On Sampling Social Networking Services

Baiyang Wang[*]

Abstract

This article aims at summarizing the existing methods for sampling social networking services and proposing a faster confidence interval for related sampling methods. Social networking services (SNSs), such as Facebook and Twitter, are an important part of current Internet culture. Collecting samples from these services should be a key process to learn more about sociological or psychological issues. However, typical sampling methods for networks, such as node-based or link-based methods, are not always feasible for social networking services. Alternate approaches such as random walk (RW) or Breadth-First-Search (BFS) are applied to gather information from social networking services more efficiently. Thus it is beneficial to compare various sampling approaches for SNSs and discuss the most suitable one for each situation, which are shown later.

Moreover, to make statistical inference on the gathered information, it is necessary to apply another set of approaches to solve the related problems. One problem is the estimation of sampling probabilities, e.g. the probability for a sampled node to be a particular node v, which is denoted as $\pi(v)$. Many approaches are currently available, and their real performance will be discussed later on in terms of RDS sampling.

RDS sampling, the so-called respondent driven sampling, is invented to detect hidden population in society and has been generalized and widely discussed ever since. Many methods for estimating the sampling probabilities and constructing confidence intervals have been proposed and improved. These methods are useful, yet their properties have not been studied sufficiently, some of which will be discussed in the following. Based on these existing methods, we also propose modification of existing estimation methods and the construction of a faster confidence interval which has not been covered in current related literature.

After all, the topic of sampling on social networking services is relatively new and there is still much to be explored.

1. Networks and Social Networking Services

Social networking services form a special kind of network. The people who join these services are connected by friendship relations and they together constitute a network web which is quite huge. Yet these networks are also intangible and it often proves difficult to collect information from them. We could trace from person to person in order to get a view of these networks, but this approach is very subtle and the problems related to it will be the main focus of this paper. After all, to analyze social networking services as networks, we should first give a clear definition of them.

[*] The author thanks all the help from Professor Mark S. Handcock, UCLA.

A network is a collection of “points”, which we call *nodes*, and specific relationships between them, which we call *edges*. In our situation, a *node* refers to an account or user on an SNS; an *edge* or *link* refers to a kind of relationship (friendship on Facebook, etc.) between two accounts; the *degree* of a node refers to the number of nodes linked to it; the *volume* of a set of nodes refers to the sum of the degrees of all the nodes; the neighborhood $N(u)$ of a node refers to the set of all nodes linked to node u.

2. Sampling Methods for Social Networking Services

2.1. Node-based Sampling Methods

2.1.1 Simple node-based sampling

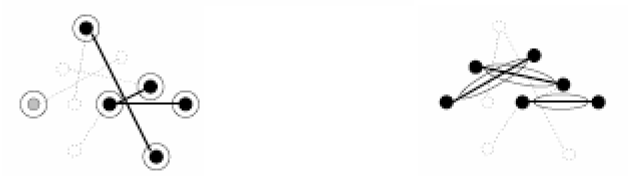
If we desire to make inference on a certain group without any other restriction, then we can apply the simplest node-based sampling method: to randomly gather a group of people on SNSs and keep the links between them (Figure 2.1a). Though this method generally preserves the topological structures of a social networking graph [1], it is not available for sampling under many circumstances. There were some available SNS datasets, such as complete available datasets for Facebook users in Harvard [2] and Caltech [3], which dated to 2008 [4]. It is possible to collect a random sample from these datasets with the node-based sampling method. Yet online social networks are highly dynamic, and typical SNSs such as Facebook can only generate a small sample each time for a given group of people, which make the node-based method of sampling difficult for SNSs.

2.1.2 Alternative node-based sampling (rejection sampling)

Another way to acquire a uniform random sample from an SNS is to generate random user IDs uniformly and then reject the IDs that do not match any user. It is proved that this kind of sampling generates the same results as uniform random sampling without replacements [4]. This approach was previously available for Facebook; but now, Facebook has stopped using number IDs, and this approach appears to be less practical.

2.2 Link-Based Sampling Method

We can also apply link-based sampling to an SNS: to randomly gather a number of links and then keep the nodes attached to them. The difference between node-based sampling and link-based sampling is illustrated as follows [1]:



(a) Node-based (b) Link-based

Figure 2.1. The difference between node-based and link-based sampling methods.

The link-based sampling method just preserves some of the topological structures of a social networking graph [1]. Besides, it shares the same weaknesses with node-based sampling, as random sampling on links in social networks is often not available. Yet when investigating on the characteristics of relationships (links) in social network services, the link-based method may be a most suitable one.

2.3 Traversal Sampling Methods

When conducting sampling on a social network graph, it is convenient to use the property that the nodes are connected from one to another. Therefore, we can sample from one node to another node (*referring*) repeatedly using the existing links between them, which may accelerate the speed of sampling on SNSs. Methods of this type are generally called *crawling methods*. Yet crawling methods can be divided into two categories. *Traversal methods* sample each node only once and include the four methods listed below; the other type allows replacements, which includes random walk methods and respondent-driven sampling in Section 2.4 and 2.5.

The following picture illustrates a typical traversal sampling:

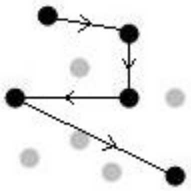


Figure 2.2. A typical traversal sampling (irrelevant links omitted).

Almost all crawling methods are biased towards nodes with higher degree, i.e. the probability of sampling nodes with higher degree is larger. The possible methods to correct this bias are discussed in Section 3.

2.3.1 Breadth-First-Search (BFS)

The Breadth-First-Search (BFS), or the *snowball sampling* method, starts with a certain node in a social network and then samples all its relevant nodes, and the samples all the relevant nodes of its relevant nodes, etc. until the total number of nodes reach a certain amount [1]. The following picture illustrates the BFS method (the number indicates the order):

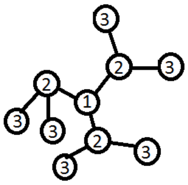


Figure 2.3. An example of the BFS method.

The BFS method is arguably the fastest method of all, and it is also feasible in current online social networks. Yet the high speed of BFS method must be weighed against its shortcomings. Many early investigations on SNSs applied the BFS method [1, 5-8], but evidence has shown that the BFS method tends to distort the topological features of a network [1], and it also suffers from other problems such as low convergence rates (the speed of the sample to approximate the actual distribution) and high bias towards high-degree nodes [9] (in other words, the probabilities of appearing in the sample for higher-degree nodes are higher).

2.3.2 Depth-First-Search (DFS)

The Depth-First-Search starts with a certain node and goes along a random route until it reaches a terminal. Then it retreats to the nearest visited node joining another branch and goes along this branch randomly until it reaches a terminal again, and so force… until a whole connected component of a social network is visited [12]. This method aims at visiting a whole graph or one connected component (such as all people that can be traced from one person in Facebook), and when it is carried out incompletely, it also introduces an unknown bias towards high-degree nodes [9]. The following picture illustrates the DFS method (the number indicates the order):

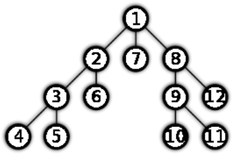


Figure 2.4. An example of the DFS method.

2.3.3 Forest Fire Sampling (FF)

Forest Fire sampling is a modification of the BFS method. In Forest Fire sampling, we start with a certain node, sample each of its neighbors with probability *p* and then sample each of the neighbors of its sampled neighbors with probability *p*, etc. until we have collected a certain number of samples. FF sampling reduces to BFS when *p*=1. This method has similar properties with the BFS method [9], and it was applied in [18].

2.3.4 Alternative Snowball Sampling (SBS)

Alternative snowball sampling, simply referred to as *snowball sampling* in [9], [19] and [20] (not to be confused with the BFS method), is another modification of the BFS method. Instead of sampling all neighbors of a node, we now just sample *n* neighbors of a node at each step, and then reject the neighbors that we have already sampled. [19] gave a crude estimation of the sampling probabilities for this approach, and used it to decrease the sampling bias of this method, but the bias correction did not always work satisfactorily.

2.4 Random Walk Sampling Methods

The random walk sampling consists of many sampling methods that are useful for sampling social networking services. A typical random walk starts with an arbitrary node, then randomly visits one of its neighbors, and then randomly visits a neighbor of the neighbor, etc. until a certain amount of nodes (with replacements) is collected. It has shown that random walk sampling is more theoretically solid than other sampling when carried out properly, though they may not be useful in terms of non-local graph properties like the graph diameter and the average shortest path length [4]. However, from our further analysis in Section 4, in terms of sampling error, random walk may not work as well as traversal methods even under some simple cases. It should be noticed that when replacements are impossible, random walk reduces to a traversal method. The variations of random walk sampling are as follows:

2.4.1 Simple Random Walk (Random Walk, RW)

Simple random walk visits at each step one of the neighbors of the previous node with the same probability. If the social network is connected and aperiodic, the sampling probability of a particular node v converges to the stationary distribution: $\pi^{RW}(v) = \frac{\deg(v)}{2\mathrm{vol}(V)}$, V be the whole network graph [13]. There is a known bias in the sampling results; therefore, we must re-weight the results using some method discussed in 4.1.

2.4.2 Metropolis-Hastings Random Walk (MHRW)

Metropolis-Hastings Random Walk visits at each step the neighbor v of the last node u with the same probability and move to the new node with probability $\min(1, \frac{\deg(u)}{\deg(v)})$. Otherwise, the node remains the same. Then the sampling probability of each node is the same: $\frac{1}{|V|}$ [11, 14]. Both RW and MHRW significantly outperform BFS in terms of convergence and topological properties, and are acceptable [4]. RW is slightly more efficient than MHRW [16, 17], but consider the higher efficiency of computing for MHRW, both methods have their merits.

2.4.3 Weighted Random Walk (WRW)

Now suppose we want to gather more information from a certain group of people in SNSs. To achieve this effect, we can place weight $w(u,v)$ on the edge (u,v) of a network so that our desired nodes incur more weight. Then we visit at each step the neighbor v of the last node u with probability $\frac{w(u,v)}{\sum_{v\in N(u)} w(u,v)}$. The sampling probability of a node v is $\pi^{WRW}(v) = \frac{w(v)}{\sum_{u\in V} w(u)}$, $w(v) = \sum_{v\in N(u)} w(u,v)$ [11].

2.4.4 Stratified Weighted Random Walk (S-WRW)

Suppose we want to gather information from several categories which differ greatly in size. To get an approximately equal number of samples from each category, some means must be taken. The S-WRW sampling, a special case of WRW, aims to solve the problem with following algorithm [11]:

(1) Run a short pilot random walk and discover categories C_i's. Let
$\hat{f}_i^{\mathrm{vol}} = \frac{1}{|S|}\sum_{u\in S}\left(\frac{1}{\deg(u)}\sum_{v\in N(u)} 1_{\{v\in C_i\}}\right)$.

(2) Let $w_{WIS}(C_i) = 1$ and $\widetilde{w}_{WIS}(C_i) = \begin{cases} w_{WIS}(C_i), & C_i \neq C_\theta \\ \tilde{f}_\theta \sum_{C\neq C_\theta} w_{WIS}(C), & C_i = C_\theta \end{cases}$. C_θ is the combination of all irrelevant categories and $\tilde{f}_\theta$ is an arbitrary number such that $0 < \tilde{f}_\theta \ll 1$.

(3) Let $\tilde{f}_i^{\mathrm{vol}} = \max\{\hat{f}_i^{\mathrm{vol}}, f_{\min}^{\mathrm{vol}}\}$, $f_{\min}^{\mathrm{vol}} = \frac{1}{\gamma}\sum_{C_i\neq C_\theta}\{\hat{f}_i^{\mathrm{vol}}\}$. γ is arbitrary, $\gamma \geq 1$ and $\tilde{f}_i^{\mathrm{vol}} = f_{\min}^{\mathrm{vol}}$ for any C_i undiscovered in the pilot random walk.

(4) Let $\widetilde{w}_e(C_i) = \frac{\widetilde{w}_{WIS}(C_i)}{\tilde{f}_i^{\text{vol}}}$, which is the weight of any edge in C_i .

(5) Let $w(C_i, C_j) = \begin{cases} \sqrt{\widetilde{w}_e(C_i)\, \widetilde{w}_e(C_j)}, C_\theta \in \{C_i, C_j\} \\ \max\{\widetilde{w}_e(C_i), \widetilde{w}_e(C_j)\}, \text{otherwise} \end{cases}$, which is the weight of any edge between C_i, C_j.

(6) An decrease in $\tilde{f}_\theta$ results in more focus on relevant categories but less convergence speed; an increase in γ results in more focus on smaller categories but larger possibility to become “trapped” in them, which means, more deviation.

2.5 Respondent-Driven Sampling (RDS)

Respondent-driven sampling was first invented as a chain-referral sampling method to detect hidden populations, such as drug addicts or sex workers, to whom traditional sampling methods were unable to get access [21]. In respondent-driven sampling, surveyees are asked to report information of whom they know, and these people may turn into later surveyees. Later this method was generalized and its characteristics were well-studied [22].

Nowadays RDS generally refers to crawling sampling methods which samples *n* nodes from one node at a time, both with and without replacements. When *n* = 1, RDS with replacements reduces to RW and when RDS is performed without replacements, it can be viewed as alternative snowball sampling. It was mentioned in [17] and [22] that the sampling probabilities of RDS is proportional to the node degree, and even so as an approximation when it is conducted without replacements. However, [9] and [30] have shown that this approximation is not satisfactory enough and there are various ways below to improve it.

3. Estimation of Sampling Probabilities

3.1 Direct Estimation

The sampling probabilities for sampling methods with replacements have already been given. For traversal methods, it is impossible to derive such formulas for sampling probabilities. However, it is still possible to derive some satisfactory estimation for traversal methods. [9] and [10] provide a direct estimation of the sampling probabilities with the following logic:

(1) First we introduce the concept of node degree distribution $p(k)$, which is the proportion of nodes with degree k inside the network. Then we consider the given network as a random network with fixed degree distribution $p(k),\ k \in \mathbb{N}$. A node with degree k is considered to have k “stubs”. We can connect all the stubs in pairs randomly to form the final random social network graph, but now we just leave them unconnected.

(2) We adopt such an algorithm: choose a node v_0, set $\mathrm{S} = \{v_0\}$. We take an element a from S and add ones of its neighbors, b, into S. We repeat this process until we are done. This algorithm implements many traversal sampling methods. It is the BFS method on a first-in first-out basis; it becomes the DFS method on a last-in first-out basis. However, we have to solve the problem of how to choose neighbors in a random graph.

(3) To solve this problem, we introduce such an approach: we assign each stub with a random number with a distribution of $U(0,1)$. Then we take v_0 with the smallest stub number, and choose the neighbor at each iteration again with the smallest stub number. This is like a "scan" process from $t = 0$ to $t = 1$.

(4) We assume the existence of this "time" variable. The expected number of sampled nodes with degree k at "time" t is: $Np(k)[1 - (1 - t)^k]$, N be the total volume. Taking the expected values as the real values, the sampling probability for a certain node v, $\pi(v) \propto 1 - (1 - t)^{\deg(v)}$.

(5) Because "time" t is unknown, we try to establish a relationship between t and sampling proportion f, which is the ratio of the sample size over the total volume. We have $f \sim \mathrm{E}f \overset{\text{def}}{=} g(t) = 1 - \sum_k p(k)(1 - t)^k$. It should be noticed that $p(k)$ can only be estimated using some other method like random walk. Letting $t = g^{-1}(f)$, the final answer we get is $\pi(v) \propto 1 - (1 - g^{-1}(f))^{\deg(v)}$. Because $\sum \pi(v) = 1$, we only need the relative $\pi(v)$.

[9] has shown that this Kurant approximation formula (for simplicity) performs quite well when the assortativity† is near 0. Otherwise, this formula yields high errors and cannot be applied.

To simplify this formula, we can take such approximation: let $g(t) = 1 - (1 - t)^{\bar{d}}$, $\bar{d}$ is the average node degree of the sample. The simplified formula does not need the estimation of node degree distribution. We will evaluate the performance of this simplification in section 3.1.1.

3.2 Gile's Successive Sampling Estimation

Gile's Successive Sampling (SS) estimates, based on simulation, are constructed as follows [30]:

(1) Initial estimation: $\pi_0(v) = \frac{k}{\sum k \frac{v_k/k}{\sum v_k/k} N} = \frac{k}{Nn} \sum \frac{v_l}{l}$, *k* is the degree of node *v*, v_k is the number of nodes with degree *k* in the sample, *n* is the sample size and *N* is the total population.

(2) For *i* = 1, ... , *r*, repeat the following steps:

i. Estimate the number of nodes with certain degrees:

$N_k^i = N \cdot \frac{v_k}{\pi_{i-1}(v)} / \sum \frac{v_k}{\pi_{i-1}(v)}$, N_k^i is the number of nodes with degree *k* in the *i*-th interation.

ii. Calculate the new sampling probabilities with simulation:

Generate a random network graph with population N and number of degree-k nodes N_k^i. Simulate M successive samples with size *n* from the network and estimate:

$\pi_i(v) = (U_k + 1)/(MN_k^i + 1)$, U_k is the total observed number of nodes with degree *k*.

(3) Take the $\pi_r(v)$ as the final $\pi(v)$.

† Assortativity is the tendency of nodes to be connected to nodes of similar degrees. It is positive if higher-degree nodes tend to be connected to higher-degree nodes; it is negative if most links are higher-degree to lower-degree connections.

[30] has shown that the Gile's SS estimates are more efficient than with-replacement estimation (sampling probabilities proportional to node degrees) when applied to traversal methods. Nevertheless, it still remains uncertain about the relative performance of the two types of estimates mentioned above. The next section will apply these estimates to various situations of statistical inference and compare their relative performances.

4. Statistical Inference for RDS Sampling Methods for SNSs

In this section, we will discuss statistical inference for RDS sampling methods. Regarding what we should estimate, we simulate networks with volume 1000 and randomly assigns each node with category A or B. We control the proportion of category A to be, say, 0.3 in the following examples. Then we will use RDS sampling to estimate the proportion of category A in terms of point estimation discussed in Section 4.1 and confidence interval estimation discussed in Section 4.2. Section 4.3 summarizes some simulation results which is related to the accuracy of many estimations for sampling probabilities. All simulation results are based on the ERGM model and averaged over 100 random network graphs with population size 1000 and 100 samples on each network without specification.

4.1 Point Estimation of Category Proportions

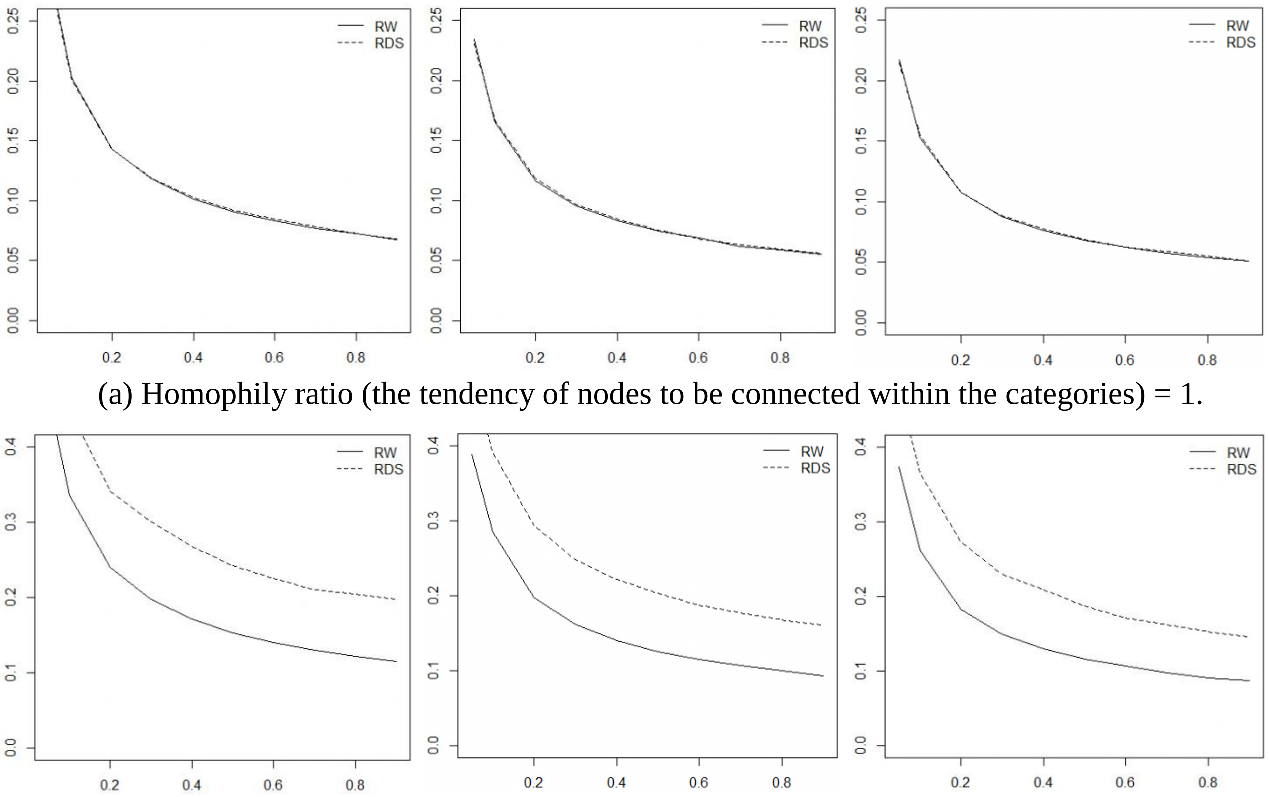
Because RDS samling methods are all biased towards higher-degree nodes, we cannot use the category proportions of the sample to replace the category proportions of the total population. The Horvitz-Thompson (or Hansen-Hurwitz, Voltz-Heckathorn) estimator [11, 15, 29] estimates population total as $\hat{x}_{total} = \frac{1}{n}\sum_{v\in S}\frac{x(v)}{\pi(v)}$, and population mean as $\hat{x}_{av} = \sum_{v\in S}\frac{x(v)}{\pi(v)} / \sum_{v\in S}\frac{1}{\pi(v)}$; $x(v)$ is the node characteristic and $\pi(v)$ is the sampling probability. This estimator is unbiased and is applied for RDS in most situations. We need only make $x(v)$ an indicator function to estimate the category proportions of the total population.

4.1.1 Sampling with Replacements

Now we investigate the accuracy of the estimator on RDS sampling with replacements and estimate the mean degree of social network graphs. We use the relative mean standard error as an indicator which is calculated from a sample of volumn n on a quadratic mean basis: $\hat{\mathrm{e}}_{\mathrm{r}} = \hat{\mathrm{e}}/p$, $\hat{\mathrm{e}} = \sqrt{\sum \mathrm{e}_{\mathrm{i}}^2/n}$, p is the category proportion. The relative mean standard errors are plotted against the sampling proportions to show how many samples we should collect in order to get a reliable estimation. There are comparisons between RDS and RW methods. We only assume that sampling probabilities are proportional to the node degrees, as Gile's SS estimates are typically applied for traversal methods.

Also, to identify the influence of node degrees on sampling results, we change a paramater called the activity ratio, which is defined as the ratio between the average node degree of category A and the average node degree of category B [29]. We set three situations from the left to the right below: activity ratio = 0.5, activity ratio = 1 and activity ratio = 2.

The relative standard errors for RW and RDS are as follows:



(a) Homophily ratio (the tendency of nodes to be connected within the categories) = 1.

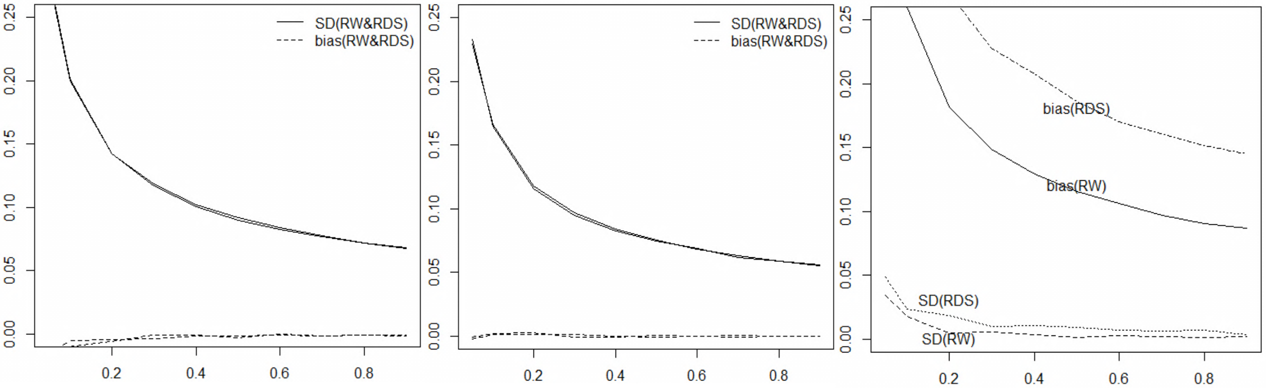
(b) Homophily ratio = 2.

Figure 4.1. The relative mean standard errors of RW and RDS (n=3) on different social network graphs; activity ratio = 0.5, 1, 2 from left to right; the horizontal axis is the sampled proportion.

All social network graphs of different activity ratios and homophily ratios indicate that the mean standard error generally decreases with more sampled nodes, though fast at first and quite slow later. Also, the patterns of RW and RDS curves were highly similar so they almost coincide and no one seems to clearly outperform another when there is no homophily. However, when the homophily ratio is large, RW tend to outperform RDS to a large extent, and all errors tend to be much larger. Meanwhile, there is a tendency that the mean standard error does not decrease rapidly with an increase in sample volume. In contrast, they seem to come close to a certain positive number (in these examples, around 0.10 and 0.20) with the volume increasing. This indicates that overly large samples may be valueless and how to further decrease the error in RW and RDS sampling should be a question of investigation.

It was also found in simulation that when the homophily ratio is high, there are some outliers that significantly change the mean standard errors, and the standard error is also higher. This is likely to be caused by the fact that the sampler is trapped in a small group of people with dense ties and does not come out. This kind of phenomenon should be avoided in both theory and practice. [9] provided an approach to eliminate this phenomenon by identifying the close social circles. However, such an identification is not always available, and more methods to eliminate this phenomenon should be discovered.

It is reasonable to hypothesize that the mean standard error does not quickly converge to zero as sample size increases, the sampler picks up some points with extreme node degrees, which increase the sample variation of the node degrees to a large extent. The following simulation results, with sampling errors decomposed into bias and standard deviation, largely prove this hypothesis, which is shown as follows:



(a) activity ratio = 0.5, homophily ratio = 1;

(b) activity ratio = 1, homophily ratio = 1;

(c) activity ratio = 2, homophily ratio = 2.

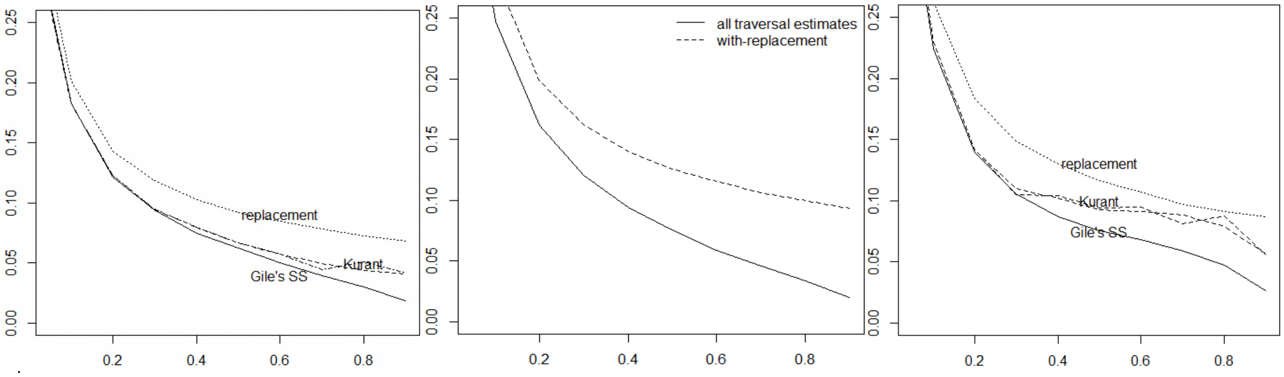
Figure 4.2. Relative biases and standard deviations (divided by proportion) for RW and RDS (n=3) methods against sampled proportion.

We can observe from the graphs that the sampling bias is almost zero, and the main contributor to the error is the standard deviation. The standard deviation decreases rather slowly as the sampling proportion increases, and remains at about 5% of the real value when there is no homophily. Therefore more work should be done on decreasing the standard deviation of the estimator. We will compare these results with the results from traversal (without-replacement) sampling below.

4.1.2 Traversal (Without-replacement) Sampling

Despite the fact that the accurate sampling probabilities of traversal methods are not available, it is still beneficial to discuss how approximate methods work, especially for RW and RDS sampling without replacements. [22] approximates the situations to sampling with replacements and [9] gives the Kurant approximation to estimate the node degree distribution, which could be modified to estimate the distribution for each node and furthermore, to estimate the degree properties. Though, the formula involves the actual node degree distribution itself, which can only be estimated through some other means and thus making methods with replacements more preferrable. Also, it might be applicable to simplify the approximation and eliminate the use the other methods, which has been discussed in Section 3.1. Moreover, Gile’s SS estimator is also available for traversal sampling. Thus it would be beneficial to discuss how these methods actually work in practice: (1) Horvitz-Thompson estimator using Kurant approximation; (2) Horvitz-Thompson estimator using simplified Kurant approximation; (3) Gile’s SS estimator.

The following graphs illustrate the relationship between relative mean standard errors and different sampling and estimation approaches:



(a) activity ratio = 0.5, homophily ratio = 1;

(b) activity ratio = 1, homophily ratio = 2;

(c) activity ratio = 2, homophily ratio = 2.

Figure 4.3. The relative mean standard errors of traversal RW, plotted against the sampling proportion; average node degree = 20; homophily ratio = 1; Kurant approximations yield almost identical errors and are not distinguished.

It can be referred from the results that all traversal estimates consistently have smaller average errors than with-replacement estimates, which shows that traversal estimates are more applicable. In the traversal estimates, Gile's SS estimates perform consistently better, but the difference is only significant when the sampled proportion is large (above 0.4) and the activity ratio is far from zero, which is quite uncommon in real practice. Considering the fact that Kurant approximations are much easier to calculate than Gile's SS estimates, it is acceptable to replace Gile's SS estimates with Kurant estimates under "normal" circumstances, where the sampling proportion is not too high. Also, it is shown that the simplified Kurant estimates almost have the same results as original Kurant estimates.

Further investigation can be made into the similarity between the Gile's SS estimates and the Kurant estimates. The correlation factors between the two estimates under different circumstances (RW & RDS; homophily ratio = 1, 2; activity ratio = 0.5, 1, 2) have been calculated and the numbers are all above 0.95 for sampling proportion <= 0.7. When the sampling proportion > 0.7 and activity ratio is far from 1, the correlation factors drop below 0.95 but still are greater than 0.5. Therefore, the two estimates are interchangeable under "regular" circumstances.

A temporary conclusion is that the traversal methods works best with approximate bias corrections, while with-replacement sampling is acceptable, though they generally yield larger errors. More investigation should be made into the nature of the errors and how to minimize them. Various factors that influence the estimation should also be identified.

4.1.3 Stability Regarding Population Size

Because online social networks are often of very large size, it is beneficial to discuss the change of the error when the population size increases. We can see from below that the errors are quite stable when the population size increases and the sampling size keeps the same. The following table indicates the relative standard errors for simplified Kurant method and Gile's SS method on RDS when sample size equals 700, 1000, 1300, 1700, 2000 and sample size remains 100:

Population size	700	1000	1300	1700	2000
Kurant	14.6%	15.0%	15.1%	15.2%	15.3%
Gile’s SS	14.6%	15.0%	15.1%	15.2%	15.3%

Table 4.1 Sampling errors under different population sizes.

From these results, we can observe that the sampling errors increase quite slowly with population size increasing, and the increasing trend tend to diminish as population increases. Therefore, the results obtained are quite stable, which provide some evidence that the results will be asymptotically stable as well. Therefore, we are sure to some extent that the RDS estimates for the characteristics of networks of large population size are also reliable.

4.2 Confidence Interval Estimation (Naive, Bootstrap and Gile’s SS Confidence Intervals)

Constructing confidence intervals for node properties should be another point of interest. While “naive” confidence intervals[‡] do not perform satisfactorily generally because of the bias of the mean in RDS search, there are currently two types of bootstrap confidence intervals which are acceptable for RDS sampling on networks.

4.2.1 Available Confidence Intervals

The first type is Salganik’s bootstrap confidence interval [23]. We first take out a sample from a network using the RDS method and then resample on the sample with replacements for N times. The resampling procedure is that if the last resampled node is of category 1, we resample the next node from all sampled nodes referred by category 1, and vice versa. Then we calculate the values $\hat{x}_1$, ..., $\hat{x}_n$ of an estimator of whichever type for each of the resampled samples and hence obtain the standard deviation (estimated) for the estimator. Bootstrap confidence intervals are then constructed using $\left[\hat{x} - z_{\alpha/2}\hat{s}_e, \hat{x} + z_{\alpha/2}\hat{s}_e\right]$. It is also suggested that taking the percentiles of $\hat{x}_1$, ..., $\hat{x}_n$ is also acceptable, which can be regarded as an alternative way of constructing bootstrap confidence intervals.

The second type is Gile’s bootstrap confidence interval [31]. Because high homophily can significantly increasing the errors of the estimates, as is shown previously, reliable confidence intervals should take the homophily factor into account, while the Salganik’s bootstrap confidence interval fails to do so directly. To solve this problem, the following set of procedures (outline) is proposed for estimating category proportions, categories denoted as 0 and 1. The essence is to mimic the process of sampling a network of high homophily ratio:

(1) Estimate the node degree distribution of the network and simulate a network with same population and estimated node degree distribution.

(2) Denote current number of nodes with category i and degree k as $\widehat{\mathbb{N}}_{i,k}$; number of edges from category i to category j as the i-j th entry of $\mathbf{H}_0$; estimate $\mathbf{H}_0(1,1) = \bar{\mathrm{d}}_1 \sum \widehat{\mathbb{N}}_{1,k}\, r_1$, $\mathbf{H}_0(0,0) = \bar{\mathrm{d}}_0 \sum \widehat{\mathbb{N}}_{0,k}\,(1 - r_0)$, $\mathbf{H}_0(1,0) = \mathbf{H}_0(0,1) = \left[\bar{\mathrm{d}}_0 \sum \widehat{\mathbb{N}}_{0,k}\, r_0 + \bar{\mathrm{d}}_1 \sum \widehat{\mathbb{N}}_{0,k}\,(1 - r_1)\right]/2$, mean degree $\bar{\mathrm{d}}_i = \sum_k \widehat{\mathbb{N}}_{i,k}\, k / \sum_k \widehat{\mathbb{N}}_{i,k}$, r_i be the sample proportion of edges from i to 1 in all edges from i.

[‡] The “naive” confidence interval is the normal confidence interval with formula $[\bar{\mathrm{x}} - \mathrm{t}_{\alpha/2,\mathrm{n}-1}\hat{\mathrm{s}}_{\mathrm{e}}/\sqrt{n}, \bar{\mathrm{x}} + \mathrm{t}_{\alpha/2,\mathrm{n}-1}\hat{\mathrm{s}}_{\mathrm{e}}/\sqrt{n}]$.

(3) Select n_0 nodes with probabilities proportional to degree, update $\mathbf{H} = \mathbf{H}_0$ · total degrees of nodes unsampled / total node degrees;

(4) While resample size < observed sample size, make the n-th sampled node active in the n-th iteration:

(5) Denote the number of nodes we wish to sample from the active node as m; set m to be the average number of referrals in the original sample; repeat m times:

(6) Select the category 1 with odds H (active node category, 1) / H (active node category, 0); sample an additional node with the desired category with probabilities proportional to degree; update H as before.

(7) Finish the algorithm and create one bootstrap sample. We calculate the bootstrap confidence intervals just as we do in the Salganik's method.

[31] has showed that Gile's procedures tend to slightly overestimate the standard variance, and therefore can be viewed as a little conservative. Yet it is also shown that the coverage probabilities of the Gile's SS confidence intervals can fall below the nominal confidence levels, especially when the sampling procedures are biased. This phenomenon can only be caused by the fact that the estimator itself can be biased, and is almost negligible as a whole according to [31].

Furthermore, Gile used the method of simulating one network instead of taking an average on different networks. To know whether the main variance comes from inside the networks or between the networks, we perform an ANOVA analysis on proportion estimation of 20 networks with size 1000 with 25 samples with size 100 on each network, with the same assumptions as before. The result shows that the right-tail probability of the F statistic is 0.3725. Therefore, it is reasonable to hypothesize that there is not a significant structural difference between different networks, and Gile's approach is reasonable and should be continued.

4.2.2 A Faster Confidence Interval for RDS Sampling

Gile's confidence intervals used the mixing matrix to mimic the process of sampling a network of higher homophily. Nevertheless, we can directly simulate a network with given mean degree, homophily ratio and activity ratio, which includes all information in the mixing matrix [32]. Our new confidence interval should be based on sampling on such a network.

It is not difficult to estimate the mean degree $\bar{d}$ using the previous simplified Kurant estimates or Gile's SS estimates. To estimate the homophily ratio, we assume that every link is sampled with the same probability, and then the proportion of links between category 1 and 0 on average should be $\frac{n_0 n_1}{N(N-1)/2}$; n_1 and n_0 are the estimated sample population of categories 1 and 0. Then, given number of referrals C and number of referrals between category 1 and 0: C_{10}, the estimated homophily ratio $\hat{h} = \frac{C n_0 n_1}{C_{10} N(N-1)/2}$. Similarly, the activity ratio $\hat{a} = \bar{d}_1 / \bar{d}_0$; $\bar{d}_1$ and $\bar{d}_0$ are the estimated mean degrees of nodes of category 1 and 0 using simplified Kurant estimates or Gile's SS estimates.

From these results, we can now construct a faster confidence interval for the RDS sampling with the following steps:

(1) Simulate a network with the same sample size and population, estimated mean degree $\bar{d}$, homophily ratio $\hat{h}$ and activity ratio $\hat{a}$;

(2) Sample from the network m times and create m samples of the simplified Kurant estimator;

(3) Compute the sample variance using the values of the estimator and hence construct the desired confidence interval as before.

To increase the speed of the computation, we use the simplified Kurant estimates, which are about three times as fast as Gile's SS estimates under R console.

4.2.3 Discussion

To investigate on the performance of different confidence intervals, we should know about their coverage probabilities. Moreover, to provide more information, the estimated standard deviation used the generate the confidence intervals are also presented. The simulation results for all three different kinds of confidence intervals are as follows:

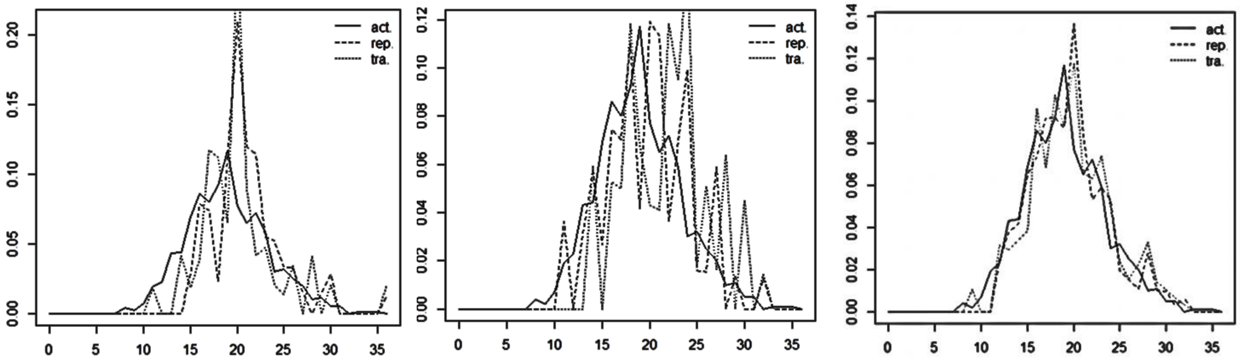
	Coverage probabilities (Salganik's)	Standard deviation (Salganik's)	Coverage probabilities (Gile's SS)	Standard deviation (Gile's SS)	Coverage probabilities (new)	Standard deviation (new)
Case A	94.8%	0.046	94.0%	0.045	95.0%	0.046
Case B	96.1%	0.032	94.2%	0.030	94.4%	0.028
Case C	91.3%	0.041	93.2%	0.043	89.2%	0.042

Table 4.2 Performance of different kinds of 95% bootstrap confidence intervals using traversal RDS (n=3), based on an average of the results from simulated networks. Case A: homophily ratio = 1, activity ratio = 1. Case B: homophily ratio = 1, activity ratio = 2. Case C: homophily ratio = 2, activity ratio = 2.

It can be observed that when the network is without specific characteristics such as activity ratio and homophily, the new confidence intervals tend to outperform the original ones. However, when there is such properties, especially homophily, the new confidence intervals are more inaccurate, also with shorter lengths. The fluctuations in the coverage probabilities indicate the complexity of this problem. Still, the three sets of confidence intervals do not differ by a great amount in terms of coverage probabilities. Therefore, the new confidence intervals con be applied when the sampled network does not specific characteristics, or the requirement for accuracy is not high and faster computational speed is desired.

4.3. Bootstrap Accuracy for RDS

We should also investigate on how well the distribution of a sample represents the total. Again we use node degree as an example. [7] showed that often hundreds of nodes had to be sampled in order to make the estimated sampling probabilities close to the sampling probabilities. We now consider the node degree distribution and have found similar results:



(a) Sample = 50. (b) Sample = 100. (c) Sample = 200.

Figure 4.4 The approximation of node degree distribution by RDS (n=3) under different sample sizes, mean network degree = 20. act: actual node degree distribution; rep: sampling with replacements; tra: traversal sampling.

It is observed that it is only with a sample size of 200 (20%) that the sample node degree distribution is relatively close to the actual node degree distribution. Therefore, both crawling methods do not converge fast enough to resemble the actual distribution. Nevertheless, there exists many ways to increase the convergence speed for random walks, which can be found in [24-28].

5. Conclusion

We have discussed a number of sampling methods for social networking services and many of these methods prove to be useful in analyzing the structure of online social networks. While some methods are typically used for some special purposes, other methods can work for general purposes. The characteristics of methods of statistical inference for RDS sampling are also discussed, and a faster confidence interval for RDS sampling is also proposed. The remaining questions are the properties of these procedures and how to refine them.

Reference

[1] S. H. Lee, P.-J. Kim, and H. Jeong. Statistical properties of Sampled Networks. *Phys. Rev. E*, 73:16102, 2006.

[2] K. Lewis, J. Kaufman, M. Gonzalez, A. Wimmer, and N. Christakis, “Tastes, ties, and time: A new social network dataset using Facebook.com,” *Social Networks*, 2008.

[3] A. L. Traud, E. D. Kelsic, P. J. Mucha, and M. A. Porter, “Community structure in online collegiate social networks,” 2008, *arXiv*:0809.0960.

[4] M. Gjoka, M. Kurant, C. T. Butts, and A. Markopoulou. Walking in Facebook: A Case Study of Unbiased Sampling of OSNs. In *INFOCOM*, 2010.

[5] A. Mislove, H. S. Koppula, K. P. Gummadi, P. Druschel, and B. Bhattacharjee. Growth of the Flickr social network. In *WOSN*, 2008.

[6] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee. Measurement and analysis of online social networks. In *IMC,* pages 29–42, 2007.

[7] A. Mohaisen, A. Yun, and Y. Kim. Measuring the mixing time of social graphs. *IMC,* 2010.

[8] C. Wilson, B. Boe, A. Sala, K. P. N. Puttaswamy, and B. Y. Zhao. User interactions in social networks and their implications. In *EuroSys*, 2009.

[9] M. Kurant, A. Markopoulou, P. Thiran. On the bias of BFS (Breadth First Search)". *International Teletraffic Congress* (*ITC 22*), 2010.

[10] M. Kurant, A. Markopoulou, P. Thiran. "Towards Unbiased BFS Sampling". *IEEE JSAC* 29 (9):1799-1809, 2011.

[11] M. Kurant, M. Gjoka, C. Butts, and A. Markopoulou. Walking on a Graph with a Magnifying Glass. *Arxiv* preprint *arXiv:1101.5463,* 2011.

[12] S. Even, *Graph Algorithms (2nd ed.),* Cambridge University Press, pp. 46–48, ISBN 9780521736534, 2011.

[13] L. Lov´asz. Random walks on graphs: A survey. *Combinatorics, Paul Erdos is Eighty*, 2(1):1–46, 1993.

[14] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculation by fast computing machines. *Journal of Chemical Physics*, 21:1087–1092, 1953.

[15] M. Hansen and W. Hurwitz. On the Theory of Sampling from Finite Populations. *Annals of Mathematical Statistics,* 14(3), 1943.

[16] A. Rasti, M. Torkjazi, R. Rejaie, N. Duffield, W. Willinger, and D. Stutzbach, "Evaluating sampling techniques for large dynamic graphs," *University of Oregon, Tech. Rep.*, Sept 2008.

[17] A. Rasti, M. Torkjazi, R. Rejaie, N. Duffield, W. Willinger, and D. Stutzbach, "Respondent-driven sampling for characterizing unstructured overlays," in *INFOCOM Mini-Conference,* April 2009.

[18] J. Leskovec and C. Faloutsos, "Sampling from large graphs," in *Proc. of ACM SIGKDD*, 2006.

[19] J. Illenberger, G. Flötteröd, and K. Nage, "An approach to correct bias induced by snowball sampling," *Sunbelt Social Networks Conference*, 2009.

[20] L. Goodman, "Snowball sampling," *Annals of Mathematical Statistics*, vol. 32, p. 148170, 1961.

[21] D. Heckathorn, "Respondent-driven sampling: A new approach to the study of hidden populations," *Social Problems,* vol. 44, p. 174199, 1997.

[22] M. Salganik and D. Heckathorn, "Sampling and estimation in hidden populations using respondent-driven sampling," *Sociological Methodology*, vol. 34, p. 193239, 2004.

[23] M. J. Salganik. *Variance Estimation, Design Effects, and Sample Size Calculations for Respondent-Driven Sampling*, Journal of Urban Health, 83:98-111, 2006.

[24] B. Ribeiro and D. Towsley. Estimating and sampling graphs with multidimensional random walks. In *IMC*, volume 011, 2010.

[25] K. Avrachenkov, B. Ribeiro, and D. Towsley. Improving Random Walk Estimation Accuracy with Uniform Restarts. In *I7th Workshop on Algorithms and Models for the Web Graph*, 2010.

[26] J. Leskovec and C. Faloutsos. Sampling from large graphs. In *KDD*, pages 631–636, 2006.

[27] M. Gjoka, C. Butts, M. Kurant, and A. Markopoulou. Multigraph Sampling of Online Social Networks. *arXiv*, (*arXiv*:1008.2565v1):1–10, 2010.

[28] S. Boyd, P. Diaconis, and L. Xiao. Fastest mixing Markov chain on a graph. *SIAM review*, 46(4):667–689, 2004.

[29] K. J. Gile, M. S. Handcock. Respondent-driven Sampling: An Assessment of Current Methodology, *Sociological Methodology*, 40, 285-327, 2010.

[30] K. J. Gile. Improved Inference for Respondent-Driven Sampling Data with Application to HIV Prevalence Estimation. *Journal of the American Statistical Association*, 106, 135-146, 2011.

[31] K. J. Gile. Supplemental Materials: Improved Inference for Respondent-Driven Sampling Data with Application to HIV Prevalence Estimation. *Journal of the American Statistical Association*, 106, 2011.

[32] M. S. Handcock, W. W. Neely, K. J. Gile, B. Draper. RDSdevelopment: Development Package for Respondent-Driven Sampling. Version Package. Project home page at http://hpmrg.org, URL http://CRAN.R-project.org/package=RDS, 2009.