Mathematical Statistics
Stockholm University

Respondent-driven sampling and an unusual epidemic

Jens Malmros
Fredrik Liljeros
Tom Britton

Research Report 2014:4

Respondent-driven sampling and an unusual epidemic

Jens Malmros[*][†]   Fredrik Liljeros[‡]   Tom Britton[§]

November 2014

Abstract

Respondent-driven sampling (RDS) is frequently used when sampling hard-to-reach and/or stigmatized communities. RDS utilizes a peer-driven recruitment mechanism where sampled individuals pass on participation coupons to at most c of their acquaintances in the community ($c = 3$ being a common choice), who then in turn pass on to their acquaintances if they choose to participate, and so on. This process of distributing coupons is shown to behave like a new Reed-Frost type network epidemic model, in which becoming infected corresponds to receiving a coupon. The difference from existing network epidemic models is that an infected individual can not infect (i.e. sample) all of its contacts, but only at most c of them. We calculate R_0, the probability of a major "outbreak", and the relative size of a major outbreak in the limit of infinite population size and evaluate their adequacy in finite populations. We study the effect of varying c and compare RDS to the corresponding usual epidemic models, i.e. the case of $c = \infty$. Our results suggest that the number of coupons has a large effect on RDS recruitment. Additionally, we use our findings to explain previous empirical observations.

Key words: Respondent-driven sampling; Epidemic model; Configuration model; Reed-Frost.

[*]Department of Mathematics, Stockholm university, SE-106 91 Stockholm, Sweden
[†]Corresponding author: jensm@math.su.se
[‡]Department of Sociology, Stockholm university, SE-106 91 Stockholm, Sweden
[§]Department of Mathematics, Stockholm university, SE-106 91 Stockholm, Sweden

1 Introduction

Hidden populations are groups of individuals which i) have strong privacy concerns due to illicit or stigmatized behaviour, and ii) lack a sampling frame, i.e., their size and composition are unknown. Examples of hidden populations include several groups that are at high risk for contracting and spreading HIV, e.g., men who have sex with men, sex workers, and injecting drug users [1, 2, 3]; it is therefore of great importance to obtain reliable sampling methods for hidden populations in order to plan and evaluate interventions in the global HIV epidemic [4, 5].

Respondent-driven sampling (RDS) [6, 7] is a sampling methodology that utilizes the relationships between individuals in order to sample from the population. By combining an effective sampling scheme and the ability to produce unbiased population estimates, RDS has become the perhaps most preferred method when sampling from hidden populations. A typical RDS study starts with the selection of a group of seed individuals. Each seed is provided with a number of coupons, typically between three to five, to distribute to his or her peers in the population. An individual is eligible for participation upon presenting a coupon at the study site. Because recruitment takes place by coupons, participants remain anonymous throughout the study, but each coupon is numbered with a unique ID to keep track of who recruited whom. Incentives are given both for the participation of an individual as well as for the participation of those to whom he or she passed coupons. After participation, which commonly includes survey questions and possibly being tested for diseases, newly recruited individuals (i.e., respondents) are also given coupons to disperse among their contacts in the population. This procedure is then repeated until the desired sample size has been reached. The sampled individuals form a tree-like structure which is obtained from tracing the coupons. Recently, online based RDS methods (webRDS), where recruitment takes place via email and a survey is filled out at a designated web site, have also been put into use [8, 9, 10]. There are several procedures available for estimating population characteristics from RDS data, most of which use a Markov model in order to approximate the actual recruitment process [11, 12, 13, 14, 15, 16]; this is not the focus of the present paper.

A frequent problem in RDS studies is the inability of the recruitment process to reach the desired sample size due to premature failure of the recruitment chains started by the seeds [17]. This is often mitigated by additional seeds that enter the study as the rate of recruitment declines; e.g., in [17], 43% of reviewed RDS studies with available data reported that additional seeds were used. Relatedly, it has been observed in webRDS studies, where recruitment is allowed to go on until it stops by itself, that the recruitment process fails to reach a large proportion of the population despite additional seeds joining in at a later time [10, 18]. While there are most

1

likely several reasons behind recruitment chain failure, such as community structure in the population causing chains to become stuck in a sub-network and/or clustering that has a similar effect, but more locally, an important reason is the limited number of coupons in the RDS recruitment process. This is the main focus of this paper. Furthermore, recruitment chain failure is highly associated with the ability of the recruitment process to start successful recruitment chains, the probability of such chains occurring, and the relative size of the population that is reached by an RDS study, all of which are related to quantities typically studied in epidemic modelling. As it turns out, it is possible to use models of infectious disease spread on social networks to describe coupon distribution in RDS, where the disease is defined as "participation in the study" and spreads by the RDS coupon distribution mechanism.

The simplest model of infectious disease spread is the Reed-Frost model, see e.g. [19, p. 11-18], where in each generation i, each infectious individual independently infects each susceptible individual with the same probability. The individuals that were infected by the individuals in generation i make up generation $i + 1$ of infectious individuals in the epidemic. After spreading the disease, the individuals in generation i are considered recovered (or dead) from the disease and are removed from the process. In the original version of the model, an infectious individual attempts to infect all susceptible individuals in the population. The model is however easily modified to the more realistic case when the structure of the population is described by a social network, hence imposing the restriction that an infectious individual only may spread the disease to his or her contacts in the social network independently of each other with the same probability. Infectious diseases are usually able to spread to all contacts of an individual, and consequently, the Reed-Frost model and other epidemic models defined on social networks do not impose any restrictions on the number of individuals that an infectious individual can infect other than those given by population structure. The RDS recruitment process differs from infectious diseases in that its spread is restricted by the limited number of coupons. Consequently, individuals with more population contacts than the number of coupons distributed to them have less capability of recruiting than if RDS recruitment were to spread in the usual manner of an epidemic, i.e. without any limitations. Depending on how the number of contacts (i.e, degrees) of population members are distributed, this may have a large effect on the capability of the RDS recruitment process to sustain and initiate recruitment. Furthermore, it may affect the ability of the recruitment process to reach a substantial proportion of the population, as the sampling procedure can limit recruitment to parts of the population.

In this paper, we model RDS as an epidemic taking place on a social network by defining a Reed-Frost type model which has an upper limit on the number of individuals that an infectious individual could infect. We will

use both infectious disease terminology and RDS terminology when referring to this model. In order to be able to specify the degree distribution of the social network, we use the configuration model [20, 21] to describe the structure of the population. We calculate the *basic reproduction number*, i.e., the number of individuals that are infected by a typical infectious individual during the early stages of the epidemic. This is often denoted by R_0. We say that there is a *major outbreak* if a non-negligible proportion of the population is infected and calculate the probability τ of such outbreaks occurring. If $R_0 \leq 1$, it is not possible for a major outbreak to occur, while if $R_0 > 1$, a major outbreak may occur. The critical value of $R_0 = 1$ is often referred to as the *epidemic threshold*. We also calculate the relative size of an outbreak in case of a major outbreak z using so-called *susceptibility sets* [22, 23]. Note that τ and z are positive only if R_0 is larger than the epidemic threshold. We compare the RDS recruitment process to corresponding epidemics with unrestricted spread and investigate the effect of varying the number of coupons and the coupon transfer probability. To our knowledge, there are no previous studies of epidemics on networks that describes behaviour similar to the present one, although the model in [24] allows for a restriction on the number of individuals that an infectious individual can infect in a homogeneously mixing population (i.e. a population without network structure).

2   Models

2.1   Network model

We consider a configuration model network consisting of n vertices. In later calculations, we will assume that $n \to \infty$. Each individual $i, i = 1, \ldots, n$, is assigned an i.i.d. number of stubs (half-edges) d_i from a prescribed distribution D having support on the non-negative integers. The network is then formed by pairing stubs together uniformly at random. If $\sum_{i=1}^{n} d_i$ is odd, an edge is added to the n:th vertex (this does not influence our results in the limit of infinite population size). This construction allows the formation of multiple edges and self-loops; it is however well known that the fraction of these is small if D has finite second moment. Specifically, the probability of the resulting graph being simple is bounded away from 0 as $n \to \infty$; see [25, Theorem 7.8] and [26, Lemma 5.3]. Hence we can condition on the graph being simple given that $E(D^2) < \infty$. Alternatively, we may proceed by removing multiple edges and self-loops from the generated graph since asymptotically this does not change the degree distribution if D has finite second moment; see [25, Theorem 7.9]. Hence, we will from now on assume that the resulting graph is simple. Moreover, the graph is locally tree-like when $E(D^2) < \infty$, meaning that it with high probability does not contain short cycles [26]. Hence, we can take advantage of the branching process

[e.g., 27] approximations that are often used for epidemics, see e.g. [19, ch. 3]. In what follows, we will assume that the degree distributions considered have finite second moment.

2.2 Epidemic model

On this graph, describing the social structure in a community, we define an epidemic model mimicking the RDS recruitment process. In this model, becoming infected corresponds to participating in the RDS study. Initially, all members of the population (vertices) are susceptible. The epidemic starts with one randomly selected individual (vertex), the index case, being infected from the outside. The infected individual uniformly selects c of his or her neighbours in the population and infects them independently of each other with the same probability p. The parameter c corresponds to the number of coupons in RDS and the parameter p to the probability of being successfully recruited to the RDS study. If the infected individual has less than c contacts, he or she infects all his or her contacts independently of each other with probability p. The newly infected individuals make up the first generation of the epidemic. After spreading the disease, the initially infected individual recovers and becomes immune (or dies) and has no further role in the epidemic. The individuals in the first generation each in turn select c of their neighbours excluding the one who infected them (which for the first generation is the index case), regardless of whether they are susceptible or not. If an individual has less than c neighbours excluding the one who infected him or her, he or she selects all of his or her neighbours. Then, they infect the selected contacts that are susceptible, independently of each other with probability p, and then recover; contacts with already infected individuals have no effect. The now infected individuals form the second generation of the epidemic. The disease continues to spread in the same fashion from the second generation and onward until there are no newly infected individuals in a generation. The individuals that were infected during the course of the epidemic make up the outbreak, and the number of ultimately infected individuals is the final size of the outbreak. Note that if we let $c = \infty$, we get the standard Reed-Frost epidemic taking place on the configuration model network [26].

Because an individual only tries to infect those he or she selected, the spread of the disease, or coupon distribution mechanism, in our model is more similar to that of webRDS than physical RDS. We discuss this further and present other possible coupon distribution mechanisms in Section 5.

3 Calculations

3.1 The basic reproduction number R_0

Assume that we have a configuration model graph G of size n, where n is large, and let the degree distribution of G be D, where $P(D = k) = p_k$. The degree of a given neighbour of an individual follow the *size-biased* degree distribution $\tilde{D}$, where $P(\tilde{D} = k) = \tilde{p}_k = k p_k / E(D)$. Assume that we have an epidemic spreading on this graph according to the description in Subsection 2.2. The degree of the index case is then distributed as D, and the degree of infected individuals in later generations during the early stages of an outbreak is distributed as $\tilde{D}$. As previously mentioned in Subsection 2.1, the graphs generated by the configuration model will with high probability not contain short cycles, meaning that we can approximate the spread of the epidemic with a (forward) branching process. Let X and $\tilde{X}$ be the offspring of the ancestor (i.e., the index case) and of the later generations in this branching process, respectively. Given that the index case has degree $k \leq c$, he or she can at most infect k neighbours. If the index case has degree larger than or equal to $c + 1$, he or she infects at most c neighbours. Because infections happens independently with the same probability p, we have that, conditionally on the degree, the probability that the index case infects j neighbours is

$$P(X = j | D = k) = \binom{c \wedge k}{j} p^j (1 - p)^{(c \wedge k) - j}, \tag{1}$$

where $j = 0, \ldots, c \wedge k$. Infectious individuals in later generations have one less contact available for infection (the one that infected them). Hence, we get that, conditionally on the degree, the probability that an infectious individual in later generations infects j neighbours is

$$P(\tilde{X} = j | \tilde{D} = k) = \binom{c \wedge (k - 1)}{j} p^j (1 - p)^{(c \wedge (k-1)) - j}, \tag{2}$$

where $j = 0, \ldots, c \wedge (k - 1)$.

Because the ability of an individual to spread the disease will depend on its degree, the offspring distributions are obtained by conditioning on the degree:

$$P(X = j) = \sum_{k=j}^{\infty} P(X = j | D = k) p_k; \tag{3}$$

$$P(\tilde{X} = j) = \sum_{k=j+1}^{\infty} P(\tilde{X} = j | \tilde{D} = k) \tilde{p}_k, \tag{4}$$

where $j = 0, \ldots, c$, and the probabilities $P(X = j | D = K)$ and $P(\tilde{X} = j | \tilde{D} = k)$ come from Eqs. (1) and (2), respectively. From standard branching process theory [27] we have that R_0 is the expected number of individuals

that get infected by an infectious individual in the second and later generations; hence

$$R_0 = E(\tilde{X}) = \sum_{j=0}^{c} j \sum_{k=1}^{\infty} P(\tilde{X} = j | \tilde{D} = k) \tilde{p}_k \tag{5}$$

$$= \sum_{j=0}^{c} j \left(\sum_{k=j}^{c-1} \binom{k}{j} p^j (1-p)^{k-j} \tilde{p}_k + \binom{c}{j} p^j (1-p)^{c-j} \left(1 - \sum_{k=1}^{c} \tilde{p}_k \right) \right).$$

The obtained R_0 is increasing in p and c, and for a fixed p, $R_0 \to R_0^{(\text{unrestricted})}$ as $c \to \infty$, where $R_0^{(\text{unrestricted})}$ is the R_0 value for the standard Reed-Frost epidemic on a configuration model network, given by [26]

$$R_0^{(\text{unrestricted})} = \left(E(D) + \frac{\text{Var}(D) - E(D)}{E(D)} \right).$$

3.2 Probability of major outbreak

When $R_0 > 1$, it is possible for a major outbreak to occur. The probability τ of such an outbreak occurring is given by the survival probability of the approximating branching process, which we get by standard techniques. We first consider a branching process with offspring distribution $\tilde{X}$ for all individuals, i.e. also for the index case. Let the extinction probability of this process be $\tilde{\pi}$. For the process to die out, all the branching processes initiated by the offspring of the ancestor must die out; hence by conditioning on the number of offspring in the first generation of the process, we get

$$\tilde{\pi} = \sum_{j=0}^{c} \tilde{\pi}^j P(\tilde{X} = j) = \tilde{\rho}(\tilde{\pi}), \tag{6}$$

where $\tilde{\rho}$ is the probability generating function of $\tilde{X}$. The solution to Equation (6) is obtained numerically. In our original branching process the ancestor has offspring distribution X and later generations have offspring distribution $\tilde{X}$. Again by conditioning on the number of individuals in the first generation, we get that the extinction probability π of the original branching process is

$$\pi = \rho(\tilde{\pi}), \tag{7}$$

where $\tilde{\pi}$ is the solution to Equation (6) and ρ is the probability generating function of X. The solution to Equation (7) is given by numerical calculations, and we obtain the probability of a major outbreak $\tau = 1 - \pi$.

Note that if we have $1 < s < \infty$ initially infected individuals in the epidemic, the probability of a major outbreak is $1 - \pi^s$, which approaches 1 as s becomes large. The number of initially infected individuals does not affect R_0 or the relative size of a major outbreak calculated in Subsection 3.3.

6

3.3 Relative size of a major outbreak

The relative size of a major outbreak in case of a major outbreak z can be obtained using susceptibility sets, constructed as follows. For each individual i, we can obtain a random list of which neighbours that i would infect given that it were to be infected. By combining the lists from all individuals in the population, it is possible to construct a directed graph with all vertices (individuals) in which there is an arc from vertex i to vertex j if j is in i:s list. The susceptibility set of an individual j consists of all individuals in this directed graph, including j itself, from which there is a directed path to j. Hence, j:s susceptibility set is such that the infection of any individual in the set would result in the ultimate infection of j. Note that j will be infected in the epidemic if and only if the initially infected individual is in j:s susceptibility set.

The susceptibility set of a randomly chosen individual, i_0 say, can be approximated with a (backward) branching process in which i_0 is the only member of the zeroth generation. We consider the number of neighbours that, if they were to be infected, would infect i_0 (as opposed to previously when we considered the number of neighbours that an individual would infect were it to be infected). Suppose that i_0 have degree d. Because all neighbours of i_0 contact him or her with the same probability θ independently of each other, the number of neighbours that contact him or her is $\mathrm{Bin}(d, \theta)$-distributed; hence, the unconditional distribution of the number of neighbours that contact him or her is a mixed binomial distribution with parameters D and θ. We now derive an equation for the contact probability θ. The degree distribution of the neighbouring individuals is $\tilde{D}$, so we obtain

$$\theta = \sum_{k=0}^{\infty} \theta_k \tilde{p}_k, \tag{8}$$

where θ_k is the probability that a neighbour with degree k contacts i_0. Because a neighbour of i_0 with degree k has to be contacted first in order to become infected, only $k - 1$ edges are available for him or her to spread the disease. Therefore, a neighbour must have at least degree two in order to first become infected and then contact i_0. If a neighbour has degree $k \geq c + 2$, he or she first selects c of the available $k - 1$ contacts and then attempts to spread the disease to them. Hence, the contact probabilities are

$$\theta_k = \begin{cases} 0, & k = 0, 1; \\ p, & k = 2, \ldots, c + 1; \\ \frac{c}{k-1} p, & k = c + 2, c + 3, \ldots. \end{cases} \tag{9}$$

The probability that a neighbour makes contact with i_0 depends on his or her degree. Hence, the degree distribution of individuals in the first generation, i.e. those neighbours of i_0 that makes contact with i_0, and of individuals in

later generations in the backward branching process is altered by the fact that they have contacted another individual. Conditionally on the event that a contact has been made, call it C, the distribution of the degree D^* of an individual in the first and later generations of the susceptibility set process is given by

$$
\begin{aligned}
P(D^* = k) &= P(\tilde{D} = k | C) \\
&= \frac{P(C | \tilde{D} = k) P(\tilde{D} = k)}{\sum_{k=0}^{\infty} P(C | \tilde{D} = k) P(\tilde{D} = k)} \\
&= \frac{\theta_k \tilde{p}_k}{\theta},
\end{aligned} \tag{10}
$$

so

$$
P(D^* = k) = \begin{cases} 0, & k = 0, 1; \\ \frac{p \tilde{p}_k}{\theta}, & k = 2, \ldots, c + 1; \\ \frac{cp \tilde{p}_k}{(k-1)\theta}, & k = c + 2, c + 3, \ldots. \end{cases} \tag{11}
$$

An individual in later generations of the process will be contacted by any of his or her neighbours independently of other neighbours with the same probability θ. Given that this individual has degree k, the number of neighbours that contact him or her is binomially distributed with parameters $k - 1$ and θ. Hence, the unconditional distribution of the number of neighbours that contact an individual in later generations is mixed binomial with parameters $D^* - 1$ and θ.

If the approximating backward branching process contains few individuals, it is unlikely that i_0 will be infected, whereas if the process reaches a large number of individuals (i.e. grows infinitely large), there is a positive probability that i_0 will not escape infection. More specifically, the probability that i_0 will be infected during a major outbreak is given by the survival probability of the backward branching process. Because i_0 is chosen randomly, we also have that the relative size of an outbreak in case of a major outbreak is given by the survival probability of the backward branching process. Let Y be the number of offspring of the ancestor and Y^* the number of offspring of individuals in later generations in the approximating branching process, respectively. Hence, $Y \sim \text{MixBin}(D, \theta)$ and $Y^* \sim \text{MixBin}(D^* - 1, \theta)$. We obtain the survival probability of the process similarly as in Subsection 3.2. Let the extinction probability of a branching

8

process with offspring distribution Y^* be π^*. We have

$$
\begin{aligned}
\pi^* &= \sum_{j=0}^{\infty} (\pi^*)^j P(Y^* = j) = E((\pi^*)^{Y^*}) \\
&= E(E((\pi^*)^{Y^*} | D^*)) = E(1 - \theta + \theta\pi^*)^{D^*-1} \\
&= \sum_{k=2}^{\infty} (1 - \theta + \theta\pi^*)^{k-1} P(D^* = k) \\
&= \frac{p}{\theta} \sum_{k=2}^{c+1} (1 - \theta + \theta\pi^*)^{k-1} \tilde{p}_k + \frac{cp}{\theta} \sum_{k=c+2}^{\infty} (1 - \theta + \theta\pi^*)^{k-1} \frac{\tilde{p}_k}{k-1}.
\end{aligned}
\tag{12}
$$

The solution to Equation (12) for π^* is obtained numerically. Let the extinction probability of the approximating branching process be π'. Then,

$$
\begin{aligned}
\pi' &= \sum_{j=0}^{\infty} (\pi^*)^j P(Y = j) = E((\pi^*)^{Y}) \\
&= E(E((\pi^*)^{Y} | D)) = E(1 - \theta + \theta\pi^*)^{D} \\
&= f_D(1 - \theta + \theta\pi^*),
\end{aligned}
\tag{13}
$$

where $f_D(\cdot)$ is the probability generating function of D and π^* is the solution to Equation (12). The solutions to Equation (13) is obtained numerically, and the relative final size of the epidemic in case of a major outbreak is $z = 1 - \pi'$.

A rigorous proof of that $z = 1 - \pi'$ is beyond the scope of this paper. It has been proved that for Reed-Frost epidemics on random intersection graphs [28] and Reed-Frost epidemics on configuration model graphs [29] that the proportion of infected during the epidemic converges in probability to the survival probability of the backward branching process. Similar arguments could also be used for our process to provide a formal proof. Additionally, we believe that the techniques described in [30] could be used to obtain stronger results for the whole epidemic process.

4 Numerical results and simulations

We now numerically examine the analytical results obtained in Section 3. In particular, we examine the relation between R_0, τ, and z and the parameters c and p, and compare the RDS recruitment process with unrestricted epidemics. We use two different degree distributions in our calculations, the Poisson degree distribution and a variant of the power-law degree distribution with exponential cut-off given by $p_k \propto k^{-\alpha} \exp(-k/\kappa)$, $k = 1, 2, \ldots$, where α is the power-law exponent and κ refers to the exponential cut-off [e.g. 31].
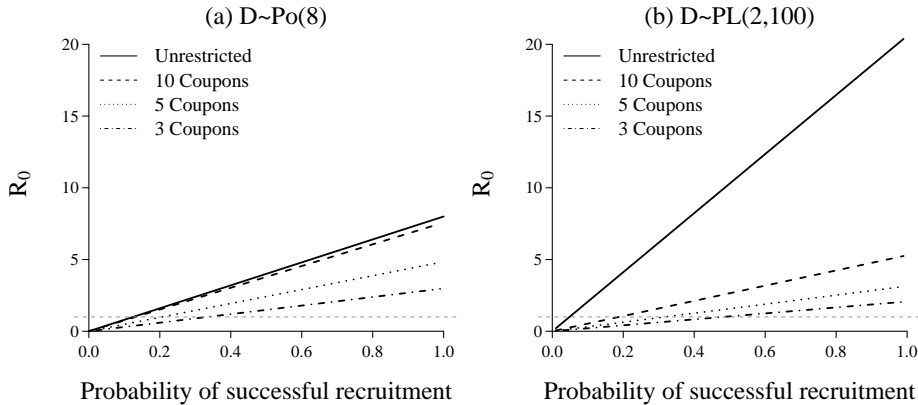
Figure 1: Comparison of R_0 for unrestricted epidemics and RDS recruitment processes with 10, 5, and 3 coupons and $p \in [0, 1]$. Plot (a) show the results for the Poisson degree distribution with parameter $\lambda = 8$ and plots (b) show the results for the power-law degree distribution with parameters $\alpha = 2$ and $\kappa = 100$. The dashed horizontal lines shows the threshold value $R_0 = 1$.

In Figure 1, we show the R_0 values for the RDS recruitment process with $c = 3, 5, 10$ and the unrestricted epidemic for $p \in [0, 1]$. Figure 1 (a) shows the results for the Poisson degree distribution with parameter $\lambda = 8$ and Figure 1 (b) shows the results for the power-law degree distribution with parameters $\alpha = 2$ and $\kappa = 100$. For both degree distributions and a fixed value of p, the limitation imposed by the number of coupons on disease spread yields smaller R_0 values for the RDS recruitment process when compared to the unrestricted epidemic for all values of c. Especially for the power-law degree distribution, all values of c give much smaller R_0 values than those of the unrestricted epidemic, and the value of p for which R_0 becomes larger than 1 (i.e., the epidemic threshold) is larger than that of the unrestricted epidemic for all values of c.

Figure 2 shows the values of τ and z for the RDS recruitment process with $c = 3, 5, 10$ and the unrestricted epidemic for $p \in [0, 1]$. Figures 2 (a) and 2 (b) show the results for τ and z, respectively, for the Poisson degree distribution with parameter $\lambda = 8$ and Figures 2 (c) and 2 (d) show the results for τ and z, respectively, for the power-law degree distribution with parameters $\alpha = 2$ and $\kappa = 100$. The relative size of a major outbreak is always smaller than the probability of a major outbreak for both degree distributions. For both degree distributions, the probability of a major outbreak for the RDS recruitment process is smaller than that of the unrestricted epidemic for small values of p and approaches that of the unrestricted epidemic when $p \to 1$. For the power-law degree distribution, the size of a major outbreak is much smaller than that of the unrestricted epidemic for all values of c and p.
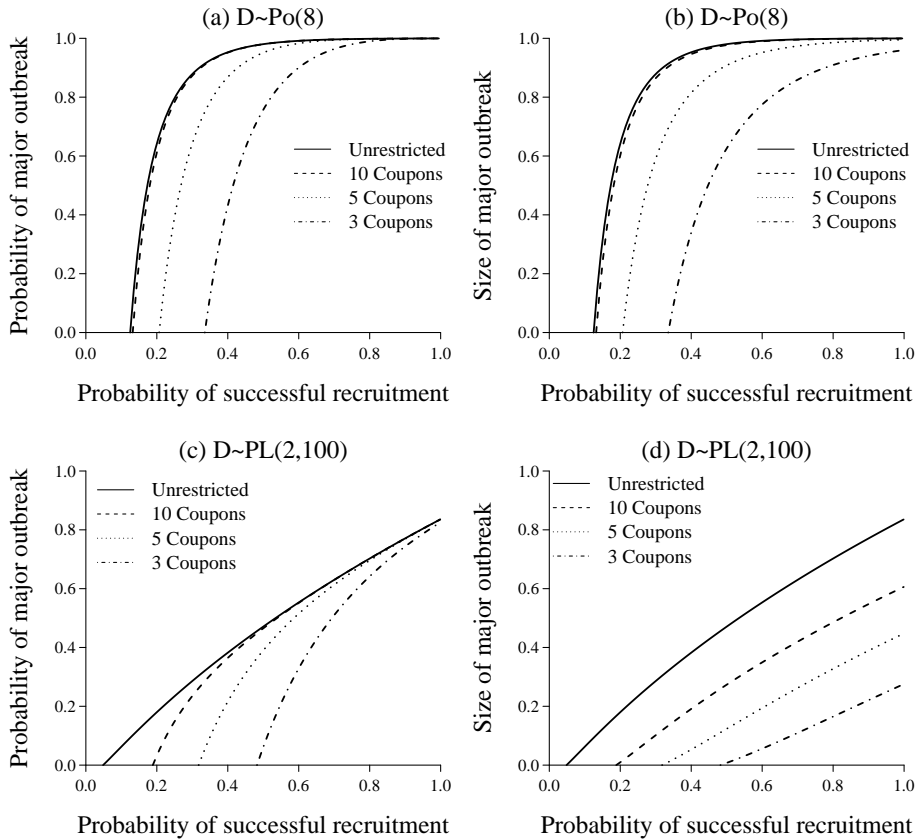
10

Figure 2: Comparison of the asymptotic probability of a major outbreak and relative size of a major outbreak for unrestricted epidemics and RDS recruitment processes with 10, 5, and 3 coupons and $p \in [0, 1]$. Plots (a) and (b) show the results for the Poisson degree distribution with parameter $\lambda = 8$ and plots (c) and (d) show the results for the power-law degree distribution with parameters $\alpha = 2$ and $\kappa = 100$.

We also make a brief evaluation of the adequacy of our asymptotic results in finite populations by means of simulations. From simulated RDS recruitment processes (as described by the model), we estimate the probability of a major outbreak and the relative size of a major outbreak in case of a major outbreak by the relative proportion of major outbreaks and the mean relative size of major outbreaks, respectively. Given a degree distribution and number of coupons c, let p_c be the smallest value of p for which the process is above the epidemic threshold. Each simulation run consists of generating a network of size 5000 by an erased configuration model approach [32], for which we make use of the iGraph R package [33]. Then, RDS recruitment processes are run on the generated network for values of $p \in [p_c, 1]$. In Figure 3, we show the estimated probability of a major outbreak $\hat{\tau}$ and estimated relative size of a major outbreak in case of a major outbreak $\hat{z}$ for varying p and the corresponding asymptotic results. Figure 3 (a) shows the results for the Poisson degree distribution with parameter $\lambda = 12$ from 5000 simulations runs of RDS recruitment processes with 3 coupons. Figure 3 (b) shows the results for the power-law degree distribution with parameters $\alpha = 2.5$ and $\kappa = 50$ from 5000 simulation runs of RDS recruitment processes with 10 coupons. In both Figures 3 (a) and (b), we show error bars for the estimates based on ± 2 standard errors, where the standard error for $\hat{\tau}$ is estimated as $SE(\hat{\tau}) = (\hat{\tau}(1 - \hat{\tau})/m))^{1/2}$, where m is the number of simulations, and the standard error for $\hat{z}$ is estimated as $SE(\hat{z}) = (\hat{\sigma}^2/m_{maj})^{1/2}$, where $\hat{\sigma}^2$ is the sample variance of the relative final sizes of major outbreaks and m_{maj} is the number of simulations resulting in a major outbreak.

Note that it is not well defined what constitutes a major outbreak in small, finite populations. Usually, the threshold for when an outbreak constitutes a major outbreak is determined by inspecting the distribution of outbreak sizes. Typically, this distribution is bimodal with modes at 0 and z, corresponding to small and major outbreaks. In our model, outbreak sizes will depend on p. For p close to p_c, where "close" depends on the degree distribution, small and major outbreaks are indistinguishable. Consequently, it is difficult to estimate τ and z for such values of p. In Figure 3, we have chosen to set the (relatively small) threshold for major outbreaks to 2% of the population over the whole interval $[p_c, 1]$. This yields fairly correct estimates for p close to p_c and does not affect estimates for p further away from p_c.

We see that both the estimated probability of a major outbreak and the estimated relative size of major outbreak in case of a major outbreak are very well approximated by the asymptotic results for both the evaluated degree distributions. As pointed out in [34], the relative size of the epidemic is more efficiently estimated than the probability of a major outbreak because each simulation yields many (correlated) observations of the backward process and only one observation of the forward process.
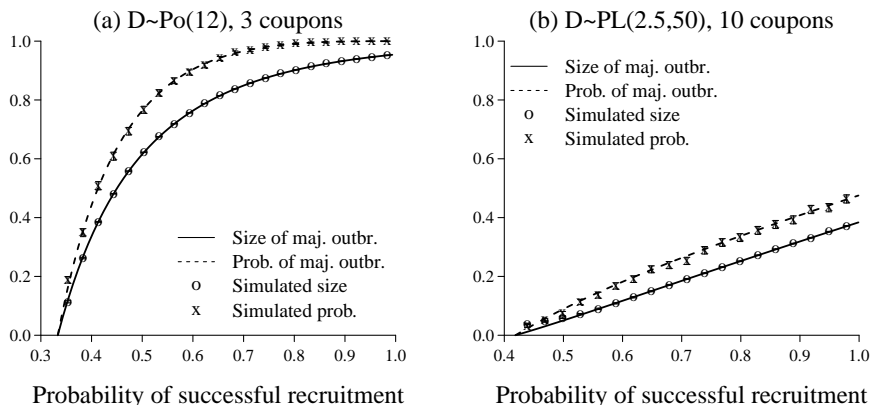
Figure 3: Comparison of results from simulations of RDS recruitment processes and the asymptotic probability and relative size of a major outbreak. Plot (a) shows the results for the Poisson degree distribution with parameter $\lambda = 12$ for processes with $c = 3$ and plot (b) shows the results for the power-law degree distribution with parameters $\alpha = 2.5$ and $\kappa = 50$ for processes with $c = 10$. Note that the error bars for the simulated relative size are very narrow and not visible for most simulated values. Also note that the horizontal scales are different.

5 Discussion and conclusions

When the RDS recruitment process is compared to the corresponding unrestricted epidemic, it is clear that the limited number of coupons has a large impact on R_0 and the value of p_c corresponding to the epidemic threshold, the probability of a major outbreak, and the relative size of a major outbreak in case of a major outbreak. This is especially true for the power-law degree distribution, for which in particular R_0 and z is much smaller than for the corresponding unrestricted epidemic. In social networks with power-law degree distribution, the vast majority of individuals will have small degrees. For these individuals, the probability of being infected in an epidemic will be small. Also, such an individual will, once infected, have few or no contacts to spread the disease to. Hence, the spread of an epidemic in such networks will be highly dependent on a few individuals with very large degrees that have the capacity to infect many of their (small degree) neighbours. Because of the relatively small value of c, the potential of large degree individuals to spread the disease is much impaired in RDS compared to an unrestricted epidemic with the same p, hence impairing the spread of the epidemic as a whole.

The impact of the number of coupons on the RDS recruitment process may in part explain why some RDS studies experience difficulties in obtaining the desired sample size and/or recruiting a substantial proportion of the

study population. Given p, the number of coupons will be crucial to whether R_0 is above or below the epidemic threshold for the recruitment process; in the latter case all recruitment chains will eventually fail. Moreover, the proportion of the population recruited by the RDS recruitment process may be small even given that p is relatively large and a major outbreak occurs. For some parameter combinations, the proportion reached can be very small; this is especially important to consider in webRDS. We illustrate this by considering the webRDS studies in [10] and [18]. In both studies, each respondent were allowed to make 4 recruitments. In the latter study, 66% of started recruitment chains had a depth of one generation (i.e. index case and one generation of recruitments) and 11% had a depth of three generations or more. This indicates that R_0 is below the epidemic threshold for this study and therefore, recruitment never takes off. In the former study, the majority of recruitments come from long recruitment chains, implying that R_0 is above the epidemic threshold. Still, recruitment eventually declined and stopped completely before reaching a large part of the population despite additional seeds joining the study. As we see in Section 4 however, relatively many parameter combinations with $R_0 > 1$ yields small z values, which could explain the observed behaviour. For both studies, heterogeneity in network structure, such that, locally $R_0 < 1$, may also be an explanation. It would be of interest to find proper inference procedures for our model to be used in further evaluation of actual RDS studies with respect to the quantities studied in this paper.

One might consider other ways to distribute coupons. The coupon distribution mechanism in our model, where a respondent selects some of his or her neighbours for attempted coupon transfer while ignoring those neighbours that were not selected, is most similar to a webRDS process. In a physical RDS study where coupons are handed over from person to person, a respondent may attempt to distribute a coupon to another neighbour if the originally intended recipient declines (here, distributing a coupon implies study participation). This modified mechanism is given as follows. A respondent first attempts to give a coupon to a randomly chosen neighbour. If the coupon is rejected, the respondent may try to distribute the same coupon to another neighbour, randomly chosen among those who previously have not been offered a coupon. When the coupon is accepted, the procedure is repeated starting by randomly selecting among those neighbours that have not been offered a coupon. When there are no more neighbours and/or coupons left, no further distribution attempts are made. The offspring probabilities in the branching process are the same as previously for individuals with degree less than the number of coupons, but the distribution of the number of coupons given out by an individual with degree larger than c will be tilted towards larger values compared to the previous model. The probabilities in
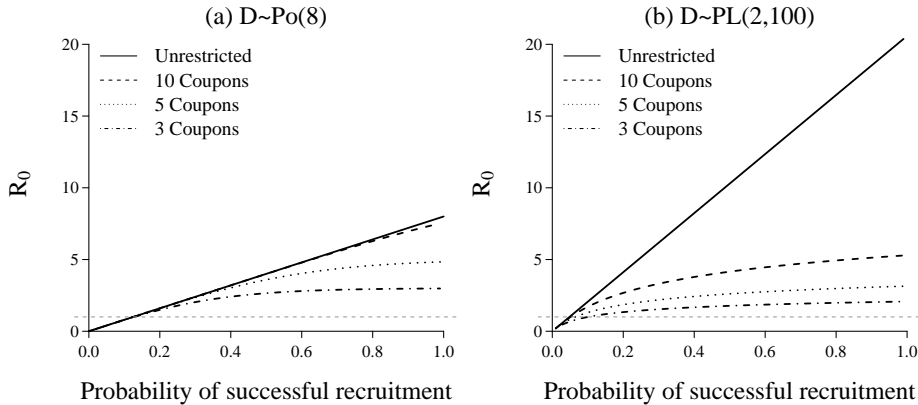
Figure 4: Comparison of R_0 for unrestricted epidemics and RDS recruitment processes where a recruiter tries to distribute a coupon until success. Plot (a) shows the results for the Poisson degree distribution with parameter $\lambda = 8$ and plot (b) shows the results for the power-law degree distribution with parameters $\alpha = 2$ and $\kappa = 100$.

Eq. (1) now become

$$P(X = j | D = k) = \begin{cases} \binom{k}{j} p^j (1-p)^{k-j}, & j < c; \\ \sum_{i=c}^{k} \binom{k}{i} p^i (1-p)^{k-i}, & j = c. \end{cases} \quad (14)$$

It is straightforward to calculate R_0 and τ using the same techniques as in Sections 3.1 and 3.2. Figure 4 shows the R_0 values for the modified RDS recruitment process with $c = 3, 5, 10$ and the unrestricted epidemic for $p \in [0, 1]$. In Figure 4 (a), we show the results for the Poisson degree distribution with $\lambda = 8$ and in Figure 4 (b) we show the results for the power-law degree distribution with parameters $\alpha = 2$ and $\kappa = 100$. It is clear that R_0 is larger for the modified recruitment process for all p compared to the process described in Subsection 2.2 and the p value corresponding to the epidemic threshold is considerably smaller. When $p \to 1$, the R_0 values converges to those seen in Figure 1. Because the modified process has similar epidemic threshold values in terms of p for different c, the corresponding τ values (not shown) are close to those for the unrestricted epidemic when $R_0 > 1$. For the final size of the epidemic, the calculations are much harder to derive and is thus out of the scope of this paper. There are several other complications that could be considered in terms of coupon distribution. E.g., it is not likely that all coupon distribution attempts of a respondent will have the same success probability, both because the respondent may act differently depending on how many attempts he or she has previously made and because the relations to his or her neighbours may be different. Other complications include different respondent behaviour depending on

15

(measurable) individual characteristics, geographical variations, and time dependence.

Overall, our results indicate that RDS studies which experience difficulties with respect to recruitment chain failure could benefit from an increased number of coupons, which would reduce the number of additional seeds needed. Furthermore, the longer recruitment chains obtained as a result of an increased number of coupons are more likely to reach remote parts of the population and meet equilibrium criteria for inference. As the recruitment potential of RDS increases from an increased number of coupons, the time to reach the desired sample size is shortened. Additionally, the study time is not subject to unexpected prolongation due to the addition of seeds. Hence, an increased number of coupons may result in lower and more predictable study costs. For webRDS studies in particular, the increase in the proportion of the population reached due to increasing the number of coupons facilitates larger sample sizes. We therefore advise that the recruitment potential of a planned RDS study should be considered beforehand so that the number of coupons could be chosen large enough to facilitate sustained recruitment and an acceptable sample size. Other factors may also increase recruitment potential. The coupon transfer probability p could be increased by e.g. larger incentives or improved information about the study; this has an immediate effect on R_0, τ, and z. Additionally, the selection of seeds could also affect recruitment capability, see e.g. [35] where different seed selection methods produce very different recruitment scenarios. In general, it is of interest to further study why certain RDS studies are more successful in reaching the desired sample size with a modest number of seeds.

The presented epidemic model is a novel contribution to the area of stochastic epidemic models and although many results from Reed-Frost epidemics on configuration model networks are expected to hold for this model, several properties of it remain to be studied. There are a number of extensions that can be considered, e.g. different recruitment probabilities through unequally weighted edges, controlling for network structural properties, e.g. clustering, and modifying the coupon distribution mechanism as previously described.

Acknowledgements

References

[1] Beyrer C, Baral SD, van Griensven F, Goodreau SM, Chariyalertsak S, Wirtz AL, Brookmeyer R. 2012 Global epidemiology of HIV infection in men who have sex with men. *The Lancet* **380**, 367–377. (doi: 10.1016/S0140-6736(12)60821-6)

[2] Kerrigan D, Wirtz A, Semini I, N'Jie N, Stanciole A, Butler J, Oelrichs R, Beyrer C. 2012 *The Global HIV Epidemics among Sex Workers*. The World Bank. (doi:10.1596/978-0-8213-9774-9)

[3] Aceijas C, Stimson GV, Hickman M, Rhodes T. 2004 Global overview of injecting drug use and HIV infection among injecting drug users. *AIDS* **18**, 2295–2303. (doi:10.1097/00002030-200411190-00010)

[4] Magnani R, Sabin K, Saidel T, Heckathorn D. 2005 Review of sampling hard-to-reach and hidden populations for HIV surveillance. *AIDS* **19**, 67–72. (doi:10.1097/01.aids.0000172879.20628.e1)

[5] Lamptey PR, Dirks RG. 2008 HIV/AIDS, reaching high-risk populations. In *Encyclopedia of Social Problems* (ed. VN Parrillo). 443–447. CA: SAGE Publications, Inc. (doi:10.4135/9781412963930)

[6] Heckathorn DD. 1997 Respondent-driven sampling: a new approach to the study of hidden populations. *Soc. Probl.* 174–199. (doi: 10.2307/3096941)

[7] Heckathorn D. 2002 Respondent-driven sampling II: Deriving valid population estimates from chain-referral samples of hidden populations. *Soc. Probl.* **49**, 11–34. (doi:10.1525/sp.2002.49.1.11)

[8] Wejnert C, Heckathorn DD. 2008 Web-based network sampling - Efficiency and efficacy of respondent-driven sampling for online research. *Sociol. Method Res.* **37**, 105–134. (doi:10.1177/0049124108318333)

[9] Wejnert C. 2009 An empirical test of respondent-driven sampling: Point estimates, variance, degree measures, and out-of-equilibrium data. *Sociol. Methodol.* **39**, 73–116. (doi:10.1111/j.1467-9531.2009.01216.x)

[10] Bengtsson L, Lu X, Nguyen QC, Camitz M, Hoang NL, Nguyen TA, Liljeros F, Thorson A. 2012 Implementation of web-based respondent-driven sampling among men who have sex with men in vietnam. *PLoS ONE* **7**. (doi:10.1371/journal.pone.0049417)

[11] Salganik MJ, Heckathorn DD. 2004 Sampling and estimation in hidden populations using respondent-driven sampling. *Sociol. Methodol.* **34**, 193–240. (doi:10.1111/j.0081-1750.2004.00152.x)

[12] Volz E, Heckathorn DD. 2008 Probability based estimation theory for respondent driven sampling. *J. Off. Stat.* **24**, 79–97

[13] Gile KJ. 2011 Improved inference for respondent-driven sampling data with application to hiv prevalence estimation. *J. Am. Stat. Assoc.* **106**. (doi:10.1198/jasa.2011.ap09475)

[14] Gile KJ, Handcock MS. 2011 Network model-assisted inference from respondent-driven sampling data. *arXiv preprint arXiv:1108.0298*

[15] Lu X, Malmros J, Liljeros F, Britton T. 2013 Respondent-driven sampling on directed networks. *Electron. J. Stat.* **7**, 292–322. (doi: doi:10.1214/13-EJS772)

[16] Malmros J, Masuda N, Britton T. 2013 Random walks on directed networks: Inference and respondent-driven sampling. *arXiv preprint arXiv:1308.3600*

[17] Malekinejad M, Johnston LG, Kendall C, Franco Sansigolo Kerr LR, Rifkin MR, Rutherford GW. 2008 Using respondent-driven sampling methodology for HIV biological and behavioral surveillance in international settings: A systematic review. *AIDS Behav* **12**, 105–130. (doi: 10.1007/s10461-008-9421-1)

[18] Stein ML, van Steenbergen JE, Chanyasanha C, Tipayamongkholgul M, Buskens V, van der Heijden PGM, Sabaiwan W, Bengtsson L, Lu X, Thorson AE, *et al.*. 2014 Online Respondent-Driven Sampling for Studying Contact Patterns Relevant for the Spread of Close-Contact Pathogens: A Pilot Study in Thailand. *PLoS ONE* **9**. (doi: 10.1371/journal.pone.0085256)

[19] Andersson H, Britton T. 2000 *Stochastic epidemic models and their statistical analysis.* vol. 151. New York: Springer. (doi:10.1007/978-1-4612-1158-7)

[20] Molloy M, Reed B. 1995 A critical-point for random graphs with a given degree sequence. *Random Struct. Algor.* **6**, 161–179. (doi: 10.1002/rsa.3240060204)

[21] Molloy M, Reed B. 1998 The size of the giant component of a random graph with a given degree sequence. *Comb. Probab. Comput.* **7**, 295–305. (doi:10.1017/S0963548398003526)

[22] Ball F, Lyne OD. 2001 Stochastic multi-type SIR epidemics among a population partitioned into households. *Adv. Appl. Probab.* **33**, 99–123. (doi:10.1239/aap/999187899)

[23] Ball F, Neal P. 2002 A general model for stochastic SIR epidemics with two levels of mixing. *Math. Biosci.* **180**, 73–102. (doi:10.1016/s0025-5564(02)00125-6)

[24] Martin-Löf A. 1986 Symmetric sampling procedures, general epidemic processes and their threshold limit theorems. *J. Appl. Probab.* **23**, 265–282. (doi:10.2307/3214172)

[25] van der Hofstad R. 2009 Random graphs and complex networks. *Available on http://www.win.tue.nl/rhofstad/NotesRGCN.pdf*

[26] Britton T, Janson S, Martin-Löf A. 2007 Graphs with specified degree distributions, simple epidemics, and local vaccination strategies. *Adv. Appl. Probab.* **39**, 922–948. (doi:10.1239/aap/1198177233)

[27] Athreya K, Ney P. 1972 *Branching Processes*. Grundlehren der mathematischen Wissenschaften. Berlin: Springer. (doi: 10.1007/springerreference_60215)

[28] Ball FG, Sirl DJ, Trapman P. 2014 Epidemics on random intersection graphs. *Ann. Appl. Probab.* **24**, 1081–1128. (doi:10.1214/13-aap942)

[29] Ball F, Sirl D. 2013 Acquaintance vaccination in an epidemic on a random graph with specified degree distribution. *J. Appl. Probab.* **50**, 1147–1168. (doi:10.1239/jap/1389370105)

[30] Barbour AD, Reinert G. 2013 Approximating the epidemic curve. *Electron. J. Probab* **18**, 1–30. (doi:10.1214/ejp.v18-2557)

[31] Newman ME. 2002 Spread of epidemic disease on networks. *Phys. Rev. E* **66**, 016 128. (doi:10.1103/physreve.66.016128)

[32] Britton T, Deijfen M, Martin-Löf A. 2006 Generating simple random graphs with prescribed degree distribution. *J. Stat. Phys.* **124**, 1377–1397. (doi:10.1007/s10955-006-9168-x)

[33] Csardi G, Nepusz T. 2006 The igraph software package for complex network research. *InterJournal* **Complex Systems**, 1695

[34] Ball F, Sirl D, Trapman P. 2009 Threshold behaviour and final outcome of an epidemic on a random network with household structure. *Adv. Appl. Probab.* **41**, 765–796. (doi:10.1239/aap/1253281063)

[35] Wylie JL, Jolly AM. 2013 Understanding recruitment: outcomes associated with alternate methods for seed selection in respondent driven sampling. *BMC Med. Res. Methodol.* **13**. (doi:10.1186/1471-2288-13-93)