Estimators for Respondent Driven Sampling

BACHELOR THESIS

Author:
Margriet Spoorenberg
5730988
Utrecht University

Supervised by:
Martin Bootsma
Utrecht University

January 12, 2019

Utrecht University

Abstract

When random sampling is not possible because of lack of a sampling frame, a different approach is needed to aquire the statistical information that is desired. Respondent driven sampling is a series of methods to sample hard to reach populations that don't provide a sampling frame. It might possible to find reasonable estimators that allow for statistical inference from the data that can be collected. Bias is one of the main problems that can occur and to find and correct for a these biases can be challenging.
In this thesis we derive two asymptotically correct estimators. One for the population prevalence and one for the homophily of the population. The difference in the degree between two parts of the population causes a bias that affects the estimates. In this thesis the coverage of the 95% confidence interval is usually below 95% for finite samples. However, this thesis does give a not too complicated example of an RDS-simulation that can be understood and adjusted.

Contents

1 Introduction

When researchers are studying a population, it is usualy impossible to make direct observations of every single individual. So instead, they try to take a representative subset of individuals from which they collect their data. With the information from this sample they make inferences about the entire population. But the population in question would have to be defined before a random sample can be drawn. For example by stating every individual in the population has a specific characteristic that is of interest to the researchers. To draw a sample, all individuals would have to be known. There would have to be a sampling frame, which is a list that consists of all the individuals of the population. In so called hidden populations there is no list because the individuals are unknown. Chain referral sampling does not need a sampling frame. It is used as a nonprobabilistic mean to investigate the population. The researcher finds some individuals from the population that are willing to participate. These participants then help the researcher find new recruits. The researcher then decides whether these recruits are included in the sample.

In 1979 Bonnie H. Erickson wrote a paper [1] in which she explained why chain referral sampling methods could not be used to make reliable estimates. Because chain referral sampling makes use of convenience samples, inference like from a random sample is not possible. The introduction of new chain referral sampling methods called Respondent Driven Sampling (RDS) led to the use of other estimators then the sampling mean. In RDS the social network of the sampled individuals is taken into account and population estimates are based on information about this network. The extra information that is collected is used to correct for bias in the sample when making inference about the population.

To investigate properties of hard to reach groups in a population we can use Respondent Driven Sampling (RDS) instead of random sampling. If initial samples are asked to recruit new samples themselves the proces will resemble a Markov Chain. The initial samples, which are not random, will, after some time, be of little influence to the result of the sampling. The aim of this thesis is to see how a bias, caused by differences in the network size of the participating groups, and homophily, the preference for a certain type of inidividuals to form ties with similar individuals, could influence the reliability of the estimator for the prevalence.

2 Theory

For clairity, an overview of the notation of all variables used in this thesis can be found in table 1.

Table 1: Notation used in this thesis. The subindices a and b refer to population a and population b.

Quantity	**variables**	**estimators**
Number of group members	(not applicable)	n_A, n_B
Recruitment count	(not applicable)	k_{aa}, k_{ab}, k_{ba}, k_{bb}
Population/prevalence	P_a, P_b	$\tilde{p}_a$, $\tilde{p}_b$
Selection proportion	p_{aa}, p_{ab}, p_{ba}, p_{bb}	$\tilde{p}_{aa}$, $\tilde{p}_{ab}$, $\tilde{p}_{ba}$, $\tilde{p}_{bb}$
Homophily	H	$\tilde{h}$
Mean Degree	N_a, N_b	$\tilde{n}_a$, $\tilde{n}_b$

2.1 Chain referral sampling

2.1.1 Prevalence

In epidemiology, prevalence is the proportion of a particular population that has a specific trait, like, for instance, an infection. A natural way to express prevalence is as a fraction. If a is a trait that the part of the population has and not having trait a means having trait b, then the prevalence P of trait a is;

$$P_a = \frac{\text{number of individuals with trait } a}{\text{number of individuals with trait } a + \text{number of individuals with trait } b}$$

which is the proportion of the population with trait a.

One way of estimating the prevalence is taking a random sample of the total population and making statistical inference about the proportion of the population that has the trait.
But suppose now that taking a random sample is impossible due to lack of a sampling frame. Without knowledge of the size and boundaries of a population, another way of estimating the prevalence is required. Another problem could be the unwillingness of members of the group to cooperate. Membership of the group can for instance imply stigmatizing or illegal traits which causes the members to protect their privacy [2]. Groups that have these characteristics are called hidden populations. Examples of such groups are drug addicts, prostitutes or, surprisingly, jazz-musicians, like Douglas D. Heckathorn and Joan Jeffri have described in their paper "Social Networks of Jazz Musicians" [3].

2.1.2 Seeds and Waves

If a random sample can not be taken from the population, a range of methods called Chain-Referral Sampling can be used to try to find an estimate for the prevalence. These methods don't require a sampling frame and are suitable for hard to reach or hidden populations [4]. The initial participants, recruits that can be reached and are willing to cooperate, give the researcher information that leads to new recruits inside the group. These initial participants are called 'seeds'. The seeds are not randomly chosen, but they are rather a convenience sample. These seeds then provide information that leeds to new recruits for the sample. Each new round of recruits is called a wave. For each member of the population to be reachable, the population must be connected by one or more links between individuals.

2.1.3 Bias

As mentioned before, the initial seeds may be biased, and not a random sample from the population. Response rates can be low and recruits may not speak out frankly because of the reasons mentioned above. The fact that they could be reached is itself a proof that they are unlike their unreachable peers. This bias will be difficult to estimate and possibly affect the bias in the following wave. Also variance can be influenced by the recruiting techniques. Analizing the results from the sample like a random sample will lead to population estimates that differ from the true parameter values of the population.

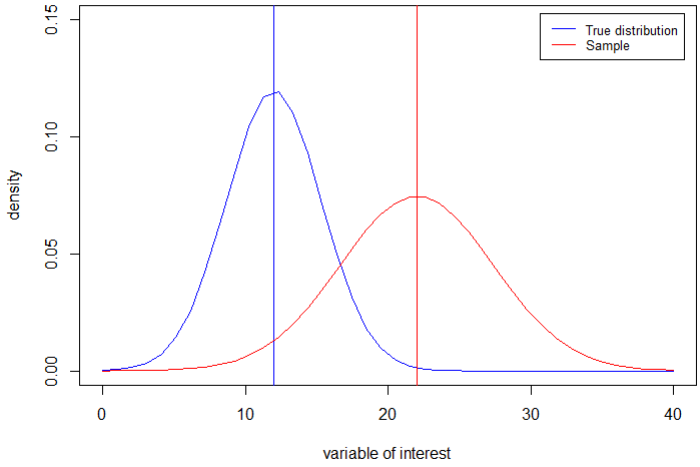


Figure 1: The difference between the blue and red vertical lines is the bias of the sample, the width of the distribution indicates a different variance between the distribution and the sample.

2.2 Respondent Driven Sampling

The shortcomings of chain-referral sampling have led to numerous improved methods for collecting and analyzing chain-referral data that are all referred to as Respondent-driven sampling (RDS). Although RDS is a family of methods, it will be considered one method for the rest of this thesis. If specifications are treated the source will be referred to.

2.2.1 Coupons

In contrast to chain-referral sampling, seeds are given a specified number of coupons that can be used to recruit new participants among their peers. The investigator does not collect information about their relations, like names for instance, but keeps track of the coupons. They are numbered and connected to the information of the participants. This way information about the social network is included in the sample. With this data, estimates about the network can be made and this information is then used to derive proportions of different groups. [5].
The number of coupons provided to each participant can vary from 1 to many. If responses are low or time is limited, having a high number of coupons per participant may be the best option. On the other hand, recruiting many peers from one recruit might not add significantly more information to the sample. If financial constraints limit the number of participants, many coupons per recruits will result in a low amount of waves. The sample might still contain the potential biases that were in the initial seeds. Few coupons per recruit on the other hand can allow for a more divers sample. The influence of the initial seeds can be smaller because to have a sample of the same size as in the many coupon case, with the same financial means, there should be more waves and therefore more recruits using their own personal network to recruit new participants. In theory, the sample could reach further into the population, giving less weight to the bias of the initial seeds. The results of a comparison of many versus few coupons is in chapter 4.5.

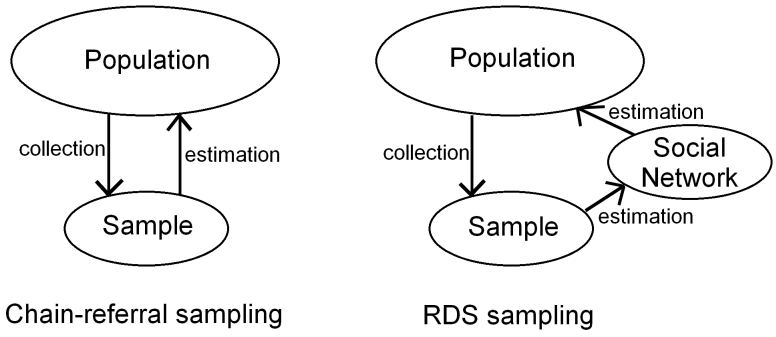


Figure 2: The difference between chain-referral and RDS sampling methods. In RDS sampling the characteristics of the network are taken into account [5].

2.2.2 Incentives

The general assumption is that the sample will be biased towards more cooperative members of the group. In RDS methods incentives are used to increase the response rate and the quality of the gathered information. A reward of some sort can be used to prevent the recruitment-chain from ending before the recuired amount of participants is met. In chain referral sampling an incentive can be given for participating in the survey. In RDS an additional incentive is given to the recruiter for recruiting a peer that participates. The later is more effective [6], [7].

2.2.3 Degree

There are rules for the type of relationship that counts as RDS relationship. The relationship between seed and new recruit has to be reciprocal. The seeds knows the new recruit and also knows the new recruit has the trait, and the new recruit knows the seed, and also knows the seed has the trait. To prevent misuse of the incentives every new recruit will have to be checked for the trait.
Each recruit reports the amount of relationships that comply with these rules. This amount is called the degree of the participant, n_i. We assume that the respondents can report their degree accurately. The new recruit will then become the new recruiter and select new recruits for the next wave. This will be repeated until the desired size of the sample is reached.

The average degree of a group is the mean of all degrees of the individuals. This average degree of, for example, group a is denoted by N_a, the mean degree of a. The number of individuals in group a is A.

$$N_a = \frac{1}{A} \sum_{i=1}^{A} n_i \tag{1}$$

The average degree of the participants will probably be higher than the average degree of the group. Members with larger networks will be oversampled and members with small networks might not be sampled at all. Two possible strategies to compensate for the oversampling are:

1. weighing the respondents inversely to their degree,
2. using extra incentives to favor recruits that have traits associated with having a low degree [2].

2.2.4 Markov process

We assume that every new recruit actualy participates, in other words, the chain does not break before the desired amount of waves is reached. Only the characteristics of the last recruiter are affecting the choice of the new recruits [8]. That makes the next probabilities dependent only on the last recruits. When a stochastic process meets the criterium

$$\mathbb{P}(X_n = x_n | X_{n-1} = x_{n-1}, X_{n-2} = x_{n-2}, ..., X_0 = x_0) \quad = \quad \mathbb{P}(X_n = x_n | X_{n-1} = x_{n-1})$$

it is called a Markov Process, a random process whose future probabilities are determined by its most recent values.
When there are two groups to choose from, there is a probability that a participant from group a recruits someone from group a. This probability we call p_{aa}. Otherwise someone from group b is recruited. These two probabilities sum up to one, so, $1 - p_{aa} = p_{ab}$. The same goes for a participant from group b. These probabilities form a probablity-matrix.

Table 2: The probability matrix with the selection proportions

group	***a***	***b***
a	p_{aa}	p_{ab}
b	p_{ba}	p_{bb}

We assume that non of the probabilities is zero. This means that any state in the process can re-occur. The process cannot become trapped inside one of the groups, because as stated earlier, all members are connected to all members. The process is therefore ergodic, which makes the recruitment process a regular Markov process. The chain will reach an equilibrium that is independent of the seeds.

2.2.5 Reciprocity

One of the assumptions for RDS is that, in our sample, all relations are reciprocal. So if T_{ab} is the number of ties from group a to group b than T_{ab} is equal to T_{ba}, the number of ties from group b to group a.

$$T_{ab} = T_{ba} \tag{2}$$

When a sample is taken the number of links from group a to group b is counted. This is most likely not going to be equal to the number of links from group b to group a, but for infinetly large samples this will be true. For finite samples this will only be an approximation and for small samples this assumption can not be used. The number of ties from group a to group b is product of the probability that a group a member will form a tie with a member of group b with the mean degree of group a weighted by the population proportion of group a. In the same manner the number of ties from group b to a is calculated. This provides a way to compare the variables of these two products,

$$T_{ab} = P_a N_a p_{ab} \qquad \text{and} \qquad T_{ba} = P_b N_b p_{ba} \tag{3}$$

so that;

$$P_a N_a p_{ab} = P_b N_b p_{ba}. \tag{4}$$

For large sample-size we can derive the prevalences. Because $P_a + P_b = 1$ we can solve the equation for P_a and P_b:

$$P_a = \frac{N_b p_{ba}}{N_a p_{ab} + N_b p_{ba}} \qquad \text{and} \qquad P_b = \frac{N_a p_{ab}}{N_a p_{ab} + N_b p_{ba}} \tag{5}$$

2.2.6 Homophily

Members of a group do not randomly choose their contacts among the entire population. There can be many reasons not to do so. Geographical distance or obstacles for example. Some people live closeby and others live on the other side of town, or the other side of a river. This can make some groups more isolated from the rest of the population. But most importantly, people may have preferences for connections with similar people.
Unfortunately, not all researchers on RDS agree on the definitions of homophily. An intuitive definition of homophily is given by David Easley and Jon Kleinberg [9]. Suppose there are two groups, a and b with proportions p and q respectively. Group members of a and b have the same average degree. When these two groups have connections, or links, the probability of a link being from group a to group a is p^2, and the probability from group b to b is q^2. If there is no homophily, then the expected fraction of cross-group relations should be close to twice the probability of having someone from group a having a link to someone from group b which is $2pq$. Or, as Easley and Kleinberg state in their homophily-test; If the fraction of cross-group relations is significantly less than $2pq$, then there is evidence for homophily.

In the approach by Douglas D. Heckathorn (2002) [8] he assumes that a fixed proportion of the contacts type a individuals are with individuals of type a because of homophily. The other 'free' contacts are randomly chosen from both groups. If p_{aa} is the selection probability that a contact from type a is with a type a individual, and H_a is the probability that selection is controlled by homophily, then p_{aa} is the sum of this probability and the probability that selection is not controlled by homophily weighted by the fraction of the population P_a.

$$p_{aa} = H_a + (1 - H_a)P_a \tag{6}$$

In Heckathorn's definition, a and b have different homophily. His definition does not take the number of contacts that the members have into account. But the chance of a contact being from group a to group a depends on the total number of contacts that are available, rather than the number of people being available. If one group has high mean degree and the other has low mean degree, contacts are more likely to be to the group with high degree.

If we agree that the chance of someone being recruited depends on the number of contacts that someone has, it would seem more logical that the selection probability depends on the avarage number of contacts of the group. So instead of choosing free contacts dependent of the fraction of the population P_a, they also depend on the mean degree of the group N_a. The prevalence is replaced by the fraction of contacts for group a or b.

$$p_{aa} = H_a + (1 - H_a)\frac{P_a N_a}{P_a N_a + P_b N_b}$$

With this definition, there is only one value for homophily that is the same for both groups, so $H_a = H_b = H$. How H depends on other variables can be derived from equation 7.

$$
\begin{aligned}
p_{aa} &= H + (1-H)\frac{P_a N_a}{P_a N_a + P_b N_b} \qquad (7)\\
p_{aa} - \frac{P_a N_a}{P_a N_a + P_b N_b} &= H\left(1 - \frac{P_a N_a}{P_a N_a + P_b N_b}\right)\\
H &= \frac{p_{aa} - \frac{P_a N_a}{P_a N_a + P_b N_b}}{(1 - \frac{P_a N_a}{P_a N_a + P_b N_b})}\\
&= \frac{(1-p_{ab}) - (1 - \frac{P_b N_b}{P_a N_a + P_b N_b})}{\frac{P_b N_b}{P_a N_a + P_b N_b}}\\
&= \frac{\frac{P_b N_b}{P_a N_a + P_b N_b} - p_{ab}}{\frac{P_b N_b}{P_a N_a + P_b N_b}}\\
&= 1 - \frac{p_{ab}(P_a N_a + P_b N_b)}{P_b N_b}
\end{aligned}
$$

Then fill the prevalences, P_a and P_b, as derived in equation 5.

$$
\begin{aligned}
H &= 1 - \frac{p_{ab}(\frac{N_b p_{ba}}{N_a p_{ab} + N_b p_{ba}} N_a + \frac{N_a p_{ab}}{N_a p_{ab} + N_b p_{ba}} N_b)}{\frac{N_a p_{ab}}{N_a p_{ab} + N_b p_{ba}} N_b}\\
&= 1 - \frac{p_{ab} N_b p_{ba} N_a + p_{ab} N_a p_{ab} N_b}{N_a N_b p_{ab}}\\
&= 1 - p_{ba} - p_{ab} \qquad (8)
\end{aligned}
$$

The degree just fell out of the equation. If we have random mixing then the chance of a link from a to a will be equal to the chance of a link between a and b, and likewise for links from group b to a and b. Then $p_{ba} + p_{ab} = 1$ and homophily would be zero, as expected, regardless of the degree.
The maximum value of H is 1. Then there are only connections within the population and no connections between the groups so $p_{aa} = 1$ and, because of reciprocity, also $p_{bb} = 1$. When homophily is zero, all contacts are chosen randomly from both groups. But members of group a might also prefer contacts outside their own group, which is referred to as heterophily. In the extreme case there are no in-group contacts, which means $p_{ab} = 1$ and $p_{ba} = 1$. Then $H = 1 - 1 - 1 = -1$.

2.2.7 Assumptions for RDS

Here follows a summary of the assumptions that are being used for estimators in RDS.

1. The population must be linked by their contacts. The ties must be reciprocal and the members must be able to identify each other as member of the population.

2. Because the all members of the population are linked, the minimal degree a participant can have is 1. Otherwise this individual cannot be reached and is not part of the linked population.

3. The population is infinite so the sample should be small compared to the size of the population. This reduces the risk of sampling the same recruit twice. In the estimation process, the possibility of recruiting the same participant twice is ignored.

4. Because of the previous assumption, new recruits can be drawn with replacement. That way the Markov process can be used.

5. Each respondent reports their number of contacts that meet assumption 1. This number is assumed to be accurate.

For this model we add the assumption that every coupon leads to a new recruit. The chain does not break. In practice this is almost impossible, but incentives can help.

3 Estimators

3.1 Recruitment count

The recruitment count is the number of times that an individual from group a recruited someone from group a etcetera. It is used to make the other estimates. This recruitment count, which is an estimator itself, is denoted by k_{aa}, k_{ab}, k_{ba} and k_{bb}.

3.2 Estimator for the selection proportion

The estimator for the selection proportions, $\tilde{p}_{aa}$, $\tilde{p}_{ab}$, $\tilde{p}_{ba}$ and $\tilde{p}_{bb}$, are derived from the recruitment count. For example, the selection proportion $\tilde{p}_{ab}$ is derived from k_{aa} and k_{ab};

$$\tilde{p}_{ab} = \frac{k_{ab}}{k_{aa} + k_{ab}}. \tag{9}$$

For large values of k_{aa} and k_{ab}, the selection proportion $\tilde{p}_{aa}$ is normally distributed with mean p_{aa}.

$$\tilde{p}_{aa} \sim N\left(p_{aa}, \frac{p_{aa}(1-p_{aa})}{k_{aa} + k_{ab}}\right) \tag{10}$$

3.3 Estimator for the mean-degree

As assumed, the respondents report their number of RSD-approved contacts, or their 'degree'. If we take the mean of the sample degree as our estimator, most likely our estimate will be to high because recruits with high degree tend to be oversampled [1]. The recruits are chosen with a probability proportional to their degree. From the sample follows a sample degree distribution. These two ingredients provide information to estimate the population degree distribution, which can then be used to estimate the mean degree of the groups [5].
For two individuals of which one has twice as many contacts as the other, the chance for the first one to be in the sample is twice as large as for the second one to be sampled. The weight of the first individual is only half of the weight of the second individual. If A is the size of group a individuals in the sample, and n_i is the degree of individual i, the estimator for the mean degree of group a individuals is:

$$\tilde{n}_a = \frac{\sum_{i=1}^{A} \frac{1}{n_i} n_i}{\sum_{i=1}^{A} \frac{1}{n_i}} = \frac{A}{\sum_{i=1}^{A} \frac{1}{n_i}} \tag{11}$$

Because the harmonic mean is equivalent to applying the arithmetic mean to the reciprocals of the degrees, we will use $\tilde{\mu}_a = \frac{1}{\tilde{n}_a}$ to calculate the variance:

$$\tilde{\mu}_a = \frac{1}{A} \sum_{i=1}^{A} \frac{1}{n_i} \tag{12}$$

The sample variance for $\tilde{\mu}_a$ is $\bar{\sigma}_a^2$:

$$\bar{\sigma}_a^2 = \frac{1}{A}\sum_{i-1}^{A}\left(\frac{1}{n_i} - \mu_a\right)^2 \tag{13}$$

The variance of a ratio can be calculated using a Taylor approximation, and the variance of $\frac{1}{\tilde{\mu}_a}$ is approximately:

$$\widetilde{Var\left(\frac{1}{\mu_a}\right)} \approx \left(\frac{1}{\tilde{\mu}_a}\right)^2\left(\frac{Var(1)}{1^2} - 2\frac{cov(1,\mu_a)}{\tilde{\mu}_a^2} + \frac{Var(\mu_a)}{\tilde{\mu}_a^2}\right) = \frac{\bar{\sigma}_a^2}{\mu_a^4} \tag{14}$$

An unbiased estimator for the variance of the degree of the population is the corrected sample variance. When we assume that the sample values are statisticaly independent, then with the sample mean $\tilde{n}_a$, the unbiased estimate for the variance of the mean degree of group a is;

$$\widetilde{Var(n_a)} = \frac{1}{A-1}\frac{\bar{\sigma}_a^2}{\mu_a^4} \tag{15}$$

Because we assume that the population is infinite, if we take a large enough sample (much larger than 200), the mean degree is normaly distributed.

$$\tilde{n}_a \sim N\left(n_a \quad , \frac{1}{\mu_a^4}\frac{\sigma_a^2}{A}\right) \tag{16}$$

3.4 Estimation of the prevalence

The estimator for the prevalence, $\tilde{p}_a$ can be derived from the estimates of the selection proportion and the degree. If there is homophily in the population, then the number of in-group contacts tends to be oversampled. When estimating the prevalence, homophily should be taken into account to compensate this oversampling. Also the mean degree of each group is part of the estimator for the prevalence because the amount of ties that an indivudual has influences the probability of being chosen as recruit. If members of group a have a higher mean degree then members from group b then the chance of a link being from a to group a is greater than from group a to group b. Group a will be oversampled.

As mentioned in section 2.2.5, when a sample is large enough, we can assume that the number of recruitments from group a to b is equal to the number of recruitments from group b to a. When we use the selection proportions from the sample and the estimated mean degree, the estimate for the prevalence can be derived from this assumption.

$$\tilde{p}_a = \frac{\tilde{n}_b\tilde{p}_{ba}}{\tilde{n}_a\tilde{p}_{ab} + \tilde{n}_b\tilde{p}_{ba}} \tag{17}$$

If we want to calculate the error of variance of the estimator of the prevalence, $\tilde{p}_a$, we consider the function f of the variables $\tilde{n}_a, \tilde{n}_b, \tilde{p}_{ab}$ and $\tilde{p}_{ba}$.

$$f(\tilde{n}_a, \tilde{n}_b, \tilde{p}_{ab}, \tilde{p}_{ba}) = \frac{\tilde{n}_b\tilde{p}_{ba}}{\tilde{n}_a\tilde{p}_{ab} + \tilde{n}_b\tilde{p}_{ba}}$$

For the variance of $\tilde{p}_a$, we approximate the estimator in point $\theta = (\mu_1, \mu_2, \mu_3, \mu_4)$ and add a specific error to all of these variables. In the appendix is the full calculation of the variance of estimator $\tilde{p}_a$. The result of this calculation follows here;

$$Var(\tilde{p}_a) = \left(-\frac{\tilde{p}_a\tilde{p}_b}{\tilde{n}_a}\right)^2 Var(\epsilon_1) + \left(\frac{\tilde{p}_a\tilde{p}_b}{\tilde{n}_b}\right)^2 Var(\epsilon_2) + \left(-\frac{\tilde{p}_a\tilde{p}_b}{\tilde{p}_{ab}}\right)^2 Var(\epsilon_3) + \left(\frac{\tilde{p}_a\tilde{p}_b}{\tilde{p}_{ba}}\right)^2 Var(\epsilon_4) \tag{18}$$

So ϵ_1 and ϵ_2 are the errors that belong to the mean degree of group a and b respectively and ϵ_3 and ϵ_4 are the errors of $\tilde{p}_{ab}$ and $\tilde{p}_{ba}$. For the variances of the ϵ_1 and ϵ_2, the variances of the mean degrees of a and b are used. The variance of the mean degree is calculated in section 3.3 (see equation (16)). Therefore;

$$Var(\epsilon_1) = \frac{1}{\mu_a^4}\frac{\sigma_a^2}{A} \quad \text{and} \quad Var(\epsilon_2) = \frac{1}{\mu_b^4}\frac{\sigma_b^2}{B} \tag{19}$$

with σ_A and σ_B the standard deviation of the reciprocal of sample mean degree of a and B respectively. And A and B the number of group a and group b recruits.
To estimate ϵ_3 and ϵ_4 we look at the distribution of $\tilde{p}_{ab}$ and $\tilde{p}_{ba}$.

$$\tilde{p}_{ab} \sim N\left(\frac{k_{ab}}{k_{aa}+k_{ab}}, \frac{\tilde{p}_{aa}\tilde{p}_{ab}}{k_{aa}+k_{ab}}\right)$$

The error of the selection proportion of the sample ϵ_3 is normaly distributed with the same variance as $\tilde{p}_{ab}$.

$$Var(\epsilon_3) = \frac{\tilde{p}_{aa}\tilde{p}_{ab}}{k_{aa}+k_{ab}} \quad \text{and, likewise} \quad Var(\epsilon_4) = \frac{\tilde{p}_{ba}\tilde{p}_{bb}}{k_{ab}+k_{bb}} \tag{20}$$

3.5 Estimation of the homophily

From the data gathered in the sample and the estimates that are derived from that information, the estimation for the homophily is calculated as before. Equation (7) becomes;

$$\tilde{p}_{aa} = \tilde{h} + (1-\tilde{h})\frac{\tilde{p}_a\tilde{n}_a}{\tilde{p}_a\tilde{n}_a + \tilde{p}_b\tilde{n}_b} \tag{21}$$

and equation (8);

$$1 - \tilde{p}_{ba} - \tilde{p}_{ab}. \tag{22}$$

We create a function f of $\tilde{p}_{ba}$ and $\tilde{p}_{ab}$ to calculate the variance of $\tilde{h}$.

$$f(\tilde{p}_{ab}, \tilde{p}_{ba}) = 1 - \tilde{p}_{ab} - \tilde{p}_{ba}$$

We approximate estimator $\tilde{h}$ in point $\theta = (\mu_3, \mu_4)$.

$$f(\mu_3+\epsilon_3, \mu_4+\epsilon_4) = 1 - (\mu_3+\epsilon_3) - (\mu_4+\epsilon_4) \tag{23}$$

The variance of $\tilde{h}$ is then;

$$Var(\tilde{h}) = Var(1 - \mu_3 - \epsilon_3 - \mu_4 - \epsilon_4)$$

and because of independence

$$Var(\tilde{h}) = Var(1) + (-1)^2Var(\mu_3) + (-1)^2Var(\epsilon_3) + (-1)^2Var(\mu_4) + (-1)^2Var(\epsilon_4)$$

$$Var(\tilde{h}) = Var(\epsilon_3) + Var(\epsilon_4)$$

From the distribution of the selection proportion follows that the variance of the homophily is;

$$Var(\tilde{h}) = \frac{\tilde{p}_{aa}\tilde{p}_{ab}}{k_{aa}+k_{ab}} + \frac{\tilde{p}_{ba}\tilde{p}_{bb}}{k_{bb}+k_{ba}} \tag{24}$$

4 Simulation

4.1 The parameters

A simulation for a single coupon RDS sample was made in R to investigate the reliability of the estimators $\tilde{p}_a$ and $\tilde{h}$. The initial values for the amount of seeds n, the recuired number of waves m, the homophily H, the fraction a of the population P_a and the average degrees of group a, N_a, and b, N_b, are entered into the simulation program. With this input the rest of the necessary parameters like the fraction of members of b of the population, P_b can be calculated. Also the probabilities p_{aa}, p_{ab}, p_{ba} and p_{bb} for participants from group a and b choosing a new recruit to fill the probability matrix (see table 3) for the Markov chain are calculated according to equation (7).

4.2 The mean degree of the sample

To estimate the prevalence, the degree of the recruits has to be simulated as well. This can be done separately from the recruit-simulation.

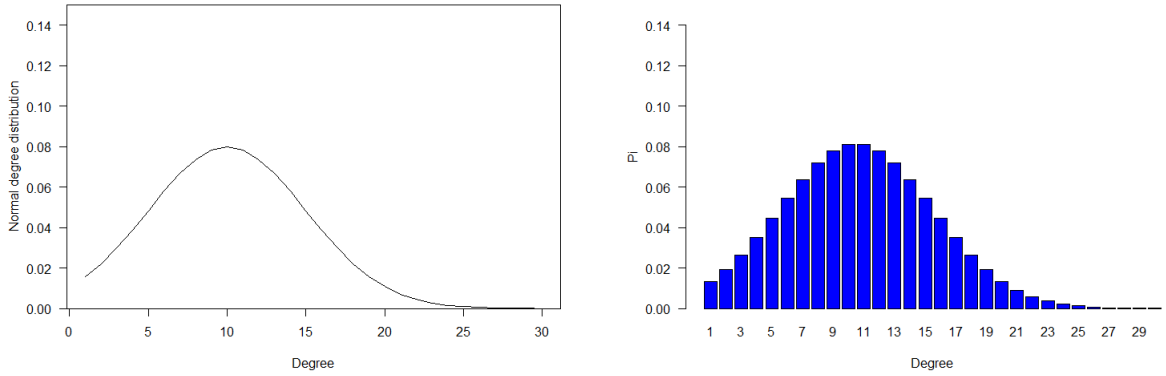
A normal distribution (Figure 3(a)) is generated for each group with standard deviation of one-half of the degree of this group, and mean equal to the mean degree of this group.

$$\text{Normal degree distribution of group } i \sim N\left(N_i, \frac{N_i}{2}\right) \tag{25}$$

Maximum number of contacts is set on three times the mean degree of the group. There is no particular theory behind these choises, other than these seem like reasonable assumptions. The distribution is then discretized into Pi (Figure 3(b)) so that every integer between 0 and $3N_i$ has a probability assigned to it. Normalisation makes sure the sum of all probabilities is 1. Pi is used to make a weighted degree distribution, PPi.

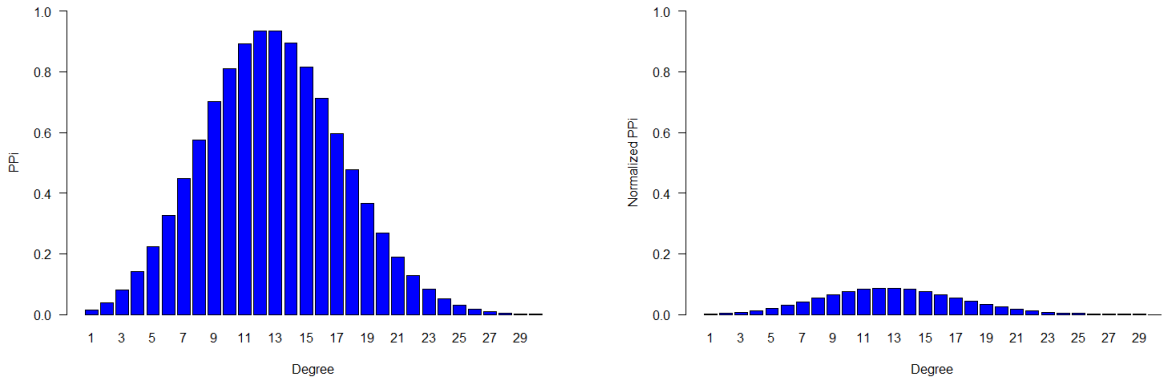
$$PPi = \frac{iPi}{\sum_{j=1}^{3N_a} jPj} \tag{26}$$

This distribution is also normalised. Now we have a weighed degree distribution from which, for every recruit from group a in the sample, a number is randomly drawn to simulate the degree of these recruits. Then the mean degree per group is calculated which is used to estimate the prevalence.
The R code for the simulation of the degrees is in listing 3.



(a) (b)

Figure 3: In (a) is a normal distribution with mean equal to the mean degree of the group and standard deviation that is one half of the the mean degree of the group, see equation (25). In (b), Pi is the discretisized version of the distribution.



(a) (b)

Figure 4: In (a), PPi is the weighted distribution of Pi as in equation (26). This distribution is then normalised so that the sum of all probabilities is 1.

4.3 Recruitments

The homophily and mean degree are used to determine the probabilities and p_{ab}, the probabilitie of someone from group a not choosing someone from group a but from group b has to be $p_{ab} = 1 - p_{aa}$, as stated in chapter 2.2.4.

Table 3: The probability matrix

group	***a*** 🔵	***b*** 🔴
a 🔵	p_{aa}	p_{ab}
b 🔴	p_{ba}	p_{bb}

For each participant, a random number is drawn between 0 and 1 which is compared to the probabilities of the recruitment matrix. For each wave, the simulation creates new recruits with chances of being a or b according to the entries of the Markov probability matrix. An example of a simulation with one coupon per participant is found in table 4 and 5.

Table 4: The results of the simulation with respondents from group a (blue) and respondents from group b (red). The recruits are represented by blue and red dots for seeds and all of the 10 waves.

Seed \Wave	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
🔵	🔵	🔵	🔵	🔵	🔴	🔴	🔴	🔵	🔴	🔴
🔵	🔵	🔵	🔵	🔵	🔵	🔵	🔵	🔵	🔵	🔵
🔵	🔵	🔵	🔵	🔴	🔵	🔵	🔴	🔴	🔴	🔵
🔵	🔵	🔴	🔴	🔵	🔴	🔴	🔴	🔴	🔵	🔵

Table 5: Example of simulation results

	🔵	🔴	Row total
Recruitment count 🔵	$k_{aa} = 20$	$k_{ab} = 6$	26
Selection proportion 🔵	$\tilde{p}_{aa} = 0.77$	$\tilde{p}_{ab} = 0.23$	1
Recruitment count 🔴	$k_{ba} = 5$	$k_{bb} = 9$	14
Selection proportion 🔴	$\tilde{p}_{ba} = 0.36$	$\tilde{p}_{bb} = 0.64$	1
Total distr. of recruits M	25	15	40
Sample distribution f	0.625	0.375	1

The R-code for the making of the simulation is in listing 1.

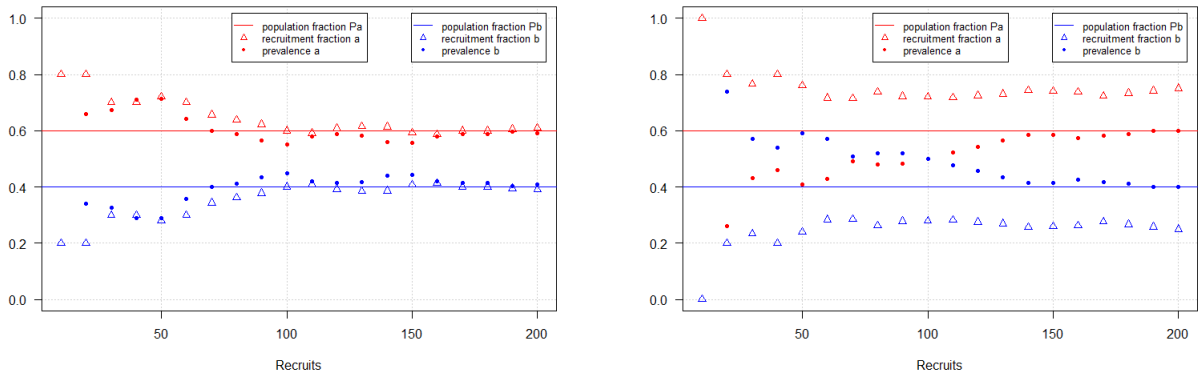
4.4 Plots of single coupon simulation

4.4.1 Samples without homophily

To vizualize the results of the sample, the amount of type a recruits per wave is accumulated and normalized. That way the amount of recruits can be plotted agains the sample distribution. Firstly the sample is run with equal mean degree for group a and b, and no homophily. The actual population fraction of group a is set to 60%. The equilibria are reached after about 10 waves in figure 5(a). Without the effect of homophily, and both groups having the same mean amount of contacts, there is no oversampling of either of the groups and the recruitment fraction as well as the estimated prevalence of the groups reach an equilibrium around the actual population fraction. The estimated prevalence is calculated for the accumulated recruitment fraction for every wave.

When mean degrees of the populations are not equal, but still without homophily, the expectation is that the group with the highest mean degree will be oversampled. In figure 5(b) the sampling results are plotted for a population with a mean degree for a that is twice as large as the mean degree for b, $Na = 20$ and

$Nb = 10$. Indeed the recruitment fraction of a is oversampled, but the prevalence estimators are at the same level as the population fraction after about 15 waves.



(a) (b)

Figure 5: Sample with 10 seeds and 20 waves, without homophily and a total population that consists for 60% of type a individuals. In (a) the accumulated sample count (percentages in triangles) as well as the estimated prevalence approach an equilibrium in about 10 waves. The degree of a and b are the same ,$Na = 10$ and $Nb = 10$. There is no oversampling of either of the two groups. In (b) the degree of a and b are different, $Na = 20$ and $Nb = 10$. Group a is oversampled but the estimator for the prevalence is good after about 15 waves.

Table 6: Results of a one-coupon sample with 10 seeds, 40 waves and no homophily. The population fraction of a individuals is 60% and the mean degree of group a is twice as high as the mean degree of group b, $Na = 20$ and $Nb = 10$.

	Ndata	a	b
1	Recruitment count k	227.00000	73.00000
2	Selection probability P	0.75000	0.25000
3	Selection proportion Sample p	0.75667	0.24333
4	Recruitment count k	78.00000	22.00000
5	Selection probability P	0.75000	0.25000
6	Selection proportion Sample p	0.78000	0.22000
7	Total distr. of recruits	305.00000	95.00000
8	Sample distribution SD	0.76250	0.23750
9	Mean degree N	20.00000	10.00000
10	Estimated mean degree	21.29895	10.20987
11	Error est. mean degree	0.55795	0.30892
12	Homophily H	0.00000	0.00000
13	Estimated homophily	-0.02333	-0.02333
14	Error estimated hom.	0.00242	0.00242
15	Prevalence P	0.60000	0.40000
16	Estimated prevalence	0.60577	0.39423
17	Error estimate prevalence	0.00153	0.00153

If we consider the 95 % confidence intervals for the homophily and the prevalence we see that the interval in this sample contains the variables. For the prevalence of group a, we can say with 95% confidence that the variable p_a lies within the interval;

$$\begin{array}{rcccl} \tilde{p}_a - 1.96\sqrt{Var(\tilde{p}_a)} & < & p_a & < & \tilde{p}_a - 1.96\sqrt{Var(\tilde{p}_a)} \\ 0.5291 & < & p_a & < & 0.6824 \end{array}$$

The interval for the prevalence of a is (0.5291, 0.6824) in this sample. For the homophily the 95 % confidence interval is (-0.1198, 0.0731), which contains $H = 0$. At first glance, the estimators seem reasonable.

4.4.2 Samples with homophily

When homophily is introduced in the population, the estimate of the prevalence is less accurate for the first waves. In figure 6 the sample $\tilde{p}_a$ takes longer to reach equilibrium then in the samples without homophily of figure 5. This is caused by a less effective sample size. Because when there is homophily, part of the recruitment is fixed, i.e. group a chooses a fixed part of recruits from group a, it takes longer to reach an equilibrium. Each wave generates less information than in the case without homophily. When there are more waves, the estimate for the degree will become better, so it will not take exactly twice as long to reach equilibrium when homophily is set to $H = 0.5$.

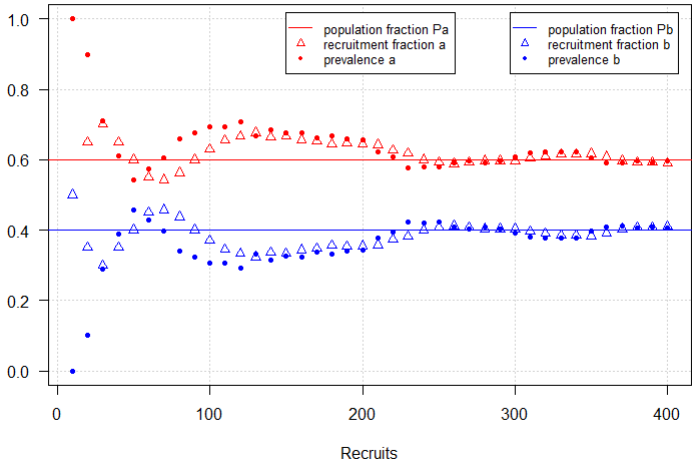


Figure 6: Sample with 10 seeds and 40 waves. Homophily is set to $H = 0.5$ and the total population consists for 60% of type a individuals. The degree of a and b are the same ,$Na = 10$ and $Nb = 10$. The recruitment fractions reach the equilibrium after about 20 waves. The estimator for the prevalence varies around the actual population fraction.

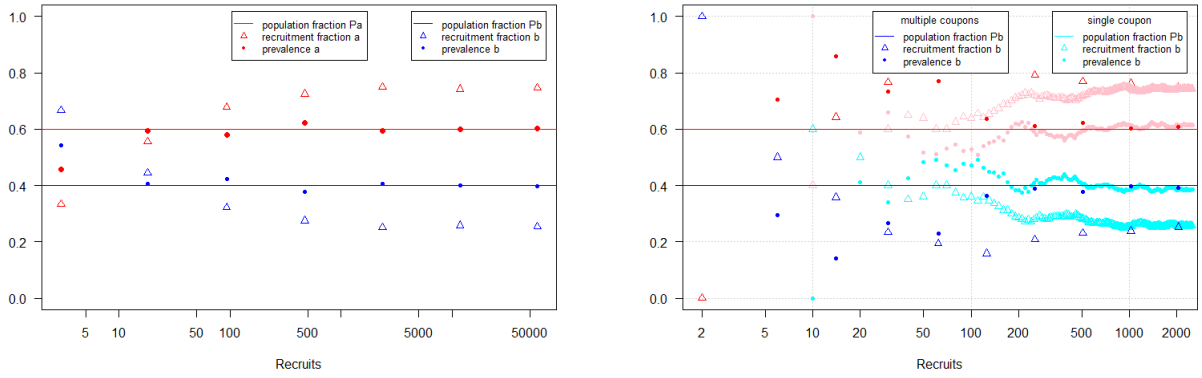
Table 7: Results of a sample with 10 seeds, 30 waves and with homophily $H = 0.5$. The population fraction of group a is 0.6 and the degree of both groups is 10.

	Ndata	a	b
1	Recruitment count k	164.00000	32.00000
2	Selection probability P	0.80000	0.20000
3	Selection proportion Sample p	0.83673	0.16327
4	Recruitment count k	31.00000	73.00000
5	Selection probability P	0.30000	0.70000
6	Selection proportion Sample p	0.29808	0.70192
7	Total distr. of recruits	195.00000	105.00000
8	Sample distribution SD	0.65000	0.35000
9	Mean degree N	10.00000	10.00000
10	Estimated mean degree	10.47197	10.81278
11	Error est. mean degree	0.34225	0.35585
12	Homophily H	0.50000	0.50000
13	Estimated homophily	0.53866	0.53866
14	Error estimated hom.	0.00269	0.00269
15	Prevalence P	0.60000	0.40000
16	Estimated prevalence	0.65340	0.34660
17	Error estimate prevalence	0.00335	0.00335

Confidence intervals are for the homophily (0.4370, 0.6403) and for the prevalence of a (0.5399, 0.7669). The variables for homophily and prevalence lie within the confidence intervals.

4.5 Varying the number of coupons

When each of the respondents is given more than one coupon, the amount recruits will grow exponentially. In Figure 7(a) the number of recruits is plotted on a logarithmic scale against the sample distributions of the simulation. The logarithmic scale makes the multi-coupon sample easier to read of the plot. For three seeds the prevalence is plotted for 6 waves. Recruits are given 5 coupons each. Group a is oversampled because the mean degree of group a is larger than the mean degree of group b, $N_a = 20$ and $N_b = 10$. The population fraction of group a is 60%. An equilibrium for the prevalence is reached almost immediately. The R-code for the multi-coupon simulation is in listing 4. In Figure 7(b) both a one- and two-coupon sample are plotted on a logarithmic scale. Homophily, population fractions and mean degrees are as in Figure 7(a). The amount of seeds is not equal for the two simulations, 2 seeds for the s coupon sample and 10 seeds for the 1 coupon sample. Both simulations reach an equilibrium for the prevalence at the same amount of about 200 recruits.

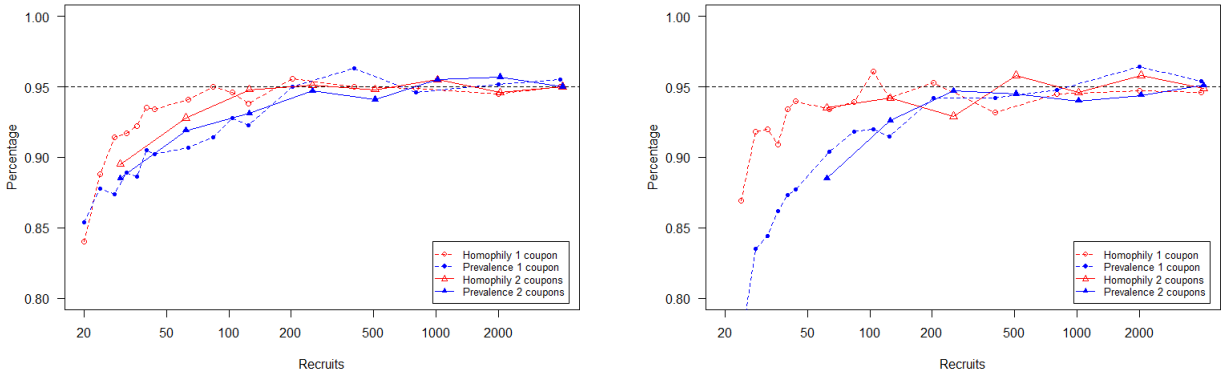


(a) (b)

Figure 7: In (a) are the recruitment fraction and prevalence for a five-coupon sample with 3 seeds and 7 waves, with homophily $H = 0,2$ and a total population that consists for 60% of type a individuals. The accumulated sample count (percentages in triangles) as well as the estimated prevalence approach an equilibrium after a short number of waves. The degree of a and b are $Na = 20$ and $Nb = 10$ respectively. There is oversampling of group a. In (b), a one-coupon sample is compared with a two-coupon sample. The mean degrees of group a and b are different, $Na = 20$ and $Nb = 10$ and the population proportion of group a is 60%. Both samples start with a different amount of seeds, 2 vs 10 for multi and single coupon sample and the number of waves varies to compare the prevalences per number of recruits more easily.

4.6 Confidence intervals

To see if the optimistic results from the samples above indeed do confirm that these estimators are reliable, 1000 simulations were run with the same parameters for a different amount of waves. In each of these samples the confidence intervals were calculated. The results of the simulations with 10 seeds, mean degree $N_a = 10$, $N_b = 10$ and a population fraction of a of 60% is in Figure 8. In Figure 8(a) the homophily is $H = 0.1$ and in Figure 8(b) the homophily is $H = 0.5$.



(a) (b)

Figure 8: The results for different numbers of waves, expressed in number of recruits on a logarithmic scale. Simulations with 4 seeds in the one coupon samples and 2 seeds in the 2 coupon sample, a population fraction of a of 60% and mean degrees $N_a = 10$ and $N_b = 10$ were run 1000 times. In (a) the homophily is $H = 0.1$ and in (b) $H = 0.5$. For each number of recruits, the percentage of variables H and P_a that lie in the calculated confidence interval are represented. The dotted lines with the circles represent the 1 coupon samples and the continuous lines with the triangles represent the two coupon samples. Red is the homophily and blue is the prevalence.

Both estimators for homophily as well as prevelence converge to the values of the corresponding variables of the population. As predicted, the prevalence of both samples in Figure 8(b) take longer to become reliable when there is homophily in the population. The consistancy of both estimators is not significantly different in the two simulations.

5 Discussion

The results of the simulations are overall in line with the predictions. The estimators for the prevalence and the homophily could, in combination with some adaption of the degree estimate, be improved.

The biased seeds are part of the estimator calculation, regardless of the amount of waves. The influence becomes less as waves are added. We could wonder if it would make a difference to exclude seeds or maybe even the first few waves from the calculation. Krista J. Gile and Mark S. Handcock studied this idea in their paper 'Respondent-Driven Sampling: An Assessment of Current Methodology', 2010 [10]. The conclusion of their investigation was that if only seeds are excluded, estimators indeed become better, but exclusion of further waves leeds to a bias in the opposite direction. The result might differ for different RDS methods, but is worth studying.

The graphs in Figure 8 (a) and (b) don't have data for the first waves. This is because for small samples, when the number of seeds and waves is low, there is a risk that k_{aa} as well as k_{ab} are zero. In that case the variance of $\tilde{p}_{ab}$ can not be calculated. It is clear that the denominator becomes zero in equation 20.
A similar problem occurs when the variance of the prevalence estimator is calculated with expression 18. This is prevented by rewriting the expression in terms of $\tilde{n}_a$, $\tilde{n}_b$, $\tilde{p}_{ab}$ and $\tilde{p}_{ba}$.

6 Appendix

To calculate the variance of the estimator we consider $\tilde{p}_a$ a function f of the variables $\tilde{n}_a, \tilde{n}_b, \tilde{p}_{ab}$ and $\tilde{p}_{ba}$.

$$f(\tilde{n}_a, \tilde{n}_b, \tilde{p}_{ab}, \tilde{p}_{ba}) = \frac{\tilde{n}_b \tilde{p}_{ba}}{\tilde{n}_a \tilde{p}_{ab} + \tilde{n}_b \tilde{p}_{ba}}$$

We approximate the estimator in point $\theta = (\mu_1, \mu_2, \mu_3, \mu_4)$.

$$\begin{aligned}
f(\mu_1+\epsilon_1, \mu_2+\epsilon_2, \mu_3+\epsilon_3, \mu_4+\epsilon_4) &= \frac{(\mu_2+\epsilon_2)(\mu_4+\epsilon_4)}{(\mu_1+\epsilon_1)(\mu_3+\epsilon_3)+(\mu_2+\epsilon_2)(\mu_4+\epsilon_4)} \\
&= \frac{\mu_2\mu_4 + \mu_4\epsilon_2 + \mu_2\epsilon_4 + \mathcal{O}(\epsilon_2\epsilon_4)}{\mu_1\mu_3 + \mu_3\epsilon_1 + \mu_1\epsilon_3 + \mathcal{O}(\epsilon_1\epsilon_3) + \mu_2\mu_4 + \mu_4\epsilon_2 + \mu_2\epsilon_4 + \mathcal{O}(\epsilon_2\epsilon_4)} \\
\text{(Assume that } \mu_1\mu_3+\mu_2\mu_4 = \triangle\text{)} & \\
&= \frac{1}{\triangle} * \frac{\mu_2\mu_4 + \mu_4\epsilon_2 + \mu_2\epsilon_4 + \mathcal{O}(\epsilon_2\epsilon_4)}{1 + \frac{\mu_3\epsilon_1}{\triangle} + \frac{\mu_1\epsilon_3}{\triangle} + \frac{\mu_4\epsilon_2}{\triangle} + \frac{\mu_2\epsilon_4}{\triangle} + \mathcal{O}(\epsilon^3)} \\
&= \frac{1}{\triangle}(\mu_2\mu_4 + \mu_4\epsilon_2 + \mu_2\epsilon_4 + \mathcal{O}(\epsilon_2\epsilon_4)) \left(1 - \frac{\mu_3\epsilon_1}{\triangle} - \frac{\mu_1\epsilon_3}{\triangle} - \frac{\mu_4\epsilon_2}{\triangle} - \frac{\mu_2\epsilon_4}{\triangle} - \mathcal{O}(\epsilon^3)\right) \\
\text{(Ignoring all terms } < \mathcal{O}(\epsilon^2)\text{)} & \\
&\approx \frac{\mu_2\mu_4}{\triangle} + \frac{\mu_4\epsilon_2}{\triangle} + \frac{\mu_2\epsilon_4}{\triangle} - \frac{\mu_2\mu_4}{\triangle}\left(\frac{\mu_3\epsilon_1}{\triangle} + \frac{\mu_1\epsilon_3}{\triangle} + \frac{\mu_4\epsilon_2}{\triangle} + \frac{\mu_2\epsilon_4}{\triangle}\right) \\
\text{(Collecting all } \epsilon \text{ terms)} & \\
&= \frac{\mu_2\mu_4}{\triangle} + \epsilon_1\left(-\frac{\mu_3\mu_2\mu_4}{\triangle^2}\right) + \epsilon_2\frac{\mu_4}{\triangle}\left(1 - \frac{\mu_2\mu_4}{\triangle}\right) + \epsilon_3\left(-\frac{\mu_1\mu_2\mu_4}{\triangle^2}\right) + \epsilon_4\frac{\mu_2}{\triangle}\left(1 - \frac{\mu_2\mu_4}{\triangle}\right) \\
\text{(Assume that } \frac{\mu_2\mu_4}{\triangle} = c\text{)} & \\
&= c + \epsilon_1 c_1 + \epsilon_2 c_2 + \epsilon_3 c_3 + \epsilon_4 c_4 + \mathcal{O}(\epsilon^3)
\end{aligned}$$

Now we take the variance on both sides of the equation.

$$Var(\tilde{p}_a) = Var(c + \epsilon_1 c_1 + \epsilon_2 c_2 + \epsilon_3 c_3 + \epsilon_4 c_4).$$

Because of independance;

$$Var(\tilde{p}_a) = c_1^2 Var(\epsilon_1) + c_2^2 Var(\epsilon_2) + c_3^2 Var(\epsilon_3) + c_4^2 Var(\epsilon_4).$$

Because all c_i's in this equation are known constants, only the variance of the ϵ_i's needs to be found. Replacing all c_i's gives;

$$Var(\tilde{p}_a) = \left(-\frac{\tilde{p}_a\tilde{p}_b}{\tilde{n}_a}\right)^2 Var(\epsilon_1) + \left(\frac{\tilde{p}_a\tilde{p}_b}{\tilde{n}_b}\right)^2 Var(\epsilon_2) + \left(-\frac{\tilde{p}_a\tilde{p}_b}{\tilde{p}_{ab}}\right)^2 Var(\epsilon_3) + \left(\frac{\tilde{p}_a\tilde{p}_b}{\tilde{p}_{ba}}\right)^2 Var(\epsilon_4)$$

Listing 1: The R-code for the Sample from the simulation of population a and b with one coupon

```
# specification of the parameters
n = 10 # n = number of seeds
m = 100# m = number of waves
H =  0.5  # homophily
Pa =  0.6 # fraction a of the population
Na=10;Nb=10 #average degree of contacts of group a and b
Bet <- 2.576 # 90% = 1.645, 95% = 1.96, 99% = 2.576

# calculation of the other parameters
Pb = 1 - Pa
# selection proportions
Paa = H +((1-H)*((Pa*Na)/((Pa*Na)+(Pb*Nb)))); Pab = 1-Paa
Pbb = H +((1-H)*((Pb*Nb)/((Pa*Na)+(Pb*Nb)))); Pba = 1-Pbb
m = m+1  #correction to make m suitable for the program

# Simulating a sample
Sample <- c(0)
# collecting all 'a' to 'a' recruitments etc.
atoa = 0; atob = 0 ;btoa = 0; btob = 0
AtoA = c(0);AtoB = c(0);BtoA = c(0);BtoB = c(0)
Sample[1] <- sample(1:n,1) #Sample collects all 'a''s per wave
for (s in 2:m){
        teller = 0
        for (t in 1:Sample[s-1]){
                if ((Sample[s-1]) > 0){
                        a = runif(1,0,1)
                        if (a >= Paa) {atob = atob + 1}
                        if (a < Paa) {teller = teller + 1 ;atoa = atoa+1}}}
        for (t in 1:(n - (Sample[s-1]))){
                if ((n - (Sample[s-1]))>0)
                        {b = runif(1,0,1)
                        if (b < Pbb) {btob =btob + 1}
                        if (b >= Pbb) {teller = teller + 1 ; btoa = btoa + 1}}}
Sample[s]=teller;AtoA[s]=atoa;AtoB[s]=atob;BtoA[s]=btoa;BtoB[s]=btob
}

Acum <- c(0) # Accumulationvector of number of a recruits
for (t in 1:m){
        Acum[t+1]=Acum[t]+Sample[t]
}
#Number of recruits per group (input for program 'degree')
TotalA <- Acum[m+1]
TotalB <- ((n*m)-TotalA)
```

Listing 2: The R-code for the calculations of the estimates

```
#simulation of degrees for all recruits a and b with program 'degree'.
MeandegreeA <- degree(Na, TotalA)
MeandegreeB <- degree(Nb, TotalB)

#atoa, atob, btoa en btob are used to fill the dataframe
# recruitment count
kaa <-atoa; kab <-atob; kba <-btoa; kbb <-btob

# estimated selection proportions
paa <-kaa/(kaa+kab); pab <-kab/(kaa+kab)
pba <-kba/(kba+kbb); pbb <-kbb/(kba+kbb)

#Estimated mean degree
na <- MeandegreeA$MeanDegree; nb <- MeandegreeB$MeanDegree

#Error Estimated mean degree
ena <- MeandegreeA$error; enb <- MeandegreeB$error

#Estimated prevalence
pa <- (nb*pba)/((nb*pba)+(na*pab))
pb <- (na*pab)/((na*pab)+(nb*pba))

#Errors
Ve1<-ena #error of mean degree a
Ve2<-enb # error of mean degree b
Ve3<-((kaa/(kaa+kab))*(kab/(kab+kaa)))/(kaa+kab) #error of pab
Ve4<-((kba/(kba+kbb))*(kbb/(kba+kbb)))/(kba+kbb) #error of pba

#Variance of the prevalence
delta <- ((na*pab)+(nb*pba))
ep <- (((pab*nb*pba)/(delta)^2)^2)*Ve1 +
(((pba/delta) - ((nb*pba*pba)/(delta^2)))^2)*Ve2 +
(((na*nb*pba)/(delta)^2)^2)*Ve3 + (((nb/delta) -
  ((nb*nb*pba)/(delta^2)))^2)*Ve4

# Estimator for homophily
h <- 1 - pab - pba

#Variance of the homophily
eh = ((pba*pbb)/(kbb+kab))+((pab*paa)/(kaa+kab))
```

Listing 3: The R code for the simulation of the degrees

```
1
2  degree <- function(Ni=20,Totali=100){
3
4  # Distribution for P(i) (Input Na,Nb)
5  sd = Ni/4;  #standard deviation
6  mean = Ni;  #mean
7  UB <- 2*Ni  #maximum number of contacts
8
9  x <- seq(1,UB,length=UB)
10
11 #SumPi is used to normalise Pi
12 SumPi <- abs(pnorm(0, mean, sd) - pnorm(UB,mean,sd))
13 Pi <- c(0)
14 for (t in 1:UB){
15     Pi[t] <- ((abs(pnorm((t-1), mean, sd) - pnorm((t),mean,sd)))/SumPi)
16 }
17
18 #plot(x,Pi,type="l",ylim = c(0,0.2))
19
20 #weighted degree distribution
21 PPi <- c(0)
22 for (t in 1:UB){
23     PPi[t] <- (t*Pi[t])
24 }
25
26 SumPPi<-sum(PPi) #SumPi is used to normalise Pi
27 for (t in 1:UB){
28     PPi[t]<-(PPi[t]/SumPPi)
29 }
30
31 # Sample to estimate the mean degree
32 prob<-PPi
33 S <- sample(1:(UB),Totali,replace=T,prob)
34
35 #Average degree by harmonic mean
36 Average <- 0
37 for (t in 1:n){
38          Average <- (Average + (1/S[t]))
39 }
40 HMeanDegree <- length(S)/Average
41
42 # Calculation of the variance of the mean degree
43 Mu <- ((1/(length(S)-1))*Average)
44 Sum <- 0
45 for (t in 1:n){
46          Sum <- (Zum + ((1/(S[t])-Mu)^2))
47 }
48 VarNi <- ((1/length(S))*Zum)
49
50 Varni <- (VarNi/(length(S)*Mu^4))
```

Listing 4: The R-code for multi-coupon simulations

```
# specification of the parameters
nn = 2# n = number of seeds
mc = 7 # m = number of waves
Hc =  0.5  # homophily of group a
Pac =  0.6 # fraction a of the population
Nac=10;Nbc=10 #average degree of contacts of group a and b
c = 2 #amount of coupons
Bet <- 1.96 # 90% = 1.645, 95% = 1.96, 99% = 2.576

# calculation oof the other parameters
Pbc = 1 - Pac
# Selection proportions
Paac = Hc +((1-Hc)*((Pac*Nac)/((Pac*Nac)+(Pbc*Nbc)))); Pabc = 1-Paac
Pbbc = Hc +((1-Hc)*((Pbc*Nbc)/((Pac*Nac)+(Pbc*Nbc)))); Pbac = 1-Pbb
mc = mc+1  #correction to make m suitable for the program


# Simulating a sample
Totalitoi<-c(nn); Samplec <- c(0)
atoac = 0; atobc = 0 ;btoac = 0; btobc = 0
AtoAc = c(0);AtoBc = c(0);BtoAc = c(0);BtoBc = c(0)
Samplec[1] <- sample(0:nn,1)
for (s in 2:mc){
        teller = 0
        for (t in 1:(c*Samplec[s-1])){
          if ((Samplec[s-1]) > 0){
                a = runif(1)#for all a's in Samplec[s-1],choose random number
                if (a >= Paac) {atobc = atobc + 1}
                if (a < Paac) {teller = teller + 1 ;atoac = atoac + 1}}}
        for (t in 1:(c*((nn*(c^(s-2)))- (Samplec[s-1])))){
      if (((nn*(c^(s-2))) - (Samplec[s-1]))>0){
                b = runif(1)
                if (b < Pbbc) {btobc =btobc + 1}
                if (b >= Pbbc) {teller = teller + 1 ; btoac = btoac + 1}}}
Samplec[s]=teller;AtoAc[s]=atoac;AtoBc[s]=atobc;BtoAc[s]=btoac;BtoBc[s]=btobc
Totalitoi[s] <- nn+AtoAc[s]+AtoBc[s]+BtoAc[s]+BtoBc[s]
}

Acumc <- c(0) # Making of the accumulationvector of number of a recruits
for (t in 1:mc){
        Acumc[t+1]=Acumc[t]+Samplec[t]
}

#Number of recruits per group (input for program 'degree')
Totala <- Acumc[mc+1]
Totalb <- (tail(Totalitoi,1)-Totala)
```

References

[1] Some Problems of Inference from Chain Data, Bonnie H. Erickson, Sociological Methodology, Vol. 10 (1979), pp. 276-302

[2] Respondent-Driven Sampling: A New Approach to the Study of Hidden Populations, Douglas D. Heckathorn, University of Connecticut, 1997.

[3] Social Networks of Jazz Musicians,Douglas D.Heckathorn and Joan Jeffri, 2003

[4] Probability Based Estimation Theory for Respondent Driven Sampling, Erik Volz and Douglas D. Heckathorn, Journal of Official Statistics, Vol. 24, No. 1, 2008, pp. 7997

[5] Sampling and Estimation in Hidden Populations Using Respondent-Driven Sampling, Matthew J. Salganik; Douglas D. Heckathorn, Sociological Methodology, Vol. 34. (2004), pp. 193-239.

[6] Collective sanctions and compliance norms: A formal theory of group-mediated social control, Heckathorn, Douglas D. 1990, American Sociological Review 55:366-384.

[7] Dynamics and Dilemmas of Collective Action, Douglas D. Heckathorn, 1996 American Sociological Review 61:250- 278.

[8] Respondent-Driven Sampling II: Deriving Valid Population Estimates from Chain-Referral Samples of Hidden Populations Douglas D. Heckathorn, Cornell University, 2002.

[9] Networks, Crowds, and Markets: Reasoning about a Highly Connected World. By David Easley and Jon Kleinberg. Cambridge University Press, 2010. Chapter 4.1

[10] Respondent-Driven Sampling: An Assessment of Current Methodology. Krista J. Gile, Mark S. Handcock, 2010.