

Modelling the Effect of Differential Recruitment on the Bias of Estimators for Respondent-Driven Sampling

Amber Tomas*

Department of Statistics
University of Oxford

January 18, 2011

*With thanks to Krista J. Gile for many helpful suggestions and conversations.

Abstract

Respondent Driven Sampling has previously been modelled as a random walk on a network. In this document we show that this model can be used to encompass within-group differential recruitment, and examine the implications for bias of several common estimators.

1 Introduction

Respondent Driven Sampling (RDS) (Heckathorn, 1997) is currently a widely used method for sampling from hidden populations. The basic method used to select a respondent-driven sample is as follows: Initially, a number of individuals from the population of interest are selected as *seeds*. Seeds are selected from a group of individuals in the population who are known to the researcher. Each seed is given a number of *coupons*, each of which has a unique bar-code, and is asked to pass on the coupons to other people they know within the population. When an individual has received a coupon, they are asked to report to a study centre where information of interest is collected by the researcher (such information is also collected from the seeds). A small monetary reward is often offered at this stage to encourage response. The responders are then themselves given coupons, and are asked to hand them on to others they know within the population, usually only to those who have not yet been recruited. In this manner, after the initial selection of seeds, the sampling is driven by the respondents. Those who report to the study centre

are known to those who have already been selected, and recruitee-recruiter relationships can be determined from the bar-codes of the coupons. The information available on which to base an estimate is therefore information collected from the respondents and from the recruitment patterns. Respondents are usually asked how many people they know within the population of interest. This provides an estimate of *degree*, as described later.

The original RDS paper (Heckathorn, 1997) suggested using the sample proportion as an estimator of population proportion (we refer to this as the “Naïve” estimator), and showed that this estimator is unbiased under very strong assumptions about the sampling process. Subsequent papers (Salganik and Heckathorn, 2004; Volz and Heckathorn, 2008; Heckathorn, 2007) have relaxed some of these assumptions and proposed several alternative estimators. We will refer to these estimators as the *Salganik-Heckathorn (SH) estimator* (Salganik and Heckathorn, 2004), the *Volz-Heckathorn (VH) estimator* (Volz and Heckathorn, 2008) and the *Heckathorn (H) estimator* (Heckathorn, 2007).

Although the estimators are easy to implement, the nature of their behaviour is not well understood. The main reason for this is that several of the assumptions which underpin the theoretical frameworks in which the estimators are derived and analysed are not met in practice. For example, it is generally assumed that sampling is with replacement or that seeds are selected randomly, whereas in practice these conditions almost never hold.

Another assumption used to derive the estimators and which is unlikely to hold in practice is that sampled individuals recruit uniformly at random from their acquaintances in the population. When this doesn’t hold, we say there exists differential recruitment. In this paper we extend the theoretical framework used in Goel and Salganik (2009) and show how it can be used to incorporate some types of differential recruitment. We then investigate the effect that differential recruitment will have on the behaviour of the estimators.

In the remainder of this document we first introduce the generalised network model which forms the basis for derivation of the RDS estimators. We then present and briefly discuss the form of the estimators used in this study. This allows us to use the network model to analyse the effect of differential recruitment on the bias of the estimators.

2 Network Models and Notation

The original network model used to model the RDS process was an undirected, unweighted graph (Heckathorn, 1997). Each individual in the population is represented as a node in the graph, and an edge between nodes i and i' , say, indicates that there is a non-zero

probability that i will recruit i' and that i' will recruit i ¹. It is usually assumed that recruits are chosen uniformly at random from among the recruiter's neighbours.

Extensions to this model have been proposed, including the use of an undirected (Goel and Salganik, 2009) or directed (Neely, 2009) weighted graph to represent non-uniform recruitment probabilities. In this paper we use a directed, weighted graph as a model for the population.

An edge exists from node i to i' if and only if there is a non-zero probability that i will recruit i' if given a coupon. The weight of the edge from node i to node i' is denoted $w_{ii'}$, and the set of children of i by $n(i)$, for all nodes i in the network. Under this model it is assumed that if i is given a coupon, then the probability i will recruit i' is given by

$$\begin{aligned} \frac{w_{ii'}}{w_i} & \text{ if } i' \in n(i) \\ 0 & \text{ if } i' \notin n(i), \end{aligned}$$

where w_i denotes the total weight on edges from node i to its children, i.e. $\sum_{k \in n(i)} w_{ik}$. In the special case that all edges have weight one, w_i is equal to the *degree* of node i , i.e. the number of neighbours of i .

Every node in the population has at least two properties associated with it: a value of the response variable y_i , and degree d_i . In this paper we assume that the response variable is binary, and that y_i represents the infection status of the i th individual.

The node and edge sets can be partitioned into mutually exclusive subsets, or "groups". *Infection groups* are defined by the values of y_i , and are denoted by A (infected) and B (uninfected). *Degree groups* are defined by a partition of the values of d_i , and will be relevant when discussing the H- estimator.

The set of nodes selected in the sample is denoted by s , and s_g denotes the intersection of s and the group g . Following standard sampling notation, we denote population totals by upper-case letters and sample quantities by lower-case letters. Hence

$$\begin{aligned} W_g & = \sum_{i \in g} w_i, \text{ and} \\ \overline{W}_g & = W_g/N_g, \end{aligned}$$

where N denotes the population size. Similarly,

$$W_{gg'} = \sum_{i \in g} \sum_{i' \in g'} w_{ii'}.$$

Finally, we denote the population proportion of nodes in group g by

$$P_g = \frac{N_g}{N}.$$

¹The label of a node will be used to refer both to that node in the network and to the individual in the population represented by that node.

We assume throughout this paper that the goal of sampling and estimation is to estimate P_A , the population proportion of infected individuals.

3 The Sampling Process

In order to make inferences about a population from a respondent-driven sample it is necessary to make some assumptions about the sampling process. In this section we briefly state the modelling assumptions which are central to the derivation of the estimators we consider in section 4.

Although the estimators we consider were originally derived in terms of the unweighted network model, for ease of generalisability we state the assumptions in terms of the more general weighted network model described in section 2.

3.1 Common Assumptions

The estimators considered in this paper are based on the following common assumptions:

Assumption 1. *Sampling is with replacement.*

Assumption 2. *All respondents recruit exactly one neighbour, and all recruits respond.*

Assumption 3. *There is a directed path between any two nodes in the network.*

Assumption 4. *The probability that individual i is selected as a seed is proportional to w_i .*

Assumption 5. *The probability that an individual i' is recruited by individual i when i has a coupon is equal to*

$$\begin{aligned} & \frac{w_{ii'}}{w_i} && \text{if } i' \in n(i) \\ & 0 && \text{if } i' \notin n(i). \end{aligned}$$

Assumption 6. *The sum of the weights on edges leading into node i is equal to the sum of weights of edges leading away from node i , for all nodes i .*

The following assumption implies that all edges are reciprocated:

Assumption 7. *If the probability is non-zero that node i will, when given a coupon, recruit node i' , then the probability is also non-zero that node i' will, when given a coupon, recruit node i . In other words, all edges are reciprocated.*

3.2 The Sampling Process as a Markov Chain on Nodes

In the case that there is only one seed and assumptions 1 and 2 hold, the sampling process can be modelled as a Markov Chain. In this case the state-space is equal to the set of nodes and the transition probabilities are given by assumption 5. The progression of time is indexed by the wave of sampling, and the original state is determined by the choice of seed. If assumption 3 also holds, then the chain is irreducible. In the case that assumptions 4, 6 and 7 hold and the chain is aperiodic (satisfied if there is at least one triangle in the graph), the following theorems apply (Goel and Salganik, 2009):

Theorem 1. *The probability π_i that individual i is recruited in wave k of the sample is proportional to w_i , for all k .*

Theorem 2. *The probability $\pi_{i'}$ that individual i recruits individual $i' \neq i$ in wave k of the sample is proportional to $w_{i'}$, for all k and i, i' .*

The proofs of these theorems follow from standard results for random walks on weighted digraphs. Even if assumption 4 is not met, Theorems 1 and 2 hold asymptotically as the number of waves is increased, i.e. $\{w_i\}$ is the stationary distribution of the Markov chain. The result of Theorems 1 and 2 are fundamental to the estimators considered in the next sections.

3.3 Violations of the Assumptions

In practice, at least some of the assumptions 1 to 7 are likely to be violated. For example, sampling is nearly always implemented without replacement, which violates assumption 1. If the sampling fraction is small, models of the sampling process based on assumption 1 are likely to be good approximations to reality, so the effect of this violation is often assumed to be negligible. However Gile and Handcock (2010) showed via simulation that even in moderately sized populations sampling without replacement can result in the estimates having substantial bias. There have been several good discussions of how the other assumptions might be violated in practice, for example Heckathorn (2002).

In the following section we discuss the estimators compared in this paper, and the assumptions upon which they are based. We then explain how differential recruitment violates the assumptions of the estimators, and investigate the effect this has on estimates of group proportion.

4 Estimators of a Population Proportion

In this section we briefly present the form of five common RDS estimators: the Naïve estimator, the Volz-Heckathorn estimator, the Successive Sampling estimator, the Salganik-Heckathorn estimator, and the Heckathorn estimator. In addition to the assumptions presented in section 3.1, the estimators are often based on the following additional assumption:

Assumption 8. *Every edge in the network has weight equal to 1.*

For completeness we derive the estimators in the more general case that assumption 5 holds, and apply assumption 8 only to derive the final form of the estimators. The more general form of the estimators will be important for our analysis of differential recruitment in section 5.

All the estimators described in this paper are functions of Hansen-Hurwitz estimators, or, in the case of the SS estimator, on the closely related Horvitz-Thompson estimators (Hansen and Hurwitz, 1943). The *Hansen-Hurwitz estimator* of a population total Y is unbiased and given by (Hansen and Hurwitz, 1943)

$$\hat{Y} = \sum_{i \in s} \pi_i^{-1} y_i, \quad (1)$$

where π_i is the wave selection probability of unit i . In the context of respondent-driven sampling, the units may be nodes or edges. The *generalised Hansen-Hurwitz estimator* of a population mean, \hat{Y}/\hat{N} , is not unbiased but is asymptotically unbiased. We refer to any ratio of Hansen-Hurwitz estimators as a generalised Hansen-Hurwitz estimator.

4.1 Naïve Estimator

The naïve estimate (Heckathorn, 1997) is equal to the sample proportion of infected individuals, i.e.

$$\widehat{P}_A^N = \frac{n_A}{n}.$$

Under assumptions 1 to 3, if the true node sampling probabilities are given by $\pi_i, i = 1, \dots, N$, then \widehat{P}_A^N is a generalised Hansen-Hurwitz estimator of

$$\frac{\Pi_A}{\Pi_A + \Pi_B}, \quad (2)$$

where Π_g denotes the subpopulation total $\sum_{i \in g} \pi_i$ (an explanation is given in appendix A.1).

Equation (2) shows that if the sampling probabilities of all nodes are equal, then \widehat{P}_A^N is a generalised Hansen-Hurwitz estimator of N_A/N . Thus, under very restrictive assumptions, \widehat{P}_A^N is asymptotically unbiased for P_A .

4.2 Volz-Heckathorn Estimator

The Volz-Heckathorn (VH-) estimator (Volz and Heckathorn, 2008) is given by

$$\widehat{P}_A^{\text{VH}} = \frac{\sum_{i \in s_A} 1/\tilde{d}_i}{\sum_{i \in s} 1/\tilde{d}_i}, \quad (3)$$

where \tilde{d}_i denotes the reported degree of node i . Using the same approach as in section 4.1, if the true node selection probability of node i is π_i , $i = 1, \dots, N$, it can be shown that $\widehat{P}_A^{\text{VH}}$ is a generalised H-H estimator of

$$\frac{\sum_{i \in A} \pi_i / \tilde{d}_i}{\sum_{i \in A} \pi_i / \tilde{d}_i + \sum_{i \in B} \pi_i / \tilde{d}_i}. \quad (4)$$

If Theorem 1 and assumption 8 hold, then $\pi_i \propto d_i$. In this case $\pi_i / d_i \propto 1$ for all i , so under restrictive assumptions, and if $\tilde{d}_i \propto d_i$ for all i , the VH-estimator is asymptotically unbiased for P_A .

4.3 Salganik-Heckathorn Estimator

The Salganik-Heckathorn (SH-) estimator makes use of the fact that it is possible to measure the proportion of within-group and cross-group recruitments in the sample by keeping track of the barcodes of coupons distributed and returned.

The SH- estimator is based on stronger assumptions than the Naïve or VH- estimators. In particular, as well as the assumptions required for Theorem 2 to hold, the derivation of the SH- estimator uses the following assumption:

Assumption 9. *The weight on edge $w_{i'}$ is equal to the weight on edge $w_{i,i}$, for all pairs i, i' .*

That is, it is assumed that the weighted directed network model is symmetric. Under these conditions it can be shown that

$$P_A = \frac{C_{\text{BA}}}{C_{\text{BA}} + C_{\text{AB}} \frac{W_A}{W_B}}, \quad (5)$$

where $C_{gg'}$ can be thought of as the probability a randomly selected recruit in group g recruits from group g' (details are given in appendix A.2). The Salganik-Heckathorn estimator is given by

$$\widehat{P}_A^{\text{SH}} = \frac{\widehat{C}_{\text{BA}}}{\widehat{C}_{\text{BA}} + \widehat{C}_{\text{AB}} \frac{\widehat{D}_B}{\widehat{D}_A}}, \quad (6)$$

where \widehat{C}_{AB} denotes the proportion of all individuals recruited by members of group A who are members of group B , and \widehat{D}_A is an estimate of mean degree given by

$$\widehat{D}_g \stackrel{\text{def}}{=} \frac{n_g}{\sum_{i \in s_g} 1/\tilde{d}_i}. \quad (7)$$

Thus $\widehat{P}_A^{\text{SH}}$ is a plug-in estimator of P_A based on expression (5), and assuming that $w_i \propto \tilde{d}_i$. Details of the derivation of expression (5) are given in appendix A.2.

Because the bias of the SH-estimator depends on the π_i , the $\pi_{ii'}$ and assumption 9, it is relatively difficult to analytically study the effect that violations of the assumptions may have on the bias of the SH-estimator, compared to the Naïve or VH-estimators.

4.4 Heckathorn Estimator

This estimator is an extension of the SH-estimator, and was motivated by the need to control for biases introduced by differential recruitment and recruitment effectiveness (Heckathorn, 2007). Rather than plug-in \widehat{D}_A for \overline{W}_A as in (6), an “adjusted” degree estimate

$$\widehat{AD}_A = \frac{\sum_{i \in s_A} \text{RCD}_i}{\sum_{i \in s_A} \left(\frac{1}{\tilde{d}_i} \text{RCD}_i \right)} \quad (8)$$

is used. RCD_i is called the “recruitment component of degree” for individual i , and is formed based on *degree* groups defined by the reported degree of the respondents. RCD_i is defined as the ratio of the degree group sample proportion and the estimated degree group equilibrium probabilities. That is,

$$\text{RCD}_i = \frac{\hat{E}_g}{\frac{n_g}{n}}, \text{ for } i \text{ in degree group } g, \quad (9)$$

where n_g is the number of respondents in degree group g and \hat{E}_g is the estimated equilibrium probability of being in degree group g . An additional assumption of the H-estimator over the SH- estimator is therefore as follows:

Assumption 10. *The sampling process satisfies the Markov property on groups, i.e. the group of the recruit in wave $k + 1$ depends only on the group of the recruit in wave k .*

This assumption is consistent with assumption 5 above only if the weights are such that every individual from group g has the same probability of recruiting from group g' , for all g, g' . That is, without more restrictive assumptions, a Markov chain on nodes does not necessarily imply a Markov chain on groups.

For all degree groups g, g' , the probability of transition for the Markov chain from g to g' is estimated by the proportion of recruitments from g to g' , $\widehat{C}_{gg'}$. The estimated

equilibrium probability of being in group g , \hat{E}_g , is then calculated from the matrix of estimated transition probabilities (Heckathorn, 2007, pg 171).

The Heckathorn estimator uses the adjusted degree estimates in place of the unadjusted degree estimates in the SH-estimator, which gives

$$\widehat{P}_A^H = \frac{\widehat{AD}_B \widehat{C}_{BA}}{\widehat{AD}_A \widehat{C}_{AB} + \widehat{AD}_B \widehat{C}_{BA}}. \quad (10)$$

If $\text{RCD}_i = 1$ for all i , then it can be seen that $\widehat{AD}_g = \widehat{D}_g$ for all g , in which case $\widehat{P}_A^H = \widehat{P}_A^{\text{SH}}$. If the chain begins in equilibrium then n_g/n is an unbiased estimate of the equilibrium probability $E_g, \forall g$. Hence it can be seen from (9) that the H-estimate is expected to differ from the SH-estimate only if the chain does not begin in equilibrium.

4.5 Summary

Although the VH- and SH-estimators are asymptotically unbiased if all the assumptions 1 to 8 hold, in practice this won't be the case. Of interest is how the estimators compare when the assumptions are violated. One thing these three estimators have in common is that they substitute the reported degree \tilde{d}_i for the sampling probability π_i . In contrast, the Naïve estimator assumes that the π_i are equal for all i . We have shown that if the assumption of uniform recruitment implied by assumption 8 is relaxed, then $\pi_i \propto w_i \not\propto d_i$. This introduces a source of bias for $\widehat{P}_A^{\text{VH}}$ and \widehat{D}_A (as estimators of P_A and \overline{W}_A respectively), but not for \widehat{C}_{AB} as an estimator of cross-group recruitment C_{AB} . How much this approximation to the sampling weights will affect the estimates is likely to depend on how great the difference is between the w_i and \tilde{d}_i . Neely (2009) showed that even small deviations of d_i/D from the true sampling probabilities π_i can result in substantial bias, though Wejnert (2009) claims otherwise. In section 5 we investigate analytically the effect of differential recruitment on the expected bias of these estimators.

5 The Effect of Differential Recruitment

Whereas the sample design is determined by the survey organiser, differential recruitment is a property of respondent behaviour which is not possible for the survey organiser to observe directly. Therefore, it is important to consider how robust the estimators are to the presence of differential recruitment.

In this section we show that the bias of all the estimators is significantly affected by differential recruitment. The direction of this bias can be explained by considering how differential recruitment affects the probabilities of selection π_i .

Before investigating further, we first define the types of differential recruitment which we will consider in this paper.

Definition 1 (Differential recruitment). *Suppose the proportion of neighbours of node i which belong to group g is equal to p_{gi} , for all $g \in G$. Then differential recruitment exists if and only if the probability i recruits from group g is not equal to p_{gi} for any $g \in G$ and any i .*

In other words, differential recruitment exists if a respondent is more (or less) likely to pass on a coupon to neighbours of some group than if he were choosing uniformly at random from his neighbours. We will consider two types of differential recruitment:

1. Within-group differential recruitment, when nodes preferentially recruit neighbours from within their own group, and
2. Between-group differential recruitment, when all nodes preferentially recruit nodes from a particular group.

It is possible to represent within-group differential recruitment in the network model of section 2 by assigning within-group edges a different weight to cross-group edges. The resulting network is consistent with the assumptions required for Theorems 1 and 2 to hold, which allows us to use this model to draw inferences about the resulting bias of the estimators. This is the approach we take in the following section. However, between-group differential recruitment can not be modelled using this framework because the network will not be consistent with assumption 6 regarding equal total weight on edges leading to and from a node. Preferentially recruited nodes will have more weight on incoming edges than on outgoing edges, which means that Theorems 1 and 2 will not hold.

For each type of differential recruitment we distinguish between infection-group differential recruitment and degree-group differential recruitment. It is useful to note that either type of differential recruitment can induce the other if the degree distributions of infected and uninfected nodes are not the same.

We now investigate analytically the effect of within-group differential recruitment on the bias of the estimators. The effect of both types of differential recruitment is considered via simulation studies presented in Tomas and Gile (2010).

5.1 Modelling Within-group Differential Recruitment

In order to model within-group differential recruitment we make the following assumption:

Assumption 11. *All edges between nodes in group g have weight proportional to K_g . All cross-group edges have weight proportional to 1.*

This is similar to the two-group model of Goel and Salganik (2009), except that we allow the strength of differential recruitment to vary by group and there is not necessarily an edge between every pair of nodes. As discussed above, if the assumptions of Theorems 1 and 2 hold the additional assumption 11 will not change this. Under these assumptions the probability of selection for node i is proportional to w_i . Only if all $K_g = 1$, i.e. there is no differential recruitment, will w_i be equal to d_i for all i .

From any RDS it is possible to calculate the proportion of cross-group recruitments \widehat{C}_{AB} . Suppose the proportion of recruitments from group A to group B , $\widehat{C}_{AB} = 0.5$. Then it is not possible to distinguish between whether recruitment was uniformly at random and approximately 50% of children of group A nodes were from group B , or if group A individuals preferentially recruited from group A , for example, and the proportion of children of group A nodes in group B is greater than 0.5. That is, it is not always possible to distinguish if differential recruitment exists, because it is intertwined with *homophily* - the proportion of within-group edges. To see this more clearly, under the assumptions of the network model described above it can be shown that

$$w_i = d_i(1 + p_{iW}(K_g - 1)), \quad (11)$$

where p_{iW} is the proportion of within group ties from node i to its children (see appendix B). Hence for fixed $K_g \neq 1$, the difference between w_i and d_i increases with increasing homophily, i.e. as p_{iW} increases. Furthermore, if it were possible to compare d_i and w_i , it would only be possible to estimate the factor $p_{iW}(K_g - 1)$ and not K_g itself.

To simplify the analysis of the influence of differential recruitment on the (asymptotic) bias of the estimators considered in the previous section, we assume that all of assumptions 1 to 6 hold, and hence that the true probability of selection $\pi_i \propto w_i$ for all i .

We now consider the case of within-infection group differential recruitment, and how this is likely to affect the bias of the estimators.

5.2 Effect of Within-Infection-Group Differential Recruitment on the Naïve Estimator and VH-estimators

Under the assumption that $\pi_i \propto w_i$, it can be seen from equations (2) and (11) that the Naïve estimator is a Hansen-Hurwitz estimator of

$$\frac{\sum_{i \in A} d_i(1 + p_{iW}(K_A - 1))}{\sum_{i \in A} d_i(1 + p_{iW}(K_A - 1)) + \sum_{i \in B} d_i(1 + p_{iW}(K_B - 1))},$$

which can be approximated by

$$\frac{N_A}{N_A + N_B \frac{\overline{D_B} (1 + \overline{p_{BW}}(K_B - 1))}{\overline{D_A} (1 + \overline{p_{AW}}(K_A - 1))}}, \quad (12)$$

where $\overline{p_{gW}}$ is the mean of the p_{iW} for $i \in g$. The approximation (12) is exact if $\sum_{i \in g} d_i p_{iW} = \overline{p_{gW}} \sum_{i \in g} d_i$, for $g \in \{A, B\}$. Clearly \widehat{P}_A^N will be approximately asymptotically unbiased for P_A if the multiplier of N_B in expression (12) is equal to one. This shows that the effect of differential recruitment on the bias of the estimator is intertwined both with homophily and with *differential activity* - the ratio of the mean degree of infected nodes to the mean degree of uninfected nodes, i.e. $\overline{D_A}/\overline{D_B}$. Thus if there is no differential recruitment (so $K_A = K_B = 1$), \widehat{P}_A^N is asymptotically unbiased only if there is no differential activity, as was previously shown by Heckathorn (2002) and Volz and Heckathorn (2008). If there is no differential activity \widehat{P}_A^N is biased if $K_A \neq K_B$. In this case, increasing differential recruitment in group A will mean that \widehat{P}_A^H is expected to increase. If there is differential activity and differential recruitment, the bias introduced by each can either reinforce or counter each other.

A similar relationship between bias and differential recruitment exists for the VH-estimator. From equations (4) and (11), if $\pi_i \propto w_i$ then \widehat{P}_A^{VH} is a generalised Hansen-Hurwitz estimator of

$$\frac{N_A}{N_A + N_B \frac{1 + \overline{p_{BW}}(K_B - 1)}{1 + \overline{p_{AW}}(K_A - 1)}}. \quad (13)$$

Thus, unlike the Naïve estimator, the effect of differential recruitment on the bias of the VH-estimator does not depend on differential activity. However, similar to the Naïve estimate, if the differential recruitment within group A is higher than that within group B , the bias of estimated proportion \widehat{P}_A^{VH} should increase.

5.3 Effect of Within-infection-group Differential Recruitment on the SH- and H- Estimators

Because within-group differential recruitment does not violate any of assumptions 1 to 6, the key relationship (5) on which the SH-estimator is based will still hold. Similarly, \widehat{C}_{AB} is still a generalised Hansen-Hurwitz estimator of C_{AB} . Therefore, the influence of differential recruitment on the SH-estimator will be through the effect that differential recruitment has on the estimates \widehat{D}_A and \widehat{D}_B . Recall from expression (6) that

$$\widehat{P}_A^{SH} = \frac{\widehat{C}_{BA}}{\widehat{C}_{BA} + \widehat{C}_{AB} \frac{\widehat{D}_A}{\widehat{D}_B}}, \quad (14)$$

so the effect of differential recruitment on \widehat{P}_A^{SH} will be via its effect on the ratio $\widehat{D}_A/\widehat{D}_B$. Under the assumption that $\pi_i \propto w_i$, $\widehat{D}_A/\widehat{D}_B$ is an asymptotically unbiased estimator of

$$\frac{1 + \overline{p_{BW}}(K_B - 1) \overline{W}_A}{1 + \overline{p_{AW}}(K_A - 1) \overline{W}_B},$$

(see appendix B for the derivation). Thus if differential recruitment K_A within group A increases relative to K_B , $\widehat{D}_A/\widehat{D}_B$ is also expected to decrease for fixed $\overline{W}_A/\overline{W}_B$ and hence, from (14), $\widehat{P}_A^{\text{SH}}$ is expected to increase.

As discussed in section 4.4, the H-estimator will be different to the SH-estimator if the Markov chain on degree groups does not begin in equilibrium. If this is the case we might expect that increased differential recruitment will mean that it takes longer for the chain to reach equilibrium (Goel and Salganik, 2009), and hence for differential recruitment to result in larger differences between the SH- and H- estimators.

5.4 Summary of the Effect of within-infection-group Differential Recruitment

For all of the estimators considered, the influence of within-infection-group differential recruitment on the estimator of P_A depends on the ratio

$$\frac{1 + \overline{p}_{BW}(K_B - 1)}{1 + \overline{p}_{AW}(K_A - 1)}. \quad (15)$$

This shows that if there is no differential recruitment, i.e. $K_A = K_B = 1$, then homophily does not affect the bias of the estimator. However, if $K_g \neq 1$, then the influence of this differential recruitment on the estimator will depend on the extent of homophily. If K_A increases relative to K_B , so the ratio (15) decreases, the expected values of \widehat{P}_A^{N} , $\widehat{P}_A^{\text{VH}}$ and $\widehat{P}_A^{\text{SH}}$ will decrease. If this ratio increases, then the estimates will also increase. If differential recruitment exists but its relation to homophily is such that the ratio (15) is close to 1, then there will be no added bias. In this way homophily can either reinforce or counter the bias introduced by differential recruitment.

It can be seen from equations (12), (13) and (14) that the rate at which changes in this ratio will affect \widehat{P}_A^{N} , $\widehat{P}_A^{\text{VH}}$ and $\widehat{P}_A^{\text{SH}}$ is a function of $N_B\overline{D}_B/N_A\overline{D}_A$, N_B/N_A and $\widehat{C}_{AB}/\widehat{C}_{BA}$ respectively. Hence the estimates may be affected differently by similar changes in differential recruitment, although the direction of change will be the same.

5.5 Simulation Results

The design of the simulations is the same as in Tomas and Gile (2010), to which readers are referred for details. Briefly, the population is modelled as an undirected network with 1000 nodes, 200 of which are infected. There is a moderate amount of *homophily* in all the networks, and differential activity is varied to take values in $\{0.5, \mathbf{1}, 1.8\}$. Networks were simulated using the **R** package `statnet` (Handcock et al., 2003, 2008). From every network a respondent driven sample was taken, and the value of the estimates computed.

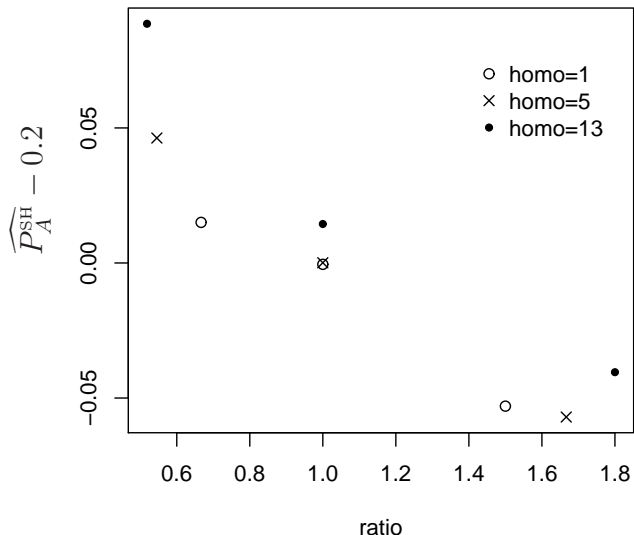


Figure 1: The estimated bias of the SH-estimator against the ratio of uninfected to infected $1 + \overline{p_{gW}}(K_g - 1)$, for the values of homophily $\overline{p_{gW}}$ such that the ratio of an infected-infected edge to an infected-uninfected edge $\in \{1, 5, 13\}$, and for differential recruitment K_g considered in the simulations. There is no differential activity.

For the sampling, ten seeds were selected at random with probability proportional to degree, every respondent was given two coupons, and the maximum sample size was 200.

Figure 1 plots the bias of the SH-estimator against the ratio $(1 + \overline{p_{AW}}(K_A - 1))/(1 + \overline{p_{BW}}(K_B - 1))$ for all simulations which were performed with seeds selected randomly with probability proportional to degree, no differential activity and perfect response rate and recruitment effectiveness. The trend for $\widehat{P}_A^{\text{SH}}$ to decrease as the ratio increases is clear, though there is substantial variation. Some of this variation seems to be due to an additional effect of homophily. The trends for the other estimators are very similar, so are not shown.

Results of the simulations for three levels of differential activity are shown in Figure 2. Note that the counts at the top of the boxplot denote the number of samples for which the estimate was equal to 1. Counts at the bottom of the boxplot denote the number of samples for which an estimate could not be calculated².

From looking at the Naïve estimates in Figure 2 it can be seen that if uninfected nodes

²The SH- and H- estimators will return a value of 1 if recruitments were made from infection group A to B , but not from B to A . In every case that an estimate could not be computed, it was due to there being no recruitments at all from at least one of the infection groups.

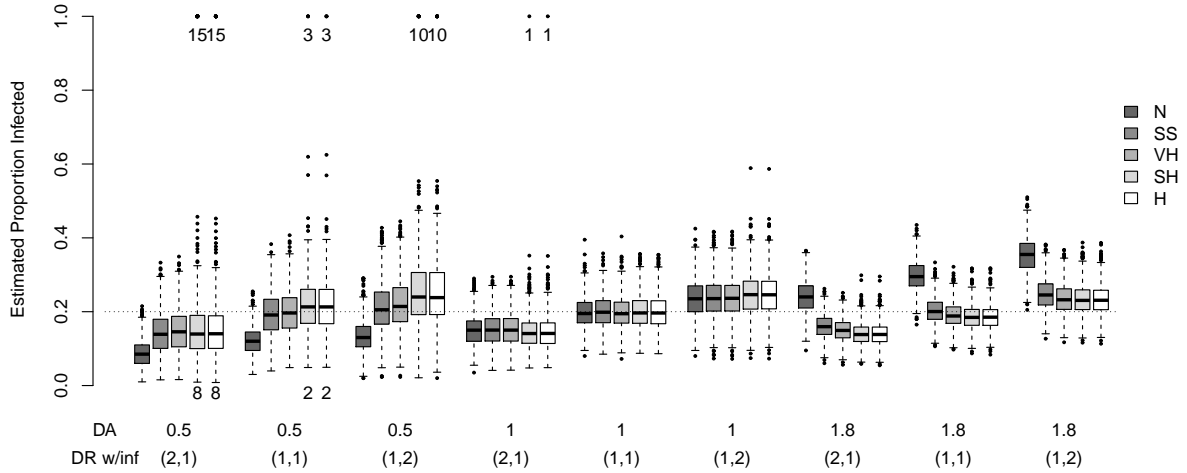


Figure 2: Simulation results for varying levels of within-infection-group differential recruitment and differential activity. Differential recruitment is coded as (K_B, K_A) .

are twice as likely to recruit from within-group as infected nodes (parameter level $(2, 1)$), the sample proportion of infected nodes is reduced relative to the case where there is no differential recruitment. For within-group differential recruitment of $(1, 2)$ the sample proportion of infected nodes is increased. This is true for all levels of differential activity. It can also be seen that the VH-, SH- and H- estimates are significantly higher or lower than the sample proportion for differential activity levels of 0.5 and 1.8 respectively. It can be seen from the instances with no differential recruitment that this is an adjustment for differential activity, because the VH-, SH- and H- estimators are significantly less biased than the Naïve estimator in these cases. The VH- estimator is slightly less biased than the SH- and H- estimators. However, there is no evidence that any of the estimators are adjusting for differential recruitment, because the estimates increase or decrease by approximately the same amount as the sample proportion as the differential recruitment level is varied for a given level of differential activity. This behaviour is as we would expect from the theoretical results of sections 5.2 and 5.3.

A Derivation of Expressions for the Estimators

A.1 Hansen-Hurwitz estimators

Consider the Hansen-Hurwitz estimator of Π_A ,

$$\begin{aligned}\widehat{\Pi}_A &= \sum_{i \in s_A} \pi_i^{-1} \pi_i \\ &= n_A.\end{aligned}$$

Thus, taking the ratio of Hansen-Hurwitz estimators for Π_A and $\Pi_A + \Pi_B$ shows that $n_A/(n_A + n_B)$ is a generalised Hansen-Hurwitz estimator of $\Pi_A/(\Pi_A + \Pi_B)$.

A.2 Derivation of Salganik-Heckathorn Estimator

The Salganik-Heckathorn (SH-) estimator makes use of the fact that it is possible to measure the proportion of within-group and cross-group recruitments in the sample by keeping track of the barcodes of coupons distributed and returned.

Denote the proportion of all individuals recruited by members of group A who are members of group B by

$$\widehat{C}_{AB} = \frac{t_{AB}}{t_{AA} + t_{AB}}, \quad (16)$$

where $t_{gg'}$ denotes the number of times in the sample an individual from group g recruited someone from group g' . Then \widehat{C}_{AB} is a generalised H-H estimator of $\Pi_{AB}/(\Pi_{AA} + \Pi_{AB})$, where $\Pi_{gg'} = \sum_{i \in g} \sum_{i' \in g'} \pi_{ii'}$. Hence under the conditions in which theorem 2 holds, so $\pi_{ii'} \propto w_{ii'}$, \widehat{C}_{AB} is a generalised Hansen-Hurwitz estimator of

$$\frac{W_{AB}}{W_{AA} + W_{AB}} = \frac{W_{AB}}{N_A \overline{W}_A} = C_{AB}, \text{ say.} \quad (17)$$

It follows from assumption 9 that

$$W_{AB} = W_{BA}, \quad (18)$$

where $W_{gg'}$ is the total weight on edges between nodes in group g and nodes in group g' . This is a generalisation of what Heckathorn (2007) refers to as the ‘‘reciprocity’’ assumption. Thus equations (17) and (18) imply that

$$\begin{aligned}C_{AB} N_A \overline{W}_A &= C_{BA} N_B \overline{W}_B, \text{ so} \\ P_A &= \frac{\overline{W}_B C_{BA}}{\overline{W}_B C_{BA} + \overline{W}_A C_{AB}} = \frac{C_{BA}}{C_{BA} + C_{AB} \frac{\overline{W}_A}{\overline{W}_B}}.\end{aligned} \quad (19)$$

B Modelling Differential Recruitment and Non-response

For i in group g

$$\begin{aligned}
 w_i &= \sum_{i' \in n(i)} w_{ii'} \\
 &\propto \sum_{i' \in n(i): i' \in g} K_g + \sum_{i' \in n(i): i' \notin g} 1, \text{ (from assumption 11),} \\
 &= d_i(1 + p_{iW}(K_g - 1))
 \end{aligned}$$

as claimed.

Under the assumption that $\pi_i \propto w_i$, $\widehat{D}_A/\widehat{D}_B$ is an asymptotically unbiased estimator of

$$\begin{aligned}
 &\frac{\sum_{i \in A} w_i}{\sum_{i \in A} w_i/d_i} \frac{\sum_{i \in B} w_i/d_i}{\sum_{i \in B} w_i} \\
 &= \frac{W_A}{N_A(1 + \overline{p_{AW}}(K_A - 1))} \frac{N_B(1 + \overline{p_{BW}}(K_B - 1))}{W_B} \\
 &= \frac{1 + \overline{p_{BW}}(K_B - 1) \overline{W}_A}{1 + \overline{p_{AW}}(K_A - 1) \overline{W}_B}.
 \end{aligned}$$

References

- Krista J. Gile and Mark S. Handcock. Respondent-driven sampling: An assessment of current methodology. *Sociological Methodology*, 40(1):285–327, 2010.
- Sharad Goel and Matthew J. Salganik. Respondent-driven sampling as markov chain monte carlo. *Statistics in Medicine*, 28(17):2202–2229, 2009.
- Mark S. Handcock, David R. Hunter, Carter T. Butts, Steven M. Goodreau, and Martina Morris. *statnet: Software Tools for the Statistical Modeling of Network Data*. Statnet Project <http://statnetproject.org/>, Seattle, WA, 2003. URL <http://CRAN.R-project.org/package=statnet>. R package version 2.0.
- Mark S. Handcock, David R. Hunter, Carter T. Butts, Steven M. Goodreau, and Martina Morris. statnet: Software tools for the representation, visualization, analysis and simulation of social network data. *Journal of Statistical Software*, 24(1), 2008. <http://www.jstatsoft.org/v24/i01/>.
- M.H. Hansen and W.N. Hurwitz. On the theory of sampling from finite populations. *Annals of Mathematical Statistics*, 14:333–362, 1943.
- Douglas D. Heckathorn. Respondent-driven sampling: A new approach to the study of hidden populations. *Social Problems*, 44:174–199, 1997.

- Douglas D. Heckathorn. Respondent-driven sampling ii: Deriving valid population estimates from chain-referral samples of hidden populations. *Social Problems*, 49(1):11–34, 2002.
- Douglas D. Heckathorn. Extensions of respondent-driven sampling: Analyzing continuous variables and controlling for differential recruitment. *Sociological Methodology*, 37(1): 151–207, 2007.
- W. Whipple Neely. Statistical theory & respondent-driven sampling (under review), 2009.
- Matthew J. Salganik and Douglas D. Heckathorn. Sampling and estimation in hidden populations using respondent-driven sampling. *Sociological Methodology*, 34:193–239, 2004.
- Amber Tomas and Krista J. Gile. The effect of differential recruitment, non-response and non-recruitment on estimators for respondent driven sampling, 2010. <http://arxiv.org/abs/1012.4122v1>.
- Erik Volz and Douglas D. Heckathorn. Probability based estimation theory for respondent driven sampling. *Journal of Official Statistics*, 24(1):79–97, 2008.
- Cyprian Wejnert. An empirical test of respondent-driven sampling: Point estimates, variance, degree measures, and out-of-equilibrium data. *Sociological Methodology*, 39 (1):73–116, 2009. URL www.respondentdrivensampling.org.