	4	1	0.205	0.018	0.608	4	5	0.330	0.036	4.261
$\hat{\mu}_{Lu}$	4	1	0.214	0.013	0.603	4	5	0.327	0.025	4.096
$\hat{\mu}_{Lu.dr}$	4	1	0.203	0.013	0.420	4	5	0.236	0.030	1.472

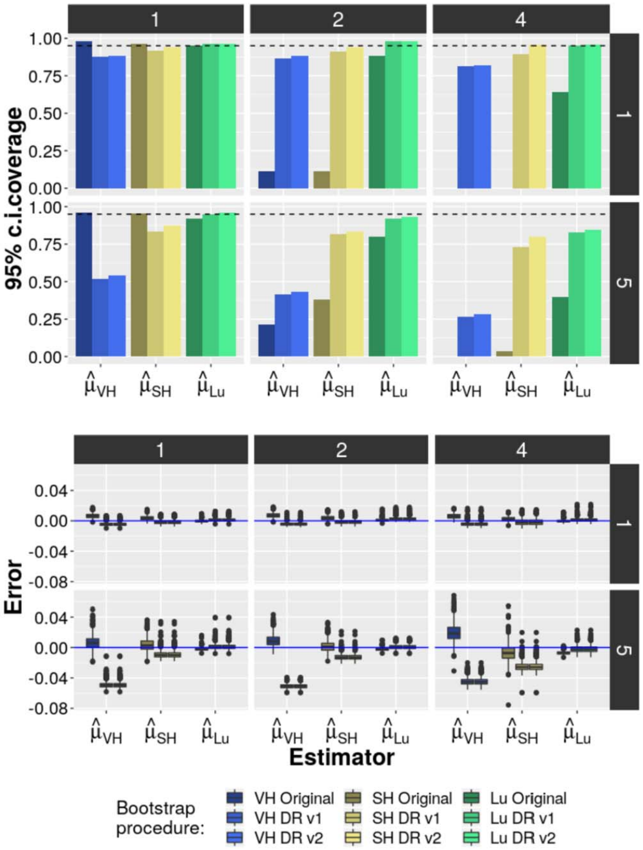


Fig 8. *95% confidence interval coverage rates, where the coverage rates are the percentage of the intervals including the true population proportion μ of 35%. The dashed line is set at 95%. Bias of the standard deviation estimates calculated as $\bar{\hat{\sigma}} - s$, where $\bar{\hat{\sigma}}$ is the average estimated standard deviation under a bootstrap methodology and s is the sample standard deviation.*

A.5. Simulation study results

In this Section, we present results from the simulation study, which correspond to the results presented in Figure 3. Also, we present additional results.

Figure 7 and 8 correspond to the point estimate and variance estimation simulation study results when the object of inference is the prevalence of the DR characteristic x. This variable is also assumed to be the one inducing DR.

Figure 9 compares the point estimate results for the three forms of differential recruitment for a moderate level of DR ($\phi = 2$) and for two levels of network homophily.

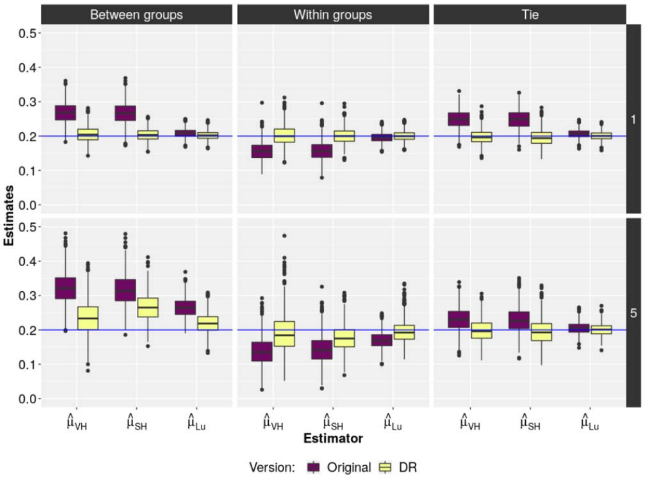


Fig 9. *Point estimates produced at varying levels of network homophily, that is, $\tau \in \{1, 5\}$, (horizontal panels) under three forms of DR on* $\mathbf{x}$ *with a parameter $\phi = 2$: between group, within group and tie DR (vertical panels). Estimators are presented in the following order: $\hat{\mu}_{VH}$, $\hat{\mu}^*_{VH.dr}$, $\hat{\mu}_{SH}$, $\hat{\mu}^*_{SH.dr}$ $\hat{\mu}_{Lu}$ and $\hat{\mu}^*_{Lu.dr}$. The blue horizontal line represents the true population prevalence for the variable* $\mathbf{z}$.

References

Beaudry, I. S., Gile, K. J. and Mehta, S. H. (2017). Inference for respondent-driven sampling with misclassification. *The Annals of Applied Statistics* **11** 2111–2141. MR3743290

Crawford, F. W., Aronow, P. M., Zeng, L. and Li, J. (2018). Identification of Homophily and Preferential Recruitment in Respondent-Driven Sampling. *American Journal of Epidemiology* **187** 153–160.

Currarini, S., Jackson, M. O. and Pin, P. (2009). An Economic Model of Friendship: Homophily, Minorities, and Segregation. *Econometrica* **77** 1003–1045. MR2547067

Frank, O. and Strauss, D. (1986). Markov Graphs. *Journal of the American Statistical Association* **81** 832–842. MR0860518

Frost, S. D. W., Brouwer, K. C., Cruz, M. A. F., Ramos, R., Ramos, M. E., Lozada, R. M., Magis-Rodriguez, C. and Strathdee, S. A. (2006). Respondent-Driven Sampling of Injection Drug Users in Two U.S.-Mexico Border Cities: Recruitment Dynamics and Impact on Estimates of HIV and Syphilis Prevalence. *Journal of Urban Health* **83** 83–97.

Gallagher, K. M., Sullivan, P., Lansky, A., and Onorato, I. M. (2007). Behavioral Surveillance among People at Risk for HIV Infection in the U.S.: The National HIV Behavioral Surveillance System. *Public Health Reports* **122** 32–38. MR3823208

Gile, K. J. and Handcock, M. S. (2010). Respondent-Driven Sampling: An Assessment of Current Methodology. *Sociological Methodology* **40** 285–327.

Gile, K. J., Johnston, L. G. and Salganik, M. J. (2015). Diagnostics for respondent-driven sampling. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **178** 241–269. MR3291770

Goel, S. and Salganik, M. J. (2009). Respondent-driven sampling as Markov chain Monte Carlo. *Statistics in medicine* **28** 2202–2229. MR2751515

Handcock, M. S., Fellows, I. E. and Gile, K. J. (2015). RDS: Respondent-Driven Sampling, Los Angeles, CA R package version 0.7-2.

Handcock, M. S., Hunter, D. R., Butts, C. T., Goodreau, S. M., Krivitsky, P. N., Bender-deMoll, S. and Morris, M. (2015). statnet: Software Tools for the Statistical Analysis of Network Data The Statnet Project (`http://www.statnet.org`) R package version 2015.6.2.

Hansen, M. H. and Hurwitz, W. N. (1943). On the Theory of Sampling from Finite Populations. *The Annals of Mathematical Statistics* **14** 333–362. MR0009832

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. MR3363437

Heckathorn, D. D. (1997). Respondent-Driven Sampling: A New Approach to the Study of Hidden Populations. *Social Problems* **44** 174–199.

Hunter, D. R., Goodreau, S. M. and Handcock, M. S. (2008). Goodness of Fit for Social Network Models. *Journal of the American Statistical Association* **103** 248–258. MR2394635

Hunter, D. R. and Handcock, M. S. (2006). Inference in Curved Exponential Family Models for Networks. *Journal of Computational and Graphical Statistics* **15** 565–583. MR2291264

Iguchi, M. Y., Ober, A. J., Berry, S. H., Fain, T., Heckathorn, D. D., Gorbach, P. M., Heimer, R., Kozlov, A., Ouellet, L. J., Shoptaw, S. and Zule, W. A. (2009). Simultaneous Recruitment of Drug Users and Men Who Have Sex with Men in the United States and Russia Using Respondent-Driven Sampling: Sampling Methods and Implications. *Journal of Urban Health* **86** 5–31.

Johnston, L. G., Malekinejad, M., Kendall, C., Iuppa, I. M. and Rutherford, G. W. (2008). Implementation Challenges to Using Respondent-Driven Sampling Methodology for HIV Biological and Behavioral Surveillance: Field Experiences in International Settings. *AIDS and Behavior* **12** 131–141.

Kandel, D. B. (1978). Homophily, Selection, and Socialization in Adolescent Friendships. *American Journal of Sociology* **84** 427–436.

Lansky, A., Abdul-Quader, A. S., Cribbin, M., Hall, T., Finlayson, T. J., Garfein, R. S., Lin, L. S. and Sullivan, P. S. (2007). Developing an HIV behavioral surveillance system for injecting drug users: the National HIV Behavioral Surveillance System. *Public Health Reports* 48–55.

Li, X., Rohe, K. et al. (2017). Central limit theorems for network driven sampling. *Electronic Journal of Statistics* **11** 4871–4895. MR3733297

Liu, H., Feng, T., Liu, H., Feng, H., Cai, Y., Rhodes, A. G. and

Grusky, O. (2009). Egocentric Networks of Chinese Men Who Have Sex with Men: Network Components, Condom Use Norms, and Safer Sex. *AIDS Patient Care and STDs* **23** 885–893.

Liu, H., Li, J., Ha, T. and Li, J. (2012). Assessment of Random Recruitment Assumption in Respondent-Driven Sampling in Egocentric Network Data. *Social networking* **1** 13–21.

Lu, X. (2013). Linked Ego Networks: Improving estimate reliability and validity with respondent-driven sampling. *Social 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. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **175** 191–216. MR2873802

Malekinejad, M., Johnston, L. G., Kendall, C., Kerr, L., Rifkin, M. and Rutherford, G. (2008). Using Respondent-Driven Sampling Methodology for HIV Biological and Behavioral Surveillance in International Settings: A Systematic Review. *AIDS and Behavior* **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 (Cambridge, Mass.)* **23** 138–47.

McLaughlin, K. R. (2016). Modeling Preferential Recruitment for Respondent-Driven Sampling, PhD in Statistics, University of California, Los Angeles. MR3527298

McPherson, M., Smith-Lovin, L. and Cook, J. M. (2001). Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology* **27** 415–444.

Montealegre, J., Johnston, L. G., Murrill, C. and Monterroso, E. (2013). Respondent Driven Sampling for HIV Biological and Behavioral Surveillance in Latin America and the Caribbean. *AIDS and Behavior* **17** 2313–2340.

Rohe, K. (2015). Network driven sampling; a critical threshold for design effects. *ArXiv preprint*. arXiv:1505.05461. MR3909942

Ross, S. M. (2014). *Introduction to probability models*. Academic press. MR0328973

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

Salganik, M. J. and Heckathorn, D. D. (2004). Sampling and estimation in hidden populations using respondent-drive sampling. *Sociological Methodology* **34** 193–239.

Shi, Y., Cameron, C. J. and Heckathorn, D. D. (2019). Model-based and design-based inference: reducing bias due to differential recruitment in respondent-driven sampling. *Sociological Methods & Research* **48** 3–33. MR3903764

R Core Team (2018). R: A Language and Environment for Statistical Computing R Foundation for Statistical Computing, Vienna, Austria.

Tomas, A. and Gile, K. J. (2011). The Effect of Differential Recruitment, Non-response and Non-recruitment on Estimators for Respondent-Driven Sampling. *Electronic Journal of Statistics* **5** 899–934. MR2831520

UNAIDS (2014). The gap report.

Verdery, A. M., Merli, M. G., Moody, J., Smith, J. A. and Fisher, J. C. (2015). Respondent-driven sampling estimators under real and theoretical recruitment conditions of female sex workers in China. *Epidemiology* **26** 661–665.

Volz, E. and Heckathorn, D. D. (2008). Probability based estimation theory for Respondent Driven Sampling. *The Journal of Official Statistics* **24** 79–97.

Wang, J., Carlson, R. G., Falck, R. S., Siegal, H. A., Rahman, A. and Li, L. (2005). Respondent-driven sampling to recruit MDMA users: a methodological assessment. *Drug and Alcohol Dependence* **78** 147–157.

Wejnert, C. and Heckathorn, D. D. (2008). Web-Based Network Sampling: Efficiency and Efficacy of Respondent-Driven Sampling for Online Research. *Sociological Methods and Research* **37** 105–134. MR2516739

Yamanis, T. J., Merli, M. G., Neely, W. W., Tian, F. F., Moody, J., Tu, X. and Gao, E. (2013). An Empirical Analysis of the Impact of Recruitment Patterns on RDS Estimates Among a Socially Ordered Population of Female Sex Workers in China. *Sociological Methods & Research* **42** 392–425. MR3190735