J. Appl. Prob. **53,** 518–530 (2016)
*doi:*10.1017/jpr.2016.17
© *Applied Probability Trust* 2016

RESPONDENT-DRIVEN SAMPLING AND AN UNUSUAL EPIDEMIC

J. MALMROS, * **
F. LILJEROS *** AND
T. BRITTON,* **** *Stockholm University*

Abstract

Respondent-driven sampling (RDS) is frequently used when sampling from hidden populations. In RDS, sampled individuals pass on participation coupons to at most c of their acquaintances in the community ($c = 3$ being a common choice). If these individuals choose to participate, they in turn pass coupons on to their acquaintances, and so on. The process of recruiting is shown to behave like a new Reed–Frost-type network epidemic, in which 'becoming infected' corresponds to study participation. We calculate R_0, the probability of a major 'outbreak', and the relative size of a major outbreak for $c < \infty$ in the limit of infinite population size and compare to the standard Reed–Frost epidemic. Our results indicate that c should often be chosen larger than in current practice.

Keywords: Stochastic epidemic model; respondent-driven sampling; configuration model; Reed–Frost

2010 Mathematics Subject Classification: Primary 92D30
Secondary 91D30

1. Introduction

Respondent-driven sampling (RDS) [15], [16] is commonly used when sampling from hidden populations, i.e. populations for which there are no sampling frames. RDS utilizes the social network of the population through a peer-driven recruitment mechanism in which previously sampled individuals recruit their neighbours in the population. An RDS study begins with the selection of a group of seed individuals. Each seed is provided with a fixed 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. 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. Recently, online-based RDS (webRDS), where recruitment takes place via email and a survey is filled out at a designated web site, have also been put into use [9], [26], [27]. 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 [13], [14], [17], [23], [25]; this is not the focus of this paper.

Received 6 January 2015; revision received 1 July 2015.
* Postal address: Department of Mathematics, Stockholm University, SE-106 91 Stockholm, Sweden.
** Email address: jensm@math.su.se
*** Postal address: Department of Sociology, Stockholm University, SE-106 91 Stockholm, Sweden.
Email address: fredrik.liljeros@sociology.su.se
**** Email address: tomb@math.su.se

https://doi.org/10.1017/jpr.2016.17 Published online by Cambridge University Press

The RDS recruitment process behaves similarly to an epidemic on a social network, where the disease corresponds to 'participation in the study' and spreads by the RDS coupon distribution mechanism. 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, where c corresponds to the number of coupons in the RDS study. In this paper we define an epidemic model for RDS on a configuration model network [20], [21] by a Reed–Frost-type mechanism, in which an upper limit on the number of individuals that an infectious individual could infect is imposed. We will use both infectious disease terminology and RDS terminology when referring to this model. In the original Reed–Frost model, see, e.g. [1, pp. 11–18], each infectious individual in generation i 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. The model is 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 may only spread the disease to his or her contacts in the social network. In our model, we then impose the additional restriction that an infectious individual only may spread the disease to at most c contacts. Previous models for RDS utilize a random walk approximation of the recruitment process in order to estimate population properties when the random walk is in stationarity. Because equilibrium is not reached in the present model, it is not useful for such inference. However, because respondents may recruit multiple contacts in our model, it is possible to examine the recruitment capability of RDS. A practical motivation to study our model is to explain the frequently observed inability of the RDS recruitment process to reach the desired sample size due to premature failure of the recruitment chains started by the seeds [18], and especially its relation to the number of coupons.

For this model, in the limit of infinite population size, we calculate the *basic reproduction number*, i.e. the number of individuals that are given coupons by a typical respondent during the early stages of RDS recruitment. This is often denoted by R_0. We say that there is a *major outbreak* if a nonnegligible proportion of the population become recruited (i.e. receive a coupon) 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 the case of a major outbreak, i.e. the asymptotic final proportion of the population that become recruited, using so-called *susceptibility sets* [3], [4]. We denote this by z. Note that τ and z are positive only if R_0 is larger than 1. We compare the RDS recruitment process to corresponding epidemics with unrestricted spread, i.e. standard Reed–Frost epidemics, for which $c = \infty$, and investigate the effect of varying the number of coupons c and the coupon transfer probability. To the best of the authors' knowledge, there are no previous studies of epidemics on networks that describe behaviour similar to the present one, although the model in [19] 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).

This paper is organised as follows. In Sections 2.1 and 2.2, we define the configuration model and the epidemic model, respectively. In Sections 3.1, 3.2, and 3.3, we calculate R_0, τ, and z, respectively. We illustrate our results by numerical calculations in Section 4, where we also evaluate our findings in finite populations by means of simulations. In Section 5 we describe an alternative recruitment process and some calculations for the corresponding epidemic model. Finally, our results are discussed in Section 6.

https://doi.org/10.1017/jpr.2016.17 Published online by Cambridge University Press

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 independent and identically distributed number of stubs (half-edges) d_i from a prescribed distribution D having support on the nonnegative 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 nth 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 [24, Theorem 7.12] and [11, Lemma 5.3]. Hence, we can condition on the graph being simple given that $\mathbb{E}(D^2) < \infty$. 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 [24, Theorem 7.10]. Hence, we will from now on assume that the resulting graph is simple. Moreover, the graph is locally tree-like when $\mathbb{E}(D^2) < \infty$, meaning that with high probability it does not contain short cycles [11]. Hence, we can take advantage of the branching process (e.g. [2]) approximations that are often used for epidemics (e.g. [1, Chapter 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 [11].

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

https://doi.org/10.1017/jpr.2016.17 Published online by Cambridge University Press

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 $\mathbb{P}(D = k) = p_k$. The degree of a given neighbour of an individual follows the *size-biased* degree distribution $\widetilde{D}$, where $\mathbb{P}(\widetilde{D} = k) = \widetilde{p}_k = kp_k/\mathbb{E}(D)$. Assume that we have an epidemic spreading on this graph according to the description in Section 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 $\widetilde{D}$. As previously mentioned in Section 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. To make this approximation fully rigorous, a coupling argument is used to connect (sequences of) the epidemic process with the approximating branching process; see, e.g. [1] and [6]. Let X and $\widetilde{X}$ be generic offspring random variables for 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 happen independently with the same probability p, we have, conditionally on the degree, the probability that the index case infects j neighbours is

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

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, analogously, we have

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

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

We obtain the distributions of X and $\widetilde{X}$ by conditioning on the degree. From standard branching process theory [2], it follows that R_0 is the expected number of individuals that get infected by an infectious individual in the second and later generations, i.e. $R_0 = \mathbb{E}(\widetilde{X})$. By this, we have the following theorem.

Theorem 1. *For the RDS epidemic model on the configuration model graph,*

$$R_0 = \sum_{j=0}^{c} j \left(\sum_{k=j}^{c-1} \binom{k}{j} p^j (1-p)^{k-j} \widetilde{p}_{k+1} + \binom{c}{j} p^j (1-p)^{c-j} \left(1 - \sum_{k=1}^{c} \widetilde{p}_k \right) \right).$$

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 $R_0^{(\text{unrestricted})} = p(\mathbb{E}(D) + (\text{var}(D) - \mathbb{E}(D))/\mathbb{E}(D))$ is the R_0 value for the standard Reed–Frost epidemic on a configuration model network; see, e.g. [11].

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,

https://doi.org/10.1017/jpr.2016.17 Published online by Cambridge University Press

which we obtain by standard techniques. We first consider a branching process with offspring distribution $\widetilde{X}$ for all individuals, i.e. also for the index case. Let the extinction probability of this process be $\widetilde{\pi}$. Standard branching process theory says that $\widetilde{\pi}$ is the smallest solution in $[0, 1)$ of $\widetilde{\pi} = \widetilde{\rho}(\widetilde{\pi})$, where $\widetilde{\rho}$ is the probability generating function of $\widetilde{X}$. This solution is obtained numerically. In our original branching process the ancestor has offspring distribution X and later generations have offspring distribution $\widetilde{X}$. Let the extinction probability of the original process be π. Analogously, it follows that π is given by $\pi = \rho(\widetilde{\pi})$, where ρ is the probability generating function of X. We obtain the following theorem.

Theorem 2. *The probability of a major outbreak is $\tau = 1 - \pi$, where π is given by $\pi = \rho(\widetilde{\pi})$ and $\widetilde{\pi}$ is the solution of $\widetilde{\pi} = \widetilde{\rho}(\widetilde{\pi})$.*

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

3.3. Relative size of a major outbreak

The relative size of a major outbreak in the case of a major outbreak z can be obtained using susceptibility sets, constructed as follows. For each individual i, we obtain a random list of neighbours that i would infect if it were to be infected itself. 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 has 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 $\text{bin}(d, \theta)$-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 $\widetilde{D}$, so we obtain $\theta = \sum_{k=0}^{\infty} \theta_k \widetilde{p}_k$, 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

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

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

https://doi.org/10.1017/jpr.2016.17 Published online by Cambridge University Press

make 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

$$\mathbb{P}(D^* = k) = \mathbb{P}(\widetilde{D} = k \mid C) = \frac{\mathbb{P}(C \mid \widetilde{D} = k)\mathbb{P}(\widetilde{D} = k)}{\sum_{k=0}^{\infty} \mathbb{P}(C \mid \widetilde{D} = k)\mathbb{P}(\widetilde{D} = k)} = \frac{\theta_k \widetilde{p}_k}{\theta},$$

so

$$\mathbb{P}(D^* = k) = \begin{cases} 0, & k = 0, 1, \\ \dfrac{p\widetilde{p}_k}{\theta}, & k = 2, \ldots, c+1, \\ \dfrac{cp\widetilde{p}_k}{(k-1)\theta}, & k = c+2, c+3, \ldots. \end{cases}$$

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.

Let Y and Y^* be generic offspring random variables for the ancestor and for individuals in later generations in the approximating branching process, respectively. Hence, $Y \sim \text{mix bin}(D, \theta)$ and $Y^* \sim \text{mix bin}(D^* - 1, \theta)$. We obtain the survival probability of the process similarly as in Section 3.2. Let the extinction probability of a branching process with offspring distribution Y^* be π^*. We have

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

The solution to (2) for π^* is obtained numerically. Let the extinction probability of the approximating branching process be π'. Analogously, we obtain $\pi' = f_D(1 - \theta + \theta\pi^*)$, where $f_D(\cdot)$ is the probability generating function of D and π^* is the solution to (2).

If the backward branching process contains few individuals, it is unlikely that i_0 will be infected in the epidemic. 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, which is in fact given by the survival probability of the branching process. Because i_0 is chosen randomly, we have $z = 1 - \pi'$. These results have been proved for SIR epidemics with general infectious period distribution on random intersection graphs [7] and on configuration model graphs [5]. We expect that similar arguments could also be used for our process in order to provide a formal proof. Additionally, we believe that the techniques described in [8] could be used to obtain stronger results for the whole epidemic process. This motivates the following theorem, where T_n denotes the final number of infected individuals in the epidemic.

https://doi.org/10.1017/jpr.2016.17 Published online by Cambridge University Press

Theorem 3. *If $R_0 \leq 1$ then $T_n/n \xrightarrow{\mathbb{P}} 0$. If $R_0 > 1$ then T_n/n converges in probability to a random variable T for which $\mathbb{P}(T = 0) = 1 - \tau$ and $\mathbb{P}(T = z) = \tau$, where $z = 1 - \pi'$ and τ is defined in Theorem 2.*

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; see, e.g. [22]. Given a degree distribution and c, we let p_c be the smallest value of p for which the process is above the epidemic threshold.

In Figure 1 we show the R_0 values for the RDS recruitment process with $c = 3, 5, 10$ and the unrestricted epidemic for $p \in [0, 1]$. In Figure 1(a) we show the results for the Poisson degree distribution with parameter $\lambda = 8$ and in Figure 1(b) the results for the power-law degree distribution with parameters $\alpha = 1.6$ and $\kappa = 220$, which has mean 8.00 and variance 721.10. For both degree distributions and a fixed value of p, the limitation imposed by c yields smaller R_0 values when compared to the unrestricted epidemic. Especially for the power-law degree distribution, the R_0 values are much smaller than those of the unrestricted epidemic, and p_c is larger than that of the unrestricted epidemic for all c.

In Figure 2 we show the values of τ and z for the RDS recruitment process with $c = 3, 5, 10$ and the unrestricted epidemic for $p \in [0, 1]$. In Figures 2(a) and 2(b) we show the results for τ and z, respectively, for the Poisson degree distribution with parameter $\lambda = 8$ and in Figures 2(c) and 2(d) the results for τ and z, respectively, for the power-law degree distribution with parameters $\alpha = 1.6$ and $\kappa = 220$. We see that z is always smaller than τ for both degree distributions. The τ values for the RDS recruitment process are smaller than those of the unrestricted epidemic for all c. When $p \to 1$, the τ values for the RDS recruitment process

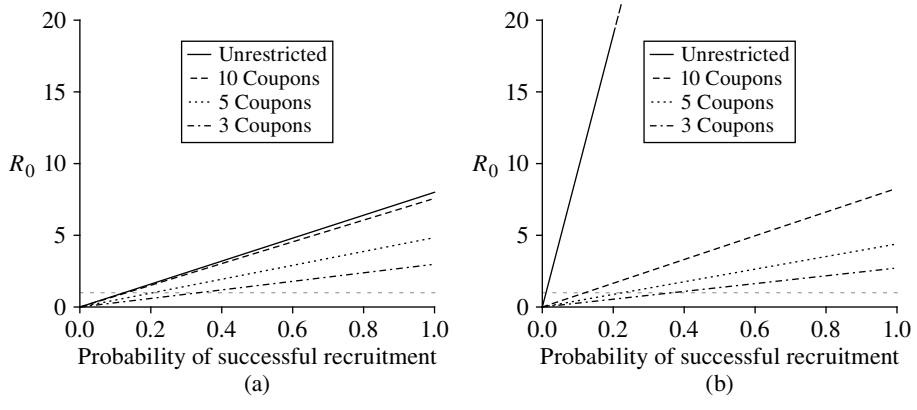


FIGURE 1: Comparison of R_0 for unrestricted epidemics and RDS recruitment processes with 10, 5, and 3 coupons and $p \in [0, 1]$. (a) The results for the Poisson degree distribution with parameter $\lambda = 8$ and (b) the results for the power-law degree distribution with parameters $\alpha = 1.6$ and $\kappa = 220$. The dashed horizontal lines show the threshold value $R_0 = 1$. In (b), the R_0 values for the unrestricted epidemic follow the indicated line outside of the vertical axis, with $R_0 \approx 94.8$ at $p = 1$.

https://doi.org/10.1017/jpr.2016.17 Published online by Cambridge University Press

approach the upper limit given by the unrestricted epidemic since almost all recruitment chains survive in this case. The z values are much smaller than those of the unrestricted epidemic for all c and p for the power-law degree distribution, whereas in the Poisson case, z is close to τ, except for when $c = 3$, where it is visibly smaller.

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 τ and z by the relative proportion of major outbreaks and the mean relative size of major outbreaks, respectively. Each simulation run consists of generating a network of size 5000 by an erased configuration model approach [10], for which we make use of the iGraph R package [12]. Then, RDS recruitment processes are run on the generated network for $p \in [p_c, 1]$. In Figure 3 we show the resulting estimates $\hat{\tau}$ of τ and $\hat{z}$ of z, and the corresponding asymptotic results. In Figure 3(a) we show the results for the Poisson degree distribution with parameter $\lambda = 8$ from 5000 simulations runs of RDS recruitment processes with three coupons and in Figure 3(b) 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. We use different parameter values for the power-law degree distribution from before in order for

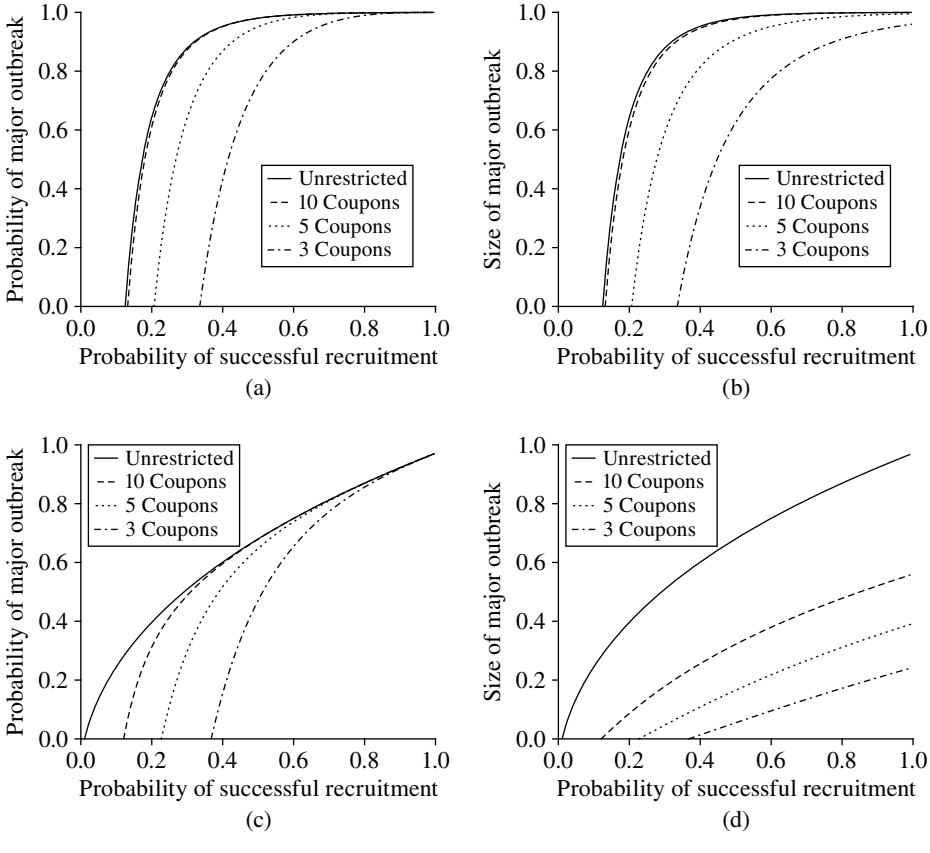


FIGURE 2: Comparison of τ and z for unrestricted epidemics and RDS recruitment processes with 10, 5, and 3 coupons and $p \in [0, 1]$. In (a) and (b) the results are shown for the Poisson degree distribution with parameter $\lambda = 8$ and (c) and (d) the results for the power-law degree distribution with parameters $\alpha = 1.6$ and $\kappa = 220$.

https://doi.org/10.1017/jpr.2016.17 Published online by Cambridge University Press

the degrees of vertices in the generated (finite) networks to be less affected by erasures when forming the network.

In Figures 3(c) and 3(d) we show normalized estimates $\hat{z}/z$ and $\hat{\tau}/\tau$ with normalized error bars based on ± 2 standard errors, for the results shown in Figures 3(a) and 3(b), respectively. The standard error (se) for $\hat{\tau}$ is estimated as $\text{se}(\hat{\tau}) = (\hat{\tau}(1-\hat{\tau})/m))^{1/2}$, where m is the number of simulations, and the standard error for $\hat{z}$ is estimated as $\text{se}(\hat{z}) = (\hat{\sigma}^2/m_{\text{maj}})^{1/2}$, 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. Two adjacent values (starting from the left) correspond to the estimates at one value of p. In general, $\hat{z}$ and $\hat{\tau}$ are well approximated by the asymptotic values with smaller differences for large p for both degree distributions. For the Poisson degree distribution, $\hat{\tau}$ and $\hat{z}$ are close to the asymptotic values for $p > 0.4$ and the normalized estimates show relatively small variation. For the Power-law degree distribution, the variation of the normalized estimates is larger and $\hat{z}$ is relatively close to z for $p > 0.5$ while τ is slightly underestimated for most p. As pointed out in [6], z is more efficiently estimated

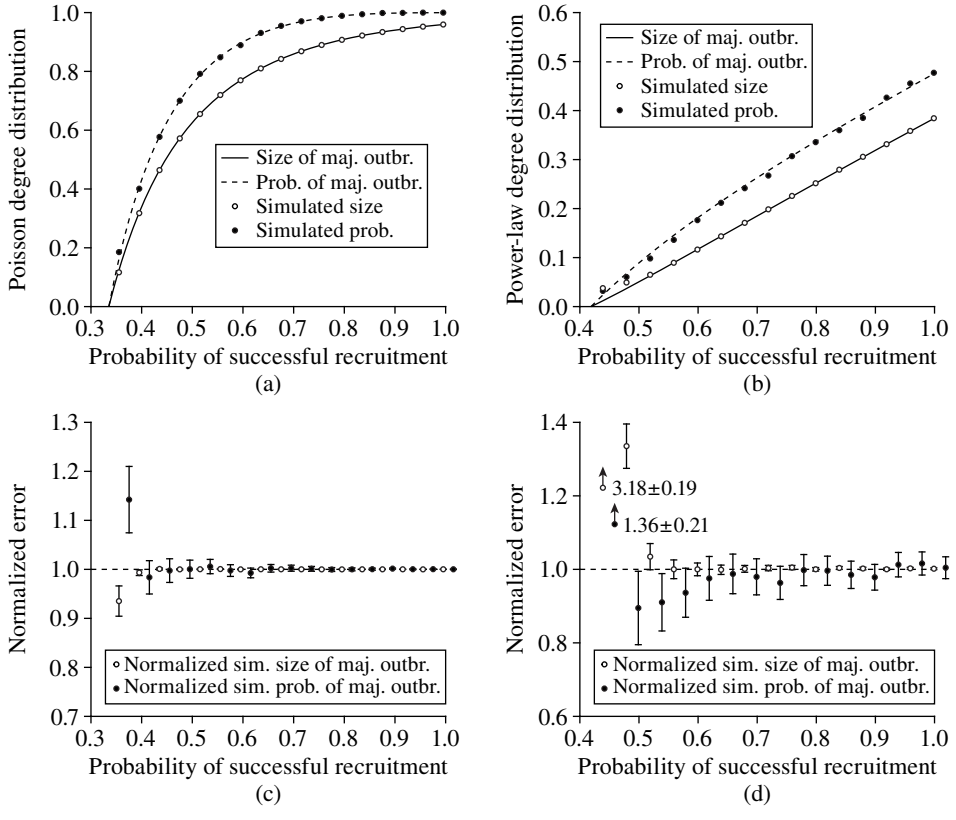


FIGURE 3: Comparison of estimated values from 5000 simulations of RDS recruitment processes and asymptotic values of τ and z. (a) The results for the Poisson degree distribution with parameter $\lambda = 8$ for processes with $c = 3$ and (b) the power-law degree distribution with parameters $\alpha = 2.5$ and $\kappa = 50$ for processes with $c = 10$. In (c) and (d), we show normalized estimates and error bars for the results shown in (a) and (b), respectively. In (d), some values are much larger than the others and are shown as numbers. Note that the horizontal scales are different.

https://doi.org/10.1017/jpr.2016.17 Published online by Cambridge University Press

than τ because each simulation yields many (correlated) observations of the backward process and only one observation of the forward process.

Note that it is not well defined as to 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, 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 reasonable estimates for p close to p_c and does not affect estimates for p further away from p_c.

5. Modified recruitment process

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. This modified mechanism is modelled as follows. A respondent first attempts to give a coupon to a randomly chosen neighbour. The coupon may be rejected either because the neighbouring individual already has received a coupon, which happens with small probability during the early stages of the epidemic in a large population, or because he or she refuses to participate. 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 (1) now become

$$\mathbb{P}(\widetilde{X} = j \mid \widetilde{D} = k) = \begin{cases} \binom{k-1}{j} p^j (1-p)^{k-1-j}, & j < c, \\ \sum_{i=c}^{k-1} \binom{k-1}{i} p^i (1-p)^{k-1-i}, & j = c. \end{cases}$$

It is straightforward to calculate R_0 and τ using the same techniques as in Sections 3.1 and 3.2. In Figure 4 we show 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 $\alpha = 1.6$ and $\kappa = 220$. It is clear that the R_0 values are larger than the corresponding values shown in Figure 1 for $c = 3, 5, 10$ and that p_c is very close to the value of p corresponding to the epidemic threshold for the unrestricted epidemic for both degree distributions and for all c. When $p \to 1$, the R_0 values converge to those seen in Figure 1. For $c = 3, 5, 10$, the τ values for the modified RDS recruitment process and the unrestricted epidemic are very close for both degree distributions and all p and are therefore not shown. Because individuals can get infected by neighbours even if they do not belong to

https://doi.org/10.1017/jpr.2016.17 Published online by Cambridge University Press

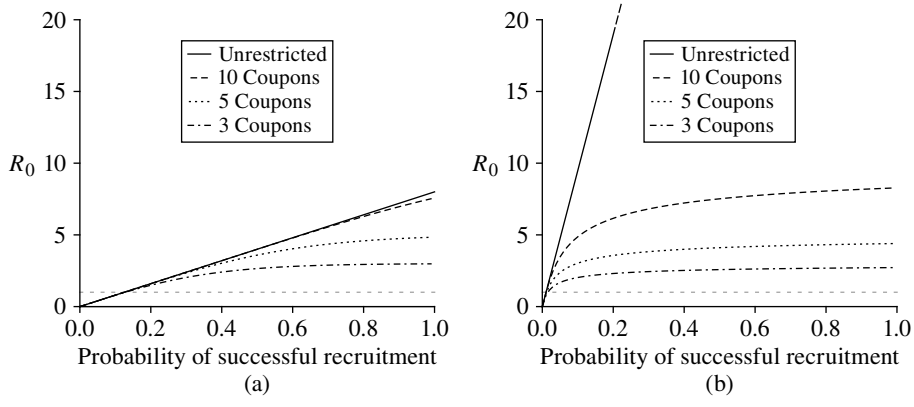


FIGURE 4: Comparison of R_0 for unrestricted epidemics and modified RDS recruitment process. In (a) we show the results for the Poisson degree distribution with parameter $\lambda = 8$ and (b) show the results for the power-law degree distribution with parameters $\alpha = 1.6$ and $\kappa = 220$. The dashed horizontal lines and R_0 values for the unrestricted epidemic in (b) are the same as in Figure 1.

the initial group of coupon receivers, the susceptibility set can not be derived as in Section 3.3. In particular, the corresponding quantity for θ_k will depend on in what stage of the epidemic a neighbour is infected. Later in the epidemic, it is more likely that an infected neighbour experience coupon rejections from previous participants; hence, a susceptible individual is at greater risk of receiving a coupon from an infected neighbour at this stage. As a consequence, different techniques must be used when deriving the final size.

6. Discussion

In this paper we have analysed RDS using a Reed–Frost-type epidemic model on the configuration model graph. The model incorporates an upper limit c on the number of neighbours that a respondent can distribute coupons to, which affect R_0, τ, and z. The proposed model best resembles online-based RDS, since respondents select possible coupon recipients once. We also presented a modified model, where respondents may attempt to distribute a coupon to several recipients, for which we calculated R_0 and τ. Our numerical calculations show that the effect of a finite c on the quantities considered can be large when compared to the unrestricted case, especially for heavy-tailed degree distributions, which may explain problems encountered in actual RDS studies.

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, p_c, τ, and z. This is especially true for the power-law degree distribution, for which in particular R_0 and z are 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 receiving a coupon will be small. Also, such an individual will, after participation, have few or no contacts to distribute coupons to. Hence, the spread of RDS in such networks will be highly dependent on a few individuals with very large degrees that have the capacity to distribute coupons to many of their (small degree) neighbours. Because of the relatively small value of c, the potential of large degree individuals to effectively distribute coupons is much smaller in RDS compared to an unrestricted epidemic with the same p, hence impairing the spread of the RDS recruitment process.

https://doi.org/10.1017/jpr.2016.17 Published online by Cambridge University Press

In practice, many RDS studies experience difficulties with reaching the desired sample size due to early failure of the recruitment chains started by the seeds. In webRDS, where there typically is no given sample size, this corresponds to not recruiting a substantial proportion of the population before recruitment stops. Because an increase in c results in increased R_0, τ, and z for a fixed p, our results indicate that (web)RDS studies that suffer from the previously mentioned difficulties may benefit from an increased number of coupons. Additionally, this may have other positive effects such as improved estimation due to prolonged recruitment chains and a shorter and more predictable study time.

In the standard Reed–Frost model on the configuration model graph, $\tau = z$, since the forward and backward branching processes coincide in this case. Because the restriction imposed by a finite c affects the forward and backward branching processes rather differently, this does not hold for our model when $c < \infty$. In this case, the forward process essentially behaves like in the unrestricted case except with fewer options to spread RDS. For the backward process however, both the contact probability and the degree distribution of individuals in later generations change when $c < \infty$, which makes it quite different from the forward process in this case.

The presented model for RDS 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. Possible extensions of the model are, e.g. different recruitment probabilities through unequally weighted edges, controlling for network structural properties, e.g. clustering, and modifying the coupon distribution mechanism by, e.g. introducing time dependence or letting the behaviour of respondents depend on individual characteristics. It could also be of interest to investigate the final size of the epidemic given by the modified recruitment process described in Section 5.

Acknowledgements

J.M. was supported by the Swedish Research Council (grant no. 2009-5759). T.B. and F.L. are grateful to Riksbankens jubileumsfond (contract P12-0705:1) for financial support. The authors are grateful to the anonymous referee for the careful reading and the constructive comments which greatly improved on the manuscript.

References

[1] Andersson, H. and Britton, T. (2000). *Stochastic Epidemic Models and Their Statistical Analysis* (Lecture Notes Statist. **151**). Springer, New York.
[2] Athreya, K. B. and Ney, P. E. (1972). *Branching Processes*. Springer, New York.
[3] Ball, F. and Lyne, O. D. (2001). Stochastic multitype SIR epidemics among a population partitioned into households. *Adv. Appl. Prob.* **33,** 99–123.
[4] Ball, F. and Neal, P. (2002). A general model for stochastic SIR epidemics with two levels of mixing. *Math. Biosci.* **180,** 73–102.
[5] Ball, F. and Sirl, D. (2013). Acquaintance vaccination in an epidemic on a random graph with specified degree distribution. *J. Appl. Prob.* **50,** 1147–1168. (Correction: **52** (2015), 908.)
[6] Ball, F., Sirl, D. and Trapman, P. (2009). Threshold behaviour and final outcome of an epidemic on a random network with household structure. *Adv. Appl. Prob.* **41,** 765–796.
[7] Ball, F. G., Sirl, D. J. and Trapman, P. (2014). Epidemics on random intersection graphs. *Ann. Appl. Prob.* **24,** 1081–1128.
[8] Barbour, A. D. and Reinert, G. (2013). Approximating the epidemic curve. *Electron. J. Prob.* **18,** 30 pp.
[9] Bengtsson, L. *et al.* (2012). Implementation of web-based respondent-driven sampling among men who have sex with men in Vietnam. *PLoS ONE* **7,** e49417.
[10] Britton, T., Deijfen, M. and Martin-Löf, A. (2006). Generating simple random graphs with prescribed degree distribution. *J. Statist. Phys.* **124,** 1377–1397.
[11] Britton, T., Janson, S. and Martin-Löf, A. (2007). Graphs with specified degree distributions, simple epidemics, and local vaccination strategies. *Adv. Appl. Prob.* **39,** 922–948.

https://doi.org/10.1017/jpr.2016.17 Published online by Cambridge University Press

[12] Csardi, G. and Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal Complex Systems* 1695.
[13] Gile, K. J. (2011). Improved inference for respondent-driven sampling data with application to HIV prevalence estimation. *J. Amer. Statist. Assoc.* **106,** 135–146.
[14] Gile, K. J. and Handcock, M. S. (2015). Network model-assisted inference from respondent-driven sampling data. *J. R. Statist. Soc. A* **178,** 619–639.
[15] Heckathorn, D. D. (1997). Respondent-driven sampling: a new approach to the study of hidden populations. *Social Problems* **44,** 174–199.
[16] Heckathorn, D. D. (2002). Respondent-driven sampling II: deriving valid population estimates from chain-referral samples of hidden populations. *Social Problems* **49,** 11–34.
[17] Lu, X., Malmros, J., Liljeros, F. and Britton, T. (2013). Respondent-driven sampling on directed networks. *Electron. J. Statist.* **7,** 292–322.
[18] Malekinejad, M. *et al.* (2008). Using respondent-driven sampling methodology for HIV biological and behavioral surveillance in international settings: a systematic review. *AIDS Behavior* **12,** 105–130.
[19] Martin-Löf, A. (1986). Symmetric sampling procedures, general epidemic processes and their threshold limit theorems. *J. Appl. Prob.* **23,** 265–282.
[20] Molloy, M. and Reed, B. (1995). A critical point for random graphs with a given degree sequence. *Random Structures Algorithms* **6,** 161–179.
[21] Molloy, M. and Reed, B. (1998). The size of the giant component of a random graph with a given degree sequence. *Comb. Prob. Comput.* **7,** 295–305.
[22] Newman, M. E. J. (2002). Spread of epidemic disease on networks. *Phys. Rev. E (3)* **66,** 016128.
[23] Salganik, M. J. and Heckathorn, D. D. (2004). Sampling and estimation in hidden populations using respondent-driven sampling. *Sociological Methodol.* **34,** 193–239.
[24] Van der Hofstad, R. (2014). Random graphs and complex networks. Vol. I. Available at http://www.win.tue.nl/~hofstad/NotesRGCN.html.
[25] Volz, E. and Heckathorn, D. D. (2008). Probability based estimation theory for respondent driven sampling. *J. Official Statist.* **24,** 79–97.
[26] Wejnert, C. (2009). An empirical test of respondent-driven sampling: point estimates, variance, degree measures, and out-of-equilibrium data. *Sociological Methodol.* **39,** 73–116.
[27] Wejnert, C. and Heckathorn, D. D. (2008). Web-based network sampling: Efficiency and efficacy of respondent-driven sampling for online research. *Sociological Meth. Res.* **37,** 105–134.

https://doi.org/10.1017/jpr.2016.17 Published online by Cambridge University Press