Respondent-driven sampling bias induced by community structure and response rates in social networks

Luis E. C. Rocha,

Karolinska Institutet, Stockholm, Sweden, and Université de Namur, Belgium

Anna E. Thorson,

Karolinska Institutet, Stockholm, Sweden

Renaud Lambiotte

Université de Namur, Belgium

and Fredrik Liljeros

Stockholm University, Sweden

[Final revision October 2015]

Summary. Sampling hidden populations is particularly challenging by using standard sampling methods mainly because of the lack of a sampling frame. Respondent-driven sampling is an alternative methodology that exploits the social contacts between peers to reach and weight individuals in these hard-to-reach populations. It is a snowball sampling procedure where the weight of the respondents is adjusted for the likelihood of being sampled due to differences in the number of contacts. The structure of the social contacts thus regulates the process by constraining the sampling within subregions of the network. We study the bias induced by network communities, which are groups of individuals more connected between themselves than with individuals in other groups, in the respondent-driven sampling estimator. We simulate different structures and response rates to reproduce real settings. We find that the prevalence of the estimated variable is associated with the size of the network community to which the individual belongs and observe that low degree nodes may be undersampled if the sample and the network are of similar size. We also find that respondent-driven sampling estimators perform well if response rates are relatively large and the community structure is weak, whereas low response rates typically generate strong biases irrespectively of the community structure.

Keywords: Complex networks; Network sampling; Public health; Respondent-driven sampling bias

1. Introduction

To estimate the prevalence of diseases, traits or behaviours in particular social groups or even in the entire society, researchers typically rely on samples of the target population. A carefully selected sample may generate satisfactorily low standard errors with a bonus of optimizing research resources and time. A common challenge is to obtain a significant and unbiased sample

Address for correspondence: Luis E. C. Rocha, Department of Public Health Sciences, Karolinska Institutet, Tomtebodavägen 18A, Stockholm 17177, Sweden.
E-mail: luis.rocha@ki.se

of the target population. This is particularly difficult if this population of interest is somehow segregated, stigmatized, or in some other way difficult to reach such that a sampling frame cannot be well defined. These so-called hidden (or hard-to-reach) populations may be for example men who have sex with men, sex-workers, injecting drug users, criminals, homeless or minority groups (Sudman *et al.*, 1988; Magnania *et al.*, 2005).

In 1997, Heckathorn introduced a new methodology to sample hidden populations named respondent-driven sampling (RDS) (Heckathorn, 1997). RDS exploits the underlying social network structure to reach the target population through the participants' own peers. The method consists in a variation of snowball sampling where the statistical estimators have weights to compensate the non-random nature of the recruiting process, i.e. that individuals with many potential recruiters have a higher chance of being sampled. In RDS, researchers select seeds to start the recruitment. A seed person then invites a number of other individuals to participate in the survey by passing a coupon to them. Those who are successfully recruited respond to a survey and receive new coupons to invite a number of other individuals within their own social network, and the process is repeated until enough participants have been recruited. Successful recruitment and participation in the survey are both financially rewarded. A fundamental assumption is that each participant knows the number of his or her own acquaintances in the target population or, in network jargon, his or her own degree. This information is used as weights to estimate the prevalence of the variable of interest in the study population.

The most popular RDS statistical estimator is due to Volz and Heckathorn, who devised a Markov process whose equilibrium distribution is the same as that of the target population (Volz and Heckathorn, 2008). This estimator is derived after a series of assumptions regarding both the underlying network structure and the recruitment process *per se*. The assumptions are generally reasonable but sometimes relatively strict for a realistic setting, e.g. the uniformly random selection of peers, persistent successful recruitment and sampling with replacement (Semaan, 2010). These and other assumptions have been scrutinized in previous theoretical studies and the estimator has performed satisfactorily in different scenarios using both synthetic (Abdul-Quader *et al.*, 2006; Salganik, 2006; Gile and Handcock, 2010) and real networks (Lu *et al.*, 2012; Verdery *et al.*, 2014). A number of real life studies have also concluded that RDS is an effective sampling method for various categories of hidden populations (see for example McKnight *et al.* (2006), Robinson *et al.* (2006), Abdul-Quader *et al.* (2006), Abramovitz *et al.* (2009) and Iguchi *et al.* (2009)).

Social networks are, however, highly heterogeneous in the sense that the structure of connections cannot be represented by characteristic values, such as for the number of contacts per

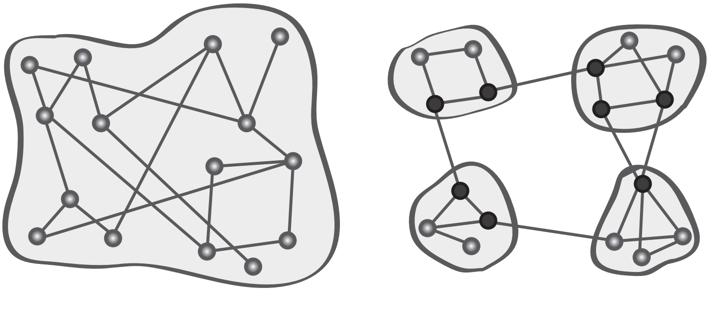


Fig. 1. Schematic network (a) without communities and (b) with four network communities: ●, bridging nodes

individual (Newman, 2010; Costa *et al.*, 2011). Since the RDS dynamics are constrained by the network structure (Fig. 1(a)), we may expect that different patterns of connectivity affect the recruitment chains. For example, the network structure may be such that a recruitment tree grows only in one part of the network (Martin *et al.*, 2003; Burt *et al.*, 2010; McCreesh *et al.*, 2011). In realistic settings using sampling without replacement, even if all individuals are willing to participate, trees may simply die out because a network has been locally exhausted and bridging nodes block further propagation of coupons to other parts of the network (Johnston *et al.*, 2013). Such a situation is not unlikely in highly clustered subpopulations where coupons may simply move around the same group of people. Previous theoretical studies have addressed some of these network constraints by studying the RDS performance on either synthetic structures (Salganik, 2006; Gile and Handcock, 2010) or samples of real networks (Lu *et al.*, 2012; Verdery *et al.*, 2014). Each approach to model social networks has its own advantages and limitations. On one hand, simple synthetic structures and sampling processes are unrealistic but allow some mathematical tractability and thus intuitive understanding. On the other hand, samples of real networks may suffer biases themselves due to their own sampling and thus potential incompleteness of data (Lee *et al.*, 2006; Latapy and Magnien, 2008).

Network clustering is particularly important in the context of social networks. Clustering may refer to various connectivity patterns but here we associate it with network communities, which are groups of individuals who are more connected between themselves than with individuals in other groups (Fig. 1(b)). This means that one may find hidden subpopulations within the study population. Examples include social groups with particular features (e.g. wealth, foreigners and ethnic minorities) embedded in the target population (Johnston *et al.*, 2013), transsexuals in populations of men who have sex with men or geographically sparse populations (Burt *et al.*, 2010). Although these subpopulations may potentially be removed by defining a more strict sampling frame, social groups (or network communities) are inherent in social and other human contact networks irrespectively of the observation scale (Wasserman and Faust, 1994; Costa *et al.*, 2011). In practice, however, the details of the social network structure may be difficult to know. Network clustering is not the same as homophily, i.e. the tendency of similar individuals to associate, but one may enhance the other. For example, individuals may share social contacts because they live geographically close, share workplaces or are structured in organizations (potentially leading to network clustering) but may be completely different in other aspects (low homophily in wealth, health status, gender, infection status and so on).

In this paper, we use computational algorithms to generate synthetic networks with network communities of various sizes, aiming to reproduce structures that are observed in real social networks. Using realistic parameters, we simulate an RDS process and quantify the performance of the RDS estimator in different scenarios of the prevalence of an arbitrary variable of interest. The paper is organized such that we first analyse how the community structure affects the RDS when it comes to size of transmission trees and generation of recruitment. Then we investigate the RDS II estimator as a function of different willingness to participate (response rates). We also test scenarios where the variable under study is correlated with the number of contacts per individual and the size of the network community. Finally, we study the consequences of the biased selection of seeds and the bias that is induced by network structure in samples of real social networks.

2. Materials and methods

2.1. Study networks

A social network is defined by a set of nodes representing the population and a set of links representing the social contacts, such as acquaintance or friendship, between two individuals. The network structure can be characterized by different network quantities (Wasserman and Faust, 1994; Newman, 2010). The most fundamental is the degree k that represents the number of links of a node or equivalently the number of contacts of an individual. The network community is defined as a group of nodes (i.e. the individuals) that are more connected between themselves than with nodes (individuals) in other groups. A fundamental property of the community structure is that only a few nodes link (or bridge) different communities. These nodes are also known as bottlenecks because they constrain the diffusion of the sampling process. If there are only a few bridging nodes, we say that the community structure is strong, whereas many bridging nodes weaken the community structure. A typical example is someone who works part time in two companies or has friends or family in two villages. This person is essential if we want a recruitment chain to move beyond one of the locations.

2.1.1. Synthetic networks

We use a computational algorithm that can generate synthetic networks with tunable community structure (Lancichinetti and Fortunato, 2009). This algorithm is not expected to reproduce a particular social network but to generate structures that are observed in social networks more realistically than those in previous studies (Salganik, 2006). We start by choosing the distribution of degrees and community sizes. We choose $P(k) \propto k^{-2.5}$, which is not expected to be the most appropriate distribution of contacts in real populations but captures the degree of heterogeneity that is typically observed in social groups (Newman, 2010; Costa *et al.*, 2011; Lu *et al.*, 2012). In fact, if no or very small costs are associated with keeping links alive, power law functions are reasonable models for empirical distributions; otherwise we usually observe broad scale distributions that are not necessarily power law like. The heterogeneity means that the majority of nodes have only a few contacts whereas a small number of them have several. We then choose $P(C) \propto C^{-1.0}$ to reproduce the heterogeneity in the size C of the communities (Newman, 2010; Costa *et al.*, 2011). We limit the sizes between 10 and 1000 nodes to guarantee that a sufficient number of communities are large and enough small sized communities are represented. Values of the exponent that are smaller than -1 would result in relatively more small sized communities. These exponents are also constrained by the number of links and the level of overlapping of communities (see below), and thus are chosen to generate a network with a single connected component. Overlapping means that a number of nodes belong to more than one community (these are the bridging nodes, e.g. the black nodes in Fig. 1(b)) whereas the rest of the nodes belong only to single communities. In this algorithm, one may further select a mixing parameter μ to add random links between the bridge nodes and randomly chosen communities to weaken the community structure. Therefore, small overlapping and small mixing generate stronger community structures. We set $\mu = 0$ and select 100 overlapping nodes in five communities to generate strong community structure. For weak community structure, we set $\mu = 0.3$, and 1000 overlapping nodes in five communities as well. Using the same set of parameters, we generate 10 realizations of each network with 10000 nodes (the size of the target or study population). In the reference network, all nodes simply belong to the same community.

2.1.2. Empirical networks

We use five samples of real life contact networks representing different forms of human social relationships. Three data sets correspond to e-mail communication, two between members of two distinct universities in Europe (sets EMA1 (Guimera *et al.*, 2003) and EMA2 (Eckmann *et al.*, 2004)) and one between employees of a company (set ENR) (Leskovec, 2014)). In these

Table 1. Summary statistics of the empirical networks used in this study

	Results for the following data sets:				
	EMA1	*ADH*	*EMA2*	*POK*	*ENR*
Number of nodes	1133	2539	3186	28295	36692
Number of links	5451	10455	31856	115335	183831
Number of communities†	57	200	71	2615	2441
Size of smallest community	2	1	1	1	1
Size of largest community	151	222	1205	2621	1481

†According to the MapEquation algorithm (Rosvall *et al.*, 2008).

data sets, nodes correspond to people and social ties are formed between those who have sent or received at least one e-mail during a given time interval. One data set corresponds to friendship ties between US high school students (set ADH) (Moody, 2001). The last data set corresponds to on-line communication between members of an on-line dating site (set POK) (Holme *et al.*, 2004). Similarly to the e-mail networks, if two members have exchanged an on-line message, a link is made between the respective nodes. Although some of these data sets do not correspond to social networks in which RDS would take place, they serve as realistic settings capturing the network structure of actual social relationships. We have selected data sets with diverse sizes and network structure to cover various contexts and configurations (Table 1).

2.2. *Respondent-driven sampling model*

We simulate the sampling by using a stochastic process reproducing several features of realistic RDS dynamics. Our model further adds a continuous time framework and the response rate can be controlled. We use similar parameters to those typically used in the literature (Lu *et al.*, 2012; Gile *et al.*, 2015). We start by uniformly selecting (unless otherwise stated) 10 random nodes as seeds for the recruitment. We assume that the network structure is known and that the recruitment follows a Poisson process, which generates exponential waiting times. Therefore, after a time t, sampled from an exponential distribution, each seed chooses uniformly three of its contacts and passes one coupon to each of them. The average waiting time is set to 5, meaning that a node waits on average five time steps (e.g. 5 days) before inviting its contacts. After waiting t time steps, and with probability p, which represents the probability of participation (or response rate, i.e. 1 minus the probability of not returning a coupon), each of these contacts recruits three of their own contacts that have not been invited yet (i.e. sampling without replacement). A node that agrees to participate is included in the sample. We assume that, if a node refuses to participate once, it becomes unavailable for recruitment by others. The process continues until all possibilities of new recruitments are exhausted or when a specific sample size has been reached. Note that this continuous time model is equivalent to a discrete time model in which randomly chosen nodes update their status sequentially.

2.3. *Prevalence of the study variable*

In RDS studies, we are interested in quantifying the prevalence of some variable A in the target population. This variable may represent, for instance, being tested positively for a given disease, being male or female, the ethnicity or having a particular physical trait. In this paper, to simplify the notation, we say that an individual and its respective node are infected with

A or not infected with A. Assuming that the network structure is known, which is unlikely in practice, we use different protocols to infect a fraction of 25% of the network nodes with A. The remaining nodes are thus assumed to be non-infected.

The reference case (case RI) corresponds to uniformly selecting the nodes within the target population, i.e. the infection A is uniformly distributed in the network. The preferential case (case PI) corresponds to selecting nodes in decreasing order of degree. We start at nodes with the highest degree and infect them until 25% of the nodes receive A. To add noise (case PRI), we select 20% of the infected nodes, cure them and redistribute these infections uniformly such that the total number of infected nodes remains fixed. The other two cases consist of infecting nodes according to the community structure. In the first case (case SI), we initially infect nodes in the smallest communities until 25% of them become infected. In the second case (case BI), we infect nodes in the largest communities until the same fraction of 25% become infected. To reduce homophily, we add noise by selecting 40% of the infected nodes and redistributing these infections as done in case PRI (these configurations are named SRI for small and BRI for large communities).

2.4. *Statistics*

To analyse the recruitment trees, we measure the total number of participants Ω (i.e. the sample size), and the size S_i and the number of generations (or waves) W_i of each recruitment tree, starting from a seed node i. The proportion of individuals in the population with A ($\hat{P}_A$) is estimated by using the RDS II estimator (Volz and Heckathorn, 2008):

$$\hat{P}_A = \frac{\sum_{i \in A \cap N} k_i^{-1}}{\sum_{i \in N} k_i^{-1}}, \tag{1}$$

where k_i is the reported degree of an individual i in the social network. We thus define

$$\theta = \sum_{j=1}^{m} \frac{\hat{P}_A^j}{m}, \tag{2}$$

as the average estimate of the prevalence of A for m simulations with the same set of parameters, with standard deviation given by σ. Complementary, we define the average bias δ, i.e. the difference between the estimate of the prevalence of A and the true prevalence P_A of A, for m simulations, as

$$\delta = \sum_{j=1}^{m} \frac{|\hat{P}_A^j - P_A|}{m}, \tag{3}$$

In the results, for convenience, we show the relative bias with respect to the true value of the prevalence, i.e. $\Delta = \delta/0.25$. The design effect DE (Lohr, 2009) is defined as

$$\mathrm{DE} = \frac{\mathrm{var}(\hat{P}_A)_{\mathrm{RDS}}}{\mathrm{var}(\hat{P}_A)_{\mathrm{SRS}}}, \tag{4}$$

where $\mathrm{var}(\hat{P}_A)_{\mathrm{RDS}}$ is the variance of the estimator $\hat{P}_A$ by using RDS and $\mathrm{var}(\hat{P}_A)_{\mathrm{SRS}}$ is the variance of the same estimator $\hat{P}_A$ by using simple uniform sampling, i.e. the same number of nodes (as in the RDS sample) is uniformly selected in the study population. The design effect thus measures the number of the sample cases that is necessary to obtain the same statistics as if a simple random sample was used. In our study, $m = 500$ (50 RDS simulations for each of the 10 generated networks with fixed parameters, and 500 RDS simulations for each of the empirical networks).

3. Results

3.1. Recruitment trees

We first look at the statistics of the recruitment trees in the case that the entire target population can potentially be recruited, i.e. the recruitment only stops if no new subject or if everyone is recruited. Since the population is fixed to 10 000 individuals, this limiting case provides us with the maximum possible coverage of the sampling for a given configuration of the RDS. In the reference case, only the degree distribution is fixed and no community structure exists. In this case, if every recruited individual responds to the survey, i.e. $p = 1.0$, nearly all the population is recruited (Fig. 2(a)). The recruitment dynamics, however, are not robust to variations in the response rate; for example, if $p = 0.7$, only about 60% of the population is recruited, and this percentage falls to negligible values if $p < 0.4$ (Newman, 2002; Malmros *et al.*, 2014). Successful recruitment occurs only if $p > 0.35$. We also observe a broad distribution in the size of the recruitment trees (Figs 2(b) and 2(c)). There is a relatively high chance for the recruitment trees to break down quickly and thus to contain only a few individuals. This typically happens when a recruitment tree reaches a high degree node. High degree nodes are easily reachable because they have many connections. As soon as the first recruitment tree passes through a high degree node, it becomes unavailable. Consequently, the recruitment trees arriving afterwards simply die out as soon as they reach the same node. However, a few trees persist sufficiently long, potentially sampling large parts of the network.

In contrast, in the case of strong community structure, we observe lower sample sizes; for example, a maximum of about 85% of the population may be recruited if $p = 1.0$ (Fig. 2(f)). This is a result of the bridging nodes connecting the various communities and limiting the sampling trees to explore the network further. These bottlenecks can be removed by making more connections between communities, i.e. weaken the community structure. The results also indicate that the response rate should be higher if networks have strong community structure for the recruitment trees to take off and to gather sufficient participants. A response rate below $p \sim 0.5$ is insufficient to generate large samples in our example. This is a fundamental issue in realistic settings, meaning that highly clustered (or, in other words, highly segregated or marginalized) populations need higher compensation to achieve the same sampling size as we would obtain if studying less segregated groups. We also identify a broad variance in the number of waves (Figs 2(i) and 2(j)), suggesting that seeds sample the network inhomogeneously. This is related to the fact that the communities have different sizes (or number of nodes) and thus the bridging nodes (connecting communities) are reached at different times by different recruitment trees. In the absence of communities, however, these bottlenecks disappear and the trees are more similar (Figs 2(d) and 2(e)). The number of waves is important because few waves may not be sufficient for the stochastic process to forget the initial conditions and thus to reach the stationary state (Klafter and Sokolov, 2011), the condition in which the RDS II estimator is expected to be unbiased (Volz and Heckathorn, 2008).

3.2. Respondent-driven sampling estimates and structure-induced bias

In the reference scenario, i.e. when A is uniformly spread in the network, the estimator θ (equation (2)) performs well but with a substantial standard deviation σ and bias Δ (equation

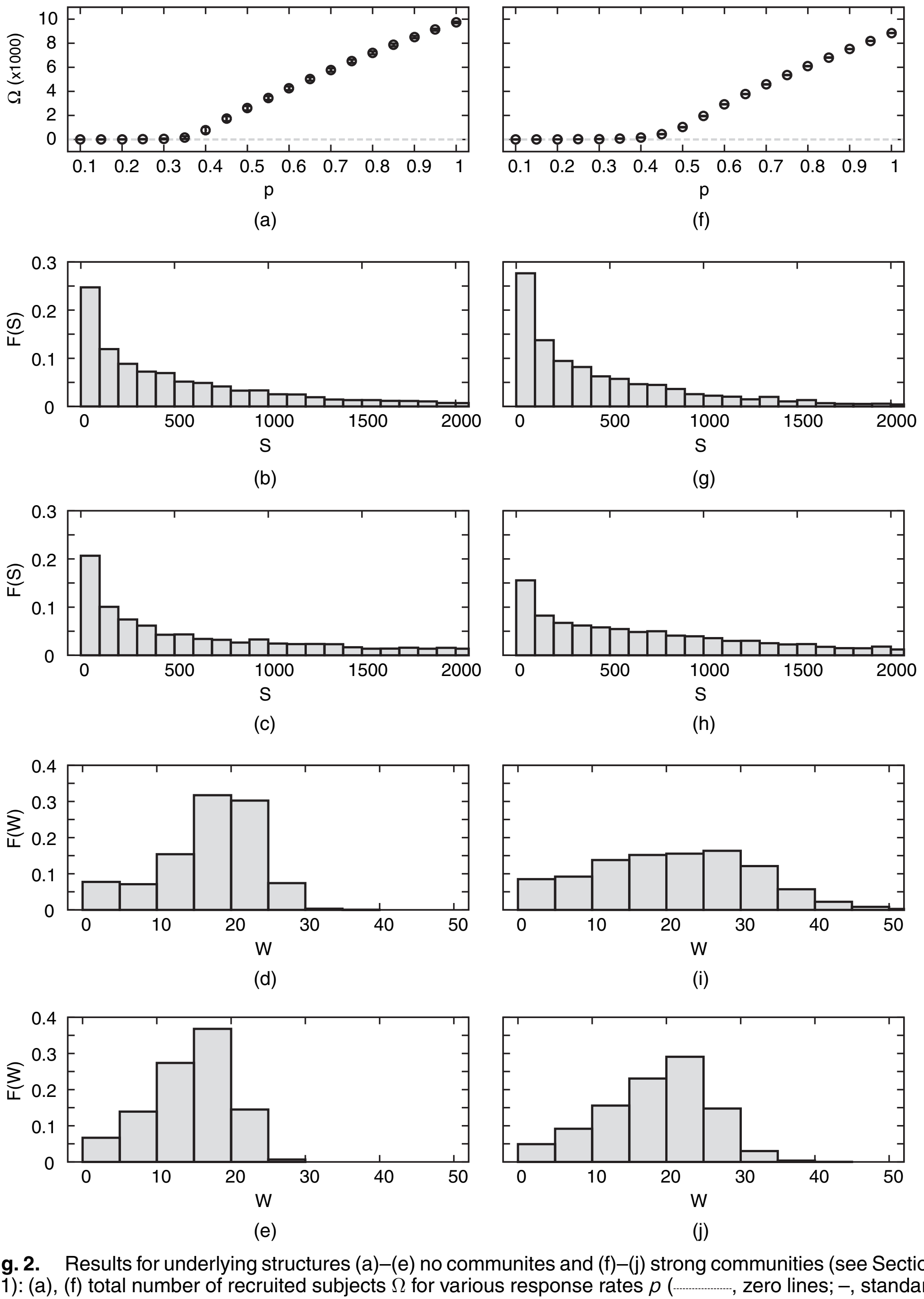


Fig. 2. Results for underlying structures (a)–(e) no communites and (f)–(j) strong communities (see Section 2:1): (a), (f) total number of recruited subjects Ω for various response rates p (………, zero lines; –, standard error); distribution of size of recruitment trees S per seed for response rates (b), (g) $p = 0.7$ and (c), (h) $p = 1.0$ (histogram bin size 100); distribution of number of waves W per seed for response rates (d), (i) $p = 0.7$ and (e), (h) $p = 1.0$ (histogram bin size 5)

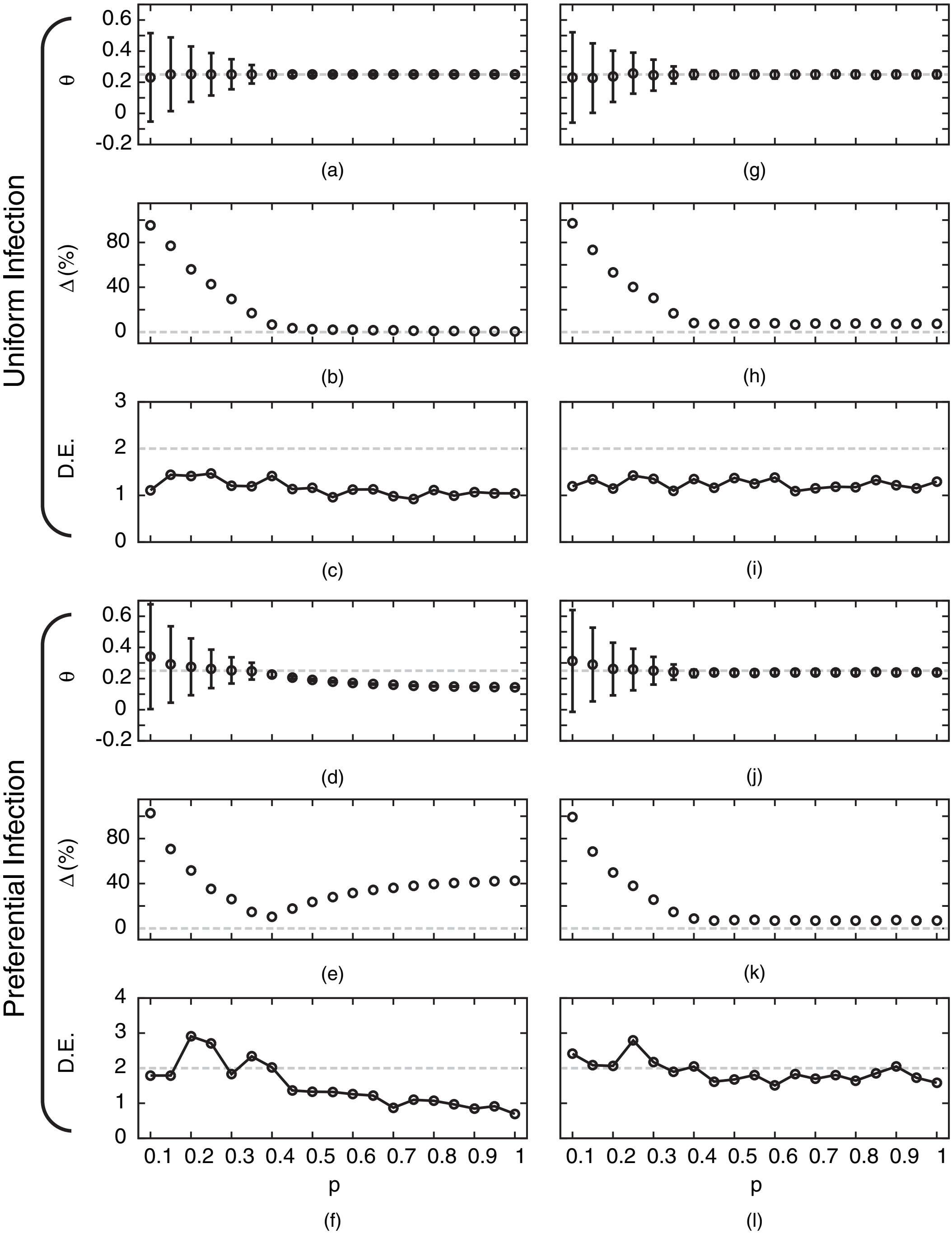


Fig. 3. Results for underlying networks having no community structure and recruitment limited to (a)–(f) 10000 participants and (g)–(l) 500 participants (in all cases, 25% of the population is infected with *A*, either following the protocal RI, i.e. infections are uniformly spread, or the protocal PRI, i.e. infections occur preferentially in high degree nodes (see Section 2.3; - - - - -, guides for the eye): (a), (d), (g), (j) RDS estimator θ (equation (2)) and standard deviation σ (I), (b), (e), (h), (k) average bias Δ (equation 3)) and (c), (f), (i), (l) design effect DE (equation (4)) with respect to the response rate *p*

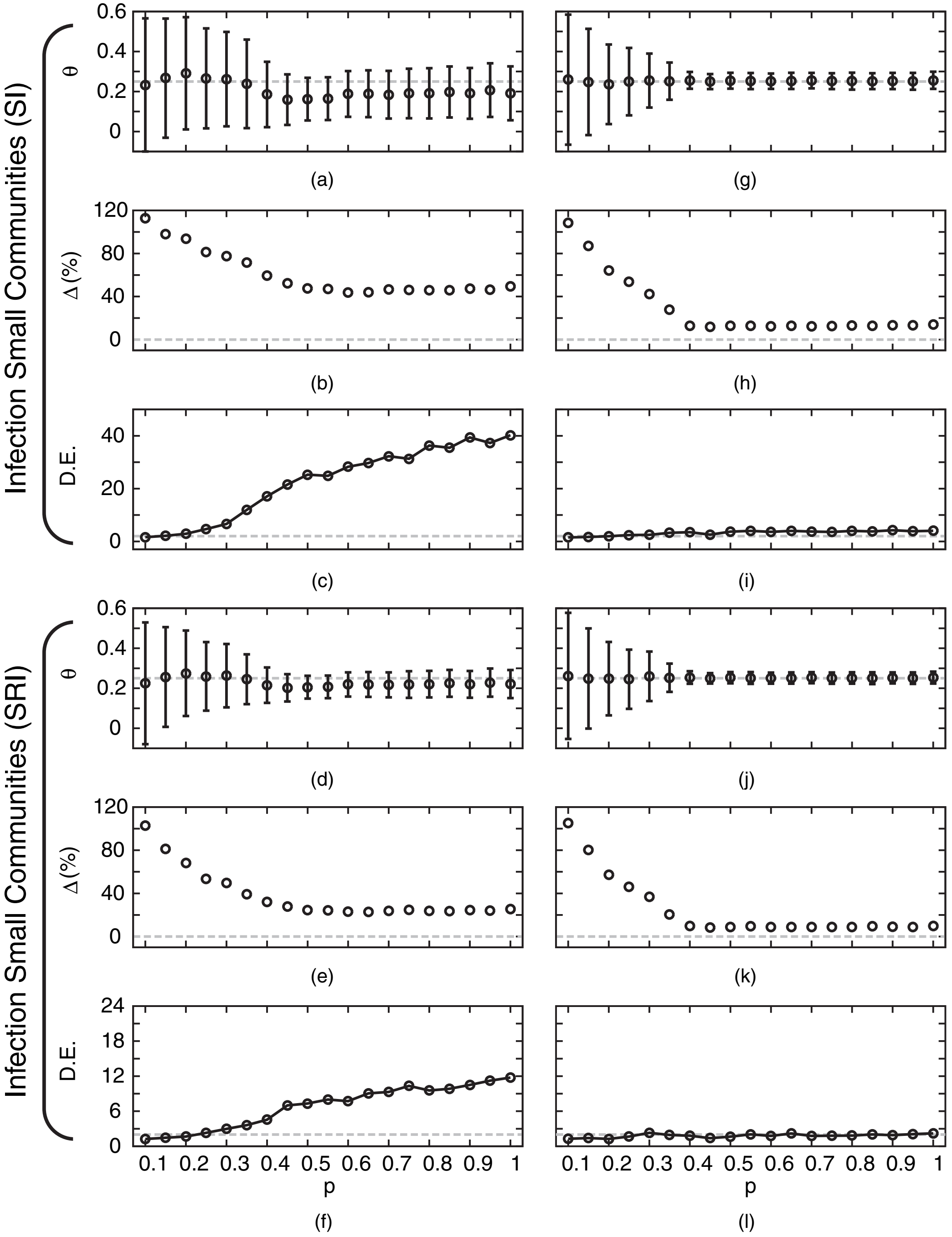


Fig. 4. Results for underlying networks having various levels of community structure (see Section 2.1) and recruitment limited to 500 participants (in all cases, 25% of the population is infected with *A*, following either protocol SI, BI, SRI or BRI (see Section 2.3)): (a)–(f), (m)–(r) strong communities; (g)–(l), (s)–(x) weak communities; (a), (d), (g), (j), (m), (p), (s), (v) RDS estimator θ (equation (2)) and standard deviation σ (I), (b), (e), (h), (k), (n), (q), (t), (w) average bias Δ (equation (3)) and (c), (f), (i), (l), (o), (r), (u), (x) design effect DE (equation (4)) with respect to the response rate p

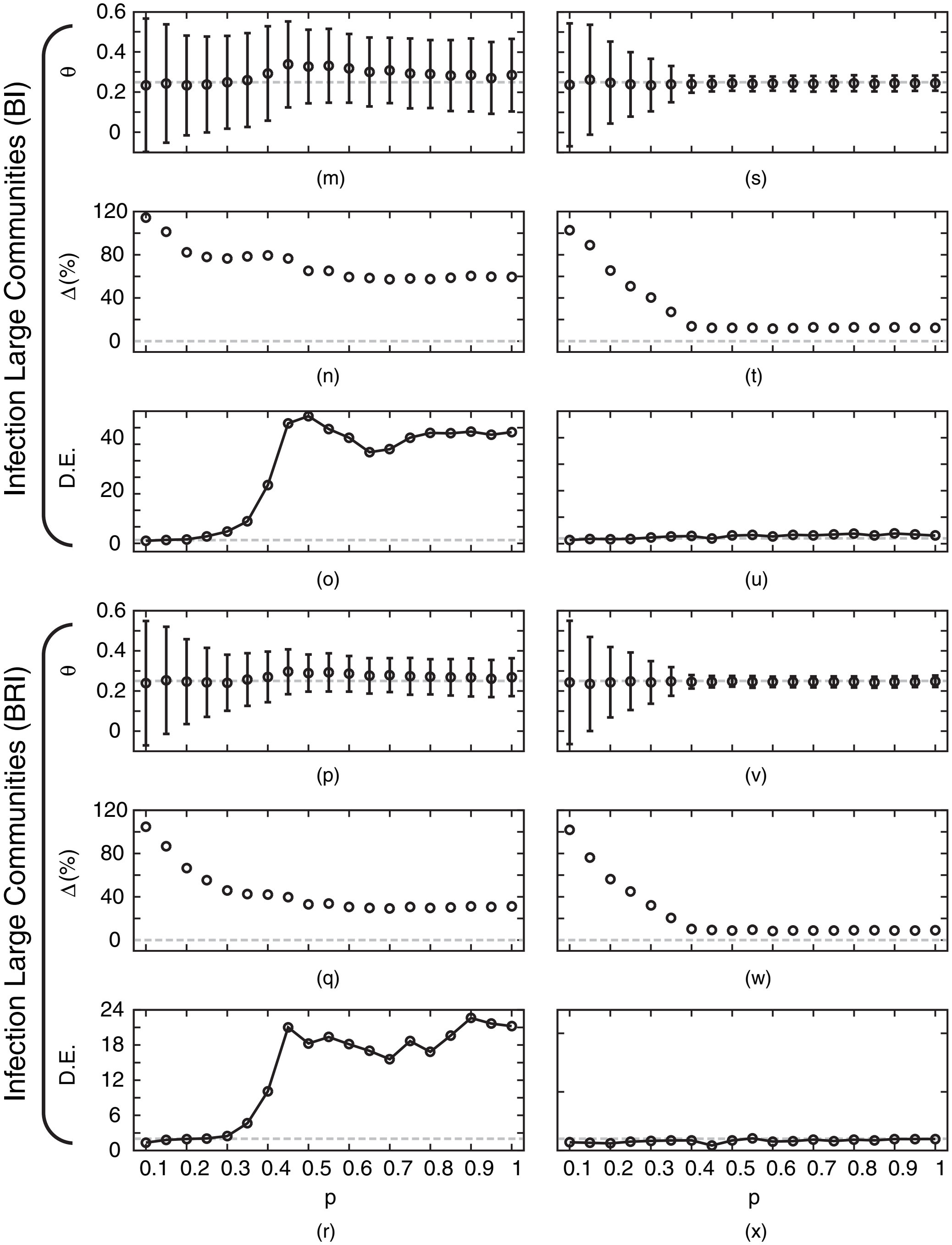


Fig. 4 (*continued*)

(3)) for $p < 0.4$. This deviation occurs independently of the maximum size of the target population (Figs 3(a)–3(c) and 3(g)–3(i)) and is a consequence of the insufficient sample size generated with low response rates. We now test a hypothetical scenario where A is concentrated in high degree nodes (see protocol PRI in Section 2.3). Individuals with a large number of contacts are more likely to acquire an infection or to propagate a piece of information (Newman, 2010). If the target sample size is 10000, A is underestimated for response rates $p > 0.3$ and the precision is worse for $p < 0.35$ (Fig. 3(d)). As in the reference scenario, the poor accuracy for low response rates is a result of the RDS not recruiting sufficient participants. The underestimation of the prevalence for larger response rates, however, indicates low degree nodes are not sufficiently sampled (see equation (1)). In fact, it becomes increasingly more difficult to sample low degree nodes as the sample size grows close to the network size due to the finite size effect. A substantial bias is also observed (Fig. 3(e)) and the design effect varies between 1 and 3 (Fig. 3(f)). If the target sample size is limited to 500 individuals, i.e. 5% of the total population, the performance of the estimator θ and the average bias Δ improves substantially even though A remains slightly underestimated for $p > 0.35$ (Figs 3(j) and 3(k)). This improvement is observed because undersampling of low degree nodes is not observed. These results are in accordance with previous recommendations that the sample size should be much smaller than the size of the target population (Gile and Handcock, 2010) to achieve good estimates by using the RDS II estimator. Some caution, however, should be pointed out since it is not straightforward to know in advance the size of the target population and thus to estimate the optimal sample size with respect to the target population. If too many subjects are recruited, relatively to the size of the target population, saturation occurs and the network structure induces biases in the estimator due to finite size effects (see also the on-line supporting information).

We now simulate scenarios where A is concentrated in specific communities, irrespectively of the degree of the nodes. This is a reasonable assumption since an infection (or other quantity) may affect only the population of some geographical region or, for example, a particular group of injecting drug users among men who have sex with men may be sharing contaminated paraphernalia. We first select 25% of the nodes that are associated with the smallest communities and infect them with A (see Section 2.3). In this setting, the prevalence is underestimated with relatively large deviations for strong community structure (Fig. 4(a)) but estimation improves for weaker community structure (Fig. 4(g)). Even for a weak community structure, the minimum average bias is about 15% (Fig. 4(h)), being at least 45% in the case of strong communities (Fig. 4(b)) and large response rates. The bias becomes substantially larger for lower response rates, as is generally the case. The design effect is also significantly affected by the community structure (Fig. 4(c)). This means, for example, that, for strong communities, to have the same statistics as if a standard simple random sample was performed, the RDS needs up to 40 times the same sample size. Furthermore, if we add noise to reduce homophily, the statistics improve but still significant biases are observed (Figs 4(d)–4(f) and 4(j)–4(l)).

However, we can assume that A is unlikely to occur in small communities because, for example, nodes that are associated with these communities are simply less likely to acquire an infection owing to isolation. Social control is also often higher in small groups. It may therefore be easier to behave in certain ways in larger groups. People who want to or who have particular behaviours or traits may thus decide to move to larger groups (however, if public health interventions or criminal justice sanctions would be applied at the group or community level, one may still prefer not to aid the researchers' estimation). To simulate this hypothetical scenario, we now infect 25% of the nodes in the largest communities (see Section 2.3). Fig. 4(m) shows that A is overestimated for $p > 0.35$ for strong community structure. These estimates improve for weaker communities, also resulting in smaller standard deviations (Fig. 4(s)) for larger response rates. The standard deviations are generally larger in this case in comparison with the case where A is concentrated in the small communities. Similarly to the previous scenarios, the bias and design effect are relatively high for strong community structure (Figs 4(n) and 4(o)), even if homophily is reduced (Figs 4(q) and 4(r)). These results show the key difference between clustering and homophily, and why it is important to distinguish them. In both experiments, the network community structure is the same and homophily is high; however, the effect on the RDS II estimator depends on where homophily occurs in the network (i.e. in large or small communities).

3.3. *Respondent-driven sampling estimates and seed-induced bias*

We have assumed so far that seeds are uniformly chosen within the target population. Although this is a reasonable standard assumption in theoretical studies, it is rarley met in real contexts because of the inherent fact that the study population is hard to reach and seed selection is non-trivial (Wylie and Jolly, 2013). If seeds are selected only between subjects who are associated with small communities (here defined as communities with fewer than 200 members), recruitment trees are generally unable to reach beyond those communities and thus the prevalence is overestimated when the infection is concentrated in the smaller communities (Figs 5(a)–5(c) and 5(g)–5(i)). In contrast, the prevalence is underestimated if the infection is concentrated in the larger communities (Figs 5(d)–5(f) and 5(j)–5(l)). The mismatch in the estimators and associated biases are particularly significant if the community structure is stronger. If we select the seeds in the largest communities (here defined as communities with more than 500 members), recruitment trees tend to stay within the largest communities, which leads to an underestimation of the prevalence and relatively high biases if the infection is mostly prevalent in the small communities (Figs 5(m)–5(o) and 5(s)–5(u)). The prevalence is overestimated, however, if the infection is mostly prevalent in the largest communities (Figs 5(p)–5(r) and 5(v)–5(x)). Note that in these experiments homophily is relatively weak since we use protocols SRI and BRI. These resuts are in remarkable contrast with the scenarios when seeds are uniformly sampled (Fig. 4).

3.4. *Respondent-driven sampling estimates on empirical networks*

We have studied the effect of the community structure in RDS estimates in contact networks generated by using theoretical models. Although the algorithm that was used to generate the synthetic networks includes several properties of real life networks, empirical networks, with their own sampling and scope limitations, contain correlations that may be challenging to reproduce theoretically and go beyond community structure or degree heterogeneity. We now analyse the RDS performance by using real life human contact networks to be able to extend the conclusions to more realistic scenarios. Previous studies report that, in real settings, response rates may vary between 0.3 (for female sex workers) and 0.7 (for men who have sex with men), with mean and median at about 0.5 (Gile *et al.*, 2015). We thus study three scenarios for the response rates $p = 0.4, 0.5, 0.6$. We find that the RDS II estimator has a different performance according to the infection protocol, response rate and contact network (Fig. 6). In general, we observe large standard deviations, and average biases that are larger than 10% in most experiments. Biases are typically larger for $p = 0.4$. In contrast, the design effect is generally somewhere between 1 and 3 (except for date set EMA2), a result that is in line with previous suggestions that 2 may be used as a general guideline on unknown populations (Salganik, 2006). A detailed comparative analysis between these networks is beyond the scope of this study. Nevertheless, we see that, if the infection is concentrated in

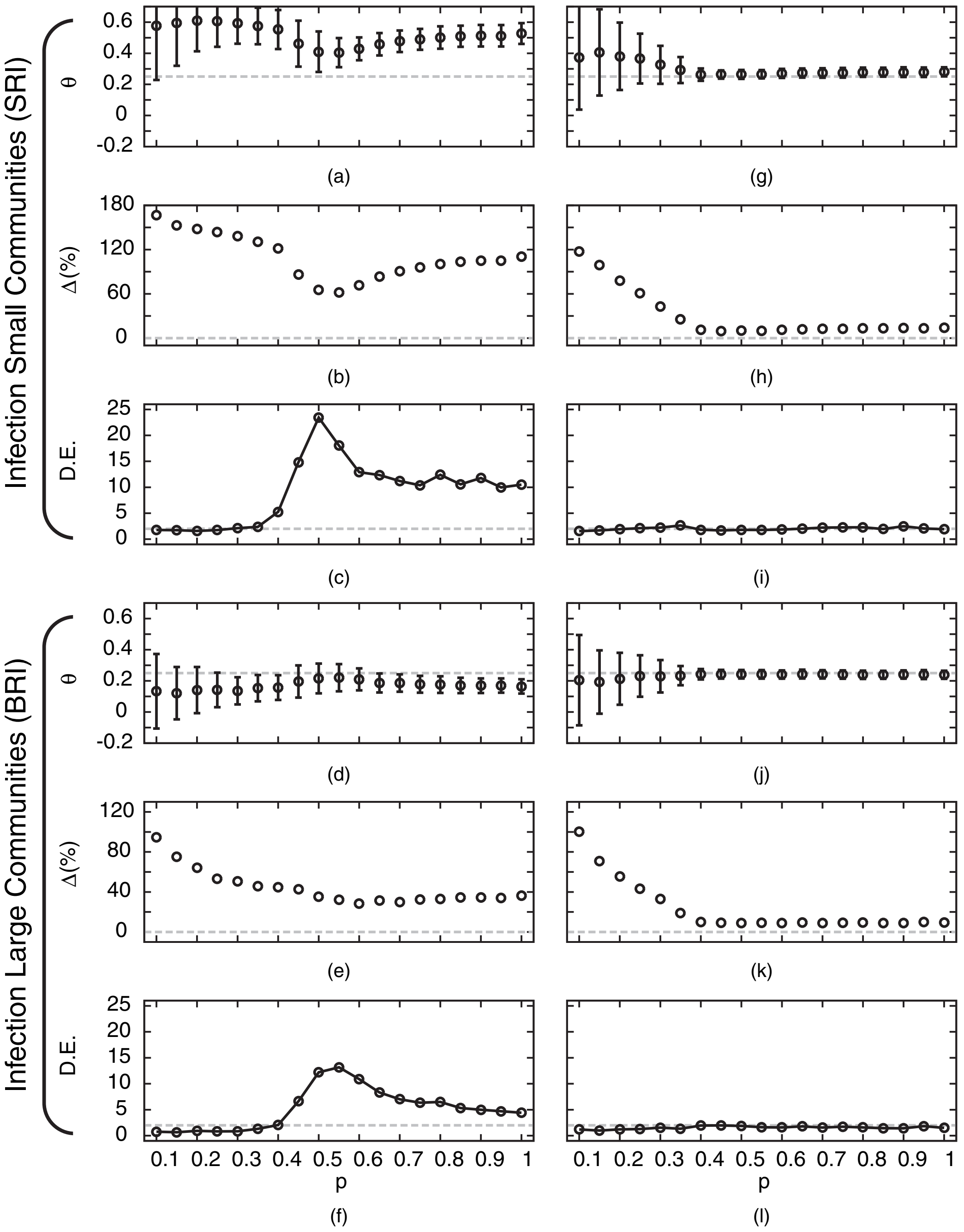
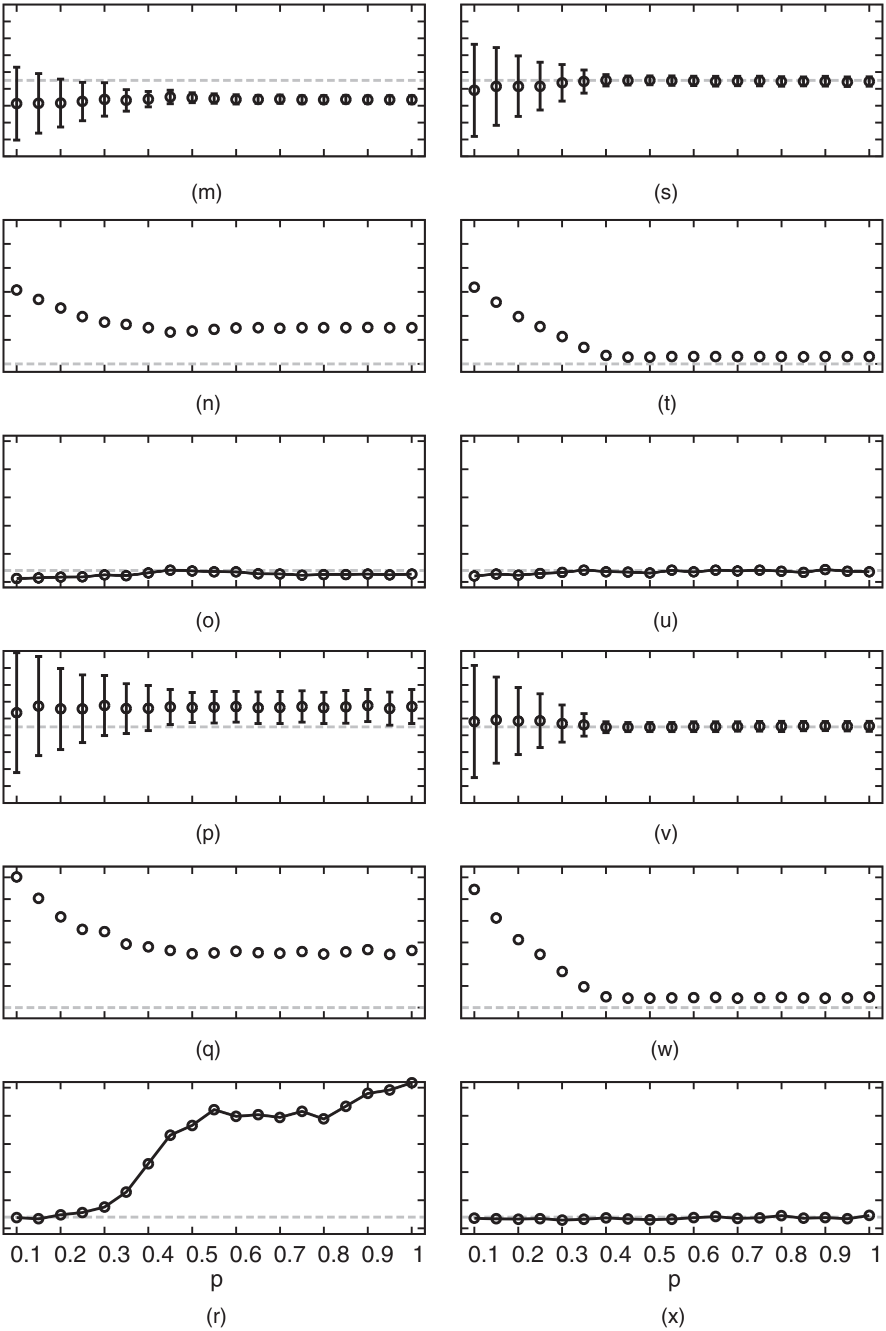


Fig. 5. Results for underlying networks having various levels of community structure (in all cases, 25% of the population is infected with A): (a)–(f) strong communities and seeds selected inside the smallest communities; (g)–(l) weak communities and seeds selected inside the smallest communities; (m)–(r) strong communities and seeds selected inside the largest communities; (s)–(x) weak communities and seeds selected inside the largest communities; (a), (d), (g), (j), (m), (p), (s), (v) RDS estimator θ (equation (2)) and standard deviation σ, (I), (b), (e), (h), (k), (n), (q), (t), (w) average bias Δ (equation (3)) and (c), (f), (i), (l), (o), (r), (u), (x) design effect DE (equation (4)) with respect to the response rate p



(m) (s)

(n) (t)

(o) (u)

(p) (v)

(q) (w)

(r) (x)

Fig. 5 (*continued*)

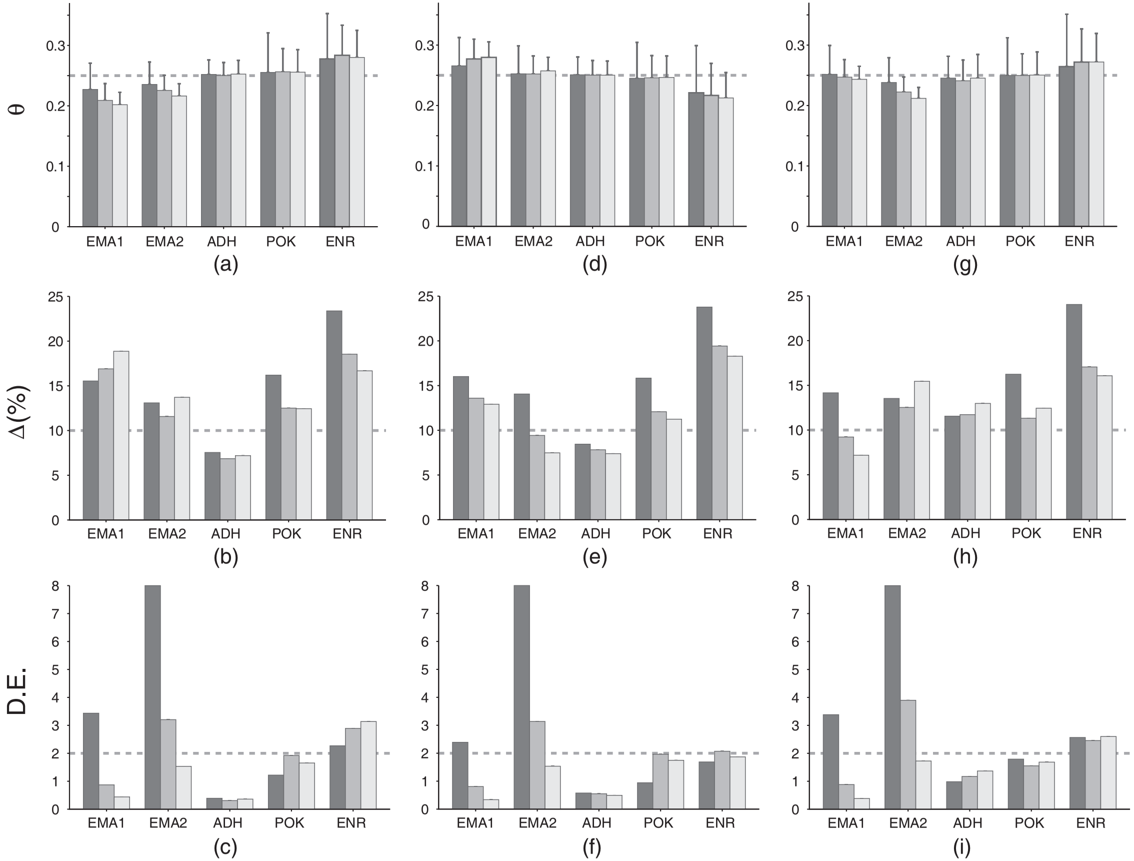


Fig. 6. Results for contact networks gathered empirically and corresponding to various types of social relationships (see Section 2.1) and recruitment related to 500 participants (the response rate p covers realistic values 0.4 (■), 0.5 (■) and 0.6 (■)): in all cases, 25% of the population is infected with A, preferentially choosing (a)–(i) the high degree nodes (PRI), or the nodes in (d)–(f) the largest (BRI) or (g)–(i) the smallest communities (SRI) (see Section 2.3); (a), (d), (g) RDS estimator θ (equation (2)) and standard deviation σ (I), (b), (e), (h) average bias Δ (equation (3)) and (c), (f), (i) design effect DE (equation (4))

the high degree nodes, RDS underestimates the prevalence in both data set EMA1 and data set EMA2, possibly because there is a relatively lower number of communities in these networks and the sample size is relatively large with respect to the size of the network (Table 1), similarly to results reported in Fig. 3(d). The overestimation in the case of data set ENR suggests that high degree nodes are more connected between themselves, and thus more often visited, than with low degree nodes. The other results indicate that sampling trees tend more often to exploit larger communities in the case of sets EMA1 and EMA2 whereas smaller communities are favoured in the ENR case. This is possibly because the largest communities correspond to a larger fraction of the network in the former cases whereas the ENR network has several relatively smaller large communities (Table 1).

4. Discussion

RDS has been proposed as an effective methodology to estimate the prevalence of variables of interest in hard-to-reach populations. The approach exploits information on the social contacts for both recruitment and weighting to generate accurate estimates of the prevalence. Social networks, however, are not random but contain patterns of connectivity that may constrain the cascade of sampling. In this paper, we have studied the bias that is induced by community structure in RDS by using both synthetic and empirical network data. We have also analysed the effect of the response rate. Altogether, we have identified that community structure generates significant biases in the RDS II estimator irrespectively of the response rate. The estimator performs relatively well if the response rate is sufficiently high, the community structure is weak and the prevalence of the variable of interest is not much concentrated in some parts of the network (low homophily). This is a result of the high heterogeneity of the network communities that constrains the sampling trees to certain parts of the network. Some parts of the network may only be accessed through key individuals who bridge the small well-hidden subgroups with the rest of the population. If these bridging nodes are not willing to participate in the recruitment or once they have been recruited, recruitment trees become trapped within a group of nodes, oversample them and as a consequence generate biases in the estimators. The structure of empirical networks may vary in different contexts. Consequently, the expected biases may be also lower or higher for specific social networks. In particular, biases should increase for sparser networks because fewer paths connect the nodes (fewer bridging nodes) and thus the recruitment becomes more sensitive to lower response rates. In contrast, lower biases are expected in denser networks since more redundant paths exist between different parts of the network. The number of network communities and the distribution of community sizes may be also different from those which we consider. Many small communities have a significant effect in the sampling, increasing the biases, because they rely on the existence of many bridging nodes and higher chances to divert or break down the recruitment. In this study, we have adopted the sampling-with-replacement framework, i.e. an invited person who refuses to participate may be invited again. By contrast, one may consider sampling without replacement, i.e. a person who refuses to participate will not be invited again. This comparison is important because in real settings a varying participation rate may be observed after several invitations of the same person. We have performed the same analysis by using both frameworks (results for sampling without replacement are not shown) and observed that this feature has little effect on the estimators that remain qualitatively similar. The major effect is on the size of the sample, particularly for intermediate values of response rates, where the number of recruited individuals is significantly lower for sampling without replacement. These discrepancies may increase for sparser networks but we have not studied this issue in detail.

To understand the effect of the participation probability p, we may consider the simple case where a single coupon is exchanged between individuals and the sampling is done with replacement (Volz and Heckathorn, 2008). In that case, the stochastic process is equivalent to a random-walk process if $p=1$. The probability p_i of finding a coupon with person i is driven by the rate equation

$$\dot{p}_i = \sum_j \frac{A_{ij}}{k_j} p_j - p_i, \tag{5}$$

where A_{ij} is the adjacency matrix of the social network (Newman, 2010). In the case of undirected and unweighted networks, where each link is reciprocated and carries the same importance, the element (i, j) of the matrix is equal to 1 if there is a link between i and j and is 0 otherwise. The study of this stochastic process has a long tradition in applied mathematics and statistical physics (see for example Klafter and Sokolov (2011), Delvenne *et al.* (2010) and Lambiotte and Rosvall (2012)). Relevant to our results, it is known that the system converges to equilibrium if the underlying network is connected (Volz and Heckathorn, 2008). In this regime, nodes would be visited by coupons with a probability proportional to their degree and the whole network is explored, independently of the initial conditions. Equilibrium is reached after a characteristic timescale τ defined as $1/\lambda_2$, where λ_2 is the first non-zero eigenvalue of the Laplacian matrix driving p in equation (5). This timescale is associated with the presence of a bottleneck (the bridging nodes) between two strongly connected communities in the network. For times that are smaller than τ, the random walk has essentially explored almost uniformly one single community but has not sufficiently explored the other. This timescale therefore provides us with a way to estimate the minimal value of p that is needed for the whole graph to be sampled, i.e. $1-p<\lambda_2$.

Although the full structure of the network is not typically known in practice, the results of our numerical exercise suggest some general recommendations for studies in real settings as follows.

(a) Experimental researchers should be aware of the potential critical bridge nodes in the study population, which may vary according to the characteristics of the population.

(b) Experimental researchers should aim to have response rates at least above 0.4 to reduce the associated biases and uncertainty of the estimates. This recommended response rate may be decreased if more coupons are used.

(c) Attention should be taken to selecting the seeds as uniformly as possible, particularly aiming to avoid many seeds either in the small or in the large groups (typically the most approachable individuals). The temptation to start all seeds within well-hidden groups, or single locations, may cause the recruitment not to move beyond these groups.

(d) Restarting the seeds (to obtain larger sample sizes) during the on-going recruitment should be generally avoided. A better strategy may be either to start the experiment with more seeds or to increase response rates to avoid dropouts.

Acknowledgements

LECR is a *Chargé de recherches* of Fonds de la Recherche Scientifique.

References

Abdul-Quader, A., Heckathorn, D., McKnight, C., Bramson, H., Nemeth, C., Sabin, K., Gallagher, K. and Jarlais, D. D. (2006) Effectiveness of respondent-driven sampling for recruiting drug users in New York City: findings from a pilot study. *J. Urb. Hlth*, **83**, 459–476.

Abramovitz, D., Volz, E. M., Strathdee, S. A., Patterson, T. L., Vera, A., Frost, S. D. and Proyecto, E. (2009) Using-respondent-driven sampling in a hidden population at risk of HIV infection: who do HIV-positive recruiters recruit. *Sex. Transm. Dis.*, **26**, 750–756.

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

Costa, Jr, L. F., Oliveira, O. N., Travieso, G., Rodrigues, F. A., Boas, P. R. V., Antiqueira, L., Viana, M. P. and Rocha, L. E. C. (2011) Analyzing and modeling real-world phenomena with complex networks: a survey of applications. *Adv. Phys.*, **60**, 329–412.

Delvenne, J.-C., Yaliraki, S. and Barahona, M. (2010) Stability of graph communities across time scales. *Proc. Natn. Acad. Sci. USA*, **107**, 12755–12760.

Eckmann, J.-P., Moses, E. and Sergi, D. (2004) Entropy of dialogues creates coherent structures in e-mail traffic. *Proc. Natn. Acad. Sci. USA*, **101**, 14333–14337.

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

Gile, K. J., Johnston, L. G. and Salganik, M. J. (2015) Diagnostics for respondent-driven sampling. *J. R. Statist. Soc.* A, **178**, 241–269.

Guimera, R., Danon, L., Diaz-Guilera, A., Giralt, F. and Arenas, A. (2003) Self-similar community structure in a network of human interactions. *Phys. Rev.* E, **68**, article 065103R.

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

Holme, P., Edling, C. R. and Liljeros, F. (2004) Structure and time-evolution of an internet dating community. *Socl Netwrks*, **26**, 155–174.

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. S. (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. *J. Urb. Hlth*, **86**, 5–13.

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

Klafter, J. and Sokolov, I. M. (2011) *First Steps in Random Walks: from Tools to Applications*. Oxford: Oxford University Press.

Lambiotte, R. and Rosvall, M. (2012) Ranking and clustering of nodes in networks with smart teleportation. *Phys. Rev.* E, **85**, article 056107.

Lancichinetti, A. and Fortunato, S. (2009) Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities. *Phys. Rev.* E, **80**, article 016118.

Latapy, M. and Magnien, C. (2008) Complex network measurements: estimating the relevance of observed properties. In *Proc. 27th Conf. Computer Communications*. Phoenix: Institute of Electrical and Electronics Engineers.

Lee, S. H., Kim, P.-J. and Jeong, H. (2006) Statistical properties of sampled networks. *Phys. Rev.* E, **73**, article 016102.

Leskovec, J. (2014) Enron email network. Stanford University, Stanford. (Available from `http://snap.stanford.edu/data/email-Enron.html`.)

Lohr, S. L. (2009) *Sampling: Design and Analysis*. Boston: Cengage Learning.

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

Magnania, R., Sabinb, K., Saidela, T. and Heckathorn, D. (2005) Review of sampling hard-to-reach and hidden populations for HIV surveillance. *AIDS*, **19**, S67–S72.

Malmros, J., Liljeros, F. and Britton, T. (2014) Respondent-driven sampling and an unusual epidemic. *Preprint arXiv:1411.4867*. Department of Mathematics, Stockholm University, Stockholm.

Martin, J. L., Wiley, J. and Osmond, D. (2003) Social networks and unobserved heterogeneity in risk for AIDS. *Popln Res. Poly Rev.*, **22**, 65–90.

McCreesh, N., Johnston, L. G., Copas, A., Sonnenberg, P., Seeley, J., Hayes, R. J., Frost, S. D. W. and White, R. G. (2011) Evaluation of the role of location and distance in recruitment in respondent-driven sampling. *Int. J. Hlth Geog.*, **10**, 1–12.

McKnight, C., Jarlais, D. D., Bramson, H., Tower, L., Abdul-Quader, A. S., Nemeth, C. and Heckathorn, D. (2006) Respondent-driven sampling in a study of drug users in New York City: notes from the field. *J. Urb. Hlth*, **83**, 54–59.

Moody, J. (2001) Peer influence groups: identifying dense clusters in large networks. *Socl Netwrks*, **23**, 261–283.

Newman, M. (2010) *Networks: an Introduction*. New York: Oxford University Press.

Newman, M. E. J. (2002) The spread of epidemic disease on networks. *Phys. Rev.* E, **66**, article 016128.

Robinson, W., Risser, J., McGoy, S., Becker, A., Rehman, H., Jefferson, M., Griffin, V., Wolverton, M. and Tortu, S. (2006) Recruiting injection drug users: a three-site comparison of results and experiences with respondent-driven and targeted sampling procedures. *J. Urb. Hlth*, **83**, 29–38.

Rosvall, M. and Bergstrom, C. T. (2008) Maps of random walks on complex networks reveal community structure. *Proc. Natn. Acad. Sci. USA*, **105**, 1118–1123.

Salganik, M. J. (2006) Variance estimation and design effects and sample size calculations for respondent-driven sampling. *J. Urb. Hlth*, **83**, i98–i110.

Semaan, S. (2010) Time-space sampling and respondent-driven sampling with hard-to-reach populations. *Methodol. Innovns Online*, **5**, 60–75.

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

Verdery, A. M., Mouw, T., Bauldry, S. and Mucha, P. J. (2014) Network structure and biased variance estimation in respondent driven sampling. Submitted to *PLOS One*.

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

Wasserman, S. and Faust, K. (1994) *Social Network Analysis: Methods and Applications*. Cambridge: Cambridge University Press.

Wylie, J. L. and Jolly, A. M. (2013) Understanding recruitment: outcomes associated with alternate methods for seed selection in respondent driven sampling bmc. *Med. Res. Methodol.*, **13**, 1–11.