Subsampling for Chain-Referral Methods

Konstantin Avrachenkov, Giovanni Neglia, Alina Tuholukova

Subsampling for Chain-referral Methods

Konstantin Avrachenkov, Giovanni Neglia, Alina Tuholukova

November 23, 2016

Abstract

We study chain-referral methods for sampling in social networks. These methods rely on subjects of the study recruiting other participants among their set of connections. This approach gives us the possibility to perform sampling when the other methods, that imply the knowledge of the whole network or its global characteristics, fail. Chain-referral methods can be implemented with random walks or crawling in the case of online social networks. However, the estimations made on the collected samples can have high variance, especially with small sample size. The other drawback is the potential bias due to the way the samples are collected. We suggest and analyze a subsampling technique, where some users are requested only to recruit other users but do not participate to the study. Assuming that the referral has lower cost than actual participation, this technique takes advantage of exploring a larger variety of population, thus decreasing significantly the variance of the estimator. We test the method on real social networks and on synthetic ones. As by-product, we propose a Gibbs like method for generating synthetic networks with desired properties.

1 Introduction

Online social networks (OSNs) are thriving nowadays. The most popular ones are: Google+ (about 1.6 billion users), Facebook (about 1.28 billion users), Twitter (about 645 million users), Instagram (about 300 million users), LinkedIn (about 200 million users). These networks gather a lot of valuable information like users' interests, users' characteristics, etc. Great

part of it is free to access. This information can facilitate the work of sociologists and give them modern instrument for their research. Of course, real social networks continue to be of great interest to sociologists as well as online social networks. For example, the Add Health study [2] has built the networks of the students at selected schools in the United States, which served as the basis of much further research [10].

The network, besides being itself an object of study, is also an instrument for collecting data. Starting just from one individual that we observe we can reach other representatives of this network. The sampling methods that use the contacts of known individuals of a population to find other members are called *chain-referral methods*. Crawling of online social networks can be viewed as automatisation of chain-referral methods. Moreover, it is one of the few methods to collect information about *hidden populations*, whose members are, by definition, hard to reach. A lot of research has targeted the study of HIV prevalence in hidden populations like drug users, female sex workers [11], gay men [12]. Another study [9] considered the population of jazz musicians. Even if jazz musicians have no reasons to hide them, it is still hard to access them with the standard sampling methods.

The problem of the chain-referral methods is that they do not achieve independent sampling from the population. It is frequently observed that friends tend to have similar interests. It can be the influence of your friend that leads you to listening the rock music or the opposite: you became friends because you were both fond of it. One way or another, social contacts influence each other in different ways. The fact that people in contact share common characteristics is usually observed in real networks and is called *homophily*. For instance, the study [6] evaluated the influence of social connections (friends, relatives, siblings) on obesity of people. Interestingly, if a person has a friend who became obese during some fixed interval of time, the chances that this person becomes obese are increased by 57%.

The population sample obtained through chain-referral methods is different from the ideal uniform independent sample and, because of homophily, leads to increased variance of the estimators as we are going to show. The main contribution of this paper is the proposed chain-referral method that allows to decrease the dependency of the collected values by subsampling. Subsampling is done via asking/inferring only contact details of some users without taking any further information.

As by-product of our numerical studies, we develop a Gibbs-like method for generating synthetic attributes' distribution over networks with desired

properties. This approach can be used for extensive testing of methods in social network analysis and hence can be of independent interest.

The paper is organized as follows. In Sec. 2 we discuss different estimators of the population mean and the problem of correlated samples. Sec. 3 presents the subsampling method, that can help to reduce the correlation. In Sec. 4 we evaluate the subsampling method formally, starting from the simple, but intuitive example of a homogeneous correlation (Sec. 4.1), and then moving to the general case (Sec. 4.2). The theoretical results are then validated by the experiments in Sec. 5. Sec. 5.0.3 presents also the method for generating synthetic networks that we used for the experiments together with the real data.

2 Chain-referral Methods and Estimators

Chain-referral methods take advantage of the individuals connections to explore the network: each study participant provides the contacts of other participants. The sampling continues in this way until the needed size of participants is reached.

In order to study formally chain-referral methods we will model the social network as a graph, where the individuals are represented by nodes and a contact between two individuals is represented by an edge between the corresponding nodes. We will make the following assumptions:

1. One individual can refer exactly another individual, selected uniformly at random from his contacts;

2. The same individual can be recruited multiple times;

3. If individual A knows individual B then individual B knows A as well (the network can be represented as an undirected graph);

4. Individuals know and report precisely their number of connections (i.e. their degree);

5. Each individual is reachable from any other individual (the network is connected).

Under these assumptions the referral process can be regarded as a *random walk* on the graph. For the real social networks some of these assumptions

are arguable. There can be inaccuracy in the reported degree, and the choice of the contact to refer can be different from uniform. The sensitivity to violation of some assumptions is studied in [7]. However, it is simpler to design chain-referral methods for online social networks, that satisfy all these assumptions. For example, the individual may be asked to disclose his whole list of contacts (if not already public) and the next participant can then be selected uniformly at random from it.

The random walk is represented by the transition matrix P with elements:

$$p_{ij} = \begin{cases} \frac{1}{d_i} & \text{if } i \text{ and } j \text{ are neighbors,} \\ 0 & \text{if } i \text{ and } j \text{ are not neighbors,} \\ 0 & \text{if } i = j, \end{cases}$$

where d_i is the degree of the node i.

We denote as g_j the value of interest at node j. We are interested to estimate the population average $\mu = \frac{\sum_{i=1}^{m} g_i}{m}$, where m is the population size.

Moreover, let us denote the value that is observed at step i of the random walk as y_i. Some estimators were developed in order to draw conclusions about the population average μ from the collected sample $y_1, y_2, ...y_n$. The simplest estimator of the population mean is the **Sample Average** (SA) estimator:

$$\widehat{\mu}_{SA} = \frac{y_1 + y_2 + ... + y_n}{n}.$$

This estimator is biased towards the nodes with large degrees. Indeed the individuals with more contacts are more likely to be sampled by the random walk. In particular, the probability at a given step to encounter node i is proportional to its degree d_i. To correct this bias the **Volz-Heckathon** (VH) estimator, which was introduced in [13], weights the responses from individuals according to their number of contacts:

$$\widehat{\mu}_{VE} = \frac{1}{\sum_{i=1}^{n} 1/d_i} \sum_{i=1}^{n} \frac{y_i}{d_i}.$$

2.0.1 Problem of Samples Correlation

Due to the way the sample was collected the variance of both estimators will be increased in comparison to the case of independent sampling. Our

theoretical analysis will focus on the SA estimator, as for the VH estimator it becomes too complicated and we leave its analysis for future research. However, we consider the VH estimator in the simulations.

The variance of the estimator in the case of independent sampling with replacement is approximated by σ^2/n for large population size, where σ^2 is the population variance. If samples are not independently selected, then a correlation factor $f(n, \mathcal{S})$ should be considered as follows:

$$\sigma^2_{\hat{\mu}_{\mathcal{S}}} = \frac{\sigma^2}{n} f(n, \mathcal{S}). \tag{1}$$

This correlation factor $f(n, \mathcal{S})$ depends on the sampling method $\mathcal{S}$ as well as on the size of the sample. We observe that $f(n, \mathcal{S})$ is an increasing function of n bounded by 1 and n. The less the samples obtained through the sampling method $\mathcal{S}$ are correlated, the smaller we expect $f(n, \mathcal{S})$ to be.

In what follows we consider chain-referral methods when only one individual out of k is asked for his value. Among these methods the correlation factor $f(n, \mathcal{S})$ will be a function of the number of values collected, n, and of k, so we can write $f(n, k)$. We expect $f(n, k)$ to be decreasing in k.

3 Subsampling Technique

In order to reduce correlation between sampled values we will try to decrease the dependency of the samples. Our idea is to thin out the sample. Indeed, the farther are the individuals in the chain from each other, the smaller is the dependency between them. Imagine to have contacted an even number h individuals, but to ask the value of interest only to every second of them. We can use then the $n = h/2$ values. It should be observed that, while we reduce in this way the correlation factor (because $f(h/2, 2) < f(h, 1)$), we also reduce by 2 the number of samples used in the estimation. Then while $f(n, k)$ becomes smaller in Eq. (1) because of the reduction of the correlation, it is not clear if $\frac{f(n,k)}{n}$ becomes smaller.

Another potential advantage originates from the fact that the cost of the referring is less than the cost of the actual sampling. For example, the information about the friends in Facebook is generally available, thus you can serf through the Facebook graph by writing a simple crawler. On the contrary retrieving the information of interest can be more costly and one may need to provide some form of incentives to participants to encourage

them to answer some questionnaires. In other context, one may need to pay the users also to reveal the identity of one of his contacts.

Among the individuals in the collected chain some of them will be asked both: to participate in the tests and provide the reference, let us call them *participants*. Some of them will be asked only to recruit other participants, let us call them *referees*. We will look at the strategy when only each k-th individual in the chain is a participant. Thus between 2 participants there are always $k-1$ referees. We will call this approach *subsampling with step* k. Let C_1 be the payment for providing the reference and C_2 the payment for the participation in the test. In this way, every referee receives C_1 units of money and every participant receives $C_1 + C_2$ units of money (C_1 for the reference and C_2 for the test). In this way, for a fixed budget B, if $C_2 > 0$, the subsampling decreases less in the number of samples.

It is evident that the bigger is k, the lower is the correlation between the selected samples. However the choice of the k is not evident: if we take it too small the dependency can be still high; if we take it too big the sample size will be inadequate to make conclusions. It also depends on the level of homophily in the network: with the low level of homophily the best choice would be to take k equal to 1, what means no referees only participants. In the following section we formalize the qualitative results derived here and we determine the value k, such that the profit from the subsampling is maximal.

4 Analysis

In this section we study formally the effect of subsampling. We start with a case when the collected samples are correlated in a known and homogeneous way. While being a too simplified model for the chain-referral methods, it illustrates the main idea of subsampling. We proceed then with the general case, when the samples are collected through the random walk on a general graph.

4.1 Simple Example: Variance with Geometric Correlation

First we will quantify the variance of the estimator for a simple case with defined correlation between the samples in the chain. We will assume that collected samples $Y_1, Y_2, ..., Y_n$ are correlated in the following way:

$$\text{corr}(Y_i, Y_{i+l}) = \rho^l.$$

In this way the nodes that are at the distance 1 in the chain have correlation ρ, at distance 2 have correlation ρ^2 an so on[1]. We will refer to this model as the *geometric model* [2]. If the population variance is σ^2, then we can obtain the variance of the SA estimator in the following way:

$$\sigma^2_{\hat{\mu}_{SA}} = \text{Var}\left[\bar{Y}\right] = \text{Var}\left[\frac{Y_1 + Y_2 + ... + Y_n}{n}\right] = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\text{Cov}(Y_i, Y_j) =$$

$$= \frac{\sigma^2}{n^2}\left(n + 2\sum_{i=1}^{n-1}(n-i)\rho^i\right) = \frac{\sigma^2}{n^2}\left(n + 2n\sum_{i=1}^{n-1}\rho^i - 2\sum_{i=1}^{n-1}i\rho^i\right) =$$

$$= \frac{\sigma^2}{n}\left(n + 2n\frac{\rho - \rho^n}{1-\rho} - 2\rho\left(\frac{\rho - \rho^n}{1-\rho}\right)'\right) = \frac{\sigma^2}{n^2}\frac{n - n\rho^2 - 2\rho + 2\rho^{n+1}}{(1-\rho)^2}.$$

From here we can get that correlation factor as:

$$f(n,1) = \frac{1 - \rho^2 - 2\rho/n + 2\rho^{n+1}/n}{(1-\rho)^2}.$$

It can be shown that this factor $f(n)$ is an increasing function of $n \in \mathbb{N}$ and it achieves its minimum value 1 when $n = 1$. It is clear, when there is only one individual there is no correlation, because we consider single random variable Y_1. When new participants are invited, the correlation increases due to homophily as we explained earlier.

Let us consider what happens to the correlation factor when n goes to infinity:

$$f(n,1) = \frac{1 - \rho^2 - 2\rho/n + 2\rho^{n+1}/n}{(1-\rho)^2} \quad \xrightarrow{n\to\infty} \quad \frac{1-\rho^2}{(1-\rho)^2} = \frac{1+\rho}{1-\rho},$$

[1]We are ignoring here the effect of resampling

[2]It could be adopted to model the case where nodes are on a line and social influences are homogeneous.

and then $f(n,1) \leq \frac{1+\rho}{1-\rho} \quad \forall n$. Using this upper bound the expression for the SA estimator variance can be bounded as $\sigma^2_{\hat{\mu}_{SA}} \leq \frac{\sigma^2}{n}\frac{1+\rho}{1-\rho}$.

This bound is very tight when n is large enough, so that it can be used as a good approximation:

$$\sigma^2_{\hat{\mu}_{SA}} \simeq \frac{\sigma^2}{n}\frac{1+\rho}{1-\rho}.$$

Figure 1 compares the approximated expression with original one, when the parameter ρ is 0.6. As it is reasonable to suppose that the sample size is bigger than 50, we can consider this approximation good enough in this case. The reason to use this approximation is that the expression becomes much simpler to illustrate the main idea of the method.

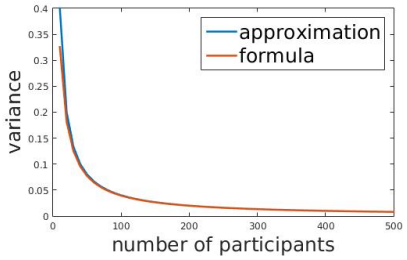


Figure 1: $\rho = 0.6$

4.1.1 Variance for subsampling

Here we will quantify the variance of the SA estimator on the subsample. For simplicity let us take $h = nk$, where the collected samples $Y_1, Y_2, Y_3, ..., Y_{nk}$ have again geometric correlation. We will take each k sample and look at the variance of the following random variable:

$$\bar{Y}^k = \frac{Y_k + Y_{2k} + Y_{3k} + ... + Y_{nk}}{n}.$$

Let us note that the correlation between the variables Y_{ik} and $Y_{(i+l)k}$ is:

$$\text{corr}(Y_{ik}, Y_{(i+l)k}) = \rho^{kl}.$$

Using the result of Sec. 4.1, we obtain:

$$\text{Var}\left[\bar{Y}^k\right] = \frac{\sigma^2}{n}\frac{1 - \rho^{2k} - 2\rho^k/n + 2\rho^{k(n+1)}/n}{(1-\rho^k)^2}.$$

or the approximate form:

$$\mathrm{Var}\left[\bar{Y}^k\right] \simeq \frac{\sigma^2}{n}\frac{1+\rho^k}{1-\rho^k}. \tag{2}$$

4.1.2 Limited Budget

Equation (2) gives the expression for the variance of the subsample, where the number of actual participants is n and two consecutive participants in the chain are separated by $k-1$ referees. It is evident that in order to decrease the variance, one needs to take as many participants as possible separated by as many referees as possible. However both of them have their cost. If limited budget B is available, then a chain of length $h = nk$ with n participants is restricted by the following equality:

$$B \geq hC_1 + nC_2,$$

where each reference costs C_1 units of money and each test costs C_2 units of money. We can express the maximum length of the chain as: $h = \frac{kB}{kC_1+C_2}$, where the number of actual participants is $n = \frac{h}{k} = \frac{B}{kC_1+C_2}$.

The approximate variance of SA estimator becomes as follows:

$$\sigma^2_{\hat{\mu}_{SA}}(k) = \frac{\sigma^2}{\frac{B}{kC_1+C_2}}\frac{1+\rho^k}{1-\rho^k}. \tag{3}$$

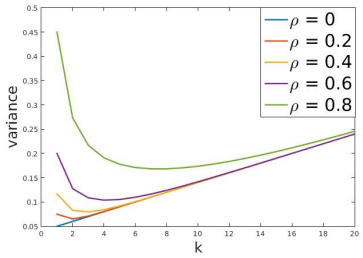


Figure 2: Variance with Equation 3 when $B = 100, C_1 = 1, C_2 = 4$

Let us observe what happens to the factors of the variance when we increase k. The first factor in (3) increases in k: the variance increases due to smaller sample size. The second factor decreases in k: the observations

become less correlated. Finally, the behavior of the variance depends on which factor is "stronger."

We can observe the trade-off in Figure 2: initially increasing the subsampling step k can help to reduce the estimator variance. However, after some threshold the further increase of k will only add to the estimator variance. Moreover, this threshold depends on the level of correlation, that is expressed here by the parameter ρ. We observe from the figure that the higher is ρ the higher is the desired k. This coincides with our intuition: the higher is the dependency, the more values we need to skip. Finally we see, that in case of no correlation ($\rho = 0$) skipping nodes is useless.

4.2 General Case

Even if the geometric model is not realistic, it allowed us to better understand the potential improvement from subsampling. This section will generalize this idea to the case where the samples are collected through a random walk on a graph with m nodes. We consider first the case without subsampling ($k = 1$).

Let $g = (g_1, g_2, ..., g_m)$ be the values of the attribute on the nodes $1, 2, ..., m$. Let P be the transition matrix of the random walk.

The stationary distribution of the random walk is:

$$\pi = \left(\frac{d_1}{\sum_{i=1}^{n} d_i}, \frac{d_2}{\sum_{i=1}^{n} d_i}, ..., \frac{d_n}{\sum_{i=1}^{n} d_i} \right),$$

where d_i is the degree of the node i.

Let Π be the matrix that consists of m rows, where each row is the vector π. If the first node is chosen according to the distribution π, then variance for any sample Y_i [3] is the following:

$$\mathrm{Var}(Y_i) = < g, g >_\pi - < g, \Pi g >_\pi, \text{where } < a, b >_\pi = \sum_{i=1}^{m} a_i b_i \pi_i.$$

and covariance between the samples Y_i and Y_{i+l} is the following [5, chapter 6]:

$$\mathrm{Cov}(Y_i, Y_{i+l}) = < g, (P^l - \Pi) g >_\pi,$$

[3] Note that $Y_i = g_j$ if the random walk is on node j at the i-th step.

Using these formulas we can write the formula for the variance of the estimator as:

$$\mathrm{Var}\left[\bar{Y}\right] = \frac{1}{n^2}\left(n\mathrm{Var}(Y_i) + 2\sum_{i=1}^{n}\sum_{j|i<j}^{n}\mathrm{Cov}(Y_i, Y_j)\right) =$$

$$= \frac{1}{n^2}\left(n(< g, g >_\pi - < g, \Pi g >_\pi) + 2\sum_{i=1}^{n}\sum_{j|i<j}^{n} < g, (P^{j-i} - \Pi)g >_\pi\right) \quad (4)$$

Eq. (4) is quite cumbersome: computing large powers of the m by m matrix P can be unfeasible. Using the spectral theorem for diagonalizable matrices:

$$\mathrm{Var}\left[\bar{Y}\right] = \frac{1}{n}\sum_{i=2}^{m}\frac{1 - \lambda_i^2 - 2\frac{\lambda_i}{n} + 2\frac{\lambda_i^{n+1}}{n}}{(1-\lambda_i)^2} < g, v_i >_\pi^2, \quad (5)$$

where $\lambda_i, v_i, u_i (i = 1..m)$ are respectively eigenvalues, right eigenvectors and left eigenvectors of the auxiliary matrix P^* [4], defined as $P^* \triangleq D^{\frac{1}{2}}PD^{-\frac{1}{2}}$, where D is the $m \times m$ diagonal matrix with $d_{ii} = \pi_i$.

In the case of subsampling similar calculation can be carried on leading to:

$$\mathrm{Var}\left[\bar{Y}^k\right] = \frac{1}{\frac{B}{kC_1+C_2}}\sum_{i=2}^{m}\frac{1 - \lambda_i^{2k} - 2\frac{\lambda_i^k}{\frac{B}{kC_1+C_2}} + 2\frac{\lambda_i^{k\left(\frac{B}{kC_1+C_2}+1\right)}}{\frac{B}{kC_1+C_2}}}{(1-\lambda_i)^{2k}} < g, v_i >_\pi^2 . \quad (6)$$

As in the geometric model Eq. (6) can be approximated as follows:

$$\sigma^2_{\hat{\mu}_{SA}} = \mathrm{Var}\left[\bar{Y}^k\right] = \frac{1}{\frac{B}{kC_1+C_2}}\sum_{i=2}^{m}\frac{1+\lambda_i^k}{1-\lambda_i^k} < g, v_i >_\pi^2 .$$

Interestingly, the expression for the variance in the general case has the same structure as for the geometric model. Therefore, the interpretation of the formula is the same. There are two factors, that "compete" with each

[4] Matrix P^* is always diagonalizable for RW on undirected graph.

other. If we try to decrease the first factor, we will increase the second one and the opposite. In order to find the desired parameter k we need to find the minimum of the estimator function for variance. Even if it is difficult to obtain the explicit formula for k, the fact that k is integer allows us to find it through binary search.

The quality of an estimator does not depend only on its variance, but also on its bias:

$$\text{Bias}(\hat{\mu}_{SA}) = E[\hat{\mu}_{SA}] - \mu =< g, \pi > -\mu. \tag{7}$$

Then the mean squared error of the estimator, $MSE(\hat{\mu}_{SA})$, is:

$$MSE(\hat{\mu}_{SA}) = \text{Bias}(\hat{\mu}_{SA})^2 + \text{Var}(\hat{\mu}_{SA}). \tag{8}$$

This bias can be non-null if the quantity we want to estimate is correlated with the degree. In fact, we observe that the random walk visits the nodes with more connections more frequently. Subsampling has no effect on such bias, hence minimizing the variance leads to minimizing the mean squared error.

5 Numerical Evaluation

To validate our theoretical results we performed numerous simulations. We considered both real datasets from the Project 90 [3] and Add health [2], as well as synthetic datasets, obtained through the Gibbs sampler. Both the Project 90 and the Add health datasets contain the graph describing the social contacts as well as information about the users.

5.0.1 Data from the Project 90

Project 90 [3] studied how the network structure influences the HIV prevalence. Besides the data about social connections the study collected some data about drug users, such as race, gender, whether he/she is a sex worker, pimp, sex work client, drug dealer, drug cook, thief, retired, housewife, disabled, unemployed, homeless. For our experiments we took the largest connected component from the available data, which consists of 4430 nodes and 18407 edges.

5.0.2 Data from the Add Health Project

The National Longitudinal Study of Adolescent to Adult Health (Add Health) is a huge study that began surveying students from the 7-12 grades in the United States during the 1994-1995 school year. In general 90,118 students representing 84 communities took part in this study. The study kept on surveying students as they were growing up. The data include, for example, information about social, economic, psychological and physical status of the students.

The network of students' connections was built based on the reported friends by each participant. Each of the students was asked to provide the names of up to 5 male friends and up to 5 female ones. Then the network structure was built to analyze if some characteristics of the students indeed are influenced by their friends.

Though these data are very valuable, they are not freely available. However a subset of the data can be accessed through the link [1] but only with few attributes of the students, such as: sex, race, grade in school and, whether they attended middle or high school. There are several networks available for different communities. We took the graph with 1996 nodes and 8522 edges.

5.0.3 Synthetic Datasets

To perform extensive simulations we needed more graph structures with node attributes.

There is no lack of available real network topologies. For example, the Stanford Large Network Dataset Collection [4] provides data of Online-Social Networks (we will use part of Facebook graph), collaboration networks, web graphs, Internet peer-to-peer network and a lot of others. Unfortunately, in most of the cases, nodes do not have any attribute.

At the same time random graphs can be generated with almost arbitrary characteristics (e.g. number of nodes, links, degree distribution, clustering coefficient). Popular graph models are Erdős-Rényi graph, random geometric graph, preferential attachments graph. Still, there is no standard way to generate synthetic attributes for the nodes and in particular providing some level of homophily (or correlation).

In the same way we can generate numerous random graphs with desired characteristics, we wanted to have mechanism to generate the values on the nodes of the given graph which will represent needed attribute, which will

satisfy the following properties:

1. Nodes attributes should have the property of homophily

2. We should have the mechanism to control the level of homophily

These properties are required to evaluate the performance of the sub-sampling methods. In what follows we derive a novel (to the best of our knowledge) procedure for synthetic attributes generation.

First we will provide some definitions. Let us imagine that we already have a graph with m nodes. It may be the graph of a real network or a synthetic one. Our technique is agnostic to this aspect. To each node i, we would like to assign a random value G_i from the set of attributes $V, V = \{1, 2, 3, ..., L\}$. Instead of looking at distributions of the values on nodes independently, we will look at the joint distribution of values on all the nodes.

Let us denote $(G_1, G_2, ..., G_m)$ as $\dot{G}$. We call $\dot{G}$ a *random field on graph*. When random variables $G_1, G_2, ..., G_m$ take respectively values $g_1, g_2, ..., g_m$, we call $(g_1, g_2, ..., g_m)$ a *configuration* of the random field and we denote it as $\dot{g}$. We will consider random fields with a Gibbs distribution [5].

We can define the *global energy* for a random field $\dot{G}$ in the following way:

$$\varepsilon(\dot{G}) \triangleq \sum_{i \sim j, i \leq j} (G_i - G_j)^2,$$

where $i \sim j$ means that the nodes i and j are neighbors in the graph.

The *local energy* of node i is defined as:

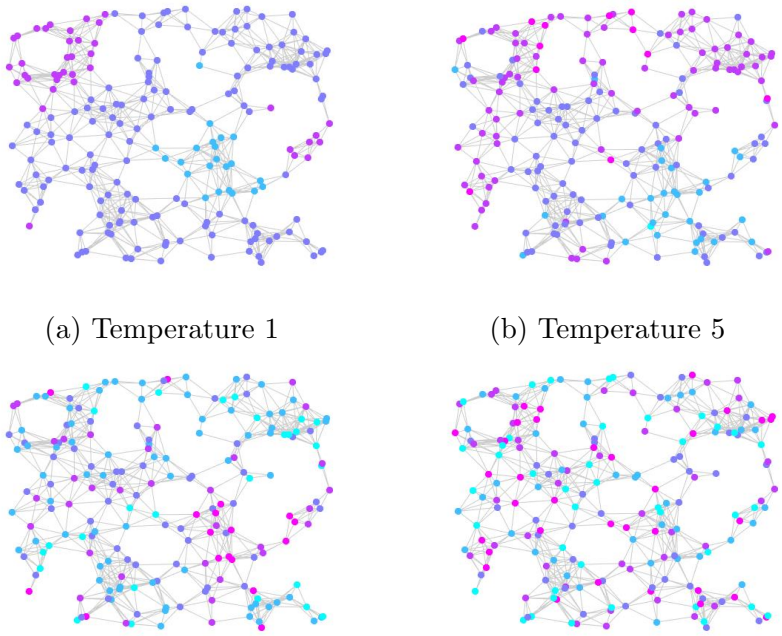
$$\varepsilon_i(G_i) \triangleq \sum_{j | i \sim j} (G_i - G_j)^2.$$

According to the Gibbs distribution, the probability that the random field $\dot{G}$ takes the configuration $\dot{g}$ is:

$$p(\dot{G} = \dot{g}) = \frac{e^{-\frac{\varepsilon(\dot{g})}{T}}}{\sum\limits_{\dot{g}' \in |V|^m} e^{-\frac{\varepsilon(\dot{g}')}{T}}}, \tag{9}$$

where $T > 0$ is a parameter called the temperature of the Gibbs field.

The reason why it is interesting to look at this distribution follows from [5, theorem 2.1]: *when a random field has distribution* (9) *then the probability*



(a) Temperature 1 (b) Temperature 5

(c) Temperature 20 (d) Temperature 1000

Figure 3: RGG(200, 0.13) with generated values for different temperature

that the node has particular value depends only on the values of its neighboring nodes and does not depend on the values of all other nodes.

Let N_i be the set of neighbors of node i. Given a subset L of nodes, we let $\dot{G}_L$ denote the set of random variables of the nodes in L. Then the theorem can be formulated in the following way:

$$p(G_i = g_i|\dot{G}_{N_i} = \dot{g}_{N_i}) = p(G_i = g_i|\dot{G}_{\{1,2,\ldots,m\}\setminus i} = \dot{g}_{\{1,2,\ldots,m\}\setminus i}).$$

This property is called *Markov property* and it will capture the homophily effect: the value of a node is dependent on the values of the neighboring nodes. Moreover, for each node i, given the values of its neighbors, the probability distribution of its value is:

$$p(G_i = g_i) = \frac{e^{-\frac{\varepsilon_i(g_i)}{T}}}{\sum\limits_{g' \in V} e^{-\frac{\varepsilon_i(g')}{T}}}.$$

The temperature parameter T plays a very important role to tune the homophily level (or the correlation level) in the network. Low temperature gives us network with highly correlated values. Increasing temperature we can add more and more "randomness" to the attributes.

In Figure 3 we present the same random geometric graph with 200 nodes and radius 0.13, $RGG(200, 0.13)$ where the values $V = \{1, 2, ..., 5\}$ are chosen according to the Gibbs distribution and depicted with different colors. From the figure we can observe that the level of correlation between values of the node changes with different temperature. When temperature is 1 we can distinguish distinct clusters. When the temperature increases ($T = 5$ and $T = 20$), the values of neighbors are still similar but with more and more variability. When the temperature is very high then the values seem to be assigned independently.

5.1 Experimental Results

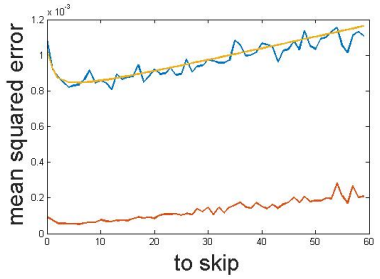
We performed simulations for two reasons: first, to verify the theoretical results; second, to see if subsampling gives improvement on the real datasets and on the synthetic ones.

The simulations for a given dataset are performed in the following way. For the fixed budget B, rewards C_1 and C_2, we first collect the samples through the random walk on the graph for the subsampling step 1. We estimate the population average with the SA and VH estimators. Then we repeat this operation in order to have multiple estimates for the subsampling step 1, that we can count the mean squared error of the estimator. The same process is performed for different subsampling steps. In this way we can compare the mean squared error for different subsampling steps and choose the optimal one.

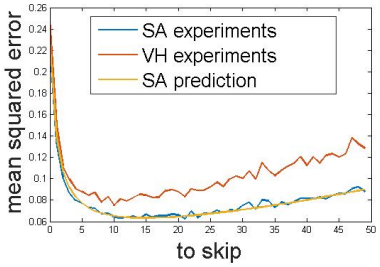
Figure 4 presents the experimental mean squared error of the SA and VH estimators and also the mean squared error of the SA obtained through Eqs. (6), (7), (8) for different subsampling steps. From the figure we can observe that the experimental results are very close to the theoretical ones. We can notice that both estimators gain from subsampling. Another observation is that the best subsampling step differs for different attributes. Thus, for the same graph from Add health study, we observe different optimal k for the attributes grade, gender and school (middle or high school) . The reason is that the level of homophily changes depending on the attribute, even if the graph structure is the same. We obtain the similar results for the synthetic

datasets. We see that for the Project 90 graph the optimal subsampling step for the temperature 100 (low level of homoplily) is lower than for the temperature 10 (high level of homophily).

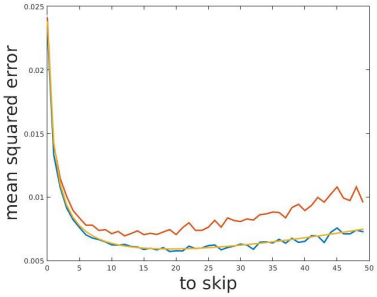
From our experiments we also saw that there is no estimator that performs better in all cases. As stated in [8] the advantage to use VH appears only when the estimated attribute depends on the degree of the node. Indeed, our experiments show the same result.



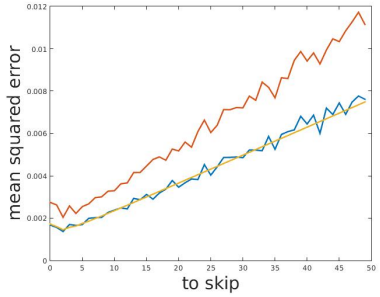
(a) Project 90: pimp



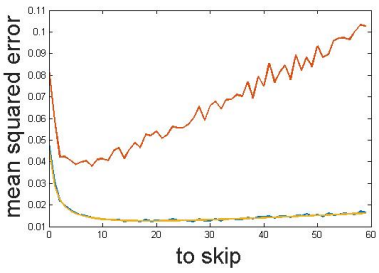
(b) Add health: grade



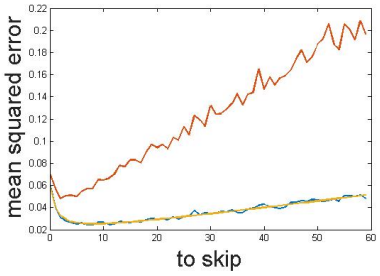
(c) Add health: school



(d) Add health: gender



(e) Project 90: Gibbs values with temperature 10



(f) Project 90: Gibbs values with temperature 100

Figure 4: Experimental results

6 Conclusion

In this work we studied the chain-referral sampling techniques. The way of sampling and the presence of homophily in the network influence the estimator error due to the increased variance in comparison to independent sampling. We proposed *subsampling technique* that allows to decrease the mean squared error of the estimator by reducing the correlation between samples. The key-factor of successful sampling is to find the optimal subsampling step.

We managed to quantify exactly the mean squared error of SA estimator for different steps of subsampling. Theoretical results were then validated with the numerous experiments, and now can help to suggest the optimal step. Experiments showed that both SA and VH estimators benefit from subsampling.

A challenge that we encountered during the study is the absence of mechanism to generate network with attributes on the nodes. In the same way that random graphs can imitate the structure of the graph we developed a mechanism to assign values to the nodes that imitates the property of homophily in the network. Created mechanism allows one to control the homophily level in the network by tuning a temperature parameter. This model is general and can also be applied in other tests.

Acknowledgements. This work was supported by CEFIPRA grant no. 5100-IT1 "Monte Carlo and Learning Schemes for Network Analytics," Inria Nokia Bell Labs ADR "Network Science," and Inria Brazilian-French research team Thanes.

References

[1] Linton C. Freeman, Research Professor, Department of Sociology and Institute for Mathematical Behavioral Sciences School of Social Sciences, University of California, Irvine. `http://moreno.ss.uci.edu/data.html`. Accessed: 2015-07-01.

[2] The National Longitudinal Study of Adolescent to Adult Health. `http://www.cpc.unc.edu/projects/addhealth`. Accessed: 2015-07-01.

[3] The Office of Population Research at Princeton University. `https://opr.princeton.edu/archive/p90/`. Accessed: 2015-07-01.

[4] Stanford Large Network Dataset Collection. `https://snap.stanford.edu/data/`. Accessed: 2015-07-01.

[5] Pierre Brémaud. *Markov chains: Gibbs fields, Monte Carlo simulation, and queues*, volume 31. Springer Science & Business Media, 2013.

[6] Nicholas A Christakis and James H Fowler. The spread of obesity in a large social network over 32 years. *New England journal of medicine*, 357(4):370–379, 2007.

[7] Krista J Gile and Mark S Handcock. Respondent-driven sampling: An assessment of current methodology. *Sociological methodology*, 40(1):285–327, 2010.

[8] Sharad Goel and Matthew J Salganik. Assessing respondent-driven sampling. *Proceedings of the National Academy of Sciences*, 107(15):6743–6747, 2010.

[9] Douglas D Heckathorn and Joan Jeffri. Jazz networks: Using respondent-driven sampling to study stratification in two jazz musician communities. In *Unpublished paper presented at American Sociological Association Annual Meeting*, 2003.

[10] Kwon Chan Jeon and Patricia Goodson. US adolescents' friendship networks and health risk behaviors: a systematic review of studies using social network analysis and Add Health data. *PeerJ*, 3:e1052, 2015.

[11] Helgar Musyoki, Timothy A Kellogg, Scott Geibel, Nicholas Muraguri, Jerry Okal, Waimar Tun, H Fisher Raymond, Sufia Dadabhai, Meredith Sheehy, and Andrea A Kim. Prevalence of HIV, sexually transmitted infections, and risk behaviours among female sex workers in Nairobi, Kenya: Results of a respondent driven sampling study. *AIDS and Behavior*, 19(1):46–58, 2015.

[12] Jesus Ramirez-Valles, Douglas D Heckathorn, Raquel Vázquez, Rafael M Diaz, and Richard T Campbell. From networks to populations: the development and application of respondent-driven sampling among IDUs and Latino gay men. *AIDS and Behavior*, 9(4):387–402, 2005.

[13] Erik Volz and Douglas D Heckathorn. Probability based estimation theory for respondent driven sampling. *Journal of official statistics*, 24(1):79, 2008.